

# **3D QSAR IN DRUG DESIGN**

## **LIGAND-PROTEIN INTERACTIONS AND MOLECULAR SIMILARITY VOLUME 2**

**Edited by**

**Hugo Kubinyi, Gerd Folkers and Yvonne C. Martin**



**KLUWER ACADEMIC PUBLISHERS**

# **3D QSAR in Drug Design**

**Ligand–Protein Interactions and Molecular  
Similarity**

QSAR =Three-Dimensional Quantitative Structure Activity Relationships

---

VOLUME 2

---

*The titles published in this series are listed at the end of this volume.*

# **3D QSAR in Drug Design Volume 2**

## **Ligand–Protein Interactions and Molecular Similarity**

Edited by

**Hugo Kubinyi**

ZHF/G, A30, BASF AG, D-67056 Ludwigshafen, Germany

**Gerd Folkers**

ETH-Zürich, Department Pharmazie, Winterthurer Strasse 190, CH-8057 Zürich, Switzerland

**Yvonne C. Martin**

Abbott Laboratories, Pharmaceutical Products Division, 100 Abbott Park Rd., Abbott Park, IL 60064-3500, USA

**KLUWER ACADEMIC PUBLISHERS  
New York / Boston / Dordrecht / London / Moscow**

eBook ISBN: 0-306-46857-3  
Print ISBN: 0-792-34790-0

©2002 Kluwer Academic Publishers  
New York, Boston, Dordrecht, London, Moscow

Print ©1998 Kluwer Academic Publishers  
London

All rights reserved

No part of this eBook may be reproduced or transmitted in any form or by any means, electronic, mechanical, recording, or otherwise, without written consent from the Publisher

Created in the United States of America

Visit Kluwer Online at: <http://kluweronline.com>  
and Kluwer's eBookstore at: <http://ebooks.kluweronline.com>

relationships in cases where the biological targets, or at least their 3D structures, are still unknown.

This project would not have been realized without the ongoing enthusiasm of Mrs. Elizabeth Schram, founder and former owner of ESCOM Science Publishers, who initiated and strongly supported the idea of publishing further volumes on 3D QSAR in Drug Design. Special thanks belong also to Professor Robert Pearlman, University of Texas, Austin, Texas, who was involved in the first planning and gave additional support and input. Although during the preparation of the chapters Kluwer Academic Publishers acquired ESCOM, the project continued without any break or delay in the work. Thus, the Editors would also like to thank the new publisher, especially Ms. Maaïke Oosting and Dr. John Martin, for their interest and open-mindedness, which helped to finish this project in time.

Lastly, the Editors are grateful to all the authors. They made it possible for these volumes to be published only 16 months after the very first author was contacted. It is the authors' diligence that has made these volumes as complete and timely as was Volume 1 on its publication in 1993.

**Hugo Kubinyi**, BASF AG, Ludwigshafen, Germany

*October 1997*

**Gerd Folkers**, ETH Zürich, Switzerland

**Yvonne C. Martin**, Abbott Laboratories, Abbott Park, IL, USA

# Contents

<b>Preface</b>	vii
 <b>Part I. Ligand–Protein Interactions</b>	
<b>Progress in Force-Field Calculations of Molecular Interaction Fields and Intermolecular Interactions</b>	3
<i>Tommy Liljefors</i>	
<b>Comparative Binding Energy Analysis</b>	19
<i>Rebecca C. Wade, Angel R. Ortiz and Federico Gigo</i>	
<b>Receptor-Based Prediction of Binding Affinities</b>	35
<i>Tudor I. Oprea and Girland R. Marshall</i>	
<b>A Priori Prediction of Ligand Affinity by Energy Minimization</b>	63
<i>M. Katharine Holloway</i>	
<b>Rapid Estimation of Relative Binding Affinities of Enzyme Inhibitors</b>	85
<i>M. Rami Reddy, Velarkad N. Viswanadhan and M. D. Erion</i>	
<b>Binding Affinities and Non-Bonded Interaction Energies</b>	99
<i>Ronald M.A. Knegtel and Peter D.J. Grootenhuis</i>	
<b>Molecular Mechanics Calculations on Protein-Ligand Complexes</b>	115
<i>Irene T. Weber and Robert W. Harrison</i>	
 <b>Part II. Quantum Mechanical Models and Molecular Dynamics Simulations</b>	
<b>Some Biological Applications of Semiempirical MO Theory</b>	131
<i>Bernd Beck and Timothy Clark</i>	
<b>Density-Functional Theory and Molecular Dynamics: A New Perspective for Simulations of Biological Systems</b>	161
<i>Wanda Andreoni</i>	
<b>Density-functional Theory Investigations of Enzyme-substrate Interactions</b>	169
<i>Paolo Carloni and Frank Alber</i>	

## Preface

Significant progress has been made in the study of three-dimensional quantitative structure–activity relationships (3D QSAR) since the first publication by Richard Cramer in 1988 and the first volume in the series. *3D QSAR in Drug Design. Theory, Methods and Applications*, published in 1993. The aim of that early book was to contribute to the understanding and the further application of CoMFA and related approaches and to facilitate the appropriate use of these methods.

Since then, hundreds of papers have appeared using the quickly developing techniques of both 3D QSAR and computational sciences to study a broad variety of biological problems. Again the editor(s) felt that the time had come to solicit reviews on published and new viewpoints to document the state of the art of 3D QSAR in its broadest definition and to provide visions of where new techniques will emerge or new applications may be found. The intention is not only to highlight new ideas but also to show the shortcomings, inaccuracies, and abuses of the methods. We hope this book will enable others to separate trivial from visionary approaches and me-too methodology from innovative techniques. These concerns guided our choice of contributors. To our delight, our call for papers elicited a great many manuscripts. These articles are collected in two bound volumes, which are each published simultaneously in two related series: they form Volumes 2 and 3 of the *3D QSAR in Drug Design* series which correspond to volumes 9–11 and 12–14, respectively, in *Perspectives in Drug Discovery and Design*. Indeed, the field is growing so rapidly that we solicited additional chapters even as the early chapters were being finished. Ultimately it will be the scientific community who will decide if the collective biases of the editors have furthered development in the field.

The challenge of the quantitative prediction of the biological potency of a new molecule has not yet been met. However, in the four years since the publication of the first volume, there have been major advances in our understanding of ligand–receptor interactions, molecular similarity, pharmacophores, and macromolecular structures. Although currently we are well prepared computationally to describe ligand–receptor interactions, the thorny problem lies in the complex physical chemistry of intermolecular interactions. Structural biologists, whether experimental or theoretical in approach, continue to struggle with the field's limited quantitative understanding of the enthalpic and entropic contributions to the overall free energy of binding of a ligand to a protein. With very few exceptions, we do not have experimental data on the thermodynamics of intermolecular interactions. The recent explosion of 3D protein structures helps us to refine our understanding of the geometry of ligand–protein complexes. However, as traditionally practiced, both crystallographic and NMR methods yield static pictures and relatively coarse results considering that an attraction between two non-bonded atoms may change to repulsion within a tenth of an Ångström. This is well below the typical accuracy of either method. Additionally, neither provides information about the energetics of the transfer of the ligand from solvent to the binding site.



With these challenges in mind, one aim of these volumes is to provide an overview of the current state of the quantitative description of ligand–receptor interactions. To aid this understanding, quantum chemical methods, molecular dynamics simulations and the important aspects of molecular similarity of protein ligands are treated in detail in Volume 2. In the first part ‘Ligand–Protein Interactions,’ seven chapters examine the problem from very different points of view. Rule- and group-contribution-based approaches as well as force-field methods are included. The second part ‘Quantum Chemical Models and Molecular Dynamics Simulations’ highlights the recent extensions of *ab initio* and semi-empirical quantum chemical methods to ligand–protein complexes. An additional chapter illustrates the advantages of molecular dynamics simulations for the understanding of such complexes. The third part ‘Pharmacophore Modelling and Molecular Similarity’ discusses bioisosterism, pharmacophores and molecular similarity, as related to both medicinal and computational chemistry. These chapters present new techniques, software tools and parameters for the quantitative description of molecular similarity.

Volume 3 describes recent advances in Comparative Molecular Field Analysis and related methods. In the first part ‘3D QSAR Methodology, CoMFA and Related Approaches’, two overviews on the current state, scope and limitations, and recent progress in CoMFA and related techniques are given. The next four chapters describe improvements of the classical CoMFA approach as well as the CoMSIA method, an alternative to CoMFA. The last chapter of this part presents recent progress in Partial Least Squares (PLS) analysis. The part ‘Receptor Models and Other 3D QSAR Approaches’ describes 3D QSAR methods that are not directly related to CoMFA, i.e., Receptor Surface Models, Pseudo-receptor Modelling and Genetically Evolved Receptor Models. The last two chapters describe alignment-free 3D QSAR methods. The part ‘3D QSAR Applications’ completes Volume 3. It gives a comprehensive overview of recent applications but also of some problems in CoMFA studies. The first chapter should give a warning to all computational chemists. Its conclusion is that all investigations on the classic corticosteroid-binding globulin dataset suffer from serious errors in the chemical structures of several steroids, in the affinity data and/or in their results. Different authors made different mistakes and sometimes the structures used in the investigations are different from the published structures. Accordingly it is not possible to make any exact comparison of the reported results! The next three chapters should be of great value to both 3D QSAR practitioners and to medicinal chemists, as they provide overviews on CoMFA applications in different fields, together with a detailed evaluation of many important CoMFA publications. Two chapters by Ki Kim and his comprehensive list of 1993–1997 CoMFA papers are a highly valuable source of information.

These volumes are written not only for QSAR and modelling scientists. Because of their broad coverage of ligand binding, molecular similarity, and pharmacophore and receptor modelling, they will help synthetic chemists to design and optimize new leads, especially to a protein whose 3D structure is known. Medicinal chemists as well as agricultural chemists, toxicologists and environmental scientists will benefit from the description of so many different approaches that are suited to correlating structure–activity

**This Page Intentionally Left Blank**

**Part I**

**Ligand–Protein  
Interactions**

**This Page Intentionally Left Blank**

# Progress in Force-Field Calculations of Molecular Interaction Fields and Intermolecular Interactions

Tommy Liljefors

Department of Medicinal Chemistry, Royal Danish School of Pharmacy, Copenhagen, Denmark

## 1. Introduction

In the force-field (molecular mechanics) method [1], a molecule is considered as a collection of atoms held together by classical forces. These forces are described by parameterized potential energy functions of structural features like bond lengths, bond angles, torsional angles, etc. The energy of the molecule is calculated as a sum of terms (Eq. 1).

$$E = E_{\text{stretching}} + E_{\text{bending}} + E_{\text{torsion}} + E_{\text{van der Waals}} + E_{\text{electronic}} + E_{\text{hydrogen band}} + \text{other terms} \quad (1)$$

The first four terms in Eq. 1 describe the energies due to deviations of structural features and non-bonded distances from their 'ideal' or reference values,  $E_{\text{electrostatic}}$  is the energy contribution due to attraction or repulsion between charges (or dipoles) and some force-fields use a special hydrogen-bonding term,  $E_{\text{hydrogen Bond}}$ . Additional terms are required — e.g. for calculations of vibrational frequencies and thermodynamic quantities.

Force-field calculations [1–3] are used in many research areas aiming at an understanding, modelling and subsequent exploiting of structure–activity/property relationships. Such areas include conformational analysis, pharmacophore identification, ligand docking to macromolecules, *de novo* ligand design, comparative molecular field analysis (CoMFA) and identification of favorable binding sites from molecular interaction fields. Although *ab initio* quantum chemical computational methodology [4] today is competitive with experiments in determining a large number of molecular properties, force-fields are commonly used due to the prohibitive amounts of computer time required for high-level *ab initio* calculations on series of drug-sized molecules and on large molecular systems as those involved in ligand–protein interactions. For calculations of, for example, molecular structures and conformational energies force-field calculations may give results in excellent agreement with experiments, provided that the force-field parameters involved in the calculations have been accurately determined [1,5].

The force-fields used for calculations of molecular interaction fields and intermolecular interactions vary from very simple force-fields as those commonly used in CoMFA 3D QSAR studies [6] to the more sophisticated force-field used for the calculation of molecular interaction fields by the GRID method [3,7–11], and the complex force-fields required for calculations of complexation energies and geometries of general intermolecular interactions between complete molecules [2].

The aim of this chapter is to review and discuss some recent developments and evaluations of force-fields used for the calculations of molecular interaction fields and intermolecular energies and geometries. The review is not meant to be exhaustive, but some

recent studies have been selected to illustrate important current directions in the development of force-fields and in their use in connection with 3D QSAR studies and calculations of molecular interaction fields and intermolecular interactions. Thus, relationships between the quality of the force-field and the results of a CoMFA 3D QSAR study will be discussed in the light of some recent investigations. New developments in the computations of molecular interaction fields by the GRID method will be described, and recent developments in the calculations of intermolecular energies and geometries for the biologically important cation- $\pi$  and  $\pi$ - $\pi$  systems will be reviewed.

## 2. CoMFA QSAR: Is there a Force-Field Dependence?

The force-field commonly used in a CoMFA QSAR study is very simple and includes only two terms, a Lennard-Jones 6-12 potential for the van der Waals (vdW) interactions and a Coulomb term for the electrostatic interactions with the probe. Considering the large number of successful CoMFA QSAR studies which have been reported [12], these two terms seem to be sufficient in most cases. It may be expected that a hydrophobic field and/or an explicit hydrogen-bond potential may, in some cases, be advantageous. However, more experience with the inclusion of such fields is necessary before a firm conclusion may be drawn.

No systematic comparison of different force-fields in connection with CoMFA studies has been undertaken. However, some interesting case studies have recently been reported. Folkers et al. have reviewed the results of CoMFA QSAR studies employing the GRID force-field and the standard (Tripos) CoMFA force-field [13]. The two force-fields are significantly different. For instance, the atomic charges employed in the calculation of the electrostatic interactions are significantly different. The GRID force-field also includes a sophisticated hydrogen-bonding potential [8–10]. Folkers et al. concluded that the GRID force-field generally gives better results in terms of  $Q^2$  and standard error of prediction than the standard CoMFA force-field for an uncharged methyl probe in cases where only the steric field contributes to the correlation. When steric and electrostatic fields contribute more equally to the correlation, the force-fields tested give very much the same results.

Two recent studies, discussed below, further illuminate the question of the force-field dependence in CoMFA QSAR studies.

### 2.1. *The steric field: Is an explicit vdW potential required?*

Instead of using the standard Lennard-Jones 6-12 potential, Kroemer and Hecht [14], mapped all atoms in the target molecules directly onto a predefined grid and for each atom checked if the center of the atom is inside or outside cubes defined by the grid. Depending on the outcome of this check, different values (0.0 or 30.0) were assigned to the grid point at the corresponding lattice intersection. These atom-based indicator variables correspond to what is obtained by using a simple hard-sphere approximation (with energy cutoff) for the calculations of vdW interactions. The indicator variables were used as the 'steric field' in a CoMFA QSAR study on five sets of dihydrofolate reduc-

tase inhibitors and the results were compared to those obtained by using the standard CoMFA method. A similar approach has previously been reported by Floersheim et al. [15]. However, in this study, the vdW surfaces of the molecules were used for the computations of a 'shape potential' and values of 0 or 1 were assigned to grid points depending on whether the points were located within or outside the surface.

In terms of  $Q^2$  and predictive  $r^2$ , the study by Kroemer and Hecht shows that the use of atom-based indicator variables gives results at least as good as those obtained by using the 6-12 Lennard-Jones potential in the standard CoMFA method. Thus, in the context of a CoMFA QSAR analysis, a step function as the one employed by Kroemer and Hecht, gives a description of the variation in the shapes of the molecules in the dataset of similar quality (or usefulness) as the Lennard-Jones 6-12 potential. An interesting point is that, in contrast to standard CoMFA QSAR, the use of a finer grid ( $< 2 \text{ \AA}$ ) in conjunction with the indicator variables improved the results significantly. This study indicates that in general, there seems to be little to gain from fine-tuning the Lennard-Jones 6-12 potential for CoMFA QSAR studies or introducing more accurate functions as, for instance, exponential functions [1] for the calculation of the vdW field. However, considering the results obtained by Folkers et al. for the GRID versus Tripos CoMFA force-fields described above, an exception to this may be cases for which the van der Waals field is the strongly dominating contributor to the correlation.

## *2.2. The electrostatic field: what quality of the charge distribution is required?*

The accurate calculation of the electrostatic contribution is clearly the most problematic part in the calculation of intermolecular energies and geometries by force-field methods [16]. The results of such calculations, in general, very much depend on the quality of the charge distribution employed and also on a proper balance between the electrostatic term and the rest of the force-field. Is there a similar strong dependence of the charge distribution used on the results of a CoMFA QSAR study?

In a CoMFA analysis of 49 substituted benzoic acids, Kim and Martin found that AMI partial charges performed better than STO-3G ESPFIT charges in reproducing Hammett  $\sigma$  constants [17]. The results were also found to be superior to those obtained by using regression analysis of the charges. Folkers et al. [13] studied the influence of the charge of the probe and the charge distribution of the target molecules for a set of 24 N<sup>2</sup>-phenylguanines. The probes used were a sp<sup>3</sup> carbon probe with charge +1, and oxygen probes with charges -0.85, -0.5 and -0.2, respectively. Three sets of atom-centered partial charges were used for the phenylguanines: (i) Gasteiger-Marsili charges [18] (default Tripos CoMFA charges); (ii) Mulliken charges calculated by semi-empirical quantum chemical methodology (the Hamiltonian was unfortunately not reported); and (iii) charges obtained by linear least-squares fitting of atom-centered point charges to the electrostatic potential calculated from the semiempirical wave function (ESP-charges [19]). The vdW field was calculated by the standard Lennard-Jones 6-12 function. The use of the different probes gave significantly different  $Q^2$  values and contributions of the fields, but the different atomic charge schemes resulted in similar statistical parameters.

Recently, Kroemer et al. [20] have extended this type of study to include a large variety of different methods to calculate the electrostatic field of 37 ligands at the benzodiazepine inverse agonist site. A total number of 17 different charge schemes was evaluated, including empirical charges (Gasteiger-Marsili), charges obtained from semiempirical quantum chemical methods (MNDO, AM1, PM3) and from *ab initio* quantum chemical methods at the Hartree-Fock level, employing three different basis sets (HF/STO-3G, HF/3-21G\*, HF/6-31G\*). The semiempirical and *ab initio* atom-centered partial charges were calculated by a Mulliken population analysis, as well as by linear least-squares fitting of atom-centered point charges to reproduce the calculated electrostatic potential (ESP-charges). In addition, the molecular electrostatic potentials calculated by using the HF/STO-3G, HF/3-21G\* and HF/6-31G\* basis sets were directly mapped onto the CoMFA grid. In all cases, a  $sp^3$  carbon probe with a charge of +1 was used. The vdW field was calculated in the standard way by a Lennard-Jones 6-12 potential and the results were compared to standard CoMFA QSAR.

All 17 charge schemes resulted in good models in terms of  $Q^2$  (0.61–0.77) and standard error of prediction (0.76–0.94). Although the different charge distribution schemes were obtained at very different levels of theory, in the context of CoMFA QSAR studies and the resulting statistical parameters there is hardly any significant difference. The electrostatic fields calculated by the various methods were, in many cases, shown to be strongly correlated. However, even with a low correlation between fields — e.g. between the semiempirically calculated Mulliken charges and the directly mapped electrostatic potentials employing the STO-3G basis set ( $r^2 = 0.62–0.66$ ) — the results in terms of  $Q^2$  are very similar (0.61–0.72).

In contrast to the similarity of the statistical parameters in the study discussed above, the contour maps obtained by using different charge distribution schemes may differ significantly. This has consequences for the use of contour maps in terms of a physicochemical interpretation of intermolecular interactions and for the use of such maps in the design of new compounds. Different charge distribution schemes may give contour maps which lead the design process in different directions. The charge scheme which is the ‘best’ one in this respect cannot unambiguously be selected on the basis of the statistical parameters obtained.

### 3. Recent Developments of the GRID Method for the Calculation of Molecular Interaction Fields

The GRID method developed by Goodford [7–11] is designed for calculation of interactions between a probe (a small molecule or molecular fragment) and a macromolecular system of known structure in order to find energetically favorable sites for the probe. A large number of probes including multi-atom probes are provided. The GRID method is very carefully parameterized by fitting experimental data for proteins and small molecule crystals. In addition to its primary use to find favorable probe sites in macromolecules, the interaction fields calculated by the GRID method have also been used extensively in 3D QSAR studies as a replacement of the electrostatic and steric fields of standard Tripos CoMFA. Wade has reviewed the GRID method and its use for ligand



design, ligand docking and 3D QSAR [21]. Recently, Goodford has reviewed the GRID force-field and its use in multivariate characterization of molecules for QSAR analysis [11].

In calculations of molecular interactions fields, a static target is generally used. With some exceptions (see below), this has also been the case for calculations using the GRID method. In the most recent version of GRID (version 15) [22], new features taking target flexibility into account have been introduced. These important new features are discussed below. In addition, a new hydrophobic probe developed by Goodford and included in the GRID method is described. It should be noted that these additions to the GRID method are all very new and no publications in which these features have been used have yet appeared.

### 3.1. Identification of energetically favorable probe sites in hydrogen-bond interactions

Hydrogen bonding is extremely important in ligand-protein interactions and therefore Goodford and co-workers have spent much effort in developing a sophisticated and carefully parameterized methodology for the calculation of hydrogen-bonding interactions [8–10]. GRID has always taken torsions about the C–O bond in aliphatic alcohols into account. Thus, in the interactions between an aliphatic alcohol and a probe which can accept or donate a hydrogen-bond, the hydrogen atom and the lone pairs of the hydroxyl group are allowed to move without any energy penalty in order to find the most favorable binding energy between the probe and the target. For instance, for the interaction between methanol and a water probe, virtually identical interaction energies are calculated for a staggered and eclipsed probe position with respect to the methyl hydrogens in methanol (Fig. 1).

However, the energy difference between eclipsed and staggered methanol is calculated to be 1.4 kcal/mol by *ab initio* HF/6-31G\* calculations [23]. If an eclipsed probe position results in an eclipsed conformation of methanol, there should be a significant energy penalty for the eclipsed arrangement shown in Fig. 1, relative to the staggered one. To investigate this, *ab initio* calculations (HF-6-31G\*) were undertaken for the two hydrogen-bonded complexes by locking the O–O–C–H dihedral angle (indicated by asterisks in Fig. 1) to 180 (staggered) and 0 degrees (eclipsed), respectively, but optimizing all other degrees of freedom, including the position of the hydrogen

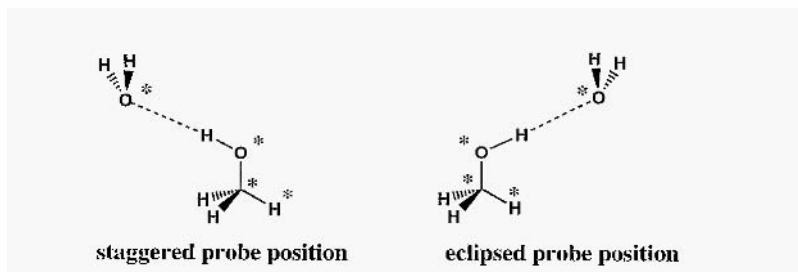


Fig. 1. Staggered and eclipsed positions of H<sub>2</sub>O in its hydrogen-bond interaction with methanol as the hydrogen-bond donor.

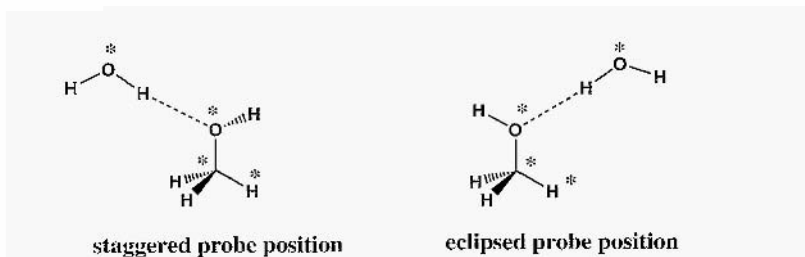


Fig. 2. Staggered and eclipsed positions of  $\text{H}_2\text{O}$  in its hydrogen-bond interaction with methanol as the hydrogen-bond acceptor. Note that methanol prefers to be in a staggered conformation in both complexes. The asterisks mark the atoms of the dihedral angle locked in the calculations.

atom involved in the hydrogen-bond [23]. The energy difference between the eclipsed and staggered arrangements in Fig. 1 was calculated to be 1.1 kcal/mol, only slightly lower than that for the corresponding conformations in methanol itself. Due to the strong directionality of the hydrogen-bond in this case, an eclipsed position of the water oxygen also leads to an eclipsed conformation of the methanol part of the complex. Thus, it can be concluded that there is an energy penalty of approximately 1 kcal/mol for the hydrogen-bonded eclipsed complex in Fig. 1 relative to the staggered one. If the hydrogen atom marked by an asterisk in the methanol part is replaced by a methyl group, this energy penalty increases to about 2 kcal/mol. The results are very similar for the interaction between a carbonyl group and an aliphatic alcohol as hydrogen-bond donor.

When methanol is acting as a hydrogen-bond acceptor, the situation is different. The two hydrogen-bonding arrangements shown in Fig. 2 are calculated to have very similar energies [23]. The reason for this is that due to the delocalized character of the lone pairs on the methanol oxygen atom, methanol can be staggered in both hydrogen-

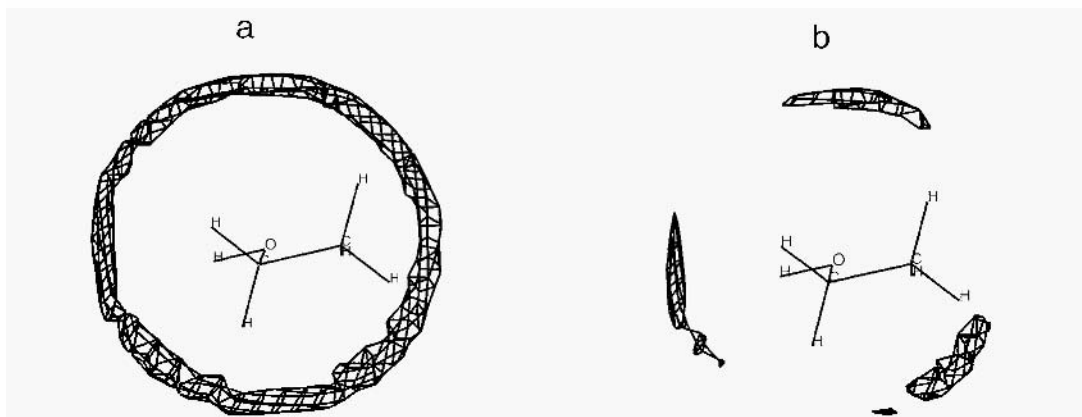


Fig. 3. GRID maps for ethonol interacting with a carbonyl oxygen probe. The left map (a) shows favorable interaction sites calculated by GRID version 14 or earlier, whereas (b) shows the corresponding map calculated by GRID version 15. The hydroxy hydrogen atom is shown in one of its rotameric slates.

bonded complexes. Thus, the energy increasing eclipsing in the methanol part of the complex is avoided.

GRID version 15 includes a new energy function which takes the findings discussed above into account. Probes which can both accept and donate a hydrogen-bond (e.g. H<sub>2</sub>O) may in the new version turn around compared to earlier versions of GRID and give rise to a new interaction geometry. For interactions with probes which cannot turn around (e.g. the carbonyl probe), the hydrogen-bond energy may be significantly diminished in unfavorable probe positions and thus give significant changes to the GRID contour map compared to earlier GRID versions. This is illustrated in Fig. 3 which displays GRID maps for the interaction between ethanol and a carbonyl oxygen probe calculated by GRID with version number  $\leq 14$  (Fig. 3a) and GRID version 15 (Fig. 3b).

Recently, Mills and Dean have reported the results of an extensive investigation of hydrogen-bond interactions in structures in the Cambridge Structural Database [24]. This study provides 3D distributions of complementary atoms about hydrogen-bonding groups. Scatterplots and cumulative distribution functions clearly demonstrate that hydrogen-bond accepting groups interacting with a hydroxyl group prefer a staggered position, as in the left structure in Fig. 1, while hydrogen-bond donating groups are much less localized. This is in nice agreement with the findings discussed above.

### 3.2. Target flexibility

An important limitation in calculations of molecular interaction fields is that flexibility of the target is not taken into account. In 3D QSAR studies, ‘side-chain’ conformations in the target molecules are often more or less arbitrarily assigned when the bioactive conformations are unknown and this may give misleading results. In ligand-protein interactions, amino acid side chains may adopt different conformations in order to better accommodate or better interact with a ligand. Of course, the analysis (GRID or CoMFA) may be repeated for several different conformations, but this is prohibitively time-consuming, especially in the case of side-chain conformations in proteins.

As discussed above, the GRID method has always taken some flexibility of the target into account. However, side chains in the target have remained fixed in their input conformations. A very interesting new development of the GRID method is the implementation of algorithms which take conformational flexibility of amino acid side chains into account, allowing them to be attracted or repelled by the probe as the probe is moving around [22]. The new algorithm is primarily intended for proteins but it may, in some cases, also be used for ‘side-chains’ in non-protein molecules. The amino acids currently supported are arginine, aspartate, asparagine, glutamate, glutamine, isoleucine, leucine, lysine, methionine, serine, threonine and valine.

The algorithm works by dividing the target molecule into an *inflexible core* and a *flexible side chain* on an atom basis. This is automatically done by the program allowing for differences in the local environment of the side chain. However, the user may override the default by forcing atoms into the core or out of the core into the flexible side-chain part.

So far, there is only limited experience with the ‘flexible side-chain’ option, but it is clear that this new feature in the GRID method is a very important step forward in the calculations of molecular interaction fields. This feature should significantly improve the description of energetically favorable probe locations and should be very useful in, for instance, ligand design and ligand docking.

### 3.3. The hydrophobic probe

To the library of GRID probes a hydrophobic probe has recently been added (probe name DRY [22]). The hydrophobic probe is designed to find locations near the target surface where the target molecule may favorably interact with another molecule in an aqueous environment. The energy expression for the hydrophobic probe is shown in Eq. 2.

$$E = E_{ENTROPY} + E_{LJ} - E_{HB} \quad (2)$$

The basic assumptions behind the construction of the probe is that the water ordering responsible for the entropic contribution to the hydrophobic effect is due to hydrogen-bonds between water molecules at nonpolar (undisturbed) target surfaces. On binding of a hydrophobic molecule, the ordered water molecules are displaced and transferred into less ordered (higher entropy) bulk water. This is an energetically favorable process ( $E_{ENTROPY}$ ). Dispersion interactions between the two hydrophobic molecules adds to the favorable energy ( $E_{LJ}$ ). The ordering of water at the nonpolar surface may be disturbed by polar target atoms which form hydrogen-bonds to water molecules. This decreases the order (increases the entropy) of the water molecules at the surface and, consequently, diminishes the hydrophobic effect ( $E_{HB}$ ). In addition, there are breaking of hydrogen-bonds which is enthalpically unfavorable.  $E_{ENTROPY}$  is calculated from the assumption that the ordered water molecules at a nonpolar surface will form on the average three out of the theoretically four possible hydrogen-bonds per oxygen atom. This gives four permutations for three out of four possible hydrogen-bonds and  $E_{ENTROPY}$  may thus be calculated by Eq. 3.

$$E_{ENTROPY} = -RT \ln 4 \quad (3)$$

$R$  is the gas constant ( $1.987 \times 10^{-3}$  kcal/mol/K) and  $T = 308$  K. This entropic contribution is assumed to be constant at an undisturbed surface.

Dispersion interactions ( $E_{LJ}$ ) are calculated by using the Lennard-Jones function and a water probe.  $E_{HB}$  which measures the hydrogen-bond interactions between water molecules and polar functional groups of the target is calculated by using the hydrogen-bond function of the GRID force-field.

The hydrophobic probe, in general, gives wide and shallow minima. This implies that the variance of the hydrophobic energies is small. Therefore, PLS methods which cluster grid points into chemically meaningful regions [25] should be employed if the hydrophobic fields are to be used as input to CoMFA/PLS. Scaling of the fields should not be done. As the energies obtained by using different GRID probes are already scaled, further scaling of GRID fields is inappropriate [26].

#### 4. Force-Field Calculations of Cation- $\pi$ and $\pi$ - $\pi$ Complexes

Aromatic ring systems are of immense importance in drugs. Bemis and Murcko analyzed 5210 known drugs in the Comprehensive Medicinal Chemistry database [27]. Among the 41 most common frameworks, 29 contain aromatic rings and benzene (phenyl) was found to be the most common one.

The benzene ring system, the prototypical aromatic system, has very unique properties. It does not have a permanent dipole moment and is, in that sense, a nonpolar molecule. However, it has a strong quadrupole moment. The electrostatic potential of benzene leads to strong attraction of a cation to the  $\pi$ -face of the ring system (cation- $\pi$  interactions: Fig. 4). Thus, benzene and related aromatic systems may be considered as ‘hydrophobic anions’ [28].

Amino acids containing aromatic side-chains as phenylalanine, tyrosine and tryptophan play an important role in protein structure and function. Thus, the understanding of intermolecular interactions involving aromatic systems and the ability to model such interactions are of crucial importance for computational studies of ligand-receptor complexes. These types of interactions have received considerable attention during recent years. Significant progress in the understanding of the nature of cation- $\pi$  interactions and  $\pi$ - $\pi$  interactions and the computational problems involved in the force-field calculations of such interactions have recently been made.

##### 4.1. Cation- $\pi$ interactions

The binding of cations to the  $\pi$ -face of benzene involves large binding energies. For instance, the binding enthalpy of  $\text{Li}^+$ -benzene in the gas phase is 38 kcal/mol and the corresponding binding enthalpy of  $\text{NH}_4^+$ -benzene is 19 kcal/mol [28]. This strong interaction with cations in fact makes the benzene ring competitive with a water molecule in binding to cations. The  $\text{K}^+$ - $\text{H}_2\text{O}$  and the  $\text{K}^+$ -benzene complexes have binding enthalpies of 18 and 19 kcal/mol, respectively [28]. It has been demonstrated that cyclophane hosts made up of aromatic rings are able to strongly bind cations including quaternary ammonium ions in aqueous solution [29].

For a long time, it was generally believed that the quaternary positively charged ammonium group of acetylcholine was interacting with an anionic site in acetylcholine esterase. However, it is now clear that acetylcholine in its binding to its esterase interacts *via* cation- $\pi$  interactions with aromatic ring systems, in particular to tryptophan

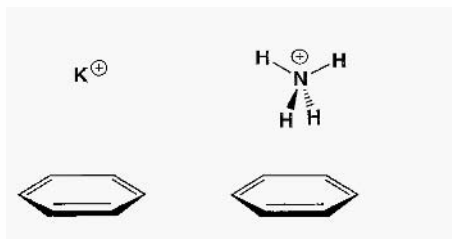


Fig. 4. Cation- $\pi$  interactions

(W84) [30]. The importance of cation- $\pi$  interactions in chemistry and biology has recently been reviewed by Dougherty [28].

Standard high-quality force-field methods seriously underestimate cation- $\pi$  interactions. For example, the interaction enthalpy of  $\text{NH}_4^+$ -benzene is underestimated by as much as 5.8 kcal/mol even if high-quality atom-centered electrostatic potential derived charges are used, charges which accurately reproduce the quadrupole moment of benzene [31].

Recently, high-level *ab initio* calculations have provided insight into the nature of cation- $\pi$  interactions. On the basis of MP2/6-311+G\*\* calculations, which give very good agreement with experiments (binding enthalpies as well as free energies) for the ammonium ion-benzene complex. Kim et al. [32] conclude that, in addition to charge-quadrupole interactions, correlation effects (dispersion energies) and polarization of the benzene electron distribution by the cation give very important contributions to the complexation energy. The failure of standard molecular mechanics force-fields to handle cation- $\pi$  interactions can, to a large part, be attributed to the fact that polarization effects are not taken into account.

In general, molecular mechanics force-fields include only two-body additive potential functions. Non-additive effects as polarization have only recently been included. Caldwell and Kollman have developed a molecular mechanics force field which explicitly includes polarization and also includes non-additive exchange-repulsion [33]. This force-field excellently reproduces the complexation enthalpy of alkali cations-benzene and ammonium ion-benzene. Thus, as has previously been shown in other cases, for instance in the calculations of hydration free energies of ions [34], the inclusion of non-additive effects such as polarization is necessary for force-field calculations of intermolecular interactions involving ions.

In a recent study, Mecozzi et al. [35] show that the variation in cation ( $\text{Na}^+$ ) binding abilities to a series of aromatic systems surprisingly well correlates to the electrostatic potential at the position of the cation in the complex. Thus, virtually all *variation* in binding energy is reflected in the electrostatic term.

Cation- $\pi$  interactions are not only limited to full cations, as ammonium ions and alkali cations but also polar molecules as  $\text{H}_2\text{O}$ ,  $\text{NH}_3$  and other molecules with partial positive charges display this type of interaction with the  $\pi$ -face of benzene, albeit with weaker interaction energies [36,37]. This makes the cation- $\pi$  type of interactions of great importance for the understanding of ligand-protein interactions.

The strong attractive interactions between cations and the  $\pi$ -face of benzene and related aromatic ring systems have been used as an argument for a proposed stabilization of the putative ion-pair interaction between the ammonium group of aminergic neurotransmitters and an aspartate side-chain in their receptors. In models of the binding site of these receptors, the aspartate residue is surrounded by highly conserved aromatic residues [38]. However, it has recently been shown on the basis of *ab initio* calculations that the stabilization of an ion pair by benzene is very much smaller than the stabilization of an isolated cation [39]. In fact, the stabilization provided is not sufficient to prevent hydrogen transfer from the ammonium ion to the carboxylate ion giving the intrinsically more stable amine-carboxylic acid complex. However, a pro-

perly located water molecule in conjunction with a dielectric continuum may provide the required stabilization [40].

The progress made in the understanding of cation- $\pi$  interactions during recent years has provided developers of molecular interaction fields and force-fields for calculations of inermolecular interactions with much valuable insight.

#### 4.2. $\pi$ - $\pi$ interactions

The benzene-benzene interaction, the prototypical  $\pi$ - $\pi$  interaction, is of great importance due to its role in the stability of proteins and in ligand-protein binding. Although it was pointed out more than 20 years ago that benzene crystal data could not be fitted without introducing electrostatics [41], in the context of force-field calculations benzene continued for a long time to be considered as a nonpolar molecule interacting with other molecules or molecular Cragments only via non-bonded vdW interactions. However, computational problems with such a model became increasingly evident [42].

A T-shaped type of arrangement of aromatic rings (Fig. 5) is strongly preferred in protein structures [43-45]. High-level *ab initio* calculations show the T-shaped complex to be significantly more stable than the stacked one [46]. However, a 'parallel-displaced' or tilted 'parallel-displaced' structure may be slightly more stable than the T-shaped one [42,47]. Although attractive vdW non-bonded interactions (dispersion) favor the stacked structure, the electrostatic interactions (quadrupole-quadrupole) which are attractive for the T-shaped arrangement but repulsive for the stacked one determine the preference for the T-shape.

Recently, Chipot et al. [46] employed potential of mean force calculations on the benzene dimer and toluene dimer in the gas phase and in water. They find the T-shaped benzene dimer in gas phase to be lower in free energy than the stacked structure. Interestingly, in the toluene case, their simulations indicate that the stacked arrangement is slightly preferred. The results of the force-field calculations are supported by high-level *ab initio* calculations. The same difference in orientational preferences are found

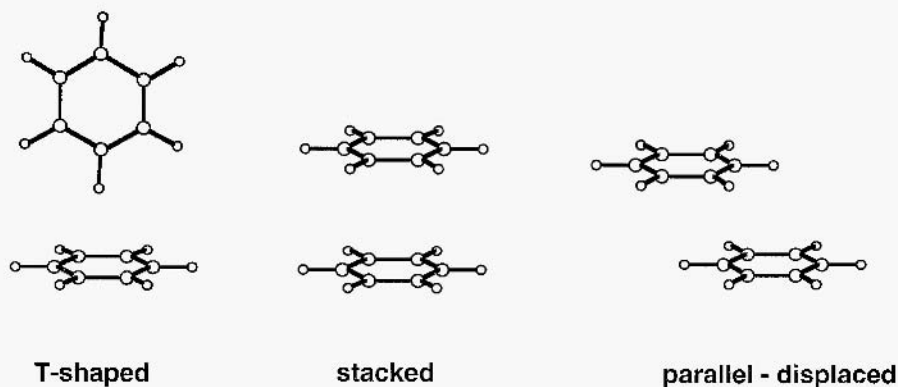


Fig. 5. Geometries of the benzene dimer.

in simulations for aqueous solution. This leads to the provocative question if the benzene dimer really is a good model for  $\pi$ - $\pi$  interactions in proteins. The authors conclude that the rarity of stacked arrangements of phenylalanine side-chains in protein structures should be explained by other factors than quadrupole-quadrupole interactions. They propose that steric and other interactions with neighboring functional groups should additionally be considered.

From a force-field point of view, a very interesting point in this study is the demonstration that atom-centered point charges obtained by least squares fitting to *ab initio* calculated (6-31G\*\*) electrostatic potentials accurately reproduce quadrupole moments and even higher-order multipole moments of benzene and toluene. Thus, such charges are sufficiently accurate to treat quantitatively the important quadrupole-quadrupole interactions in the dimers. If this also holds for other types of aromatic systems remains to be studied.

## 5. Summary and concluding remarks

Recent case studies on the force-field dependence of the results obtained by the CoMFA 3D QSAR methodology indicate that, in general, the use of a higher-quality force-field does not seem to lead to a significantly better 3D QSAR model in terms of statistical parameters ( $Q^2$  and standard error of prediction). In particular, the statistical parameters seem to be quite insensitive to the quality of the charge distribution used for the calculations of the electrostatic field. However, the contour plots derived from the analysis may show significant differences. Whether a force-field based on a higher level of theory also produces contour plots better suited for the design of new analogs remains to be studied.

Significant developments of the force-fields in Goodford's GRID method for the calculation of molecular interactions fields have recently been made. The most important new feature is that the flexibility of target side-chains may optionally be taken into account. In addition, the calculation of directional preferences in hydrogen-bonding to aliphatic alcohols has been improved and a hydrophobic probe has been included in the GRID library of probes.

Progress in the understanding of intermolecular interactions of the biologically important cation- $\pi$  and  $\pi$ - $\pi$  type has demonstrated that such interactions can be quantitatively modelled by force-field calculations. However, explicit inclusion of polarization is required in the cation- $\pi$  case. It has also been demonstrated that using atom-centered point charges obtained by least-squares fitting to *ab initio* calculated electrostatic potentials, quadrupole moments and also higher-order multipole moments of benzene and toluene can be accurately reproduced. Atom-centered point charges, can in this case, be used to quantitatively calculate dimer properties for which quadrupole-quadrupole interactions are important.

## Acknowledgements

I thank Dr Peter Goodford for valuable discussions on the GRID method. This work was supported by grants from the Danish Medical Research Council and the Lundbeck Foundation, Copenhagen.



## References

1. Burkert, U. and Allinger, N.L., *Molecular mechanics*, ACS Monograph 177. American Chemical Society. Washington D.C., 1982.
2. Siebel, G.L. and Kollman, P.A., *Molecular mechanics and the Modeling of drug structures*, In *Comprehensive medicinal chemistry*. Vol. 4, Hansch C., Sammes P.G., Taylor. J.B. and Ramsden. C.A. (Eds.), Pergamon Press, Oxford, 1990. pp. 125–138.
3. Goodford, P., *The properties of force fields*. In Sanz, F., Giraldo, J. and Manual, F. (Eds.) *QSAR and molecular modeling: Concepts, computational tools and biological applications*, Prous Science Publishers. Barcelona. 1995, pp. 199–205.
4. Hehre, W.J., Radom, L., v.R. Schleyer. P. and Pople, J.A., *Ab initio molecular orbital theory*, John Wiley & Sons. New York, 1986.
5. (a) Gundertofte. K., Liljefors, T., Norrby, P.-O. and Pettersson, I. *A comparison of conformational energies calculated by several molecular mechanics methods*. *J. Comput. Chem.*, 17 (1996) 429–449.  
(b) Pettersson, I and Liljefors, T., *Molecular Mechanics calculated conformational energies of organic molecules*, In Lipkowitz, K.B. and Boyd. D.B. (Eds.) *Reviews in Computational Chemistry*, Vol. 9. VCH Publishers, Inc., New York, 1996. pp. 167–189.
6. Cramer, III, R.D., Patterson, D.E. and Bunce, J.D., *Comparative molecular field analysis (CoMFA): I. Effect of shape on binding of steroids to carrier proteins*, *J. Am. Chem. Soc.*, 110 (1988) 5959–5967.
7. Goodford, P.J., *A computational procedure for determining energetically favorable binding sites on biologically important macromolecules*, *J. Med. Chem.*, 28 (1985) 849–857.
8. Boobbyer, D.N.A., Goodford, P.J., McWhinnie, P.M. and Wade, R.C., *New hydrogen-bond potentials for use in determining Energetically favourable binding sites on molecules of known structure*, *J. Med. Chem.*, 32 (1989) 1083–1094
9. Wade, R.C., Clark K. and Goodford, P.J., *Further development of hydrogen-bond functions for use in determining energetically favorable binding sites on molecules of known structure: 1. Ligand probe groups with the ability to form two hydrogen bonds*, *J. Med. Chem.*, 36 (1993) 140–147.
10. Wade, R.C., Clark, K. and Goodford, P.J., *Further development of hydrogen-bond functions for use in determining energetically favourable binding sites on molecule of known structure: 2. Ligand probe group with the ability to form more than two hydrogen bonds*. *J. Med. Chem.*, 36 (1993) 148–156.
11. Goodford, P., *Multivariate characterization of molecules for QSAR analysis*, *J. Chemometrics*. 10 (1996) 107–111.
12. Cramer, III, R.D., DePriest, S.A., Patterson, D.E. and Hecht, P., *The developing practice of comparative molecular field analysis*, In Kubinyi, H. (Ed.) *3D QSAR in drug design: Theory, methods and applications*, ESCOM Science Publishers, Leiden. 1993. pp. 443–485.
13. Folkers, G., Merz, A. and Rognan, D., *CoMFA: Scope and limitations* In Kubinyi, H. (Ed.) *3D QSAR in drug design: Theory, methods and applications*. ESCOM Science Publishers. Leiden, 1993. pp. 583–618.
14. Kroemer, R.T. and Hecht, P., *Replacement of steric 6–12 potential-derived interaction energies by atom-based indicator variables in CoMFA leads to models of higher consistency*, *J. Comput.-Aided Mol. Design*. 9 (1995) 205–212.
15. Floersheim, P., Nozulak, J. and Weber, H.P., *Experience with comparative molecular fields analysis*, In Wermuth, C.G. (Ed.) *Trends in QSAR and molecular modelling 92* (Proceedings of the 9th European Symposium on Structure-Activity Relationships: QSAR and Molecular Modeling). ESCOM Science Publishers, Leiden. 1993. pp. 227–232.
16. Berendsen, H.J.C., *Electrostatic interactions*, In van Gunsteren, W.F., Weiner, P.K. and Wilkinson, A.J. (Eds.) *Computer simulation of biomolecular systems: Theoretical and experimental applications*. Vol. 2. ESCOM Science Publishers, Leiden. 1993. pp. 161–181.
17. Kim, K. H. and Martin, Y.C. *Direct predictions of linear free energy substituent effects from 3D structures using comparative molecular field analysis: 1. Electronic effects of substituted benzoic acids*, *J. Org. Chem.*, 34 (1991) 2723–2729.
18. Gasteiger, J. and Marsili, M., *Interactive partial equalization of orbital electronegativity A rapid access to atomic charges*, *Tetrahedron*, 36 (1980) 3219–3228.

19. (a) Chirlian, L.E. and Franel, M.M., *Atomic charges derived from electrostatic potentials: A detailed study*. J. Comput. Chem., (1987) 804–905.  
 (b) Besler, B.H., Merz, Jr., K.M. and Kollman, P.A., *Atomic charges derived from semiempirical methods*. J. Comput. Chem., 11 (1990) 431–439
20. Kroemer, K.T., Hecht, P. and Liedl, K.R., *Different electrostatic descriptors in comparative molecular field analysis: A comparison of molecular electrostatic and coulomb potentials*. J. Comput. Chem., 11 (1996) 1296–1308.
21. Wade, R.C., *Molecular interaction fields*, In Kubinyi, H. (Ed.) 3D QSAR in drug design, Theory, methods and applications, ESCOM Science Publishers, Leiden, 1993, pp. 486–505.
22. Goodford, P., *GRID user guide*, Edition 15, Molecular Discovery Ltd., Oxford, UK, 1997,
23. Liljefors, T. and Norrby, P.-O., unpublished results.
24. Mills, J.E.J. and Dean, P.M., *Three-dimensional hydrogen-bond geometry and probability information from a crystal survey*, J. Comput.-Aided Mol Design, 10 (1996) 607–622.
25. Clementi, S., Cruciani G., Rigan elli, D. and Valigi, R., *GOLPE: Merits and drawbacks in 3D-QSAR*, In Sanz, F., Giraldo, J. and Munaut, F. (Eds.) QSAR and molecular modelling: Concepts, computational tools and biological applications, Prous Science Publishers, Barcelona, 1996, pp. 408–414.
26. Goodford, P., personal communication.
27. Bemis, G.W. and Murcko, M.A., *The properties of known drugs: I. Molecular frameworks*, J. Med. Chem., 39 (1996) 2887–2983.
28. Dougherty, D.A., *Cation- $\pi$  interactions in chemistry and biology: A new view of benzene, Phe, Tyr, and Trp*, Science 271 (1996) 163–168.
29. Dougherty, D.A., and Stauffer, D.A., *Acetylcholine binding by a synthetic receptor: Implications for biological recognition*, Science, 250 (1990) 1558–1560.
30. (a) Sussman, J.L., Halel, M., Frolow, F., Oefner, C., Goldman, A., Toker, L. and Silman, I., *Atomic structure of acetylcholinesterase from Torpedo californica: A prototypic acetylcholine-binding protein*, Science, 253 (1991) 872–879.  
 (b) Harel, M., Quinn, D.M., Nair, H. K., Silman, I. and Sussman, J.L., *The X-ray structure of a transition state analog complex reveals the molecular origin of the catalytic power and substrate specificity of acetylcholinesterase*, J. Am. Chem. Soc., 118 (1996) 2340–2346.
31. Caldwell, J.W. and Kollman, P.A., *Cation- $\pi$  interactions: Nonadditive effects are critical in their accurate representation*, J. Am. Chem. Soc., 117 (1995) 3177–4178.
32. Kim, K.S., Lee, J.Y., Lee, S.J., Ha, T.-K. and Kim, D.H., *On binding forces between aromatic ring and quaternary ammonium compound*. J. Am. Chem. Soc., 116 (1994) 7399–7400.
33. Caldwell, J., Dung, L.X. and Kollman, P. A., *Implementation of nonadditive intermolecular potentials by use of molecular dynamics: Development of a water–water potential and water–ion cluster interactions*, J. Am. Chem. Soc., 112 (1990) 9144–9147.
34. Meng, E.C., Cieplak, P., Caldwell, J.W. and Kollman, P.A., *Accurate solvation free energies of acetate and methylammonium ions calculated with a polarizable water model* J. Am. Chem. Soc., 119 (1994) 12061–12062.
35. Mecozzi, S., West, Jr., A.P. and Dougherty, D.A., *Cation - $\pi$  interactions in simple aromatics: Electrostatics provide a predictive tool*, J. Am. Chem. Soc., 118 (1996) 2307–2308.
36. Cheney, B.V., Schultz, M.W., Cheney, J. and Richards, W.G., *Hydrogen-bonded complexes involving benzene as an H-acceptor*, J. Am. Chem. Soc. 110 (1988) 4295–4198.
37. Rodtham, D.A., Suzuki, S., Suenram, R.D., Lovas, F.J., Dasgupta, S., Goddard, III, W.A. and Blake, G.A., *Hydrogen bonding in the benzene–ammonia dimer*, Nature, 363 (1993) 735–737.
38. Trumpp-Kallmeyer, S., Hoflack, J., Briunvels, A. and Hibert, M., *Modeling of G-protein-coupled receptors: Application to dopamine, adrenaline, serotonin acetylcholine, and mammalian opsin receptors*, J. Med. Chem., 35 (1992) 3348–62.
39. Liljefors, T. and Norrby, P.-O., *Ab initio quantum chemical model calculations on the interactions between monoamine neurotransmitters and their receptors*, In Schwartz, T.W., Hjort, S.A. and Sandholm Kastrop, J. (Eds.) Structure and function of 7TM receptors, Alfred Benzon Symposium 39, Munksgaard, Copenhagen, 1996, pp. 194–207.

40. Liljefors, T. and Norrby, P.-O., *An ab initio study of the trimethylamine-formic acid and the trimethylammonium-formate anion complexes, their monohydrates and continuum solvation*, J. Am. Chem. Soc., 119 (1997) 1052–1058.
41. Williams, D.E., *Coulombic interaction in crystalline hydrocarbons*, Acta Cryst., A30 (1974) 71–17.
42. Pettersson, I. and Liljefors, T., *Benzene–benzene (phenyl–phenyl) interactions in MM2/MMP2 molecular mechanics calculations*, J. Comput. Chem., 8 (1987) 139–145.
43. Burley, S.K. and Petsko, G.A., *Aromatic–aromatic interaction: A mechanism of protein structure stabilization*, Science, 229 (1985) 23–28.
44. Burley, S.K. and Petsko, G.A., *Dimerization energetics of benzene and aromatic amino acid side chains*, J. Am. Chem. Soc., 108 (1986) 7995–8001.
45. Singh, J. and Thornton, J.M., *The interaction between phenylalanine rings in proteins*, FEBS Lett., 191 (1985) 1–6.
46. Chipot, C., Jaffe, R., Maigret, B., Pearlman, D.A. and Kollman, P.A., *Benzene dimer: A good model for  $\pi$ – $\pi$  interactions in proteins? A comparison between the benzene and the toluene dimers in the gas phase and in aqueous solution*, J. Am. Chem. Soc., 118 (1996) 11217–11224.
47. Schauer, M. and Bernstein, E.R., *Calculations of the geometry and binding energy of aromatic dimers: benzene, toluene, and toluene–benzene*, J. Chem. Phys., 82 (1985) 3722–3727.

**This Page Intentionally Left Blank**

# Comparative Binding Energy Analysis

Rebecca C. Wade<sup>a\*</sup>, Angel R. Qrtiz<sup>b</sup> and Federico Gago<sup>c</sup>

<sup>a</sup> European Molecular Biology Laboratory, Meyerhofstraße 1, Postfach 10.2209, 69012 Heidelberg, Germany

<sup>b</sup> Department of Molecular Biology, TPC-5, The Scripps Research Institute 10550 North Torrey Pines Road, La Jolla, CA 92037, U.S.A.

<sup>c</sup> Departamento de Farmacología Universidad de Alcalá, 28871 Alcalá de Henares, Madrid, Spain.

## 1. Introduction

Classical regression techniques have long been used to correlate the properties of a series of molecules with their biological activities in order to derive quantitative structure–activity relationships (QSAR) to assist the design of more active compounds [1]. This approach has been successfully extended to three dimensions by using molecular coordinates of the ligands to derive 3D QSARs [2]. However, the availability of the three-dimensional structures of many macromolecular drug targets has opened an alternative approach to drug design, namely structure-based drug design (SBDD), in which the physico-chemical interactions between the receptor and a series of ligands are used to rationalize the binding affinities [3,4]. SBDD makes use of techniques ranging from those employing simple scoring functions through molecular mechanics calculations to detailed free energy perturbation calculations employing molecular dynamics simulation [5]. Now, particularly as a result of recent developments in the design of targeted combinatorial libraries of compounds [6], it is becoming increasingly common to have data on the activities of a family of compounds *and* knowledge of the three-dimensional structure of the target macromolecule to which they bind. While the activities of these compounds could

---

\* To whom correspondence should be addressed.

### Abbreviations

CoMFA	Comparative Molecular Field Analysis
HSF-PLA <sub>2</sub>	Human synovial fluid phospholipase A <sub>2</sub>
PLS	Partial least squares
QSAR	Quantitative structure activity relationship
SBDD	Structure-based drug design
SDEP	Standard Deviation of Error of Predictions given by:

$$SDEP = \sqrt{\frac{\sum (Y - Y')^2}{N}} = \sqrt{\frac{PRESS}{N}}$$

where  $Y$  is experimental activity;  $Y'$  is predicted activity;  
and  $N$  is the number of compounds

$Q^2$  Performance metric given by:

$$Q^2 = 1 - \left[ \frac{\sum (Y - Y')^2}{\sum (Y - \langle Y \rangle)^2} \right]$$

where  $\langle Y \rangle$  is the average experimental activity

be improved using the techniques of classical QSAR, 3D QSAR or SBDD, none of these alone makes full, simultaneous and systematic use of all the available information. This is the purpose of Comparative Binding Energy (COMBINE) Analysis [7,8].

The 'COMBINE' acronym refers to combinations in terms of both data and techniques:

1. data on ligand–receptor structures and the measured activities of a series of ligands are combined;
2. molecular mechanics and chemometrics are combined for the analysis.

In outline, COMBINE analysis involves generating molecular mechanics models of a series of ligands in complex with their receptor and of the ligands and the receptor, in unbound forms, and then subjecting the computed ligand–receptor interaction energies to regression analysis in order to derive a QSAR relating ligand-binding constants or activities to weighted selected components of the ligand–receptor interaction energy. While the chemometric analysis performed is similar to that in a Comparative Molecular Field Analysis (CoMFA) [9], the data analyzed in COMBINE analysis differ by explicitly including information about the receptor-ligand interaction energies rather than only about the interaction properties of the ligands.

In contrast to free energy perturbation methods [10,11], a full sampling of phase space is not performed in COMBINE analysis: it is instead assumed that one or a few representative structures of the molecules are sufficient when experimental information about binding free energies is used for model derivation. Although any error in the modelling would introduce 'noise' into the dataset, this can be filtered out by means of the subsequent chemometric analysis.

Although occasionally there is a linear relationship between binding free energy and computed binding energy derived from molecular mechanics calculations for single conformations of the bound and unbound states of a series of ligand–receptor pairs, this is not the case in general. This is because the entropic contribution to binding can vary over a series of ligands and because sufficiently accurate modelling of a full series of compounds can be difficult to achieve. A number of authors have correlated binding free energies with a few terms, defined according to physical interaction type, of the computed binding energies by linear regression [12–16]. A physical basis for such an analysis is provided by linear response theory which relates the electrostatic binding energy to the electrostatic binding free energy [16]. The COMBINE method differs from these approaches, in that more extensive partitioning of the binding energy is considered and multivariate regression analysis is used to derive a model. This is important for two reasons: firstly, from a modelling perspective, because it is not assumed that the computed components of the binding free energy can be calculated with high accuracy. Rather, one of the foundations of COMBINE analysis is the realization that such calculations are usually noisy, and that is why only those contributions of the binding energy that present the best predictive ability are selected and weighted in the resultant model. Secondly, it is realized that binding free energy is rarely a linear function of binding energy. The extensive decomposition allows those components that are predictive of binding free energy to be detected and these may implicitly represent other physically important interactions or even entropic terms.

A QSAR model is derived for *each* target receptor studied with the COMBINE method, as the method was specifically designed for ligand optimization. Thus, a derived regression model is not applicable to *all* ligand-receptor interactions in the way that a general-purpose empirical ‘scoring function’ derived from statistical analysis of a diverse set of protein–ligand complexes is designed to be [17,18]. The philosophy is to account for peculiarities in the modelling and parameterization of a given set of compounds, so that both optimal and inexpensive predictive models can be derived.

In the next section, we describe the COMBINE analysis method. This is followed by a description of its application to two sets of enzyme inhibitors. COMBINE analysis is then discussed in terms of the quality of its predictions, its pros and cons, and its future prospects.

## 2. The COMBINE Method

### 2.1. Theory

The goal of the COMBINE analysis procedure is to derive an expression for the receptor binding free energy of a ligand,  $\Delta G$ , of the following form.

$$\Delta G = \sum_{i=1}^n w_i \Delta u_i^{sel} + C \quad (1)$$

From this expression, biological activities may be derived by assuming that these quantities are functions of  $\Delta G$ . The expression is derived by analyzing the interaction of a set of ligands with experimentally known activities or binding affinities for a target receptor. Conformations of the ligand–receptor complexes and the unbound ligands and receptor are modelled with a molecular mechanics force field. It is assumed that these are representative of the full ensemble of structures that would be sampled by these molecules. From these, ligand–receptor binding energies,  $\Delta U$ , are computed for each ligand.

$$\Delta U = E_{lr} - E_r - E_l = E^{inter}_{lr} + \Delta E_r + \Delta E_l \quad (2)$$

where  $E_{lr}$  and  $E^{inter}_{lr}$  are the total and intermolecular energies, respectively, of the ligand–receptor complex;  $E_r$  the energy of the unbound receptor  $r$ ; and  $\Delta E_r$  is the change in the potential energy of the receptor upon formation of the complex; and  $\Delta E_l$  are the corresponding energies for the ligand  $l$ .  $\Delta U$  itself will not, in general, correlate with  $\Delta G$ , but it is likely that some of its components will. Therefore,  $\Delta U$  is partitioned into components according to physical type and which of the  $n_l$  defined fragments of the ligand and  $n_r$  defined regions of the receptor are involved.

$$\begin{aligned} \Delta U = & \sum_{i=1}^{n_l} \sum_{j=1}^{n_r} u_{ij}^{vdw} + \sum_{i=1}^{n_l} \sum_{j=1}^{n_r} u_{ij}^{elec} + \sum_{i=1}^{n_l} \Delta u_i^{B,L} + \sum_{i=1}^{n_l} \Delta u_i^{A,L} + \\ & \sum_{i=1}^{n_l} \Delta u_i^{T,L} + \sum_{i<i'}^{n_l} \Delta u_{ii'}^N \sum_{j=1}^{n_r} \Delta u_j^{B,R} + \sum_{j=1}^{n_r} \Delta u_j^{A,R} + \sum_{j=1}^{n_r} \Delta u_j^{T,R} + \sum_{j<j'}^{n_r} \Delta u_{jj'}^N \end{aligned} \quad (3)$$

The first two terms on the right-hand side describe the intermolecular interaction energies between each fragment  $i$  of the ligand and each region  $j$  of the receptor. The next four terms describe changes in the bonded (bond, angle and torsion) and the non-bonded (Lennard-Jones and electrostatic) energies of the ligand fragments upon binding to the receptor, and the last four terms account for changes in the bonded and non-bonded energies of the receptor regions upon binding of the ligand.

The  $n$  terms,  $\Delta u_i^{sel}$  in Eq. 1 that correlate with  $\Delta G$  are selected from the ligand–receptor binding energy,  $\Delta U$ , and the coefficients  $w_i$  and constant  $C$  determined by regression analysis.

## 2.2. Implementation

The procedure for COMBINE analysis is outlined schematically in Fig. 1. There are essentially three steps to be followed for the derivation of a COMBINE model, namely modelling of the molecules and their complexes, measurement of the interactions between ligands and the receptor and chemometric analysis to derive the regression equation. Each of these steps will be considered in turn.

### 2.2.1. Molecular modelling

The three-dimensional models of the ligand–receptor complexes and the unbound receptor and ligands can be derived with a standard molecular mechanics program. The dependence of the results on the modelling protocol followed has not yet been investigated in detail. The use of different starting conformations for the receptor, the inclusion of positional restraints on parts of the receptor, different convergence criteria during energy minimization or different ways of treating the solute–solvent interface and the dielectric environment can all produce different regression equations. The sensitivity of COMBINE models to these factors compared to corresponding QSAR models that use the overall intermolecular interaction energies as regressors remains to be fully studied. One of the appealing characteristics of the COMBINE approach, however, is that, as a result of the decomposition of the intermolecular interaction energies on the basis of chemical fragments, artefacts in the modelled ligand–receptor complexes that could otherwise pass unnoticed can be easily detected.

In general, the limited available experience indicates that energy minimization should be mild, so that major steric clashes are eliminated while avoiding artefactual structural distortions due to inaccuracies in the modelled forces. It is particularly important to employ a suitable model of the solvent environment. When modelling explicit water molecules, inclusion of only crystallographic water molecules may not be sufficient [30] but we have found that solvation of the ligand and receptor molecules with an approximately 5 Å thick shell of water molecules produces reasonable results [7,8].

While several conformations of each molecule or complex, derived for example from conformational analysis or molecular dynamics simulations, could be used for COMBINE analysis, we have so far used only single conformations derived from energy minimization.



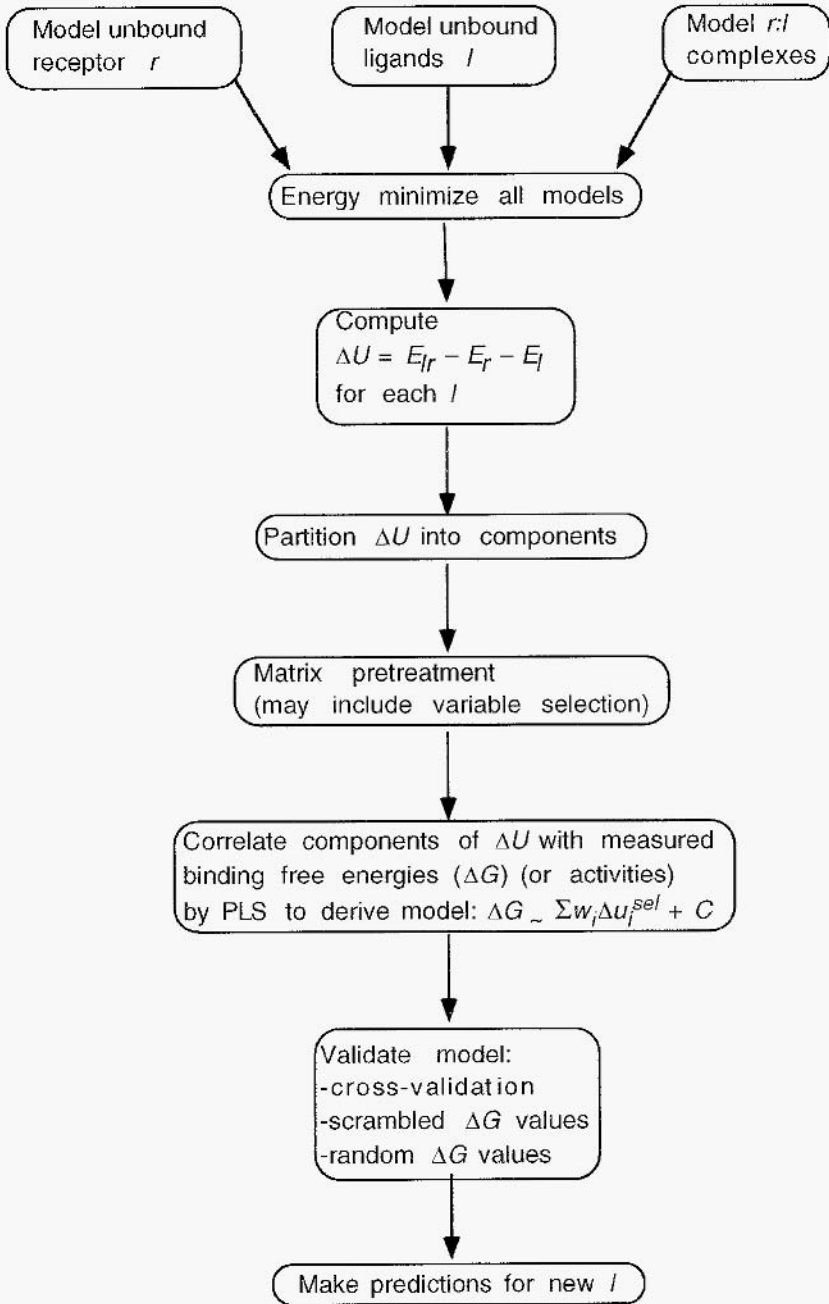


Fig. 1. Flowchart showing the stages of a COMBINE analysis

### 2.2.2. *Measurement of the interaction energies*

After modelling, ligand and receptor energies must be computed and decomposed in the form required for regression analysis. That is, a matrix is built with columns representing the energy components given in Eq. 3 and rows representing each compound in the set. A final column containing inhibitory activities is then added to the matrix.

The energy decomposition scheme must, at some point, meet the two opposing tendencies of the Scylla of detailing enough energy terms that the elements responsible for the activity differences can be isolated and the Charybdis of including so many terms that the signal-to-noise ratio is so low that the subsequent analysis fails to obtain a meaningful model. Recent investigations in our groups indicate that a reasonable compromise is to consider each residue in the receptor as contributing two interaction terms: one for van der Waals and one for electrostatic interactions. Inclusion of intramolecular energies, which are a potential source of noise and cumbersome to compute, appears to result in little improvement in the regression models [19]. For these reasons, it is probably advisable to omit them from the statistical analysis, although their importance can be expected to depend on the extent of conformational changes on binding.

In our studies, we have found that differences in the way electrostatic interactions are computed can have a considerable effect on the regression models [19] and Pérez et al. (submitted). The electrostatic energies can be given by a Coulombic expression or derived from solution of the Poisson-Boltzmann equation according to classical continuum electrostatic theory [20]. We are currently comparing these methods in COMBINE analysis and our results so far underscore the general importance of considering the desolvation free energies upon binding as additional variables. Although the information they provide may not always be 'new' as it can be implicitly contained in other intermolecular electrostatic energy terms that are highly correlated with them. This collinearity may explain why good results can be obtained when this physically relevant contribution is not included in the analysis (see section 3.2 for additional details).

Additional terms to describe entropic contributions — e.g. freezing out of side-chain rotamers on binding — could also be included in COMBINE analysis. This has not yet been tested and their influence on the models derived remains to be investigated.

### 2.2.3. *Chemometric analysis*

As a result of the large number of terms and the correlated nature of the variables, partial least squares (PLS) [21] is the technique of choice for deriving the regression equation. In PLS analysis, a model is derived by projecting the original matrix of energy terms onto a small number of orthogonal 'latent variables'. After this projection, the original energy terms are given weights according to their importance in the model. Those that do not contribute to explaining the differences in binding have negligible effects and just add 'noise'. When the ratio between the really informative variables and these 'noisy' variables is too low, the PLS method may fail to obtain a model. A sensible strategy to avoid this situation is to pretreat the data by setting very small values to zero and removing those variables that take nearly constant values in the matrix. If this pretreatment is insufficient, variable selection can be carried out, with the aim of elimi-

nating from the matrix those variables that do not contribute to improving the predictive ability of the model. To this end, we have employed the GOLPE method [22], in which the effect of the variables on the predictive ability of the models is evaluated through fractional factorial designs and advanced cross-validation techniques. Variable selection must, however, be carried out with care as it is prone to overfitting the data, particularly when selection is pursued beyond a certain limit [19].

COMBINE models can be validated by following the same principles as used in other 3D QSAR methodologies [2]. Apart from the minimum requirement of internal consistency (as evaluated by cross-validation), random exchange of the biological activities among the different molecules (permutation or *scrambling*) and the use of external test sets are strongly recommended, in order to highlight possible overfitting problems.

### 3. Applications

#### 3.1. Phospholipase $A_2$

The first application of COMBINE analysis [7,8] was done on a set of 26 inhibitors of human synovial fluid phospholipase  $A_2$  (HSF-PLA $_2$ ), an enzyme that catalyzes the hydrolysis of the *sn*-2 acyl chain of phosphoglycerides releasing arachidonic acid, the precursor of several inflammatory mediators. The enzyme is mainly alpha-helical and has about 120 amino acid residues and seven disulfide bridges. The inhibitors are transition state analogs that bind in the substrate binding site, a slot whose opening is on the enzyme surface and runs all the way through the enzyme. The key catalytic residues are His-48 and Asp-99, and a calcium ion bound to the active site is required for substrate binding.

An initial scatterplot showed a very poor correlation between biological activities and calculated binding energies ( $r = 0.21$ , Fig. 2a). However, the COMBINE model

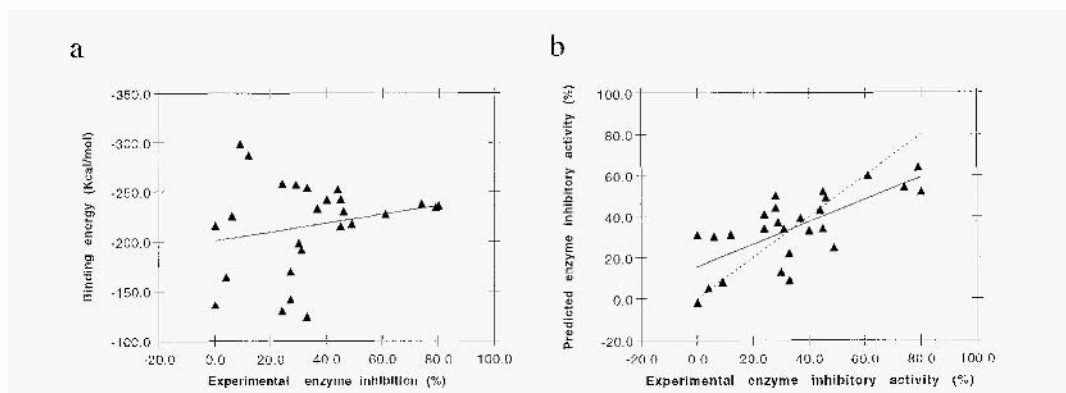


Fig. 2. (a) Total calculated binding energy of the HSF-PLA $_2$  inhibitors to the enzyme versus activity expressed as percentage inhibition ( $r = 0.21$ ). (b) Predicted versus experimental activity for the HSF-PLA $_2$  inhibitory activity on external 'blind' cross-validation. The predictive model was derived using two latent variables and yielded a fitted  $R^2 = 0.92$ , an internally cross-validated  $Q^2 = 0.82$ , and an externally cross-validated  $Q^2 = 0.52$ . The broken line corresponds to a perfect fit, and the solid line shows the regression fit ( $r = 0.71$ ).

obtained for this data set showed good fitting properties and significant predictive ability (Fig. 2b), as assessed by a value of  $Q^2 - \langle Q^2 \rangle_s$  of 0.59, that is, the difference between the estimated  $Q^2$  and the average  $Q^2$  obtained in 20 scrambled models [7].

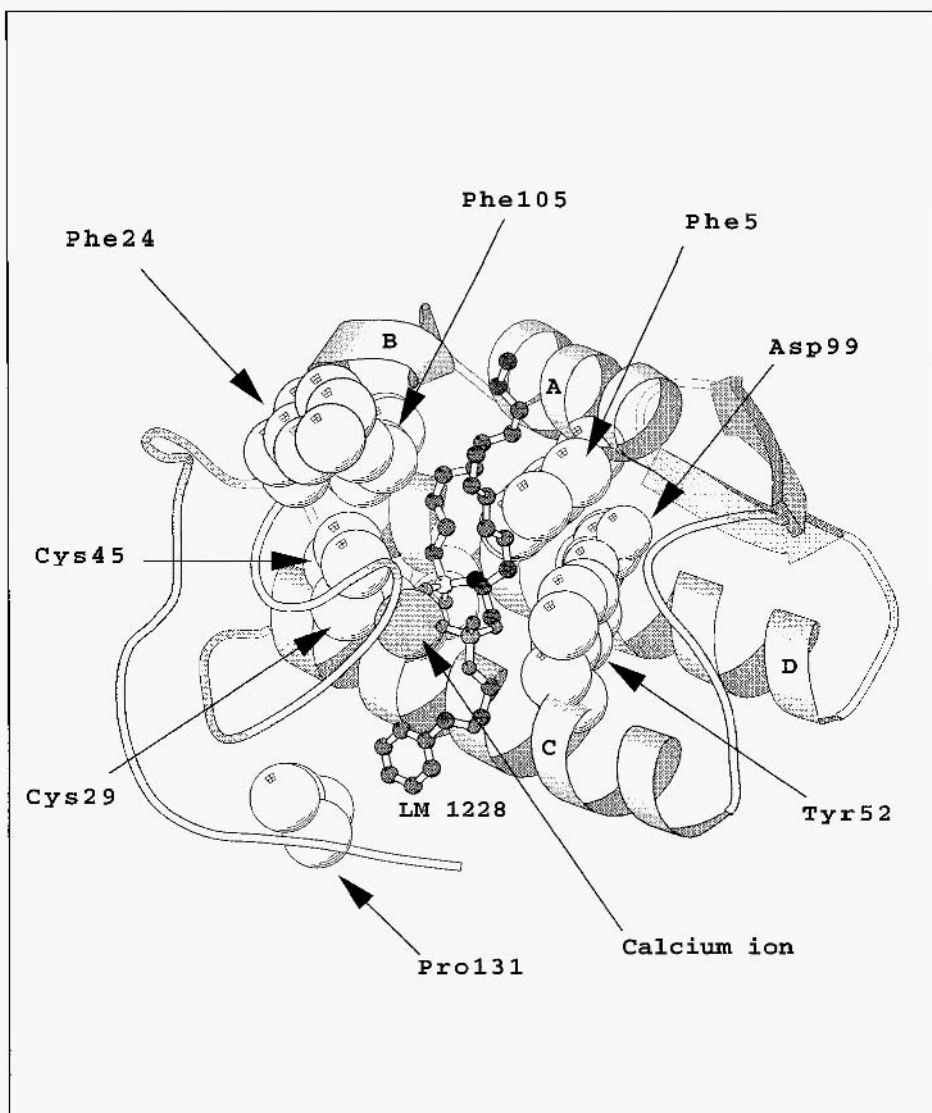


Fig. 3. Schematic diagram of HSF-PLA2 complexed with a representative inhibitor (LM1228). Spheres represent atoms of protein residues lining the binding site that are frequently selected to contribute to regression models in COMBINE analysis (see reference [7] and table 3 therein). The calcium ion in the active site (shaded sphere) makes an important contribution to COMBINE models. This diagram was generated with the molscript program [32].

From the initial energy matrix, around 50 energy terms were finally selected to obtain the regression model. These energy contributions reflect complex relationships since they may have been selected because of their correlation with other variables. For this reason, they are better regarded as ‘effective’ energies and care must be exercised to avoid misinterpretations. However, it is of interest to examine these energy contributions more closely. Most of the selected intermolecular effective energies correspond to interactions with residues in the enzyme active site. Overall, the model suggests that in this particular dataset the binding affinity is dominated by electrostatic interactions with the calcium ion located at the binding site. Several van der Waals interactions then modulate the affinity of the inhibitors. Some of the residues in the B helix (top left, Fig. 4) and the calcium-binding loop form a rigid wall sensitive to the conformation of the *sn*-2 chain. On the other side of the binding site, two aromatic residues form a pocket in which an inhibitor must fit in order to have optimal activity. Finally, the C-terminal region of the enzyme forms an additional pocket with favorable interactions for inhibitors with benzyl moieties in the *sn*-3 chain. It is noteworthy that other researchers have arrived at similar SARs for a set of indole-based compounds [23], which suggests that the energies selected by COMBINE analysis may have some physical meaning in favorable cases. On the other hand, some other interactions have no clear physical meaning and they seemed simply to be correlated with some other, physically more relevant, variables. This is the only way to rationalize many of the selected interactions between the phosphate group of the inhibitors and charged residues exposed on the enzyme surface, some of them very far away from the phosphate group.

### 3.2. HIV-1 proteinase

Perhaps the most controversial issue that arose when the COMBINE approach was first reported [7,8] was the variable selection procedure. The large number of variables used in the original paper was a consequence of splitting the inhibitors into several fragments and considering intramolecular energy terms for both the protein and the inhibitors. The phospholipase A<sub>2</sub> example, on the other hand, can be regarded as a particularly difficult case, in the sense that the initial correlation between experimental activities and calculated binding energies was rather poor. The high correlation between calculated intermolecular interaction energies (using the MM2X force field) and enzyme inhibition reported for a set of 33 HIV-1 proteinase inhibitors [14] prompted us to apply the COMBINE methodology to this same data set using the AMBER force field [24]. The coordinates of L-689,502-inhibited HIV-1 proteinase were used for the receptor, which included the water molecule that stabilizes the closed conformation of the dimeric enzyme by bridging a  $\beta$ -hairpin from each monomer to the inhibitors. Adopting the same philosophy as followed by the researchers at Merck [14], the enzyme was held fixed and only the inhibitors and the water molecule were allowed to relax on energy minimization. The intermolecular interaction energies were then calculated and related to the biological activities by means of a simple linear regression equation. For the COMBINE decomposition scheme, these interaction energies were partitioned on a per

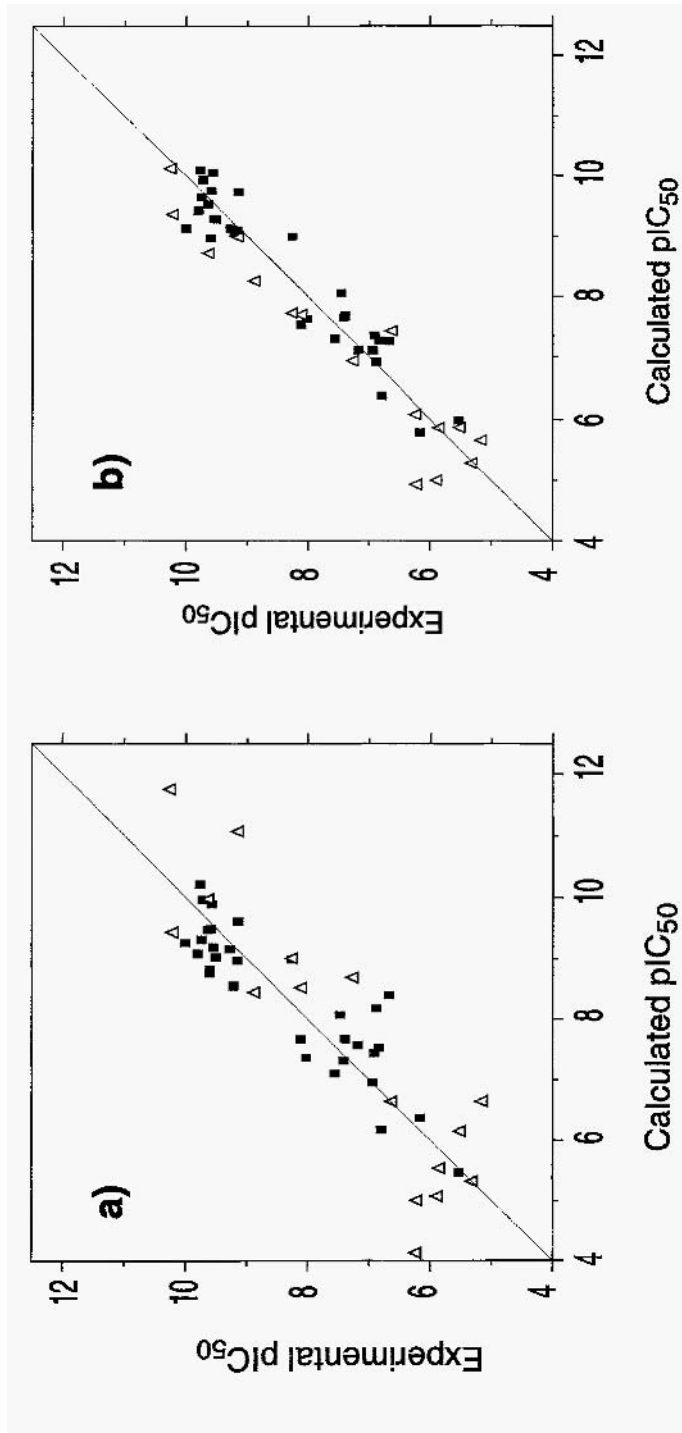


Fig. 4. Experimental versus predicted activities ( $pIC_{50}$ ) for the set of HIV-1 protease inhibitors used to derive the model (squares) and for the external set of 16 inhibitors (triangles). (a) Simple model that considers overall intermolecular interaction energies ( $R^2 = 0.79$ ;  $Q^2 = 0.77$ ; external SDEP = 1.08); and (b) COMBINE model (2 principal components) in which the electrostatic interaction energy terms have been calculated by continuum methods and desolvation effects have been included ( $R^2 = 0.90$ ;  $Q^2 = 0.72$ ; external SDEP = 0.62). One of the compounds in reference [14] has a  $pIC_{50}$  value less than 5.0 and was omitted from both analysis.

residue basis. Each inhibitor was considered as a single fragment and no intramolecular energy terms were considered. The number of variables per inhibitor was thus equal to 2 (van der Waals and electrostatic) times the number of protein residues ( $[2 \times 99 \text{ amino acids}] + 1 \text{ water molecule}$ ) = 398. No variable selection was employed. The resulting matrix was pretreated simply by zeroing those interaction energies with absolute values lower than 0.1 kcal/mol and removing any variables with a standard deviation below 0.1 kcal/mol. This pretreatment reduced the number of variables that entered the PLS analysis to around 50. It is noteworthy that the number of variables was effectively reduced in this example, without the need for variable selection, underscoring the fact that it is possible for a simple pretreatment of the original matrix to accomplish virtually the same effect.

Plots of predicted versus observed  $\text{pIC}_{50}$  values obtained for the inhibitors studied and for an additional set of 16 inhibitors not included in the derivation of the models [14] are shown in Fig. 4. While the internal cross-validation results are comparable in both cases, it is apparent that the PLS model from the COMBINE analysis (Fig. 4b) outperforms the simpler regression equation (Fig. 4a) in external predictions (Pérez, et al., submitted).

Attempts to incorporate desolvation effects into a predictive model were reportedly unsuccessful for the HIV-1 proteinase complexes [14]. The electrostatic interaction energy terms incorporated into the COMBINE model described in Fig. 4b were calculated by means of a continuum method, as implemented in the DelPhi program [25], using dielectric values of 4 and 80 to represent the molecular interiors and the surrounding solvent, respectively. The electrostatic desolvation energies of both the protein and the inhibitors were also included as two additional variables. Their incorporation into the model resulted in a slight improvement in predictive ability ( $Q^2 = 0.72$  versus  $Q^2 = 0.70$  for 2 principal components) [Pérez et al., submitted]. Interestingly, the variables whose weights were most affected by the desolvation energy correction were precisely those involving the charged residues that participate in strong electrostatic interactions between the inhibitors and the enzyme (Fig. 5).

## 4. Discussion

### 4.1. Quality of results

COMBINE models for the set of HSF-PLA<sub>2</sub>-inhibitor complexes compare favorably with CoMFA models for the same inhibitors aligned as in the modelled bound complexes [19]. Using the same dataset and the same cross-validation method, the best CoMFA model (the so-called N-T-C model [19], selected for its optimal predictive ability for external test sets) had a  $Q^2 = 0.62$  and a standard deviation of error of predictions (SDEP) = 13.5. The corresponding values obtained with COMBINE analysis were  $Q^2 = 0.82$  and SDEP = 9.3 [7]. These figures of merit suggest a better predictive performance for COMBINE, but it should be noted that no scrambling of the biological data was done in the CoMFA study, so that the methods have not been compared using the more rigorous 'excess' $Q^2$  value described in section 3.1.

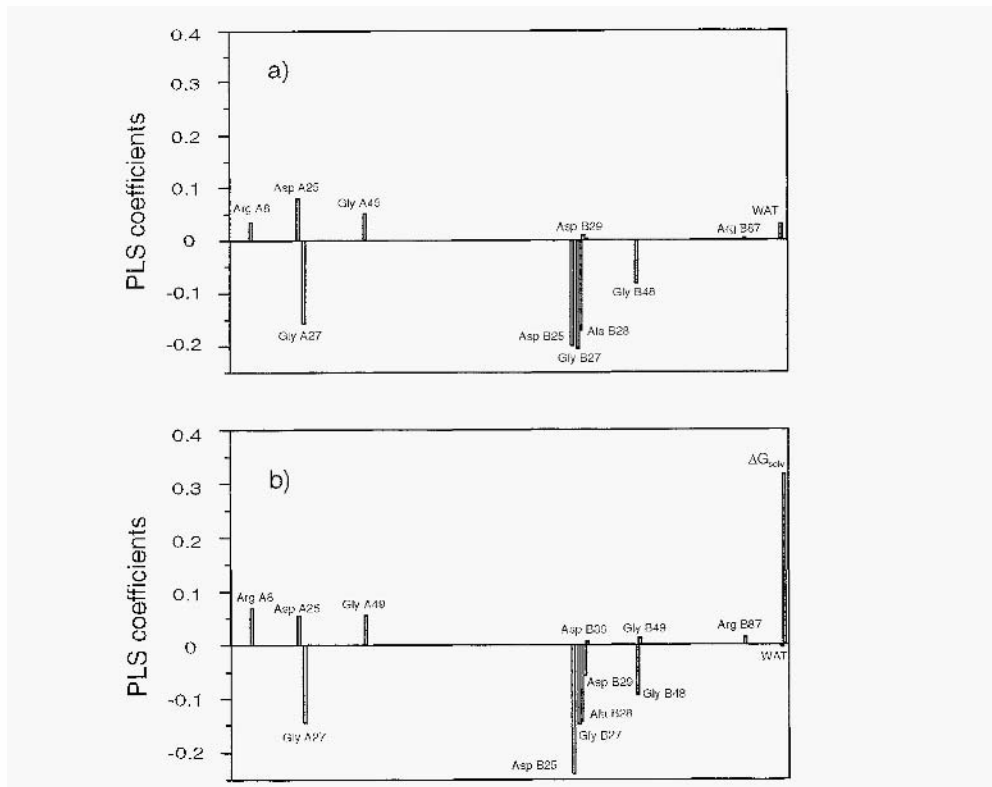


Fig. 5. (a) PLS coefficients for the electrostatic contributions of each residue from a COMBINE analysis on the set of HIV-1 proteinase-inhibitor complexes. Only the coefficients exhibiting significant variance are given non-zero values and labelled. (b) PLS coefficients after incorporation of desolvation effects. The coefficient of the electrostatic contribution to the desolvation of the inhibitors ( $\Delta G_{solv}$ ) is the largest of all and clearly modulates some of the other interactions. The electrostatic contribution to the desolvation of the protein, on the other hand, appears to be highly correlated with other variables, so that its presence in the model is not required

For the HSF-PLA<sub>2</sub> and HIV-1 protease examples, both the conventional and cross-validated squared correlation coefficients provided by COMBINE analysis compared very favorably with the ones obtained by the classical approach of using just the overall intermolecular interaction energies as the independent variables (Figs 2 and 4).

A further advantage of COMBINE analysis is that it highlights those regions in the enzyme binding site that contribute most to the differences in activity among the ligands. In the two examples reported above, this analysis allowed the identification of mechanistically important residues, and this information may guide the design of further chemical modifications on the inhibitors. For the HSF-PLA<sub>2</sub> case (see Fig. 3), the regions detected as important for activity were largely consistent with those identified by CoMFA and in studies of different sets of ligands [26].



#### 4.2. Why is COMBINE successful?

It is pertinent to consider the reason for the success of the COMBINE methodology in predicting binding affinities, given that the method is based on an incomplete model of molecular interactions and conformations. It has been suggested that successful binding free energy predictions rely on multiple cancellations among the contributions that are neglected in the model [27]. These cancellations may be fortuitous but can mostly be expected to originate from enthalpy–entropy compensation on binding in aqueous solution. An alternative explanation is that, *within a congeneric series*, it is possible that some of the neglected terms, particularly entropic ones, do not contribute to the binding free energy differences because they are similar for all the members of the series. Entropic contributions to binding free energies can be roughly divided into three separate terms: changes in translational and rotational entropy, changes in configurational entropy and changes in solvent entropy [28]. Within a series, it is likely that values are approximately constant for translational and rotational entropy changes, and perhaps also for configurational and vibrational entropy contributions, as well as for the enthalpic changes arising from changes in the conformation of the receptor. One component for which this cannot be expected to be the case in general, however, is the differential solvation free energy upon binding. It has even been suggested that, within a certain closely related series of ligands binding to a common receptor, the solvation free energies may vary more significantly than the intrinsic ligand–receptor interactions [29]. However, this does not mean that it is absolutely essential to incorporate this term explicitly in the model in order to obtain good binding affinity predictions. For example, in a study involving a series of flavonoid trypsin inhibitors [30], it was found that the inhibitory potencies could be correlated well with calculated binding energies derived using a simple distance-dependent dielectric for modelling electrostatic interactions and considering only the bound state. When a continuum model was used to compute electrostatic interactions, the accuracy of the computed differences in binding free energies was increased, but for only one inhibitor was it important to take into account the differences in solvation free energies between bound and unbound conformations. In this example, decomposition of the binding free energy components showed that the differences in the total electrostatic free energy of binding (including the desolvation contribution) were much smaller than the corresponding differences in van der Waals binding energies, so that it was this latter term that dominated the binding free energy differences. A similar scenario can thus be envisaged in some of the other cases in which reasonable correlations between *in vitro* activities and calculated binding energies are obtained without considering solvation effects.

As shown above, it is our experience that incorporation of a more rigorous treatment of electrostatic interactions by means of continuum models and consideration of desolvation effects upon binding lead to an improvement of the correlations and to more robust COMBINE models [Pérez et al., submitted]. However, the increase in predictive ability upon incorporation of desolvation contributions has not been, in the HIV protease test case, as large as we anticipated (see section 3.2). The reason is that some intermolecular interactions appear to be collinear with the desolvation free

energies of the inhibitors, so that some information about the solvent effects is implicitly included in the intermolecular energies (see sections 3.2 and 2.3 for additional details).

#### 4.3. *When to apply COMBINE antilysis*

The COMBINE method can be applied to any dataset for which the following are available:

1. experimental binding or activity measurements for a series of ligands that bind to a target macromolecular receptor; the number of measurements necessary for obtaining a good model will depend on the quality of the data but 15 is a reasonable lower limit;
2. an experimentally determined three-dimensional structure of the target macromolecular receptor complexed to a representative ligand.

As in other QSAR techniques, the method relies on the assumption that all ligands analyzed bind to the receptor at the same binding site and that the binding mode can be deduced by comparative modelling techniques. For situations in which the binding mode can alter dramatically on minor modification of the ligand [31], it should in principle be possible to include more than one model of each complex in the analysis and deduce the correct binding mode from the regression analysis [32]. However, this has yet to be investigated in a practical application.

While the method has so far been applied only to the binding of a series of small molecule ligands to a protein, it should also be applicable to interactions between macromolecules. for instance. to predict the effects of protein mutations.

## 5. Conclusion

COMBINE analysis provides a means to exploit information about the three-dimensional structure of a target macromolecule and about measured activities of a series of compounds in order to derive a model to predict their binding free energies. The applications described here to HSF-PLA<sub>2</sub> and HIV proteinase inhibitors demonstrate that predictive models that give insight into the mechanism of inhibition can be derived. The predictive performance of these models compares very favorably with that of other regression methods that make more conventional use of molecular mechanics interaction energies or other QSAR techniques such as CoMFA. Analysis of the importance of different energetic terms and the effects of different chemometric protocols has shown how robust models can be obtained. Further work should include the exploration of the effects of including additional descriptors such as those that explicitly describe entropic changes on binding. In addition, it may be possible to obtain improvements in the method by optimizing modelling protocols and tailoring the chemometric tools more specifically towards the data extracted from the molecular systems studied.

Further applications to different types of ligand-receptor datasets are necessary for a complete assessment of the capabilities of the method. However, the quality of the

COMBINE models obtained so far is probably good enough to encourage its widespread application to QSAR problems where it can assist in the design of new compounds which will provide real experimental 'blind tests' of predictive ability.

## Acknowledgements

We thank our collaborators Ana Checa and Carlos Pérez for their enthusiastic participation in parts of the work presented here; Dr. Gabriele Cruciani and Dr. Manuel Pastor for helpful comments on chemometrics and provision of the GOLPE program; Dr. Albert Palomer for bringing the problem of structure-activity relationships of PLA<sub>2</sub> inhibitors to our attention; Dr. Mayte Pisabarro for her contribution to the modelling of the PLA<sub>2</sub> inhibitors; Dr. Kate Holloway for provision of cartesian coordinates for the training set of inhibitors and L-689, 502-bound HIV-1 proteinase; and Dr. P.J. Kraulis for the MOLSCRIPT program.

A.R.O. was the recipient of a predoctoral fellowship from the Comunidad Autonoma de Madrid. This work was supported in part by Laboratorios Menarini, S.A. (Badalona, Spain) and the Spanish CICYT (SAF projects 94/630 and 96/231 to F.G.).

## References

1. Kubinyi, H., *QSAR: Hansch analysis and related approaches*, VCH, Weinheim, 1993.
2. Kubinyi, H. (Ed.), *3D-QSAR in drug design: Theory methods and applications*, ESCOM, Leiden, 1993.
3. Kuntz, I.D., *Structure-based strategies for drug designed and discovery*, *Science* 257 (1992) 1078–1082.
4. Blundell, T.L., *Structure-based drug design*. *Nature*, 381 Suppl. (1996) 23–26
5. Greer, J., Erickson, J.W., Baldwin, J.J. and Varney, M.D., *Application of the three-dimensional structures of protein target molecules in structure-based drug design*, *J. Med. Chem.*, 37 (1994) 1035–1054.
6. Hogan, J.C., *Directed combinatorial chemistry*. *Nature*, 384 Suppl. (1996) 17–19.
7. Ortiz, A.R., Pisabarro, M.T., Gago, F. and Wade, R.C., *Prediction of drug binding affinities by comparative binding energy analysis*, *J. Med. Chem.*, 38 (1995) 2681–2691.
8. Ortiz, A.R., Pisabarro, M.T., Gago, F. and Wade, R.C., *Prediction of drug binding affinities by comparative binding energy analysis: Application to human synovial fluid phospholipase A<sub>2</sub> Inhibitors*, In *QSAR and molecular modelling: Concepts, computational tools and biological applications*, Sanz, F., Giraldo, J. and Manaut, F. (Eds.) J.R. Prous, Barcelona, 1995, pp. 439–443.
9. Cramer III, R.D., Patterson, D.E. and Bunce, J.D., *Comparative molecular field analysis (CoMFA): I. Effect of shape on binding of steroids to carrier proteins*, *J. Am. Chem. Soc.*, 110 (1988) 5959–5067.
10. Straatsma, T.P. and McCammon, J.A., *computational alchemy*, *Ann. Rev. Phys. Chem.*, 43 (1992) 407–435.
11. Kollman, P., *Free energy calculation: Application to chemical and biochemical phenomena* *Chem. Rev.*, 93 (1993) 2395–2417
12. Blancy, J.M., Weiner, P.K., Dearing, A., Kollman, P.A., Jorgensen, E.C., Oatley, S.J., Burrige, J.M. and Blake, C.C.F., *molecular mechanics simulation of protein-ligand interactions: Binding of thyroid hormone analogues to prealbumin*, *J. AM. Chem. soc.*, 104 (1982) 6424–6434.
13. Menziani, M.C., De Benedetti, P.G., Gago, F. and Richards, W.G., *The binding of benzene sulfonamides to carbonic anhydrase enzyme: A molecular mechanics study and quantitative structure-activity relationship*, *J. Med. Chem.*, 32 (1989) 951–956.
14. Holloway, M.K., Wai, J.M., Halgren, T.A., Fitzgerald, P.M.D., Vacca, J.P., Dorsey, B.D., Levin, R.B., Thompson, W.J., Chen, L.J., deSolms, S.J., Gaffin, N., Ghosh, A.K., Giuliani, E.A., Graham, S.L., Guare, J.P., Hungate, R.W., Lyle, T.A., Sanders, W.M., Tucker, T.J., Wiggins, M., Wiscount, C.M.,

- Woltersdorf, O.W., Young, S.D., Darke, P.L. and Zugay, J.A. *A priori prediction of activity for HIV-1 protease inhibitors employing energy minimization in the active site*, J. Med. Chem., 38 (1995) 305–317.
15. Grootenhuis, P.D.J. and van Galen, P.J.M., *Correlation of binding affinities with non-bonded interaction energies of thrombin-inhibitor complexes*, Acta Cryst., D51 (1995) 560–566.
  16. Åqvist, J., Medina, C. and Samuelsson, J.E., *A new method for predicting binding affinity in computer-aided drug design*, Prot. Eng., 7 (1994) 385–391
  17. Böhm, H.-J., *The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure* J. Comput-Aided Mol. Design. 8 (1994) 243–256.
  18. Head, R.D. Smythe, M.L., Oprea, T.I., Waller, C.L., Green, S.M. and Marshall, G.R., *VALIDATE: A new method for the receptor- based prediction of binding affinities of novel ligands*. J. Am. Chem. Soc., 118 (1996) 3959–3969.
  19. Ortiz, A.R., Pastor, M., Palomer, A., Cruciani, G., Gago, F. and Wade, R.C., *Reliability of comparative molecular field analysis models: Effects of data scaling and variable selection using a set of human synovial phospholipase A2 inhibitors*, J. Med. Chem. (1997) 40. 1136–1148.
  - Honig, B. and Nicholls, A., *Classical electrostatics in biology and chemistry* Science, 268 (1995) 1144–1149.
  21. Wold, S., *PLS-partial least squares projections to latent structures*. In 3D-QSAR in drug design: Theory, methods and applications. Kubinyi, H. (Ed.) ESCOM. Leiden. 1993. pp. 523–550.
  22. Baroni, M., Constantino, G., Cruciani, G., Riganelli, D., Valigi, R. and Clementi, S., *Generating optimal linear PLS estimations (GOLPE): An advanced chemometric tool for handling 3DQSAR problems*, Quant. Struct.-Act. Relat., 12 (1993) 9–20.
  23. Schevitz, R.W., Bach, N.J., Carlson, D.G., Chrigadze, N.Y., Clawson, D.K., Dillard, R.D., Draheim, S.E., Hartley, L.W., Jones, N.D., Mihelich, E.D., Olkowski, J.L., Snyder, D.W., Sommers, C. and Wery, J.-P., *structure-based design of the first potent and selective inhibitor of human non-pancreatic secretory Phospholipase A<sub>2</sub>*, Nat. Struct. Biol., 2 (1995) 458–465.
  24. Cornell, W.D., Cieplak, P., Bayly, C.I., Gould, I.R., Merz, K.M., Ferguson, D.M., Spellmeyer, D.C., Fox, T., Caldwell, J.W. and Kollman, P. A., *A second generation force field for the simulation of proteins, nucleic acids, and organic molecules*. J. Am. Chem. Soc., 117 (1995) 5179–5197.
  25. Gilson, M.K., Sharp, K.A. and Honig, B.H., *Calculating the electrostatic potential of molecules in solution: Method and error assessment*, J. comput. Chem., 9 (1987) 327–335.
  26. Wheeler, T.N., Elanchar, S.G., Andrew, R.C., Fang, F., Gray- Nunez, Y., Harris, C.O., Lambert, M.H., Mehrotra, M.M., Parks, D.J., Ray, J.A. and Smalley, T.L., *substrate specificity in short- chain phospholipid analogs at the active site of human synovial phospholipase A<sub>2</sub>*, J. Med. Chem., 37 (1995) 4118–4129.
  27. Ajay and Murcko, M.A., *Computational methods to predict binding free energy in ligand-receptor complexes*, J. Med. Chem., 38 (1995) 4952–4967.
  28. Gilson, M.K., Given, J.A., Bush, B.L. and McCammon, J.A., *The statistical thermodynamic basis for computation of binding affinities: a critical review*, Biophys. J., 72 (1997) 1047–1069.
  29. Lybrand, T.P., *Ligand-protein docking and rational drug design*, Curr. Op. Str. Biol. 5 (1995) 224–228.
  30. Checa, A., Ortiz, A.R., de Pascual Teresa, B. and Gogo, F., *Assessment of solvation effects on calculated binding affinity differences: Trypsin inhibition by flavonoids as a model system for congeneric series*. Med. Chem., in press.
  31. Mattos, C. and Ringe, D., *Multiple binding modes*. In 3D QSAR in Drug design: Theory, methods and applications, Kubinyi, H., (Ed.) ESCOM. Leiden 1993, pp. 226–254.
  32. Kraulis, P.J., *MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures*, J. Appl. Crystallog., 24 (1991) 946–950.

# Receptor-Based Prediction of Binding Affinities

Tudor I. Oprea<sup>a\*</sup> and Garland R. Marshall<sup>b</sup>

<sup>a</sup>*Department of Medicinal Chemistry, ASTRA Hässle AB, S-43183 Mölndal, Sweden*

Fax # +46 31 776 3710; E-mail tudor.oprea@hassle.se.astra.com

<sup>b</sup>*Center for Molecular Design, Washington University, St. Louis, MO 63130, U.S.A.*

## 1. Introduction

The pharmaceutical industry aims at the therapeutic manipulation of macromolecular targets, collectively termed receptors, using specific ligands (drugs). These receptors are macromolecules specialized in recognizing a specific molecular pattern from the large number of surrounding molecular species with which it could interact. Several categories of macromolecules are included:

1. *pharmacological receptors*: macromolecular complexes that can be activated by specific signal molecules (e.g. agonists), process followed by a specific biological response from the cell (organ) associated with these complexes;
2. *enzymes: proteins* that catalyze specific biochemical reactions upon substrate (ligand) binding with ( $\pm$  high) specificity;
3. *antibodies*: macromolecules (receptors) that  $\pm$  specifically bind antigens (haptens), then activate cellular (immune) responses in the presence of these antigens;
4. *DNA*: nucleic acid chains that ( $\pm$ ) specifically bind drugs (e.g. antiviral or anti-mitotic) designed to block DNA replication.

A major task for today's medicinal chemistry units is to reduce research costs, since on the average one in 10 000 screened compounds may reach the market. In the past decade, efforts to reduce these costs have relied on computational and, recently, combinatorial chemistry. Computational chemists use (largely) theoretical methods and tools implemented as software to design novel compounds with optimized biological properties for a particular therapeutic target. Because vast numbers of candidate compounds can be virtually generated in the computer (Fig. 1), an important tool in the computational chemist's arsenal has become the prediction of the binding affinity of these virtual compounds to the given receptor.

Once a lead compound has been obtained via screening, medicinal chemists generate congeneric compounds, which aim at preserving the same scaffold while replacing key pharmacophoric [1] features with isosteric groups [2–7]. Van Drie et al. [5] have described a program ALADDIN for the design or recognition of compounds that meet geometric, steric or substructural criteria. whereas DOCK, a cavity-matching algorithm [8–11] has been quite successful in finding non-congeneric molecules of the correct shape to interact with a receptor [12,13]. Caflisch et al. [14] have used a fragment-based approach to map high-affinity sites on HIV-1 protease [15] with the idea of

---

\*To whom correspondence should be addressed.

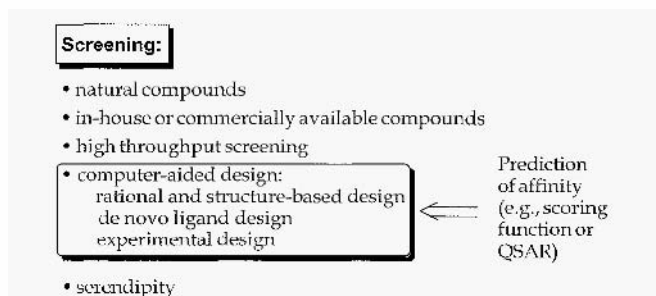


Fig. 1. The importance of affinity prediction in computational chemistry.

synthetically joining these fragments to create high-affinity ligands. A conceptually similar, experimentally based method, was used by Fesik et al. [16], who determined the relative binding modes of two low-affinity ligands by NMR, then connected them to create a high-affinity lead in their ‘SAR by NMR’ approach.

The importance of altering chemical properties of a given molecular scaffold has been recognized and used in a number of methods that are capable of generating novel structures in the computer [17–33]. All of these approaches attempt to help the medicinal chemist discover novel compounds which will be recognized at a given receptor and have proven successful in applications to HIV protease [34–39]. Micromolar leads were found based on haloperidol [10], coumarin analogs [40,41], cyclic ureas [39], and a variety of other structures. Analogs of the coumarin and cyclic urea leads which have been optimized for affinity and bioavailability have been advanced toward clinical trials.

We have recently reviewed [42] these methods in the context of affinity prediction using 3D QSAR methods. *De novo* design methods, docking and other molecular modelling techniques — e.g. computational combinatorial chemistry [43] — must handle large numbers of structures, typically in the thousands, and therefore need a sorting procedure in order to rank these virtual molecules. This is typically performed using a scoring function which evaluates the binding affinity by estimating the ligand–receptor free energy of binding,  $\Delta G_{bind}^{L-R}$ . Scoring functions [44–48] are empirical paradigms that estimate, in the molecular mechanics approximation (or some other framework that estimates non-covalent interactions), the intermolecular energies of the electrostatic and van der Waals (vdW) forces, including a hydrogen-bonding term, some estimate of the entropic and desolvation costs, as well as hydrophobic interactions. The use of free-energy perturbation (FEP) methods [49] to calculate the  $\Delta\Delta G$  of binding is not yet practical as it is limited to minor modifications of compounds with known activity and requires significant computational resources, limiting their applicability.

An intermediate method between FEP methods and scoring functions has been proposed by Åqvist and co-workers [50–51], in their linear interaction energy approach, which averages the interactions between the inhibitor and its surroundings using molecular dynamics for the bound and free states only. Another intermediate method computes the standard free energy of binding for the predominant states [52] by first

identifying the major minima of the potential energy and solvation energy functions, then evaluating their contribution to the configuration integrals of the ligand, the receptor and their complex, using a conformational energy search method [53]. This approach has been successfully applied to cyclic urea inhibitors [39] binding to HIV-1 protease [52].

The molecular mechanics interaction energy [54] has been used for a series of HIV-1 protease inhibitors with limited diversity, whereas 3D QSAR methods have been more suitable for a series with higher diversity [55–57]. Such models could also be used to evaluate software-generated compounds of diverse scaffolds for synthetic prioritization. As most of these approaches are reviewed elsewhere in this volume, we focus on the methods that use the receptor's 3D structure to derive the scoring function.

## 2. Scoring Functions: General Comments

The calculation of the free energy of binding is based on the linear free energy relationship (LFER) Formalism that relates  $\Delta G_{bind}^o$ , the standard free energy of binding, to the logarithm of the dissociation constant,  $-\log K_D$ , at thermodynamic equilibrium concentrations of the ligand,  $[L]$ , receptor,  $[R]$ , and the corresponding  $[L-R]$  complex, for the reaction  $L + R \rightarrow L-R$ :

$$\Delta G_{bind}^o = -RT \ln \frac{[L][R]}{[L-R]} = -RT \ln K_D = 2.303RT pK_D \quad (1)$$

where  $R = 8.314 \text{ J/mol/K}$ , and  $T$  is the temperature. For  $T = 310 \text{ K}$ ,  $2.303RT = 5.936 \text{ kJ/mol}$ . At  $37^\circ\text{C}$ ,  $\Delta G_{bind}^o$  is approx.  $6.0 * pK_D$  (kJ/mol), or  $1.42 * pK_D$  (kcal/mol). In fact, we model  $\Delta G_{bind}^{L-R}$ , but we assume the same reference state, so we often substitute  $\Delta G_{bind}^o$  with  $\Delta G_{bind}^{L-R}$ . The binding of a ligand to a receptor, a multi-step process, evaluated by the total  $\Delta G_{bind}^o$  (Eq. 1), includes sequential steps going from independent ligands and receptors in the surrounding physiological environment to the ligand–receptor complex. Any accessible conformation can, in principle, change into the active one during this process. Entropy is lost during reversible or irreversible binding and gained in the desolvation process because of the waters freed from the binding site and the ligand's hydration shell. Hydrophobic forces typically play an important role. One has to ascertain that no late-limiting steps occur during intermediate stages, and that non-specific binding does not obscure the experimental binding affinity. All the parameters that cannot be directly measured remain hidden (e.g. receptor-induced conformational changes of the ligand, geometric variations of the binding site, etc.), and the scoring functions use a time-sliced (frozen) model (i.e. the system is at equilibrium and time-independent). Kinetic bottlenecks in the intermediate steps may occur, and  $\Delta G_{bind}^{L-R}$  remains thermodynamic in nature (not kinetic).

Several assumptions are made when deriving a scoring function, since binding free energy is, for most cases, approximately calculated [58]:

1. The modelled compound, not its metabolite(s) or any of its derivatives, produces the observed effect.

2. The proposed (modelled) geometry is, with few exceptions, considered rigid for the receptor, and modelled as a single (bioactive) conformation for the ligand which exerts, in this single conformation, the binding effects; the dynamic nature of this process, as shown for lactate dehydrogenase, which is likely to assume different conformational states at the binding site [59], is typically ignored.
3. The loss of translational and rotational entropy upon binding is assumed to follow a similar pattern for all compounds; an additional entropic cost is considered for freezing the single-bond rotors.
4. The binding site is the same for the modelled compounds.
5. The binding free energy,  $\Delta G_{bind}^{L-R}$ , is largely explained within a molecular mechanics framework, and is prone to the inherent errors of the force field.
6. The on-offrate is similar for modelled compounds (i.e. the system is considered to be at equilibrium), and kinetic aspects are usually not considered.
7. Solvent effects, temperature, diffusion, transport, pH, salt concentrations and other factors that contribute to the overall  $\Delta G_{bind}^{L-R}$ ; are not considered.

Quite frequently the binding free energy is expressed as the sum of the free energy components, conceptually shown in the master equation (Eq. 2) [58]:

$$\Delta G_{bind}^{L-R} = \Delta G_{sol} + \Delta G_{conf} + \Delta G_{int} + \Delta G_{motion} \quad (2)$$

which accounts for contributions due to the solvent ( $\Delta G_{sol}$ ), to conformational changes in both ligand and protein ( $G_{conf}$ ), to the ligand–protein intermolecular interactions ( $\Delta G_{int}$ ) and to the motion in the ligand and protein once they are at close range ( $\Delta G_{motion}$ ). The master equation (Eq. 2) can also be written as:

$$\Delta G_{bind}^{L-R} = \Delta G_{sol} + \Delta U_{vac} - T\Delta S_{vac} \quad (3)$$

where  $\Delta G_{bind}^{L-R}$  is separated at equilibrium into solvation effects ( $\Delta G_{sol}$ ), and two components for the process in vacuum: the internal energy ( $\Delta U_{vac}$ ) and entropy ( $T\Delta S_{vac}$ ).  $\Delta G_{sol}$  can be calculated with a variety of methods [58], while  $T\Delta S_{vac}$  is often related to the number of non-methyl single bonds [47]. Both  $\Delta G_{sol}$  and  $T\Delta S_{vac}$  are assumed to have similar values for congeneric series, hence Eq. 3 is widely used in QSAR studies by expanding only the internal energy term:

$$\Delta U_{vac} = \Delta U_{vdW}^{L-R} + \Delta U_{coul}^{L-R} + \Delta U_{distort}^L + \Delta U_{distort}^R + \Delta U_{conf}^R \quad (4)$$

which includes the steric (*vdW*) and electrostatic (*coul*) aspects of the ligand–receptor interaction ( $\Delta U_{vdW}^{L-R}$  and  $\Delta U_{coul}^{L-R}$ ), the distortions (*distort*) induced by this interaction in both ligand and receptor ( $\Delta U_{distort}^L$  and  $\Delta U_{distort}^R$ ) and the ligand-induced conformational changes of the receptor ( $\Delta U_{conf}^R$ ).  $\Delta U_{conf}^R$  represents agonist-induced conformational rearrangements of the receptor that may be an important component of signal transduction and are not considered to occur upon antagonist binding to the same receptor [60].

The receptor-based scoring functions compute, using different approximations, the terms of the master equation (Eq. 2), using Eq. 1 to convert the binding-affinity data into



its free energy equivalent. We focus here on some of the recent scoring functions that aim at predicting small ligands, and not proteins. For a thorough review of methods that predict the binding free energy, the reader is referred to the work of Ajay and Murcko [58].

### 3. The LUDI Scoring Function

This regression-based scoring function [44] aims to predict fast and accurate binding affinities for *de novo* designed ligands generated by the LUDI [23] program. This approach estimates  $\Delta G_{bind}^{L-R}$  by approximating the contributions for hydrogen bonding, for entropy due to frozen rotatable bonds due to binding and for desolvation based on hydrophobic complementarity, using the following master equation:

$$\Delta G_{bind}^{L-R} = \Delta G_0 + \Delta G_{hb} \sum_{h-bonds} f(\Delta R, \Delta \alpha) + \Delta G_{ionic} \sum_{ionic\ int.} f(\Delta R, \Delta \alpha) + \Delta G_{lipo} |A_{lipo}| + \Delta G_{rot} NROT \quad (5)$$

where  $\Delta G_0$  is related to the reduction in rotational and translational entropy,  $\Delta G_{hb}$  is the free energy associated with hydrogen-bond formation,  $\Delta G_{ionic}$  is the binding energy from ionic interactions,  $\Delta G_{lipo}$  is the lipophilic interaction contribution and  $\Delta G_{rot}$  is the energy loss by freezing the internal degrees of freedom in the ligand. A penalty function,  $f(\Delta R, \Delta \alpha)$ , is introduced to track large deviations from the ideal hydrogen-bond:

$$f(\Delta R, \Delta \alpha) = f_1(\Delta R) f_2(\Delta \alpha)$$

where,

$$f_1(\Delta R) = \begin{cases} 1 & \Delta R \leq 0.2 \text{ \AA} \\ 1 - \frac{(\Delta R - 0.2)}{0.4} & \Delta R \leq 0.6 \text{ \AA} \\ 0 & \Delta R > 0.6 \text{ \AA} \end{cases}$$

$$\text{and } f_2(\Delta \alpha) = \begin{cases} 1 & \Delta \alpha \leq 30^\circ \\ 1 - \frac{(\Delta \alpha - 30)}{50} & \Delta \alpha \leq 80^\circ \\ 0 & \Delta \alpha > 80^\circ \end{cases}$$

and  $\Delta R$  is the deviation of the hydrogen-bond length  $H \cdots O/N$  from the ideal 1.9 Å value, and  $\Delta \alpha$  is the deviation of the hydrogen-bond angle  $N/O-H \cdots O/N$  from its ideal 180° value. As defined, this function tolerates small deviations of up to 0.2 Å and 30° from the ideal geometry.

Böhm [44] analyzed 45 protein–ligand complexes (affinity range = –9 to –76 kJ/mol) taken from the Protein Data Bank (PDB) [61], and found the following equation by multiple-regression analysis ( $r^2 = 0.76$ ,  $S = 7.9$  kJ/mol) and cross-validation ( $q^2 = 0.696$ ,  $S_{press} = 9.3$  kJ/mol = 2.2 kcal/mol):

$$\Delta G_{bind}^{L-R} \text{ (kJ/mol)} = 5.4\Delta G_0 - 4.7\Delta G_{hb} - 8.3\Delta G_{ionic} - 0.17\Delta G_{lipo} + 1.4\Delta G_{rot} \quad (6)$$

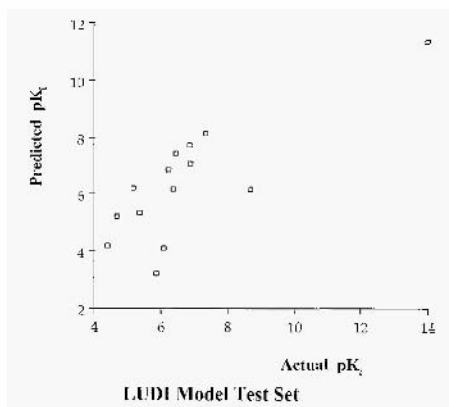


Fig. 2. predicted versus experimental  $pK_i$  value for fourteen inhibitors (see reference [44]).

The training set covered 12 orders of magnitude in binding affinity, and ligand weights ranging from 66 to 1047 daltons. The training set included 9 crystalline ligand–receptor complexes and 5 ligands docked to DHFR (see Fig. 2). This function was primarily aimed at small ligands and developed with the speed factor in mind (hence, just five adjustable parameters). Both VALIDATE [47] and Jain’s scoring function [48] have been inspired by this scoring function, as reflected throughout this chapter.

#### 4. The Wallqvist Scoring Function

Wallqvist et al. proposed a knowledge-based potential based on inter-atomic contact preferences between ligand and receptor atoms [45]. This model was parameterized by an analysis of 38 high-resolution protein crystal complexes taken from PDB [61]. For these ligand–protein complexes, molecular surfaces have been generated using the Connolly algorithm [62] and Bondi vdW radii [63]. The interface surface area for each atom was then catalogued, using a packing score,  $S^{ab}$ , calculated for all  $ab$  pairs with  $d_{ab} \leq 2.8$  Å:

$$S^{ab} = \frac{1}{d_{ab} |\theta(r_a, r_b) - \pi|} \quad (7)$$

where  $d_{ab}$  is the distance between surface elements  $a$  and  $b$ , and  $\theta(r_a, r_b)$  is the angle formed by the surface normal vectors  $r_a$  and  $r_b$  at these points. For each surface element, the area was totaled for each atom pair and used to determine the atom–atom preference score,  $P_{ij}$ , as the ratio of the total interfacial area contributed by each atom pair.  $F_{ij}$ , normalized by the product of the fractional contribution of each atom in the pair.  $F_i$  and  $F_j$ :

$$P_{ij} = \frac{F_{ij}}{F_i F_j} = \frac{A_{ij} / A_{tot}}{(A_i / A_{tot})(A_j / A_{tot})} \quad (8)$$

where the highest- and lowest-scoring preference,  $P_{ij}$ , were used to identify the most and least-observed adjacent atomic surfaces in the dataset.

$$\Delta G_{pred} = - \sum_{i \in A} \sum_{j \in B} \alpha_{ij} A_{ij} + \beta = - \sum_{i \in A} \sum_{j \in B} (\gamma + \delta \ln P_{ij}) A_{ij} + \beta \quad (9)$$

where  $\alpha_{ij}$  is the sum performed for all atoms  $i$  of molecule  $A$  and  $j$  of molecule  $B$ . Due to the paucity of data used,  $\alpha_{ij}$  is assumed to have two physical components: one that is directly proportional to the surface area, and independent of the surface atom type, and another that is specific to the buried atom. The coefficients  $\gamma$  and  $\delta$  are obtained by minimizing  $|\Delta G_{pred} - \Delta G_{exp}|$  for the 38 enzyme-inhibitor complexes whose  $\Delta G_{bind}^{L-R}$  was calculated from their dissociation constant via Eq. 1. The energy range was between  $-18$  and  $-12$  kcal/mol, and the linear fit RMS deviation was 1.5 kcal/mol ( $R = 0.74$ ). The regression analysis yields the following scoring function

$$\Delta G_{bind}^{L-R} = - \sum_{i \in A} \sum_{j \in B} \left( 10.3 \frac{\text{cal}}{\text{mol } \text{\AA}} + 30.1 \frac{\text{cal}}{\text{mol } \text{\AA}} \ln P_{ij} \right) A_{ij} + 2.39 \frac{\text{kcal}}{\text{mol}} \quad (10)$$

where  $\text{amin} = -59.0$  cal/mol/Å.

The fit, while not excellent, clearly establishes a connection between atom–atom–interfacial contacts and the  $\Delta G_{bind}^{L-R}$ . The  $|\Delta G_{pred} - \Delta G_{exp}|$  error can be attributed, besides the inherent small dataset problem, to the inter- and intra-experimental errors that occur upon measuring  $\Delta G_{exp}$ . The authors further constructed an average interfacial binding parameter  $\bar{\alpha}_i$  for each atom type in their set, largely hydrophobic in nature. Its analysis revealed the range and details of specific interactions that are deemed important at protein interfaces. This scoring function empirically estimates the buried surface of a ligand–receptor complex and is capable of specifically identifying atom–atom and residue–residue preferences in such complexes. Some prior modelling efforts are needed — e.g. docking of the ligand in the binding site and some conformational analysis. Chemical constructions based on a given optimized surface are still an open issue.

## 5. The Verkhivker Scoring Function

This approach [46] is a knowledge-based interaction potential based on Sippl’s [64] knowledge-based approach, and parameterized for HIV-1 protease [15] ligand–receptor complexes using the following master equation (Eq. 11)

$$\begin{aligned} \Delta G_{bind}^{L-R} = & \Delta G^{L-R \text{ interaction}} + \Delta G^{L-wat \text{ interaction}} + \Delta G^{R-wat \text{ interaction}} + \Delta G^{\text{nonpolar desolvation}} \\ & + \Delta G^{\text{polar desolvation}} + \Delta G^{L \text{ isomerization}} + \Delta G^{L-R \text{ rot, translation}} \\ & + \Delta G^{\text{wat rot, translation}} - T\Delta S^L \text{ conformation} - T\Delta S^R \text{ conformation} \end{aligned} \quad (11)$$

where  $\Delta G^{L-R \text{ interaction}}$  is further subdivided in  $\Delta G^{n-n}$  (the interaction between nonpolar groups only) and  $\Delta G^{n-p-p-p-n}$  (the interaction between polar and nonpolar groups). The

reader is referred to the original work [46] for a complete description of each term in the above equation: here we just give the ligand–receptor pairwise interaction potentials as an example.

Using a set of 30 pairs of ligand–protein complexes of HIV-1, HIV-2 and SIV proteases, a set of distance-dependent interaction potentials was derived for 12 atom pairs relevant for protein–ligand interactions, using Sippl's [64] original approach and the CHARMM\_19 vdW radii [65]. for a given atom pair  $a$  (ligand) and  $b$  (receptor) separated by a distance  $d$

$$\Delta G^{ab}(d) = RT \ln(1 + m_{ab}\sigma) - RT \ln\left\{1 + m_{ab}\sigma \left[\frac{f^{ab}(d)}{f(d)}\right]\right\} \quad (12)$$

where  $m_{ab}$  is the number of pairs with ligand atom of type  $a$  and receptor atom of type  $b$ ;  $\sigma$  is the weight given to each observation;  $f^{ab}(d)$  is the frequency with which this pair of atoms is observed at interatomic distance  $d$ ; and  $f(d)$  is the total number of atom pairs of all types that are separated by distance  $d$ . In this work [46]. the probabilities of observing a particular distance were computed. normalized by the frequencies observed for all types of  $L-R$  atom pairs in the training set. then translated into mean force potentials. To minimize the effect of low-occurrence data points in the parameterization, a cross-validation procedure was applied by eliminating the evaluated complex from the training set for the empirical  $\Delta G_{bind}^{L-R}$  calculations of the seven studied HIV-1 protease inhibitor complexes [46]. The authors further break down the contributions of various terms from their master equation (Eq. 11) for the studied inhibitors, providing a rationale for the variation in binding affinity of these ligands. The enthalpy–entropy compensation effect [66] appears to be supported by their model, which provides a reasonable estimate for the absolute binding free energy (Fig. 3).

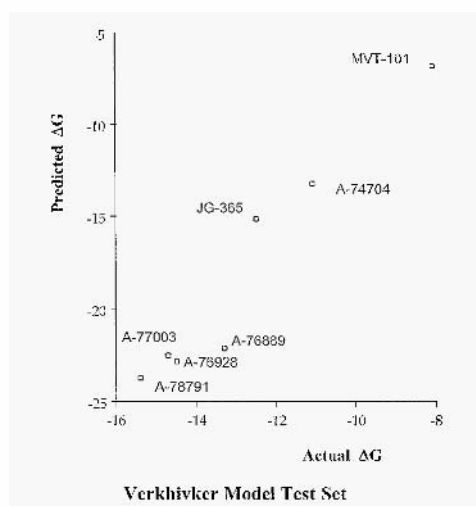


Fig. 3. Calculated versus experimental binding free energies for seven HIV-1 protease inhibitors (see reference [46]).

## 6. VALIDATE

VALIDATE [47] is a hybrid approach explicitly based on (and dependent on) a QSAR training set paradigm that calculates physico-chemical properties of both ligand and  $L-R$  complex to estimate  $\Delta G_{bind}^{L-R}$ , based on the following master equation (Eq. 13):

$$\Delta G_{bind}^{L-R} = \beta_1 E_{vdW}^{L-R} + \beta_2 E_{coul}^{L-R} + \beta_3 SF + \beta_4 H \log P + \beta_5 N r_{rot.bonds} + \beta_6 \Delta H_{bind}^L + \beta_7 CSA_{lipophilic}^{L-R} + \beta_8 CSA_{hydrophilic}^{L-R} \quad (13)$$

where  $\beta_1 - \beta_8$  are the fitted regression coefficients for the master equation terms (Eqs. 14–20), clarified below.

### 6.1. The steric and energetic intermolecular interactions and the steric fit

Structural complementarity is essential for the specific binding of a ligand to the receptor at the binding site. The non-bonded steric interaction energy is computed from the explicit sum of the Lennard-Jones potentials:

$$E_{vdW}^{L-R} = \sum_i^L \sum_j^R \varepsilon_{ij} \left( \frac{1}{R_{12}} - \frac{2}{R_6} \right) \quad (14)$$

where

$$\begin{aligned} \varepsilon_{ij} &= \sqrt{\varepsilon_i \varepsilon_j} \\ R_{12} &= \left( \frac{r_{ij}}{R_i + R_j} \right)^{12} \\ R_6 &= \left( \frac{r_{ij}}{R_i + R_j} \right)^6 \end{aligned}$$

and  $r_{ij}$  is the distance between atom center  $i$  and atom center  $j$ ,  $R_i, \varepsilon_i$  is the vdW radius, epsilon value of atom  $i$ , and  $R_j, \varepsilon_j$  is the vdW radius, epsilon value of atom  $j$ .

The electrostatic interaction energy is the explicit sum of the Coulombic potentials

$$E_{coul}^{L-R} = \frac{1}{4\pi\varepsilon_0} \sum_i^L \sum_j^R \frac{q_i q_j}{r_{ij}} \quad (15)$$

using partial atomic charges on the ligand ( $q_l$ ) and the receptor ( $q_r$ ) from the implementation of the Amber force field within the MacroModel program.

Additionally, we have included a steric complementarity fit ( $SF$ ) to describe the packing of a ligand in the receptor binding site. The steric fit is computed by summing the number of ‘good contacts’ for each atom of the ligand which is contained in the active site. For example, antibody–steroid complexes and chymotrypsin-binding ligands actually contain a considerable percentage of their atoms outside of the active site,

while HIV protease inhibitors are largely surrounded by the receptor. VALIDATE defines a good contact as an instance where the vdW surface of a ligand atom is within a modifiable parameter,  $\epsilon$ , of the vdW surface of a receptor atom:

$$SF = \frac{\sum_{i,j} C_{ij}}{N} \quad (16)$$

where

$$C_{ij} = \begin{cases} 1, & d_{ij} \leq |r_i + r_j \pm \epsilon| \\ 0, & d_{ij} > |r_i + r_j \pm \epsilon| \end{cases}$$

and  $N$  is the number of ligand atoms contained in the active site:  $r_i, r_j$  is the van der Waals radii of atoms  $i$  and  $j$ , and  $d_{ij}$  is the distance from atom center  $i$  to atom center  $j$ . For  $SF$ , we investigated  $\epsilon^*$  values ranging from 0.1 Å to 0.3 Å, but we reported results for  $\epsilon^* = 0.3$  Å only. The steric fit is also normalized by the number of ligand atoms within the active site.

## 6.2. Ligand transfer from solution into the binding site

The lipophilicity of a hydrophobe is estimated by the energy needed to create a cavity in the aqueous solvent in which that molecule can fit. When hydrophobic ligands bind to the receptor, the energy of cavity creation is released, entropically favoring the process.  $\log P$ , the partition coefficient for octanol–water [67], gauges the ligand's preference for the active site of the receptor versus the aqueous solvent. We use the fragment-based  $H \log P$  method in Hint 1.1 [68] to compute the ligand's partition coefficient. With the partition coefficient, a negative value indicates a preference for a polar (hydrophilic) environment and a positive value indicates a preference for a nonpolar (lipophilic) environment. Based on direct observations of the nature of the binding site, for example, for HTV-1 protease (predominantly lipophilic) and L-arabinose sugar-binding protein (predominantly hydrophilic), we compute the amount of hydrophilic and lipophilic surface area as ratios to the total surface area of the receptor active site. The final value of the partition coefficient is then modified based on this information:

$$PC = RC * HlogP\_PC \quad (17)$$

where  $HlogP\_PC$  is the partition coefficient as computed by Hint 1.1, and  $RC = 1$  if the receptor active site is predominantly lipophilic and  $-1$  if the receptor active site is predominantly hydrophilic.

The sign of the coefficient  $RC$  is determined using the sums of the lipophilic and hydrophilic surface areas of the active site. For a hydrophilic receptor active site with  $RC = -1$ , one of the following criteria must be true:

1. If less than five ligands are available for this active site, all calculations must yield that at least 55% of the total surface of the active site is hydrophilic.

2. If five or more ligands are available, at least half of the calculations must yield that 55% or more of their total surface area is hydrophilic, while the remaining calculations must yield that at least a majority of the total surface area is hydrophilic. This decision is taken by determining the hydrophobic/hydrophilic preference of the active site based on surface area calculations [69] (see section 6.4), defining surface types in the same way as Böhm [44]: any carbon that is covalently bound to no more than one non-carbon is considered lipophilic, and any hydrogen connected to such a carbon is also lipophilic. All other atoms are considered hydrophilic. For protein-protein systems, a different treatment for this calculation was performed, as a large portion of the protein inhibitor remains freely accessible to the solvent and is not bound at the active site. Since only the active site, a small portion of the protein ligand, is desolvated by binding to the receptor, only the  $H\log P$  for this region is relevant. Therefore, this region was extracted from the protein and the calculation was done only on this part of the molecule.

### 6.3. Conformational entropy and enthalpy

Changes in conformational entropy occur when the freely rotating side chains of the dissociated components are forced to adopt more rigid conformations on complex formation. VALIDATE estimates the change in conformational entropy by counting the number of rotatable bonds.  $Nr_{rot. bonds}$ . All non-terminal single bonds (except methyl groups) are included. For non-aromatic ring systems, the number of degrees of freedom is of the order  $n - 4$ , where  $n$  is the number of bonds in the ring. The rotatable bond count is:

$$Nr_{rot. bonds} = Nr_{ntsb} + \sum_i (n_i - 4) \quad (18)$$

where  $Nr_{ntsb}$  is the number of non-terminal single bonds, and  $n_i$  is the number of single bonds in ring  $i$ . As for the  $H\log P$  calculations, an exception was made for protein-protein systems, because a large portion of the protein inhibitor is not bound and remains freely accessible to the solvent. VALIDATE counts only the rotatable bonds at the active site interface.

The change in conformational enthalpy,  $\Delta H_{bind}^L$ , is approximated here by the amount of energy required for the ligand to adopt the receptor-bound conformation, defined by:

$$\Delta H_{bind}^L = |E_{bind, site}^L - E_{sol}^L| \quad (19)$$

where  $E_{bind, site}^L$  is the energy of the ligand's receptor-bound conformation, and  $E_{sol}^L$  is the energy of the ligand in solvent at its nearest local minimum, calculated by comparing the energy of the receptor-bound conformation of the ligand to the nearest local minimum of the unbound ligand using the GB/SA [70] solvation model with the Amber all-atom force-field implementation in MacroModel.

#### 6.4. Contact surface areas: the ligand–receptor interface

VALIDATE computes four components to surface complementarity. These are lipophilic complementarity (nonpolar/nonpolar), hydrophilic complementarity (polar/polar, opposite charge), lipophilic/hydrophilic (polar/nonpolar) non-complementarity, and hydrophilic (polar/polar, like charge) non-complementarity. VALIDATE uses 256 evenly distributed points, obtained from the SASA program [69], placed on the vdW surface of each receptor atom whose vdW surface is within 5 Å of the atom center of any ligand atom. If a point on this surface is within a mean solvent radius (1.4 Å for water) of the vdW surface or a ligand atom, it is considered a contact point. Its type is based upon the determination of the polar/nonpolar nature of both atoms and the criteria discussed above.

VALIDATE has two different types of CSA calculation. In the first method, the absolute surface area between ligand and receptor (similar to Bohm [44]) is computed using the type of contact observed for each point on the receptor surface. Lipophilic complementarity is counted only once, even if that point is within the distance limit described above of more than 1 ligand atom's vdW surface. The total surface on each atom for each type of contact is computed by dividing the number of contact points of that type by 256 (the total number of points possible) and then multiplying by the total surface area of the atom:

$$CSA = \sum_i^R \frac{(4\pi r_i^2 * CP_i)}{256} \quad (20)$$

where  $CP_i$  is the number of contact points on atom  $i$  and  $r_i$  is the vdW radius of atom  $i$ .

The second method is a pairwise summation, similar to Hint 1.1 [68]. For a single point of surface contact of a receptor atom within the distance of a ligand, we record a sum of  $n$ , instead of 1, as described above.

VALIDATE included both descriptors, as their combination improved both the fitted model and its predictivity. Due to the structural diversity in the training set, we observed that different receptor–ligand pairs had different CSA values. We, therefore, scaled the computed surface areas by dividing the total surface area of HIV-1 protease (chosen as the most representative receptor in the dataset) to the largest total surface area computed for a given receptor's active site for any of its known ligands. CSA values were then multiplied with the scaling factor.

#### 6.6. VALIDATE results

VALIDATE was trained on 51 receptor–ligand co-crystallized structures [47] available from PDB [61] (see Table 1). This training set included 15 HIV-1 protease–inhibitor, 9 thermolysin–inhibitor, 12 endothiapepsin–inhibitor, 8 L-arabinose binding protein–sugar, 4 antibody–steroid, 4 subtilisin–Novo–protein and 2  $\beta$ -trypsin–protein complexes, respectively. Ligand size ranged from 24 atoms (Leu-NHOH) to 1512 atoms (SSI M70G M73K), and the  $pK_i$  was between 2.47 and 14.0. The crystal-structure based test set (Table 2) included 14 inhibitors which were obtained from PDB. Neither ligands



Table 1 Receptor–ligand list for the VALIDATE training set

Protein–inhibitor	PDB code	$pK_i$
HIV – AG1001	N/A	6.44
HIV – AG1002	N/A	6.26
HIV – AG1004	N/A	6.64
HIV – Roche	N/A	6.91
HIV – MVT101	4hvp	6.12
HIV – SC52964	N/A	10
HIV – JG365	7hvp	9.62
HIV – Acetylpepstatin	5hvp	7.85
HIV – GRI 16624X	N/A	8.27
HIV – U75875	1hiv	9
HIV – L-689,502	N/A	8.95
HIV – A74704	9hvp	8.35
HIV – A77003	1hvi	10.02
HIV – Hydroxypethylene	laaq	7.74
HIV – L-700,417	4phv	9.15
Thermolysin – Phosphoramidon	1tlp	7.55
Thermolysin – N-(1 carboxy-3phenyl)-L-LeuTrp	1tmn	7.47
Thermolysin – N-phosphoryl-L-letricineamide	2tmn	4.1
Thermolysin – ValTryp	3tmn	5.9
Thermolysin – Lcu-NHOH	4tln	3.72
Thermolysin – ZFPLA	4tmn	10.19
Thermolysin – ZGp(NH)LL	5tmn	8.04
Thermolysin – ZGp(O)LL	6tmn	5.05
Thermolysin – CH2CO-Leu-OCH3	7tln	2.47
Endothiapepsin – PD 125754	leed	4.9
Endothiapepsin – L-364.099	2er0	6.4
Endothiapepsin – H 256	2er6	7.2
Endothiapepsin – H 261	2er7	9
Endothiapepsin – L-363,564	2er9	7.4
Endothiapepsin – CP 71,362	3er3	7.1
Endothiapepsin – PD 125967	4er1	6.6
Endothiapepsin – H 142	4er4	6.8
Endothiapepsin – CP 69,799	5er2	6.6
L-arabinose Bind. Prot. – L-arabinose	lahe	6.5
L-arabinose Rind. Prot. – D-fucose	labf	5.2
L-arabinose Bind. Prot. P254G –D-fucose	labp	5.8
L-arabinose Bind. Prot. P254G –L-arabinose	lhap	6.9
L-arabinose Rind. Prot. P254G –D-galactose	9abp	8
L-arabinose Bind. Prot. MI 08L –L-arabinose	6abp	7
L-arabinose Bind. Prot. MI OXL –D-fucose	7abp	5.4
L-arabinose Bind. Prot. M 108L –D-galactose	8abp	6.6
beta-Trypsin – UPTI	1tpa	14
Beta-Trypsin – PTI	2pte	13.3
DB3 – progesterone 11a-ol hemisuccinate	1dbm	9.44
DB3 – 5a-prepnane -3b-ol hemisuccinate	2dbl	8.7
DB3 – Etiocholanolone	1dhj	7.62
DB3 – Progesterone	1dbb	9
Subtilisin-Novo – Eglin c L45R	1sbn	10.3
Subtilisin-Novo – CI-2	2sni	11
Subtilisin-Novo – SSI M73K	3sic	10.2
Subtilisin-Novo – SSI M70G M73K	5sic	10.2

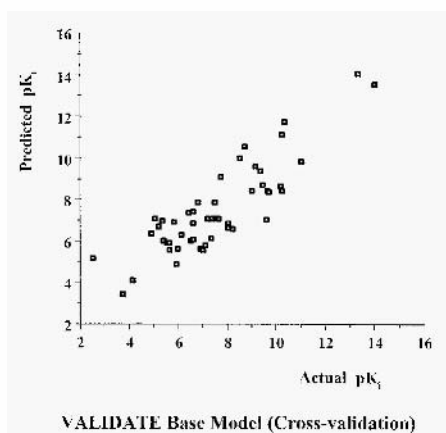


Fig. 4. Cross-validated PLS analysis of 51 complexes in the training set (see Table 1).  $q^2 = 0.776$  and standard error ( $S_{press}$ ) = 1.139 for the six-components model (similar results obtained with SONNIC, a neural network program)

Table 2 Crystal structures used as test set for the VALIDATE base model

	PDB code	$pK_i$	
		Actual	Predicted
DHFR – Folate	1dhf	7.1	7.29
DHFR – Methotrexate	1dds	8.3	6.4
Penicillipepsin – IvaVVLYSta-OEt	1apt	9.4	7.71
Penicillipepsin – Iva VVSta-OEt	1apv	7.7	8.04
Carboxypeptidase – L-benzylsuccinate	1cbx	6.3	5.92
Carboxypeptidase – GlyTyr	3cpa	4	5.14
Carboxypeptidase – LAGp(0)F	8cpa	9.1	9.39
Alpha-Thrombin – MDI. 28050	1ths	7.1	7
Alplia-Thrombin – NAPAP	1dwd	8.2	8.51
Trypsinogen – IleVal	2tpi	3.3	3.35
Trypsinogen – ValVal	4tpi	2.0	3.39
DNA – Daunomyein	1da0	6.5	6.1
DNA – Netropsin	121d	8.8	9.59
DNA – 4-6-diamidine-2-phenyl indole	1d30	6.3	5.04

nor the specific receptors in this test set were included in the training set. Included were 2 DHFR, 2 penicillipepsin, 3 carboxypeptidase, 2  $\alpha$ -thrombin, 2 trypsinogen and 3 DNA complexes. Actual versus predicted affinities are shown in Fig. 4 (cross-validation results on the training set) and Fig. 5 (test set), respectively.

An extensive discussion of the VALIDATE results was presented in the initial publication of this work. In the base model, the electrostatic interaction energy contributed only 2.7% to the final PLS model which may, in part, be explained by the inadequate approximations of the partial charge representation in our model derivation. The steric

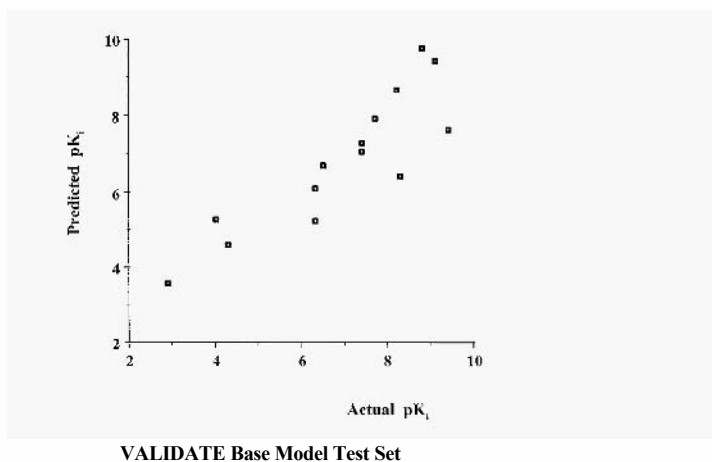


Fig. 5. Prediction of affinities of 14 crystalline complexes using the VALIDATE hose model: predictive  $r^2 = 0.806$  and absolute average error = 0.697.

fit parameter also failed to give a significant contribution to the overall model, but it proved to be a major contributor in a series of 22 steroids binding to DB3, the monoclonal antibody against progesterone [71]. In this particular case, the surface of the steroid ligands is approx. 45% exposed to the aqueous solvent, therefore a significant number of ligand atoms are not in the binding site. Several limitations of VALIDATE are addressed in VALIDATE II, discussed below.

#### 6.7. VALIDATE II: a scoring function for predicting hiv-1 protease inhibitors

Compared to the LUDI scoring function [44] (see Eq. 5), VALIDATE did not explicitly consider hydrogen-bonds (Eq. 13), as initial attempts to incorporate this parameter into

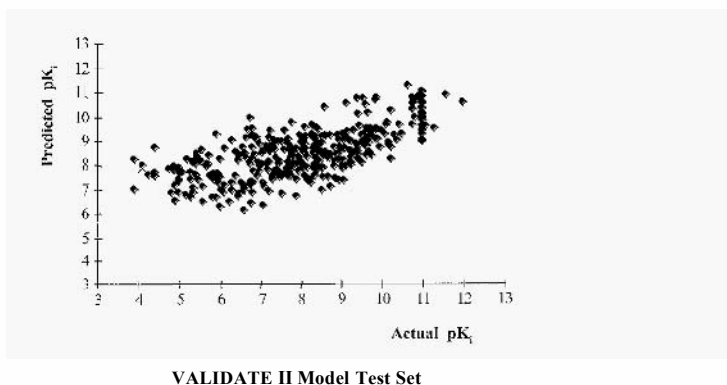


Fig. 6. Prediction of affinities of 363 HIV-1 protease inhibitors using the VALIDATE II model: predictive  $r^2 = 0.48$  and absolute average error = 1.05.

Table 3 Receptor–ligand list for the VALIDATE II training set

Protein–inhibitor	PDB code	$pK_i$
HIV – A79285	1dif	10.66
HIV – SB203238	1hbv	6.37
HIV – SKF108738	1hef	8.8
HIV – SKF107457	1heg	7.93
HIV – CGP53820	1hih	5.53
HIV – SB204144	1hos	8.55
HIV – SB206343	1hps	9.22
HIV – VX477	1hpv	9.22
HIV – KNI272	1hpx	8.22
HIV – L735524	1hsg	9.42
HIV – GR123976	1hte	7
HIV – GR126045	1htf	8.19
HIV – GR137615	1htg	9.78
HIV [GGSSG linked] – A76928	1hvc	10.96
HIV – A78791	1hvj	11.4
HIV – A76928	1hvk	10.96
HIV – A76889	1hvl	9.95
HIV – XK263	1hvr	9.51
HIV [V82A] – A77003	1hvs	10.3
HIV – cyc(Phe-Ile-Val)	1mtr	8.4
HIV – SB203386	1sbg	7.74
HIV – U8554SE	8hvp	8
HIV – DMP323 [NMR average]	1bvg	9.57
HIV – U100313	2upj	10.39
HIV – A9888	1pro	11.3

Note: The other HIV-1 protease inhibitor complexes used in VALIDATE II are listed in Table 1.

the base model reduced its accuracy. Since hydrogen-bonds are important in the receptor–ligand interaction, however, this was reflected in VALIDATE II, a scoring function developed by Ragno and Marshall [manuscript in preparation]. VALIDATE II incorporates 15 additional descriptors, including 3 related to hydrogen bonds: the number of ideal hydrogen bonds for the ligand in water, the number of hydrogen bonds between the ligand and receptor and the difference between the above two. Other descriptors were the AMSOL-derived polarization free energy of solvation, the cavity formation free energy, the ligand’s dipole moment in solvent and the HOMO energy. These descriptors were included in an attempt to more accurately describe the electrostatic contribution to ligand binding, which was less than 3% in the VALIDATE base model.

VALIDATE II model was trained using 39 HIV-1 protease inhibitor crystallographic complexes (see Table 3) and 29 explicative variables (14 from the VALIDATE base model). The two-PC model explains 91 % of the variance, and yields a  $r^2 = 0.74$  (standard deviation of calculation, SDEC = 0.73), and a  $q^2 = 0.52$  (standard deviation of prediction, SDEP = 0.98). This was better compared to using only the original VALIDATE descriptors:  $r^2 = 0.55$  (standard deviation of calculation, SDEC = 0.96) and a  $q^2 = 0.37$  (standard deviation of prediction, SDEP = 1.12), which is not surprising considering that the number of descriptors is double in the first method. The external test set

included 363 HIV-1 protease inhibitors, which were docked into the binding site using the nearest crystalline analog as template structure, using MacroModel. The affinity range was between 6.11 and 11.4 for the test set, and 3.92 and 11.96 for the training set. Actual versus predicted affinities are shown in Fig. 6 for the 363 compounds (test set).

VALIDATE II shows a proof of concept — i.e. a scoring function that can be constructed from crystal complexes that have no ambiguity with respect to the binding mode. This function can be readily used to predict the affinities of compounds for which the binding mode is not known and has to be modelled, although in its current version the error is greater than one would like ( $\pm 1.3$  kcal/mol). It was hoped that such generic models could embody the physical chemistry of binding. The fact that the errors of prediction for the low-affinity ligands (not included in the training set) are higher than the errors of prediction for the high-affinity ligands suggests that this was not accomplished. It should be pointed out that test-set ligands with low affinity have, on average, a higher experimental error than those in the high-affinity range.

VALIDATE tested the ability of a scoring function to generalize in the case where the binding modes were experimentally defined, with associated errors from different assay procedures and no internal standards for reference — in other words, a basic test of how well we can do if we get the correct binding mode of the ligand. Here one has to deal with different errors in calculations on different receptors, which do not necessarily cancel. By contrast, VALIDATE II was focused on an individual protein, thus leading to some cancellation of errors in modelling the target. This scoring function tested the ability to evaluate ligands in the situation where the binding mode is modelled in a crude manner (with errors): ligands were docked into the binding site in an approximate manner, by similarity to the binding mode of the corresponding transition state isostere. Overall, the results show that VALIDATE II is tolerant to inaccuracies in the binding mode (albeit less accurate). It was not surprising that its predictive ability is greater than VALIDATE's, since it was specifically trained for HIV-1 protease.

The choice of parameters remains arbitrary due to significant cross-correlations between parameters. For example, in the VALIDATE II model, the HOMO energy, one of the most significant parameters, can be omitted and another model of similar predictability can be derived from the remaining parameters. It is possible that finding the correct binding mode may constitute a major drawback for the external prediction of this model. The influence of this problem was significantly reduced for VALIDATE, because we used known complexes for both the training and the test set. It may be more difficult to predict the effects of minor changes in structure than in getting the relative affinities of diverse complexes approximately right. The VALIDATE II results motivate the development of an improved scoring function to allow accurate prediction of binding modes as we believe much of the error in our predictions comes from imprecise orientations of the inhibitors in the complexes.

## **7. The Jain Scoring Function**

To overcome the problems given by incorrect orientations in the binding site, Jain proposes [48] a regression-based scoring function that is both fast and tolerant to inaccurate ligand orientations. This approach has a master equation (Eq. 21) that includes terms

tuned with neural network-based functions that include a sigmoidal ( $s$ ), a gaussian ( $g$ ) (Eq. 22), and a distance ( $d$ ) (23):

$$F = \sum_{i,j} f_0(d(i, j)) + \sum_{i,j} f_1(d(i, j), i, j) + \sum_{i,j} f_2(d(i, j), i, j) + (l_5 \cdot phbe) + (l_6 \cdot lhbe) \quad (21)$$

$$+ (l_7 \cdot n\_rot) + (l_8 \cdot \log(mol.weight))$$

$$g(x, \mu, \sigma) = e^{-(x+\mu)^2/\sigma} \text{ and } s(x, \mu) = 1/(1 + e^{10(x-\mu)}) \quad (22)$$

$$d(i, j) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2} - r_i - r_j \quad (23)$$

where  $f_0$  is the hydrophobic term,  $f_1$  is the polar contribution,  $f_2$  estimates the repulsive term, both  $(l_5, phbe)$  and  $(l_6, lhbe)$  represent solvation and  $(l_7, n\_rot)$  and  $(l_8, \log(mol.weight))$  are the entropic term, respectively. Tunable linear parameters are denoted  $l_i$ , whereas nonlinear ones are denoted  $n_i$ . In what follows (Eqs. 24 and 26–28), — we have attempted to eliminate what we believe were typos from the original paper by Jain [48]. The reader is advised to compare the two versions.

The hydrophobic complementarity :

$$f_0(x) = l_0 g(x, n_0, n_1) + l_1 s(x, n_2 + n_1) \quad (24)$$

is composed of a gaussian  $g$  — that captures the positive portion of atomic contacts, and a sigmoidal  $s$  — that captures the portion due to steric overlap.

The polar complementarity (Eq. 25) is summed over all pairs of polar atoms using the charge–charge interactions ( $c_i, c_j$ ):

$$f_i(x, i, j) = f_{1a}(x) f_{1b}(i, j) (1 + n_6 c_i) (1 + n_6 c_j) \quad (25)$$

where

$$f_{1a}(x) = l_2 g(x, n_3, n_4) + l_3 s(x, n_2 + n_4) \quad (26)$$

computes the effect of hydrogen-bonds and salt bridges, and:

$$f_{1b}(i, j) = s(-(b_{ij} v_i)(b_{ij} v_j) - n_5) \quad (27)$$

is a correction term for hydrogen-bond directionality, defined using three vectors:  $v_i$  ('out' direction for atom  $i$ ),  $v_j$  ('in' direction for atom  $j$ ) and a normalized vector  $b_{ij}$  (from atom  $i$  to atom  $j$ ), respectively. Switching atoms  $i$  and  $j$  yields the same value for this correction term, as  $f_{1b}$  is symmetric. The unfavorable charged contacts (Eq. 28) are accounted for by summing contacts for all pairs of polar atoms of the same sign:

$$f_2(x) = l_4 s(x, n_7) f_{1a}(x) f_{1b}(i, j) (1 + n_6 c_i) (1 + n_6 c_j) \quad (28)$$

Solvation effects are accounted for by computing the difference between the total and actual number of 'hydrogen-bond equivalents' for the protein ( $phbe$ ) and ligand ( $lhbe$ ),

using the tunable parameters  $l_5$  and  $l_6$ , respectively (see Eq. 21). Entropic costs are estimated by counting the number of freely rotatable bonds in the ligand ( $n_{rot}$ ) and by the log 10 of the molecular weight of the ligand ( $\log(mol.weight)$ ), using the tunable parameters  $l_7$  and  $l_8$ , respectively.

The function  $F$  was trained on 34 crystal structures from PDB [61], which included several proteases (e.g., 6 thermolysin–inhibitor, 6 trypsin–inhibitor and 5 thrombin–inhibitor complexes), two cytochrome P450–inhibitor complexes, binding proteins to fatty acids, galactose and retinol, among others. The affinity ranged between 2.82 and 14.0 on the  $pK_d$  scale. When trained using ligand minimization with pose optimization (defined as the orientation of the ligand into the binding site prior to docking),  $F$  yields a good pair rank-correlation coefficient (0.95 out of 1.0) and a good root-mean-squared error of 0.72. The  $l$  and  $n$  values were obtained after convergence (five iterations) (Table 4).

Table 4  $l$  and  $n$  values after convergence.

parameter	Value	Parameter	Value
$l_0$	0.0898	$n_0$	0.6213
$l_1$	-0.0841	$n_1$	0.1339
$l_2$	1.2338	$n_2$	0.1880
$l_3$	-0.1796	$n_3$	0.3234
$l_4$	-0.0500	$n_4$	0.6313
$l_5$	-0.1539	$n_5$	0.6139
$l_6$	0.0000	$n_6$	0.5000
$l_7$	-0.2137	$n_7$	0.5010
$l_8$	-1.0406		

Additional constraints were imposed by introducing a steric overlap term

$$K_{ij} = -10.0 (d_{ij} + \delta)^2 \quad (29)$$

where  $\delta$  was set to 0.7 for complementary polar contacts and 0.1 for others (thus, a steric overlap of 0.5 Å yields a penalty of 1.6  $pK_d$  units). This penalty was introduced for ligand orientation optimization during docking, because  $F$  was trained on native crystal structures, where unfavorable ligand–protein contacts are scarce, if any. However,  $K_{ij}$  was not included in the final score.

The expressions used in this scoring function are, in principle, a different functional form from the ‘classical’ equations from molecular mechanics (e.g. Eqs. 15 and 25 are different representations of the charge–charge interaction). However, this approach yields a result that is quite similar to VALIDATE [47], in that hydrophobic contacts (expressed as the lipophilic CSA in VALIDATE and as  $f_0$  in  $F$ ) are deemed most important (18.5% in VALIDATE and 44% in  $F$ , respectively). Electrostatic interactions rank second in importance (26%), unlike in VALIDATE (less than 3%).  $F$  compares well [48] to Böhm’s function [44], as the breakdown terms from the master equations yield similar values for the trypsin–benzamidine complex [48]. This function appears to be rapid, accurate and tolerant to ligand pose, and to produce good estimates for the binding affinity.

## 8. The HTS Approach

For a computational high-throughput screening (HTS) objective, Rose [72] defined a scoring function based on the following master equation (Eq. 30):

$$\Delta G_{bind}^{L-R} = \Delta G_{h-bond} + \Delta G_{metal\ cr.} + \Delta G_{vdW} + \Delta G_{desolv..polar} + \Delta G_{desolv..non-polar} + \Delta G_{rot.int} + \Delta G_{rot/trans} \quad (30)$$

where  $\Delta G_{bind}^{L-R}$  the hydrogen bond energy.  $\Delta G_{metal\ cr.}$  is the metal center bond energy,  $\Delta G_{vdW}$  is the Lennard-Jones interaction energy,  $\Delta G_{desolv..polar}$  is the desolvation energy for unsatisfied donor/acceptor terms.  $\Delta G_{desolv..non-polar}$  is the desolvation energy for non-polar atoms,  $\Delta G_{rot.int}$  is the entropy loss due to restrictions in internal degrees of freedom and  $\Delta G_{rot/trans}$  is the loss of rotational and translational entropy. The HTS effective hydrogen bond potential,  $\Delta G_{h-bonds}$  is defined as:

$$\Delta G_{h-bonds} = f(d_{surface}) \cdot \cos \theta \cdot f(H...A) \cdot f(D-H...A) \cdot f(A) \cdot f(D) \cdot f(q_A, q_D)$$

where the distance of an atom from the protein surface becomes essential: if  $d_{surface} > 4 \text{ \AA}$ , then  $\Delta G_{h-bond} = -0.5$  to  $-1.0$  kcal/mol. which also assigns a maximum for the  $\Delta G_{desolv..polar}$  term. If the same atom is on the surface, then  $\Delta G_{h-bond} = 0.0$  to  $-0.3$  kcal/mol, and  $\Delta G_{desolv..polar} = 0.0$ . In Eq. 31,  $H$  is the hydrogen atom,  $A$  is the acceptor atom and  $D$  is the donor,  $\theta$  is the angle for the donor out of the acceptor plane. while  $q_A$  and  $q_D$  are the partial charges for the acceptor and donor atoms, respectively. In the HTS function, aromatic carbons are treated as hydrogen-bond acceptors, whereas aromatic C-Hpairs are treated as donors. In a similar fashion the effective metal center bond potential is defined:

$$\Delta G_{metal\ cr} = f(d_{surface}) \cdot \cos^3 \lambda \cdot f(M...A) \cdot f(q_A)$$

where  $\lambda$  is the angle of the metal out of the acceptor plane.  $M$  is the metal atom,  $A$  is the coordinating acceptor atom and  $q_A$  are formal charges.

The HTS effective desolvation terms are:

$$\Delta G_{desolv..non-polar} = \sum f(d_{surface}) \cdot \Delta A_i \cdot 0.007 \text{ kcal/mol/\AA}^2$$

which is applied for all nonpolar atoms. and  $\Delta A_i$  is the change in the solvent accessible surface area upon ligand binding, and:

$$\Delta G_{desolv..polar} = \sum f(d_{surface}) \cdot \Delta A_i \cdot \sigma_i \quad (34)$$

which is applied to all unsatisfied donor and acceptor atoms, with  $\sigma_i$  as an atom-type specific solvation parameter fitted to the  $\Delta G_{sol}$  of small molecules.

The entropy terms are defined as:

$$\Delta G_{rot.int} = \sum f(d_{surface}) \cdot \Delta RT \ln(3)$$



scoring functions work at all. It is well known, for example, in the field of HIV-1 protease inhibition, that different pharmaceutical companies measure the  $K_i$  or  $IC_{50}$  values at pH values ranging from 4.5 to 7, at  $Na^+$  concentrations ranging from 0.001 M to 0.1 M. Both factors are known to influence  $K_M$  values for the protease up to 100-fold, hence significantly different  $K_i$  values are expected for the same ligand! In view of the above facts, interested parties should combine their efforts by pooling together binding affinities for all known crystal structures of a given protein, and have them tested with the same procedure, ideally at different temperatures as well (to observe the influence of enthalpy–entropy compensation and thus better calibrate the model).

It is evident that attempts to calibrate scoring functions rely upon comparison of calculated (predicted) and experimental values. As such, the experimental values remain the only reality check. With the increasing diversity of bioassays, different laboratories yield different affinity constants for the same reference compound, sometimes by as much as 3  $pK_i$  units. We, therefore, believe it is timely to propose the definition of internal standards for biological targets of major interest, both at the substrate/agonist and inhibitor/antagonist levels. Peer-reviewed scientific journals should instruct their referees to condition acceptance for publication only upon disclosure of experimental results for such compounds deemed as internal standards. Such compounds could then be used for relative potency comparisons. We expect improvements in the performance of scoring functions relying on such internal standards, thus leading to more accurate predictions. While this request may seem an overstatement, we note that two recent reviews do not even mention the need for more accurate biological assays, despite an elegant treatment of physico-chemical and computational aspects of the ligand–receptor binding [52,73].

The second requirement, speed, is obviously dependent on the number of terms in the master equation, the quality of the software implementation and the difficulty of estimating each parameter, besides hardware performance. While benchmarks are not available for these scoring functions, one is inclined to believe that Jain's or Böhm's scoring functions would be faster than VALIDATE on the same set of compounds. None the less, VALIDATE has the highest diversity of the training set, which can be both an advantage (intuitively) and a drawback (as compared to VALIDATE II on the HIV-1 protease crystals). There is obviously a trade between speed and accuracy. One should be concerned more with rapidly screening out inactives' without too much effort and, as such, a steric overlap penalty (Eq. 29) should be an explicit part of every scoring function.

The presence of solvent molecules in the binding site is also an important aspect which has not been explicitly accounted for in the above scoring functions. While the importance of water in mediating ligand–protein interactions has been amply described in the literature, none of the scoring functions considers docking water molecules in the binding site in the same run as the ligand. Such a procedure could be improved if one considers that methods that rapidly estimate the mean residence time of a water molecule [74] at the interface with biomolecules are available [75]. In combination with such methods, the direct influence of the solvent in the binding process would be better accounted for.

The third requirement, tolerance to inaccuracies of the ligand's orientation in the binding site, appears to be a promising feature, provided that docking calculations were extremely fast. Such inaccuracies are also present in the structural data, and thermal factors should be checked prior to deriving the scoring function. The binding affinity represents the statistical ensemble averaged over  $10^{10}$  of ligand–receptor pairs, hence the ability to predict it based on the calculations for a single conformation of both ligand and receptor remains questionable [76].

Last, but not least, one should question the need to perform the scoring of ligands at all. Most such procedures are aimed at *de novo* design software, which is used with the explicit intention to explore the virtual chemical space in search of different, diverse, novel ligands. These virtual ligands are then ranked according to the scoring function, then later inspected by the computational chemist alone or in combination with synthetic chemists to evaluate synthetic feasibility of such ligands. Since the aim of this procedure is to explore molecular diversity, one should perform a cluster analysis (perhaps with the use of an experimental design procedure) on all the novel ligands and cluster them according to some similarity (or diversity) measure. Getting representative ligands from each cluster may prove more effective than predicting their binding affinity! In a combined procedure, one could envision a scoring function applied to D-optimal design-selected cluster representatives prior to individual compound inspection/selection.

Overall, the above results show that the receptor-based scoring function methodology holds promise and has room for improvement. With the proper selection of the training set (which one could bias for the therapeutic target of choice), and if thermodynamic aspects are accounted for (perhaps with the use of microcalorimetry), one can expect quite good results from applying the scoring function. None the less, one has to bear in mind that novel chemistry (in novel ligands) may behave in unpredicted manner, hence one should always expect (and welcome) the unexpected.

### Acknowledgements

One of the authors (T.I.O.) wishes to acknowledge useful discussions with Dr. Regine Bohacek (Ariad, Cambridge, MA), Dr. Peter Rose (Agouron Pharmaceuticals, San Diego, CA), Dr. Jeff Blaney (Chiron Corporation, San Francisco, CA) and Dr. Ludovic Kurunczi (Faculty of Pharmacy, Timisoara, Romania). We thank the co-authors of VALIDATE for a stimulating scientific environment during the development of this scoring function.

### References

1. Gund, P., *Three-dimensional pharmacophoric pattern searching*, Prog. Mol. Subcell. Biol., 11 (1997), 117–143.
2. Jakes, S.E., Watts, N., Willett, P., Bawden, D., and Fisher, J.D., *Pharmacophoric pattern matching in files of 3D chemical structures: Evaluation of search performance*, J. Mol. Graph., 5, (1987), 41–48,
3. Jakes, S.E., and Willett, P., *Pharmacophoric pattern matching in files of 3D chemical structures: Selection of interatomic distance screens*, J. Mol. Graphics, 4, (1986) 12–20.

4. Sheridan, R.P., Rusinko, A., III, Nilakantan, R., Venkataraghavan R., *Searching for pharmacophores in large coordinate data bases and its use in drug design*, Proc. Natl. Acad. Sci. USA, 86, (1989), 8165–8169.
5. Van Drie, J.H., Weininger, D., and Martin, Y.C., *ALADDIN: An integrated tool for computer-assisted molecular design and pharmacophore recognition from geometric steric and substructure searching of three-dimensional molecular structures*, J. Comput.-Aided Mol. Design 3 (1989) 225–251
6. Martin, Y.C., *3D Database searching in Drug Design*, J. Med. Chem., 35 (1992) 2145–2154.
7. Martin, Y.C., Bures, R.I.G., Danaher, E.A., and DeLazzer, J., *New strategies that improve the efficiency of the 3D design of bioactive molecules*, In Trends in QSAR and molecular modelling 92, Wermuth, C.G., (Ed.) ESCOM, Leiden, 1993. P. 20–27.
8. Kuntz, I.D., Blaney, S.M., Oatley, S.J., Langridge, R. and Ferrin, T.E., *A geometric approach to macro-molecule-ligand interactions*, J. Mol. Biol. 161 (1982) 269–288.
9. DesJarlais, R.L., Seibel, G.L., Dixon, J.S., and Kuntz, I.D., *Using shape complimentary as an initial screen in designing ligands for a receptor binding site of known three-dimensional structure*, J. Med. Chem., 31 (1990) 722–729.
10. DesJarlais, R.L., Seibel, G.L., Kuntz, I.D., Fruth, P.S., Alvarez, J.C., de Montellano, P.R. O., DeCamp, D.L., Babe, L.M., Craik, C.S., *Structure-based design of nonpeptide inhibitors specific for the human immunodeficiency virus I protease*, Proc. Natl. Acad. Sci. USA, 87, (1990) 6644–6648.
11. Oshiro, C.M., Kuntz, I.D., Dixon, J.S., *Flexible ligand docking using a generic algorithm*, J. Comput.-Aided Mol. Design. 9 (1995) 113–130.
12. Ring C. Sun E., McKerrow S., Lee G., Rosenthal P., Kuntz I., and Cohen F., *Structure-based inhibitor design by using protein models for the development of antiparasitic agents*, Proc. Natl. Acad. Sci. USA. 90. (1993), 3583–3587.
13. Shoichet, B.K., Stroud, R.M., Santi, D.V., Kuntz, I.D., and Perry, K.M., *Structure-based discovery of inhibitors of thymidylate synthase*, Science. 259. (1993 ), 1445–1450.
14. Cafilisch, A., Miranker, A. and Karplus, M., *Multiple copy simultaneous search and construction of ligands in binding sites application to inhibitors of HIV-1 aspartic proteinase*, J. Med. Chem., 36, (1993), 2142–2167.
15. Oprea, T.I., Waller, C.L., and Marshall, G.R., *Viral proteases: Structure and function*. In Cellular proteolytic systems. Ciechanover, A. and Schwartz, A. (Eds.). Wiley-Liss, Inc. New York. 1994. pp. 183–221.
16. Shuker, S.B., Hajduk, P.J., Meadows, R.P. and Fesik, S.W., *Discovering high affinity ligands for proteins: SAR by NMR*, Science, 274. (1996), 1531–1534.
17. Lewis, R. and Dean, P., *Automated site-directed drug design: the concept of spacer skeletons for primary structure generation*, Proc. R. Soc. Lond. [Biol.], 236 (1989) 125–140.
18. Lewis, R. and Dean, P., *Automated site-directed drug design: The formation of molecular templates in primary structure generation*, Proc. R. Soc. Lond. [Biol.], 236 (1989) 141–162.
19. Ho, C.M.W. and Marshall, G.R., *Cavity search; An algorithm for the isolation and Display of cavity-like binding regions*, J. Comput.-Aided Mol. Design. 4 (1990) 337–354.
20. Miranker, A. and Karplus, M., *Functionality maps of binding sites: A multiple copy simultaneous search method*, Proteins: Struct. Funct. Genet. 11 (1991) 29–34.
21. Moon, J.B., and Howe, W.J., *Computer design of bioactive molecules: A method for receptor-based de novo ligand design*, Proteins: Struct. Funct. Genet. 11 (1991) 314–328.
22. Nishibata, Y. and Itai, A., *Automatic creation of drug candidate structures based on receptor structure Starting point for artificial lead generation*, Tetrahedron. 47 (1991 ) 8985–8990.
23. Böhm, H.-J., *The computer program LUDI: A new method for the de novo design of enzyme inhibitors*. J. Comput.-Aided Mol. Design. 6 (1992) 61–78.
24. Borman, S., *New 3-D search and de novo design techniques aid drug development*, C&EN. (1992) 18–26.
25. Chau, P.L., and Dean, P.M., *Automated site-directed drug design: An assessment of the transferability of atomic residual charges (CNDO) for molecular fragments*, J. Comput.-Aided Mol. Design. 6 (1992) 407–426.
26. Chau, P.L., and Dean, P.M., *Automated site-directed drug design: Searches of the Cambridge Structural Database for bond lengths s in molecular fragments to be used for automated structure assembly*, J. Comput.-Aided Mol. Design, 6 (1992) 397–406.

27. Chau. P.L.. and Dean. P.M., *Automated site-directed drug design: The generation of a basic set of fragments to be used for automated structure assemble*, J. Comput.-Aided. Mol. Design. 6 (1992) 385–396.
28. Ho. C.M.W. and Marshall. G.R., *FOUNDATION: A program to retrieve subsets of query elements, including active site region accessibility, from three-dimensional ional databases*, J. Coinput.-Aided Mol. Design, 7 (1993) 3–22.
29. Ho. C.M.W., and Marshall, G. R. *SPLICE: A program to assemble partial query solutions from three-dimensional database searches into novel ligands*, J. Cornput.-Aided Mol. Design, 7 (1 993) 623–647.
30. Rotstein. S.H., and Murcko. M.A., *Group Build: A fragment-based method for de novo drug design*, J. Med. Chem., 36. (1993) 1700–1710.
31. Ho. C.M.W. and Marshall. G.R., *DBMAKER: A program to generate 3D databases based upon user specified criteria*, J. Comput.-Aided Mol. Design. 8 (1994) 65–86.
32. Ho. C.M.W. and Marshall, G.. *De novo design of ligands*, In Proceedings of the twenty-seventh annual Hawaii International Conference on system sciences, Hunter, L. (Ed.) IEEE Computer Society Press, Washington. DC. Vol. 5 1994, pp. 213–220.
33. Bohacek, R.S. and McMartin C., *Multiple highly diverse structures complimentary to enzyme binding sites: Results of extensive application of a de novo design method incorporating combinatorial growth*, J. Am. Chem. Soc., 116 (1994) 5560–5571.
34. Roberts, N.A., Martin, J.A., Kinchington, D., Broadhurst, A.V., Craig, J.C., Duncan I. B., Galpin, S.A., Handa, B.K., Kay, J., Krohn, A., Lambert, R.W., Merrett, J.H., Mills J. S., Parkes K.E.B., Redshaw, S., Ritchie, A.J., Taylor, D.L., Thomas. G.J., and Machin, P. J, *Rational design of peptide-based HIV proteinase inhibitors*, Science. 248 ( 1990) 358–361.
35. Tomasselli, A.G., Hui, J.O., Sawyer. T.K., Thaisrivongs, S., Hester, J.B. and Heinrikson R.L., *The evaluation of non-viral substances of the HIV protease us leads in /he design of inhibitors for AIDS therapy*, Adv. Exp. Med. Biol., 306 (1991) 469–482.
36. Thaisrivongs, S., Tomasselli, A., Moon, J., Hui, J., McQuade, T., Turner, S., Stronbach, J., Howe. J. Tarpley, W., and Heinrikson, R., *Inhibitors of the protease from human immunodeficiency virus: Design and modeling of a compounds containing a dihydroxyethylene isostere insert with a high binding affinity and effective antiviral activity*, J. Med. Chem., 34 (1991) 2344–2356.
37. Thanki, N., Rao, J., Foundling, S., Howe. W., Moon, J., Hui, J., Tomasselli, A., Heinrikson, R., Thaisrivongs, S., and Wlodawer, A., *Crystal structure of a complex of HIV- I protease with a dihydroxyethylene-containing inhibitor: comparison with molecular modeling*, Protein Science, 1 (1992) 1061–1072.
38. Thaisrivongs, S., Turner. S., Strohbach. J., TenBrink. R., Tarpley, W., McQuade T., Heinrickson., R., Tomasselli, A., Hui, J., and Howe, W., *Inhibitors of the protease for the HIV: Synthesis, enzyme inhibition and antiviral activity of a series of compounds containing the dihydroxyentylene transition-state isostere*, J. Med. Chem., 36 (1993) 941–952.
39. Lam, P.Y., Jadhve. P.K., Eyermann, C.J., Hodge, C.N., Ru, Y., Bacheler, L.T., Meek, J.L., Otto, M.J., Rayner, N.M., Wong, Y.N., et al., *Rational design of potent, bioavailable, nonpeptide cyclic ureas as HIV protease inhibitors*, Science, 263 ( 1994) 380–384.
40. Lunney, E.A., Hagen, S.E., Domagala, J.M., Humblet, C., Kosinski, J., Tait, R.D., Warmus, J.S., Wilson, M., Ferguson. D. et al., *A novel nonpeptide HIV I protease inhibitor: elucidation of the binding mode and its application in the design of related analogs*, J. Med. Chem., 37 (1994) 2664–2667.
41. Thaisrivongs, S., Tomich, P.K., Watenpaugh, K.D., Chong, K.T., Howe, W.J., Yang, C.P., Strohbach. J.W., Turner. S.R., McGrath, J.P., Bohanon, M.J., et al. *Structure-based design of HIV protease inhibitors: 4-hydroxycoumarins and 4-hydroxy-2-pyrones as non-peptidic inhibitors*, J. Med. Chem., 37 (1994) 3200–3204.
42. Oprea. T.I., Ho, C.M.W.. and Marshall, G.R., *De novo design: Ligand construction and prediction of affinity*, In Application of computer-aided molecular design: Agrochemicals, materials and pharmaceuticals, Reynolds, C.H., Holloway M.K., and Cox. H. (Eds.) ACS. Washington DC. 1995. pp. 64–81.
43. Caffisch, A., *Computational combinatorial ligand design: Application to human alpha-throbin*, J. Comput.-Aided Mol. Design. 10 (1996) 372–396.

44. Böhm, H.-J.. *The development of a simple empirical wiring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure*, J. Comput.- Aided Mol. Design, 8 (1994) 243–256.
45. Wallqvist. A., Jernigan, R.L., and Covell. D.G., *A preference-based free-energy parameterization of enzyme-inhibitor binding: Applications to HIV-1 protease inhibitor design*. Protein Sci.. 4. (1995) 1881–1903.
46. Verkhivker. G., Appelt, K., Freer. S.T., and Villafranca, J.E.. *Empirical free energy calculations of ligand-protein crystallographic complexes: I. Knowledge-based ligand-protein Interaction Potentials Applied to the Prediction of HIV- 1 Protease Binding Affinity*, Protein Eng., 8 (1995) 677-691.
47. Head. R.D., Smythe, M.L.. Oprea, T.I.. Waller. C.L.. Greene S.M.. and Marshall. G.R., *VALIDATE: A new method for the receptor-based prediction of binding affinities of novel ligands*, J. Ani. Chem. Soc., 118 (1996) 3959–3969.
48. Jain, A.. *Scoring non-covalent pretein-ligand interactions: A continuous differentiable function tuned to compute binding affinities*, J. Comput.-Aided Mol. Design. 10 (1996) 427–440.
49. Kollman P.. *Free energy calculations: Application to chemical and biochemical phenomena*, Chem., Rev., 93 (1993) 2395–2417.
50. Åqvist, J., Medina C. and Samuelsson, J.-E., *A new method for predicting binding affinity in c computer.-aided drug design*, Protein Eng., 7 (1994) 385–391.
51. Åqvist. J., and Mowbray, S.L., *Sugar recognition by a glucose/galactose receptor: Evaluation of binding energetics from molecular dynamics simulations*, J. Biol. Chem., 270. (1995) 9978–9981.
52. Gilson, M.K.. Given J.A.. and Head, M.S., *A new class of models for computing receptor-ligand binding affinities*, Chem, Biol.. 4 (1997) 87–92.
53. Head, M.S., Given J.A.. and Gilson M.K.. *Mining minima Direct computation of conformational free energy*, J. Phys. Chem, A, 101 (1997) 1609–1618.
54. Holloway. M.K.. *Structure-based design of human immunodeficiency virus-1 protease inhibitors: correlating calculated energy with activity in Application of computer-Aided molecular design: Agrochemicals, material\ and pharmaceuticals*. Reynolds, C.H., Holloway M.K., and Cox, H. (Eds.) ACS, Washington DC. 1995. p. 82.
55. Waller. C.L.. Oprea. T.I.. Giolitti, A., and Marshall. G.R., *3D QSAR of human immunodeficiency virus-1 protease inhibitor: I. A CoMFA study employing experimentally-determined alignment rules*, J. Med. Chem. 36 (1993) 4152–4160.
56. Oprea, T.I.. Waller. C.L.. and Marshall. G.R., *3D-QSAR of human immunodeficiency virus-1 protease inhibitors: II. Predictive power using limited exploration of alternate binding modes*, J. Med. Chem.. 37 (1994) 2206–2215.
57. Oprea. T.I., Waller. C.L.. and Marshall. G.R., *3D-QSAR of human immunodeficiency virus-1 protease inhibitor: III. Interpretation of CoMFA results*, Drug Des. Discov., 12 (1994) 29–51
58. Ajay. and Murcko. M.. *Computational methods to predict binding free energy in ligand-receptor complexes*, J. Med. Chem.. 38 (1995) 4953–4967.
59. Xue. Q., and Yeung, E.S., *Differences in the chemical reactivity of individual molecules of an enzyme*, Nature, 373 (1995) 681–683.
60. Searle, M.S., and Williams. D.H., *The cost of conformational order: Entropy changes in molecular associations*, J. Am. Chem. Soc. 114 (1992) 10690–10697.
61. Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanonchi. T. and Tasumi M., *The Protein Data Bank: A computer-based archival file for macromolecular structures*, J. Mol., Biol., 112 (1977) 535–542.
62. Connolly, M., *Solvent-accessible surfaces of protein and nucleic acids*, Science. 221 (1983) 709–713.
63. Bondi. A., *Van der Waals volumes and radii*, J. Phys. Chem., 68 (1964) 441–449.
64. Sippl, M.J., *Boltzmann's principle, knowledge-based mean fields and proreïn folding: An approach to the computational determination of protein structures*, J. Comput.-Aided Mol. Design, 7 (1993) 473–501.
65. Brooks. B.R., Bruccolieri, R.E., Olafson, D., States, D., Swaminathan, S. and Karplus, M., *CHARMM: A program for macromolecular energy, minimization, and dynamics calculation*, J. Comput. Chem., 4 (1983) 181–186.

66. Gilli, P.; Feretti, V., Gilli G. and Borea, P.A., *Enthalpy-entropy compensation in drug receptor binding*, J. Phys. Chem., 98 (1994) 1515–1518.
67. Leo, A., *Estimating  $\log P_{oct}$  from structures*, Chem. Rev., 5 (1993) 1281–1306.
68. Kellogg, G.E., Semus, S.F., and Abraham, D.J., *HINT: A new method of empirical hydrophobic field calculation for CoMFA*, J. Comput.-Aided Mol., Design, 5 (1991) 545–552.
69. Le Grand, S.M. and Merz, K.M., *Rapid approximation to molecular surface area via the use of Boolean logic and look-up tables*, J. Comput. Chem., 14 (1993) 349–352.
70. Still, W.C., Tempczyk, A., Hawley R.C., and Hendrickson, T., *Semianalytical treatment of solvation for molecular mechanics and dynamics*, J. Am. Chem. Soc., 112 (1990) 6127.
71. Oprea, T.I., Head, R.D., and Marshall, G.R., *The basis of cross-reactivity for a series of steroids binding to a monoclonal antibody against progesterone (DB3): A molecular modeling and QSAR study*, In QSAR and molecular modelling: Concepts, computational tools and biological applications, Sanz, F., Giraldo, J., and Manaut, F. (Eds.) Prous Science Publishers, Barcelona. 1995, pp. 451–455.
72. Rose, P.W., *Scoring methods in ligand design*, 2nd USCF Course in Computer-Aided Molecular Design, San Francisco, CA, 1997.
73. Gilson, M.K., Given, J.A., Bush, B.L., and McCammon, J.A., *The statistical-thermodynamic basis for computation of binding affinities: A critical review*, Biophys. J., 72 (1997) 1047–1069.
74. Garcia, A.E., and Stiller, L. *Computation of the mean residence time of water in the hydration shells of biomolecules*, J. Comput. Chem. 14 (1993) 1396–1406.
75. Hummer, G., Garcia, A.E., and Soumpasis, D.M., *Hydration of nucleic acid fragments: Comparison of theory and experiment for high-resolution crystal structures of RNA, DNA, and DNA-drug complexes*, Biophys. J., 68 (1995) 1639–1652.
76. Oprea, T.I., and Waller, C.L., *Theoretical and practical aspects of three dimensional quantitative structure-activity relationships*. In Reviews in Computational Chemistry, vol. 11, Lipkowitz, K.B., and Boyd, D.B. (Eds), Wiley, New York, 1997, pp. 127–182.

**This Page Intentionally Left Blank**

# ***A Priori* Prediction of Ligand Affinity by Energy Minimization**

**M. Katharine Holloway**

*Molecular Design and Diversity, Merck Research Laboratories, West Point, Pennsylvania 19486, U.S.A.*

Quantitative *a priori* prediction of the affinity of a ligand for its receptor is, in a sense, the holy grail of structure-based molecular design, ideally allowing one to optimize the number and kind of compounds designed for a program prior to any chemical synthesis. Historically, several approaches have been employed to calculate and/or predict binding affinity.

The free energy of binding can be calculated directly via Free Energy Perturbation (FEP) calculations [1]. The relative binding energies for pairs of inhibitors are determined using a thermodynamic cycle in which the structure of one inhibitor is perturbed into the structure of another, both in the receptor site and in solvent. This approach has been reported to yield relative free energies that are accurate to  $\pm 1$  kcal/mol with respect to experiment. However, due to the amount of computer time required, these calculations are impractical for routine assessment of the binding affinity of proposed compounds.

Other more efficient approaches have included: Comparative Molecular Field Analysis (CoMFA) [23], the Hypothetical Active Site Lattice (HASL) [4], HINT hydrophobicities [5], solvent accessibility [6] or solvent-induced interactions [7], atom-atom contact preferences [8,9], a general mean field model [10,11] and scoring algorithms for database search or *de novo* design methods [12–16], as well as energy minimization methods.

This chapter will focus on computational studies which employ energy minimization in an X-ray or modelled active site as the means to predict the affinity of a ligand for its receptor. These studies fall into two categories: (a) energy-component approaches — i.e. those that incorporate receptor–ligand energy values as one term in a sum of binding energy contributions or as part of a 3D QSAR; and (b) energy-only approaches — i.e. those that employ energy minimization methods alone to predict ligand affinity.

## **1. Energy-component Approaches**

Recent approaches that incorporate energy minimization as part of a larger binding energy equation or 3D QSAR have included Marshall's VALIDATE [17], Ortiz and Wade's COMBINE [18] and the empirical free energy evaluation method of DeLisi [19,20].

The VALIDATE method incorporates 12 physico-chemical and energetic parameters, including the electrostatic and steric interaction energy between the receptor and ligand computed using the AMBER force field in MacroModel. A predictive 3D QSAR was derived for a diverse training set of 51 crystalline complexes, including HIV-1 protease, thermolysin, endothiaepsin,  $\beta$ -trypsin and subtilisin-Novo inhibitors; antibody(DB3)-



bound steroids; and L-arabinose binding protein-bound sugars. The ligands ranged in size from 24–1512 atoms and the  $pK_i$  values ranged from 2.47 to 14.0. The best fit equation, using PLS analysis, yielded an  $r^2 = 0.849$  with a standard error of 1.006 log units and a cross-validated  $r^2 = 0.776$ . This QSAR was found to be predictive for at least two of three test sets of enzyme–inhibitor complexes: 14 structurally diverse crystalline complexes (predictive  $r^2 = 0.81$ , absolute average error = 0.70 log units), 13 HIV protease inhibitors (predictive  $r^2 = 0.57$ , absolute average error = 0.73 log units) and 11 thermolysin inhibitors (predictive  $r^2 = 0.72$ , absolute average error = 1.48 log units). For more details of this approach see the chapter by T.I. Oprea and G.R. Marshall in this volume.

The COMBINE approach utilizes only the intermolecular interaction energy between the receptor and its ligand, but employs a unique method that partitions the energy among receptor and ligand fragments and subjects these energy components to statistical analysis. This is proposed to enhance contributions from mechanistically important interaction terms and to tune out noise due to inaccuracies in the potential energy functions and molecular models. For a set of 26 synovial fluid phospholipase A<sub>2</sub> inhibitors, the direct correlation between interaction energies, computed using the cff91 DISCOVER force field, and percent enzyme inhibition was very low,  $r = 0.212$ . However, with the COMBINE approach, employing PLS fitting and the GOLPE variable selection procedure, good correlations with percent inhibition were observed ( $r^2 = 0.92$ , cross-validated  $r^2 = 0.82$ ). For more details of this approach see the chapter by Wade et al. in this volume.

DeLisi and co-workers employ an empirical free energy evaluation that computes the binding free energy as a sum of various free energy contributions, including the interaction energy between the receptor and ligand computed with the CHARMM force field. The binding of nine serine endopeptidase inhibitors [19]; biotin, iminobiotin and thio-biotin to streptavidin [19]; five peptide antigens to MHC Class I receptors [19]; and six peptide-based HTV-1 protease inhibitors [20] have been examined using this methodology. The calculated empirical free energies are in good agreement with experiment and/or FEP calculations for the first three cases. For the HIV-1 protease inhibitors, the method is used to guide docking of the ligands via minimization of the free energy function, leading to better agreement with the observed bound structure (RMSD = 1.21 Å) than minimization of the CHARMM energy function (RMSD = 1.69 Å), although both correlated well with  $\ln K_i$  values (empirical energy function,  $r = 0.970$ ; CHARMM energy function,  $r = 0.967$ ).

## 2. Energy-only Approaches

Recent approaches that primarily employ energy minimization methods to predict ligand affinity have included studies of the binding of (a) influenza sialidase inhibitors [21], (b) aldehyde substrates for aldose reductase [22], (c) thrombin inhibitors [23], (d) serine proteinase inhibitors [24], and HIV-1 protease inhibitors [25–29].

Taylor and von Itzstein [21] performed calculations on the mechanism of sialoside cleavage by influenza virus sialidase and its inhibition by transition-state analogs such

as 2-deoxy-2,3-didehydro-N-acetylneuraminic acid (Neu5Ac2en). Minimized enzyme-inhibitor complexes for Neu5Ac2en and two analogs were generated employing a combination of molecular dynamics and molecular mechanics calculations with the CVFF force field. The computed intermolecular binding enthalpies correlated qualitatively with the observed  $K_i$  values.

Subsequently, De Winter and von Itzstein [22] performed calculations on the binding of substrates (D-xylose, L-xylose and D-lyxose) to wild-type human aldose reductase and two site-directed mutants, H110A and H110Q. Three different protonation states of His<sub>110</sub> were examined: (1) Nε<sub>2</sub>-H or HisI, (2) Nδ<sub>1</sub>-H or HisII, and (3) Nε<sub>2</sub>-H and Nδ<sub>1</sub>-H or His+. Minimum energy conformations for each enzyme-substrate complex were generated via a combination of molecular dynamics sampling and molecular mechanics minimization with the AMBER program. The solvation energy of these complexes was evaluated using DELPHI. An excellent correlation was observed between the calculated average interaction enthalpy, both with ( $r = 0.94$ ) and without solvation ( $r = 0.87$ ), and the measured  $\log(K_m)$  values for the HisI wild-type and H110A and H110Q mutant models, while there was no correlation for the other two wild-type models, HisII and His+. Thus it is proposed that the key residue His<sub>110</sub> is neutral and protonated at Nε<sub>2</sub> when an aldehyde substrate is bound to human aldose reductase. Further, the  $\log(K_m)$  of a fourth substrate, D-glucose, was correctly predicted (predicted = -0.93; observed = -0.94) from the correlation equation which included solvation,  $\log(K_m) = 2.72 + 0.241$  (average interaction enthalpy).

Grootenhuis and van Galen [23] examined the binding of 35 non-covalently bound inhibitors of human thrombin in the argatroban, TAPAP and NAPAP series with  $pK_i$  values ranging from 3 to 8. Energy minimization of the inhibitors in the X-ray structure of thrombin was performed using the CHARMM force field. Various electrostatic models were examined ( $\epsilon = 1, r, 4r$ ), as well as flexibility of the enzyme active site and the inclusion of solvation corrections. The correlation between the CHARMM interaction energy and  $pK_i$  was best ( $n = 32, r = 0.81, \text{cross-validated } r^2 = 0.60, \text{ standard deviation} = 0.97 \text{ log units}$ ) when the enzyme active site was held rigid, with a distance dependent dielectric constant ( $\epsilon = r$ ) employed during energy evaluation but not during energy minimization. The equation of the best fit line was  $E_{inter} = -4.23 (pK_i) - 54.74$ . Preliminary investigation of solvation effects using Eisenberg's solvation function did not improve the correlation. For more details of this work see the chapter by R.M.A. Knegtel and P.D.J. Grootenhuis in this volume (p. 99 ff).

Kurinov and Harrison [24] performed calculations, using the AMMP molecular mechanics program, on a series of 18 compounds containing hydrophobic groups and basic amines in order to predict their ability to inhibit bovine trypsin. Fifteen of the compounds were predicted to be inhibitors and three compounds were predicted not to be inhibitors, based on the presence or absence of a low-energy binding geometry within 2 Å of the known benamidine binding site. The binding energies calculated for the compounds predicted to inhibit trypsin were found to correlate ( $n = 15, r = 0.755; n = 14, r = 0.857$ ) with the observed  $\log K_i$  values, which ranged from -1.15 to -4.00. In addition, the compounds that were predicted not to inhibit trypsin did not inhibit trypsin. The slope of the correlation line was observed to be less than  $RT$ ; this was

attributed to neglect of solvation and entropic terms. It was found that an all-atom potential with no cutoff radius was essential to obtain a good correlation and that the flexibility of the ligand must be included (here via molecular dynamics) in order to properly represent the distribution of conformations. Subsequent X-ray structure determination of six of the inhibitors co-crystallized with trypsin indicated good overall agreement between the predicted and observed binding modes. Interestingly, the enzyme structure varied little from one X-ray structure to another, leading to the conclusion that trypsin is relatively rigid in its inhibited form, thus only flexibility of the inhibitor, not the enzyme, need be considered in predicting binding modes and energies.

HIV-1 protease inhibitor binding has been examined by four different groups. Weber and co-workers initially examined [25] three HIV-1 protease inhibitors, MVT-101, JG365, and U85548e, for which X-ray structures had been determined. These structures were minimized using the CVFF force field, using a distance dependent dielectric model to mimic solvation effects. A 'high' dielectric model,  $\epsilon = 4r$ , was found to produce minimized complexes and binding energies that were in better agreement with the X-ray structure and  $K_i$  measurements than did a 'low' dielectric model,  $\epsilon = r$ . Using this protocol, the calculated interaction energy of the three inhibitors with HIV-1 protease ranged from -53 to -56 kcal/mol, but were not correlated with observed  $K_i$ .

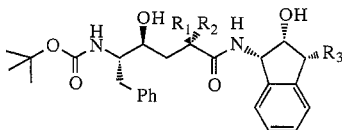
However, subsequent examination [26] of a set of 21 modelled peptide substrates of HIV-1 protease with single amino acid substitutions at  $P_4$  to  $P'_3$  led to calculated interaction energies for the corresponding tetrahedral intermediates that correlated well with the observed  $k_{cat}/K_m$  ( $n = 21$ ,  $r = 0.64$ ;  $P_1 - P'_1$  only,  $n = 8$ ,  $r = 0.93$ ;  $P_2 - P'_2$  only,  $n = 14$ ,  $r = 0.86$ ). These calculations were performed using the AMMP molecular mechanics program and a constant dielectric model ( $\epsilon = 1$ ). Side-chain conformations for the peptide substrate were determined via systematic search. The catalytic mechanism and factors influencing the catalytic efficiency of the different substrates were discussed in relation to the models: in particular, it was proposed that the ordered water molecule observed in many X-ray structures between the inhibitor and enzyme 'flaps' acts as the nucleophile rather than the more traditional hypothesis of a water activated by the catalytic aspartic acids, Asp<sub>A25</sub> and Asp<sub>B225</sub>. For more details of this work see the chapter by I.T. Weber and R.W. Harrison in this volume (p. 115 ff).

Miertus and co-workers [27] examined a set of eight peptide-based HIV-1 protease inhibitors, modelled on the hexapeptide MVT-101 structure. Enzyme-inhibitor calculations were performed using the CVFF force field: short molecular dynamics runs were performed to relieve strain in manually constructed inhibitor models: solvation effects were evaluated with the Polarizable Continuum Method, with  $\epsilon = 80$  representing a polar environment. Summation of total complexation and solvation energies led to values that were used to prioritize inhibitors for synthesis. e.g. incorporation of negatively charged residues at  $P_2$  or  $P'_2$  improved the energy and was consistent with improved inhibition [28]. Varying the protonation state of the basic amine in these reduced peptide inhibitors led to uniform changes in the complexation energy, but some variability in the solvation energy by modulating the overall charge on the inhibitor. Variation of the central transition state mimic was also examined; inclusion of a second hydroxyl or amine to the U-85545e reference structure was predicted to be favorable.

Viswanadhan and Reddy [29] examined a series of 11 peptidomimetic HIV-1 protease inhibitors with structural variation at the  $P'_2$  position. Energy minimizations were performed with the AMBER force field and solvation energies were computed using an explicit water shell. The binding enthalpy resulting from these calculations compared favorably to FEP calculations for seven of the inhibitor pairs. Hydrophobic interaction energies were also computed but found to correlate less strongly ( $n = 7$ ,  $r = 0.72$ , standard deviation = 1.05 kcal/mol) than the calculated binding enthalpies ( $n = 7$ ,  $r = 0.92$ , standard deviation = 0.57 kcal/mol) with the observed binding energy.

We have also examined the binding of HIV-1 protease inhibitors using energy minimization methods [30,31], with the goal of predicting activities and prioritizing synthesis for proposed inhibitors *a priori*. For this purpose, a training set of 33 inhibitors with structural variation at the  $P'_1$  and  $P'_2$  positions was employed to derive a correlation between the computed enzyme–inhibitor interaction energy and the observed  $IC_{50}$  values. The inhibitors employed in the training set are shown in Tables 1 and 2. They were selected based on variety of structure and activity: 16 inhibitors (2–17) contained

Table 1 Experimental  $IC_{50}$  values and calculated enzyme–inhibitor interaction energies for the  $P_1$ , training set of HIV-1 protease inhibitors.

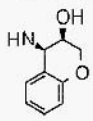
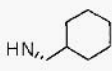
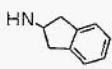
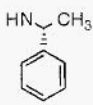
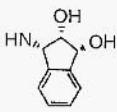
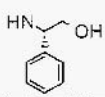


No.	$R_1$	$R_2$	$R_3$	$IC_{50}$ (nM)	p $IC_{50}$	$E_{inter}$ (kcal/mol)
1	CH <sub>2</sub> Ph	H	H	0.25	9.602	-145.1
2	CH <sub>2</sub> Ph	CH <sub>3</sub>	H	7.7	8.114	-140.4
3	CH <sub>2</sub> CH <sub>2</sub> CH <sub>2</sub> Ph	H	OH	0.19	9.721	-146.6
4	CH <sub>2</sub> -4-CF <sub>3</sub> Ph	H	H	0.26	9.585	-149.6
5	(E)-CH <sub>2</sub> CH=CHPh	H	H	0.23	9.638	-147.1
6	CH <sub>2</sub> C <sub>6</sub> F <sub>5</sub>	H	H	0.6	9.222	-149.4
7	CH <sub>2</sub> -4-CH <sub>3</sub> Ph	H	H	0.29	9.538	-146.5
8	CH <sub>2</sub> -4-NH <sub>2</sub> Ph	H	H	0.31	9.509	-146.1
9	CH <sub>2</sub> -4-NO <sub>2</sub> Ph	H	H	0.27	9.569	-151.4
10	H	H	H	2934	5.533	-129.2
11	CH <sub>2</sub> -4-OHPh	H	H	0.16	9.796	-149.7
12	CH <sub>2</sub> CH=CH <sub>2</sub>	H	H	27.5	7.561	-137.8
13	CH <sub>2</sub> -4-IPh	H	H	0.72	9.143	-148.4
14	CH <sub>2</sub> C(O)Ph	H	H	5.42	8.266	-150.3
15	CH <sub>2</sub> -4-pyridyl	H	H	0.53	9.276	-144.9
16	CH <sub>2</sub> SPh	H	H	0.25	9.602	-146.0
17	CH <sub>2</sub> -4-t-butylPh	H	H	0.17	9.770	-150.9

Table 2 Experimental  $IC_{50}$  values and calculated enzyme-inhibitor interaction energies for the  $P_2$  training set of HIV-1 protease inhibitors.

No.	R	$IC_{50}$ (nM)	$pIC_{50}$	$E_{inter}$ (kcal/mol)
18		114	6.943	-131.5
19		9.53	8.021	-132.9
20		34.25	7.465	-135.5
21		690	6.161	-134.1
22		161	6.793	-131.3
23		66.3	7.179	-139.3
24		212.42	6.673	-144.0
25		121.8	6.914	-134.2
26		0.7	9.155	-146.7
27		0.18	9.745	-145.5

Table 2 (continued)

No.	R	IC <sub>50</sub> (nM)	pIC <sub>50</sub>	E <sub>inter</sub> (kcal/mol)
28		40.5	7.393	-134.5
29		30 000	4.523	-124.1
30		130	6.886	-129.8
31		146	6.836	-134.1
32		0.1	10.000	-149.1
33		38.6	7.413	-138.4

modifications in  $P'_1$  and 16 inhibitors (**18–33**) contained modifications in  $P'_2$ ; pIC<sub>50</sub> values ranged from 4.523 to 10.000.

An initial model of **1** was constructed based on the X-ray structures of renin inhibitors bound in the active sites of fungal aspartyl proteases such as Endothiapepsin and *Rhizopus* pepsin. Models of **2–33** employed the model of **1** as a template. The inhibitor models were minimized in the active site of the L-689,502 inhibited HIV-1 protease X-ray structure [32] using the MM2X force field [30]. In all calculations the inhibitor was completely flexible and the enzyme was completely rigid. Dielectric constants of 1.5 for intramolecular interactions and 1.0 for intermolecular interactions were employed. All titratable HIV-1 protease residues were charged with the exception of Tyr<sub>59</sub> and one of the pair of catalytic aspartates, Asp<sub>A25</sub>, which was protonated on O<sub>δ1</sub>. The latter protonation state was chosen based on pH rate profiles which suggest that the catalytic aspartates of the fungal aspartyl proteases Penicillopepsin and *Rhizopus* pepsin [33] and the HIV-1 protease [34] share one negative charge.

The computed total energy,  $E_{tot}$  is a sum of the intramolecular energies of the enzyme,  $E_{enz}$ , and the inhibitor,  $E_{inh}$ , and the intermolecular (or interaction) energy,  $E_{inter}$  between the enzyme and the inhibitor. Since the enzyme is held fixed,  $E_{enz}$  is constant and is not computed.  $E_{inter}$  corresponds to the sum of the van der Waals ( $E_{vdw}$ ) and electrostatic ( $E_{elec}$ ) interactions between the inhibitor and the enzyme.

$$E_{tot} = E_{enz} + E_{inter} + E_{inh}$$

$$E_{inter} = E_{vdw} + E_{elec}$$

$E_{inter}$  should correspond to one component of the overall free energy of binding,  $\Delta G_{bind}$ , with other significant contributions being: the solvation/desolvation of the enzyme, inhibitor and enzyme–inhibitor complex; the flexibility of the enzyme; the energy associated with achieving the bioactive conformation of the enzyme and inhibitor; and the entropy changes associated with binding. However, one might expect  $E_{inter}$  to correlate with  $\Delta G_{bind}$  if other effects are not dominant — i.e. if the enzyme structure does not vary much from one inhibited complex to the other, if the inhibitors occupy the same binding sites and have similar conformation and character, and if entropy changes are relatively constant. Indeed, the computed  $E_{inter}$  values listed in Tables 1 and 2 correlate well with the experimental observation,  $\text{pIC}_{50}$ , as shown in Fig. 1, with the following relationship:

$$\text{pIC}_{50} = -0.169(E_{inter}) - 15.707 \quad (1)$$

$$n = 33, r^2 = 0.783, \text{ cross - validated } r^2 = 0.770, s = 0.675$$

Separation of  $E_{inter}$  into the van der Waals,  $E_{vdw}$ , and electrostatic,  $E_{elec}$ , components, as depicted in Fig. 2, indicated that neither correlated as well individually with  $\text{pIC}_{50}$  as the sum of the two,  $E_{inter}$ . This is consistent with the observation that both hydrophobic binding — i.e. steric complementarity — and hydrogen-bonding interactions — i.e. electrostatic complementarity — are critical to the activity of HIV-1 protease inhibitors.

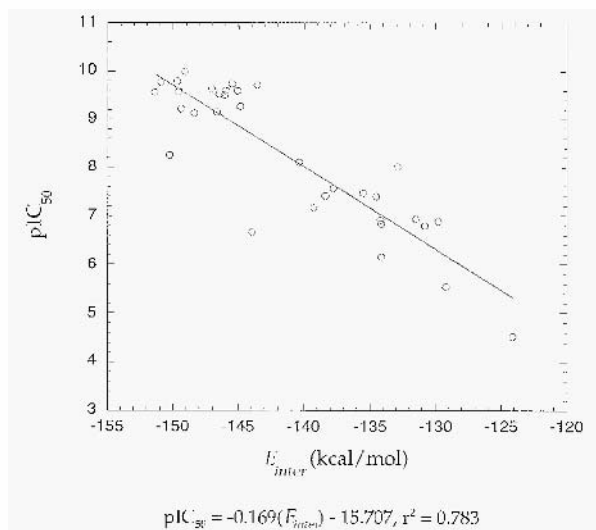


Fig. 1. Calculated enzyme–inhibitor interaction energy ( $E_{inter}$ ) versus experimental enzyme inhibition ( $\text{pIC}_{50}$ ) for the training set of inhibitors.

Variation of the protonation state and protonation site of the catalytic aspartic acids, Asp<sub>A25</sub> and Asp<sub>B225</sub>, indicated that the observed correlation between  $E_{inter}$  and  $pIC_{50}$  was essentially independent of protonation state. As shown in Table 3, the correlation

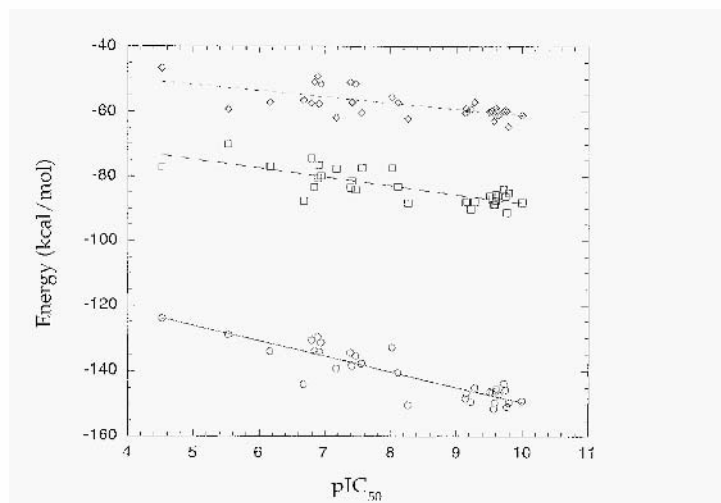
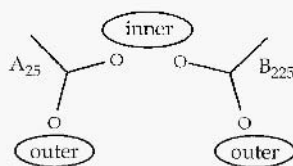


Fig 2. Electrostatic (diamonds) versus van der Waals (squares) contributions to the interactions energy,  $E_{inter}$  (circles), for the training set of inhibitors 1–33. Equations of the best-fit line and correlation coefficients are:  $E_{elec}$ :  $pIC_{50} = -0.23.3(E_{elec}) - 5.1988$ ,  $r_2 = 0.436$ ;  $E_{vdw}$ :  $pIC_{50} = -0.221(E_{vdw}) - 10.270$ ,  $r_2 = 0.602$ ;  $E_{inter}$ :  $pIC_{50} = -0.169(E_{inter}) - 15.707$ .  $r^2 = 0.783$ .

Table 3 effect of aspartic acid ( $A_{25}$  and  $B_{225}$ ) protonation state on the correlation between observed  $pIC_{50}$  and calculated  $E_{inter}$  for 1–33.



Charge state	Proton position	$r^2$
-1	A <sub>25</sub> outer O	0.784
-1	A <sub>25</sub> inner O	0.784
-1	B <sub>225</sub> outer O	0.784
-1	B <sub>225</sub> inner O	0.777
0	A <sub>25</sub> outer O, B <sub>225</sub> inner O	0.804
0	A <sub>25</sub> inner O, B <sub>225</sub> outer O	0.771
0	A <sub>25</sub> outer O, B <sub>225</sub> outer O	0.783
2	none	0.710



coefficient is similar for all protonation states with the possible exception of the completely ionized species in which both aspartates are negatively charged. This is consistent with calculations by Tossi [28] (see above), which indicate that varying the protonation state of the secondary amine in reduced peptide inhibitors leads to uniform changes in the complexation energies within a series.

However, the observed correlation was not independent of the force field employed as shown in Table 4. Similar calculations with the CHARMM force field [35] led to an  $r^2$  of 0.520 between  $E_{inter}$  and  $\text{pIC}_{50}$ . We postulated that the better correlation obtained with the MM2X force field might be due to the consistent charging scheme employed. Indeed, incorporation of the MM2X charges in the CHARMM calculations led to a significant improvement in the correlation ( $r^2 = 0.683$ ). Thus, the MM2X force field and, in particular, the MM2X charges, appear superior to the CHARMM force field and charges for this specific application. This is interesting in light of some of the other studies of this kind which employed the CHARMM force field (see above).

We also examined the contribution of other key factors in binding — e.g. the flexibility of the enzyme active site, the difference in energy between the solution and bound conformations of the inhibitor and the solvation/desolvation of the inhibitor and the enzyme. Including the flexibility of the enzyme active site did improve the correlation in the context of CHARMM minimization [31]; unfortunately, these calculations were not possible with the MM2X force field.

Including the difference in energy between the solution and bound conformations of the inhibitor was attempted by subtracting from  $E_{inter}$  the difference in intramolecular energy between the free and bound inhibitor conformations. Two different dielectric constants ( $\epsilon = 1$  and  $\epsilon = 50$ ) were employed to model the free inhibitor conformation that was the closest local minimum to the bound conformation. Neither dielectric model led to energy differences that had a significant effect on the correlation [31]. This suggests that either there is no real energy penalty paid for the inhibitor to achieve the bioactive conformation or, more likely, that the penalty is similar for the inhibitors in the training set.

Table 4 *The effect of computational parameters such as force field, charge model and solvation, on the correlation with  $\text{PIC}_{50}$  for 1–33.*

Computed value	$r^2$
$E_{inter}$ (MM2X)	0.784
$E_{inter}$ (CHARMM)	0.520
$E_{inter}$ (CHARMM, MM2X charges)	0.683
$E_{solv,l}$ (BMIN)	0.079
$E_{solv,l}$ (BMIN), excluding 9	0.156
$E_{solv,Tot}$ (BMIN)	0.118
$E_{inter}$ (MM2X) & $E_{solv,l}$ (BMIN)	0.786
$E_{inter}$ (MM2X) & $E_{solv,Tot}$ (BMIN)	0.789
Inhibitor surface area	0.303
Inhibitor volume	0.319

The effects of solvation were included using the BATCHMIN GB/SA [36] continuum solvation method with the MM2 force field [37]. Solvation energies were computed for the rigid enzyme active site, for each free inhibitor, and for each enzyme–inhibitor complex. However, neither the solvation energy of the inhibitor ( $E_{solv,i}$ ) nor the total solvation energy ( $E_{solv,Tot} = E_{solv,E-I} - E_{solv,E} - E_{solv,I}$ ) improved the correlation when employed in a multiple linear regression with the  $E_{inter}$  values as shown in Table 4. This is consistent with a low correlation between the computed solvation energy and the observed activity ( $r^2 = 0.079$  and  $0.118$  for  $E_{solv,I}$  and  $E_{solv,Tot}$ , respectively). Even when the outlier **9** was excluded from the regression between  $E_{inter}$  and  $E_{solv,I}$ , the correlation was only slightly improved. It is interesting to note that the surface area and volume computed for the bound inhibitor conformation correlated better with  $pIC_{50}$  than the BATCHMIN GB/SA solvation energies. Clearly, solvation/desolvation effects are important to the overall binding process. However, these results indicate that either solvation effects vary little for this set of inhibitors or that we have not properly approached the computation of solvation effects. The inability to improve the correlation with the inclusion of a solvation term is in agreement with the results of Grootenhuis and van Galen [23] who employed an Eisenberg solvation function for a series of thrombin inhibitors.

Thus, we have discovered nothing that improves the predictivity of our original correlation between  $E_{inter}$  and  $pIC_{50}$ , despite the fact that the other energetic factors are clearly key contributors to binding affinity.

In order to demonstrate that the observed correlation is not specific to this training set of inhibitors, we examined two other sets of HIV-1 protease inhibitors — i.e. cyclic urea inhibitors **34–37** from DuPont-Merck [38] and hydroxycoumarin and hydroxypyronone inhibitors **38–68** from Upjohn [39]. Both sets of inhibitors differ significantly from the Merck set, in that they are water-displacing templates in which an oxygen of the inhibitor occupies the binding site of a critical ordered water molecule located between the flaps of the enzyme and the inhibitor in the X-ray structures of HIV-1 protease/inhibitor complexes. Additionally, the activity measure for these datasets are  $K_i$ , not  $IC_{50}$ . These cannot be compared directly since  $K_i$  is an intrinsic property, while  $IC_{50}$ , varies as a function of the assay conditions — e.g. as a function of substrate, pH and salt concentration. Thus, we can derive correlations between the calculated  $E_{inter}$  and the observed  $pK_i$  values, given in Tables 5–8, for each of the additional training sets (DuPont-Merck,  $r^2 = 0.967$ ; Upjohn,  $r^2 = 0.606$ ) or for the combination of the two ( $r^2 = 0.759$ ). The individual correlations are illustrated in Fig. 3 for the DuPont-Merck inhibitors and Fig. 4 for the Upjohn inhibitors. However, it would be ill-advised i.e. comparing apples to oranges) to attempt to correlate  $E_{inter}$  with the observed  $pIC_{50}$  or  $pK_i$  values for the entire set of Merck, DuPont-Merck and Upjohn inhibitors since there is no standard of reference for their activity.

Note that the correlation for the Upjohn series of inhibitors, while very good, is not as strong as that for the Merck or DuPont-Merck series. This may result from the fact that these inhibitors were assayed as diastereomeric mixtures, while the modelling was performed on a single structure that was consistent with the more active diastereomer, **67**, of the separate mixture **65–68**.

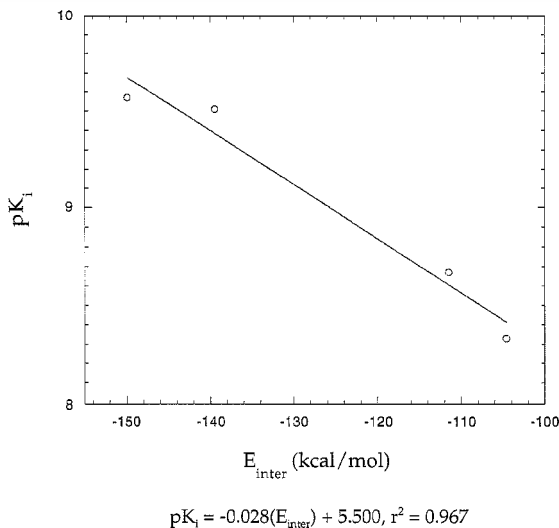


Fig. 3. Calculated enzyme-inhibitor interaction energy ( $E_{inter}$ ) versus experimental enzyme inhibition ( $pK_i$ ) for the DuPont-Merck inhibitors.

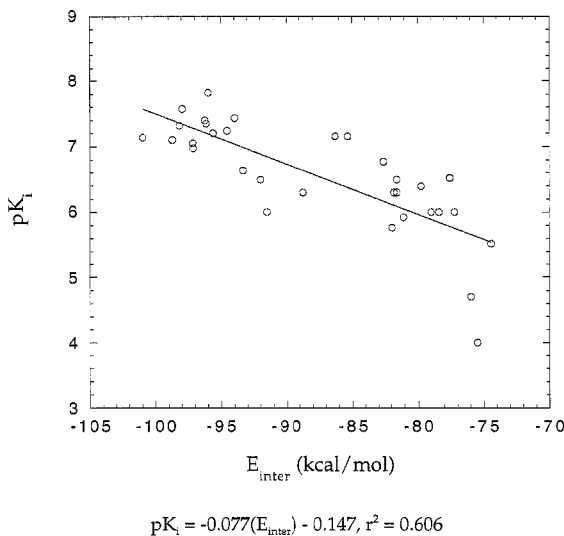


Fig. 4. Calculated enzyme-inhibitor interaction energy ( $E_{inter}$ ) versus experimental enzyme inhibition ( $pK_i$ ) for the Dupont-Merck inhibitors.

Employing the initial correlation for the Merck series of inhibitors described in Eq. 1, we were able to make predictions of activity for inhibitors prior to synthesis — i.e. true predictions and not *post hoc* explanations of activity. A set of compounds, **69–79**, for

which activities were predicted, is shown in Table 9 with the computed  $E_{inter}$  and predicted and observed  $pIC_{50}$  values. The accuracy of these predictions is illustrated graphically in Fig. 5. Here, the line is one of unit slope, not a correlation line. The average unsigned error for the predicted set of compounds is 0.59 log units across a range of 5.10 log units — i.e. only about a factor of 4. In many cases, this is comparable to the error inherent in the measurement of a biological activity.

It must be emphasized that we made many more predictions than are shown in Table 9; these are only examples. In addition, when our prediction for a compound was unfavorable, it was frequently not synthesized; or when a prediction was favorable, the exact compound that was modelled was not synthesized, but rather an analog.

As can be seen in Table 9, this approach was able to distinguish small differences in structure/activity between two diastereomers, **71** and **72**, as well as gross differences in structure/activity between the inverted pair of inhibitors **70** (Ro-31-8959) and **73**, which occupy the same binding sites but in the opposite direction. The 6-membered lactam ring in **69** was correctly predicted to fit poorly in the active site, although the analogous 5-membered lactam was a potent inhibitor ( $IC_{50} = 37$  nM) [40].

However, the most significant prediction, for **76**, involves an interesting feature of HIV-1 protease-inhibitor complexes. Due to the symmetrical nature of the enzyme, some inhibitors, such as Ro-31-8959, **70** [41], are observed to bind in the active site in two directions in the X-ray structure with respect to the flaps, that in their closed H-bonded form introduce the asymmetry that is the direction marker. This bidirectional binding is illustrated in Fig. 6. The hypothesis that **1** and **70** might bind in the active site in opposite directions, as illustrated in Fig. 7, could explain the preference for inverse stereochemistry at the transition state hydroxyl. Based on this hypothesis, the hybrid inhibitor **76**, which contained the C-terminal halves of both **1** and **70**, was designed. It was predicted ( $IC_{50} = 2.8$  nM), and subsequently observed ( $IC_{50} = 7.6$  nM), to be a potent HIV-1 protease inhibitor.

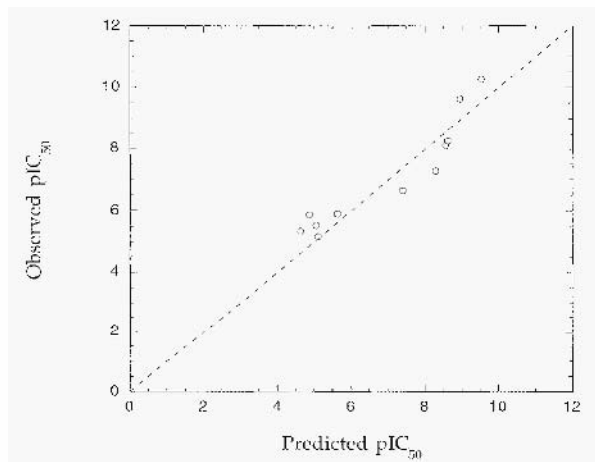


Fig. 5. Plot of predicted  $pIC_{50}$  versus observed  $pIC_{50}$  values for the inhibitor whose activity was predicted a priori (prior to assay). The line is one of unit slope, i.e. predicted  $pIC_{50} =$  observed  $pIC_{50}$ .



Fig. 6. An illustration of the symmetrical nature of the HIV-1 protease. The X-ray structures [41] of both orientations of Ro-31-8959, **70**, in cyan and yellow, are shown in the enzyme active site, represented by a green  $\alpha$ -carbon ribbon. The  $\beta$ -hairpin structures which loop across the front of the image are the flaps. The ordered water molecule which sits between the flaps and the inhibitor is displayed in the foreground as a ball-and-stick figure; the catalytic aspartic acids are displayed in the background as stick figures.

It was the first in a novel series of inhibitors that led to CRIVAN<sup>®</sup>, **80**, an HIV-1 protease inhibitor which was approved for the treatment of AIDS in March 1996. A

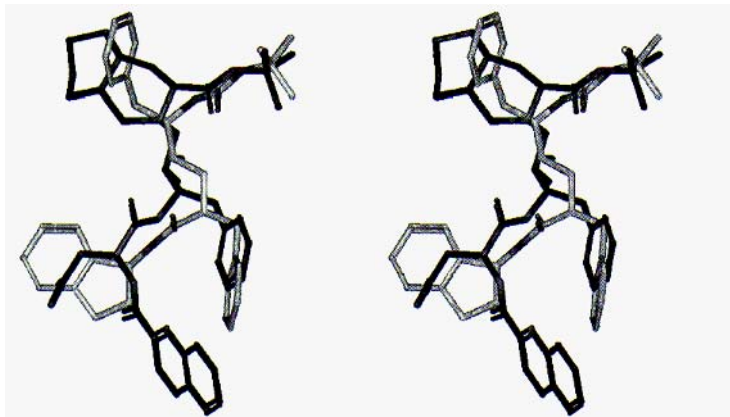


Fig. 7. A comparison of the model of **1** in light gray oriented in an  $N \rightarrow C$  fashion and the X-ray structure (reference [41]) of **70** in dark gray oriented in a  $C \rightarrow N$  fashion in the HIV-1 protease active site.

Table 5 Observed  $K_i$ ,  $pK_i$  and calculated enzyme-inhibitor interaction energies for Dupont-Merck cyclic urea HIV-1 protease inhibitors (reference [38]).

No.	R	$K_i$ (nM)	$pK_i$	$E_{inter}$ (kcal/mol)
34		4.7	8.328	-104.60
35		2.14	8.670	-111.46
36		3.1	9.509	-139.50
37		2.7	9.569	-149.92

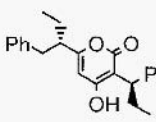
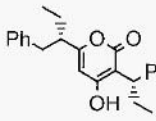
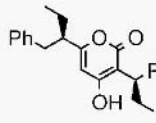
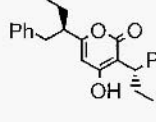
Table 6 Observed  $K_i$ ,  $pK_i$  and calculated enzyme-inhibitor interaction energies for Upjohn 4-hydroxycoumarin HIV-1 protease inhibitors (reference [39]).

No.	$R_1$	$R_2$	$R_3$	$K_i$ ( $\mu$ M)	$pK_i$	$E_{inter}$ (kcal/mol)
38	Me	Ph	H	3	5.523	-74.44
39	Et	Ph	H	1	6.000	-77.20
40	n-Pr	Ph	H	0.4	6.398	-79.75
41	n-Bu	Ph	H	0.5	6.301	-81.59
42	I-Pr	Ph	H	1	6.000	-78.41
43	cyclo-Pr	Ph	H	0.3	6.523	-77.56
44	t-Bu	Ph	H	1	6.000	-78.99
45	H	CH <sub>2</sub> CH <sub>2</sub> Ph	H	20	4.699	-75.98
46	H	CH <sub>2</sub> CH <sub>2</sub> CH <sub>2</sub> Ph	H	1.7	5.770	-81.96
47	Et	Ph	OMe	0.5	6.301	-81.80
48	cyclo-Pr	Ph	OMe	0.17	6.770	-82.60

Table 7 Observed  $K_i$ ,  $pK_i$  and calculated enzyme-inhibitor interaction energies for Upjohn 4-hydroxy-2-pyrone HIV-1 protease inhibitors (reference [39]).

No.	$R_1$	$R_2$	$K_i$ ( $\mu\text{M}$ )	$pK_i$	$E_{\text{inter}}$ (kcal/mol)
49	PhCH <sub>2</sub> CH <sub>2</sub>	Et	0.5	6.301	-88.75
50	PhCH <sub>2</sub> CH <sub>2</sub> CH <sub>2</sub>	Et	1	6.000	-91.50
51	n-Pr	Et	1.2	5.921	-81.08
52	PhCH <sub>2</sub> CH <sub>2</sub>	cyclo-Pr	0.32	6.495	-81.62
53	PhCH <sub>2</sub> CH <sub>2</sub> CH <sub>2</sub>	cyclo-Pr	0.23	6.638	-93.31
54	n-Pr	cyclo-Pr	0.32	6.495	-92.01
55	CH <sub>2</sub> (CH <sub>2</sub> (CH <sub>3</sub> ) <sub>2</sub> )	Et	0.07	7.155	-85.33
56	CH <sub>2</sub> (CH <sub>2</sub> (CH <sub>3</sub> ) <sub>2</sub> )	cyclo-Pr	0.07	7.155	-86.28
57	CH(CH <sub>2</sub> CH <sub>3</sub> )CH <sub>2</sub> Ph	cyclo-Pr	0.04	7.398	-96.24
58		cyclo-Pr	0.074	7.131	-100.94
59		cyclo-Pr	0.027	7.569	-97.95
60		cyclo-Pr	0.048	7.319	-98.14
61		cyclo-Pr	0.037	7.432	-93.97
62		cyclo-Pr	0.058	7.237	-94.54
63		cyclo-Pr	0.063	7.201	-95.61
64		cyclo-Pr	0.079	7.102	-98.69

Table 8 Observed  $K_i$ ,  $pK_i$  and calculated enzyme-inhibitor interaction energies for Upjohn 4-hydroxy-2-pyrone HIV-1 protease inhibitors (reference [39]).

No.	Structure	$K_i$ (nM)	$pK_i$ (nm)	$E_{int.}$ (kcal/mol)
65		105	6.979	-97.12
66		89	7.051	-97.16
67		45	7.347	-96.12
68		15	7.824	-95.98

comparison of the modelled structure of **76** and the X-ray structure [42] of **80** is depicted in Fig. 8.

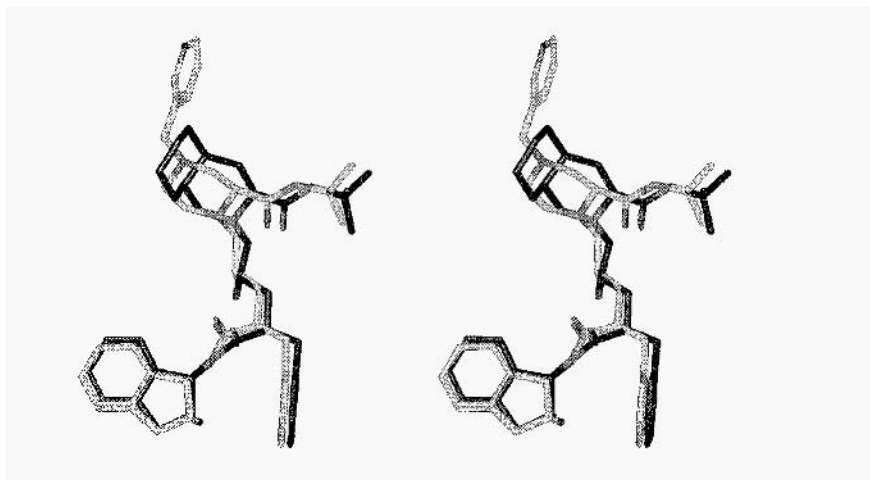


Fig. 8. A comparison of the model of **76** in dark gray and the X-ray structure (reference [42]) of **80** (CRIXIVAN<sup>®</sup>) in light gray as bound in the HIV-1 protease active site



Table 9 Calculated enzyme-inhibitor interaction energies and predicted and observed  $pIC_{50}$  values for the predicted set of HIV-1 protease inhibitors.

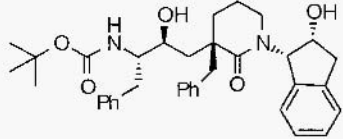
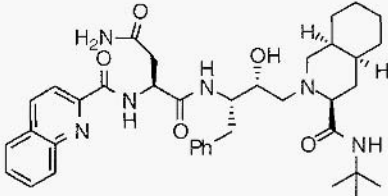
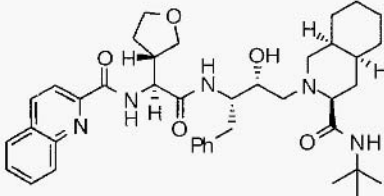
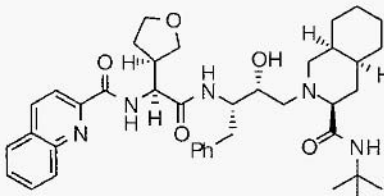
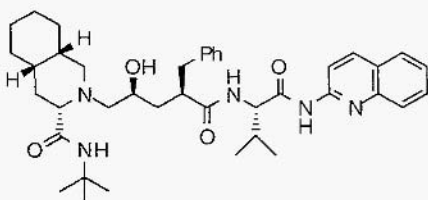
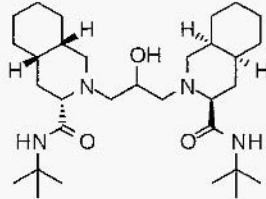
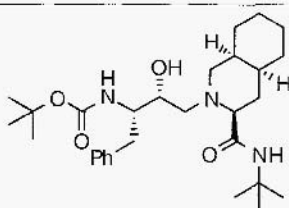
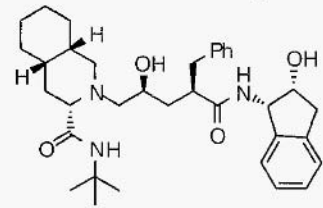
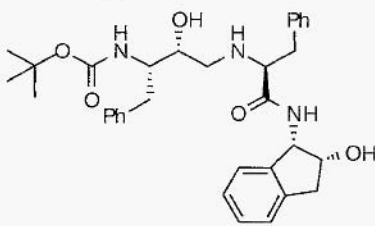
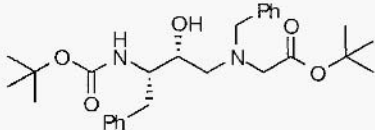
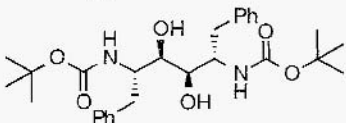
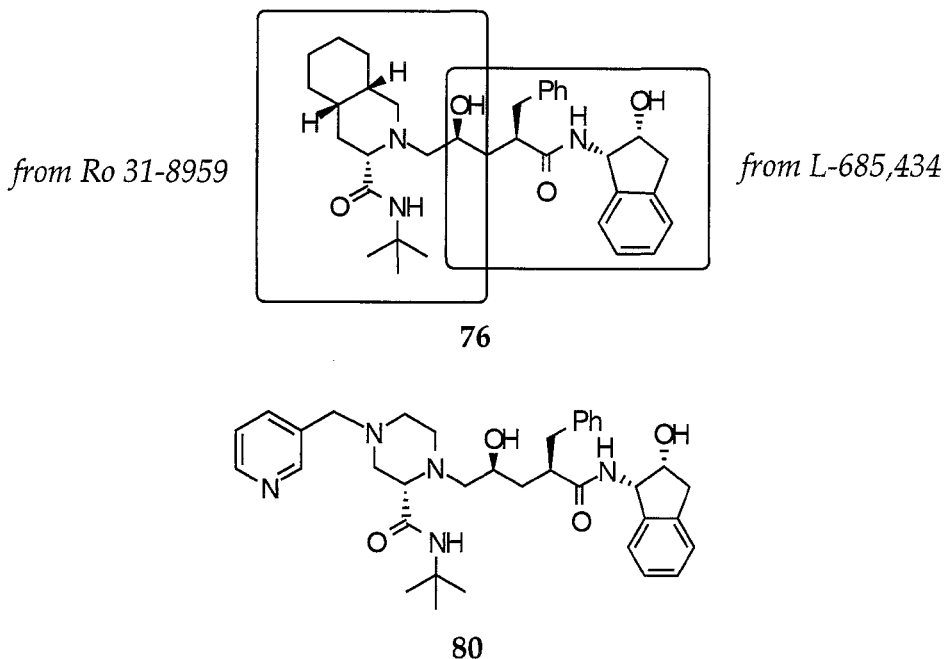
No.	Structure	$E_{inter}$ (kcal/mol)	Predicted $pIC_{50}$	Observed $pIC_{50}$
69		-125.9	5.628	5.897
70		-145.5	8.950	9.638
71		-143.6	8.628	8.268
72		-148.9	9.526	10.268
73		-141.6	8.289	7.277
74		-122.8	5.103	5.168

Table 9 (Continued)

No.	Structure	$F_{\text{inter}}$ (kcal/mol)	Predicted $\text{pIC}_{50}$	Observed $\text{pIC}_{50}$
75		-122.5	5.052	5.523
76		-143.2	8.560	8.116
77		-136.3	7.390	6.640
78		-120.0	4.628	5.328
79		-121.4	4.865	5.862

### 3. Conclusion

Direct computation of a ligand's affinity for its receptor — e.g. via free-energy perturbation calculations — can be costly and time-consuming. However, simple rapid computational alternatives can be employed to predict the affinity of a ligand for its receptor. These approaches include energy minimization of the ligand in a modelled or crystallographically determined receptor site, either alone or as one component of an energy equation or 3D QSAR. They can allow the prediction of activity *a priori* (i.e. prior to synthesis and assay), thus permitting pre-screening of ideas and prioritization of ligands for chemical synthesis. Due to their speed, they are amenable to use with large numbers



and types of compounds for a given receptor target and a given assay system. The obvious shortcomings of most energy minimization methods are neglect of solvation/desolvation, receptor flexibility, ligand flexibility and entropic effects. In general, they appear to be most successful when employed for a set of ligands that occupy similar sites on a receptor whose structure does not vary greatly from one ligand/receptor complex to another. While this sounds restrictive, it is often the case in the optimization of a ligand structure, either via single compound synthesis or the design of a modular or combinatorial library. We have demonstrated the utility of such an approach in the rational design of HIV-1 protease inhibitors, which facilitated the design of CRIXIVAN®, an HIV-1 protease inhibitor indicated for the treatment of AIDS.

## References

1. Kollman, P., *Free energy calculation: Applications to chemical and biochemical phenomena*, Chem. Rev., 93 (1993) 2395–2417.
2. Oprea, T.I., Waller, C.L. and Marshall, G.R., *3-dimensional quantitative structure-activity relationship of human immunodeficiency-virus-(1) protease inhibitors: 2. Predictive power using limited exploration of alternate binding modes.*, J. Med. Chem., 37 (1994) 2206–2215.
3. Waller, C.L., Oprea, T.I., Giolitti, A. and Marshall, G.R., *3-dimensional QSAR of human immunodeficiency-virus-(1) protease inhibitors: 1. A COMFA study employing experimentally-determined alignment rules*, J. Med. Chem., 36 (1993) 4152–4160.
4. Doweyko, A.M., *Three-dimensional pharmacophores from binding data*, J. Med. Chem., 37 (1994) 1769–1778.

5. Meng, E.C., Kuntz, I.D., Abraham, D.J. and Kellogg, G.E., *Evaluating docked complexes with the HINT exponential function and empirical atomic hydrophobicities*, J. Comput.-Aided Mol. Design. 8 (1994) 299–306.
6. Nauchitel, V., Villaverde, M.C. and Sussman, F., *Solvent accessibility as a predictive tool for the free-energy inhibitor binding to the HIV-1 protease*, Protein Science. 4 (1995) 1356–1364.
7. Wang, H. and Ben-Naim, A., *A possible involvement of solvent-induced interactions in drug design*, J. Med. Chem., 39 (1996) 1531–1539.
8. Wallqvist, A., Jernigan, R.L. and Covell, D.G., *A preference-based free-energy parameterization of enzyme-inhibitor binding: Applications to HIV 1 protease inhibitor design*, Protein Science. 4 (1995) 1881–1903.
9. Wallqvist, A. and Covell, D.G., *Docking enzyme-inhibitor complexes using a preference-based free-energy surface*, Proteins: Struct., Funct. Gene., 25 (1996) 403–419.
10. Verkhivker, G., Appelt, K., Freer, S.T., and Villafranca, J.E., *Empirical free energy calculations of ligand-protein crystallographic complexes: I. Knowledge-based ligand-protein interaction potentials applied to the prediction of human immunodeficiency virus 1 protease binding affinity*, Protein Eng., 8 (1995) 677–691.
11. Verkhivker, G.M. and Rejto, P.A., *A mean field model of ligand- protein interaction Implication for the structural assessment of human immunodeficiency virus type I protease complexes and receptor-specific binding*, Proc. Natl. Acad. Sci. USA. 93 (1996) 60–64.
12. Meng, E.C. Shoichet, B.K. and Kuntz, I.D., *Automated docking with grid-based energy evaluation*, J. comput. Chem., 13 (1992) 505.
13. Verlinde, C.L.M.J., Rudenko, G., and Wim, G.J.H., *In search of new lead compounds for trypanosomiasis drug design: a protein structure-based linked-fragment approach*, J., comput.-Aided Mol. Design, 6 (1992) 131–147.
14. Rotstein, S.H. and Murcko, M.A., *Groupbuild: A fragment-based method for de novo drug design* J. Med. Chem., 36 (1993) 1700–1710.
15. Böhm, H.-J., *The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure*, J. Comput.-Aided Mol. Design, 8 (1994) 243–256.
16. Bohacek, R.S.: McMartin, C., *De novo designed of highly diverse structures complementary to enzyme binding sites: Application to thermolysin*, In Rey nolds, C.H., Holloway, M.K. and Cox. H.K., (Eds.) Computer-aided molecular design: Applications in agrochemicals, materials and pharmaceuticals, ACS Symposium series 589. American Chemical Society, Washington, DC, 1995. pp. 82–97.
17. Head, R.D., Smythe, M.L., Oprea, T.I., Waller, C.L., Green, S.M. and Marshall, C.R., *VALIDATE: A new method for the receptor- based prediction of binding affinities of novel ligands*, J. Am. Chem. Soc., 118 (1996) 3959–3969.
18. Ortiz, A.R., Pisabarro, M.T., Gago, F. and Wade, R.C., *Prediction of drug binding affinities by comparative binding energy analysis*, J. Med. Chem., 38 (1995) 2681–2691.
19. Vajda, S., Weng, Z., Rosenfeld, R. and DeLisi, C., *Effect of conformational flexibility and solvation on receptor-ligand binding free energies*, Biochemistry, 33 (1994) 13977–13988.
20. King, B.L., Vajda, S. and Delisi, C., *Empirical-free-energy as a target function in docking and design: Application to HIV-1 protease inhibitors*, FEBS Lett., 383 (1996) 87–91.
21. Taylor, N.R. and von Itzstein, M., *Molecular modeling studies on ligand-binding to sialidase from influenza virus and the mechanism of catalysis*, J. Med. Chem., 37 (1994) 616–624.
22. De Winter, H.L., and von Itzstein, M., *Aldose reductase cis a target for drug design-Molecular modeling calculations on the binding of acyclic sugar substrates to the enzyme*, Biochemistry, 34 (1995) 8299–8308.
23. Grootenhuis, P.D.J. and van Galen, P.J.M., *Correlation of binding affinities with nonbonding interaction energies of thrombin-inhibitor complexes*, Acta Cryst., D51 (1995) 560–566.
24. Kurinov, I.V. and Harrison, R.W., *Prediction of New Serine Proteinase Inhibitors*, Structural Biology, 1 (1994) 735–743.
25. Sansom, C.E., Wu, J. and Weber, I.T., *Molecular mechanics analysis of inhibitor binding to HIV-1 protease*, Protein Eng., 5 (1992) 659–667.

26. Weber, I.T., Harrison, R.W., *Molecular mechanism calculations on HIV-1 protease with peptide-substrates correlate with experimental data*, Protein Eng., 9 (1996) 679–690.
27. Miertus, S., Furlan, M., Tossi, A. and Romeo, D., *Design of new inhibitors of HIV-1 aspartic pretease*, Chem. Phys., 204 (1996) 173–180.
28. Tossi, A., Furlan, M., Antcheva, N., Romeo, D. and Miertus, S., *Efficient inhibition of HIV-1 aspartic protease by synthetic, computer designed peptide mimetics*, Minerva Biotech., 8 (1996) 165–171.
29. Viswanadhan, V.N., Reddy, M.K., Wlodawer, A., Varney, M.D. and Weinstein, J.N., *An approach to rapid estimation of relative binding affinities of enzyme inhibitors: Application to peptidomimetic inhibitors of the human immunodeficiency virus type I protease*, J. Med. Chem., 39 (1996) 705–712.
30. Holloway, M.K., Wai, J.M., Halgren, T.A., Fitzgerald, P.M.D., Vacca, J.P., Dorsey, B.D., Levin, R. B., Thompson, W.J., Chen, L.J., Desolms, S.J., Gaffin, N., Ghosh, A.K., Giuliani, E.A., Graham, S.L., Guare, J.P., Hungate, R.W., Lyle, T.A., Sanders, W.M., Tucker, T.J., Wiggins, M., Wiscourt, C.M., Woltersdorf, O.W., Young, S.D., Darke, P.L., and Zugay, J.A., *A priori prediction of activity for HIV-1 protease inhibitors employing energy minimization in the active site*, J. Med. Chem., 38 (1995) 305–317.
31. Holloway, M.K. and Wai, J.M., *Structure-based design of human immunodeficiency virus-1 protease inhibitor: Correlating calculated energy with activity*, In Reynolds, C.H., Holloway, M.K., and Cox, H.K. (Eds.) Computer-aided molecular design: Applications in agrochemicals, materials, and pharmaceuticals, ACS Symposium series 589. American Chemical Society, Washington, DC. 1995, pp. 36–50.
32. Thompson, W.J., Fitzgerald, P.M.D., Holloway, M.K., Emini, E.A., Darke, P.L., McKeever, B.M., Schleif, W.A., Quintero, J.C., Zugay, J.A., Tucker, T.J., Schwering, J.E., Homnick C.F., Nunberg, J., Springer, J.P. and Huff, J.R., *Synthesis and antiviral activity of a series of HIV-1 protease inhibitors with functionality tethered to the PI or PI, phenyl substituents: X-ray crystal structure assisted design*, J. Med. Chem., 35 (1992) 1685–1701.
33. Hofmann, T., Hodges, R.S. and James, M.N.G., *Effect of pH on the activities of Penicillopepsin and Rhizopus pepsin and a proposal for the productive substrate binding mode in Penicillopepsin*, Biochemistry, 23 (1984) 635–643.
34. Hyland, L.J., Tomaszek, T.A., Jr. and Mcek, T.D., *Human immunodeficiency virus-1 protease: 2. Use of pH rate studies and solvent Kinetic isotope effects to elucidate details of chemical mechanism* Biochemistry, 30 (1991) 8454–8463.
35. CHARMM version 21.1.7b.; available from Molecular Simulations, Inc., Burlington, MA, U.S.A.
36. Available from W. Clark Still, Department of Chemistry, Columbia University, New York, U.S.A..
37. Allinger, N.L., *Conformational analysis 130. MM2: A hydrocarbon force field utilizing  $V_1$  and  $V_2$  torsional terms*, J. Am. Chem. Soc., 99 (1977) 8127.
38. Lam, P.Y.S., *Rational design of potent, bioavailable, nonpeptide cyclic ureas as HIV protease inhibitors*, Science, 263 (1994) 380–384.
39. Thaisrivongs, S., *Random and Rational: Lead Generation via Rational Drug Design and Combinatorial Chemistry*, New York, 19–20 October 1994.
40. Vacca, J.P., Fitzgerald, P.M.D., Holloway, M.K., Hungate, R.W., Starbuck, K.E., Chen, L.J., Darke, P.L., Anderson, P.S., and Huff, J.R., *Conformationally constrained HIV-1 protease inhibitors*, Bioorg. Med. Chem. Lett., 4 (1994) 499–504.
41. Ghosh, A.K., Thompson, W.J., Fitzgerald, P.M.D., Culberson, J.C., Axel, M.G., McKee, S.P., Huff, J.R. and Anderson, P.S., *Structure based design of HIV-1 protease inhibitor: Replacement of two amides and a  $10\pi$ -aromatic system by a fused bis-tetrahydrofuran*, J. Med. Chem., 37 (1994) 2506–2508.
42. Chen, Z., Li, Y., Chen, E., Hall, D.L., Darke, P.L., Culberson, J.C., Shafer, J. and Kuo, L.C., *Crystal structure at 1.9-Å resolution of human immunodeficiency virus (HIV) II protease complexed with L-735, 524; An orally bioavailable inhibitor of the HIV protease*, J. Biol. Chem., 269 (1994) 26344–26348.

# Rapid Estimation of Relative Binding Affinities of Enzyme Inhibitors

M. Rami Reddy<sup>a</sup>, Velarkad N. Viswanadhan<sup>b</sup> and M.D. Erion<sup>a</sup>

<sup>a</sup>*Metabasis Therapeutics, Inc., 9390 Towne Centre Drive, San Diego, CA 92121, U.S.A.*

<sup>b</sup>*Amgen, Inc., 1840 Deltavilland Drive, Mail Stop 14-2-8, Thousand Oaks, CA 91360, U.S.A.*

## 1. Introduction

Computer-assisted molecular modeling (CAMP) comprises a variety of computational methodologies intended to quantitatively or qualitatively describe molecular properties. In some cases, the field has advanced to a state where accurate predictions are possible (e.g. geometric and electronic properties of small molecules). In other cases, however, the properties are complex and require either advances in theory or substantial increases in computational power (e.g. protein folding). One application of CAMP that has received considerable attention over the past two decades entails its use as an aid in drug-design. Ideally, CAMP would provide rapid and accurate prediction of drug-target binding affinities such that a large and structurally diverse population of potential targets could be evaluated and thereby prioritized prior to chemical synthesis. In reality, methodologies have been advanced that either provide qualitative rank ordering of a large number of molecules in a relatively short period of time or, at the other extreme, generate quantitatively accurate predictions of relative binding affinities for structurally related molecules using substantial computing power. Consequently, techniques that increase speed without greatly compromising accuracy (or vice versa) are of value to drug-discovery programs.

Advances in protein crystallography and molecular simulations have greatly aided computer-assisted drug-design paradigms and the accuracy of their binding affinity predictions [12]. Methods of inhibitor design range from graphical visualization of the ligand in the binding site to calculation of relative binding affinities using molecular dynamics simulations in conjunction with the TCP approach [3,4]. Figure 1 shows a flowchart employed by drug-discovery groups for structure-based drug-design. Typically, the process begins by generating a working computational model from crystallographic data which includes the development of molecular mechanics parameters for non-standard residues, building any missing segments, assigning the protonation states of histidines, and orientation of carbonyl and amide groups of Asn and Gln amino acid residues based upon neighboring donor/acceptor groups. Inhibitor design is then aided by a variety of visualization tools. For example, hydrophobic and hydrophilic regions of the active site are readily identified by calculating the electrostatic potential at different surface grid points. Frequently, the information gained on the characterization of the active site is supplemented by analyses of ligand conformational energies.

These studies are often followed by an estimation of binding affinity which is performed at three different levels or complexity (Fig. 1) depending on computational power, time and resources, namely: (i) qualitative predictions based on docking/ visualization and molecular mechanics calculations; (ii) quantitative predictions based on

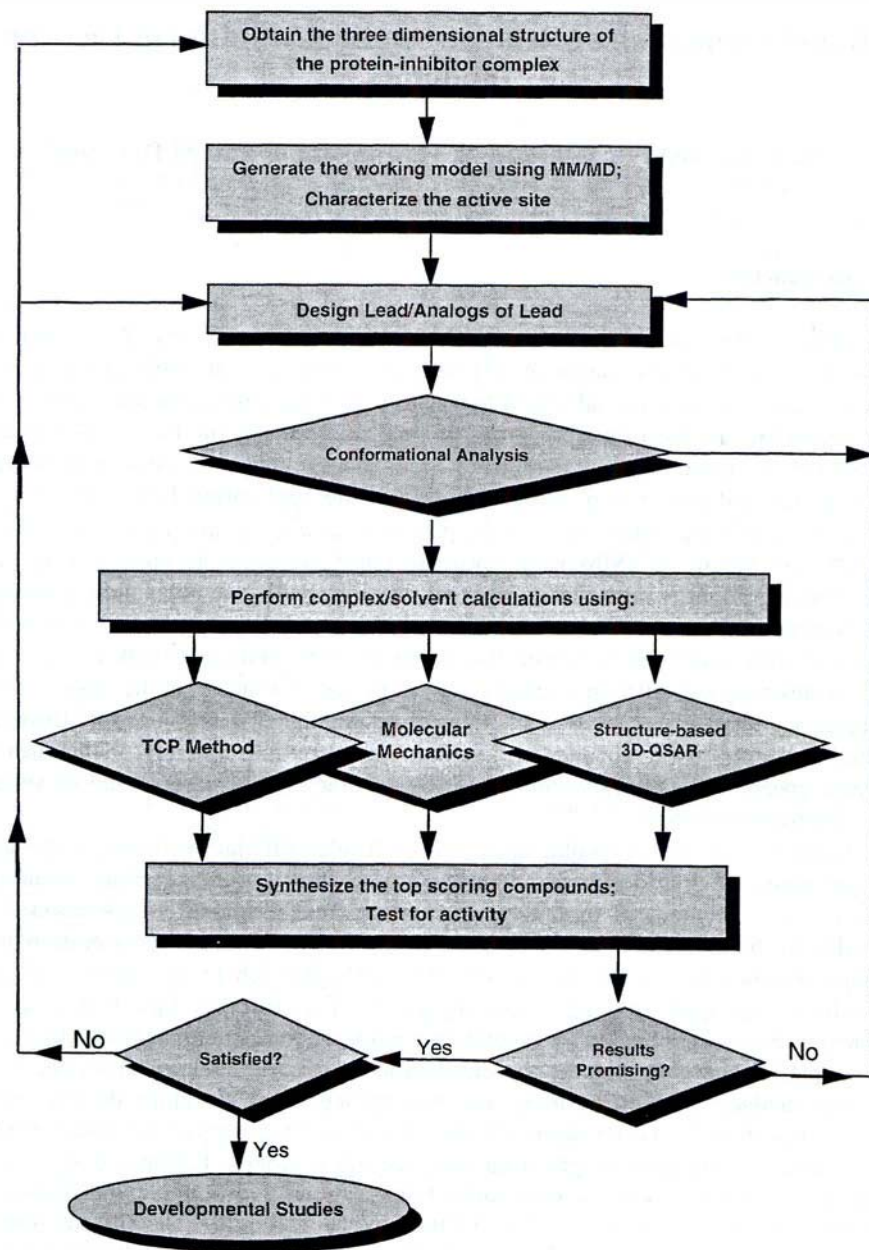


Fig. 1. Flowchart for the structure-based drug-design Paradigm.

TCP calculations: and (iii) semi-quantitative predictions based on regression methods that incorporate interaction variables (intra- and intermolecular interaction energies) and ligand properties (desolvation, log P, etc.). Based on the predicted binding affinities, top-scoring compounds are synthesized and tested for activity. In this review, applications of the TCP methods for predicting the binding affinity quantitatively are first discussed, followed by a review of methods used to predict ligand binding affinity qualitatively. These methods are then compared with more recent structure-based 3D QSAR approaches.

## **2. Free Energy Perturbation Methods**

Thermodynamic cycle perturbation (TCP) has been reported to predict accurately the relative binding affinities between two related inhibitors. Accurate prediction of ligand binding affinity will be dependent on the accuracy of each factor that effects binding. One important factor often neglected in most approximate methods used to determine relative binding affinities is desolvation. The importance of desolvation to binding affinity is clearly evident in calculations using the TCP method. For example, TCP calculations on transition state inhibitors bound to thermolysin carried out by Merz and Kollman [5] predicted that the replacement of an  $-NH$  group with a methylene group would not be detrimental to binding affinity, despite the loss of a good hydrogen bond between the NH and a backbone amide carbonyl. The principal reason shown clearly by the calculation was due to differences in inhibitor desolvation. This prediction, which was made ahead of biochemical measurements, was later confirmed experimentally [5].

More recently, several research groups used the TCP method and calculated relative binding differences of HIV-1 protease inhibitors. Reddy et al. [6] evaluated the relative binding affinities of JG365 and its analog (JC365A) lacking the penultimate valine residue. The calculated difference in the relative binding free energy ( $3.3 \pm 1.1$  kcal/mol) is in good agreement with the experimental value of  $3.8 \pm 1.3$  kcal/mol. This calculation showed that the loss of main-chain hydrogen bonds of the valine and its side-chain hydrophobic interactions with protein residues leads to considerable loss of potency for the JG365A. The observed binding preference for JG365 was explained by the stronger ligand–protein interactions, which dominate over an opposing contribution arising from the larger desolvation of JG365 relative to the truncated analog (JG365A). Ferguson et al. [7] computed the relative binding affinities of the S and R enantiomers of JG365 and reported good agreement with experiment. Tropshaw and Hermans [8] employed the TCP slow-growth method to estimate the relative binding affinities of the S and R enantiomers of JG365 and showed good agreement with experiment. Calculations by Rao et al. [9] for other HIV-1 protease inhibitors showed good agreement with experiment. Reddy et al. [10] evaluated a large set of related analogs using computer-assisted drug-design methods that combine molecular mechanics, dynamics, TCP calculations, inhibitor design, synthesis, biochemical testing and crystallographic structure determination of the protein–inhibitor complexes. The calculated relative binding free energies were successfully incorporated into the design of novel HIV-1 protease inhibitors. This study involved a large set of molecules whose relative binding



affinities were predicted using TCP method prior to synthesis, and were later confirmed by experimental measurements.

Though the TCP method is theoretically more accurate, it suffers from some practical limitations. Firstly, it requires substantial computational resources for the development of high-quality force-field parameters, as well as for the molecular dynamics/Monte Carlo calculations (a typical mutation takes about 1 week of CPU time on IBM RS6000/590 workstation). Secondly, the technique is limited to analogs that differ by relatively small structural changes, because it is often difficult to obtain well-converged results for larger mutations. Thirdly, exploring the available conformation space within a reasonable amount of time is often difficult for a molecule with many rotatable bonds (e.g. peptidomimetic inhibitors of HIV-1 protease). Consequently, the TCP method has not been widely used in drug-design. It should be noted, however, that with increased computational resources and improved force-field parameters, this method offers virtually unlimited potential.

### **3. Estimation of Binding Affinities**

While the TCP method is useful in the prediction of relative binding affinities of structurally similar inhibitors, real-life drug-design problems involve the calculation of relative binding affinities for inhibitors with a greater degree of structural dissimilarity. Hence, faster methods that can accommodate structural diversity are being developed. In some cases, predictions of inhibitor binding has been based solely on a visual analysis of structures without any calculations of binding energies. These methods [11] relied on graphical analysis of features such as steric, hydrophobic and electronic complementarity of the docked inhibitor to the target protein, the extent of buried hydrophobic surface and the number of rotatable bonds in the inhibitor. Quantitative descriptors, based on molecular shape [12] and grid-based energetics [13], have also proved to be useful. More advanced methods have used empirical scoring functions [14], derived from crystal structure data and experimental binding affinities. Though molecular mechanics methods would be expected to enhance the accuracy of these studies, results from a variety of studies have shown only modest success [15], which has been attributed to the large approximations involved in the analysis (e.g. solvent model used, lack of entropic terms, etc.).

Recently, some improvements have been realized through inclusion of an analysis of the binding conformations of new analogs by Monte Carlo (MC)/energy minimization (EM) methodology, and qualitative prediction of relative binding affinities using molecular mechanics calculations to evaluate interaction energies and ligand strain. For example, purine nucleoside phosphorylase (PNP) inhibitors [16–18] were designed and evaluated prior to synthesis using MC/EM techniques to derive the binding conformations and interaction energies that were used in predictions of relative binding affinities. The energy differences between inhibitors were thought to reflect relative binding affinities, since the structures were not highly dissimilar and contained few rotatable bonds, both of which suggested that differences in solvation and entropy would contribute minimally to binding affinity. Although the binding conformations

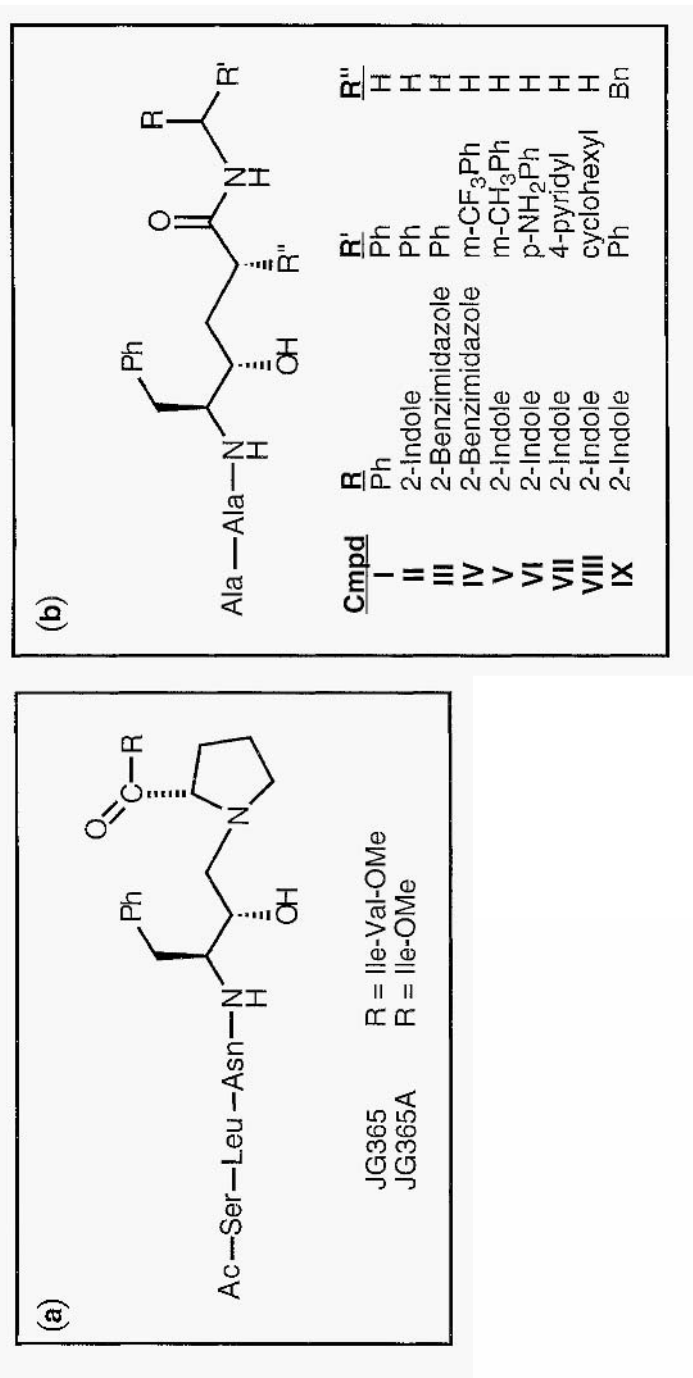


Fig. 2. Structures of the HIV-1 protease inhibitors considered in this work.

were accurately predicted in this study, interaction energies for some inhibitors proved to be less informative (data unpublished), presumably because of unaccounted factors such as desolvation and entropy. Another example using these techniques [19] was reported for the design of HIV-1 protease inhibitors. In this case, the binding modes of the putative/proposed inhibitors were obtained by carefully aligning them with the known crystal structures of inhibitors in the active site of the enzyme. These inhibitors, which are shown in Fig. 2, were then evaluated by performing minimization calculations both in solvent and in complex using the AMBER [20] force field.  $\Delta E$  scores are shown in Table 1 and were computed from two sets of minimizations.  $\Delta E_{intra}$  the difference in intramolecular interaction energies for the energy-minimized inhibitor in the bound and solution conformations.  $\Delta E_{inter}$  is the corresponding difference in intermolecular interaction energies. Since these energies are associated with significant uncertainties, the results were not expected to match the quality of results from TCP simulations, hence only a qualitative agreement in the overall trends with experimental results was expected using this method. Nevertheless, analysis of the energy differences offered a rationale for the preferential binding of JG365 and AG1007 to the HIV-1 protease relative to their respective analogs. The interaction energies of JG365 and AG1007 in the protein complex are greater than the corresponding interaction energies in the solvent state, and is presumably the reason these inhibitors have high affinity despite their larger desolvation energies (relative to their analogs, JG365A and AG1006). However, as shown in Table 1, these energy differences do not agree quantitatively with experimental binding free energies.

Table1 Differences in the calculated interaction energies of HIV-1 protease inhibitors (in kcal/mol) in the complex and solvated states.  $\Delta E_{intra}$  is the difference in intramolecular interaction energy between the complexed and solvated states of an inhibitor; and  $\Delta E_{inter}$  is the corresponding difference in intermolecular interaction energy. The sum of the two numbers ( $\Delta E_{tot}$ ) represents the total interaction energy of an inhibitor in the complex state relative to the solvated state. Score for the hydrophobic interaction (P) is also shown for each ligand complex (calculated using equation 10)

System	$\Delta E_{intra}$	$\Delta E_{inter}$	$\Delta E_{tot}$	P
JG36S	7.6	-74.5	-70.9	-2.63
JG365A	8.2	-59.9	-51.7	-2.41
<i>AG1007 series</i>				
I (AG1006)	2.7	-63.0	-60.3	-1.40
II (AG1007)	2.4	-65.0	-62.6	-1.53
III	4.9	-62.6	-57.7	-1.50
IV	3.9	-67.3	-63.4	-1.78
V	4.4	-65.7	-61.3	-1.86
VI	7.4	-67.0	-59.6	-1.43
VII	4.8	-66.6	-61.8	-1.38
VIII	5.9	-62.9	-57.0	-1.78
IX	7.2	-66.9	-59.7	-1.95

## **4. Regression Methods**

### *4.1. Introduction*

The methods discussed above either provide an accurate prediction of relative binding affinities (TCP), but with some significant constraints, or provide only qualitative trends for relative binding affinities across a more structurally diverse set of compounds. Ideally, methods that combine both of those features would greatly enhance the utility of computational methods to drug-design. Increased structural diversity, however, requires accurate calculation of additional factors that significantly impact the compounds' binding affinity. For example, the larger the difference in structure, the greater the chance that solvation, hydrophobic effects, conformational flexibility, etc. will influence relative binding affinities. Understanding the magnitude of each contribution is key to an accurate prediction. Since an equation that accurately incorporates each factor has not been derived, we cannot expect to calculate absolute binding free energies. One strategy for increasing the accuracy of traditional 3D QSAR approaches [2-26] would be to include crystal structure information from protein complexes [27,28]. For example, in the study by Marshall and co-workers [26], the use of crystal structure data in the determination of alignment rules and field-fit minimization is shown to enhance CoMFA [21] predictions. However, even these methods do not take full advantage of the crystallographic information available, since they do not include scoring functions that incorporate interaction variables.

Recently, the results of molecular mechanics calculations on protein complexes have been used in regression-based approaches for prediction of relative binding affinities in conjunction with other molecular properties. Unlike the 3D QSAR extensions mentioned earlier, where crystallographic information is used mainly for obtaining alignments for a molecular similarity analysis, these approaches use the information to aid in energetic calculations. Holloway and co-workers [29] predicted the binding affinities of a number of HIV-1 protease inhibitors using the Merck force field [30] for calculations of intermolecular interaction energies. This study, however, did not include other relevant molecular properties (such as desolvation and entropy contributions) that are likely to be important to binding affinity. Head et al. [31] developed a regression-based method (VALIDATE) for the receptor-based prediction of binding affinities of novel ligands. In this method, energy variables calculated by molecular mechanics using the generalized Born/ surface area (GB/SA) implicit water model are used as enthalpy of binding and properties such as complementary hydrophobic surface area are used to estimate the entropy of binding through heuristics. In this study, 51 diverse crystalline complex structures were used to assemble the training set and the predictive models employed both conventional regression and artificial neural networks. This study shows that a regression-based approach can qualitatively assess a diverse set of compounds and targets in a single model.

### *4.2. Computational Models and Energy Minimization Methods*

In an effort to develop a regression-based method for semi-quantitative prediction of relative binding affinities, we have examined a set of HIV-1 protease inhibitors (Fig. 2)

[19]. In our approach [19], the energy variables (intra and inter) calculated by performing molecular mechanics calculations both in complex and solvated states using an explicit water model, the strength of hydrophobic interaction and number of buried rotatable bonds of inhibitor are used to derive a regression equation for predicting relative binding free energy differences.

In the method, the protein, inhibitors (Fig. 2) and solvent waters were modelled using the AMBER [20] all-atom force field. The computational model of the JG365A (Fig. 2a) complex with HIV-1 protease was derived from the X-ray structure of the JG365 complex. Similarly, crystal structures of HIV-1 protease complexes of AG 1006 and AG I007 (Fig. 2b) were used to obtain the computational models of HIV-1 protease complexes with other inhibitors in the series. The hydrogens of the crystallographic water, the inhibitor and the protein dimer were added using the EDIT module of AMBER. Hydroxyl hydrogens were oriented to obtain the best possible geometry for hydrogen bonding. Electrostatic charges and parameters for the standard residues of the inhibitor model were taken from the AMBER database. For non-standard residues of all the inhibitors, electrostatic charges were fitted with CHELP [32] from *ab initio* [33] (single-point) HF/6-31G\* wave functions, using structures optimized at HF/3-21G\* level. One of the aspartic acids in the catalytic dyad (Asp 124) was protonated in all calculations. All equilibrium bond lengths, bond angles and dihedral angles for non-standard residues were calculated from *ab initio* (HF/3-21G\*) optimized structures. Missing force-field parameters were estimated from similar chemical species in the AMBER database. For the protein complex calculations, the solute was immersed in a large spherical water bath (of radius 25 Å) constructed from repeated cubes of extended simple-point charge (SPC/E) [34] water molecules which represented a snapshot from an MD simulation of liquid water [35]. The SPC/E rigid geometry model potential was used to model explicitly the solvent water. For the solvent calculations, the solute was solvated with SPC/E water in a rectangular box whose dimensions allowed a 10.0 Å layer of water to surround the solute atoms. Water molecules located less than 2.5 Å from the solute atom were removed.

Molecular mechanics calculations (energy minimizations) on all the structures were performed using the BORN module of the AMBER program. A four-stage protocol was set up for energy minimizations of the protein–inhibitor complexes, as well as solvated inhibitors. All the technical details are described in our earlier paper [19].

The minimized structures, both in the complexed and in the solvated states, were used for calculating the internal strain on the ligand upon binding ( $\Delta E_{bind}(intra)$ ) and relative interaction energy of the ligand in the complex versus solvent ( $\Delta E_{bind}(inter)$ ), as given by

$$\Delta E_{bind}(intra) = E_{com}(intra) - E_{sol}(intra) \quad (1)$$

where  $E_{com}(intra)$  refers to the intramolecular interaction energy of the ligand in the bound state, and  $E_{sol}(intra)$  refers to the intramolecular interaction energy of the ligand in aqueous medium.

Similarly,  $\Delta E_{bind}(inter)$  is calculated, using

$$\Delta E_{bind}(intra) = E_{com}(intra) - E_{sol}(inter) \quad (2)$$

where  $E_{com}(inter)$  refers to the intermolecular interaction energy of the ligand in the bound state. and  $E_{sol}(inter)$  refers to the intermolecular interaction energy of the ligand in the aqueous medium. Electrostatic and van der Waals interactions may be weighted differently by separating  $\Delta E_{bind}(inter)$  into electrostatic and van der Waals components. Scores for intra- and intermolecular components of relative differences in interaction energies for a pair of ligands L1 and L2 are given by

$$\Delta\Delta E_{bind}(intra : L1 \rightarrow L2) = \Delta E_{bind}(intra : L2) - \Delta E_{bind}(intra : L1) \quad (3)$$

$$\Delta\Delta E_{bind}(inter : L1 \rightarrow L2) = \Delta E_{bind}(inter : L2) - \Delta E_{bind}(inter : L1) \quad (4)$$

$\Delta\Delta E_{bind}(tot : L1 \rightarrow L2)$  from Eq. 5 is used to score the total relative difference in the binding energies of L1 and L2 to the protein

$$\Delta\Delta E_{bind}(tot : L1 \rightarrow L2) = \Delta\Delta E_{bind}(intra : L1 \rightarrow L2) + \Delta\Delta E_{bind}(inter : L1 \rightarrow L2) \quad (5)$$

This energy difference is used to calculate the relative binding free energy difference ( $\Delta\Delta G_{bind}$ ) between inhibitors L1 and L2. Earlier TCP calculations [6,10] on these compounds offered explanations for the relative binding affinities of several pairs of inhibitors, as shown in Table 2. Scores for relative intra- and intermolecular interaction energy differences (Eqs. 3–5) are listed in Table 2, along with corresponding TCP-calculated results. In order to understand the effect of different N-terminal groups on the final minimization results, energy calculations were performed using both asparaginequinoline and Ala-Ala as N-terminal groups for the compounds **2**, **4**, **5**, **7** and **9**. No significant energy differences were found between these two sets of calculations. Therefore, results presented in Tables 1 and 2 are consistent with our earlier published work [19]. These scores were used in developing the regression equations and calculating correlations.

Table 2 *Relative differences in the free energy of binding (kcal/mol) as observed experimentally ( $\Delta\Delta G_{bind}$  (expt.)) and by TCP simulations ( $\Delta\Delta G_{bind}$  (TCP)), are compared with the scores of relative enthalpic differences  $\Delta\Delta E_{bind}$  (calc.), and changes in hydrophobic interactions strength ( $\Delta P_{bind}$  (calc.)) for pairs analogous inhibitors of the HIV-1 protease*

'Mutation'	$\Delta\Delta G_{bind}$ (expt.)	$\Delta\Delta G_{bind}$ (TCP)	$\Delta\Delta E_{bind}$ (calc)	$\Delta P_{bind}$ (calc.)
JG365 $\rightarrow$ JG36SA	3.80	3.3	19.2	0.22
<i>AG1007 series</i>				
<b>II</b> $\rightarrow$ <b>I</b>	1.30	1.9	2.3	0.13
<b>II</b> $\rightarrow$ <b>III</b>	1.95	1.3	4.9	0.03
<b>II</b> * $\rightarrow$ <b>IV</b> *	-0.16	0.2	-0.8	-0.25
<b>II</b> * $\rightarrow$ <b>V</b> *	-0.06	0.4	1.3	-0.33
<b>II</b> $\rightarrow$ <b>VI</b>	-	1.1	3.0	0.10
<b>II</b> $\rightarrow$ <b>VII</b>	-	0.8	0.8	0.15
<b>II</b> * $\rightarrow$ <b>VIII</b> *	2.03	-	5.6	-0.25
<b>II</b> * $\rightarrow$ <b>IX</b> *	0.86	-	2.9	-0.42

\*Experimental values for these molecules are based on a different N-terminal group. an asparagine–quinoline moiety replacing Ala–Ala in the compounds **II**, **IV**, **V**, **VIII** and **IX**.

### 4.3. Results and discussion

In order to identify the variables that are important for the semi-quantitative prediction of HIV-1 protease inhibitors' binding affinity, a variety of regression equations were evaluated. Initially, intermolecular interaction energy of the inhibitor to the protein was used as the only regression variable. For a 'mutation' L1 to L2, this variable may be defined as

$$\Delta E_{bind}(inter : L1 \rightarrow L2) = \Delta E_{com}(inter : L2) - E_{com}(inter : L1) \quad (6)$$

A meaningful regression model, however, could not be obtained. presumably because the model lacked solvent contributions. By taking the role of solvent (desolvation) into account (Eq. 4), a good regression model was obtained

$$\begin{aligned} \Delta \Delta G_{bind}(expt) &= 0.17 \Delta \Delta E_{bind}(inter) + 0.90 \\ n &= 7; r = 0.89; RMS = 0.68 \end{aligned} \quad (7)$$

Inclusion of ligand strain (Eq. 5;  $\Delta \Delta E_{bind}(tot)$ ) further improved the regression model

$$\begin{aligned} \Delta \Delta G_{bind}(expt) &= 0.19 \Delta \Delta E_{bind}(tot) + 0.41 \\ n &= 7; r = 0.92; RMS = 0.57 \end{aligned} \quad (8)$$

By treating the scores for intermolecular (Eq. 4) and intramolecular (Eq. 3) interaction energies as independent variables in a multiple linear regression (MLR) model, slightly better statistics were obtained

$$\begin{aligned} \Delta \Delta G_{bind}(expt) &= 0.20 \Delta \Delta E_{bind}(inter) + 0.24 \Delta \Delta E_{bind}(intra) + 0.29 \\ n &= 7; r = 0.93; RMS = 0.64 \end{aligned} \quad (9)$$

It is clear that the present dataset is too small to develop MLR models with more than two independent variables, since MLR models need to be cross-validated. For larger datasets, further enhancement of the MLR model may be possible by separating the interaction energy score into electrostatic and van der Waals terms. and by inclusion of other related variables to represent the hydrophobic interactions. For example, hydrophobic interactions for a given ligand 'L' binding to a protein, as applied to the present set of HIV-1 protease inhibitors [19], can be represented by

$$P(int) = -k \sum_{i=1}^n \sum_{j=1}^m a_i b_j \exp(-D_{ij}) \quad (10)$$

where.  $P(int)$  is the score for hydrophobicity interaction. constants  $a_i$  and  $b_j$  represent atomic hydrophobicity constants assigned to atoms  $i$  and  $j$  in the ligand and protein, respectively, and  $D_{ij}$  is the distance between them. An exponential function is used to account for distance dependence of hydrophobic interactions. Earlier studies have demonstrated the usefulness of the exponential functional form in scoring ligand-protein hydrophobic interactions [18]. The difference in hydrophobic interaction energy

for a given modification of a ligand,  $L1 \rightarrow L2$ , was calculated in these studies using the following expression

$$\Delta P(L1 \rightarrow L2) = P(int:L2) - P(int:L1) \quad (11)$$

where  $P(int:L2)$  and  $P(int:L1)$  are the scores for hydrophobic interaction (Eq. 10) for ligands  $L2$  and  $L1$ , respectively. Using this variable (Eq. 11) alone, the following regression equation was obtained

$$\begin{aligned} \Delta\Delta G_{bind}(expt.) &= 2.77\Delta P + 1.78 \\ n &= 7; r = 0.72; RMS = 1.05 \end{aligned} \quad (12)$$

The best two-variable model (in terms of correlation coefficient) is obtained when  $\Delta\Delta E_{bind}(tot)$  (Eq. 5) and  $\Delta P$  (Eq. 11) are used as independent variables

$$\begin{aligned} \Delta\Delta G_{bind}(expt.) &= 0.16\Delta\Delta E_{bind}(tot) + 1.20\Delta P + 0.71 \\ n &= 7; r = 0.94; RMS = 0.58 \end{aligned} \quad (13)$$

For this model, a leave-one-out cross-validation gave  $r = 0.84$  and  $RMS = 0.81$ , indicating a reasonable predictive power for this model. When three independent variables ( $\Delta\Delta E_{bind}(inter)$  (Eq. 4),  $\Delta\Delta E_{bind}(intra)$  (Eq. 3) and  $\Delta P$  (Eq. 11)) are used, an almost perfect correlation ( $r = 0.99$ ;  $RMS = 0.38$ ) is obtained, though with 7 data points the dataset is too small for a three-variable model. However, it is worth noting that a leave-one-out cross-validation of the three-variable model gave better predictive statistics ( $r = 0.93$ ;  $RMS = 0.56$ ).

The above results show that interaction energy scores based on present molecular mechanics calculations are quite useful in developing regression equations for predicting the relative binding affinity, semi-quantitatively, of inhibitors to the HTV-1 protease enzyme. The predicted results correlated well with experimental measurements when solvation effects were included. Adding hydrophobicity variable,  $\Delta P$ , produced slight improvement in  $r$  ( $r = 0.92$  for the one-variable model versus 0.94 for the two-variable model) for this dataset. These results suggest that regression models offer a rapid way of semi-quantitatively predicting relative binding affinities of inhibitors within a congeneric series and, therefore, a viable alternative to the TCP method. Recently, a similar procedure was used to develop a multi-variable regression equation ( $r = 0.92$  and leave-one-out cross-validation gave  $r = 0.81$ ) for 25 inhibitors of Fructose-1,6-bisphosphatase, and applied successfully for designing and optimizing new inhibitors prior to synthesis [36].

## 5. Conclusion

The TCP method remains the most accurate method for calculating relative binding affinity of inhibitors to an enzyme. However, due to its relative complexity and computation-intensive nature, practical applications are restricted to analysis of structurally related inhibitors. Accordingly, there is a need for methods that enable rapid assessment of a



large number of structurally unrelated molecules in a reasonably accurate manner. Regression models, using energy variables based on molecular mechanics calculations in both explicit solvent and complex states, the strength of hydrophobic interactions and the number of rotatable bonds of the inhibitor, were shown to be valuable in the rapid estimation of the relative binding free energy differences semi-quantitatively between two inhibitors. As shown with HIV-1 protease [19] and later with Fructose-1,6-bisphosphatase inhibitors [36], multivariate models that account for these properties are useful as a rapid computational alternative to TCP calculations. These regression-based models will continue to evolve and become more accurate as: (1) force-field parameters become more refined. (2) other variables important for binding are included. (3) methods for estimating relative binding entropy changes improve. (4) docking and scoring procedures improve and (5) average molecular dynamics simulations are used to obtain energy variables.

## References

1. Appelt, K., Bacquet, R.J., Bartlett, C.A. Booth. C.L., Freer. S.T., Fuhry, M.A.M., Gehring, M.R., Hermann. S.M. Houlard, E.F., Janson, C.A., Janes, T.R., Kan. C., Kathardekar, V., Lewis, K.K., Marzoni. G.P., Matthew. D.A., Molhr, C W.M.E., Morse, C.A., Oatley, S.J., Opden, R.O., Reddy, M.R., Reich, S.H., Schoettlin, W.S., Smith. W.W., Varney, M.D., Villafranca, J.E., Ward. R.W., Webber, S., Webber, S.E., Welsh, K.M. and White. J.. *Design of enzyme inhibitors using iterative protein crystallographic analysis*, J. Med. Chem., 34 (1991) 1925–1934.
2. Montgomery, J.A., Niwas, S., Rose. J.D., Secrist, J.A., Sudhakar Babu. Y., Bugg. C.E., Erion, M.E., Guida. W.C. and Ealick, S.E., *Structure based design of inhibitors of purine nucleoside phosphorylase. I. 9-(arylmethyl) derivatives of 9-deazaguanine*, J. Med. Chem., 36 (1993) 55–69.
3. Beveridge, D.L. and DiCapua. F.M., *Free energy via molecular simulation*, Ann. Rev. Biophys. Biophys. Chem., 18 (1989) 431–492.
4. McCammon, J.A., *Free energy from simulation*. Current Opinion in Structural Biology, 1 (1991) 196–200.
5. Mezc, K.M. and Kollman, P.A., *Free energy perturbation simulation of the inhibition of thermilysin: prediction of the free energy of binding of a new inhibitor*, J. Am Chem. Soc., 111 (1989) 5649–5658.
6. Rami Reddy, M., Viswanadhan. V.N. and Weinstein, J.N., *Relative free energy differences in the binding free energies of human immunodeficiency virus 1 protease inhibitors: A thermodynamic cycle perturbation approach*. Proc. Natl. Acad. Sci. USA. 88 (1991) 10287–10291.
7. Ferguson. D.M., Radmer, R.J. and Kollman, P.A., *Determination of the relative binding free energies of peptide inhibitors to the HIV-1 protease*, J Riled. Chem., 34 (1991) 2654–2659.
8. Tropshaw. A.J. and Hermans, J.. *Application of free energy simulations to the binding of a transition state analogue inhibitor to HIV-1 protease*, Prot. Eng., 5 (1992) 29–33
9. Rao, B.G., Tilton, R.F. and Singh, U.C., *Free energy perturbation studies on inhibitor binding to HIV-1 protease*. J. Am. Chem. Soc., 114(1992) 4447–4452
10. Rami Reddy, M., Varney. M.D., Kalish, V., Viswanadhan V.N. and Appelt, K., *Calculation of relative difference in the binding freeenergies of HIV-1 protease inhibitors: A thermodynamic cycleper turbation approach*, J. Med. Chem., 114 ( 1994) 10117–10122.
11. Bohacek, R.S. and McMartin, C., *Definition and display of steric, hydrophobic, and hydrogen-bonding properties of ligand binding sites in proteins using Lee and Richards accessible surface: Validation of a high-resolution graphical tool for drug design*, J. Med. Chem., 35 (1992) 1671–1684.
12. Kurtz, J.D., Meng, E.C. and Shoichet, B.K., *Structure-based molecular design*. Ace. Chem. Res., 27 (1994) 117–123.
13. Goodford, P.J., *A computational procedure for determining energetically favorble binding sites on biologically important macromolecules*, J. Med. Chem., 28 ( 1985) 849–857.

14. Boehm, H.-J., *The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure*, J. Comput.-Aided Mol. Design, 8 (1994) 243–256.
15. Sansom, C.E., Wu, J. and Weber, I.T., *Molecular mechanics analysis of inhibitor binding to HIV-1 protease*. Protein Eng., 5 (1992) 659–667.
16. Montgomery, J.A., Erion, M.D., Niwas, S., Rose, J.D., Secrist, J.A., Babu, S.Y., Bygg, C.E., Guida, W.C. and Ealick, S.E., *Structure-based design of inhibitors of purine nucleoside phosphorylase. 2. 9-(armenthyl) derivatives of 9-deazaguanini*, J. Med. Chem., 36 (1993) 55–69.
17. Erion, M.D., Montgomery, J.A., Niwas, S., Rose, J.D., Ananthan, S., Allen, M., Secrist, J.A., Babu, S.Y., Bugg, C.E., Guida, W.C. and Ealick, S.E., *Structure-based design of inhibitors of purine nucleoside phosphorylase. 3. 9-artmenthyl derivatives of 9-deazaguanine substituted on the methylene-group*, J. Med. Chem., 36 (1993) 3771–3783.
18. Secrist, J.A.I., Niwas, S., Rose, J.D., Babu, S.Y., Bugg, C.E., Erion, M.D., Guida, W.C., Ealick, S.E. and Montgomery, J.A., *Structure-based design of inhibitors of purine nucleoside phosphorylase. 2. 9-allycyclic and 9-heteroallycyclic derivatives of 9-deazaguanine*, J. Med. Chem., 36 (1993) 1847–1854.
19. Viswanadhan, V.N., Rami Redy, M., Wlodawer, A., Varney, M.D. and Weinstein, J.N., *An approach to rapid estimation of relative binding affinities of enzymes inhibitors: application to peptidomimetic inhibitors of the human immunodeficiency virus type I protease*, J. Med. Chem., 39 (1996) 705–712.
20. Singh, U.C., Weiner, P.K., Caldwell, J.K. and Kollman, P.A., *AMBER 3.3*. University of California, San Francisco, CA, U.S.A., 1986.
21. Cramer, R.D., Pattercon, D.E. and Bunce, J.D., *Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins*, J. Am. Chem. Soc., 110 (1988) 5959–5967.
22. Viswanadhan, V.N., Ghose, A.K. and Weinstein, J.N., *Mapping the binding site of the nucleoside transporter protein: A 3D-QSAR study*, Biochim. Biophys. Acta, 1039 (1991) 356–366.
23. Debnath, A.K., Hansch, C., Kim, K.H. and Martin, Y.C., *Mechanistic interpretation of the genotoxicity of nitrofurans (antibacterial agents) using quantitative structure-activity relationships and comparative molecular field analysis*, J. Med. Chem., 36 (1993) 1007–1016.
24. Depriest, S.A., Mayer, D., Naylor, C.B. and Marshall, G.R., *3D QSAR of angiotensin-converting enzyme and thermolysin inhibitors: A comparison of CoMFA models based on deduced and experimentally determined active site geometries*, J. Am. Chem. Soc., 115 (1993) 5372–5384.
25. Kubinyi, H., *3D QSAR in drug design: Theory, methods and applications*, ESCOM Science publishers B.V., Leiden, 1993.
26. Martin, Y.C., *A fast new approach to pharmacophore mapping and its application to dopaminergic and benzodiazepine agonists*, J. Comput.-Aided Mol. Design, 7 (1993) 83–102.
27. Waller, C., Oprea, T.I., Giolitti, A. and Marshall, G.R., *Three-dimensional QSAR of human immunodeficiency virus (I) protease inhibitors: 1. A study employing experimentally-determined alignment rules*, J. Med. Chem., 36 (1993) 41 52–41 60.
28. Oprea, T.I., Waller, C.L. and Marshall, G.R., *3D-QSAR of human immunodeficiency virus (I) protease inhibitors III: Interpretation of CoMFA results*. Drug Design and Discovery. 12 (1994) 29–51.
29. Holloway, K., Wai, J.M., Halgren, T.A., Fitzgerald, P.M., Vacca, J.P., Dorsey, B.D., Levin, R.B., Thompson, W.J., Chen, J.L., deSolms, J.S., Gaffin, N., Ghosh, A.K., Giuliani, E.A., Graham, S.L., Guare, J.P., Hungate, R.W., Lyle, T.A., Sanders, W.M., Tucker T.J., Wiggins, M., Wiscourt, C.M., Wolterdorf, O.W., Young, S.D., Darke, P.L. and Zugay, J.A., *A priori prediction of activity for HIV-1 protease inhibitors employing energy minimization in the active site*, J. Med. Chem., 38 (1995) 305–317.
30. Halgren, T.A., *MM forcefield*, J. Comput. Chem., 17 (1996) 490.
31. Head, R.D., Smythe, M.L., Oprea, T.I., Waller, C.L., Green, S.M. and Marshall, G.R., *VALIDATE: A new method for the receptor-based prediction of binding affinity of novel ligands*, J. Am. Chem. Soc., 118 (1996) 3059–3969.
32. Chirlian, L.E. and Franel, M.M., *Atomic charges derived from electrostatic potentials: A detailed study*, J. Comp. Chem., 8 (1987) 894–905.
33. Frisch, M.J., Head-Gordon, M., Schelegel, H.B., Raghavachari, K., Binkley, J.S., Gonzales, C., Defrees, D.J., Fox, D.J., Whiteside, R.J., Seeger, R., Melius, C.F., Baker, J., Martin, R., Kahn, L.R., Stewart, J.J., Fluder, E.M., Topiel, S. and Pople, J.A., *GAUSSIAN 88: Gaussian*. Pittsburgh, PA, U.S.A., 1988.

34. Berendsen, H.J.C., Grigera, J.R. and Straatsma, T.P., *The Missing Term in Effective Pair Potentials*, J. Phys. Chem., 91 (1987) 6269–6271.
35. Rami Reddy, M. and Berkowitz, M., *Dielectric constant of SPC/E water*, Chem. Phys. Lett., 155 (1989) 173–176.
36. Rami Reddy, M., Viswanadhan. V.N. and Erion, M.D., *Rapid estimation of relative binding affinities of enzyme inhibitors: Application to inhibitors of Fructose-1,6-bisphosphatase*, J. Med. Chem. (to be submitted).

# Binding Affinities and Non-Bonded Interaction Energies

Ronald M.A. Knegt<sup>1</sup> and Peter D.J. Grootenhuys<sup>2</sup>\*

*Department of Computational Medicinal Chemistry,  
N. V. Organon, P.O. Box 20, 5340 BH Oss, The Netherlands*

## 1. Introduction

The association of two molecules to form a stable complex involves a delicate interplay of electrostatic interactions, such as hydrogen bonds and ionic contacts, the matching of hydrophobic surfaces and entropic effects [1]. A schematic representation of the factors that play a role in molecular association is given in Fig. 1. These factors determine the change in free energy ( $\Delta G$ ) upon complexation which is the resultant of a change in enthalpy ( $\Delta H$ ) and entropy ( $\Delta S$ ). The logarithm of the dissociation or inhibition constant ( $pK_i$ ) of the formed complex is linearly related to the change in binding free energy:

$$\Delta G = \Delta H - T\Delta S = RT \ln(K_i) = -2.303 RT pK_i \quad (1)$$

in which  $T$  denotes the absolute temperature and  $R$  the gas constant. The linear relationship between  $\Delta G$  and  $pK_i$  poses the challenge to correlate energetic terms, which can be calculated from physical principles, with experimentally determined binding constants [2,3]. Especially in cases where the actual synthesis and evaluation of large numbers of compounds is too time-consuming or expensive, theoretical approaches capable of accurately predicting binding constants would be extremely useful [4]. Even if this is not the case, such methodology may be used to prioritize the order of screening and synthesis. However, the derivation of such a 'scoring function' has proven to be a rather challenging task. In fact, the lack of reliable methods for predicting binding affinities still presents one of the major bottlenecks in structure-based drug design [2,3,5].

In this chapter, recent research will be reviewed aimed at predicting binding constants and energies of molecular complexes for which the three-dimensional (3D) structure is known. Applications using comparative molecular field analysis (CoMFA) [6,7] usually do not utilize experimental structural information regarding the receptor and thus fall outside of the scope of this review. It has been demonstrated, however, that in some cases the use of experimentally determined conformations of bound ligands for superpositioning can yield CoMFA models with improved predictive properties (see e.g. reference [8]). Although the methodology discussed here can be applied generally, we focus in this review on complexes between biopolymers, such as proteins and DNA. on the one hand, and small organic molecules or polypeptides, on the other hand. Such systems are particularly relevant for the discovery of enzyme inhibitors and receptor (ant)agonists with therapeutical applications. In the search for novel drug molecules, the

---

\* To whom correspondence should be addressed.

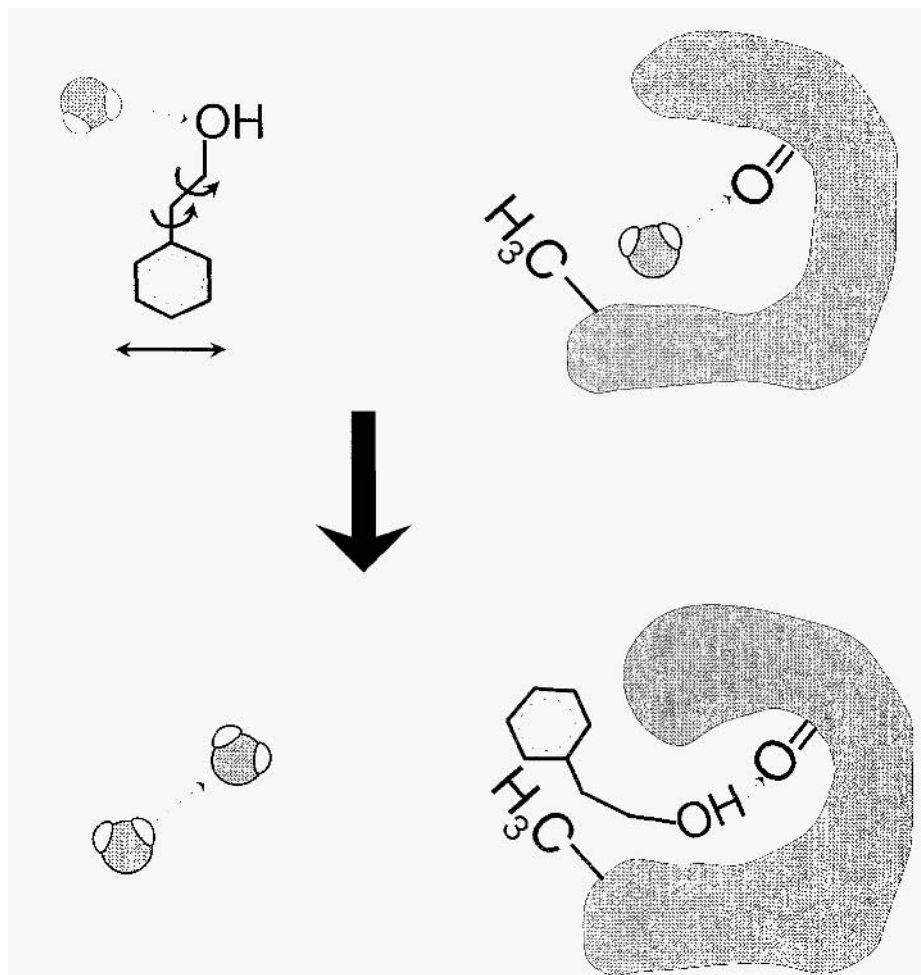


Fig. 1. Schematic representation of the events and thermodynamic factors involved in ligand-receptor complexation. Upon binding to its receptor, the translational and rotational entropy of the entire ligand is reduced and part of its solvent accessible surface becomes buried. Interactions with the solvent are lost and replaced by hydrogen-bond and van der Waals interactions at the newly formed protein-ligand interface. This results in the (partial) freezing of the internal rotational degrees of freedom of the ligand. Similarly, part of the receptor surface is desolvated and interactions with the ligand are established possibly resulting in (local) conformational changes in the receptor. Finally, entropy is gained by the release of ordered solvent molecules from the interacting ligand and receptor surfaces.

accurate prediction of the binding properties of lead compounds by computational techniques can potentially reduce and guide experimental work. General applicability, computational speed and, of course, the ability to deliver accurate predictions of the binding free energy are the desired characteristics of such a scoring function.

The most commonly used technique for the simulation of intermolecular complexation at the atomic level utilizes molecular mechanics force-fields, since quantum

mechanical calculations (although they have been reported [9]) are mostly unpractical for larger systems. Force-field based approaches have a solid foundation in physical chemistry and the potential energy equations and parameters used to estimate interaction enthalpies have been optimized to model quantum chemical and experimental results on the energetics and dynamics of various systems. Similarly, numerical solutions to the Poisson-Boltzmann equation [10] can provide estimates of binding free energies of systems involving charged molecules. Recently, a number of more empirical approaches have been developed. In some cases, such approaches cannot easily be interpreted on a physical level; they are solely aimed at providing the best possible fit between experimental and calculated binding energies. In the following sections we will discuss the methodology, results, advantages and shortcomings of such strategies.

## **2. Force-Field Based Methods**

Free energy perturbation (FEP) calculations are potentially the most accurate theoretical approach to the calculation of binding constants and free energies from structures of receptor-ligand complexes [11]. Unfortunately, the calculation of a single binding constant requires several molecular dynamics simulations with long computation times. Moreover, since FEP calculations can be fairly sensitive to the starting geometry and parameters of the complex at hand as well as the precise protocol parameters, their practical use requires careful preparation of input data and analysis of the results. Although good results have been obtained for the design of HIV protease inhibitors [12], FEP simulations are, in general, still too time-consuming to be practically applicable in the prediction of binding constants of more than only a handful of chemically related ligands. Research in this field continues. However, and approximate methods have been reported allowing the estimation of relative binding free energies from only one or two molecular dynamics simulations [13,14]. Although these protocols provide a significant reduction of the time required to calculate binding free energies, they remain too slow for the evaluation of thousands of compounds and, in addition, only structurally minor changes can be evaluated.

Since the direct calculation of binding free energies is still cumbersome, approximations are needed to allow fast scoring of ligand-binding modes using molecular mechanics force-fields. Of the terms constituting the free energy, the entropy is still the most difficult term to be predicted due to its intricate dependence on the dynamics and surface properties of the molecules under study [1,15]. The omitting of entropic factors eliminates the need to generate trajectories of conformations with molecular dynamics and yields a much more tractable scoring function. Even though this reduces the problem to energy minimization and evaluation of a single ligand-receptor complex, various parameters can still be varied in order to improve the correlation with measured binding data. Recently, a number of studies have been reported that systematically investigated the correlation between the interaction energy of small molecule inhibitors with proteins and their experimentally determined binding constants. These will be discussed in more detail in the following paragraphs.

An early application of force-field energies in the ranking of receptor-bound small molecules was the incorporation of a grid-based representation of the AMBER force-

field [16] in the molecular docking program DOCK [17]. Originally, only a simple function was used to account for van der Waals bumps between ligand and receptor [17], which was improved later to reflect shape Complementarity [18] and eventually chemical complementarity by means of force-field scoring [19]. In order to optimize and evaluate docked molecules, a rigid body minimization of the ligand is performed and ranking is based on the minimum interaction energy. Rather surprisingly, even approaches with relatively crude approximations, such as evaluating rigid ligands and a rigid receptor using a grid representation of the force-field, can still reproduce correct binding orientations. A typical example is the application of DOCK to a homology model structure of cercarial elastase [20]. Here force-field scoring ranked the best two inhibitors identified in the fine chemicals directory of 55.313 compounds as 85th ( $K_i = 3 \mu\text{M}$ ) and 627th ( $K_i = 6 \mu\text{M}$ ), while three other inhibitors with  $K_i$ 's  $< 100 \mu\text{M}$  were ranked 122nd, 918th and 561st. This indicates that considerable manual filtering is still required to select compounds for testing from DOCK output. Although the correlation between force-field score and binding constant appears to be low, the successful identification of diverse micromolar inhibitors given the inaccuracies of a homology model for the enzyme and rigid structures for the ligands remains an impressive result.

In order to rationalize the selection of parameters for force-field based scoring of enzyme inhibitors, Grootenhuis and van Galen [21] applied several different energy minimization and evaluation protocols to a set of 35 non-covalently bound thrombin inhibitors. Using starting coordinates from crystal structure determinations or hand-built models, each complex was energy minimized using the CHARMM force-field [22] and its non-bonded interaction energy was evaluated. The influence of protein flexibility and the balance between van der Waals and electrostatic contributions on the correlation between calculated and experimentally determined binding constants was tested for both the minimization and interaction energy evaluation steps. Protein flexibility was varied between a completely flexible enzyme and a completely rigid protein via protocols where harmonic constraints of increasing strength were applied to the active site, while the rest of the protein remained fixed. In all cases, the ligand remained flexible. The balance of the electrostatic and van der Waals contributions to the total force-field score was modified by using different effective dielectric functions  $D (= 4\epsilon r)$  in the CHARMM non-bonded potential energy function:

$$E_{nonb} = \sum(A/r^{12} - B/r^6 + q_1q_2/Dr) \quad (2)$$

where  $A$  and  $B$  denote the Lennard-Jones parameters,  $q_1$  and  $q_2$  partial atomic charges and  $r$  the interatomic distance. The electrostatic interactions were either switched off or stepwise increased to vacuum conditions ( $\epsilon = 1$ ) via two distance-dependent dielectric functions for the energy minimization and interaction energy evaluation:  $\epsilon = 4r$  and  $\epsilon = r$ , respectively. It was found that a protocol in which the entire enzyme was held fixed during minimization without electrostatics, followed by evaluation using a dielectric function  $\epsilon = r$ , gave the best fitted regression coefficient ( $r^2$ ) of 0.66 with a standard deviation ( $SD$ ) of 0.97  $pK_i$  units between energies and  $pK_i$ 's. Interestingly, receptor flexibility and electrostatic interactions appear to be of minor importance

during the minimization phase in which an optimal fit between protein and ligand needs to be found. Scaled-down electrostatics were found to be essential, however, in the evaluation phase.

A similar analysis was reported by Kurinov and Harrison [23], who used molecular dynamics to search the conformational space of phenylethylamine analogs. Scoring was done in vacuum using the UFF force-field [24] and Fourier-Green functions to accelerate the computations. For the 15 compounds examined, a  $r^2$  of 0.58 with the experimentally determined  $pK_i$ 's was found, which improved to 0.74 when one outlier was removed. It was found that especially the use of an all-atom representation and the evaluation of all electrostatic interactions without the use of a distance cutoff were critical to obtaining a good correlation.

Several groups have attempted to expand the molecular mechanics force-fields that are ordinarily used to include terms for desolvation and hydrophobic interactions. Luty and co-workers [25] added a desolvation term to a grid representation of the AMBER [16] force-field of the form:

$$E_{desolvation} = \sum \sum -S_i f_i \exp(-r_{ij}^2 / 2\sigma^2) \quad (3)$$

where the summation is performed over all ligand and receptor atom pairs. In Eq. 3,  $S_i$  and  $f_i$  are the respective solvation parameter and fragmental volume of a mobile atom  $i$  which are multiplied by a Gaussian envelope function of width  $\sigma$  describing the volume of displaced water. Although results were shown for the trypsin–benzamidine complex only, the modified force-field was able correctly to dock benzamidine into the partially flexible active site. Results comparing the calculated and measured binding properties of protease inhibitors have not yet been reported. Yet another approach was described by Viswanadhan et al. [26], who added an additional function describing hydrophobic interactions to the AMBER [16] force-field:

$$P_{hydrophobic} = -k \sum \sum a_i b_j \exp(-r_{ij}) \quad (4)$$

where  $P$  is a scoring function which ranks hydrophobic contacts between receptor and ligand atoms described by the constants  $a_i$  and  $b_j$  respectively, at a mutual distance  $r_{ij}$ . Its contribution to the overall score is scaled by a factor  $k$ . By comparing the energies of solvated and complexed states of 11 HIV-protease inhibitors relative binding energies were obtained which were compared with experimental results. Interestingly, energies obtained using only the standard force-field energies already yielded a  $r^2$  of 0.85 ( $SD = 0.57$  kcal/mol), which only slightly increased to 0.88 ( $SD = 0.58$  kcal/mol) when the hydrophobicity term was included. However, the hydrophobicity score can be useful in predicting the outcome of modifications of lead compounds which reduce or increase the hydrophobic character of the molecule.

An interesting extension of the use of molecular mechanics energies in the prediction of binding properties was reported by Ortiz et al. [27] and termed comparative binding energy (COMBINE) analysis (see also the chapter by R.C. Wade in this volume, p. 19 ff). In this approach, inhibitors of human synovial fluid phospholipase A<sub>2</sub> (HSP-PLA<sub>2</sub>) were divided in small sections of which the bonded and non-bonded interaction



with individual residues in the binding site was evaluated. Also the intramolecular interactions among inhibitor fragments and active site residues were calculated and all energies were stored in a matrix. A column containing the experimentally determined activities (in percentages) was added to the matrix which was subsequently reduced using partial least squares and D-optimal design techniques [28]. This procedure allows for the selection of only those energy terms that contribute significantly to differences in binding free energy. The final model contained 50 energy terms (only 2% of the original matrix) which yielded a  $r^2$  of 0.92 ( $SD = 6.23\%$  activity). Direct correlation of binding energies with activities yielded only a correlation coefficient of 0.21 which is partially due to errors in modelling the inhibitor conformations, an inaccurate description of the highly charged inhibitors and active site by the force-field and the fact that the percentage activity might not be linearly related to the binding free energy. By selecting only the relevant energy terms, this method provides useful insights into important protein–ligand interaction. An additional advantage is that intramolecular contributions to binding energies can also be accounted for. Current problems involve the relatively arbitrary breakdown of the inhibitors into fragments and the selection of relevant variables from sets which are often highly collinear.

The use of force-field derived energies in the prediction of binding characteristics has proven to be useful in several cases. For instance, the development of Merck's HIV-1 protease inhibitor Crixivan was guided by manually docking novel inhibitors to the crystal structure of the enzyme and evaluating the resulting complexes using energy minimization [29]. The observed correlations ( $r^2 = 0.580.78$ ) between binding energy and IC50s for known ligands allowed for the successful prediction of novel ligands prior to actual synthesis and testing. Similarly, the interaction energies obtained after energy minimization of inhibitors docked to purine nucleoside phosphorylase (PNP) had enough predictive power to allow for the successful selection of inhibitors with improved binding properties [30]. Force-field interaction energies also qualitatively described the observed trend in binding energies of peptide-derived transition state analog inhibitors of thrombin [31].

Generally speaking, it seems that the force-field based methods do surprisingly well, even though entropic effects and (de)solvation are usually neglected. Apparently, when the change in entropy upon receptor binding is similar for a set of molecules, the binding energy itself can have predictive power. Most of the studies mentioned above, however, report one or more outliers which notably decrease the correlation between calculated and measured binding energies and examples are known where the overall correlation appears to be quite low [27]. This casts some doubt on the general applicability, especially in cases where the ligands do not belong to a congeneric series, or when the selectivity of a compound with respect to two different proteins is an issue. Finally, most force-field based methods require in the order of minutes to minimize and evaluate ligand–receptor complexes, which can be detrimental when very large ( $> 10^4$ ) numbers of compounds need to be screened.

### 3. Methods for Solving the Poisson-Boltzmann equation

Several schemes have been proposed that utilize the linearized Poisson-Boltzmann equation for estimating binding free energies [10]. Usually the electrostatic con-

tributions are separated into a Coulombic term equal to that used in molecular mechanics force-fields (Eq. 2) and a polarization term due to interactions between polar atoms of the solvent and of the ligand and receptor. The energy due to polarization is a sum over all charges in the system  $\sum \phi(r)q_i$ , where  $q_i$  is the partial charge of an atom and  $\phi(r)$  the electrostatic potential, which is obtained by solving:

$$\nabla[\varepsilon(r) \nabla\phi(r)] + 4\pi\rho(r) = 0 \quad (5)$$

where  $\phi(r)$ ,  $\rho(r)$  and  $\varepsilon(r)$  are the electrostatic potential, the charge density and the dielectric function, respectively. This equation can be linearized and solved numerically using finite difference methods and a grid representation of the respective fields. The finite difference Poisson-Boltzmann (FDPB) method [10] is aimed at estimating the polar (electrostatic) contribution to the hydration free energy. The apolar contribution to the free energy is usually estimated based on the solvent accessible surface that is buried upon complexation, often using a single solvation parameter  $\sigma$  ( $\sim 20$ – $25$  cal/mol/Å<sup>2</sup>) for all atom types, although values of 5–40 cal/mol/Å<sup>2</sup> have been reported too.

The FDPB method has been applied successfully to a number of systems. For instance, a  $r^2$  of 0.85 (with errors  $< 1$  kcal/mol) was found for binding free energy differences calculated with the FDPB method for two receptor systems, arabinose (ABP) and sulphate-binding protein (SBP) [32]. In the case of ABP, five different ligands and for SBP three point mutations were studied and calculated free energies deviated with less than 1 kcal/mol from the experimental values. Similar results ( $x^2$  goodness of fit 0.32, average deviation from experiment within 0.5 kcal/mol) were reported by Zhang and Koshland [33], who examined 63 pairs of nine mutant proteins with seven R-malate substrate derivatives. The accuracy of their predictions appeared not to be highly sensitive to small changes in the charges used for the protein. A similar result was obtained for the ligands charges, which were derived from the CHARMM [22] force-field parameters for amino acids with similar functional groups as the substrate analogs. Nevertheless, again one outlier remained unexplained. The nonpolar contribution to the free energy was clearly required to achieve the best correlation with experiment. Qualitative agreement with experimental data can, in some cases, be achieved when the nonpolar term is entirely omitted, as was demonstrated by the results of Jedrzejewski et al. [34] on inhibitors of neuramidase. In addition, work has been reported on the introduction of molecular flexibility in FDPB calculations by means of Monte Carlo conformational searches [35]. Recent work in our department [Engels et al., preparation] complemented the FDPB method with two fitted terms for the nonpolar contribution and the loss of translational and rotational entropy for the ligand, respectively. This adaptation yielded a  $r^2$  of 0.74 and an error of 1 pK<sub>i</sub> unit for calculated and measured binding free energies of 36 serine protease–inhibitor complexes.

The FDPB method appears to be quite suitable for systems which are highly charged but has a number of potential difficulties which make routine application non-trivial; unfortunately, no definite protocol exists for the choice of dielectric constants for the protein and ligand, the treatment of dielectric boundaries in the system and the derivation of partial charges. A more detailed discussion of the sources and magnitude of errors in FDPB calculations is given by Shen and Wendoloski [32]. In addition, the

nonpolar contribution is rather crudely estimated based on the accessible surface, which makes the method less reliable for systems where a large hydrophobic component is part of the binding free energy.

## 4. Empirical Scoring Functions

### 4.1. Models derived by fitting procedures

Since scoring schemes based upon physical principles such as molecular mechanics force-fields and methods based on solving the Poisson-Boltzmann equation are limited in their predictive power, alternative scoring schemes have been devised that no longer attempt to derive the interaction energies from first principles. Such methods attempt to find appropriate mathematical functions that can be fitted to experimental binding data. Although in most cases the scoring function is still interpretable from a physical point of view, often some of the atomic detail is lost and transferred to surface properties and additional terms aimed at modelling entropic effects such as ligand flexibility, desolvation and hydrophobicity.

Probably the most widely used empirical scoring scheme is the one proposed by Böhm [36] which is aimed at rapid calculation of binding energies for use in molecular docking and *de novo* design. The binding free energy function consists of five different terms which are listed below:

$$\Delta G_{binding} = \Delta G_0 + \Delta G_{HB} + \sum_{H-bonds} f(\Delta r, \Delta \alpha) + \Delta G_{ionic} \sum_{ionic\ int.} f(\Delta r, \Delta \alpha) \quad (6)$$

$$+ \Delta G_{lipo} |A_{lipo}| + \Delta G_{rot} NROT$$

where  $f(\Delta r, \Delta \alpha)$  denotes a penalty function which accounts for deviations from ideal hydrogen-bond geometries characterized by the deviations of the ideal donor acceptor distance ( $\Delta r$ ) and angle ( $\Delta \alpha$ ).  $\Delta G_0$  denotes a constant that can be interpreted as the loss of translational and rotational entropy of the ligand upon binding.  $\Delta G_{HB}$  and  $\Delta G_{ionic}$  represent the contributions of unperturbed hydrogen bonds and ionic interactions, respectively.  $\Delta G_{lipo}$  is the contribution of lipophilic interactions assumed to be proportional to the lipophilic contact surface  $A_{lipo}$ , which is estimated using a coarse grid method. Finally,  $\Delta G_{rot}$  accounts for the loss of entropy due to bond rotations within the ligand which become fixed upon complex formation and is multiplied by  $NROT$ , the number of rotatable bonds.

This relatively simple function was fitted to 45 protein–ligand complexes taken from the protein database (PDB) for which the  $K_i$  (ranging from 25 mM to 40 fM) was known. The best fit was obtained when all five  $\Delta G$  parameters were adjustable during fitting and yielded a  $r^2$  of 0.762 with a standard deviation of 8.7 kJ/mol — i.e. an uncertainty of 1.4 orders of magnitude in  $K_i$ . Interestingly, with a minimum correlation coefficient of 0.821 among five different versions, the scoring function does not appear to be very sensitive to changes in its composition such as making  $\Delta G_{HB}$  equal to  $\Delta G_{ionic}$ , setting  $\Delta G_0$  to a fixed value or removing the hydrogen-bond perturbation penalty function. On the basis of complexes for which the binding constant could not be accurately

predicted, several possible sources of error were suggested. For instance, the scoring function is not sensitive to differences in the strengths of hydrogen bonds and does not account for more exotic interactions such as those between quaternary ammonium groups and aromatic rings [37]. Secondly, any conformational strain induced in either the protein or the ligand is not taken into account. Also interactions involving solvent molecules are not implemented in the scoring function.

The VALIDATE approach [38] presents a hybrid approach by combining force-field derived energies with terms for the octanol–water partition coefficients, steric fit, rotatable bonds, ligand strain energy and four terms describing polar/nonpolar (non)complementarity. In particular, the terms penalizing for surface noncomplementarity are of interest since this aspect of scoring is rarely accounted for. Using a training set of 51 complexes fitted by neural network and partial least-squares methods, a  $r^2$  of 0.85 was found with an error of 1.01 log units. The model was subsequently tested on three sets of protein–ligand complexes and yielded good results for crystal structures but less good results for modelled HIV-protease inhibitors and thermolysin inhibitors, in particular. Interestingly, the electrostatic energy contributed only 3% to the model, which could account for the disappointing results for the thermolysin data (where a zinc ion and phosphates are involved in binding). Also other terms like steric fit and nonpolar (non)complementarity contributed only marginally to the scoring function, independent on the method of fitting. The sensitivity of the method to starting geometries remains a serious problem.

In order to overcome possible dependence on the starting geometries of the ligand, scoring functions have been designed aimed at simplifying the energy landscape of a ligand that is to be placed in the binding pocket of its receptor. To increase the speed and reproducibility of placing and optimizing the conformation of a bound ligand, the energy function has to be smooth and contain unique minima corresponding to the native binding mode. A simple piecewise linear potential energy function was used in a genetic algorithm docking approach by workers at Agouron Pharmaceuticals [39] for flexible docking of HIV-1 protease and FKBP-12 complexes. It was shown that using a scoring function that in its simplicity resembles a square-well potential comparable to that used in macromolecular docking [40], flexible ligands could be docked to conformations close to those observed in crystal structures. The selection of meaningful parameters for the repulsive part of the energy function turned out to be critical, since small values enhance the possibilities of the search method to overcome energy barriers, while higher values reproduce steric effects much more realistically. Although this simplified scoring function clearly aids in finding correct binding orientations for complex flexible ligands, it does not allow for an accurate comparison of binding energies of different ligands or the prediction of binding energies.

An attempt at solving both these problems — i.e. improve search performance while correlating the resulting scores with binding affinities — was published recently [41]. Here a continuous differentiable scoring function was obtained by combining nonlinear Gaussian-like and sigmoidal functions to describe hydrophobic and polar interactions. Directionality and charge effects are taken into account for hydrogen-bonding interactions. Additional terms include solvation effects (estimated using the difference in the

total number of potential hydrogen bonds and the actual number of hydrogen bonds in both ligand and receptor) and entropic factors (estimated using the number of rotatable bonds and logarithm of the molecular weight of the ligand). A calibration data set of 34 complexes was used to determine the 17 adjustable parameters of the model. Best results were obtained when the ligands were energy-minimized *in vacuo* to remove conformational strain and their pose in the binding site was optimized during fitting of the parameters. The final scoring function achieved a  $r^2$  of 0.90 and a mean error of 0.72 log units. Decomposition of the scores showed that 44% and 26% can be attributed to hydrophobic and polar interactions, respectively, 25% to entropic effects and only 5% to the solvation term. The decomposition of this scoring function is very similar to that of Böhm's function [36], although the contribution of the loss of translational and rotational entropy is estimated to be higher. Estimates for this term based on experimental results vary wildly, however, and the interpretation of differences in this parameter is not straightforward. The strong dependence on basic hydrophobic and electrostatic terms agrees with the observation that force-field based scoring methods can also still account for a large body of binding data and suggests that they could be improved by accounting for the entropic effects.

#### 4.2. Solvent accessible surface-based methods

A different approach to predicting binding affinities from structural data acknowledges that all the important thermodynamical quantities involved are highly correlated with the size and composition of the surface that is buried upon complexation [42]. Thus, a correlation was established by Bohacek and McMartin [43], who noted a quantitative relationship between the complementarity of polar and nonpolar solvent accessible surfaces and inhibitor potency in thermolysin–inhibitor complexes. Extending the pioneering work of Eisenberg and McLachlan [44], Horton and Lewis [45] demonstrated that the free energy of binding can be correlated to the solvent accessible surfaces of polar and nonpolar atoms by fitting the equation:

$$\Delta G = \alpha \Delta G_{\text{apolar}} + \beta \Delta G_{\text{polar}} + \Delta G_{\text{rot/trans}} \quad (7)$$

where  $\Delta G_{\text{apolar}}$  and  $\Delta G_{\text{polar}}$  are the free energies of solvation for apolar and polar atoms, respectively, as derived from the difference in solvent accessible surface area between the bound ( $A_c$ ) and unbound species ( $A_u$ ) summed over each atom  $i$  and multiplied by an atomic solvation parameter  $\Delta\sigma_i$  [44]:

$$\Delta G = \sum \Delta\sigma_i (A_c - A_u) \quad (8)$$

The  $\Delta G_{\text{rot/trans}}$  term is a parameter which represents the free energy change due to loss of rotational and translational freedom of the ligand, which, like  $\alpha$  and  $\beta$ , is adjusted during the fitting procedure. Fitting of Eq. 7 to the experimental binding free energies of 15 rigid enzyme–inhibitor complexes yielded a  $r^2$  of 0.92. Even though the test set used is not very diverse and it is not clear how sensitive the method is to small changes in inhibitor and receptor, the results encouraged further study of the role of buried surface

area in molecular association for a large number of systems [46,47]. A more detailed analysis of protein–ligand interfaces was provided by Wallqvist et al. [48], who examined 38 protein–protein and protein–ligand complexes taken from the PDB and extracted atom–atom preference scores using the buried interfacial area. For every atom–atom combination (using 21 extended atom types), its preference of occurrence at an interface  $P_{ij}$  is calculated as the ratio of the fraction of the total interfacial area  $A_{tot}$  contributed by each atom pair  $i$ – $j$ , normalized by the product of the contributions of each atom separately:

$$P_{ij} = (A_{ij}/A_{tot}) / ((A_i/A_{tot}) * (A_j/A_{tot})) \quad (9)$$

Atom–atom pairs with a high  $P_{ij}$  value occur with high frequency at adjacent surfaces in the complexes studied and such combinations are therefore assumed to be thermodynamically favorable. Using these atomic preference parameters, an estimate for the free energy of binding  $\Delta G_{pred}$  can be obtained by fitting:

$$\Delta G_{pred} = -\sum_{i \in A} \sum_{j \in B} \alpha_{ij} A_{ij} + \beta \quad (10)$$

with

$$\alpha_{ij} = \gamma + \delta \ln P_{ij} \quad (11)$$

to the experimentally determined free energy changes by varying the parameters  $\beta$ ,  $\delta$  and  $\gamma$ . A fit with an rms deviation of 1.5 kcal/mol and a  $r^2$  of 0.55 was obtained and used to interpret the binding properties of several HIV-1 protease inhibitors. The derived scoring function has also been successfully applied in molecular docking [49]. An

Table 1 Overview of force-field based and empirical scoring schemes. The biological system on which the model was tested, the number of complexes studied ( $N_{comp}$ ), the obtained fitted regression coefficient ( $r^2$ ), the methodology used and the number of adjustable parameters are listed

System	$N_{comp}$	$r^2$	Method	#adj. param.	Ref.
<i>Force-field based scoring schemes</i>					
Thrombin	32	0.66	CHARMm	2	[21]
Trypsin	14	0.74	UFF	2	[23]
HIV-protease	11	0.88	AMBER + hydroph.	3	[26]
HSF-PLA <sub>2</sub>	26	0.92	COMBINE	2	[27]
<i>Empirical scoring schemes</i>					
Various	45	0.76	LUDI	5	[36]
Various	51	0.85	VALIDATE	12	[38]
Various	34	0.90	Gaussian/sigmoidal	17	[41]
Various	38	0.55	Pref.-based surface	3	[49]
HIV-protease	7	NR	Knowledge-based	2 ( $\Delta G_{rot/trans}$ )	[51]
Mostly proteases	24	0.92	Surfaces & $\Delta G_{rot/trans}$	3	[45]

attractive characteristic of this approach is that the contribution to the free energy of binding by interfacial atoms can be displayed on the solvent accessible surface. Even though the fit between measured and predicted free energies is not perfect, it allows regions in the ligand to be identified which do not yet have an optimal interaction with the protein and are thus suitable for modification, as well as residues in the protein which are crucial to ligand binding. A comparable method allowing the visualization of binding properties by analyzing the spatial distribution of different atom types around 3-atom fragments for 83 high resolution crystal structures was recently published [50], although no direct correlation with binding affinities was made.

Also the work of Verkhivker [51] is based on the analysis of protein-ligand interatom distances and allows for the detailed decomposition of the free energy of binding. In this case, the scoring function consists of knowledge-based ligand-protein interaction potentials, similar to those used in protein-folding simulations. This function is complemented by terms for the desolvation of different types of surfaces based on Eq. 8 and terms accounting for the loss of rotational and translational entropy. Distance-based potentials were derived from seven HIV-1 protease/inhibitor crystal structures by counting the frequencies with which certain contacts occurred and for this set of inhibitors binding free energies were calculated. A good correlation with experimental data was found and a detailed analysis of the different contributions to total binding affinity was given. The analysis is, however, limited to HIV-1 protease inhibitors, which were also used for the derivation of the ligand-protein pairwise potentials.

A similar approach was described by DeWitte and Shakhnovich [52], who derived knowledge-based potentials for a *de novo* design approach (SMoG). A total of 125 protein-ligand complexes were used for counting atom-atom contacts within 5 Å of each other, while using the average probability of all possible contacts as the reference state. Correlation coefficients obtained by comparing calculated binding constants with experimental binding data were in the range 0.77–0.81 for 3 protein-ligand systems, which is an encouraging result given the fact that no fitting was applied. Although the use of techniques from protein-folding simulations is very interesting, application to a broader range of ligand-protein complexes is required for a proper evaluation of this method. An overview of the methodology discussed in sections 2 and 4 is given in Table 1.

## 5. Conclusion

Progress has been made in the development of methodologies capable of predicting binding affinities of protein-ligand complexes over the last few years. Simple force-field minimization and energy evaluation perform reasonably well for series of similar compounds. Finite difference Poisson-Boltzmann methods are preferred for highly charged systems. The development of empirical scoring functions, capable of accounting for different entropic contributions and often with improved search profiles, is a field of intense activity and holds considerable promise for the future. Although a universal and highly accurate scoring function may still be out of reach, current approaches are often capable of yielding reasonable correlations for sets of similar ligands. Areas of

possible improvement include the correct treatment of electrostatics and bound solvent molecules. ligand and receptor flexibility and (de)solvation. A clear bottleneck for scoring function development is the rather limited number of high-resolution crystal structures for which accurate thermodynamic binding data are available.

We feel that in the practice of structure-based drug design the error in predicting the binding of a ligand to a receptor is often significantly larger than the errors typically reported in the literature. It seems that a good statistical performance of a scoring function provides no guarantee for a high predictive or extrapolative power. In order to assess realistically the predictive power of a scoring function and to monitor the progress being made with the development of scoring functions, we would advocate to extend the CASP (Critical Assessment of Techniques for Protein Structure Prediction) -trials [53,54] with test cases relevant to receptor–ligand scoring. Such a test case could comprise a set of high-resolution structures of diverse protein–ligand complexes for which preferably accurate microcalorimetric data should be available. The (in)capacity to predict correctly the binding affinity of truly novel inhibitors would provide a clear criterion to judge the qualities of different scoring methods. In addition, the availability of microcalorimetric data would allow for a more detailed comparison between the experimentally determined changes in enthalpy and entropy and those suggested by decomposition of predicted binding free energies and could be used to improve the existing scoring schemes.

## Acknowledgement

The authors wish to thank Drs. M. Engels and V.J. van Geerestein for their useful suggestions and critical reading of the manuscript.

## References

1. Janin, J., *Elusive affinities*, Proteins: Struct. Funct. Genet., 21 (1995) 30–39.
2. Ajay and Murcko. M.A., *computational methods to predict binding free energy in Ligand-receptor complexes* J. Med. Chem., 38 (1996)4953–4967.
3. Böhm, H.-J. and Klebe, G., *What can we learn from molecular recognition in protein-ligand complexes for the design of new drugs?*, Angew. Chem. Int. Ed. Engl. 35 (1996) 2588–2614.
4. Verlinde C.L.M.J. and Hol, W.G.J., *Structure-based drug design: Progress, results and challenges*, Structure 2 (1994) 577–587.
5. Böhm, H.-J., *Current computational tools for de novo ligand design* Curr. Opin. Biotech., 7 (1996) 433–436.
6. Clark, M., Cramer, R.D., III, Jones, D.M., Patterson, D.E. and Simeroth, P.E., *comparative molecular field analysis (CoMFA): 1. Effect of shape on binding of steroids to carrier proteins*, J. Am. Chem. Soc., 110 (1988) 5959–5967.
7. Clark, M., Cramer, R.D., III, Jones, D.M., Patterson, D.E. and Simeroth, P.E., *comparative molecular field analysis (CoMFA): 2. Towards its use with 3D-structural databases*, Tetrahedron Comput. Methodol., 3 (1990)47–59.
8. Grootenhuis, P.D.J. and van Helden, S.P., *Rational approaches towards protease inhibition: Predicting the binding of thrombin inhibitors*, In Wipff, G. (Ed.) Computational approaches in supramolecular chemistry, Kluwer Academic Publishers, Dordrecht (NI), 1994, 137–149.



9. Perakyla, M. and Pakkanen, T.A., *Model assembly study of the ligand binding by p-hydroxybenzoate hydroxylase: Correlation between the calculated binding energies and the experimental dissociation constants*, Proteins: struct. Funct. Genet., 21 (1995) 22–29.
10. Cramer, C.J. and Truhlar, D.G., *Continuum solvation models: Classical and quantum mechanical implementations*. In Lipkowitz, K.B. and Boyd, D.B. (Eds.) Reviews in computational chemistry, 6, VCH Publishers Inc., New York, 1995, pp. 1–72.
11. Kollman, P., *Free energy calculations: Applications to chemical and biochemical phenomena*, Chem. Rev., 93 (1993) 2395–2417.
12. Rao, B.G., Tilton, R.F. and Singh, U.C., *Free energy perturbation studies on inhibitor binding to HIV-1 proteinase*, J. Am. Chem. Soc., 114 (1992) 4447–4452.
13. Aqvist, J. and Mowbray, S.L., *Sugar recognition by a glucose/galactose receptor: Evaluation of binding energetics from molecular dynamics simulations*, J. Biol. Chem., 270 (1995) 9978–9981.
14. Liu, H.Y., Mark, A.E. and van Gunsteren, W.F., *Estimating the relative free energy of different molecular states with respect to a single reference state*, J. Phys. Chem., 100 (1996) 9485–9494.
15. Finkelstein, A.V. and Janin, J., *The price of lost freedom: Entropy of biomolecular complex formation*, Protein Eng., 3 (1989) 1–3.
16. Weiner, S.J., Kollman, P.A., Case, D.A., Singh, U.C., Ghio, C., Alagona, G., Profeta, S. and Weiner, P.K., *A new force field for molecular mechanics simulation*, J. Am. Chem. Soc., 106 (1984) 765–784.
17. Kuntz, I.D., Blancy, J.M., Oatley, S.J., Langridge, R. and Ferrin, T.E., *A geometric approach to macromolecule–ligand interactions*, J. Mol. Biol., 161 (1982) 269–288.
18. DesJarlais, R.L., Sheridan, R.P., Seibel, G.L., Dixon, J.S., Kuntz, I.D. and Venkataraghavan, R., *Using shape complementarity as an initial screen in designing ligands for a receptor binding site of known three-dimensional structure*, J. Med. Chem. 31 (1988) 722–729.
19. Meng, E.C., Shoichet, B.K. and Kuntz, I.D., *Automated docking using grid-based energy evaluation*, J. Comp. Chem., 13 (1992) 505–524.
20. Ring, C.S., Sun, E., McKerrow, J.H., Lee, G.K., Rosenthal, P.J., Kuntz, I.D. and Cohen, F.E., *Structure-based inhibitor design by using protein models for the development of antiparasitic agents*, Proc. Natl. Acad. Sci. USA, 90 (1993) 3583–3587.
21. Grootenhuis, P.D.J. and van Galen, P.J.M., *Correlation of binding affinities with non-bonded interaction energies of thrombin-inhibitor complexes*. Acta Cryst., D51 (1995) 560–566.
22. Brooks, B., Brucoleri, R., Olafson, B., States, D., Swaminathan, S. and Karplus, M., *Charmm: A program for macromolecular energy minimization and molecular dynamics calculations*, J. Comp. Chem., 4 (1983) 187–217.
23. Kurinov, I.V. and Harrison, R.W., *Prediction of new serine proteinase inhibitors*, Nature Struct. Biol., 1 (1994) 735–743.
24. Rappé, A.K., Casewit, C.J., Colwell, K.S., Goddard, W.A.I. and Skiff, W.M., *UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations*, J. Am. Chem. Soc., 114 (1992) 10024–10046.
25. Luty, B.A., Wasserman, Z.R., Stouten, P.F.W., Hodge, C.N., Zacharias, M. and McCammon, J.A., *A molecular mechanics/grid method for evaluation of ligand-receptor interactions*, J. Comp. Chem., 16 (1995) 454–464.
26. Viswanadhan, V.N., Reddy, M.R., Wlodawer, A., Varney, M.D. and Weinstein, J.N., *An approach to rapid estimation of relative binding affinities of enzyme inhibitors: Application to peptidomimetic inhibitors of the human immunodeficiency virus type 1 protease*, J. Med. Chem., 39 (1996) 705–712.
27. Ortiz, A.R., Pisabarro, M.T., Gago, F. and Wade, R.C., *Prediction of drug binding affinities by comparative binding energy analysis*, J. Med. Chem., 38 (1995) 2681–2691.
28. Mitchell, T.J., *An algorithm for the construction of 'D-optimal' experimental designs*, Technometrics, 16 (1974) 203–210.
29. Holloway, K.M., Wai, J.M., Halgren, T., Fitzgerald, P.M.D., Vacca, J.P., Dorsey, B.D., Levin, R.B., Thompson, W.J., Chen, L.J., deSolms, S.J., Gaffin, N., Ghosh, A.K., Giuliani, E.A., Graham, S.L., Guare, J.P., Hungate, R.W., Lyle, T.A., Sanders, W.M., Tucker, T.J., Wiggins, M., Wiscourt, C.M., Woltersdorf, O.W., Young, S.D., Darke, P.L. and Zugay, J.A., *A priori prediction of activity for HIV-1 protease inhibitors employing energy minimization in the active site*, J. Med. Chem., 38 (1995) 305–317.

30. Babu, Y.S., Ealick, S.E., Bugg, C.E., Erion, M.D., Guida, W.C., Montgomery, J.A. and Secrist, J.A., III, *Structure-based design of inhibitors of purine nucleoside phosphorylase*, Acta Cryst., D51 (1995) 529–535.
31. Jetten, M., Peters, C.A.M., Visser, A., Grootenhuys, P.D.J., van Nispen, J.W. and Ottenheim, H.C.J., *Peptide-derived transition state analogue inhibitors of thrombin; Synthesis, activity and selectivity*, Bioorg. Med. Chem., 3 (1995) 1099–1114.
32. Shen, J. and Wendoloski, J., *Electrostatic binding energy calculation using the finite difference solution to the linearized Poisson-Boltzmann equation: Assessment of its accuracy*, J. Comp. Chem., 17 (1996) 350–357.
33. Zhang, T. and Koshland, D.E., Jr., *Computational method for relative binding energies of enzyme-substrate complexes*, Prot. Sci. 5 (1996) 348–356.
34. Jedrzejewski, M.J., Singh, S., Brouillette, W.J., Air, G.M. and Luo, M., *A strategy for theoretical binding constant,  $K_b$ , calculations for neuramidase aromatic inhibitors designed on the basis of the active site structure of influenza virus neuramidase*, Proteins: Struct. Funct. Genet. 23 (1995) 264–277.
35. Zacharias, M., Luty, B.A., Davis, M.E. and McCammon, J.A., *Combined conformational search and finite-difference Poisson-Boltzmann approach for flexible docking: Application to an operator mutation in the lambda repressor-operator complex*, J. Mol. Biol., 238 (1994) 455–465.
36. Böhm, H.-J., *The development of a simple empiric scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure*, J. Comput.-Aided Mol. Design, 8 (1994) 243–256.
37. Dounnerty, D.A. and Stauffer, D.A., *Acetylcholine binding by a synthetic receptor: Implications for biological recognition*, Science, 250 (1990) 1558–1560.
38. Head, R.D., Smythe, M.L., Oprea, T.I., Waller, C.L., Green, S.M. and Marshall, G.R., *VALIDATE: A new method for the receptor-based prediction of binding affinities of novel ligands*, J. Am. Chem. Soc., 118 (1996) 3959–3969.
39. Verkhivker, G.M., Rejto, P.A., Gehlhaar, D.K. and Freer, S.T., *Exploring the energy landscapes of molecular recognition by a genetic algorithm: Analysis of the requirements for robust docking of HIV-1 protease and FKBP-12 complexes*, Proteins: Struct. Funct. Genet., 250 (1996) 342–353.
40. Knegtel, R.M.A., Rullman, J.A.C., Boelens, R. and Kaptein, R., *MONTY: A Monte Carlo approach to protein-DNA recognition*, J. Mol. Biol., 235 (1994) 318–324.
41. Jain, A.N., *Scoring noncovalent protein-ligand interactions: A continuous differentiable function tuned to compute binding affinities*, J. Comput.-Aided Mol. Design, 10 (1996) 427–440.
42. Novotny, J., Bruccoleri, R.E. and Saul, F.A., *On the attribution of binding energy in antigen-antibody complex MCPC 603.D1.3 and HyHEL-5*, Biochemistry, 28 (1989) 4735–4749.
43. Bohacek, R.S. and McMartin, C., *Definition and display of steric, hydrophobic and hydrogen-bonding properties of ligand binding sites in proteins using Lee and Richards' accessible surface: Validation of a high-resolution graphical tool for drug design*, J. Med. Chem., 35 (1992) 1671–1684.
44. Eisenberg, D. and McLachlan, A.D., *Solvation energy in protein folding and binding*, Nature, 319 (1986) 199–203.
45. Horton, N. and Lewis, M., *Calculation of the free energy of association for protein complexes*, Prot. Sci., 1 (1992) 169–181.
46. Krystek, S., Stouch, T. and Novotny, J., *Affinity and specificity of serine endopeptidase-protein inhibitor interactions: Empirical free energy calculations based on crystallographic studies*, J. Mol. Biol., 234 (1993) 661–679.
47. Vajda, S., Weng, Z., Rosenfeld, R. and DeLisi, C., *Effect of conformational flexibility and solvation on receptor-ligand binding free energies*, Biochemistry, 33 (1994) 13977–13988.
48. Wallqvist, A., Jernigan, R.L. and Covell, D.G., *A preference-based free energy parameterization of enzyme-inhibitor binding: Applications to HIV-1 protease inhibitor design*, Prot. Sci., 4 (1995) 1881–1903.
49. Wallqvist, A. and Covell, D.G., *Docking enzyme-inhibitor complexes using a preference-based free energy surface*, Proteins: Struct. Funct. Genet. 25 (1996) 403–419.
50. Laskowski, R.A., Thornton, J.M., Humblet, C. and Singh, J., *X-SITE: Use of empirically derived atom packing preferences to identify favorable interaction regions in the binding sites of proteins*, J. Mol. Biol. 259 (1996) 175–201.

51. Verkhivker, G., Appelt, K., Freer, S.T. and Villafranca, J.E., *Empirical free energy calculations of ligand-protein crystallographic complexes: 1. Knowledge-based ligand-protein interaction potentials applied to the prediction of human immunodeficiency virus protease 1 binding affinity*, Protein Eng. 8 (1995) 677–691
52. DeWitte, R.S. and Shakhnovich, E.I., SMOG: *De novo design method based on simple, fast, and accurate free energy estimates: 1. Methodology and supporting evidence*, J. Am. Chem. Soc., 118 (1996) 11733–11744.
53. Moult, J., *The current state of the art in protein structure prediction*, Curr. Opin. Biotech., 7 (1996) 322-127.
54. Eisenberg, D. *Into the black of night*, Nature Struct. Biol., 4 (1997)95–97.

# Molecular Mechanics Calculations on Protein-Ligand Complexes

Irene T. Weber and Robert W. Harrison

*Department of Microbiology and Immunology, Kimmel Cancer Center; Thomas Jefferson University, 233 South 10<sup>th</sup> Street, Philadelphia PA 19107, U.S.A.*

## 1. Introduction

In order to understand the structures and energetics of protein-ligand complexes, we have chosen to run relatively short, but accurate, molecular mechanics calculations on several related complexes. In contrast, the estimation of free energy differences by the free energy perturbation or thermodynamic integration methods requires extensive and time-consuming molecular dynamics simulations for each protein-ligand state (for review see [1]). Recently, we and other groups have found that simple molecular mechanics energy minimization can give intermolecular interaction energies that correlate well with trends in measured binding energies [2-5]. The aim of the molecular mechanics calculations is to reproduce the physics and chemistry of interactions in protein-ligand complexes without empirical corrections or restraints. Therefore, we have carefully evaluated the factors influencing the accuracy of the calculations. The fundamental improvement in accuracy of the new molecular mechanics and dynamics program, AMMP [6] is due to the inclusion of all long-range non-bonded energies and all hydrogen atoms. Tests of these improvements are described in references [7-9]. These improvements and other factors influencing the accuracy of the calculations are discussed. Molecular mechanics calculations with AMMP have been shown to agree with a variety of independent experimental data. The agreement verifies the accuracy of the potentials and optimization procedures. Spectral data were used to improve the potential set [2]. Minimization of protein structures was shown to result in atomic positions that are within the experimental error in the crystal structures of proteins [9]. Finally, the protein-ligand interaction energies were shown to correlate with free energy differences derived from kinetic data [2,3,7].

Molecular mechanics is a classical approximation to the inherently quantum problem of molecular chemistry. Therefore, the making and breaking of chemical bonds cannot be treated correctly in the standard formulation. However, molecular mechanics is valid for the treatment of equilibrium systems or 'states'. The choice of states will depend on the specific protein-ligand system. The choice is particularly critical for calculations on enzyme-substrate complexes where the enzyme catalyzed reaction involves a change in a chemical bond. It follows that the key problem in the application of molecular mechanics to understanding enzyme catalyzed reactions is the choice of physically valid states with intact bonds.

Two types of protein-ligand complexes have been studied. Firstly, a protein-ligand complex that is formed without large conformational changes in protein or ligand and without a change in chemical bonds. In this case, only two states need to be considered — the protein-ligand complex is compared to the free protein and ligand with the

displacement of the bound waters. Calculations were used to predict novel inhibitors of trypsin [7]. Secondly, enzyme-substrate complexes have been analyzed, where the enzyme catalyzed reaction involves a change in a chemical bond. The calculations have been successfully applied to calculation of the HIV protease-substrate interaction energies [2], and modelling the structure and energetics of glucokinase with different sugar substrates [3]. The calculated protein-ligand interaction energies were shown to correlate with the free energy differences calculated from kinetic measurements, despite the absence of bulk water in the calculation.

## 2. Accuracy in Molecular Mechanics Calculations

In order to be useful, the molecular mechanics energy must be accurately calculated. Molecular mechanics potentials typically consist of two sets of terms: one set reproduces the bond and angle geometry of the molecules, and the other set reproduces the long-range terms such as electrostatics and van der Waals forces. Ideally, the bond and angle terms are defined in a self-consistent manner, so that they have a energy minimum at an unstrained structure. Strain is introduced by the non-bonded terms; the predicted energy and structure are the result of the interplay between the two types of potentials.

Since the molecular shape is determined by the geometric arrangement of the atoms, the bond and angle terms are important for accuracy. The ability of the structures of both the protein and ligand to deform upon complex formation depends on these terms. The parameterization is especially important when protein-ligand complex formation requires changes in structure or strong interactions such as hydrogen bonds. There are many different potential sets including AMBER [10], MM3 [11], OPTIMOL [4], and UFF [12]. We started with the UFF potential set and partial charges from AMBER. The UFF potential set was chosen because its mathematical form reflects the Feynmann-Hellman theorem resulting in highly efficient parameterization, and because it is readily adaptable to a wide range of chemistry. Other choices, such as AMBER, would require that we relate a novel compound to a known amino acid or nucleotide, which is not always possible. However, we found that it was advantageous to perform minor re-parameterization. Improving the fit to molecular geometry and spectral data improved the agreement when calculating binding energy [2]. It was especially useful to use isotope effects in infra-red spectral data to check that the force terms were well parameterized. Since there are often fewer absorption and Raman bands than adjustable parameters, the ability to reproduce shifts in spectra upon deuteration confirms the potential more than does the ability to reproduce a single data set. These improvements resulted in better agreement between minimized and experimental protein structures. The improvements in AMMP also make it possible to perform stable molecular dynamics simulations on nucleic acids in the absence of special hydrogen-bonding restraints.

The ability of a ligand to fit into a binding site on a protein depends on the molecular shape and the non-bonded forces. The molecular shape is inadequately parameterized by a united atom potential where the apolar hydrogens are joined into larger united atoms. An 'all-atom' potential should be used, as shown in references [7,8]. However, the binding energy depends on the electrostatic energy, as well as the similarity in

shape. Since electrostatics are observable on the macroscopic scale, it is critically important that they be accurately calculated on the microscopic scale by including all long-range terms [6,13]. Since the number of terms in the electrostatic energy scales with the square of the number of atoms in the structure, simple implementations of this calculation are prohibitively expensive for calculations on proteins with several thousand atoms. Therefore, most calculations have introduced a distance cutoff to make the calculations faster. This cutoff is often physically inappropriate. Modern approaches increase the speed of the calculation by changing the algorithm rather than using a cutoff radius. AMMP uses a dual multipole algorithm for the efficient calculation of long-range forces [8]. This algorithm is comparable in speed to the standard treatment with a 8–10 Å cutoff radius. A local multipole expansion is used for the long-range interactions around each atom, interactions from atoms more distant than 6 Å are approximated with a quadratic expansion. The energy or force from distal atoms is calculated from the local expansion, and the energy from the local atoms is calculated explicitly. The local expansion is updated when the atoms have moved far enough. This algorithm is more efficient because the local expansion is not calculated every time the energy or force is calculated. This expansion is a solution to Laplace's equation and has a convergence radius of about 1 Å. Higher-order expansions were not found to be helpful, because the limit in convergence is due to changes in the positions of atoms at a distance rather than failure of the expansion [8]. For large structures, it is more efficient to use the Fast Multipole Method to calculate the potential and the local expansion [14,9].

The model should be optimized with the potential before calculating the interaction energy. There are many optimization strategies. Often, a high-quality structural model with experimentally determined water structure is available as a starting model for the protein-ligand complex. In this case, local optimizations like conjugate gradients or quasi-Newton methods are usually sufficient. AMMP uses conjugate gradients with the Polak-Ribiere beta and an inexact line search [8]. Convergence is monitored by the Chebyshev or  $l_\infty$  norm on the force, rather than the quadratic or  $l_2$  norm. The false appearance of convergence can occur with the quadratic norm when only a small number of atoms are in bad positions. A randomization step is used to insure that the current position is not at a stationary point, but is indeed an energy minimum [9]. The randomization is done with a short run of molecular dynamics followed by further optimization. When a good starting model is not available, other strategies should be used. These strategies include exhaustive searching with the Fourier Green's function [7,15] or 4-d embedding and closely related homotopy methods [2,9,16]. Failure to produce a model that closely approximates the correct complex will result in no agreement between the calculated and observed binding energies.

### **3. Agreement with Atomic Positions in Protein Crystal Structures**

Minimization of protein crystal structures without restrictions on atomic positions has been optimized, so that the minimized structures are essentially identical to the starting crystal structures. The minimization of the protein introduces differences, probably due

to limitations in the potentials and the search procedures. AMMP minimization of the crystal structures of proteins resulted in root mean square [RMS] differences for minimized compared to crystal structure of 0.40–0.51 Å for main chain atoms and 0.52–0.74 Å for side chain atoms [9]. These values are well within the range of 0.16–0.79 Å for RMS differences observed between main chain atoms in different crystal forms of the same protein [17,18] and less than the range 1.32–1.68 Å for the positions of side chain atoms in three different crystal forms of bovine pancreatic trypsin inhibitor [17]. These differences due to minimization with AMMP are within the range of experimental differences in protein crystal structures, and verify the accuracy of the potentials and minimization procedure.

#### 4. Prediction of Differences in Free Energy of Binding for Protein-Ligand Complexes

The differences in free energy  $\Delta G$  for formation of various protein–ligand complexes are estimated from molecular mechanics calculations and correlated with the differences in free energy derived from kinetic measurements.

$$\text{The thermodynamic free energy } \Delta G = \Delta H - T\Delta S$$

The calculated interaction energy is estimated from the total electrostatic and non-bonded energy between atoms in the ligand and the protein. As long as the structures of the protein and the ligand are not highly strained in the complex, differences in the internal energies of the protein and the ligand can be ignored. The calculated protein–ligand interaction energy is proportionate to  $\Delta H$ , and gives good correlation with  $\Delta G$  when entropy changes ( $\Delta S$ ) for formation of the different protein–ligand complexes are small (or similar for the compared ligands or proteins).

#### 5. Enzyme-Inhibitor Complexes

The treatment of enzyme–inhibitor complexes is the same as for any protein–ligand complex that does not involve chemical changes in the protein or ligand. The choice of states is relatively simple. The ‘reaction’ is the formation of the enzyme–inhibitor complex from free inhibitor and enzyme with the displacement of the bound waters from the enzyme and inhibitor. When the inhibitors are similar in size and solubility, differences in the energy of the enzyme–inhibitor complex dominate differences in the free energy of binding. The measured inhibition constant,  $K_i$ , is the ratio  $[E][I]/[EI]$ , which is an apparent equilibrium constant. Thus, the expected free energy of binding is  $-RT \ln(K_i)$ .

An example is the application of molecular mechanics to estimate the energetics of inhibitor binding to trypsin [7]. The energy of binding of 15 different small molecules to trypsin was estimated by molecular mechanics calculations using the Fourier Green’s function method [15]. The predicted binding energies agreed with measured inhibition constants (Fig. 1). The correlation coefficient between the predicted binding energy and the free energy derived from  $K_i$  values was 0.75 for 15 inhibitors. (The correlation

coefficient  $R = \frac{\sum(x-\langle x \rangle)(y-\langle y \rangle)}{(\sum(x-\langle x \rangle)^2 \sum(y-\langle y \rangle)^2)^{1/2}}$  is used throughout this chapter; literature values for  $R^2$  have been converted to  $R$  values.)

## 6. Enzyme-Substrate Complexes

The choice of states for calculations on enzyme substrates is more complicated. An enzymatic reaction proceeds through several steps. The enzyme binds substrate(s), the substrate(s) chemically react proceeding through reaction intermediates and, finally, the product(s) dissociate. Any one of these steps could be the slow step in the reaction and dominate the kinetics. The intermediates are not necessarily transition states of the reaction (energy maxima along the reaction coordinate). It is, therefore, necessary to understand the reaction mechanism, and to model the key intermediate states. However, only quasi-stable states can be modelled by the molecular mechanics approximation. Comparison of the calculated binding energy with the observed kinetics in a regression analysis will extract which reaction intermediates are related to rate-limiting steps. Predictive calculations can be made once the key intermediate states have been determined.

Differences in free energy can be related to enzyme kinetics by the equation  $\Delta G^\ddagger = -RT \ln(k_{cat}/K_m)$  from transition state theory [19]. The velocity of the reaction can be written as  $k_{cat}/K_m[E][S]$ , and the ratio of the velocities for different substrates at constant  $[E][S]$  reflects the difference in the activation energies. The use of  $-RT \ln(k_{cat}/K_m)$  for a

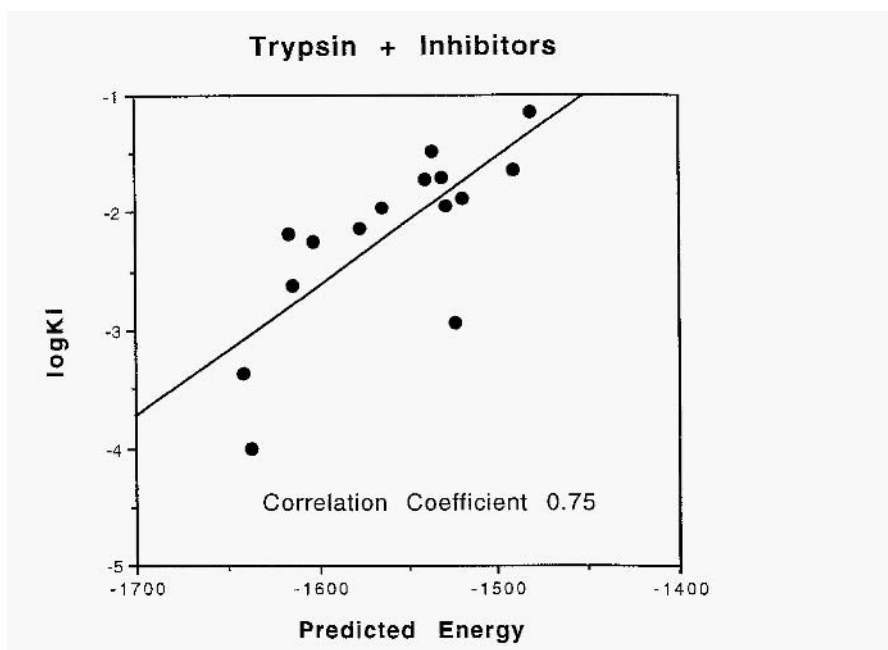
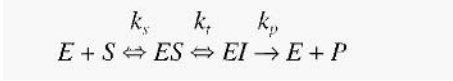


Fig. 1. Calculations for trypsin with 15 different inhibitors [7].



multi-step reaction can be justified by an extension of the Michaelis-Menten scheme, as described in Fersht [19, p. 102]. If the reaction has three steps with rate constants  $k_s$ ,  $k_t$  and  $k_p$ :



where  $E$  is enzyme,  $S$  is substrate,  $I$  is an intermediate and  $P$  is product.

The Michaelis-Menten equation is:

$$v = ((k_p k_i)/(1 + k_i)[E_0][S]) / (k_s/(1 + k_i) + [S])$$

By definition,  $K_m = (k_p k_i)/(1 + k_i)$  and  $k_{cat} = k_s/(1 + k_i)$ . Therefore,  $k_{cat}/K_m$  is  $(k_p k_i)/k_s$ , and  $\Delta G^\ddagger = -RT (\ln(k_p) + \ln(k_i) - \ln(k_s))$ . The expression is linear in the three relevant rate constants, which means that these terms can be estimated independently. When the final step is slow and rate limiting, changes in  $k_{cat}/K_m$  will reflect changes in  $k_p$ .

We have studied substrate complexes with two different enzymes: human glucokinase [3] and HIV protease [2]. Human glucokinase catalyzes the phosphorylation of sugar substrates. The glucokinase structure was modelled from the crystal structure of yeast hexokinase on the basis of 30% identity in amino acid sequence. Only the first reaction intermediate was modelled: the open conformation of glucokinase with bound sugar. The interaction energy calculated for the glucokinase-sugar complexes gave an impressive correlation coefficient of 0.99 with the kinetic data for 4 sugar substrates [3], as shown in Fig. 2. It was found to be important to include the experimentally deter-

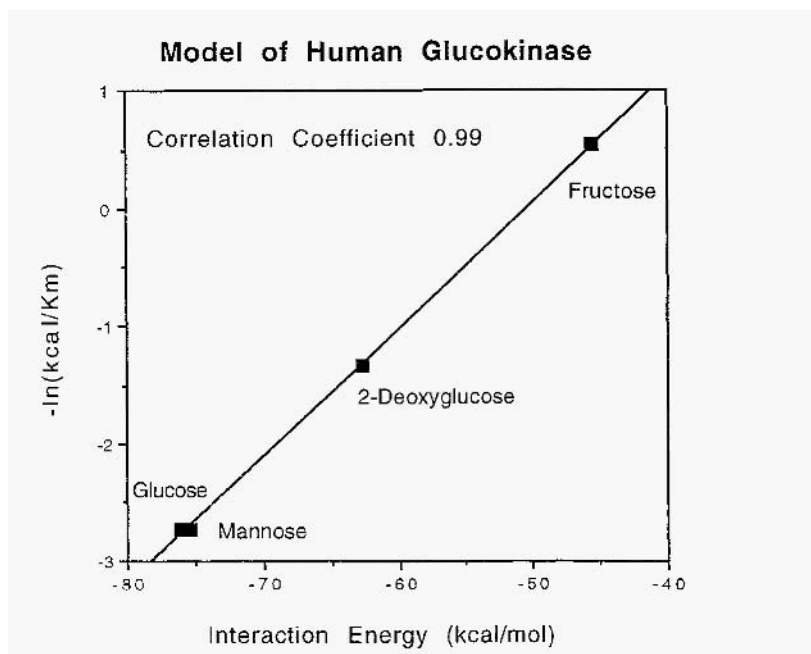
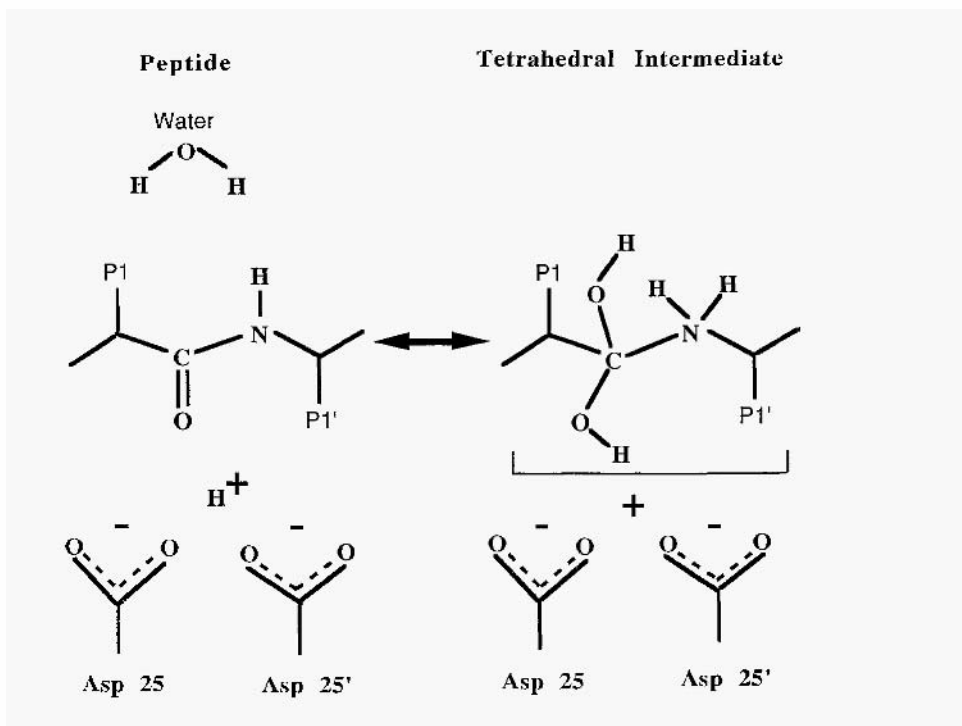


Fig. 2. Calculations for a homology model of human glucokinase showed agreement with kinetic measurements for four sugar substrates [3].

mined water structure from yeast hexokinase in this calculation because there were water-mediated hydrogen bonds to the substrate. This result illustrated the utility of the calculations even for structures that are predicted by homology with known protein structures.

The calculations on HIV protease–substrate complexes will be described in more detail to illustrate the complications involved in choosing physically valid states. HIV protease is a dimeric aspartic protease that catalyzes the hydrolysis of peptide bonds in proteins or peptides of at least 7 residues in length [20]. The peptide hydrolysis proceeds through a minimum of three steps: the binding of substrate, formation and decomposition of a tetrahedral intermediate and release of products [21,22]. The reaction intermediates were modelled from crystal structures of HIV protease with peptide-like inhibitors [2]. The protease–peptide complex and the protease complex with the tetrahedral reaction intermediate were modelled for an octapeptide (Fig. 3). These models represent two steps in the reaction. The two catalytic aspartate residues were modelled as sharing a proton, which is an average over several discrete configurations where



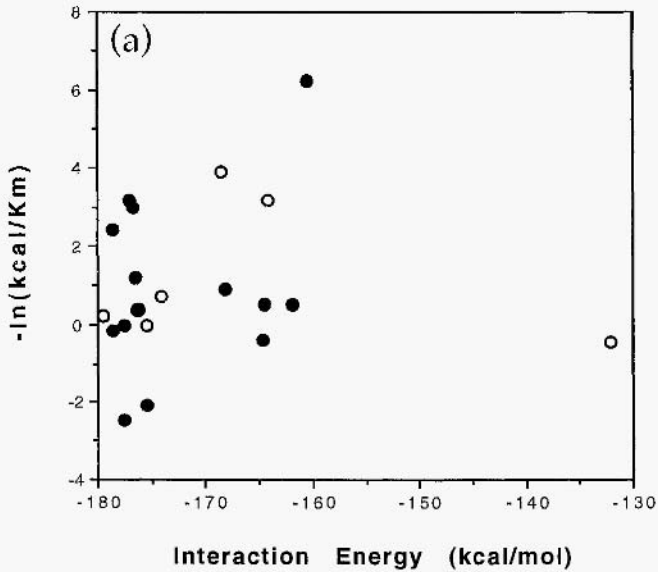
### HIV Protease Reaction Intermediates

Fig. 3. Two reaction intermediates were modeled for HIV-1 protease: the protease-peptide complex, and the protease complex with tetrahedral intermediate. The two catalytic aspartic residues were modeled as sharing a proton, which is an average over several discrete configurations where different  $\delta O$  atoms of the two aspartates are protonated. This model with a central proton was demonstrated to be stable during molecular dynamics simulations [8].

different  $\delta O$  atoms are protonated. Since it is important to model a quasi-stable equilibrium state, the stability of this proton was demonstrated by molecular dynamics simulations with peptide substrate [8]. However, similar molecular mechanics calculations by others on HIV protease inhibitors used a model in which only one of the four possible  $\delta O$  atoms was protonated [4]. Crystallographic water structure was included and there were no restraints on protein atoms, in contrast to related calculations by other groups [4,5]. The tetrahedral intermediate was modelled as a gem-diol amine where the water molecule which is seen between the flaps and the inhibitor in most crystal structures of HIV protease was incorporated into the diol. Partial charges were assigned to the gem-diol amine using the electronegativity equalization scheme in AMMP, and the total charge for the gem-diol amine was set to +1.0. The total number of atoms and total charge were conserved in this model. The calculated interaction energies were compared with kinetic measurements for peptide substrates with single amino acid substitutions at positions  $P4$  to  $P3'$ , where the scissile peptide bond is between  $P1$  and  $P1'$  (Fig. 4). The interaction energy for the HIV protease–peptide complexes had no significant correlation with the free energy derived from kinetic data. However, the interaction energy for the complexes with the tetrahedral intermediate gave significant correlation with kinetic data. The correlation coefficient was 0.93 for 8 substrates with changes in residues  $P1$  and  $P1'$  next to the scissile bond, 0.86 for 14 substrates with changes in residues  $P2$ – $P2'$  and 0.64 for all 21 substrates with changes in  $P4$ – $P3'$  positions. These correlations are significant at the 0.995–0.9995 confidence level by Student's T test, despite the absence of corrections for conformational entropy or for the effects of solvent. The higher agreement for substitutions of substrate positions  $P2$ – $P2'$  is probably because they lie within the protease and their conformations are more restricted. The lower correlation for more distal residues probably arises from their greater conformational variation and exposure to solvent on the protein surface.

The results for models of HIV protease with reaction intermediates can be compared to the results of other calculations on HIV protease–inhibitor complexes. Correlation coefficients of 0.66 to 0.71 were obtained in a recent study of three-dimensional quantitative structure–activity relationship (3D QSAR) of HIV protease inhibitors [23]. In these and other QSAR calculations, a 3D pharmacophore map is defined using molecular properties and activities for a series of ligands. The weights for each individual molecular property are empirically adjusted to maximize the correlation of predicted and observed binding constants. Head et al. [5] have developed a hybrid method using molecular mechanics minimization with the AMBER all-atom force field [10] to calculate the enthalpy of binding and heuristics to estimate the entropy of binding. Only the inhibitor atoms and the HIV protease atoms within 8 Å of the inhibitor were minimized in these calculations, which resulted in a predictive correlation coefficient of 0.755 for 13 inhibitors. Holloway et al. [4] have observed high correlation coefficients of 0.76–0.885 between the interaction energy calculated for different inhibitors of HIV protease using the OPTIMOL potential and the measured enzyme inhibition. In their calculations, the protease atoms were static and the inhibitor was minimized, while only one aspartate  $\delta O$  atom was protonated. Our molecular mechanics calculations gave similar agreement with the observed energy from kinetic parameters without applying

## HIV Protease-Peptide Substrate



## HIV Protease-Tetrahedral Intermediate

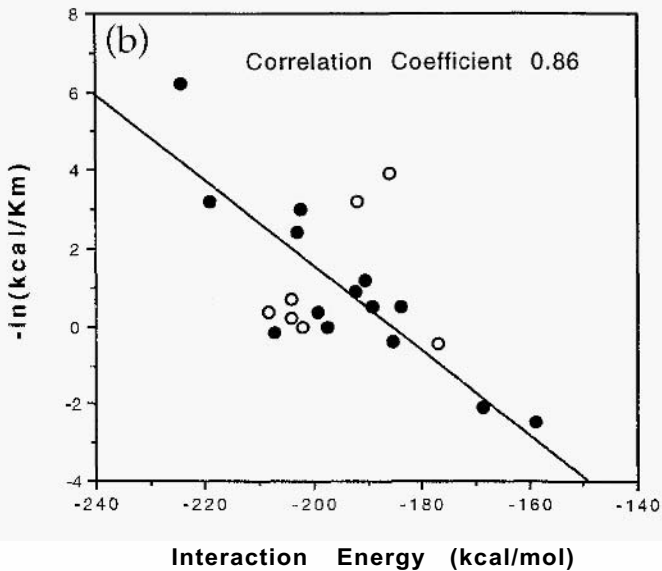


Fig. 4. Calculations for HIV protease with 21 peptide substrates [2]. (a) protease-peptide models: there was no significant correlation. The 14 peptide models with changes in substrate positions P2-P1 'closeto the scissile peptide bond' are indicated by solid circles; the seven models with changes in substrate positions P4, P3 and P3' are indicated by open circles. (b) Protease-tetrahedral intermediate models: the regression line is plotted for the 14 substrates with changes in positions P2-P2' which give a correlation coefficient of 0.86; the seven models with changes in positions P4, P3 and P3' are indicated by open circles.

any empirical corrections to the potential set or restrictions on atomic positions (see Figs. 1,2 and 4]. Therefore, significant correlation has been obtained between calculated and measured binding energies using a variety of potentials and different minimization procedures. These results show the success of applying molecular mechanics calculations on a single configuration to estimate relative binding energies for a series of protein–ligand complexes.

## 7. Statistical Mechanics Interpretation of the Molecular Mechanics Energy

The interaction energies calculated from molecular mechanics are usually much larger in magnitude than the observed free energy differences. This is the result of using a single point estimate for a thermodynamic average. Molecular mechanics and dynamics calculations are a simplistic classical approximation to an otherwise unsolvable quantum problem, and even a long ( $> 1000$  ps) molecular dynamics simulation is an insufficient statistical sample with respect to the behavior of  $10^{12}$  or more molecules over seconds to hours of time. Molecular mechanics calculations provide a point estimate of the internal energy for a specific molecular conformation. The free energy is determined by the partition function which is an ensemble average of these point estimates.

The molecular mechanics approximation of using a conformation near an energy minimum will be most accurate when one conformation dominates the partition function. Then the effect of using one point in the ensemble can be treated analytically [2]. The consequence is that molecular mechanics calculations have an apparent ideal gas constant which is different from  $R$ . It follows that the slope for the line of best correlation between  $-\ln(k_{\text{cat}}/K_m)$  and the calculated interaction energy is not equal to  $RT$ , as seen in Figs. 2 and 4. The measured differences in free energy will be lower than the differences in calculated interaction energy. Moreover, the slope gives an estimate of the average entropy of binding and is expected to vary for different protein-ligand systems.

In practical terms, the differences in the thermodynamic internal energy ( $\Delta U$ ) are calculated from the expected differences in the Hamiltonian or molecular internal energy ( $\Delta H$ ) at the energy minimum, or the average of differences over a molecular dynamics run. The calculated differences are correlated with observed differences. The correlation should be highest when the estimate for the internal energy  $\Delta U$  is most accurate. Since the Helmholtz free energy difference  $\Delta F = \Delta U - T\Delta S$ , the correlation between  $\Delta F$  and  $\Delta U$  will be best when there are only small variations in the entropy difference  $\Delta S$ . In other words, when the correlation is high, the thermodynamic internal energy terms dominate the free energy and the predictive power will be high. In contrast, when the correlation is low, the entropic terms dominate the free energy and the predictive ability will be low. Calculations on HIV protease substrates resulted in correlation coefficients that varied from 0.93 for substitutions adjacent to the scissile bond to 0.64 for distal substitutions of P4, P3 or P3' [2]. This variation is consistent with distal substrate residues occurring in several conformations in the protease complex since they are partly exposed to solvent. Therefore, the effects of conformational entropy will be

relatively large, resulting in lower correlation. The entropic contribution is reflected in the degree of correlation of the calculated interaction energies and the observed differences in free energy for protein–ligand binding or enzyme–substrate catalysis.

## **8. Future Directions**

Fundamental improvements in the computer algorithms and the integration of thermodynamic potentials with molecular potentials are required for uniformly reliable predictions. There are two basic problems with the current methodology. Firstly, while molecular mechanics potentials are good at predicting local molecular geometry, their application in a single molecular configuration results in a systematic mis-estimation of thermodynamic energies. This thermodynamic ‘sampling error’ can have insidious effects when thermodynamic observations like free energies of binding are used to parameterize molecular potentials. When the single configuration is representative of a unimodal distribution, the estimates can be rigorously proven to be overestimates of free energy differences [2]. These overestimates can be corrected by a simple scaling when they are not too large. Molecular dynamics methods like free energy integration [24] can be used to estimate the distribution of configurations and thus improve the estimation of the free energy. However, these methods are computationally expensive and ultimately suffer from the same flaw of sampling errors because the accessible phase space volume is limited.

Secondly, the solvent is often neglected or poorly treated in the calculations. Errors in the treatment of solvent lead to poor predictions for hydrophilic groups because the competition between the 55 molar water and the millimolar (or less) ligand is neglected. The conceptually simplest approach to solvent corrections is to use a discrete water model. It is important to include crystallographically determined waters in the calculations, especially since these water molecules may be critical for binding ligand or for stabilizing the protein structure. However, the crystal structure will not include bulk water. The problem for bulk water is that there are many configurations of water which enter into the estimate. Another common approach is to use a continuum model like a constant dielectric. The continuum model is physically incorrect near a protein surface, because near the surface the molecular nature of water dominates. An alternative is to use a screening potential. Screening potentials are a hybrid between a discrete model, in that they use a molecular representation, and continuum models, in that they reproduce the asymptotic behavior of the continuum solution.

The classic analytic screening potential is due to Debye (for a description of its role in thermodynamic models see [25]). The presence of mobile counterions and dipoles is treated by introducing a screening factor to the electrostatic terms. A factor of the form  $\exp(-r/r_c)$  is used where  $r$  is the radius between charged group and  $r_c$  is the correlation radius. The correlation radius is a function of the ionic strength, but it can also be treated as an adjustable parameter. Debye screening has been implemented in AMMP with a constant correlation radius. In preliminary tests, this approximation improved the correlations in energy and structure over the use of no screening [26]. However, a single correlation radius is physically unsatisfactory. Different parts of the protein will have

different exposures to solvent and therefore require different correlation radii. Therefore, multiple correlation radii are probably needed for a physically correct treatment of solvent.

## References

1. Briggs, J.M., Marrone, T.J. and McCammon, J.A., *Computational science new horizons and relevance to pharmaceutical design*, Trends Cardiovasc. Med., 6 (1 996) 198–204.
2. Weber, I.T., and Harrison, R.W., *molecular mechanics calculations on HIV-1 protease with peptide substrates correlate with experimental data*, Protein Eng., 9 (1996) 679–690.
3. Xu, L.Z., Weber, I.T., Harrison, R.W., Gidh-Jain, M. and Pilkins, S.J., *Sugar specificity of human  $\beta$ -cell glucokinase: Correlation of molecular models with kinetic measurements*. Biochemistry. 34 (1995) 6083–6092.
4. Holloway, M.K., Wai, J.M., Ialagren, T.A., Fitzgerald, P.M., Vacca, J.P., Dorsey, B.D., Levin, R.B., Thompson, W.J., Chen, L.J., deSolms, S.J., Gaffin, N., Ghosh, A.K., Giuliani, E.A., Graham, S.L., Guare, J.P., Hungate, R.W., Lyle, T.A., Sanders, W.M., Tucker, T.J., Wiggins, M., Wiscourt, C.M., Woltersdorf, O.W., Young, S.D., Darke, P.L. and Zugay, J.A. *A priori prediction of activity for HIV-1 protease inhibitors employing energy minimization in the active site*, J. Med. Chem., 38 (1 995) 305–317.
5. Head, R.D., Smythe, M.L., Oprea, T.I., Waller, C.L., Green, S.M. and Marshall, G.R., *VALIDATE: A new method for the receptor-based prediction of binding affinities of novel ligands*, J. Am. Chem. Soc., 118 (1996) 3959–3969.
6. Harrison, R.W., *Stiffness and energy conservation in the molecular dynamics: An improved integrator*, J. Comp. Chem., 14 (1993) 1112–1122.
7. Kourinov, I. and Harrison, R.W., *Prediction of novel serine proteinase inhibitors*, Nature Struc. Biol., 1 (1994) 735–743.
8. Harrison, R.W., and Weber, I.T., *Molecular dynamics simulation of HIV-1 protease with peptide substrate*, Protein Eng., 7 (1994) 1353–1363.
9. Harrison R.W., Chatterjee, D. and Weber, I.T. *Analysis of six protein structures predicted by comparative modeling techniques*, Proteins: Struct. Funct. Genet., 23 (1995) 463–471.
10. Weiner, S.J., Kollman, P.A., Nguyen, D.T. and Case, D.A., *An all atom force field for simulation of proteins and nucleic acids*, J. Comput. Chem., 7 (1986) 230–252.
11. Allinger, N.L., Yuh, Y.H. and Lii, J.-H., *Molecular mechanics: The MM3 force field for hydrocarbons*, J. Am. Chem. Soc., 111 (1989) 8551–8565.
12. Rappe, A.K., Casewit, C.J., Colwell, K.S., Goddard, W.A., III, and Skiff, W.M., *UFF: A full periodic table force field for molecular mechanics and molecular dynamics simulations* J. Am. Chem. Soc., 114 (1992) 10024–10035.
13. Schreiber, H. and Steinhauser, O., *Cutoff size does strongly influence molecular dynamics results on sloved polypeptides*, Biochemistry. 31 (1992) 5856–5860.
14. Greengard, L. and Rokhlin, V., *A fast algorithm for particle simulations*, J. Comp. Phys., 73 (1987) 325–348.
15. Harrison, R.W., Kourinov, T.V. and Andrews, L.C. *The Fourier Green's function and the rapid evaluation of molecular potentials*, Protein Eng., 7 (1994) 359–369.
16. Zeng, C.B., Aleshin, A.E., Hardie, J.B., Harrison, R.W. and Fromm, H.J., *ATP-binding site of human brain hexokinase as studied by molecular modeling and site-directed mutagenesis*, Biochemist, 35 (1996) 13157–13164.
17. Wlodawer, A., Nachman, J., Gilliland, G.L., Gallagher, W. and Woodward, C., *Structure of form III crystals of bovine pancreatic trypsin inhibitor*, J. Mol. Biol., 198 (1987) 469–480.
18. Zegera, I., Maes, D., Dao-Thi, M.-H., Poortmans, F., Palmer, R. and Wyns, L. *The structures of R Nase A complexed with 3'CMP and d(CpA): Active site conformation and conserved water molecules*. Protein Science. 3 (1994) 2322–2339.
19. Fersht, A., *Enzyme structure and mecahnism* W.H. Freeman and Company. New York, 1985.

20. Darke, P.L., Nutt, R.F., Brady, S.F., Garsky, V.M., Ciccarone, T.M., Leu, C.T., Lumma, P.K., Freidinger, R.M., Vebal, D.F. and Sigal, I.S., *HIV-1 protease specificity of peptide cleavage is sufficient for processing of gag and polpolyproteins*, *Biochem. Biophys. Res. Commun.*, 156 (1988) 297–303.
21. Polgar, L., Szeltner, Z. and Boros, I., *Substrate-dependent mechanisms in the catalysis of human immunodeficiency virus protease*, *Biochemistry*, 33 (1994) 9351–9357.
22. Hyland, L.J., Tomaszek, T.A. and Meek, T.D., *Human immunodeficiency virus-1 protease: 2. Use of pH rate studies and solvent kinetic isotope effects to elucidate details of chemical mechanisms*, *Biochemistry*, 30 (1991) 8454–8463.
23. Oprea, T.I., Waller, C.L. and Marshall, G.R., *Three-dimensional quantitative structure-activity relationship of human immunodeficiency virus (I) protease inhibitors: 2. Predictive power using limited exploration of alternate binding modes*, *J. Med. Chem.*, 37 (1994) 2206–2215.
24. Beveridge, D.L. and DiCapua, F.M., *Free energy via molecular simulation: applications to chemical and biomolecular systems*, *Ann. Rev. Biophys. Biophys. Chem.*, 18 (1989) 431–492.
25. Goldenfeld, N., *Lectures on phase transitions and the renormalization group*, Vol. 85, *Frontiers in Physics*, Addison-Wesley, Reading, MA, 1992.
26. Reed, C., Fu, Z.-Q., Wu, J., Xue, Y.-N., Harrison, R.W., Chen, M.-J. and Weber, I.T. *Crystal structure of TNF- $\alpha$  mutant R31D with greater affinity for receptor R1 compared to R2*, *Protein Eng.* 10 (1997) 101–107.



**This Page Intentionally Left Blank**

**Part II**

**Quantum Chemical Models  
and Molecular Dynamics  
Simulations**

**This Page Intentionally Left Blank**

# Some Biological Applications of Semiempirical MO Theory

**Bernd Beck and Timothy Clark**

*Computer Chemie Centrum des Instituts für Organische Chemie I der Friedrich-Alexander-Universität Erlangen/Nürnberg Niigelsbachstrasse 25, D-91056 Erlangen, Germany*

## 1. Introduction

Semiempirical molecular orbital (MO) methods were originally developed as tools for bench organic chemists to investigate typical physical organic problems on molecules of about 50 atoms or less. There has, however, been such a rapid development in the efficiency and accuracy of semiempirical methods in recent years that semiempirical MO calculations are now beginning to intrude into what used to be the realm of pure field techniques. The present review is intended to cover the new techniques that are emerging for very large molecules and to describe some significant recent applications.

What are the advantages of approximate MO techniques over force fields, which usually reproduce experimental geometries and energies considerably more accurately? The ability of quantum mechanical techniques to describe bond making and breaking processes may be critical for mechanistic studies, but generally the detailed and accurate anisotropic electron density distribution around atoms provide the major advantage of using MO techniques. The performance of the semiempirical methods is naturally critical, MNDO, for instance, was unable to describe hydrogen bonds, making it completely unsuitable for biological applications, AM1 and PM3 are better in this respect, but still have critical problems describing the geometry and the rotation barrier about peptide C–N bonds.

Another major advantage of semiempirical MO techniques is their speed and versatility. They have been used as platforms for solvent simulations, quantum mechanical/molecular mechanical (QM/MM) mixed methods and a variety of other applications. There is still a tendency to neglect the power of such methods for biological applications, and so we present an overview of some of the new literature and novel applications of these techniques.

This chapter is organized as follows: in section 2, we deal with the pure semiempirical methods, in section 3, we describe the QM/MM methods, the differences between the most commonly used methods and their abilities. In section 4, we give a short introduction into quantum-based electrostatics, and in sections 5 describe some possible applications of the QM/MM approach.

## 2. Pure Semiempirical Methods

As mentioned above, the major task of all methods presented here is to reduce the number of necessary operations during the SCF calculation in order to allow calculations on large molecular systems. In the following section, we describe several approaches developed during the last few years that allow the calculation of large molecular systems within the semiempirical approximation.

## 2.1. Strictly localized MOs

One of the first methods was published by Rivail and co-workers in 1992 [1]. As a first approximation, the electronic wave function is built up from strictly localized molecular orbitals (SLMO) [2]. These are contributions of atomic hybrid orbitals (HYO). The most common SLMO involves two atomic centers and describes the bond between these atoms. One-center SLMOs are used to describe lone pairs and multi-centre SLMOs for extended  $\pi$ -systems. In order to improve the quality of the description for the region of interest, subsystems are defined. The subsystems are then optimized in the field of the environment, which is represented by the SLMOs.

This simple approach has some major disadvantages. The computational effort for setting up the SLMOs of the environment is still very high. The SLMOs are unable to deal with charge transfer and, therefore, the quality of the results depends crucially on the definition and the size of the subsystems.

## 2.2. MOZYME

An improved method, implemented in the MOZYME program, was developed by Stewart [3] in 1994 and is also based on localized molecular orbitals (LMOs). The LMO theory is an alternative way to generate MOs that correspond to the electronic structures in Lewis molecular structures. Therefore, calculations involving LMOs can be limited to that region of space in which the LMO exists. To obtain a self-consistent field (SCF), however, the LMOs are allowed to expand. Even so, for large systems such as proteins, the size of an LMO will be small compared with the size of the entire system.

For small systems, every LMO involves all the atoms of the system. In large molecules, however, calculations involving LMOs are much more efficient. All occupied–virtual interactions involving LMOs separated by large distances will automatically be zero. Obviously the larger the system, the more important this becomes.

The calculation of the density matrix can be limited to those matrix elements that are represented by an LMO. Therefore, the computational effort for this step only depends on the number of LMOs — i.e. it increases linearly with the size of the system. For the calculation of the energies of LMOs and occupied–virtual interactions, the computational effort is independent of the size of the entire system, since the size of an LMO depends only on the local electronic structure.

On the other hand, the long-range electrostatic effect is a non-local phenomenon for which the computational effort rises with  $N^2$ . However, the calculation of this effect is by far the simplest and, therefore, the computational costs for the SCF calculation using LMOs rises nearly linearly with the size of the system ( $N^{-1}$ ) in contrast to  $N^3$  for the conventional MO–SCF procedure. The starting set of LMOs must satisfy several requirements: they should form an orthonormal set. There must be one LMO for every occupied and every unoccupied MO in the system. Ideally, they should involve one or at the most two atoms.

Conventional Lewis diagrams of molecules provide a good starting point for constructing the initial set of LMOs. For each atom with a basis set of one s and three p

atomic orbitals, a set of four hybrid orbitals is constructed (one for each atom to which the atom is bonded or for its lone pairs and p-orbitals for  $\pi$ -bonds).

Stewart also uses a distance cutoff for interactions, in order to reduce memory requirements and because many of the integrals are not necessary within the LMO approach. Therefore, only those one-electron integrals were calculated which represent interactions at less than a given distance (normally 6–8 Å). At distances greater than about 7 Å the calculation of the two-center two-electron integrals is modified. Out of the 100 integrals needed for small distances, only seven remain, mainly because the quadrupolar and higher multipolar terms are ignored. At distances larger than 30 Å, simple Coulomb repulsion is used.

The method was tested for various polypeptides, up to 264 residues, and it could be shown that for larger systems, single SCF calculations using the LMO approach are up to 160 times faster than conventional SCF for the largest system. For systems with less than 50 atoms, the conventional methods are still more efficient.

The range of the molecules that can be studied using the LMO method is limited to those systems that can be represented as non-radical Lewis structures. These structures are essential in order to be able to build up the initial LMOs. The test calculations have shown that the time dependency is still nearer  $N^{1.5}$  than  $N^1$ . Optimization of large systems is still impractical because of the heavy memory demand. Even so, MOZYME is a great step forward in the quantum chemical treatment of large molecular systems. However, more tests and some additional modifications seem to be necessary for the LMO method to become a useful tool.

### *2.3. Divide and conquer*

An alternative and also very promising approach is the so-called density matrix divide and conquer (D&C) method developed by W. Yang and co-workers [4,5] for density functional theory and recently implemented into MOPAC [6]. Dixon and Merz [7] have recently published another linear scaling semiempirical method based on the same principles.

The basic concept of the D&C approach is to divide a large molecular system into a set of relatively small subsystems. An approximate total electron density matrix is then built up from the contributions of the subsystems. The linear scaling of the method is a result of the fact that matrix diagonalization is not required on the global Fock matrix, but rather for a set of smaller subsystems. Semiempirical methods benefit from the D&C approach unless the system becomes very large, then the quadratic expense of calculation of the two-center integrals will begin to dominate. True linear scaling for large molecular systems can be achieved by the use of finite cutoffs for diatomic interactions or perhaps by employing the continuous fast multipole method [8–11].

As shown in Fig. 1, so-called buffer regions were defined at each end of the subsystems in order to reduce truncation effects. There is also an overlap between adjacent subsystems to facilitate the propagation of electronic effects throughout the molecule. The Roothaan-Hall equations for the subsystems were solved including the atoms in the buffer region, and the density matrix then constructed for a given subsystem excluding the buffer atoms.

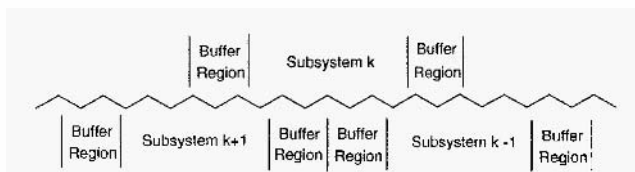


Fig. 1. Divide and conquer subsetting scheme

The D&C-SCF method can be summarized as follows. Starting from an initial guess of the global density matrix, the subsystem Fock matrices are assembled. The Roothaan-Hall equations for each subsystem are solved and their density matrices constructed. Finally, all subsystem density matrices are combined in an appropriate fashion to arrive at a new global density matrix. The results obtained from test calculations on molecules up to about 10 000 atoms [6] clearly show the linear scaling of the D&C method for the energy calculation. For the gradient calculation, a quadratic dependency ( $N^2$ ) was observed. Starting at about 250 atoms, the D&C approach is faster than the normal SCF scheme. Yang and co-workers also include solvation into their method by using the COSMO model (conductor-like dielectric continuum) [12–16]. On the other hand, it is still not possible to optimize a macromolecule fully in a reasonable time. There are problems using the cutoff [7] and the results are strongly dependent on the size of the subsystems used.

The LMO and D&C approaches are both steps forward that allow the calculations of large molecular systems like enzymes or proteins up to about 10 000 atoms within the semiempirical approximation. Despite the fact that it is still impracticable to perform full geometry optimization on these systems, it is possible to calculate Mulliken charges, molecular dipole moments and some other molecular properties. More research in this area is necessary to solve some of the problems mentioned above and to allow these methods to become more applicable in common use.

In contrast to the pure semiempirical methods, developed in the last few years, QMiMM approaches have been developed over the last two decades by combining different levels of QM methods with force fields. A survey of semiempirical QM/MM will be given in the next section.

### 3. QM/MM Methods

As mentioned before, QM/MM methods have been developed continuously over the last two decades. In 1976, Warshel and Levitt [17] presented the first QM/MM approach. Since then, several hybrid QMiMM models have been developed, combining semiempirical [18–24] density Functional [25], valence bond [26,27] or *ab initio* Hartree-Fock [28] methodology with frequently used force fields like MM3 [29] AMBER [30] or CHARMM [31]. In the meantime, these methods have become well established; for example, for the investigation of solvation phenomena [32–46] or for biochemical problems [47–51].

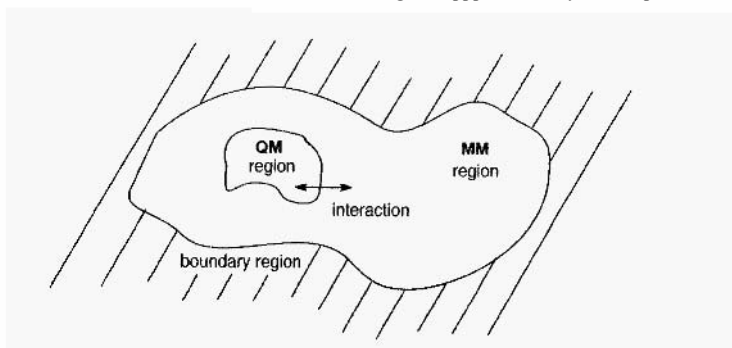


Fig. 2. Partitioning of a molecular system into QM and MM part.

The basic idea of all QM/MM methods is to treat that part of the molecule which undergoes the most important electronic changes quantum-mechanically, whereas the rest of the molecule is treated by molecular mechanics.

The methods described in the following differ in the way they include the interaction between the QM and MM parts, by the usage of link atoms or the use of an additional 'boundary region' around the molecular system to count for environmental effects. Therefore, the following section is divided into two parts. The first deals with those QM/MM methods that use so-called 'link atoms', and the second describes those which do not use them.

### 3.1. Methods with link atoms

Many research groups have published hybrid quantum mechanical/molecular mechanical approaches in which they link atoms to connect the QM with the MM parts of the molecule. See, for example, Singh and Kollman [28] or Merz and co-workers [25]. An approach very often used is that of Field, Bash and Karplus, published in 1990 [18], which will be described in more detail. The authors' aim was to implement a generally applicable method, with which a large molecular system can be studied with geometry optimization, molecular dynamics and Monte Carlo methods. Therefore, they combined the MM program CHARMM [31] with the AM1 and MNDO semiempirical molecular orbital procedures. The molecular system to be studied is divided into two parts, as shown in Fig. 2. As mentioned above, an additional boundary region was included to account for the surroundings that are neglected. The Hamiltonian for the entire system is given by

$$\hat{H}_{sys} = \hat{H}_{QM} + \hat{H}_{MM} + \hat{H}_{QM/MM} + \hat{H}_{boundary}$$

and the energy by

$$E_{sys} = E_{QM} + E_{MM} + E_{QM/MM} + E_{boundary}$$

The QM and MM terms are treated as usual. A more detailed description is given for the QM/MM and the boundary terms.



The QM/MM Hamiltonian  $\hat{H}_{QM/MM}$  consists of the interactions of the QM and MM parts represented by atomic charges and van der Waals parameters. The interaction is given by the electrostatic and van der Waals interactions and the polarization of the QM part by the atomic charges of the environment.

For the boundary term, two methods are possible: the periodic boundary [52] and the stochastic boundary approaches [53,54]. The only complication within the periodic boundary approach is that the images of the central box contain a copy of the QM atoms and that their charge distribution changes during the calculation. This problem can be avoided by choosing the periodic cells such that the QM images are far enough apart to be unimportant for the interaction energy. Another important aspect are the so-called 'link atoms', mentioned earlier. This type of atom is necessary if, for example, in an enzyme reaction some residues have to be treated quantum-mechanically, while the rest of the protein does not. In this case, there are bonds between the QM and MM parts. The 'link atoms' are used to terminate the QM electron density along these bonds. Different schemes have been proposed [55,56]. The authors treated them exactly as QM hydrogens and they are invisible to the MM atoms, because no interactions are calculated. The procedure employed in the version described is:

1. Partition a molecule into QM and MM parts.
2. Define the 'link atoms'.
3. Delete connectivities among the QM atoms. All MM internal coordinate energy terms that involve QM atoms are deleted.
4. Create the non-bond list. The MM non-bond list is generated as usual; there are two QM/MM lists: one for the vdW, and one for the electrostatic interactions.
5. Calculate the MM energies and forces.
6. Calculate the QM/MM vdW interactions using the vdW list.
7. Calculate the QM and QM/MM electrostatic interactions and forces.

The test calculations performed clearly show that the partitioning can have significant effects on the results obtained. For the definition of the 'link atoms', one should not break  $\pi$ -bonds or bonds in which conjugation effects will be important. It is also important to note that the inclusion of the 'link atoms energy' means that it is only possible to compare the energies between systems with the same number and types of 'link atoms'. The authors recommend that it is necessary that a number of different partitioning approaches should be tested for each system in order to determine the effects on the results. This method also neglects the polarization of the MM part by the QM part.

Recently, Morokuma and Mascras published a new integrated QM/MM optimization scheme for equilibrium structures and transition states [57]. In order to describe the so-called IMOMM method, we will use an example from the original paper. The 'real' system is metal complex  $M(P(CH_3)_3)_2$ , and the 'model' system in the QM calculation is  $M(PH_3)_3$ . Within IMOMM the atoms of the molecule are divided into four different sets, as shown in Fig.3. In set 1, the atoms which appear in both the 'model' QM part and the 'real' MM part are present. The second subset includes those atoms which are only present in the QM calculation, but substituted in the MM calculations by real (different) atoms. In other methods, this set is referred to as 'junction dummy atoms' [28]

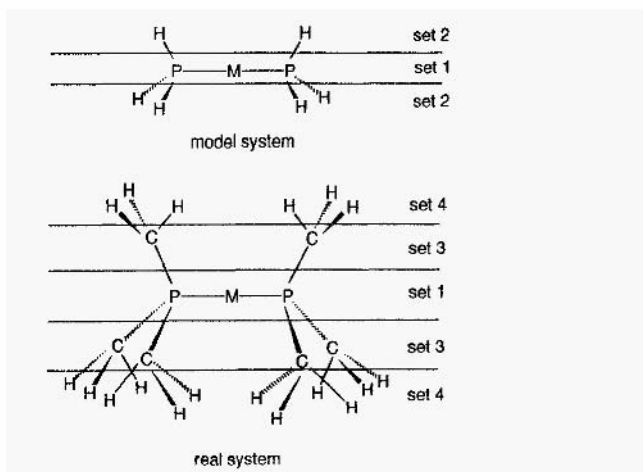


Fig. 3. Partitioning of the 'model' and the 'real' system within IMOMM.

or the 'link atoms' [18]. Set 3 atoms are only present in the MM calculation, but each of them corresponds to an atom in set 2 (generally hydrogens) in the QM calculation. Finally, set 4 contains those atoms which are only present in the MM calculation.

For each pair of atoms in sets 2 and 3, the same definition of internal coordinates was chosen. During the optimization, the coordinates of the atoms in set 3 are a function of those in sets 1 and 2. This scheme differs sharply from the approach described above, where the coordinates in set 2 ('link atoms') optimize independently from those in the MM part.

The equations for energies and gradients for the combined QM/MM description are a function of those of the independent QM and MM descriptions. In order to avoid double counting of contributions already treated in the QM part, these terms were selectively deleted in the MM calculation.

The computational algorithm starts with the input of the 4 sets of atoms. The QM calculations are then done, followed by the MM calculations, or vice versa. In the next step, the energy and gradients are calculated, followed by a check for convergence. If convergence has not been achieved, new coordinate values for sets 1 and 2 atoms are generated and the calculation starts again.

This method has been designed to perform a QM geometry optimization within an MM environment. Therefore, it is possible to use *ab initio*, DFT or semiempirical MO methods with several different force fields. Until now, results have only been published for test calculations on small systems. The major limitation of this method is the fact that the QM/MM interactions are not treated explicitly.

The last method using link atoms that we will describe is that of Thiel and Bakowies [23], published in 1996. The authors present a hierarchy of QM/MM approaches. The so-called model A only represents a mechanical embedding of the quantum mechanical region. Model B includes several interactions between the different parts and also the polarization of the QM part by the MM environment. The most

refined model C also includes the polarization of the MM part by the QM part. In contrast to earlier methods, models B and C include an explicit MM correction for interactions involving link atoms. The formal disadvantage of this is the fact that this method cannot be used in the MD simulations.

Model B includes almost the same treatment for the QM/MM interaction, as described by Field, Bash and Karplus [18]. As a refinement of earlier QM/MM methods, the authors have adopted parameterized models to calculate QM electrostatic potentials and MM partial charges. These have been calibrated against RHF/6-31G\* reference data [58]. In the majority of cases, this model successfully reproduces experimental data. Problems may arise if strongly charged MM atoms are close to the QM/MM boundary. In this case, the electrostatic interaction tends to be overestimated.

Model C includes the polarization of the MM fragment in order to remove the asymmetry for the description of non-bonded QM/MM interactions ('normal' QM/MM only includes polarization of the QM part). It adopts the treatment of Thole [50] to describe induced dipole interactions. This model only needs one parameter per element, the isotropic atomic polarizability. This method currently provides one of the most advanced treatments of polarization within semiempirical QM/MM approaches. The consideration of MM polarization seems to be crucial in applications involving charged QM parts, which generate large electrostatic fields.

The MNDO/MM computer program uses either the MNDO or the AM1 wave functions and the MM3 force field. The method was tested for various smaller organic molecules up to about 40 atoms in order to be able to compare the results with pure QM methods. For most test calculations, models B and C provide sufficiently reliable results.

### 3.2. The SLMO technique

One method which allows bonds to be shared by the QM and MM part and does not use 'link atoms' is the local self-consistent field method of Rivail and co-workers [20]. This method is a further development of the strictly localized molecular orbital (SLMO) approach [1], already described in section 2 on pure semiempirical methods. However, the environment is now treated by mechanics instead of SLMOs.

For an atom pair at the boundary between the two parts, the atom belonging to the QM part is called the frontier atom (X). The atomic orbitals of X are transformed such that hybrid orbitals are obtained which are colinear with the bonds of this atom. The hybrid orbitals belonging to bonds between the QM and MM part are excluded from the orbital basis (Fig. 4). The associated electron densities are treated as external point charges on the cationic QM fragment. Again, this approach includes electrostatic and vdW interactions, as well as the perturbation of the Fock matrix (polarization of the QM part). MM polarization is not implemented.

This method suffers from the fact that the electron density of the excluded orbital is not known unless a QM calculation has been performed for the entire system. Additional problems may occur in the calculation of the electrostatic QM/MM interactions, because the total charge of a neutral QM subsystem is not forced to be zero.

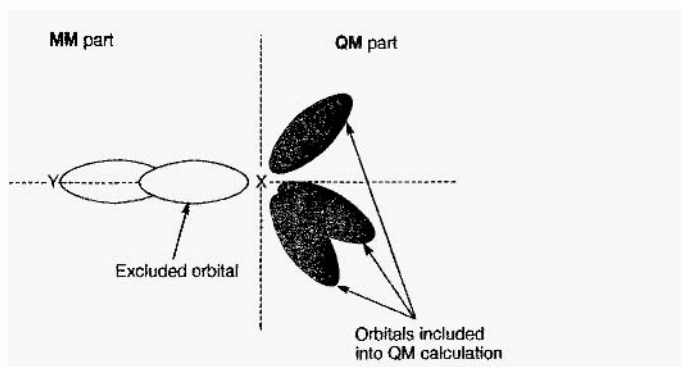


Fig. 4. Description of system partitioning in this hybrid method.

Recently, Rivail and co-workers have published a hybrid classical quantum force field based on the method described above for modelling very large molecules [60].

### 3.3. Intermolecular approaches

The following two methods assume that there are no bonds between the QM and the MM atoms. Therefore, they do not need 'link atoms' or related approaches. The first method discussed was the so-called PCM (Point Charge Model) method developed in our research group, which was originally designed to calculate heats of adsorption and reaction pathways in zeolites [24]. In PCM, the environment is fixed and the atoms are treated as point charges with assigned van der Waals radii. The interactions between the environment and the QM part were considered to consist mainly of three contributions. The first of these is the Coulomb interaction between the point charges and the QM atoms including the electronic and core contributions. Secondly, the van der Waals interaction was calculated using a Lennard-Jones [6,12] potential, whereas the potential parameters were taken from the Universal Force Field (UFF) developed by Rappé and co-workers [61]. Finally, the perturbation of the Fock matrix (polarization of the molecule within the environment) was also considered. In order to show how important it is to include the polarization into the QM/MM interactions, Fig. 5 (see p. 145) shows an example of an inhibitor optimized within the binding site of an enzyme. The result of a gas phase calculation was used as reference.

The induced dipole moment is 5.48 Debye. Test calculations have shown that the VAMP-PCM approach can be used successfully for biochemical problems. (Examples will be discussed later.) There are two problems to be solved: the flexibility of the environment, and the fact that the PCM approach neglects the MM polarization, in contrast to the QM/MMpol method published by Thompson and co-workers in 1995 [22].

The QM/MMpol approach couples the QM region (for both ground and excited states) with a polarizable MM description, according to the approach described in the earlier work of Luzhkov and Warshel [62,03]. The first step was to define a system consisting of the QM and MM components. The QM part includes the electrons ( $\psi$ ) and the

nuclei ( $Z_a$ ) or effective cores, whereas the MM compounds consist of charged atomic centres (qm) with atom centered polarizable point dipoles ( $\mu_m$ ) [64–66]. Figure 6 shows the interactions between each of the compounds.

Interaction in the QM part are included into the  $\hat{H}_{QM}$  operator. The operators  $\hat{H}_{MM}^{el,stat}$  and  $\hat{H}_{MM}^{el,pol}$  consists of interactions 4 and interaction 5 and 6, respectively.  $\hat{H}_{QMMM}^{el,stat}$  (interactions 7 and 8) and  $\hat{H}_{QMMM}^{el,pol}$  (interactions 9 and 10) are the important terms. In addition to these operators, there are also those describing the van der Waals interaction ( $\hat{H}_{QMMM}^{vdW}$ ,  $\hat{H}_{MM}^{vdW}$ ) and  $\hat{H}_{MM}^{bonded}$  containing bond terms for the MM part. The Hamiltonian for the entire system is obtained by

$$\hat{H}_{sys} = \hat{H}_{QM} + \hat{H}_{QMMM}^{el,stat} + \hat{H}_{MM}^{el,pol} + \hat{H}_{QMMM}^{vdW} + \hat{H}_{MM}^{el,stat} + \hat{H}_{MM}^{el,pol} + \hat{H}_{MM}^{bonded} + \hat{H}_{MM}^{vdW}$$

During the QM system SCF, the MM region is included as a static external potential consisting of fixed point charges and fixed atom-centered polarizable dipoles. The MM part SCF is done by a fixed QM state. A single iteration of the entire system consists of a full iteration of both the QM–SCF and MM–SCF.

The method was implemented into the semiempirical INDO/S approximation but it is also applicable to other semiempirical and ab initio Hamiltonians. As a test case, QM/MMpol was applied to the analysis of the ground and excited states of the bacteriochlorophyll S dimer (P) of the photosynthetic reaction center (RC) of *Rhodospseudomonas viridis*. During the calculations, the system consisted of 325 QM atoms embedded in 20 158 polarizable MM atoms. The results obtained are in reasonable agreement with experiment. The explicit values could be found in the original publication [22]. Recently, the method has been extended in order to use it for MD simulations [67].

Basic components needed for the evaluation of electrostatic interaction energies in combined quantum-mechanical and molecular-mechanical approaches are the electrostatic potential and the partial charges. In section 4, we will give a short summary of published approaches which deal with the calculation of these properties.

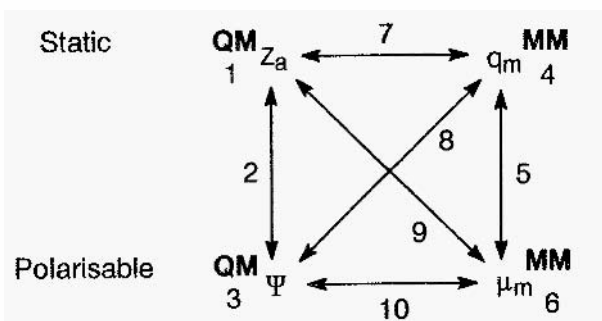


Fig. 6. Schematic view on the interactions in the QM/MMpol method.

## 4. Quantum-based Electrostatics

There are several possibilities to describe the charge distribution within a molecule. Atom-centered charges are commonly used (the one-center approach) for this purpose. Only a few methods using multi-center approaches have been published.

### 4.1. One-center approaches

The proper definition of atomic charges is very difficult. The main problem arises from the fact that partial charges are not observable quantities. Thus, any possible definition is arbitrary. The simplest method to derive atomic charges is to use the electronegativity of the atoms [68,69]. This technique is often used for MM methods. However, conjugation and polarizability effects suggest that quantum-mechanical charges are more reliable in more complex systems. The simplest approximations are the so-called Coulson [70] and Mulliken charges [71]. They are derived directly from the density matrix. A more useful approach to describing the electronic distribution in molecules may be to choose an atom-centered charge distribution that reproduces molecular properties at and outside of a surface (usually the van der Waals surface) appropriate to the range of distances at which molecules interact [72]. One such useful property is the electrostatic potential, which describes the forces exerted on a positive probe of unitary charge at the desired surface. In the following, we will describe some methods for the calculation of partial charges within the semiempirical approximation.

In 1995, Truhlar and co-workers published two new charge models, called AM1-CM1A and PM3-CM1P, based on experimental dipole moments [73]. The partial charge methods are parameterized such that the dipole moments calculated from them are as accurate as possible in comparison to the experimental ones. These charge models yield rms errors of 0.30 D for AM1 and 0.26 D for PM3 in the dipole moments of a set of 195 neutral molecules. The test set covers a variation of organic functional groups such as halides, C-S-O and C-N-O linkages. Another remarkable point is that the atomic charges computed with the CM1 model are in good agreement with high-level *ab initio* calculations for both neutral compound and ions, whereas CM1 is far less expensive with respect to the computational effort.

Bakowies and Thiel developed another efficient semiempirical approach for the computation of partial atomic charges [58]. The charges are derived from a semiempirical charge equilibration model, which is based on the principle of electronegativity equalization [68,69]. The approach is a further development of the QEq model of Rappé and Goddard [74]. During the parameterization, Bakowies and Thiel used *ab initio* potential derived atomic charges (RHF/6-31G\*) as references. Final parameters have been published for hydrogen, carbon, nitrogen and oxygen atoms. The results were compared to *ab initio* Mulliken and PD atomic charges and reproduced with an accuracy of approx. 0.005e and 0.1e, respectively. The authors claim that the larger deviation for PD charges can be traced back to statistically ill-defined PD charges of buried atoms that constitute the molecular skeleton.

The so-called potential derived (PD) or electrostatic potential (ESP) derived atomic charges used by Bakowies and Thiel are very important in QSAR or QM/MM methods. There are various methods of deriving potential-based charges using *ab initio* and semiempirical techniques. Most differ in the way they generate the grid points for fitting procedure. Chirlian and Francl (CHELP) [75] used spherical shells, and Brenenam and Wiberg (CHELPG) [76] defined a cube of points spaced 0.3-0.8 Å apart containing the molecule and including 2.8 Å of headspace on all sides. Besler, Merz and Kollman (MK) [77] used the Connolly molecular surface algorithm to define points on several concentric shells (1.4, 1.6, 1.8 and 2.0 times the vdW radius) surrounding the molecule. Spackman uses a geodesic point selection scheme [78]. Most of the semiempirical methods use modified *ab initio* algorithms [79,80]. Representatively for all these methods, we will discuss our VESPA method for deriving high-quality ESP charges using semiempirical MO methods [81,82].

The charge-fitting process begins with the calculation of the electrostatic potential in a grid around the molecule. The points were generated using a modified Marsili algorithm [83,84] with variable step size (edge length), which enables the user to control the number of the grid points. All points within 1.4 times the van der Waals radius of the molecule were eliminated, the maximum distance of the grid points is one-third of the vdW radius. Figure 7 (see p. 145) shows such a grid produced by VESPA using a step size of 0.3 Å.

The charge-fitting process begins with the calculation of the electrostatic potential for each point. The electrostatic potential at these points is then calculated using the NAO-PC model [85,86], described later. This enables us to construct a very fast algorithm for the calculation of ESP-derived atomic charges, in which the most time-consuming step is the final linear least-squares fit procedure suggested by Chirlian and Francl [75]. The only constraint during this fit is that the sum of the ESP charges have to reproduce the molecular charge, but it is also possible to use, for example, the molecular dipole moment for this purpose.

In many cases, this procedure gives charges of 6-31G\* quality, especially for AM1 and PM3. For example, for a test set of 27 organic compounds, AM1-VESPA charges correlate with the HF/6-31G\*(MK) ESP charges with  $R = 0.934$  and a standard deviation of 0.108. If the phosphorus compounds are omitted, the correlation coefficient increases to 0.954 (standard deviation of 0.104). Without the P- and S-compounds, the correlation coefficient reaches 0.960 (0.098). Using the CHELPG method [76] instead of MK to obtain the 6-31G\* values has no influence on the correlation obtained.

Kollman and co-workers [87] have described the RESP method, designed to improve the quality of the ESP charges. They use a restraint in the form of a hyperbolic penalty function in the charge-fitting procedure. This requires an iterative solution to self-consistency in  $q_i$  (the point charge at atom  $j$ ). A similar approach for the VESPA algorithm was tested, but no real improvement was achieved. In some cases, we experienced major convergence problems.

It was also shown that the VESPA approach to determining electrostatic potential-derived atomic charges using semiempirical techniques essentially is orientationally invariant. Conformational variation gives smooth and reasonable changes in both the

molecular electrostatic potential and the charges derived from it. On the other hand, Francl et al. [88] have shown that the least-squares matrix for this fitting problem may be rank deficient, and that statistically valid charges cannot always be assigned to all atoms in a molecule. This problem increases with increasing size of the molecule. We were able to show that an increasing number of grid points can improve the quality of the point charges obtained, especially for buried atoms, but does not always lead to well-defined atomic charges [82].

Since this method is so effective at generating monopoles that reproduce the electrostatic potential of a molecule, it is possible to use VESPA for large bio-organic systems [82]. In this case, however, there is a critical relationship between the structure of the molecule and the quality of the results. Atoms far away from the nearest grid points have very poorly defined charges (e.g. those inside the helix of  $\alpha$ -helical Ala<sub>28</sub>). This is a general disadvantage of all methods that use grid points at defined distances around a molecule.

#### 4.2. Multi-center approaches

Two methods are mentioned here. The first one is the distributed multiple analysis developed by Stone in 1981 [89]. Stone's aim was to achieve a reliable description of the charge distribution in a molecule that is suited to the calculation of the electrostatic potential outside the charge distribution itself. The method was defined as an extension of the Mulliken population analysis. It uses a multipole expansion instead of point charges to describe the charge distribution. A detailed description is given in the original publication. The second is the NAO-PC model, mentioned above. The natural atomic orbital-point charge model (NAO-PC) has been developed to calculate accurate molecular electrostatic potentials within the semiempirical approximation [85,86].

This model uses nine point charges (including the core charge) to represent heavy atoms, because it is impossible to fit the quantum chemical potential properly by a simple atom-centered point charge model [90]. Figure 8 illustrates the arrangement of the NAO-PCs for formaldehyde. The positions and magnitudes of the eight charges that represent the atomic electron cloud are calculated from the natural atomic orbitals (NAOs) and their occupations. Each hybrid NAO is represented by two point charges situated at the centroid of each lobe. The positions of the centroids and the magnitudes of the charges are obtained by numerical integration of the Slater-type hybrids and the results used to set up polynomials and look-up tables that replace the integration step in the actual MEP calculation.

Not only MEPs, which will be described later, but also atomic and molecular multipole moments up to the octupole moment [89,91,92], are well represented by the NAO-

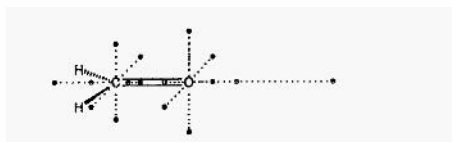


Fig. 8. Schematic representation of the NAO-PCs for the C and O atom in formaldehyde.



PC method. Molecular dipoles are calculated directly from the NAO-PCs and the results are compared to those obtained using the standard calculational method [93]. The standard deviation between NAO-PC and directly calculated dipole moments for a test set of 90 organic molecules is 0.171 Debye. The two largest deviations are for  $\text{H}_2\text{S}$  (0.78 D) and  $\text{SCl}_2$  (0.51 D), suggesting that NAO-PC performs less well for sulfur than for the other elements tested.

The calculation of the quadrupole moments uses the definition of Buckingham [94]. Calculations were performed for 20 test molecules. The standard deviations between calculated and experimental values are 1.2 and 0.7 Debye. Ångström for  $Q_{xx}$  and  $Q_{yy}$ , respectively. The correlation is even better if the experimental margins of error are included.

The surprisingly good results for molecular multipole moments indicate that NAO-PC can describe an NDDO wave function nearly completely without significant loss of accuracy. It, therefore, provides a possibility to reduce a molecular wave function to a discrete number of point charges.

The calculation of atomic multipoles is also possible within this model. Atomic dipoles especially give detailed information about electronic and structural properties. As atomic multipoles are not measurable quantities within molecules, however, the discussion was limited to molecular moments.

### 4.3. Electrostatic potential

Electrostatic interactions are known to play a key role in determining the structure and activity of biomolecules [95,96,97]. Therefore, the electrostatic potential is also important within hybrid QM/MM models. A lot of work has been directed toward calculating reliable molecular electrostatic potentials from semiempirical methods [58,85,86,98-103]. Some of these methods will be described in the following.

The MEP, in general, is given by

$$V(r) = \sum_{\alpha} \frac{Z_{\alpha}}{|R_{\alpha} - r|} - \int \frac{\rho(r') dr'}{|r' - r|}$$

where  $V(r)$  is the electrostatic potential at any point  $r$ ;  $Z_{\alpha}$  is the charge of atom  $\alpha$  located at  $R_{\alpha}$ ; and  $\rho(r')$  is the electronic density function of the molecule.

Within the monopole approximation, which is often used, this equation simplifies to

$$V_i = \sum_{j=1}^n \frac{q_j}{r_{ij}}$$

where  $n$  is the number of atoms;  $q_j$  is the atomic point charge; and  $r_{ij}$  is the distance between atom  $j$  and grid point  $i$ .

Using the NAO-PC model, it becomes

$$V(\vec{r}) = \sum_{\alpha (\alpha \in H)} \sum_{i=1}^{2N_{\alpha} + 1} \frac{q_{i\alpha}}{|\vec{r}_{i\alpha} - \vec{r}|} + \sum_{\alpha (\alpha \in H)} \frac{q_{\alpha}}{|\vec{R}_{\alpha} - \vec{r}|}$$



Fig. 5. Polarization of the inhibitor *L*-benzylsuccinate (bzs) within the active site of carboxypeptidase A. The arrow shows the induced change in the dipole moment.

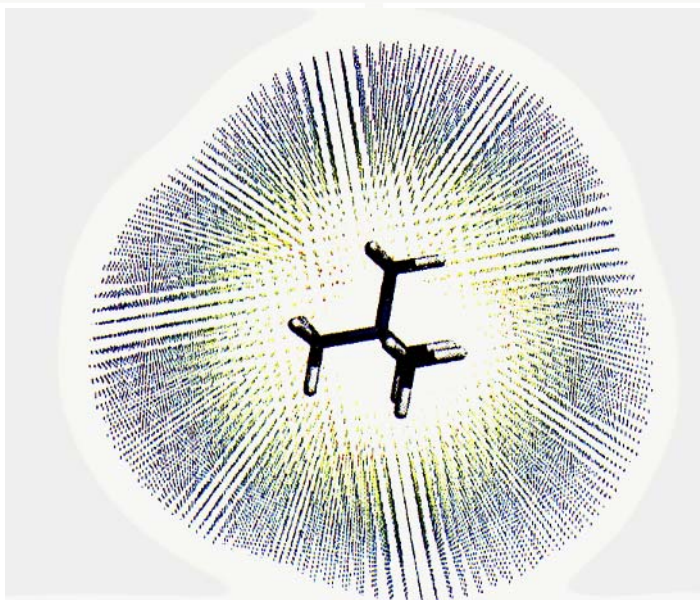


Fig. 7. The tetramethylammonium ion surrounded by a grid of points generated with VESPA. Several point layers have been removed in order to show the cavity,

where  $q_{ia}$  is the charge of the NAO-PCs located at  $r_{ia}$ ;  $q_{\alpha}$  is the charge from the hydrogen atoms located at  $r_{\alpha}$ ; and  $n_{orb}$  is the number of NAOs.

The double sum gives the contribution of the heavy atoms to the electrostatic potential, and the second term treats the hydrogen atoms. For a test set consisting of 12 molecules, the MEPs calculated using the NAO-PC approximation were compared to *ab initio* RHF/6-31 G\* results obtained with Gaussian 92 [104]. As shown in the original publication [86], the errors are between 10% and 14%. The correlation coefficients vary between 0.91 and 0.99, with the exception of sulfanilamide (0.889) and tosyl chloride (0.855). We conclude that the NAO-PC model is a reliable and very fast method for the calculation of MEPs.

Bakowies and Thiel [58] developed a parameterized method to compute the electrostatic potential using the AM1 and MNDO wave functions. The procedure is based on a previously suggested one [105]. A large set of *ab initio* RHF/6-31G\* reference data have been used to calibrate the method.

Applying the final parameters (H, C, N, O), the *ab initio* electrostatic potentials are reproduced with an average accuracy of 20% (AM1) and 25% (MNDO), respectively. Kikuchi et al. published a method for the fast evaluation of molecular electrostatic maps for amino acids, peptides and proteins by empirical functions [106]. The MEPs due to valence electrons are calculated by a set of simple empirical functions at various origins. Those due to the core electrons and nuclei are considered by a point charge approximation. The results are compared to *ab initio* STO-5G values. For amino acids, for example, correlation coefficients between 0.93 and 0.97 were obtained. In the final section, we describe some possible applications of semiempirical binding site models in 3D QSAR.

## 5. Applications

As mentioned above, the pure semiempirical methods for the calculation of large molecular systems are still under development. Therefore, no applications of these techniques have yet been published in the area of 3D QSAR. In contrast, QM/MM methods have been used within this research area. There are two main possibilities. The first is to use QM/MM methods to study reaction pathways of enzyme reactions [107,108] or for docking studies [109] or even for MD simulations [51], in order to find new lead compounds or active conformations that can be used in further QSAR studies. On the other hand, these methods can be used for the prediction of biochemically relevant properties such as absolute binding free energies [110] or the inhibition strength [111]. In the following, some of the applications mentioned above will be described in more detail.

### 5.1. Reaction mechanisms

Richards et al. [107], for example, have examined chorismate mutase catalysis. This enzyme catalyses the skeletal rearrangement of chorismate to prephenate by selecting a destabilized conformer of chorismate. It was found that the minimum energy enzyme-substrate complex has chorismate in a distorted geometry compared to the gas-phase

ground-state structure. It could also be shown that the two residues Arg90 and Glu78 are the most important residues during the reaction. This investigation used the CHARMM QM/MM method [18]. Mulholland and Karplus [108] used the same method for investigations on triose phosphate isomerase (TIM), citrate synthase (CS) and the FK506-binding protein (FKBP).

Properly applied QM/MM methods can give accurate reaction paths, they are able to quantify the interactions involved and determine the contributions of individual residues. These data are very useful in the design of new lead compounds. Another approach are docking studies, which give action conformations of ligands and binding points within the active site directly.

## 5.2. QM/MM docking

Molecular docking is one way to formulate the problem of molecular recognition computationally. Docking is simply the collision of the substrate with the binding site in the correct conformation and orientation. In order to apply docking methods successfully, the active site geometry should be known, either from X-ray crystallography or by homology modelling for proteins with known sequence but unknown structure.

A combination of a genetic algorithm [112] with our VAMP-PCM method [24,113], mentioned above, was used to approach the docking problem. While the genetic algorithm covers the conformational and orientational flexibility of the ligand, the QM/MM method refines the interactions between the substrate and the environment. We have used this approach for the prediction of the docking positions of several cyclic nucleotides within experimentally known (3GAP, 1GKY) and a modelled binding domain (1APK). The 3D structure of the RIA regulatory subunit was modelled by Weber et al. in 1987 [114] and is available from the Brookhaven Database.

In order to compare the theoretical predictions with the X-ray structures and to decide which of them should be applied to the modelled binding domain, the different approaches for the docking problem were first tested on known substrate–enzyme complexes.

### 5.2.1. 3GAP–cAMP

The cAMP binding domain of the bacterial catabolite gene activator protein CAP served as the reference protein during the homology modelling of the cAMP binding domain of RIA [114]. CAP senses the level of cAMP and regulates transcription from several operons in *E. coli*. cAMP (Fig. 9) serves as a hunger signal, both in bacteria and mammals. The crystal structure of the CAP dimer with two bound molecules of cAMP was published by McKay and Steitz in 1981 [115], and refined by Weber and Steitz in 1987 to a resolution of 2.5 Å (entry 3GAP in the Brookhaven Database) [116].

It is remarkable that VAMP-PCM, starting with the best 20 solutions of the GA, optimizes all of them to be inside the pocket (6 GA solutions were originally outside). This shows that the point charge model is able to describe the interactions between the substrate molecule and the protein adequately. The best solution obtained from the different methods within the binding site of the enzyme are shown in Fig. 10.

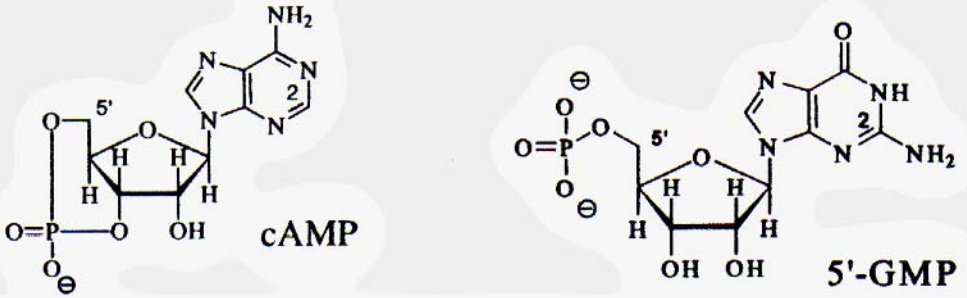


Fig. 9. Schematic structures of cyclic adenosine-monophosphate (cAMP), 5'-guanosine-monophosphate (5'-GMP) and cyclic guanosine-monophosphate (cGMP).

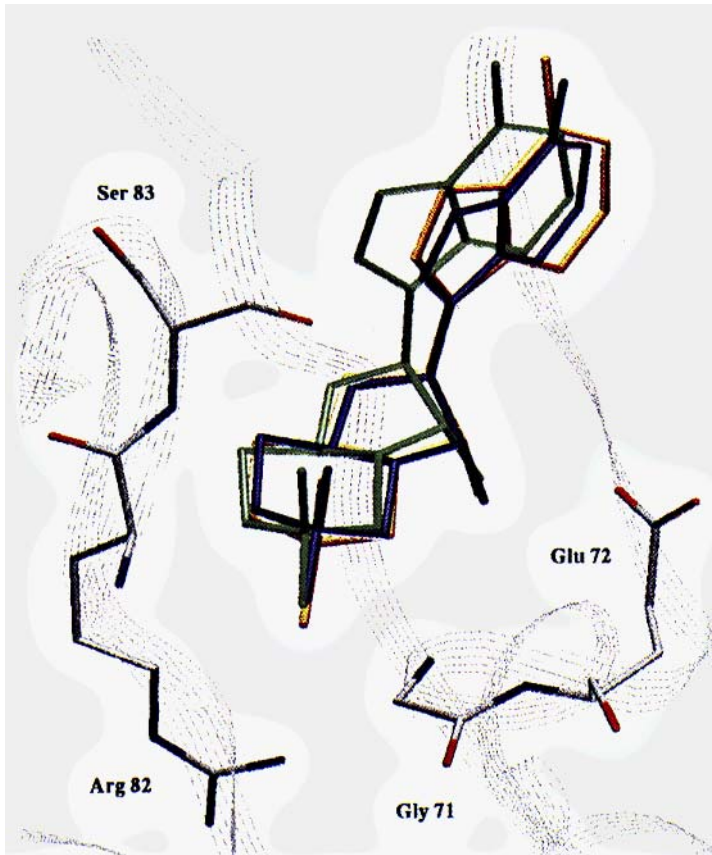


Fig. 10. Binding site of CAP complexed with cAMP. Solutions of the different methods are as follows: X-ray (blue), GA (green); VAMP-PCM (red); and GA/VAMP-PCM (yellow).

The RMS deviations compared to the X-ray structure are 0.80 Å, 0.78 Å and 0.76 Å for the GA, VAMP-PCM and GA/VAMP-PCM solutions, respectively. The results are very similar and all deviations are within the experimental resolution. Especially for the cyclo-phosphate and ribose parts of cAMP, which are located in a tight binding pocket (main residues are Arg82, Ser83, Gly71 and Glu72), the predictions are nearly identical with the X-ray structure.

For the aromatic part of cAMP, the deviations are larger because, in the case of the CAP monomer, no residue binds directly to the adenosine part of the ligand, as is also true for the CAP dimer.

### 5.2.2. 1GKY-5'-GMP

As a second test case, we chose guanylate kinase (entry 1GKY in the Brookhaven Database) complexed with guanosine-5'-monophosphate (5'-GMP). The enzyme was isolated from baker's yeast and crystallized as a complex with its substrate GMP by

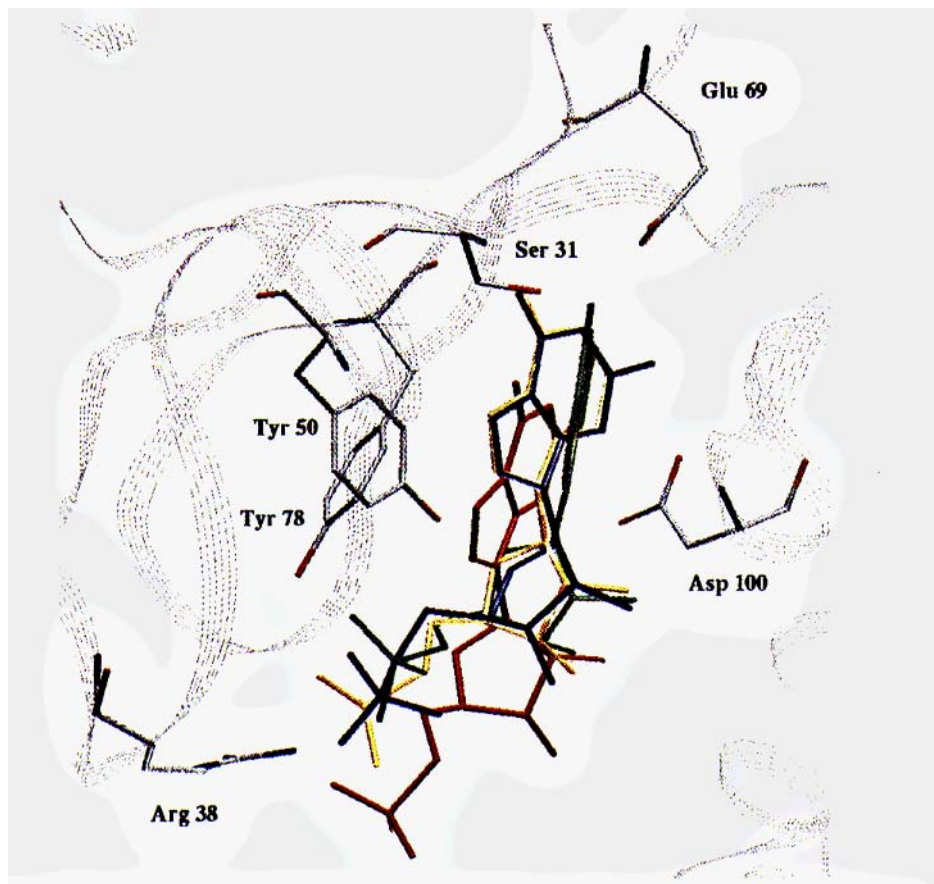


Fig. 11. 1GKY complexed with 5'-GMP. Solutions of the different methods are as follows: X-ray (blue); GA (green); VAMP-PCM started from displaced X-ray geometry (red); and VAMP-PCM started from best GA solution (yellow).

Stehle and Schulz [117]. Its X-ray structure has been refined at a resolution of 2.0 Å [118]. The enzyme catalyzes the reaction  $\text{ATP} + \text{GMP} \rightleftharpoons \text{ADP} + \text{GDP}$ . The ligand guanosine-5'-monophosphate, 5'-GMP is shown in Fig. 9. Again, the X-ray structure of 1GKY/5'-GMP was used as starting point for this investigation. The results of the different docking algorithms within the active site of 1GKY are shown in Fig. 11. The experimentally determined position of the ligand seems to be only one of several possible positions in the binding site of the enzyme. Although the binding site is very large, and therefore allows more conformational flexibility, the best predicted positions of 5'-GMP using the GA and the combined method (VAMP-PCM started with the GA solutions) are in the good agreement with the X-ray structure.

### 5.2.3. Prediction of the docking position of cAMP in the RIA binding domain

As mentioned above, the binding domain of RIA regulatory subunit was modelled by homology using the cAMP binding domain of CAP as reference [114,116]. For a better

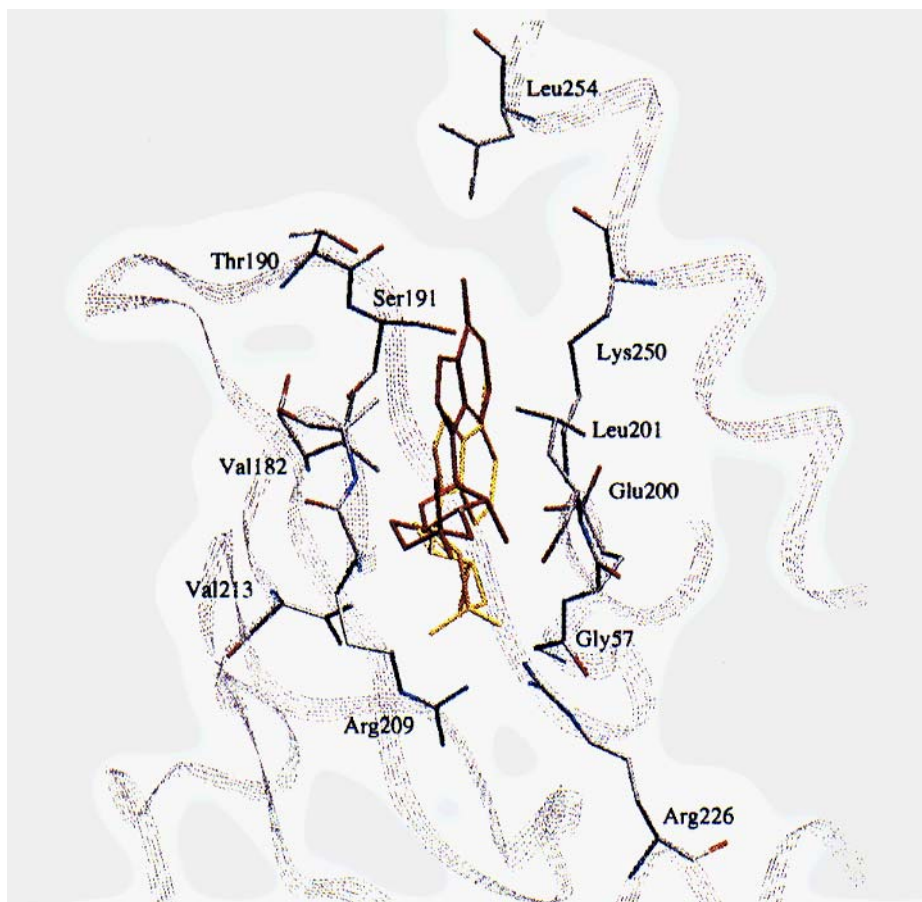


Fig. 12. Predicted docking position of cAMP in the RIA binding site (yellow). The original position of cAMP (red) is shown as reference.

orientation, the best solution of our docking approach and the original docking position (APK) within the modelled binding domain are shown in Fig. 12.

The RMS deviation between the original and our predicted positions of cAMP is 1.12 Å. Considering the theoretical refinement of 2.5 Å for the modelled structure (1apk), this is a small value. Remarkable in this modelled binding site is that, in addition

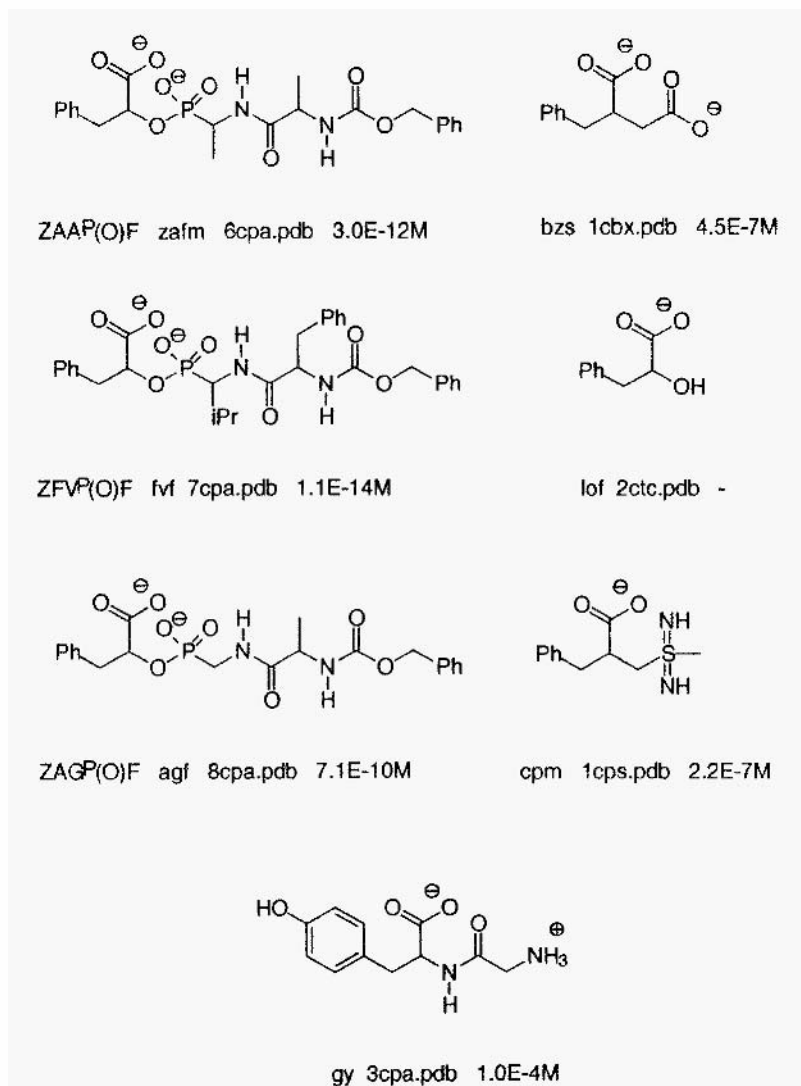


Fig. 13. Schematic view of *L*-benzylsuccinate (*bzs*), *L*-phenyllactate (*lof*), (*L*-(-)-2-carboxy-3-phenylpropyl) methylsulfodimine (*cpm*), glycyl-*L*-tyrosine, (*gy*), *O*-[[*(1R)*-[[*N*-(phenylmethoxycarbonyl)-*L*-alanyl] amino] ethyl] hydroxyphosphinyl]-*L*-3-phenyllactate (ZAA<sup>P</sup>(O)F, *zaf*), *O*-[[*(1R)*-[[*N*-(phenylmethoxycarbonyl)-*L*-phenylalanyl] amino] isobutyl] hydroxyphosphinyl]-*L*-3-phenylacetate (ZFV<sup>P</sup>(O)F, *fvf*) *O*-[[*(1R)*-[[*N*-(phenylmethoxycarbonyl)-*L*-alanyl] amino] methyl] hydroxyphosphinyl]-*L*-3-phenyllactate (ZAG<sup>P</sup>(O)F, *agf*). and their experimental  $K_i$  values.



to Arg209, Arg226 also contributes to the binding of cAMP. This is the main reason for the deviation obtained. Because of the strong electrostatic interaction between the cyclo-phosphate part of cAMP and the two arginines, the whole ligand moves toward these residues during the optimization in VAMP-PCM.

Using different dielectric constants ( $\epsilon = 1, 2, 4$ ) in the QM/MM calculations did not give significant changes in the results discussed above. The results obtained suggest that the combination of a genetic algorithm with a quantum mechanical/molecular mechanical approach provides a powerful tool for treating the docking problem. The average CPU time for this optimization on a R8000 (90 MHz) Power Challenge is around 700 s.

### 5.3. Binding energies

QM/MM methods can also be used for the prediction of absolute binding free energies [110] or for the prediction of binding affinities [111]. In the following, we describe a

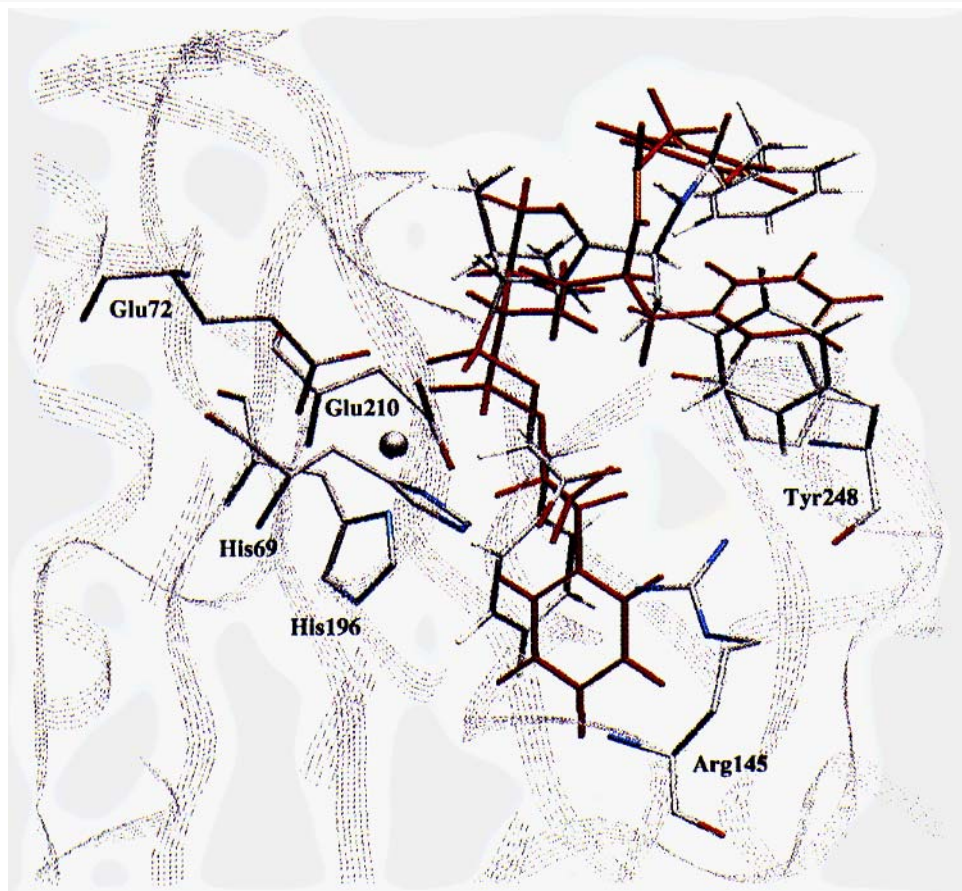


Fig. 14. X-ray (red) and optimized structure of ZFV(O)F within the binding site of CPA. The Delta value, which will be correlated to the inhibition strength, is then obtained by  $\Delta = \Delta H_r(\text{PCM}) - \Delta H_r(\text{Ligand/Bulk}) + \Delta H_r(\text{Bulk})$ .

case study for the performance of the method for docking flexible molecules into a protein-binding site. We have used the VAMP-PCM method to rank a set of seven carboxypeptidase A (CPA) inhibitors according to their inhibition strength. The binding geometries of these ligands are known crystallographically, and also the experimental  $K_i$  values for six of them [119-121]. To account for the effects of solvation, we decided to use a supermolecule approach with 28 water molecules surrounding the ligand. The schematic structures of the inhibitors are shown in Fig. 13. A representative example of an optimized phosphonate inhibitor (ZFV<sup>P</sup>(O)F) in CPA is shown in comparison with their X-ray structures in Fig. 14.

In Fig. 15, delta is plotted against  $\log K_i$  for live out of the seven inhibitors to visualize the correlation between these values. The zwitterionic gy ligand is not shown in the plot, because of a highly positive delta (191.91). This indicates that our approach in the present form may have some problems with zwitterionic structures. Nevertheless, the ligand is ranked correctly. For the lof inhibitor, no experimental inhibition strength is known; we rank this compound as the second worst inhibitor in our test set. The results above suggest that lof is a much worse inhibitor than the structurally similar bzs ligand.

Thus, using a combined QM/MM approach for the simulation of the protein environment and a supermolecule approach for solvation effects allows us to reproduce the inhibition strength ranking of the chosen set of CPA ligands correctly. In order to verify this result and to strengthen this approach, further investigations are in progress. To improve the calculation of the solvent effects, one can also use QM/MM models [32-46].

## 6. Conclusion and Outlook

The above examples demonstrate the emerging power of semiempirical MO theory in biological applications. There is still much development (not least in the quantum

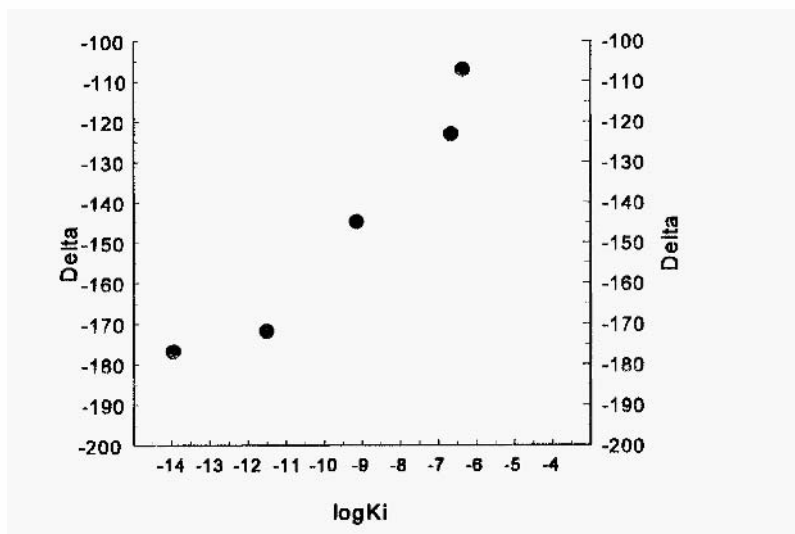


Fig. 15. Plot delta versus  $\log K_i$ .

mechanics of the methods themselves) to be done. but modern hardware and software have opened new possibilities for applying MO theory to systems that would have been treated with force fields only a few years ago. Much of the battle will be to overcome the negative image that early semiempirical methods gained within the organic community. This will only be possible by clear and detailed demonstrations of the advantages of using quantum mechanics in real biological applications.

## References

1. Ferenczy, G.G., Rivail, J-L., Surján P.R. and Náray-Szabó, G., *NDDO fragment self-consistent field approximation for large electronic systems*, J. Comput. Chem., 13 (1992) 830–837.
2. Náray-Szabó, G., *Towards a molecular orbital method for the conformational analysis of very large biomolecules*. Acta Phys. Acad. Sci. Hung., 40 (1976) 261–273.
3. Stewart, J.J.P., *Application of localized molecular orbitals to the solution of semiempirical self-consistent field equations*. Int. J. Quant. Chem., 58 (1996) 133–146.
4. Zhao, Q. and Yang, W., *Analytical energy gradients and geometry optimisation in the divide-and-conquer methods for large molecules*, J. Chem. Phys., 102 (1995) 9598–9603.
5. Yang, W. and Lee, T.-S., *A density-matrix divide-and-conquer approach for electronic structure calculations of large molecules*, J. Chem. Phys., 103 (1995) 5674–5678.
6. Lee, T.-S., York, D.M. and Yang, W., *Linear scaling semiempirical quantum calculations of macromolecules*, J. Chem. Phys., 105 (1996) 274–42750.
7. Dixon, S.L. and Merz, Jr., K.M., *Semiempirical molecular orbital calculations with linear system size scaling*, J. Chem. Phys., 104 (1996) 6643–6649.
8. White, C.A., Johnson, B.G., Gill, P.M.W. and Head-Gordon, M., *The continuous fast multiple method*, 230 (1994) 8–18.
9. Strain, M.C., Scuseria, G.E. and Frisch, M.J., *Achieving linear scaling for the electronic Coulomb problem*, Science 271 (1996) 51–53
10. Teeter, M.M., Roe, S.M. and Heo, N.H., *Atomic resolution (0.83 Å) crystal structure of the hydrophobic protein crambin at 130 K*, J. Mol. Biol., 230 (1993) 292–311.
11. Deisenhofer, J., *Crystallographic refinement of the structure of bovine pancreatic trypsin inhibitor at 1.5 Å resolution*, Acta Crystallograph. B, 31 (1975) 238–250.
12. Stewart, J.J.P., *MOPAC7 Version 2 manual*, QCPE, Bloomington, IN, 1993.
13. Klant, A. and Schürmann, G., *COSMO: A new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradients*, Perkin Trans., 2 (1993) 799–805.
14. Troung, T.N. and Stephanovich, E.V., *Analytical first and second energy derivatives of the generalized conductorlike screening model for free energy of solvation*, J. Chem. Phys., 103 (1995) 3709–3717.
15. Andzelm, J., Kölnmel, C. and Klant, A., *Incorporation of solvent effects into density functional calculations of molecular energies and geometries*, J. Chem. Phys., 103 (1995) 9312–9320.
16. York, D., Lee, T.S. and Yang, W., Chem. Phys. Lett. (submitted).
17. Washel, A. and Levitt, M., *Theoretical studies of enzymic reactions: 1. Dielectric electrostatic and steric stabilization of the carbanion ion in the reaction of lysozyme*, J. Mol. Biol., 103 (1976) 227–235.
18. Field, M.J., Bash, P.A. and Karplus, M., *A combined quantum mechanical and molecular mechanical potential for molecular dynamics simulation*, J. Comput. Chem., 11 (1990) 700–733.
19. Vasiyev, V.V., Bliznyuk, A.A. and Voityuk, A.A., *A combined quantum chemical molecular mechanical study of hydrogen bonded systems*, Int. J. Quant. Chem., 44 (1992) 897–930.
20. Théry, V., Rinaldi, D., Rivail, J.-L., Maigret, B. and Ferenczy, G.J., *Quantum mechanical computations on very large systems: The local self-consistent self method*, J. Comput. Chem., 15 (1994) 269–282
21. Thompson, M.A., Glendening E.D. and Feller, D., *The nature of K<sup>+</sup>/crown ether interactions A hybrid quantum mechanical-molecular mechanical study*, J. Phys. Chem., 98 (1994) 10465–10476.
22. Thompson, M.A. and Schenter, G.K., *Excited states of the bacteriochlorophyll 6 dimer of rhodospseudomonas viridis: A QM/MM study of the photosynthetic reaction center that includes MM polarisation*, J. Phys. Chem., 99 (1995) 6374–6386.

23. Bakowies, D. and Thiel, W., *Hybrid models for combined quantum mechanical and molecular mechanical approaches* J. Phys. Chem., 100 (1996) 10580–10594.
24. Alex, A., Beck, B., Lanig, H., Rauhut, G. and Clark, T. (paper in preparation).
25. Stanton, R.V., Hartsough, D.S. and Merz, Jr., K.M., *An examination of a density functional/molecular mechanical coupled potential* J. Comput. Chem., 16 (1996) 113–128.
26. Bernardi, F., Olivucci, M. and Robb, M.A., *Simulation of MC-SCF results on covalent organic multi-bonds reactions: Molecular mechanics with valence bond (MM-VB)*, J. Am. Chem. Soc., 114 (1992) 1606–1616.
27. Aqvist, J. and Warshel, A., *Simulation of enzyme reactions using valence bond force fields and other hybrid quantum classical approaches*, Chem. Rev., 93 (1993) 2523–2530.
28. Singh, U.C. and Kollman, P.A., *A combined ab initio quantum mechanical and molecular mechanical method for carrying out simulations on complex molecular systems: Applications to the  $\text{CH}_3\text{Cl} + \text{Cl}^-$  exchange reaction and gas phase protonation of polyethers*, J. Comput. Chem., 7 (1986) 718–730.
29. Allinger, N.L., Yuh, Y.H. and Lii, J.-H., *Molecular mechanics: The MM3 force field for hydrocarbons* I.J. Am. Chem. Soc., 111 (1989) 8551–8582.
30. Pearlman, D.A., Case, D.A., Ross, J.W., Cheatham. III. T.E., Ferguson, D.M., Seibel, G.L., Singh, U.C., Weiner, P.K. and Kollman, P.A., *AMBER 4.1*, University of California, San Francisco, CA, 1995.
31. Brooks, B.R., Burccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S. and Karplus, M., *CHARMm: A program for macromolecular energy, minimization and dynamics calculations*. J. Comput. Chem., 4 (1983) 187–217.
32. Bash, P.A., Field, M.J. and Karplus, M., *Free energy perturbation method for chemical reactions in the condensed phase: A dynamical approach based on a Combined quantum and molecular mechanics potential* J. Am. Chem. Soc., 109 (1987) 8092–8094.
33. Gao, J., *Absolute free energy of solution from Monte Carlo simulations using combined quantum and molecular mechanical potentials*, J. Phys. Chem., 96 (1992) 537–540.
34. Gao, J. and Xia, X., *A priori evaluation of aqueous polarization effects through Monte Carlo QM/MM simulations*, Science, 258 (1992) 631–635.
35. Gao, J. and Pavelites, J.J., *Aqueous basicity of the carboxylate lone pairs and the C-O barrier in acetic acids: A combined quantum and statistical mechanical study*, J. Am. Chem. Soc., 114 (1992) 1912–1914.
36. Gao, J., *Hybrid quantum and molecular mechanical simulations: An alternative avenue to solvent effects in organic chemistry*, Acc. Chem. Rea., 27 (1993) 298–305.
37. Gao, J., Luque, F.J. and Orozco, M., *induced dipole moments and atomic charges based on average electrostatic potentials in aqueous solution*, J. Chem. Phys., 98 (1993) 2975–2982.
38. Gao, J., *Potential of mean force for the isomerization of DMF in aqueous solution: A Monte Carlo QM/MM simulation study* J. Am. Chem. Soc., 115 (1993) 2930–2935.
39. Gao, J. and Xia, X., *A tow-dimensional energy surface for a type II  $\text{S}_{\text{N}}^2$  reaction in aqueous solution*, J. Am. Chem. Soc., 115 (1993) 9667–9675.
40. Stanton, R.V., Hartsough, D.S. and Merz, Jr., K.M., *Calculation of solvation free energies using a density functional molecular dynamics coupled potential*, J. Phys. Chem., 97 (1993) 11868–11870.
41. Gao, J., *Combined QM/MM simulation study of the Claisen rearrangement of allyl vinyl ether in aqueous solution*, J. Am. Chem. Soc., 116 (1994) 1563–1564.
42. Liu, H. and Shi, Y., *Combined molecular mechanical and quantum mechanical potential study of a nucleophilic addition reaction in solution*, J. Comput. Chem., 15 (1994) 1311–1318.
43. Liu, H., Müller-Plathe, F. and Van Gunsteren, W.F., *A molecular dynamics simulation study with a combined quantum mechanical and molecular mechanical potential energy function: Solvation effects on the conformational equilibrium of demethoxy ethane*, J. Chem. Phys., 102 (1995) 1722–1730.
44. Hartsough, D.S. and Merz, Jr., K.M., *Potential of mean force calculations on the  $\text{S}_{\text{N}}^1$  fragmentation of tert-butyl chloride*, J. Phys. Chem., 99 (1995) 384–390.
45. Stanton, R.V., Little, L.R. and Merz, Jr., K.M., *Quantum free energy perturbation study within a PM3/MM coupled potential*, J. Phys. Chem., 99 (1995) 483–486.
46. Thompson, M.A., *Hybrid quantum mechanical/molecular mechanical force field development for large flexible molecules: A molecular dynamics study of 18-crown-6*, J. Phys. Chem., 99 (1995) 4794–4804.

47. Bash. P.A., Field. M.J., Davenport. R.C., Petsko. G.A., Ringe. D. and Kaiplus. M., *Structure of the triosephosphate isomerase phosphoglycolohydroxamate complex: An analog of the intermediate on the reaction pathway*, *Biochemistry* 30 (1991) 5821–5826.
48. Waszkowycz B., Hiller, I.H., Gensmantel. N. and Payling, D.W., *Combined quantum mechanical–molecular mechanical study of catalysis by the enzyme phospholipase A2: An investigation of the potential-energy surface for amide hydrolysis*. *J. Chem. Soc., Perkin Trans. 2* (1991) 225–2032.
49. Vasilyev, V.V., *Tetrahedral intermediate formation in the acylation step of acetylcholinesterases A combined quantum chemical and molecular mechanical model*, *J. Mol. Struct. (THEOCHEM)*, 304 (1994) 129–141 .
50. Elcock, A.H., Lyne, P.D., Mulholland, A J., Nandra. A. and Richards, W.G., *Combined quantum and molecular mechanical study of DNA cross-linking by nitrous acid*, *J. Am. Chem. Soc.*, 117 (1995) 4706–4707.
51. Hartsough, D.S. and Merz, Jr., K.M., *Dynamic force field models: Molecular dynamics simulations of human carbonic anhydrase II using a quantum mechanical/molecular mechanical coupled potential*, *J. Phy. Chem.*, 99 (1995) 11266–11275.
52. Born, M. and Von Karman, T., *Über Schwingungen in Raumgittern*, *Physik. Z.*, 13 (1912) 297–309.
53. Brooks, III, C.L. and Kaiplus. M., *Deformable stochastic boundaries in molecular dynamics* *J. Chem. Phys.* . 79 (1983) 6312–6325.
54. Brünger. A.T., Huber, R. and Kaiplus. M , *Trysinogen-trypsin transition: A molecular dynamics study of induced conformational change in the activation domain*, *Biochemistry*. 26 (1987) 5153–5164.
55. Davis, T.D., Maggiora, G.M. and Christoffersen. R.E., *Ab initio calculations on large molecules using molecular fragments: Unrestricted Hartree fock calculations on low lying states of formaldehyde and its radical ions*, *J. Am. Chem. Soc.*, 96 (1974) 7878–7887.
56. Clementi, E., *Computational aspects for large chemical systems: Lecture notes in chemistry*, Springer, New York, 1980.
57. Mascras, F. and Morokunia, K., *IMOMM: A new integrated ab initio + molecular mechanics geometry optimisation scheme of equilibrium structures and transition states*, *J. Comput. Chem.*, 16 (1995) 1170–1179.
58. Bakowies, D. and Thiel, W ., *Semiempirical treatment of electrostatic potentials and partial charges in combined quantum mechanical, molecular mechanical approaches* *J. Comput. Chem.*, 17 (1996) 87–108.
59. Thole, B.T., *Molecular polarizabilities calculated with a modified dipole interaction*. *Chem. Phys.*, 59 (1981) 341–350.
60. Monard, G., Loos M., Théry. V., Baka, K. and Rivail, J.-L., *Hybrid classical quantum force field for modeling very large molecules*, *Int. J. Quant. Chem.* 58 (1996) 153–159.
61. Rappé, A.K., Caswit, C.J., Colwell, K.S., Goddard. III, W.A. and Skiff. W.M , *UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations*, *J. Am. Chem. Soc.*, 114 (1992) 10024–10035.
62. Luzhkov, V. and Warshel, A., *Microscopic calculations of solvent effects on absorption spectra of conjugated molecules*, *J. Am. Chem. Soc.*, 113 (1991) 4491–4499.
63. Luzhkov, V. and Warshel, A., *Microscopic models for quantum mechanical calculations of chemical processes in solution: LD/AMPAC and SCAAS/AMPAC calculations of solvation energies*, *J. Comput. Chem.*, 13 (1992) 199–213
64. Vesely, F.J ., *N-particle dynamics of polarizable Stockmayer-type molecules*, *J. Comput. Phys.*, 24 (1977) 361–371.
65. Ahlstrom, P., Wallqvist, A., Engstrom. S. and Jonsson, B., *A molecular dynamics study of polarizable water*, *Mol. Phys.*, 68 (1989) 563–581.
66. Dang, L.X., Rice. J.E., Caldwell, J. and Kollman, P.A., *Ion solvation in polarizable water: Molecular dynamics simulation*, *J. Am. Chem. Soc.*, 113 (1991) 2481–2486.
67. Thompson. M.A., *QM/MMpol: A consistent model for solute/solvent polarization: Application to the aqueous solvation and spectroscopy of formaldehyd, acetaldehyd and acetone*, *J. Phys. Chem.*, 100 (1996) 14492–14507.
68. Gasteiger. J. and Marsili, M., *Iterative partial equalization of orbital tronegativity: A rapid access to atomic charges*, *Tetrahedron*. 36 (1990) 3219–3288.

69. (a) Abraham, R.J. and Hudson, B., *Charge calculations in molecular mechanics III: Amino acids and peptides*, J. Comput. Chem., 6 (1985) 173–181. (b) Abraham, R.J. and Smith, P.E., *Charge calculations in molecular mechanics IV: A general method for conjugated systems*, J. Comput. Chem., 9 (1987) 288–297.
70. Coulson, C.A. and Longuet-Higgins, H.C., *The electrostatic structure of conjugated systems: I. General theory*, Proc. Roy. Soc., A191 (1947) 39–60.
71. Mulliken, R.S., *Electronic population analysis on LCAO-MO molecular wave functions, I*, J. Chem. Phys., 23 (1955) 1833–1840.
72. Williams, D.E., *Net atomic charges and multipole models for the ab initio molecular electric potential*, In Lipkowitz, K.B. and Boyd, D.B. (Eds.) *Reviews in computational chemistry*. Vol. 2. VCH, Weinheim, 1991, pp. 219–271.
73. Storer, J.W., Giesen, D.J., Cramer, C.J. and Truhlar, D.G., *Class IV charge models: A new semiempirical approach in quantum chemistry*, J. Comput.-Aided Mol. Design, 9 (1995) 87–110.
74. Rappé, A.K. and Goddard, III, W.A., *Charge equalization for molecular dynamics simulation*, J. Phys. Chem., 95 (1991) 3358–3363.
75. Chirlian, L.E. and Francl, M.M., *Atomic charges derived from electrostatic potentials: A detailed study*, J. Comput. Chem., 8 (1987) 894–905.
76. Breneman, C.M. and Wiberg, K.B., *Determining atom-centred monopoles from molecular electrostatic potentials: The need for high sampling density in formamide conformational analysis*, J. Comput. Chem., 11 (1990) 361–373.
77. Besler, B.H., Merz, Jr., K.M. and Kollman, P.A., *Atomic charges derived from semiempirical methods*, J. Comp. Chem., 11(1990) 431–439
78. Spackman, M.A., *potential derived charges using a grodesic point selection scheme*, J. Comput. Chem., 17(1996) 1–18.
79. Ferenczy, G.G., Reynolds, C.A. and Richards, W.G., *Semiempirical AM1 electrostatic potentials and AM1 electrostatic potential derived charges: A comparison with ab initio values*, J. Comput. Chem., 11(1990)159–169.
80. Orozco, M. and Luque, F.J., *On the Use of AM1 and MNDO wave functions to compute accurate electrostatic charges*, J. Comput. Chem., 11 (1990)909–923.
81. Beck, B., Glen, R.C. and Clark, T., *VESPA: A new, fast approach to electrostatic potential-derived atomic charges from semiempirical methods*, J. Comput. Chem., 18 (1997) 744–756.
82. Beck, B., Glen, R.C. and Clark, T., *A detailed study of VESPA electrostatic potential-derived atomic charges*, J. Mol. Model., 1 (1995) 176–187.
83. Heiden, W., Goetze, T. and Brickmann, J., *Fast generation of molecular surfaces from 3D data fields with enhanced 'marching cube' algorithm*, J. Comput. Chem., 14 (1993) 246–250.
84. Marsili, M., *Computation of volumes and surface areas of organic compounds*, In Jochum, C., Hicks, M.G. and Sunkel, J. (Eds.) *Physical property prediction in organic chemistry*, Springer Verlag, Berlin, 1988, pp. 249–251.
85. Rauhut, G. and Clark, T., *Multicenter point charge model for high-quality molecular electrostatic potentials from AM1 calculations*, J. Comput. Chem., 14 (1993) 503–509.
86. Beck, B., Rauhut, G. and Clark, T., *The natural atomic orbital point charge model for PM3: multipole moments and molecular electrostatic potentials*, J. Comput. Chem., 15 (1 994) 1064–1073.
87. Bayly, C.I., Cieplak, P., Cornell, W.D. and Kollman, P.A., *A well-behaved electrostatic potential based method using restraints for deriving atomic charges: The RESP method*, J. Phys. Chem., 97 (1993) 10269–10280.
88. Francl, M.M., Carey, C., Chirlian, L.E. and Gange, D.M., *Charge fit to electrostatic potentials: II. Can atomic charges unambiguously fit to electrostatic potentials*, J. Comput. Chem., 17 (1996) 367–383.
89. Stone, A.J., *Distributed multipole analysis or how to describe a molecular charge distribution*, Chem. Phys. Lett., 83 (1981)233–239.
90. Chipot, C., Ángyán, J., Ferenczy, G.G. and Scheraga, H.A., *Transferable net atomic charges from a distributed multipole analysis for the description of electrostatic properties: A case study of saturated hydrocarbons*, J. Phys. Chem., 97 (1993) 6628–6636.
91. Sokalski, W.A. and Sawaryn, A., *Correlated molecular and cumulative atomic multipole moments*, J. Chem. Phys., 87 (1987) 526–534.

92. Stogryn, D.E. and Stogryn, A.P., *Molecular multipole moments*, Mol. Phys., 11 (1966) 371–393.
93. Stewart, J.J.P., MOPAC: *A semiempirical molecular orbital program*, J. Comput.-Aided Mol. Design. 4 (1990) 11–12.
94. Buckingham, A.D., *Molecular quadrupole moments*. Quart. Rev., 13 (1959) 183–214.
95. Perutz, M.F., *Electrostatic effects in proteins*, Science, 201 (1978) 1187–1191.
96. Warwicker, J. and Watson, H.C., *Calculation of the electric potential in the active site cleft due to  $\alpha$ -helix dipoles*, J. Mol. Biol., 157 (1982) 671–679.
97. Warshel, A. and Russell, S.T., *Calculations of electrostatic interactions in biochemical systems in solution*, Q. Rev. Biophys., 17 (1984) 283–291.
98. Giessner-Pretre, C. and Pullman, A., *Molecular electrostatic potentials: Comparison of ab initio and CNDO results*, Theor. Chim. Acta, 25 (1972) 83–89.
99. Alhambra, C., Luque, F.J. and Orozco, M., *Comparison of NDDO and quasi-ab initio approaches to compute semiempirical molecular electrostatic potentials*. J. Comput. Chem., 15 (1994) 12–22.
100. Luque, F.J., Illas, F. and Orozco, M., *Comparative study of the molecular electrostatic potential obtained from different wavefunctions: Reliability of the semiempirical MNDO wavefunction*, J. Comput. Chem., 11 (1990) 416–430.
101. Luque, F.J. and Orozco, M., *Reliability of the AMI wavefunction to compute molecular electrostatic potentials*, Chem. Phys. Lett., 168 (1990) 269–275.
102. Alemán, C., Luque, F.J. and Orozco, M., *Suitability of the PM3-derived molecular electrostatic potentials*, J. Comput. Chem., 14 (1993) 799–808.
103. Bharadwaj, R., Windemuth, A., Sridharan, S., Honig, B. and Nicholls, A., *The fast multipole boundary element method for molecular electrostatics: An optimal approach for large systems*, J. Comput. Chem., 16 (1995) 898–913.
104. Gaussian 92. Frisch, M.J., Trucks, G.W., Head-Gordon, M., Gill, P.M.W., Wong, M.W., Foresman, J.B., Johnson, B.G., Schlegel, H.B., Robb, M.A., Replogle, E.S., Gomperts, R., Andres, J.L., Raghavachari, K., Binkley, J.S., Gonzalez, C., Martin, R.L., Fox, D.J., Defrees, D.J.C., Baker, J., Stewart, J.J.P. and Pople, J.A., Gaussian Inc., Pittsburgh, PA. 1992.
105. Ford, G.P. and Wang, B., *New approach to the rapid semiempirical calculation of molecular electrostatic potentials based on the AMI wave function: Comparison with ab initio HF 6-31G\* results*, J. Comput. Chem., 14 (1993) 1101–1111.
106. Nakajima, H., Takahashi, O. and Kikuchi, O., *Rapid evaluation of molecular electrostatic potential maps for amino acids, peptides and proteins by empirical functions*, J. Comput. Chem., 17 (1996) 790–805.
107. Lyne, P.D., Mulholland, A.J. and Richards, W.G., *Insights into chorismate mutase catalysis from a combined QM/MM simulation of the enzyme reaction*, J. Am. Chem. Soc., 117 (1995) 11345–11350.
108. Mulholland, A.J. and Karplus, M., *Simulations of enzymic reactions*, Biochem. Soc. Trans. 24 (1996) 247–254.
109. Beck, B., Lanig, H., Glen, R.C. and Clark, T., J. Med. Chem. (submitted).
110. Alex, A. and Finn, P., Fourth World Congress of Theoretically Oriented Chemists – WATOC '96, Jerusalem, Israel, 1996.
111. Lanig, H., Beck, B. and Clark, T., Poster, MGMS Meeting, York, U.K., 1996.
112. Jones, G., Willet, P. and Glen, R.C., *Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation*, J. Mol. Biol., 245 (1995) 43–53.
113. Rauhut, G., Alex, A., Chandrasekhar, J., Steinke, T., Sauer, W., Beck, R., Hutter, M., Gedeck, P. and Clark, T., VAMP6.0, Oxford Molecular Ltd., Medawar Centre, Oxford Science Park, Sandford-on-Thames, Oxford OX4 4GA, U.K.
114. Weber, I.T., Steitz, T.A., Bubis, J. and Taylor, S.S., *Predicted structures (cAMP binding domains of type I and type II regulatory subunits of cAMP-dependent protein kinase)*, Biochemistry, 26 (1987) 343–351.
115. McKay, D.B. and Steitz, T.A., *Structure of catabolite gene activator protein at 2.9 Å resolution suggests binding to left handed B-DNA*, Nature, 290 (1981) 744–749.
116. Weber, I.T., and Steitz, T.A., *Structure of a complex between catabolite gene activator protein and cyclic AMP refined at 2.5 Å resolution*, J. Mol. Biol., 198 (1987) 311–326.

117. Stehle, T. and Schulz, G.E., *Three-dimensional structure of the complex between guanylate kinase from yeast with its substrate GMP*, J. Mol. Biol., 211 (1990) 249–254.
118. Stehle, T. and Schulz, G.E., *Refined structure of the complex between guanylate kinase and its substrate GMP*, J. Mol. Biol., 224 (1992) 1127–1141.
119. Mangani, S., Carloni, P. and Orioli, P., *Crystal structure of the complex between carboxypeptidase A and the biproduct analog inhibitor L-benzylsuccinate at 2.0 Å resolution*, J. Mol. Biol., 223 (1992) 573–578.
120. Cappalonga, A.M., Alexander, R.S. and Christianson, D.W., *Structural comparison of sulfodiimine inhibitors in their complexes with zinc enzymes*, J. Mol. Biol., 267 (1992) 19192–19197.
121. Kim, H. and Lipscomb, W.N., *Comparison of the structures of three carboxypeptidase A-phosphonate complexes determined by X-ray crystallography*, Biochemistry, 30 (1991) 8171–8180.



**This Page Intentionally Left Blank**

# Density-Functional Theory and Molecular Dynamics: A New Perspective for Simulations of Biological Systems

Wanda Andreoni

*IBM Research Division, Zurich Research Laboratory, CH-8803 Rüschlikon, Switzerland*

## 1. Introduction to the Methods

Density-functional theory (DFT) is an exact theory of the ground state of a many-particle system [1,2]. The mathematical formulation was given by Hohenberg and Kohn (HK) in the mid-1960s [3,4], namely the demonstration of (i) the uniqueness of the ground-state density associated with a given external potential, and of (ii) the variational character of the density energy functional. As a consequence, (iii) the existence of a universal density functional (HK functional) — i.e. independent of the external potential and of the specific system — was established, the exact form of which, however, is unknown. The variational character of the functional makes the ground-state density and energy accessible, in principle, via minimization procedures. A precise prescription to convert this complex problem into a practical scheme for a many-electron system was given by Kohn and Sham [5], in which one-electron orbitals were introduced to describe the electron density. In this way, the solution of the energy minimum problem became formally similar to that of the Hartree and Hartree-Fock approaches, namely it was reduced to a set of equations to be solved iteratively. The underlying physics was more complete, however, being electron–electron correlation included in the HK functional. Also, the meaning of the single-particle orbitals and energies thus obtained was different [1].

There are several reasons why the power of such a method had to wait decades before becoming clear to the community of physicists, and especially to that of chemists. The first step toward a practical application was to introduce a valid approximation for the exchange–correlation energy functional. For a long time, the only practical implementation was the local density approximation (LDA), also suggested by Kohn and Sham [5], which is correct in the limit of the homogeneous electron gas and was expected to be a valid approximation for systems with slowly varying density. LDA was appreciated as physically sound and appropriate for a number of real systems by solid-state physicists who first used it for simple metals and semiconductors, and rapidly realized its validity beyond the realm for which it had originally been designed [6,1]. On the contrary, the formal analogy with jellium being more appropriate for extended systems and the picture so far away from the traditional thinking of chemists, LDA had difficulty in being accepted in the study of molecular systems. Moreover, the relatively poor performance of LDA for isolated atoms and the incorrect description of the tail of the exchange–correlation potential often yielded significant errors for cohesive energies and electron affinities. The scenario changed when gradient corrections to the LDA were introduced in the exchange and correlation functionals (GGA), which improved the agreement between these calculated values and experiment in a significant way. From a fundamental point of view, it must be said that LDA defines a specific reference model

system whose physical meaning is clear because the LDA functional is well defined. GGA suffers instead from the multiplicity of proposed prescriptions and the general lack of a fundamental justification for them. This is even more true for 'hybrid' schemes that have recently become popular in applications to chemistry.

Nowadays, DFT is the method of choice for many problems in physics and chemistry, ranging from molecular to condensed-matter systems. The advance in computational algorithms (and the progress in computer hardware) have also made tests possible on a variety of systems. It is clear by now that DFT can be applied (and successfully) to many real problems for which the size system makes traditional quantum-chemistry calculations impractical.

Further support for the use of DFT has come during the past decade from the so-called Car-Parrinello (CP) method [7,8], which combines it with molecular dynamics (MD). As such, it allows one (i) to use dynamical procedures (such as simulated annealing) to minimize the energy functional with the possibility for simultaneous optimization of both electronic and ionic variables, and (ii) to determine the time evolution of the system, kept on the Born-Oppenheimer surface, at finite temperature. On the practical side, it can be applied to new classes of problems, especially those for which MD with classical potentials fails to give an appropriate description. For instance, this is the case for covalently bonded systems whenever bonds rearrange as in structural changes and for any quantum-mechanical event such as bond breaking and (re)forming. Also, within the same scheme, one can treat systems in both molecular and condensed phases such that, for instance, six evolution patterns can be properly studied, without having to sacrifice the accuracy of the calculations on passing, for example, from the dimer to the solid state. Starting from the CP proposal of other procedures for local minimizations, a number of new procedures have been suggested for efficient optimization, either simultaneous or sequential, of geometries and electronic structure.

DFT and DFT-MD methods are currently flourishing in chemistry and materials science in general. On the other hand, their application to biochemistry and biology is only at its beginning. In the following section, I shall briefly mention some attempts made so far in this field, and mainly discuss where and why it appears worthwhile to pursue this effort. I shall also try to clarify some technical and more fundamental reasons that limit the range of applications at the moment.

## 2. DFT and DFT-MD in Biochemistry

The reason for introducing DFT in biochemistry are: (i) the need for an *ab initio* description of the interatomic interactions, in many cases, in which force fields become 'fragile', and (ii) its computational efficiency compared to that of post-Hartree-Fock methods. The experience gained so far with the available prescriptions for the exchange-correlation energy functionals in DFT shows a general ability to describe well systems and situations in which metallic, covalent or ionic bonds are dominant. This is the case for a number of ground-state properties and for the energetics. Biological matter, however, is much more complex than that of traditional inorganic and/or organic materials. A reliable description requires the ability to account correctly for other types

of interactions as well. These are hydrogen bonds, as well as van der Waals and hydrophobic interactions, whose role in biology is known to be crucial for both stability and functionality.

Static calculations on key model systems (such as DNA base pairs) ‘in vacuo’ have been performed for some time. The main purpose is generally that of examining the performance of DFT-GGA methods versus post-Hartree-Fock methods [9], in accounting for structure, dipole moments and energetics. A more systematic and critical work would certainly be useful. Regrettably, calculations of this kind often suffer from the limited accuracy of the computational scheme (e.g. small basis sets for the electronic wave functions) and are thus unable to provide reliable reference data for useful comparisons. DFT-based results on simple systems have sometimes been used for the construction of classical potentials to be applied in molecular mechanics calculations of realistic models of biological systems. Although interesting in principle, parametrizations derived in this way have no guarantee of being transferable.

More recently, it has become possible to extend DFT-GGA static calculations to complexes of a few hundred atoms — i.e. the size of biological models investigated in laboratory experiments [10-12]. The starting point is the X-ray refinement of the system (which typically exists in the crystal phase), and the outcome is the locally optimized geometry within DFT-GGA. Therefore, the results for the structure are primarily a test of the computational scheme, although they can also include some features, such as the hydrogen pattern, that are complementary to experimental observations. However, the new information they provide concerns the electronic structure. For instance, in the example reported in reference [11], where a platinum-modified nucleobase pair was studied in the crystal phase, one could capture the effects of metalation on the chemical bonding and show its effects on the electronic structure of the nucleobase pair in the complex.

Static calculations are far from being exhaustive. Still, merely from the point of view of testing of the computational scheme, local energy minimization starting from an educated guess is not sufficient and does not guarantee the absence of spurious minima and/or unrealistic conformations. More important, the study of the dynamics is vital for the understanding of the physical and chemical behavior of biomolecules [13] and is certainly essential for the correct description of the influence of water on their conformations and interactions. This is why the combination of DFT and MD, as achieved with the CP method, has strong potential in this field as well. Moreover, it has recently been shown with CP simulations that liquid water can be treated at a reasonably accurate level within at least one of the DFT-GGA schemes [14]. It is probably worth emphasizing that *ab initio* liquid water constitutes an important step forward in view of the well-known severe difficulties that classical force fields encounter to represent properly its intrinsic properties and specific interactions (see e.g. [15]). CP simulations of water and of some chemical reactions in aqueous solutions [16] indicate that hydrogen bonds can be correctly described with current exchange-correlation functionals in DFT. Still, more experience must be gained to reinforce and generalize such statements.

The ability to represent liquid water with a reasonable degree of accuracy opens up also the possibility of determining ‘structures in solution’ fully *ab initio*. This has

recently been achieved for the monostrand major cisplatin-DNA adduct [17] starting from the structure of the model system in the crystal phase.

Finally, the correct treatment of hydrogen-bonded systems may require an explicit account of the quantum nature of the nuclei. An extension of DFT-MD in this direction that combines it with the path-integral method has recently been introduced [18,19].

Dispersion forces are a bottleneck for current implementations of DFT. The inability of the latter to describe dispersion forces is an obvious consequence of the nature of the approximations made for the exchange-correlation density functionals. Because these approximated functionals are local or at most contain the density gradient (semilocal), they cannot represent the exchange-correlation energy of distant atomic systems with non-overlapping charge distributions [20]. Thus, they generally fail to predict both position and depth of the van der Waals minima, as well as the long-range behavior of the interaction potential. Although there are current attempts to make improvements and/or *ad hoc* corrections within the available schemes (see e.g. [21]), new strategies have recently been proposed within DFT as more reliable and fully *ab initio* approaches [22,23].

Increasing the size and time scales of the system under investigation is an urgent technical problem. In fact, a typical computer model in CP simulations does not contain more than 300 atoms and does not run for more than 10 ps. Solutions to the ‘size problem’ are indeed available for *ab initio* electronic structure calculations (see e.g. [24–26]) where the scaling of the algorithm is linear. These will eventually merge into the combined DFT-MD method [24].

The ‘time problem’ can, in some cases, be reduced by making use of algorithms that are standard in classical MD [27,28]. In the study of activated processes, or more generally of events that are ‘rare’ in the time scale of the computer simulation, one can make use of ‘constrained MD’ [29], namely of MD where specific reaction coordinates are imposed and constrained during the runs. In this way, one can also calculate relative free energies and, in particular, free energy barriers. The implementation in DFT-MD has opened an important new avenue for the study of chemical reactions [30]. In particular, this method has recently been applied to simulate the reaction that is considered to be the first step in the binding of cisplatin to DNA, namely the substitution of one chlorine with a water molecule in the aqueous solution [17]. Clear insights into the reaction path of the reaction, the accompanying electronic mechanisms and the effect of the solvent have been thus obtained, and also a quantitative estimate of the free energy barrier.

Hybrid quantum-classical approaches have long been proposed for simulations in biochemistry, mainly to either include quantum corrections into molecular mechanics calculations (see e.g. [31,32]) or introduce solvation effects into semiempirical quantum models (see e.g. [33]). An extension of the DFT-CP method in this direction would be of great help to overcome the technical limitations mentioned above (see e.g. [34]).

### 3. Conclusion

The application of DFT-based MD to cases of relevance in biochemistry and biology is in its ‘embryonic stage’. Indeed, it has still to acquire a number of features before establishing itself as an accurate and convenient investigation tool in this field.

So far, applications have mainly concerned the geometry and electronic structure of fragments or crystal models of biological systems and, to a lesser extent, the chemistry and dynamics of models of biochemically relevant process (see section 2 above, as well as references [35] and [36]). In particular, owing to its strength and localized nature, the study of metal ion binding to enzymes and nucleic acids suffers from the current limitations of the method to only a minor extent. Also, the role of water is often crucial in determining the mechanism and energy scale of the binding itself. Therefore, this is the area where progress can already be made [12,37,17], both the *ab initio* description and the dynamical approach being of crucial importance.

As discussed above, some technical difficulties must be overcome, and the accuracy of the exchange-correlation energy functionals for the description of all relevant interactions must also be fully established. It should be evident why such an effort is worthwhile. In fact, trying and validating a certain approximation for the universal HK functional will result in a global reliable description of the fundamental interactions, independent of the specific interacting atoms or molecules or aggregation state, independent of the size of the system under investigation and of its physical conditions (e.g. temperature, pressure). This is drastically different from trying and validating an empirical or a semiempirical approach, such as a classical potential or a quantum-mechanical model. In such cases, one has to (est a specific parametrization that can be suitable only for a subset of problems and systems, and often has little chance of being predictive.

In conclusion, DFT-based MD, assisted by intelligent modelling, has all the prerequisites to render computer simulations useful not only for the investigation of biochemical systems, but also for the design of new biologically relevant materials.

## Acknowledgements

I wish to thank Paolo Carloni and Alessandro Curioni for useful discussions.

## References

1. Dreizler, R.M. and Gross, E.K.U., *Density-functional theory*, Springer-Verlag, Berlin, 1990
2. Parr, R.G. and Yang, W., *Density-functional theory of atoms and molecules*, Oxford Science Publications., New York, 1989.
3. Hohenberg, P. and Kohn, W., *Inhomogeneous electron gas*, Phys. Rev., 136 (1964)B864–1887.
4. See also Lieb, E.H., *Density functionals for Coulomb systems*, In Dreizler, R.M. and Providencia. J. (Eds.) *Density functional methods in physics*. Plenum, New York, 1985, pp. 31-80; Levy, M. and Perdew, J.P., *The constrained search formulation of density functional theory*, *ibid.*, pp. 1-30.
5. Kohn, W. and Sham, L.J., *Self-consistent equations including exchange and correlation effects*, Phys. Rev., 140 (1965) A1133–A1138.
6. Gunnarsson, O., Jonson, M. and Lundqvist, B.I., *Description of exchange and correlation effects in inhomogeneous electron systems*, Phys. Rev. B, 20 (1979) 3136–3164.
7. Car. R. and Parrinello, M., *Unified density-functional theory and molecular dynamics*, Phys. Rev. Lett., 55 (1985) 2471–2474.
8. Car, R., *Molecular dynamics from first principles*, In Binder, K. and Ciccotti G. (Eds.) *Monte Carlo and molecular dynamics of condensed matter systems*. Italian Physical Society Publications. Bologna, Italy, 1995. pp. 601–634.

9. See e.g. Sponer, J., Leszczynski, J. and Hobza, P., *Structures and energies of hydrogen-bonded DNA base pairs: A nonempirical study with inclusion of electron correlation*, J. Phys. Chem., 100 (1996) 1965–1974.
10. Hutter, J., Carloni, P. and Parrinello, M., *Non-empirical calculations of a hydrated RNA duplex*, J. Am. Chem. Soc., 118 (1996) 8710–8712.
11. Carloni, P. and Andreoni, W., *Platinum-modified nucleobase pairs in the solid state: theoretical study*, J. Phys. Chem., 100 (1996) 17797–17800.
12. Carloni, P. and Alber, F., *Density-functional theory investigations of enzyme-substrate interactions*, this volume and references therein.
13. Karplus, M. and Petsko, G.A., *Molecular dynamics simulations in biology*, Nature. 347 (1990) 631–639.
14. Sprik, M., Hutter, J. and Parrinello, M., *Ab initio molecular dynamics simulation of liquid water: comparison of three gradient-corrected density functional*, J Chem. Phys., 105 (1996) 1142–1152.
15. Beveridge, D.L., Swaminathan, S., Ravishanker, G., Withka, J.M., Srinivasan, J., Prevost, C., Louise-May, S., Langley, D.R., DiCapua, F.M. and Bolton, P.H., *Molecular dynamics simulations on the hydration, structure and motions of DNA oligomers*, In Westhof, E. (Ed.) Water and biological macromolecules. Macmillan, London, U.K., 1993, pp. 165–225.
16. See e.g. Mejer, E.J and Sprik, M., *A density-functional study of the addition of water to SO<sub>3</sub> in the gas phase and in aqueous solution*, J. Phys. Chem. (in press).
17. Carloni, P., Sprik, M. and Andreoni, W., *Cisplatin-DNA binding mechanism: Key steps from ab initio molecular dynamics* (in preparation).
18. Marx, D. and Parrinello, M., *Ab-initio path integral molecular dynamics Basic ideas*, J. Chem. Phys., 104 (1996) 4077–4082.
19. Tuckerman, M.E., Marx, D., Klein, M.L. and Parrinello, M., *On the quantum nature of the shared proton in hydrogen bonds*. Science. 275 (1997) 817–819, and references therein.
20. Kristyán, S. and Pulay, P., *Can (semi) local density functional theory account for the London dispersion forces?* Chem. Phys. Lett., 229 (1994) 175–180.
21. Osinga, V.P., van Gisbergen, S.J.A. and Baerends, E.J., *Density functional results for isotropic and anisotropic multipole polarizabilities and C<sub>6</sub>, C<sub>7</sub> and C<sub>8</sub> van der Waals dispersion coefficients for molecules*, J. Chem. Phys. 106 (1997) 5091.
22. Kohn, W. and Meir, Y., *Van der Waals energies in density functional theory*, Phys. Rev. Lett. (submitted)
23. Cross, E.K.U., Dobson, J.F. and Petersilka, M., *Density functional theory time-dependent phenomena*, In Nalewajski, R.F. (Ed.) Density functional theory. Topics in Current Chemistry. Vol. 181. Springer, Heidelberg. 1996, pp. 81–172.
24. Mauri, F. and Galli, G., *Electronic structure calculations and molecular dynamics simulations with linear system-size scaling*, Phys. Rev. B, 50 (1994) 4316–4326.
25. Carlsson A.E., *Order-N density-matrix electronic-structure method for general potentials*, Phys. Rev. B, 51 (1995) 13935–13941.
26. Kohn, W., *Density functional and density matrix methods scaling linearly with the number of atoms*, Phys. Rev. Lett., 76 (1996) 3168–3171.
27. van Gunsteren, W.F., *Molecular dynamics and stochastic dynamics simulations: A primer*, In van Gunsteren, W.F., Weiner, P.K., Wilkinson, A.J. (Eds.) Computer simulation of biomolecular systems, Vol. 2, ESCOM, Leiden, The Netherlands. 1993, pp. 3–36.
28. Tuckerman, M.E. and Parrinello, M., *Integrating the Cur-Parrinello equations II: Multiple time scale techniques*, J. Chem. Phys., 101 (1994) 1316–1329.
29. Carter, E.A., Ciccotti, G. and Hynes, J.T., *Constrained reaction coordinate dynamics for the simulation of rare events*, Chem. Phys. Lett., 156 (1989) 472–477; Ciccotti, G., Ferrario, M. and Hynes, J.T., *Constrained molecular dynamics and the mean potential for an ion pair in a polar solvent*, Chem. Phys. 129 (1989) 241–251.
30. Curioni, A., Sprik, M., Andreoni, W., Schiffer, H., Hutier, J. and Parrinello, M., *Density-functional-theory based molecular dynamics simulation of acid catalyzed reactions in liquid trioxane*, J. Am. Chem. Soc. 119 (1997) 7218.

31. Perákylá, M. and Kollman, PA., *A simulation of the catalytic mechanism of aspartyl-glucosaminidase using ab-initio quantum mechanics and molecular dynamics*, J. Am. Chem. Soc., 119 (1997) 1189–1196.
32. Stanton, R.V., Hartsough, D.S. and Merz, K.M., Jr., *An examination of a densityfunctional/molecular mechanical coupled potential*, J. Comput. Chem., 16 (1995) 113–128.
33. See e.g. Cramer, C.J. and Truhlar, D.G., *Molecular orbital theory calculations of aqueous solvation effects in chemical equilibria*, J. Am. Chem. Soc., 113 (1991) 8552–8554; Giesen, D.J., Gu, M Z., and Truhlar, D.G., *A universal organic solvation model*, J. Org. Chem., 61 (1996) 8720–8721.
34. Wei, D. and Salahub, D.R., *A combined density functional and molecular dynamics simulation of a quantum water molecule in aqueous solution*, Chem. Phys. Lett., 224 (1991) 291–296.
35. Buda, F., deGroot, H. and Bifone, A., *Charge localization and dynamics in rhodopsin*, Phys. Rev. Lett., 71 (1996) 4474–4477.
36. Rovira, C., Ballone, P. and Parrinello, M., *A density functional study of iron–porphyrin complexes*, Chem. Phys. Lett., 271 (1997) 247–250.
37. Sagnella, D.E., Laasonen, K. and Klein, M.L., *Ab initio molecular dynamics study of proton transfer in a polyglycine analog of the ion channel gramicidin A*, Biophys. J., 71 (1996) 1172–1178.



**This Page Intentionally Left Blank**

# Density-Functional Theory Investigations of Enzyme-substrate Interactions

Paolo Carloni<sup>a</sup> and Frank Alber<sup>b</sup>

<sup>a</sup> IBM Research Division, Zurich Research Laboratory, Säumerstrasse 4, CH-8803 Rüschlikon, Switzerland and Department of Chemistry, University of Florence, Via G. Capponi 7, I-50121 Firenze Italy

<sup>b</sup> Department of Pharmacy Swiss Federal Institute of Technology (ETH), Winterthurerstrasse 190, CH-8057 Zurich, Switzerland.

## 1. Introduction

Density-functional theory (DFT) [1,2], originally developed in solid-state physics, is a valuable, versatile and efficient quantum-mechanical method for electronic structure calculations of chemical systems [3]. Recently, application of the DFT methods has also been extended to biological molecules [3,4].

Here, we present some results from recent investigations on two important enzymes. superoxide dismutase and thymidine kinase. Our goal is to understand the factors that play a crucial role in the enzyme-substrate interactions. We also briefly discuss the importance of solvation effects on the quantum-mechanical calculations [5].

## 2. Superoxide dismutase

Copper-zinc superoxide dismutase (SOD) is a dimeric enzyme containing a copper ion and a zinc ion in the active site of each subunit. The active site of the enzyme consists of the copper ion (which is essential to the SOD activity) coordinated by four histidines in a distorted square planar geometry. One of these residues, His61, acts as a bridge between the Cu and Zn ions. An Asp and two other His residues complete that co-

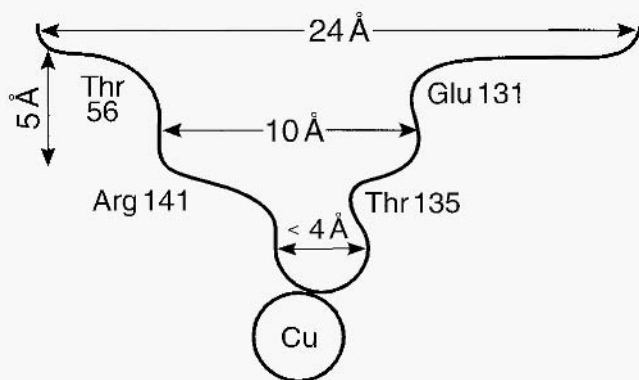


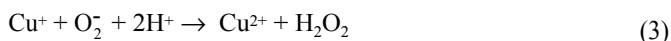
Fig. 1. Schematic view of Cu, Zn SOD active site and active site channel. (From Getzoff, E.D, Halwell, R.A. and Trainer J.A., In Inouye. M. (Ed.) *Protein Engineering: Applications in Science Industry and Medicine*, Academic Press, New York, 1986, pp. 41-69.)

ordination of Zn. The X-ray structure of the oxidized [6] and reduced [7] bovine SOD reveals that the copper–histidine complex is located at the bottom of a shallow cavity present in the protein (Fig. 1).

The physiological role of SOD is the removal of the harmful superoxide anion radical,  $O_2^-$ , produced during the oxygen metabolic cycle [8–13]. SOD dismutates  $O_2^-$  to molecular oxygen and hydrogen peroxide at a very high rate ( $2\text{--}3 \times 10^9 \text{ s}^{-1} \text{ M}^{-1}$ ; references [9,14]):

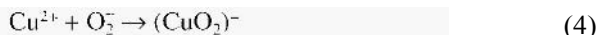


Two mechanisms have been proposed. In the most widely accepted mechanism [8], the copper ion is repeatedly reduced and oxidized in a two-step process:

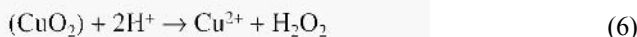


Within this scheme, the Cu–His61 bond breaks down in the first step [15]; subsequently, His61 accepts a proton from the solvent [11,16]. In the second step, this proton is transferred to a second superoxide molecule forming  $HO_2^-$ , and the Cu–His61 bond is reformed.

A second mechanism, which involves the formation of a stable copper-superoxide intermediate has been suggested on the basis of *ab initio* quantum-mechanical calculations [17–19]:



The  $(CuO_2)$  intermediate would then oxidize a second superoxide molecule:



Depending on whether an electron transfer (ET) occurs between copper(II) and superoxide (reactions 2 and 4), the two different mechanisms can be operative [20].

With the aim of understanding some of the important factors governing the superoxide–copper ET, we have undertaken DFT calculations on several models of SOD active site and its adduct with superoxide [21].

### 2.1 Computational procedure

A copper tetraimidazole complex, whose geometry has been taken from the X-ray structure of oxidized SOD [6], is our model for the active site of SOD (Fig. 2). Arg141, the only invariant residue of SOD [22], has been included in the model by modelling it as an ammoniumion. The mode of binding of  $Q_5^-$  to SOD is also shown. We have performed single-point DFT calculations on various complexes, within the local spin density approximation of DFT [1,2]. We use the LDA parameterization by Perdew and Zunger [23], which is based on Monte Carlo simulations of the free electron gas investigated by Ceperley and Alder [24]. Some calculations were done using spin-polarized techniques.

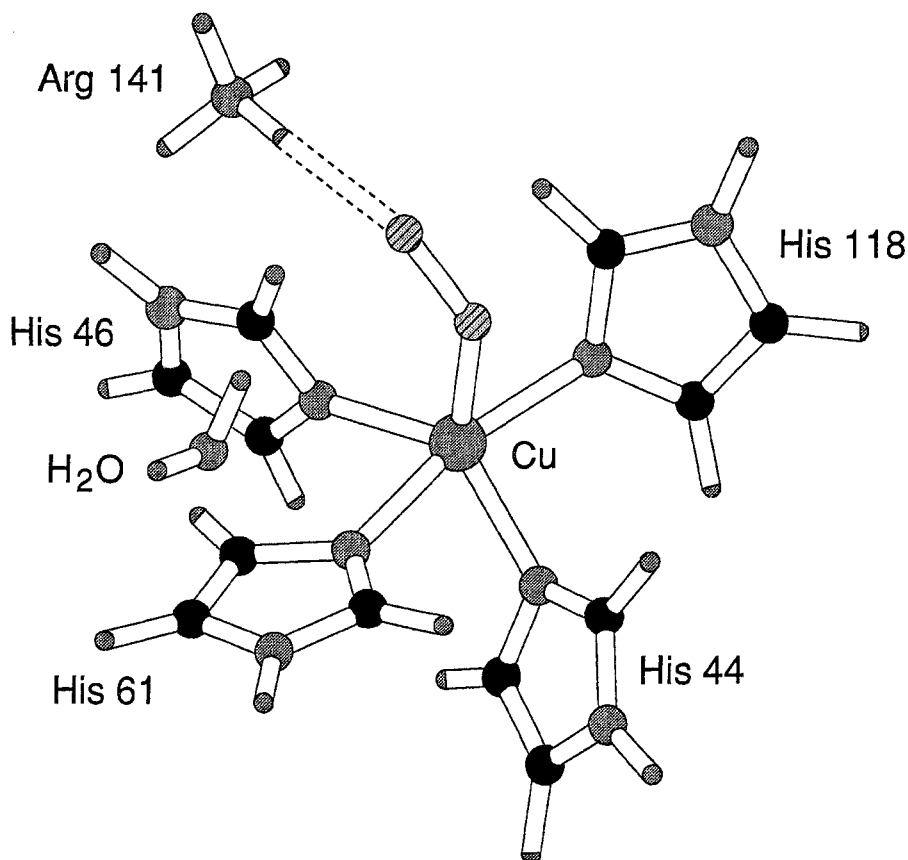


Fig. 2. Model of the SOD active site-superoxide adduct. (Reprinted with permission from Reference [21], © 1995 The American Chemical Society.)

Our basis set consists of plane waves (for a detailed description the reader is referred to reference [21]). The electronic structure of the complexes are presented in terms of density of states (DOS).

## 2.2. Results and discussion

Figure 3a shows the DOS of the oxidized SOD active site — i.e. of the copper(II)-imidazole complex. The main contributions come from the imidazole ligands. The contributions of the other molecules present in the complexes are also shown. The four imidazole nitrogen atoms bound to copper produce a square planar distorted ligand field: the copper  $d_{xy}$ ,  $d_{xz}$  and  $d_{yz}$  levels form a large peak. Two small shoulders on the main peak are due to the copper  $d_{z^2}$  orbital. At higher and lower energies lie the antibonding

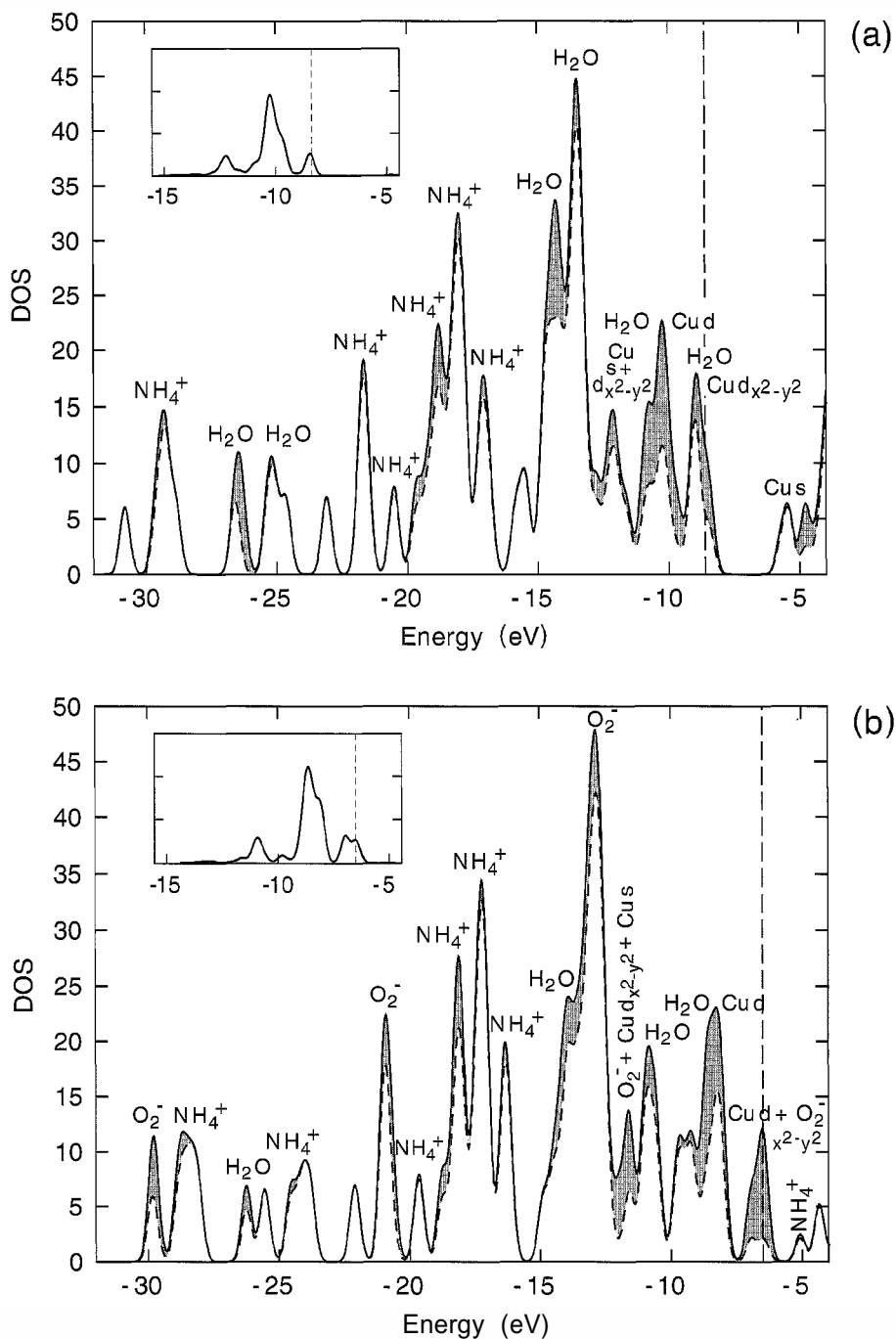


Fig. 3. DOS of the SOD active site (a) and its adduct with superoxide (b). At the top left, the Cu d-orbital splittings are also shown. (Reprinted in part, with permission from reference [21], © 1995 The American Chemical Society.)

( $\sigma^*$ ) and bonding ( $\sigma$ ) molecular orbitals formed by the copper  $d_{x^2-y^2}$  with the ligands. The bonding–antibonding splitting of the  $d_{x^2-y^2}$  derived states is 5.3 eV [25].

When the superoxide is brought into close proximity of the copper ion (Fig. 3b), a new peak of the DOS at the Fermi level is present: this peak is due to a molecular orbital formed by the  $d_{x^2-y^2}$  and the two  $\pi^*$  orbitals of superoxide. A stable Cu(II)–O $_2^-$  complex is formed, and an electron is partially transferred from the superoxide to the copper(II) ion. Consistently with the stability of the Cu(II)–O $_2^-$  complex, the binding energy against dissociation into a neutral oxygen molecule and the reduced SOD active site is found to be 99 kJ/mol [25]. The reason why the  $d_{x^2-y^2}$  orbital is substantially higher in energy here than in the oxidized SOD active site is the large electron Coulomb repulsion of the copper ion.

We now discuss the effect of an important residue in the active site of the enzyme, Arg141, which has been included in the calculations. Because of its mobility [26], Arg141 could H-bond strongly to the terminal oxygen of the substrate and hence stabilize the Cu–O $_2^-$  intermediate, favoring the mechanisms 4 and 5. By positioning the guanidinium group of Arg141 closer to the substrate — i.e. by reducing the guanidinium–superoxide H-bond from the 2.5 Å in the previous complex to 1.5 Å — we find that the electronic structure essentially does not change. The binding energy is also sizeably increased by 15 kJ/mol.

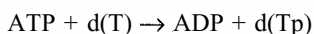
In conclusion, our calculations show that partial electron transfer process between SOD and its substrate occurs through the copper  $d_{x^2-y^2}$  and the  $\pi^*$  orbitals of the superoxide. The electronic structure of the enzyme substrate complex is mainly influenced by Coulomb repulsion and, to a lesser extent, by spin-polarization effects of the paramagnetic Cu(II) ion and of superoxide  $\pi^*$  orbitals [25]. The covalent hybridization between the Cu- $d$  and the superoxide  $\pi^*$  orbitals is, in comparison, negligible.

The electronic structure at the Fermi level changes only slightly upon reducing the Arg141–superoxide distance. Other levels are shifted in energy by the long-range Coulomb interaction with the Arg guanidinium, whose position has been modified.

Our calculations essentially agree with earlier quantum-mechanical calculations [17], in that the ET between superoxide and copper does not take place *in vacuo*. However, unlike the conclusion drawn in [17], we believe that our and previous quantum-mechanical calculations do not necessarily imply a mechanism described by reactions 4–6; the presence of the electrostatic field of the protein/water environment might alter this picture, and the Cu–O $_2^-$  complex, which *in vacuo* is stable, could easily dissociate if the effect of the protein is taken into account, consistently with the two-step mechanisms 2–3.

### 3. Herpes Simplex Type 1 Thymidine Kinase

Viral herpes simplex type 1 thymidine kinase (HSV1 TK) is a key enzyme in the metabolism of the herpes simplex virus. Its physiological role is to salvage thymidine into the DNA metabolism by converting it to thymidine monophosphate: the phosphorylation is achieved by transfer of the  $\gamma$ -phosphate group from ATP to the 5'-OH group of thymidine:



Understanding the biochemistry of this enzyme is important for application in the treatment of virus infections and for cancer chemotherapy [27–34].

We present here our recent DFT study that has focused on the HSV 1 TK nucleoside interactions [35]. Our goal is to gain a better understanding of the nature of HSV1 TK binding interactions and of its mechanism of action.

### 3.1. Computational procedure

HSV1 TK is a dimeric enzyme with 376 residues per subunit. The two subunits are related by  $C_2$  symmetry (Fig. 4). The active site is formed by an ATP and a nucleoside binding region [36]. Our modelling of HSV1 TK active site (Fig. 5) includes the residues fixing the thymine ring (Met128 and Tyr172); the guanidinium group of Arg163, modelled again by an ammonium ion, is also included because of its important electrostatic role. Several HSV1 TK-thymine complexes have been considered by protonating the residues and the substrate differently. Calculations were carried out in the DFT framework, using the Becke-Lee-Yang-Parr approximation for the exchange-correlation functionals [37,38]. The interaction between valence electrons and ionic cores is described by pseudopotentials of the Martins-Troullier type [39]. The Kohn-Sham orbitals are expanded in plane waves up to an energy cutoff of 70 Ry. The complex geometries were fully optimized. Position constraint was imposed for the  $C\beta$  of Tyr 172 and Met 128 and on the ammonium nitrogen.

### 3.2. Results and discussion

The difference density  $\Delta\rho = \rho_{\text{complex}} - \rho_{\text{fragments}} - \rho_{\text{substrate}}$  describes how the electron density  $\rho$  changes during the formation of the complex. Inspection of  $\Delta\rho$  for all complexes reveals that no charge transfer from or to the substrate is present. The O and N atoms of thymine as well as the Arg 163–Tyr172 H-bond are sizeably polarized. Interestingly, there is no evidence of polarization of the Met128 sulfur atom [40]. This indicates that sulfur plays only a minor role in binding. That the role of Met128 sulfur in the binding process is purely hydrophobic and steric has been confirmed by very recent site-directed mutagenesis experiments, which have shown that the activity is preserved when the Met residue is replaced by another hydrophobic residue such as Ile [41].

The complexes in which Tyr172 is deprotonated and thymine protonated are more stable by at least 130 kJ/mol, in terms of the binding energy, than those bearing a neutral substrate and/or neutral tyrosine. The stabilization is due to the favorable and strong Coulomb interaction between the ionic pair. By including the electrostatic environment using a standard force field [42], the energetics obtained using the quantum-mechanical calculations turn out not to be significantly affected.

We conclude that the enzyme–substrate binding is dominated by Coulombic interactions, with no sizeable charge transfer to or from the substrate. The calculations suggest the existence of a ‘proton transfer complex’ between Tyr 172 and the substrate. Such a complex could be important for the efficiency of the enzyme by enhancing the rigidity of the active site.

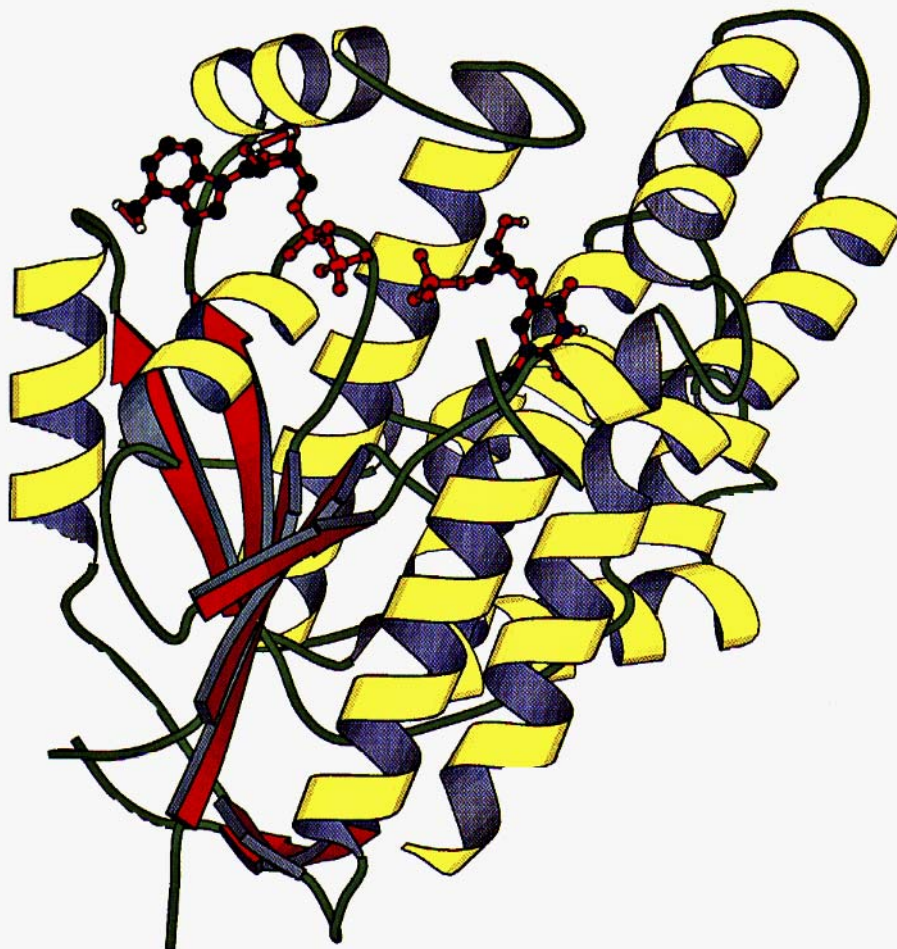


Fig. 4. Structure of one subunit of HSV1 TK.

#### 4. Conclusion

The two examples shown here illustrate that the DFT approach can be used to investigate enzyme-substrate interactions. The DFT method appears to be well suited to treat a variety of systems and chemical bonding, from a metal-based enzyme to a protein containing a 'p-complex' between two aromatic rings. These examples also indicate that, in general, the effects of the protein/water environment should be considered to obtain a more realistic description of the interactions at the active site [5]. The development of coupled *ab initio* classical molecular dynamics simulations methods [43] will further improve the theoretical investigations of the mechanism of action of enzymes, by allowing the study of the reactive processes in the presence of the dynamical effects of the environment.



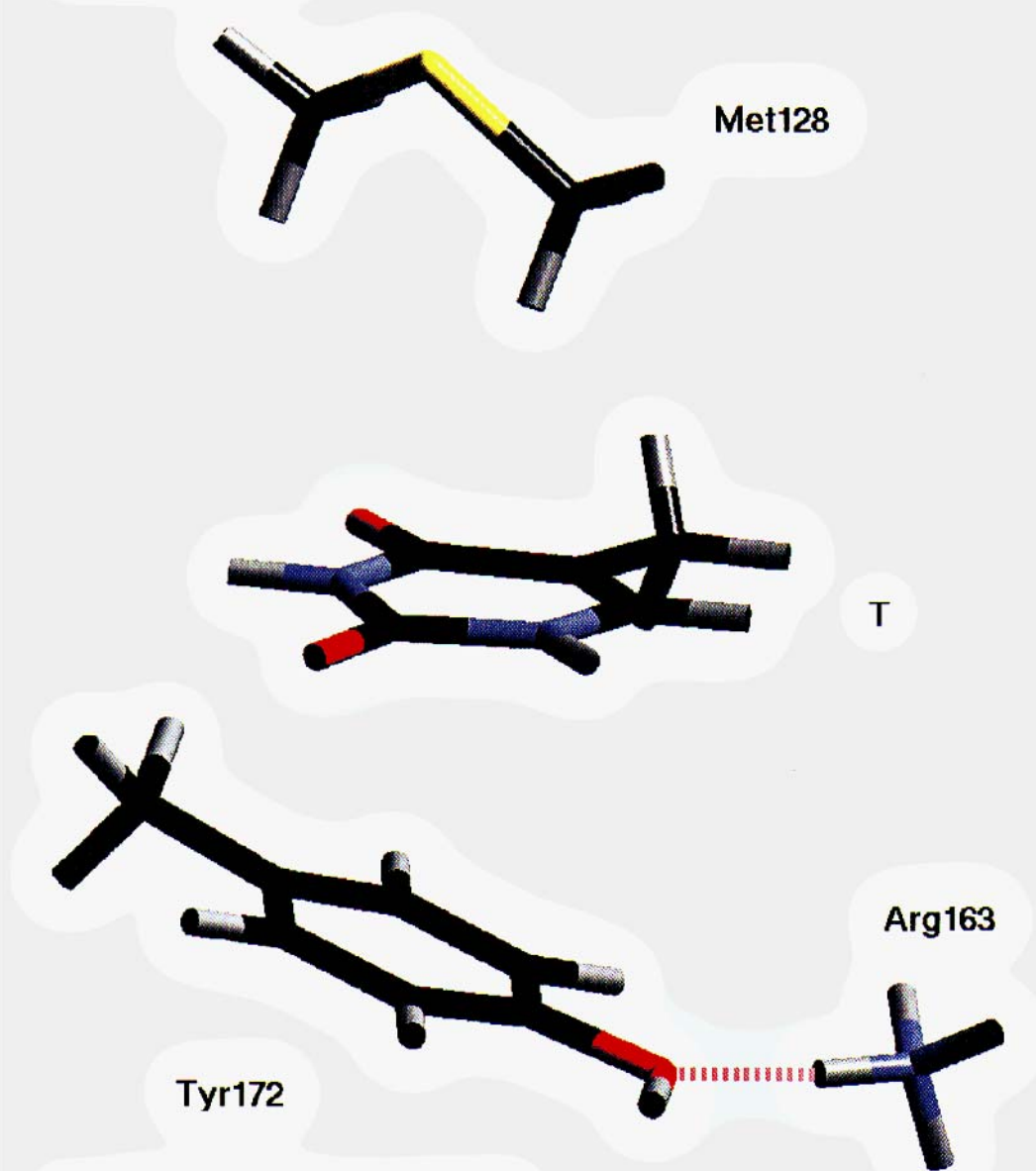


Fig. 5. Model of the HSV1 TK active site-tyamine adduct

### Acknowledgements

We thank Wanda Andreoni for her valuable suggestions.

## References

1. Hohenberg, P.C. and Kohn, W., *Inhomogeneous electron gas*, Phys. Rev., 136 (1964) B864–B887.
2. Kohn, W. and Sham, L.J., *Self-consistent equations including exchange and correlation effects*, Phys. Rev., 140 (1965) A1133–A1138.
3. For a detailed discussion of the DFT methods and applications in biochemistry, the reader is referred to Andreoni, W., *Density-functional theory and molecular dynamics: A new perspective for simulations of biological systems*, this volume.
4. See, e.g. (a) Li, Yan and Evans, J.N.S., *The hard-soft acid-base principle in enzymatic catalysis: Dual reactivity of phosphoenolpyruvate*, Proc. Natl. Acad. Sci., 93 (1996) 4612–4616 (b) Hütter, J., Carloni, P. and Parrinello, M. *Non-empirical calculations on a hydrated RNA duplex*, J. Am. Chem. Soc., 118 (1996) 8710–8712; (c) Carloni, P. and Andreoni, W., *Platinum-modified nucleobasepairs in the solid state: A theoretical study*, J. Phys. Chem., 100 (1996) 17797–17800; (d) Bernardi, F., Bottoni, A., Casadio, R., FariSELLI, P. and Rigo, A., *Ab initio study of the dioxygen binding site of hemocyanin: A comparison between CASSCF, CASPT2, and DFT approaches.*, Int. J. Quant. Chem., 58 (1996) 109–119; (e) Sagnella, D.E., Laasonen, K. and Klein, M.L., *Ab initio molecular dynamics study of proton transfer in a polyglycine analog of the ion channel gramicidin A.*, Biophys. J., 71 (1996) 1172–1178; (f) Oldziej, S. and Ciarkowski, J., *Mechanism of action of aspartic proteinases: Application of transition-state analogue theory.*, J. Comput.-Aided Mol. Design, 10 (1996) 583–588; and (g) Bajorath, J., Kraut, J., Li, Z.Q., Kitsoon, D.H. and Hagler, A.T., *Theoretical studies on the dihydrofolate reductase mechanism: Electronic polarization of bound substrates*, Proc. Natl. Acad. Sci. USA, 88 (1991) 6423–6426.
5. York, D.M., Lee, T.S. and Yang, W., *Quantum mechanical study of aqueous polarization effects on biological macromolecules*, J. Am. Chem. Soc., 118 (1996) 10940–10941, and references therein.
6. Tainer, J.A., Getzoff, E.D., Beem, K.M., Richardson, J.S. and Richardson, D.C., *Determination and analysis of the 2 Å structure of copper, zinc superoxide dismutase*, J. Mol. Biol., 160 (1982) 181–217.
7. Banci, L., Bertini, I., Bruni, B., Carloni, P., Luchinat, C., Mangani, S., Orioli, P.L., Piccioli, M., Ripnieski, W. and Wilson, K.S., *X-ray, NMR and molecular dynamics studies of reduced bovine superoxide dismutase: Implications for the mechanism*, Biochem. Biophys. Res. Commun., 202 (1994) 1088–1095.
8. McCord, J.M. and Fridovich, I., *Superoxide dismutase: An enzymatic function for an erythrocyte protein (hemocyanin)*, J. Biol. Chem., 244 (1969) 6049–6055.
9. Fee, J.A. and Bull, C., *Steady state kinetic studies of superoxide dismutases*, J. Biol. Chem., 261 (1986) 13000–13005.
10. Fee, J.A. and Di Corleto, P.E., *Observation on the oxidation-reduction properties of bovine erythrocyte superoxide dismutase*, Biochem., 12 (1973) 4893–4899.
11. Tainer, J.A., Getzoff, E.D., Richardson, J.S. and Richardson, D.C., *Structure and mechanism of copper, zinc superoxide dismutase*, Nature (London), 306 (1983) 284–287.
12. Valentine, J.S., Pantoliano, M.W., McDonnell, P.J., Burger, A.R. and Lippard, S.J., *pH dependent migration of copper (II) to the vacant zinc-binding site of zinc-free bovine erythrocyte superoxide dismutase*, Proc. Natl. Acad. Sci. USA, 76 (1979) 4245–4249.
13. Banci, L., Bertini, I., Luchinat, C. and Piccioli, M., *Spectroscopic studies on Cu<sub>2</sub>Zn<sub>2</sub>SOD: A continuous advancement of investigation tools*, Coord. Chem. Rev., 100 (1990) 67–103.
14. Fielden, E.M., Roberts, P.B., Bray, R.C., Lowe, D.J., Mautner, G.N., Rotilio, G. and Calabrese, L., *The mechanism of action of superoxide dismutase from pulse radiolysis and electron paramagnetic resonance*, Biochem. J., 139 (1974) 49–60.
15. McAdam, M.E., *A pulse-radiolysis study of the manganese-containing superoxide dismutase from bacillus stearothermophilus*, Biochem. J., 165 (1977) 71–79, and references therein.
16. Morpurgo, L., Giovagnoli, C. and Rotilio, G., *Studies of the metal sites of copper proteins. A model compound for the copper site of superoxide dismutase*, Biochim. Biophys. Acta, 322 (1973) 204–210.
17. Osman, R. and Bash, H., *On the mechanism of action of superoxide dismutase: A theoretical study*, J. Am. Chem. Soc., 106 (1984) 5710–5714.

18. Rosi, M., Sgamellotti, A., Tarantelli, F., Bertini, I and Luchinat, C., *A theoretical investigation of the copper-superoxide system: A model for the mechanism of copper-zinc superoxide dismutase*, Inorg. Chim. Acta, 107 (1985) L21–L22.
19. Rosi, M., Sgamellotti, A., Tarantelli, F., Bertini, I. and Luchinat, C., *Ab initio calculations of the  $Cu^{2+}-O_2^-$  interactions as a model for the mechanism of copper-zinc superoxide dismutase*, Inorg. Chem., 25 (1986) 1005–1008.
20. Banci et al. (reference [7]) have recently proposed that at low physiological substrate concentration the two-step mechanism would take place, whereas at higher concentrations the second mechanism would take place.
21. Carloni, P., Blöchl, P.E. and Paninello, M., *Electronic structure of the Cu, Zn superoxide dismutase active site and its interactions with the substrate*, J. Phys. Chem., 99 (1995) 1338–1348.
22. Getzoff, E.D., Tainer, J.A., Stempien, M.M., Bell, G.I. and Hallewell, R.A., *Evolution of CuZn superoxide dismutases and the Greek-key  $\beta$ -barrel structural motif*, Proteins: Struct. Funct. Gen., 5 (1989) 322–366.
23. Perdew, J.P. and Zunger, A., *Self-interaction correction to density-functional approximations for many-electron systems*, Phys. Rev. B, 23 (1981) 5048–5079.
24. Ceperley, M. and Alder, B.L., *Ground state of the electron gas by a stochastic method*, Phys. Rev. Lett., 45 (1980) 566–569.
25. To estimate spin-polarization effects, spin-polarized calculations were also carried out. For a detailed description of the spin-polarized effects, see reference [21].
26. Banci, L., Carloni, P., La Penna, G. and Orioli, P., *Molecular dynamics studies on superoxide dismutase and its mutants: The structure and functional role of Arg 143*, J. Am. Chem. Soc., 114 (1992) 6994–7001.
27. Elion, G.B., Furman, P.A., Fyfe, J.A., De Miranda, P., Beauchamp, C. and Schaeffer, H.J., *Selectivity of action of an antiherpetic agent, 9-(2-hydroxymethyl)guanine* Proc. Natl. Acad. Sci., 74 (1977) 5716–5720.
28. Schaeffer, H.J., Beauchamp, C., De Miranda, P., Elion, G.B., Bauer, D.I. and Collins, P., *9-(2-hydroxymethyl)guanine activity against viruses of the herpes group*, Nature, 212 (1978) 583–585.
29. Culver, K.W., Ram, V., Wallbridge, S., Ishii, H., Oldfield, E.H. and Blaese, R.M., *In vivo gene transfer with retroviral vector-producer cells for the treatment of experimental brain tumors*, Science, 256 (1992) 1550–1552.
30. Chen, S.-H., Shine, H.D., Goodman, J.C., Grossman, R.G. and Woo, S.L.C., *Gene therapy for brain tumors: regression of experimental gliomas by adenovirus-mediated gene transfer in vivo*, Proc. Natl. Acad. Sci., 91 (1994) 3054–3057.
31. O'Malley, B.W., Jr., Chen, S.-H., Schwartz, M.R. and Woo, S.L.C., *Adenovirus mediated gene therapy for human head and neck squamous cell cancer in a nude mouse model*, Cancer Res., 55 (1995) 1080–1085.
32. Chambers, R., Gillespie, G.Y., Soroceanu, L., Andreansky, S., Chatterjee, S., Chou, J., Roizman, B. and Whitely, R.J., *Comparison of genetically engineered herpes simplex viruses for treatment of brain tumors in a scid mouse model of human malignant glioma*, Proc. Natl. Acad. Sci., 92 (1995) 1411–1415.
33. Vile, R.G. and Hart, I.R., *Use of tissue-specific expression of the herpes simplex thymidine kinase gene to inhibit growth of established murine melanomas following direct intratumoral injection of DNA*, Cancer Res., 53 (1993) 3860–3864.
34. Caruso, M., Panis, Y., Gagandeep, S., Houssin, D., Salzmann, J.-L. and Klatzmann, D., *Regression of established macroscopic liver metastases after in situ transduction of a suicide gene*, Proc. Natl. Acad. Sci., 90 (1993) 7024–7028.
35. Alber, F., Kuonen, O., Scapozza, L., Folkers, G. and Carloni, P., *Density functional studies on herpes simplex type I thymidine kinase–substrate interactions: the role of Tyr 172 and Met 128 in thymine fixation*, Proteins: Struct. Funct. Gen., in press.
36. Wild, K., Bohner, T., Aubry, A., Folkers, G. and Schultz, G.E., *The three-dimensional structure of thymidine kinase of herpes simplex type I*, FEBS Lett., 369 (1995) 289–292.
37. Becke, A., *Density-functional exchange-energy approximation with correct asymptotic behavior*, Phys. Rev. A, 38 (1988) 3098–3100.

38. Lee, C., Yang, W. and Parr, R.G., *Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density*, Phys. Rev. B. 37 (1988) 785–789.
39. Troullier, N. and Martins, J.L., *Efficient pseudopotentials for plane-wave calculations*, Phys. Rev. B, 43 (1991) 1993–2006.
40. The sulfur atom of Met128 is 4.8 Å away from the thymine ring (see Fig. 1). Therefore, it should in principle be possible to find sizable polarization effects on the sulfur.
41. Folkers, G., Pilger, B., Alber, F. and Scapozza, L., (in preparation).
42. We have used the parameterization of the GROMOS96 force field: van Gusteren, W.F., Billeter, S.R., Eising, A.A., Hünenberger, P.H., Krüger, P., Mark, A.E., Scott, W.R.P. and Tironi, I.G., *Biomolecular simulation: The GROMOS96 manual and user guide*, BIOMOS B.V., Biomolecular Software, Laboratory of Physical Chemistry, ETH Zentrum, CH-8092 Zurich, Switzerland, 1996.
43. Car, R. and Parrinello, M., *Unified approach for molecular dynamics and density-functional theory*, Phys. Rev. Lett. 55 (1985) 2471–2474.

**This Page Intentionally Left Blank**

# Molecular Dynamics Simulations: A Tool for Drug Design

Didier Rognan

Department of Pharmacy Swiss Federal Institute of Technology CH-8057 Zürich, Switzerland

## 1. Introduction

The recent growth of recombinant DNA technology (cDNA cloning, southern blotting, PCR) [1] has boosted, in the last decade, the identification of biological macromolecules and their expression in purity and quantity adequate for structure determination. About 5000 coordinate entries are today available in the Brookhaven Protein Data Bank [2,3] and more than 10 000 are expected at the turn of the century. Experimentally determined protein structures represent by far the most promising starting point for rational drug design. There is, therefore, a need for filling the gap between 3D structures of biological targets and potential substrate/inhibitors of those macromolecules. Computational chemistry tries to rationalize this gap.

In parallel with the development of modern molecular biology, computational chemistry methods have also dramatically evolved [4–6]. The early static pictures of pharmacophores [7] or, in the best cases, of protein–ligand complexes [8] have now been supplemented by dynamic representations of drug–receptor interactions [9]. A major breakthrough came in the late 1970s from the application of molecular dynamics (MD) [10], initially developed for fluids [11], to biological macromolecules [12]. Basically, MD can be described by the numerical solution of Newton's second law of motion (Eq. 1):

$$m_i \frac{d^2 r_i}{dt^2} = -\nabla_i [V(r_1, r_2, \dots, r_N)] \quad i = 1, N \quad (1)$$

For a molecular system of  $N$  particles having a mass  $m_i$ , atomic positions  $r_i$  at a time  $t$ , are derived from the gradient of the potential energy  $V$ , which is classically obtained by molecular mechanics. As  $r_i$  is calculated within short time steps  $\Delta t$  of 1–2 fs (Eq. 2), the time history or trajectory of all atoms may be easily monitored and thus give access to dynamic properties that may be of interest for studying protein structure, folding, catalytic mechanisms and binding processes:

$$r_i(t + \Delta t) = 2r_i(t) - r_i(t - \Delta t) - \Delta t^2 \nabla_i V(t) / m_i \quad (2)$$

Early simulations were limited to simplified molecular representations (*in vacuo*) and very short time scales (a few ps) [13–14]. However, the combined development of supercomputer technology, algorithm parallelization [15–16], force fields [17–20] and time-saving techniques [21] permits nowadays the simulation of complex biological systems for longer time scales (up to a few ns) [22–23], and with higher accuracy [18,24].

It is not our aim in this chapter to review methodological advances [25] or potential applications of MD to biology [26]. As both the methodology and input 3D structures

are becoming accessible to the great majority of the scientific community, we will rather emphasize new aspects in the use of MD simulations: (i) as a tool for the rationalization of structure–activity relationships, and (ii) as a promising method for protein structure-based drug design.

## 2. Dynamics of Free Ligands: Efficient Sampling of Conformational Space

The most popular application of MD simulations is conformational sampling, especially if experimental constraints (distance, dihedral angles) are explicitly taken into account. Therefore, MD is nowadays an integral part of protein structure refinement using either NMR or X-ray diffraction constraints [27,28]. Several techniques aimed at enhancing the conformational hyperspace that can be scanned are described in the literature [29–33].

In a drug-design protocol, these methods may be very powerful to propose reliable conformations of small molecules in their free state and even to avoid conformational artefacts given by X-ray diffraction. One simple and useful application is the conformational analysis of tetra-O-methyl-(+)-catechin (Fig. 1). In the crystalline state, the two observed conformations of the benzopyran ring places the dimethoxyphenyl moiety at C2 in an equatorial position [34]. This is in disagreement with proton NMR coupling constants which suggest an interconversion of axial and equatorial conformations [35]. A 4.5 ns MD simulation *in vacuo* starting from the two crystal structures not only showed the interconversion, but was also able to reproduce the NMR-derived ratio between the two populations of axial and equatorial conformations [34].

An efficient conformational sampling of free ligands may also reveal significant differences, related to their specificity profile for closely related receptors. The conformational hyperspace accessible to deltorphin C and dynorphin A analogs could, for example, explain their selectivity for  $\delta$ ,  $\mu$  [36] and  $\kappa$  opioid receptors [37].

## 3. Synergistic Use of MD and 3D QSAR

3D QSAR models are highly dependent on the alignment of bioactive Conformations [38]. Although the complexity of conformational searching techniques does not necess-

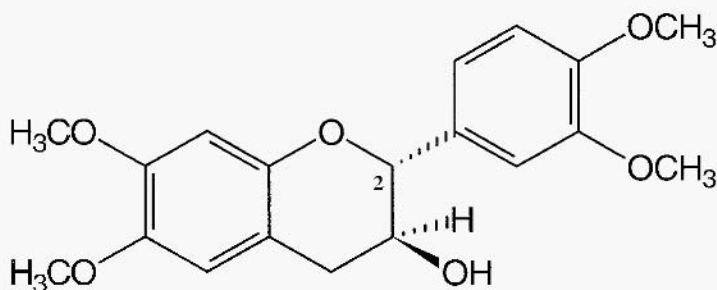


Fig. 1. Chemical structure of tetra-O-methyl-(+)-catechin.

arily correspond to their usefulness in determining bioactive conformations, another potential application of MD resides in the conformational analysis of semi-flexible molecules prior to pharmacophore mapping. Relevant conformational populations or molecular properties derived from MD may thus be readily identified and imported into a QSAR table [39–40].

Alternatively, simulations of protein–ligand complexes may explain some discrepancies in 3D QSAR analyses, due to a non-uniform binding mode of tabulated molecules and thus identify outliers. One example in our lab concerned the CoMFA study of class I MHC-binding nonapeptides [41]. Whatever the parameters used, no model was able to explain the poor activity of two compounds (Asn and Tyr analogs; see Fig. 2). Since the free ligands only were taken into account, a single binding mode had to be postulated for the 10 nonapeptides. In fact, this assumption did not depict the probable situation. When all compounds were directly modelled and simulated in the MHC binding groove, the above-described discrepancy could be easily explained by the progressive expulsion of the C-terminal Asn/Tyr side chains from a rather hydrophobic local subsite, whereas other side chains remained in the binding groove [42]. Very short MD trajectories (30 ps) were sufficient indeed for clearly distinguishing the two outliers. After their removal from the analysis, the model was found to be much more predictive ( $r_{cv}^2$  raised from 0.52 to 0.75) [38]. A clear identification of outliers is one of the quality controls that are recommended for evaluating 3D QSAR models [43]. MD simulations could be particularly well adapted for this task, notably when several binding modes or conformational heterogeneity of the tabulated ligands is suspected.

#### 4. MD as a Complementary Tool to X-Ray Structure Determination

Protein crystal structures represent one of the most attractive starting points for a rational drug-design procedure. However, X-ray structures may not be accurate enough for drug design because: (i) only few informations on the dynamics of the macromolecule can be derived [44], and (ii) electron density cannot be interpreted due to a too low resolution [45]. Here are two examples where both reasons were verified, and for which the complementary use of MD provided a reliable help.

The first case concerns the enzyme acetylcholinesterase (AChE) whose function is to hydrolyze acetylcholine in cholinergic nerves. The X-ray structure of AChE from *Torpedo californica* has been obtained at a resolution of 2.8 Å [46]. Remarkably, the active site is located far from the enzyme surface (about 20 Å) at the bottom of a deep, narrow gorge. This gorge may function as a cation pump by the combined action of a dipole gradient (aligned within the gorge axis) and aromatic side chains delimiting the walls of the gorge [47]. However, its mechanism of action and notably the high catalytic rate of the enzyme could not be fully explained from the crystal structure alone. Notably, three residual electron density peaks present in the gorge have to be attributed to either water molecules or small cationic species that may drive the substrate entry and fix its bound conformation [48]. MD simulation of AChE in presence of either three water molecules or three ammonium cations filling the extra electron density provided a plausible explanation. Simulations performed in presence of water molecules yielded to



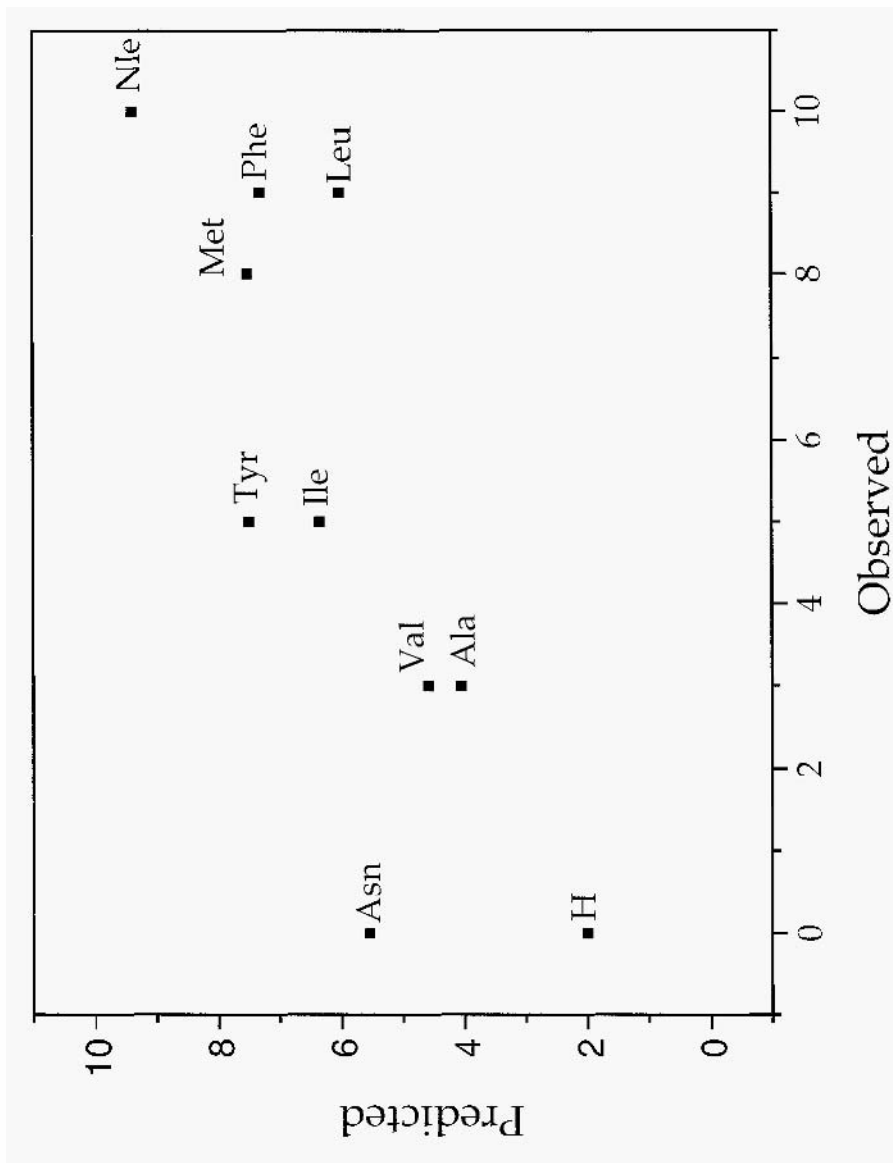


Fig. 2. Predicted versus observed biological activities (detection limit concentration, in  $-\log M$ , lending to the lysis of class I MHC H-2L<sup>d</sup>-expressing target cells by the T cell clone IE1) for a set of 10 nonpeptides (Tyr-Pro-His-Phe-Met-Pro-Thr-Asn-Xaa, Xaa = Tyr, Phe, Nle, Leu, Ile, Met, Val, Ala, Asn, H) (reference [41])

altered conformations of active site residues (rms deviations of 1.5 Å), whereas MD with explicit definition of three small cations led to structures in remarkable agreement with X-ray diffraction data (rmsd < 0.5 Å) [48]. The combined use of X-ray crystallography and molecular dynamics simulations clarifies here the dynamical behavior of AChE. It reveals the transient formation of a short channel through the active site, large enough for a water molecule [9]. A so-called ‘back-door’ hypothesis was formulated to explain substrate/product entry/elimination. Although it was supported by the electrostatic potential of the enzyme [9], it still has not been fully evidenced by recent site-directed mutagenesis studies [49].

A second possible use of MD to complement experimentally derived structures applies to protein crystal structures for which the electron density map of the ligand is too sparse to ensure an unambiguous definition of its bound conformation. This happened to the X-ray structure determination of class I major histocompatibility encoded (MHC) proteins, purified from natural sources (infected cells, for example). MHC-I proteins regulate the immune surveillance of intracellular pathogens by presenting antigenic peptides to cytotoxic T cells at the surface of infected cells [50]. As free proteins are unstable and need the presence of a bound peptide to properly fold [51], purification of MHC molecules is always accompanied by a co-purification of a peptide pool (containing up to several hundred peptides) whose electron density cannot be solved [45,52]. To propose a bound conformation of natural peptidic ligands as well as MHC-peptide interactions, we filled the residual electron density map of the HLA-A2.1 protein by a viral antigen, known to be naturally presented by this allele. After 100 ps MD simulation in water, a 3D picture of a MHC-peptide complex could be proposed [53]. It was nicely validated one year later by X-ray crystallography of the HLA-A2.1 protein in complex with the same peptide [54]. The bound conformation of the peptide was in remarkable agreement with our MD model (Fig. 4), with rms deviations of 1.2 Å on backbone atoms. Notably, the correct anchoring side chains (at position 2, 3 and 9) were found in their bioactive conformations (Fig. 3). More discrepancies were observed in the middle part of the peptide sequence (from positions 4 to 7) which is loosely bound to the protein and bulges out of the binding cleft. However, the MD model was, to the best of our knowledge, the very first realistic three-dimensional picture of a MHC-peptide complex for which the full atomic coordinates of the bound peptide were described.

## **5. Simulating Truncated Active Sites in a Virtual Fluid**

The rate of protein cloning/sequencing being by far higher than that of 3D structure determination, the majority of macromolecular targets of potential interest for the pharmaceutical community have unknown 3D structures. However, if their primary sequence is close enough to that of a protein for which a NMR or X-ray structure exists, the target protein may be modelled by homology to the existing 3D structure [55]. As a precise function is generally associated with a well-defined fold, a reliable picture of the active site may be constructed by this technique. However, if dynamical insights into the protein function are needed, one is getting into trouble. The main problem resides in

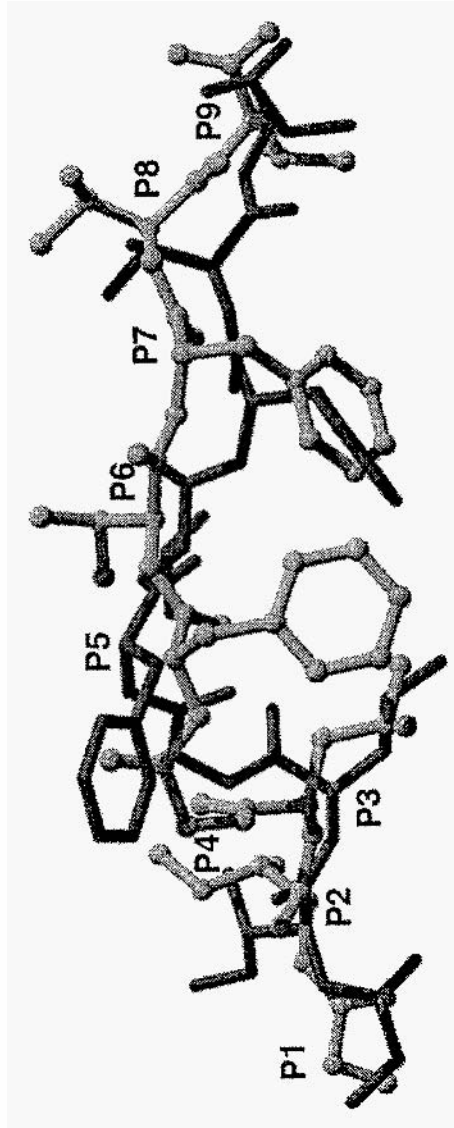


Fig. 3. X-ray versus MD model of the HLA-A2. 1-bound conformation of a T cell epitope (GILGFYFTL) from the influenza virus matrix protein. Backbone atoms of the simulated HLA-A2 have been fitted to the crystal structure and are displayed for sake of clarity. The MD and crystal structure of the bound nonapeptide are shown as gray and white ball and sticks respectively. peptide positions are labeled at the C $\alpha$  atoms, from the N-terminus (P1) to the C-terminus (P9).

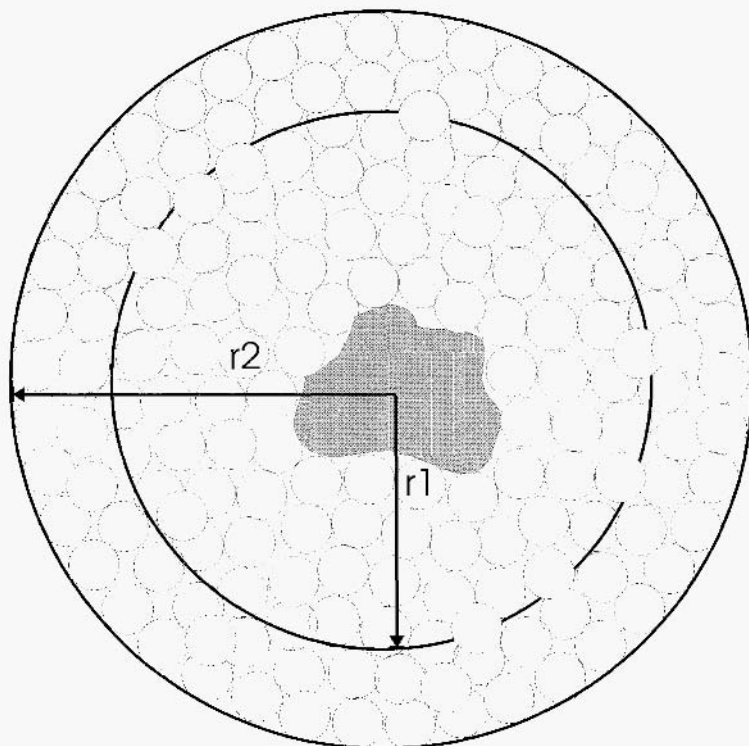


Fig. 4. Schematic display of a truncated active site, surrounded by a pseudo-particle fluid. The particles located in the outer shell ( $r1 < r < r2$ ) are harmonically constrained, whereas the particles of the inner shell ( $r < r1$ ) and the truncated active site are able to freely move.

building the outer part of the active site, for which less homology to the crystal tern-plate(s) is usually observed. A work-around is to construct a pseudo-receptor model [56], a collection of individual amino acids connected by artificial bonds or spacers, that may reproduce as closely as possible the real interaction capacities of the full protein. However, the flexibility of ligand-receptor interactions is not really taken into account by pseudoreceptor models.

To enhance this representation, we have developed an MD method within the GROMOS program [57], able to simulate free amino acids of an active site surrounded by a pseudo-particle fluid reproducing the behavior of the outer amino acids and the solvent (Fig. 4) [58]. The properties of the particle (radius, charge, dipole moment) may be chosen in order to reproduce either a hydrophobic or hydrophilic protein environment [51] (Eqs. 3–4):

$$V = V_{LJ} = 4\epsilon[(\sigma / r)^{12} - (\sigma / r)^6] \quad (3)$$

$$V = V_{LJ} + V_{EL} = V_{LJ} + (4\pi\epsilon)^{-1} \sum (q_i q_j / r_{ij}) \quad (4)$$

The potential energy  $V$  between two particles depend on their nature. Simple Lennard-Jones particles interact only through van der Waals dispersion forces ( $V_{LJ}$ ; Eq. 3) as a function of the interacting distance  $r$ , their diameter  $\sigma$  (about 6 Å) and the potential well value  $\epsilon$  at the optimal interaction distance. They reproduce pure hydrophobic environments. More sophisticated dipole particles were created to simulate polar outer parts. They consist of two pseudoatoms  $i, j$  of opposite charge  $q$  connected by a very short bond. Their dipole moment (1.44 D) was slightly larger than that of an N–H bond.

Several protein structures (adenylate kinase, retinol binding protein, HIV-1 protease, HLA-B27 human leukocyte antigen) for which a crystal structure exists could be pretty well reproduced by simulating the truncated active site–ligand complexes in a virtual fluid. Positional rms deviations, atomic fluctuations, radius of gyration and hydrogen-bonding patterns were close to that of the parent crystal structure, and nearly as accurate as the corresponding values obtained after standard MD simulation of the fully solvated complex [58]. The method was able not only to reproduce experimentally determined protein–ligand complexes, but could furthermore well explain the effect of protein mutation at the active site, or rationalize the different binding affinities of related ligands to the same macromolecule [59]. The main advantage of the method are: (i) it combines the rapidity of *in vacuo* simulations with the accuracy of computations in a full water environment, and (ii) it takes into account the only part of the protein (active site) that can be reliably derived by homology modelling.

## 6. Quantitative Analysis of Ligand Binding

One of the most exiting applications of MD to biology is the computation of free energy differences that may be related to the experiment [60]. Any biological process can be formulated by the corresponding thermodynamic cycle (Fig.5) As the free energy  $G$  is a state function ( $\Delta G$  of a closed cycle is zero), and the free energy change  $\Delta G$  describing a system in equilibrium is path-independent. experimentally determined free energy changes (horizontal processes corresponding to biological equilibria; Fig. 5) can be directly compared to computed free energy changes (vertical processes corresponding to theoretical equilibria: Fig. 5). In practice, the starting state ‘0’ is progressively converted into the final one ‘1’, by modifying a state-dependent parameter  $\lambda$  from 0 to 1 at discrete intervals in  $\lambda$ . Free energy differences  $\Delta G$  are generally calculated from the potential energy function  $V_{\lambda(i)}$  by the free energy perturbation technique (Eqs. 5 and 6) or the thermodynamic integration method (Eq. 7):

$$\Delta G = G_1 - G_0 = \sum_i G_{\lambda(i+1)} - G_{\lambda(i)} \quad (5)$$

$$G_{\lambda(i+1)} - G_{\lambda(i)} = -RT \ln \langle \exp - [V_{\lambda(i+1)} - V_{\lambda(i)}] / RT \rangle_{\lambda(i)} \quad (6)$$

$$\Delta G = G_1 - G_0 = \int_0^1 \langle \partial V / \partial \lambda \rangle_{\lambda} d\lambda \quad (7)$$

The major drawbacks of free energy calculations are: (i) an experimentally determined protein structure of high accuracy is recommended as a starting point; (ii) it is a

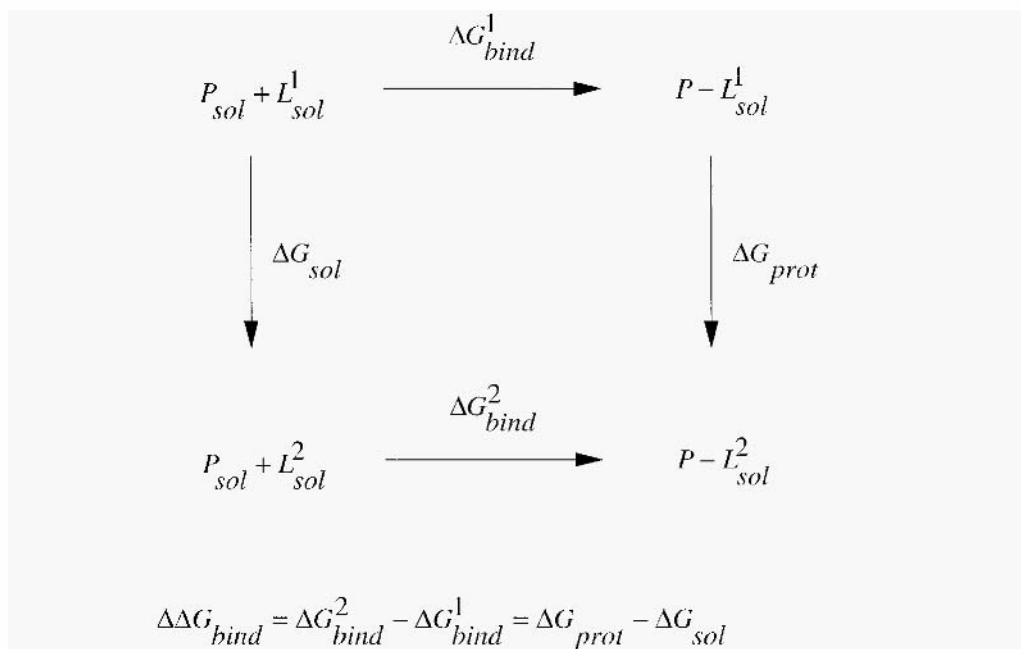


Fig. 5. Thermodynamic cycle for the formation of a protein (P)-ligand (L) complex. Horizontal equilibrium values are measured experimentally (free energy of binding of ligands L1, L2 to the protein P;  $\Delta G_{bind}^1$ ,  $\Delta G_{bind}^2$ ), whereas vertical equilibria can only be computed (free energy change upon ligand mutation in water or in the protein-bound state;  $\Delta G_{sol}$ ,  $\Delta G_{prot}$ ). Experimentally determined free energy differences ( $\Delta G_{bind}^1 - \Delta G_{bind}^2$ ) are equal to computed ones ( $\Delta G_{sol} - \Delta G_{prot}$ ). If L2 is a dummy molecular the free energy difference relates to the absolute free energy of association of ligand L1 to protein P.

time-consuming method as free energy changes need to be calculated for the free and bound states, in forward and backward directions: (iii) it necessitates an exhaustive conformational sampling within each perturbation window for all intermediate states between the starting and final one; and (iv) it better applies to amino acid mutations for which force-field parameters are generally best derived. This means that mutating a small molecular weight inhibitor into an analog needs, first of all, an accurate parameterization of the two organic molecules.

In the literature, numerous examples are provided for computing free energy changes upon protein and/or ligand mutation (for a recent review see [61]). The advantage of the method is that it allows a clear distinction of entropy and enthalpy contribution upon ligand binding [62,63] that may be used in a lead optimization program, for example. Many applications are focused on solvation free energies of small molecules [6,17], or host-guest interaction complexes [64,65]. Quantifying protein-ligand interactions is much more complex. Generally, free energy changes upon protein mutation is considered to study the protein function and mechanism of action [66,67]. However, some reports denote that it may be applied with reasonable success to the quantitative

analysis of ligand binding [68–71], notably for determining its stereospecificity [72,73] or optimizing its solubility and bioavailability [74,75].

One challenging problem is the computation of absolute free energies, like the free energy of association of a ligand to a protein [61,76], and for that task, ‘double annihilation techniques’ [77] have to be used. That means that the ligand is progressively mutated to ‘nothing’ (dummy atoms), both in solution and in the protein-bound state. The major difficulty that is encountered is to relate the drastic chemical modification of the ligand to a rather large absolute free energy (about 10–15 kcal/mol). For two different protein–ligand complexes (boitin/streptavidin, N-acetyltryptophanamide/  $\alpha$ -chymotrypsin), absolute  $\Delta\Delta G$  binding values were overestimated by 3–4 kcal/mol but remained in qualitative good agreement with experimental values [76], even if the free energy of association is very high (18 kcal/mol for the biotin–streptavidin complex). Whether  $\Delta\Delta G$  changes may be decomposed into van der Waals and electrostatic contributions (a very important parameter for lead optimization) is still a matter of debate [78–79] as free energy decomposition is path-dependent [78] and individual components converge very slowly [62].

Another interesting aspect of free energy calculations is the prediction of cavity hydrations at the protein–ligand interfaces. At least two water-mediated H-bonds were shown to be energetically necessary for hydrating a protein–ligand interfacial cavity [80]. A recent survey of 19 high-resolution protein X-ray structures provided even more drastic requirements as 80% of water molecules bridging protein–ligand interactions were involved in three or more H-bonds [81]. Taking into account accurate positions of protein-bound water molecules clearly accelerates and fastens ligand design, as recently exemplified by the protein-based design of thymidilate synthase [82] and HIV-1 protease inhibitors [83].

## 7. Qualitative Analysis of Ligand Binding

MD models are not aimed at replacing experimentally determined three-dimensional structures. In the very best cases, using time-consuming protocols, rms deviations of about 1 Å from the starting crystal structure have to be expected [18,24]. However, such accurate representations are not very realistic in a drug-design protocol. A drug-design cycle such as that proposed by Blundell [84] can only be successful if it relies on an interactive multidisciplinary approach that provides quick answers and feedback. Obviously, ns simulations of a series of protein–ligand complexes are not compatible with this rule, and would delay the prioritization of ligand synthesis and testing.

However, this does not mean that MD simulations of a macromolecular target with a set of related ligands cannot be very useful if one is interested in qualitative aspects of ligand design. MD models present the advantage to propose instantaneous or time-averaged molecular properties that can clearly discriminate high-affinity from low-affinity ligands [85] and thus may either explain [85,86] or predict binding properties [87,88]. The major difficulty of this approach is to find the best compromise between an acceptable accuracy level and a reliable force-field representation [89]. In the field we were interested in (antigen recognition by class I major histocompatibility proteins), and

which will be developed in that section, 200 ps simulation of fully solvated protein–ligand complexes were sufficient to depict predictive models. The most informative molecular properties of the hundred MHC–ligands complexes we have simulated to date will be briefly reviewed.

### 7.1. Atomic positional fluctuations of the bound ligand

Starting from a basic assumption — *the more bound, the less flexible* — we tried to relate the binding affinity of a set of six bacterial peptides to a single class I MHC protein (Table 1).

Atomic fluctuations of the bound peptides, averaged per residue, were thus computed from energy-minimized structures that had previously been averaged over 500 conformations, for each complex (Fig. 6a). It is important to recall that all complexes were simulated using identical conditions, starting from the same protein X-ray structure and the same peptide conformation (with preservation of the  $\chi^1$ ,  $\chi^2$  angles for side chains) [85]. Interestingly, the C-terminal amino acid (position 9) of the two low-affinity peptides **4**, **5** (Table 1) was highly flexible in the protein-bound state. As this position was known to be a dominant anchor to the MHC protein [90], we could reasonably relate the weak affinity of the corresponding two ligands to a probable dissociation of the C-terminal residue from the binding groove, illustrated here by its higher atomic flexibility. In about 90% of all MHC–peptide complexes simulated in our group, this molecular property was always in qualitative agreement with the observed binding affinity of the corresponding ligand. The atomic mobility of each peptide amino acid could be well related to its binding role. Strong anchoring positions (P1, P2, P3 and P9; and Pn meaning position n of the peptide) were generally much less flexible than weak anchoring positions (P4 to P8, see Fig. 6a). Furthermore, amino acid flexibility was used as a guide to enhance the free enthalpy of binding of designed non-natural peptides to the

Table 1 Binding of bacterial peptides to HLA-B\*2705 (reference [85])

Peptide no.	Sequence	Binding affinity <sup>a</sup>
<b>1</b>	RRIKAITLK <sup>b</sup>	n.d. <sup>c</sup>
<b>2</b>	QRLKEAAEK	+
<b>3</b>	RRKAMFEDI	++
<b>4</b>	ERLAKLSGG	–
<b>5</b>	LRDAYTDM	–
<b>6</b>	RRKAMFED	+

<sup>a</sup> The affinity for HLA-B\*2705 was indirectly measured by an epitope stabilization assay which titrates the ability of the ligand to induce the native fold of the HLA protein, that can be recognized by a conformation-specific monoclonal antibody (reference [91]) (++; binding affinity lower than 1  $\mu$ M; +; binding affinity between 1 and 20  $\mu$ M; and –: no detectable binding). The six peptides correspond to sequences of the 57 kD heat shock protein (GroEL) of *Chlamydia trachomatis*.

<sup>b</sup> Peptide model fitted in the extra electron density map produced by self-nonapeptides, in the crystal structure determination of HLA-B27 (reference [52]).

<sup>c</sup> Not determined.



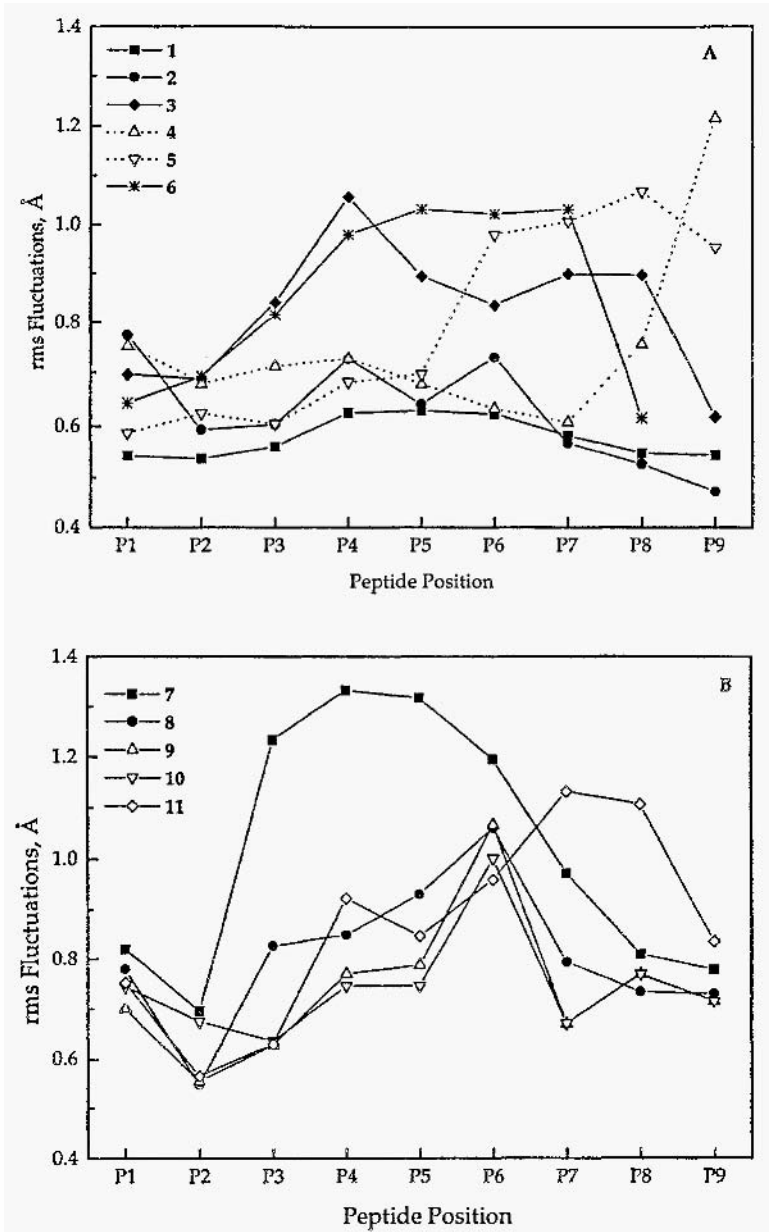


Fig. 6. Root mean square atomic fluctuations of a protein-bound ligand as an indicator of the protein-ligand complex stability. A, atomic flexibility of six HLA-B\*2705-bound peptides 1-6 (Table 1), showing variable binding affinities for the HLA protein. Fluctuations were calculated from energy-minimized conformations (1000 steps of steepest descent followed by 1000 steps of conjugate gradient energy minimization), time-averaged over the last 500 conformations [85]. B, atomic flexibility of a series of HLA-B\*2705 bound nonapeptides 7-11 for which a secondary anchor position (P3) has been varied (Table 2).

Table 2 design of nonnatural HLA-B27 ligands from a bacterial epitope (reference [87])

Peptide no.	Peptide sequence								Binding affinity
	P1-	P2-	P3-	P4-P5-	P6-	P7-	P8-	P9	
	Lys-Arg- <b>Xaa</b> -Ile-Asp-Lys-Ala-Ala-Lys								
<b>7</b>			Gly <sup>a</sup>						+
<b>8</b>			Leu						+
<b>9</b>			Hpa <sup>b</sup>						++
<b>10</b>			Ana <sup>c</sup>						++
<b>11</b>			Bna <sup>d</sup>						++

<sup>a</sup> Peptide **7** is a natural sequence (117–125) of the *Chlamydia trachomatis* GroEL protein (reference [91]).

Binding affinities were measured as described in Table 1.

<sup>b</sup> Hpa, homophenylalanine.

<sup>c</sup> Ana,  $\alpha$ -naphthylalanine.

<sup>d</sup> Bna,  $\beta$ -naphthylalanine.

same MHC protein, HLA-B\*2705 [87,88]. We substituted bulky hydrophobic residues ( $\alpha$ - and  $\beta$ -naphthylalanine, homophenylalanine; see Table 2) for the amino acid of a natural peptidic ligand at a secondary anchor position (P3). As additional interactions to the protein were desired, a better binding of the designed analogs should result in a decreased mobility of the mutated peptide position. After simulating the natural as well as the nonnatural peptides in complex with the host HLA protein, the lowered flexibility of the non-natural amino acid at P3 (Fig. 6b) led us to predict an increased binding of ligands **9–11** when compared to the two natural peptides **7, 8**. Binding assays [87], as well as thermodynamic analysis of the complex denaturation by CD spectroscopy (Fig. 7), were in perfect agreement with the above predictions, as non-natural ligands **9–11** were about 50-fold more potent binders than natural peptides **7–8** [87].

A very similar strategy was used for designing an optimal linker for trivalent thrombin inhibitors [93]. The atomic fluctuations of the linker part, derived from 150 ps MD simulations of several solvated thrombin–inhibitor complexes were also in good agreement with its contribution to  $\alpha$ -thrombin binding. In both cases, looking at atomic positional fluctuations of bound ligands provides a crude but predictive estimate of the stability of protein–ligand complexes.

## 7.2. Accessible versus buried surface areas

These molecular properties are generally correlated to the atomic flexibility of the bound ligand (*the more flexible the less buried*). In the two examples cited above, buried surface areas could also be well related to binding affinities [85,87]. Notably, for the set of non-natural analogs **9–11**, designed to increase the free enthalpy of binding to the protein (Table 2), the amount of buried surface area for the mutated position was in perfect agreement with the size of the corresponding side chain (indicating that almost the entire part of the side chain was buried upon binding to the complementary hydrophobic pocket) and, consequently, with the binding affinity (Fig. 8a).

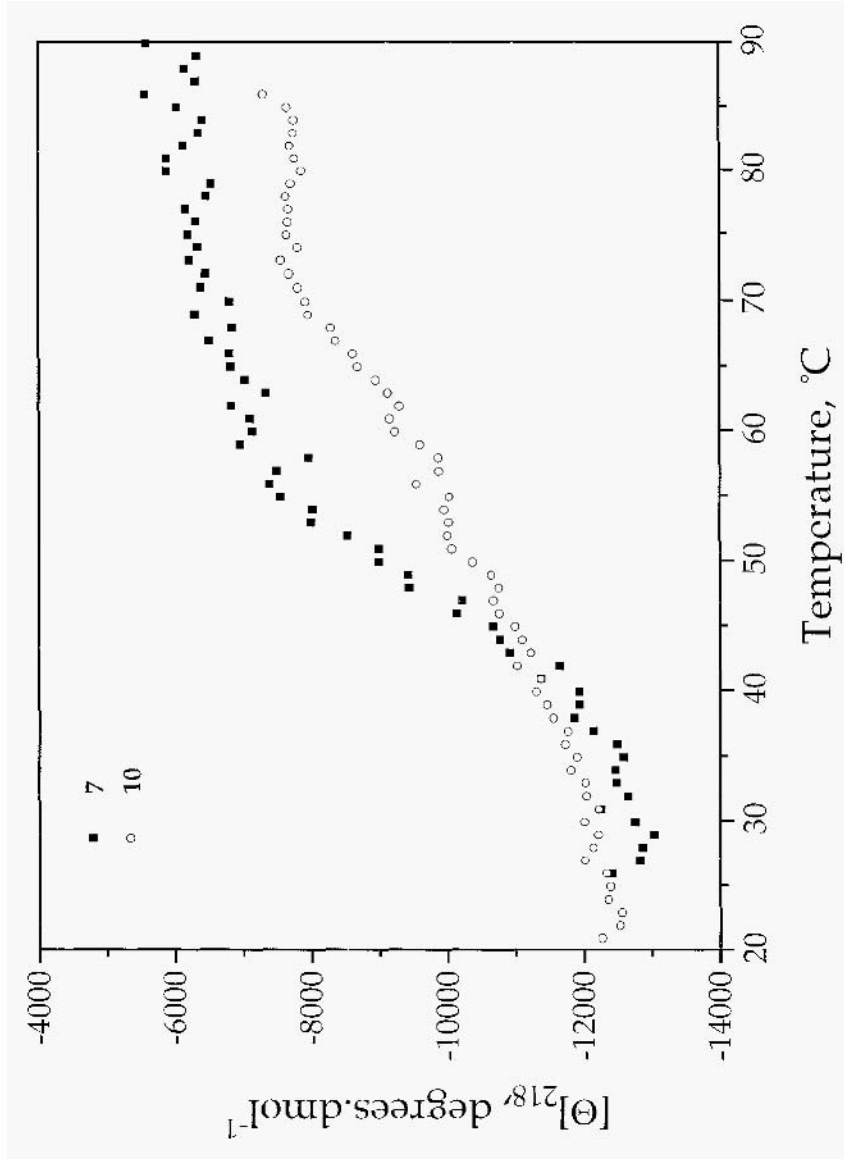


Fig. 7. Thermal denaturation of the complex between HLA-B\*2705 and two related ligands **7**, **10**. Melting temperatures ( $T_m$ ) of the complexes with peptides **7** and **10** have been estimated to 48°C and 57°C, respectively. The contribution of the non-natural chain to the free energy of binding may be estimated from  $\Delta Tm$  values to 2 kcal/mol (reference [92]).

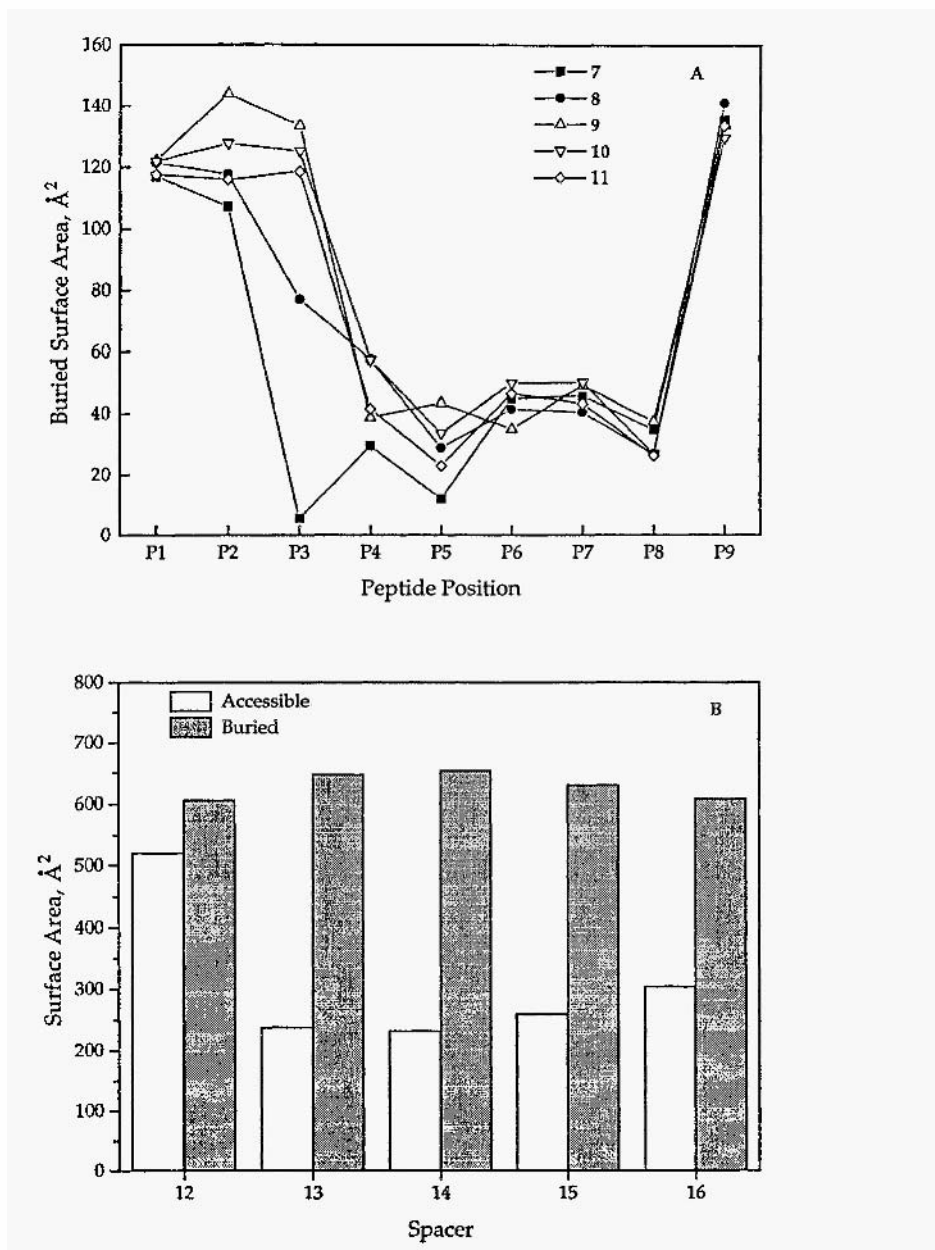


Fig. 8. Accessible/buried surface areas as indicators of the stability of protein-ligand complexes [87]. A, Buried surface areas of five HLA-B27 bound peptides (ligands 7–11, Table 2), calculated on relaxed time-averaged conformations (for the last 500 conformers, between 100 and 150 ps of simulation). Surface areas were calculated for each peptide position, using the MSprogram [94] with a 1.4 Å radius probe. B, Substitution of organic spacers for a pentapeptide core sequence (ligands 12–16, table 3) of a bacterial peptide. Buried and accessible surface areas were measured as in (A) for the whole ligand.

Table 3 Introduction of organic spacers in the sequence of a class I MHC ligand (Gln-Arg-Leu-Spacer-Lys) (reference [87])

Peptide no.	Spacer	Binding affinity
<b>12</b>	Ly-Glu-Ala-Ala-Glu <sup>a</sup>	+
<b>13</b>	Aba-Aba-Aba <sup>b</sup>	+
<b>14</b>	Aha-Aha <sup>c</sup>	+
<b>15</b>	Aua <sup>d</sup>	+
<b>16</b>	Ada <sup>e</sup>	+

<sup>a</sup> *Escherichia coli* dnaK (220–228) epitope [91].

<sup>b</sup> Aba, 4-amino-butyric acid.

<sup>c</sup> Aha, 6-amino-hexanoic acid.

<sup>d</sup> Aua, 11-amino-undecanoic acid.

<sup>e</sup> Ada, 12-amino-dodecanoic acid.

Binding affinities were measured as described in Table 1.

The predictability of that molecular property was verified in another design attempt aimed at simplifying the nonapeptide structure of natural HLA ligands [87]. Starting from a bacterial peptide known to bind to HLA-B27, we have substituted various organic spacers for a pentapeptide core sequence (from positions 4 to 8; see Table 3), suspected to interact with a T cell receptor (TcR) and not the HLA protein [52]. Basically, we wished to simplify as much as possible the peptidic structure of the ligand without impairing its affinity for the host MHC protein; and at the same time, try to alter TcR recognition. In terms of desired molecular properties, this means that the total buried surface area of the designed ligands should remain equal to that of the parent peptide (implying a conserved binding to the HLA protein), but the total accessible surface area should be considerably decreased (implying a reduced binding to the TcR). After simulating the natural as well as the designed ligands in complex with HLA-B27, we computed accessible and buried surface areas for ligands **12–16** in the bound state (Table 3). It clearly appeared that our design goal was perfectly reached, as all MHC–peptide pairs were predicted to be equally stable from the analysis of atomic trajectories. The surface area accessible to water (or a putative TcR) had been significantly reduced by 50% for all ligands bearing the designed organic spacers (Fig. 8b). However, this should not preclude for a tight binding to the HLA-B27 molecule because the total surface area buried upon MHC binding, for each non-natural analog, was similar to that of the natural peptidic ligand (about 600 Å<sup>2</sup>). All ligands were synthesized and tested in an *in vitro* binding assay. As predicted, the introduction of organic spacers did not alter the binding affinity for the HLA-B27 protein, as very similar binding affinities could be observed [87,95]. The ‘spacer effect’ was independent on the parent peptide sequence as similar modifications on three different T cell epitopes led to the same computational and experimental results [D. Rognan and J.A. López de Castro, unpublished results]. These molecules are the very first class I MHC ligands for which half of the canonical peptide structure (P4–P8) have been successfully replaced by organic spacers, and represent the first rational step towards nonpeptide TcR partial agonists or antagonists.

### 7.3. Protein-ligand non-bonded distances

Monitoring the time course of critical topological features (non-bonded distances, angles) is often used to analyze trajectories of protein–ligand complexes [96–100]. The regio- and(or) stereoselectivity of hydroxylation of nicotine and several camphor derivatives by cytochrome P450cam could have been predicted with a good accuracy by simply looking at the non-bonded distances between substrate carbon atoms and the ferryl oxygen intermediate of the heme moiety [73,96,97]. Examination of some key distances was also used to study the deacylation enantioselectivity of acylenzymes [97]. Such analyses are best suited for studying enzyme–substrate interactions for which a well-defined topology of a few atoms in the active site is often associated with a precise biochemical event.

For enzyme–inhibitor complexes or when the macromolecule–ligand interaction surface is very broad (e.g. MHC–peptide complexes), one cannot restrict the trajectory analysis to a few atom-centered non-bonded distances. For examining the fine specificity of antigen binding to two class I HLA alleles differing by only one amino acid (Table 4), we have used a slightly different approach [86], in which key distances were not measured between atoms, but center of masses (cmass). The distance between protein and peptide cmass remained constant for the most stable pairs **17a,b–18a,b**, whereas it was still increasing after 200 ps MD for the less stable complexes **19a,b** (see Fig. 9 and Table 4). Computing inter-cmass distances between individual MHC-binding amino acids and their complementary pocket allowed the identification of the peptide part (position 9) that was progressively expelled from the binding groove. The analysis was greatly facilitated by the fact that the peptide sequence could be well separated into a protein-binding substructure (P1–P2–P3 and P9) and a non-interacting part (P4–P5–P6–P7–P8). Furthermore, each MHC-binding amino acid develops allele-specific interactions with complementary pockets of the MHC proteins. Therefore, this approach is particularly well suited to examine protein–peptide interactions and monitor the binding contribution of each peptide amino acid.

### 7.4. Protein-ligand hydrogen-bonds

The distribution of protein–ligand H-bonds is also a good indicator of the complex stability. Very often, a pure quantitative H-bond analysis based on distance–angle criteria [102,103] is performed on time-averaged structures [85,93,99,100]. For rationalizing subtle structure–activity relationships for the set of peptides **17–19** (Table 4), we combined a quantitative and qualitative analysis of intermolecular H-bonding by computing the frequency of all protein–ligand H-bonds [86] (Fig. 10). The distribution of strong and medium H-bonds correlates well with the binding potency of the peptide. A similar number of strong H-bonds were found for complexes **17a**, **17b**, **18a** and **18b**, consistent with the similar binding efficiencies of peptides **17** and **18** to both MHC alleles. At the opposite, a reduced number of medium and(or) strong H-bonds (peptide **19** in complex with the two alleles) correlates with the decreased binding of this peptide. The weakest binding potency (peptide **19** to B\*2703) could be qualitatively and

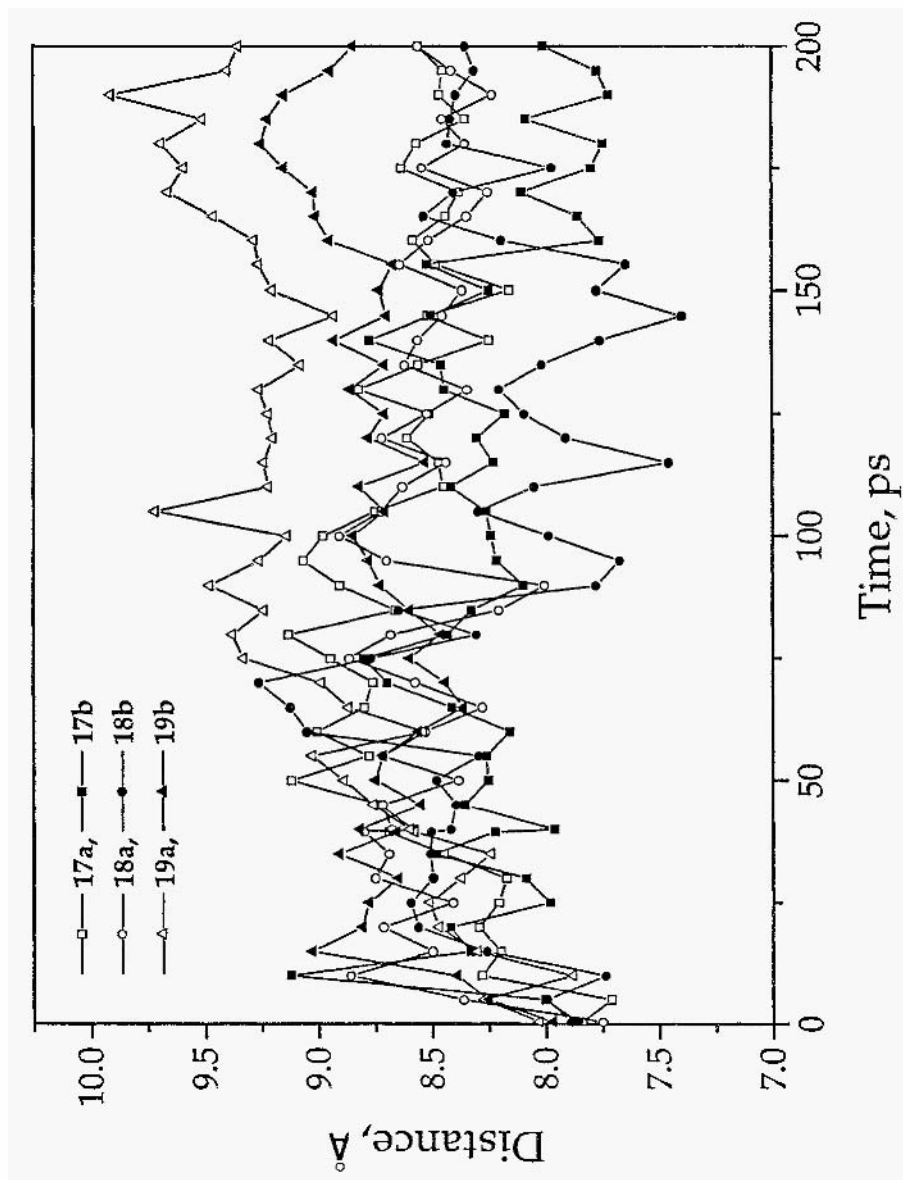


Fig. 9. Time course of the distance between the center of mass of peptides 17-19 and that of the host HLA-B27 stbht-pe (a, B\*2703; b, B\*2705). Binding affinities of the three ligands for the two alleles are listed in Table 4.

Table 4 Distance in Å, between proteins (HLA-B\*2703, HLA-B\*2705) and peptide center of masses (reference [86])

No. sequence	17 RRYQKSTEL <sup>a</sup>		18 ARYQKSTEL		19 RQYQKSTEL	
	B*2703	B*2705	B*2703	B*2705	B*2703	B*2705
Allele	B*2703	B*2705	B*2703	B*2705	B*2703	B*2705
C50 <sup>b</sup> (μM)	1.0	1.0	1.5	2.0	> 100	10
Complex	<b>17a</b>	<b>17b</b>	<b>18a</b>	<b>18b</b>	<b>19a</b>	<b>19b</b>
Inter-mass Distances, Å						
d1 <sup>c</sup>	8.2 ± 0.3	8.5 ± 0.3	8.3 ± 0.3	8.2 ± 0.4	9.3 ± 0.9	9.1 ± 0.7
d2	5.9 ± 0.3	5.9 ± 0.3	5.5 ± 0.3	5.6 ± 0.4	7.3 ± 0.5	6.6 ± 0.5
d3	11.5 ± 0.4	12.1 ± 0.5	12.3 ± 0.5	10.7 ± 0.5	12.0 ± 0.6	12.3 ± 0.6
d4	3.6 ± 0.3	2.9 ± 0.4	2.3 ± 0.3	1.5 ± 0.5	3.4 ± 0.4	2.8 ± 0.3
d5	4.7 ± 0.2	4.7 ± 0.3	4.7 ± 0.3	4.4 ± 0.3	5.1 ± 0.4	4.7 ± 0.3
d6	4.9 ± 0.6	4.8 ± 0.5	5.3 ± 0.5	5.0 ± 0.7	4.6 ± 0.5	5.2 ± 0.6
d7	2.5 ± 0.4	2.6 ± 0.3	2.9 ± 0.3	3.4 ± 0.4	7.3 ± 0.6	4.8 ± 0.5

<sup>a</sup> Human histone H3 peptide: a self peptide, naturally bound to HLA-B\*2705 [90] and B\*2703 (reference [101]).

<sup>b</sup> Binding data are expressed as the micromolar excess of peptide analog relative to the wild-type peptide **17**, at which HLA-B27 fluorescence (measured by FMC analysis with an anti-B27 monoclonal antibody) on RMA-S cells was half the maximum obtained with the wild-type peptide.

<sup>c</sup> d1, protein/peptide; d2, protein/MHC-anchors (P1-P3, P9); d3, protein/TcR-anchors (P4-P8); d4, pocket A/P1; d5, pocket B/P2; d6, pocket D/P3; d7, pocket F/P9.

quantitatively explained by the low number of strong hydrogen-bonds. Interestingly, not only the number but also the quality of the MHC–ligand interactions (strong versus medium H-bond) correlates well with the binding potency.

### 7.5. Protein–ligand non-bonded contacts

Reporting the number of non-bonded contacts between the host protein and a series, a related ligand affords a more general survey of intermolecular contacts. In our design attempt to simplify and enhance the binding of natural T cell epitopes to the HLA-B\*2705 protein, we substituted two organic polymers for the pentapeptide sequence (from position 4 to 8) of the bacterial nonapeptide **12** (Table 5). After parameterization of the organic polymer for the AMBER force field [104] and MD simulation of the corresponding three complexes, we predicted an increased binding of the tetramer analog **20** and a decreased binding affinity of the trimeric spacer **21** with respect to the natural nonapeptide [89], by simply comparing the number of protein–ligand non-bonded contacts closer than 4 Å (Fig. 11). Interestingly, the 200 ps trajectories were long enough to predict an enhanced binding role of the tetrameric polymer to the central part of the HLA binding groove (increased number of non-bonded contacts with the spacer), whereas the reduced binding of the trimeric compound was attributed to its short length and the perturbation of the interaction of the N-terminal part (less contacts to the very important P1–P3 peptide sequence, in spite of a high number of interactions in the



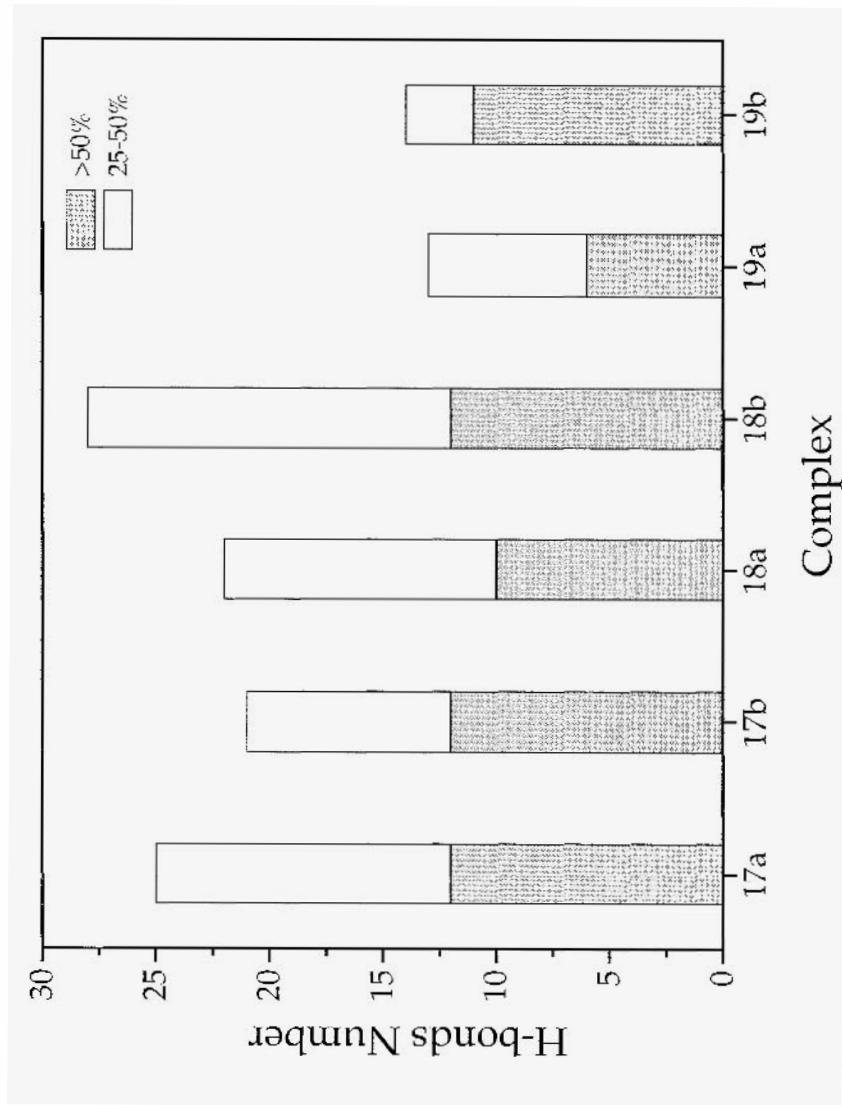


Fig. 10. MHC-peptide hydrogen-bonding frequency for peptides 17-19 in complex with HLA-B\*2703 (a) or HLA-B\*2705 (b). Binding affinities of the three ligands for the two alleles are listed in Table 4. H-bonds have been here geometrically defined by an acceptor (A) to donor (D) distance less than 3.25 Å and a D-H...A angle higher than 120 deg. Interactions were statistically monitored throughout the simulations for a total of 400 conformations per MHC-peptide complex. Two categories of H-bonds were defined: strong ones with frequencies higher than 50%, and medium ones with occurrences between 25% and 50%.

Table 5 Incorporation of two organic polymers in the sequence of a class I MHC ligand (Gln-Arg-Leu-Spacer-Lys) (reference [89])

Peptide no.	Spacer
<b>12</b>	Lys Glu-Ala-Ala-Glu <sup>a</sup>
<b>20</b>	(R) <sub>4</sub> <sup>b</sup>
<b>21</b>	(R) <sub>3</sub>

<sup>a</sup> *Escherichia coli* dnaK (220–228) epitope (reference [91]).

<sup>b</sup> The structure of the organic moiety R will be published elsewhere.

spacing part). The number of non-bonded contacts was in good agreement with the *a posteriori* observed binding results for the three compounds. When compared to the natural peptide ligand **12**, the tetrameric and trimeric compounds exhibit a 5-fold increase and a 4-fold decrease in HLA-B\*2705 binding, respectively [J.A. López de Castro. personal communication.

## 8. Ligand Design and Docking

In the past five years, tremendous research efforts have been devoted to computer algorithms able to optimize the *de novo* design of ligands from the knowledge of protein three-dimensional structures [4,105]. One of the major drawbacks of fragment build-up procedures (one of the main techniques used in *de novo* drug design, with 3D database searching) is the ‘irreversible’ nature of the ligand-design procedure. Once a fragment (substructure) complementary to one part of the active site has been designed, its structure cannot generally be perturbed (e.g. ring closure or opening) during the ongoing ligand design, in order to optimize non-bonded interactions of the final hit with the receptor active site. This problem was addressed by coupling ligand building to MD in two closely related algorithms [106,107] in which the protein active site is filled with either particles or fragments that can be covalently linked or separated in a stochastic and dynamically reversible manner. While the particle-based approach is only adequate for building apolar ligands, the fragment-based procedure takes into account electrostatic contributions to the protein–ligand binding energy. Both procedures, tested on the same set of two high-resolution crystal structures (HIV1-protease, FKBP-12), suggested inhibitors closely related to existing ligands. Unfortunately, this procedure is not optimally suited for pharmaceutical purposes, because of the complexity of the structures that are generated (none of the hits proposed in the original papers had been synthesized to test the validity of the method), and the computing effort that is asked for.

Flexible docking of small molecules to known three-dimensional protein structures will be the last application of MD to drug-design techniques that will be reviewed here. Although recently described algorithms can relatively well handle the ligand flexibility problem [108–110], the potential flexibility of the active site still remains an open question. Generally, this problem can be addressed only by saving different conformations or rotameric states of the protein active site and use these coordinates as starting points

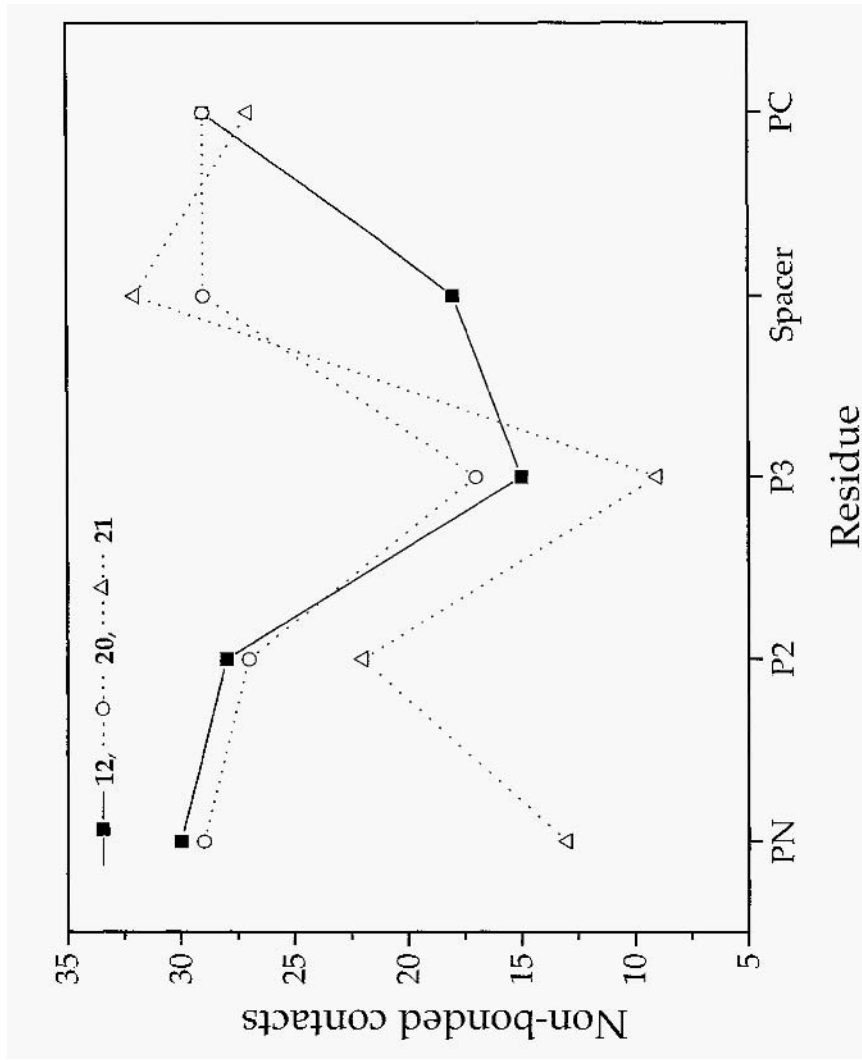


Fig. 11. Substitution of organic polymers (19, 20) for the natural pentapeptide core (Lys-Glu-Ala-Ala-Glu) of an HLA-B27-binding bacterial epitope (peptide 12, Table 5). PN, P2, P3 and PC represents the N-terminal, second, third and C-terminal residues, respectively. Intermolecular non-bonded contacts were summed up for the spacing group (Spacer). Non-bonded contacts between each ligand residue and any protein atom closer than 4 Å were calculated from time-averaged conformation based on 200 ps MD simulations in a water shell (reference [89]).

for parallel docking attempts. The only described methodology that can intrinsically provide a reliable solution is a molecular dynamics docking technique [111], in which the motion of the mass of the ligand is separated from its internal/rotational motion and from that of the receptor by three different coupling to thermal baths. When applied to the flexible docking of phosphocholine to the frozen McPC603 antibody structure, the mass of the ligand was shown to explore a very wide conformational space (about 0.8 nm<sup>2</sup>) and to land after 20 ps simulation in a crystal-like situation (rmsd of 0.5 Å for substrate skeleton atoms) [111]. No attempt to dock a flexible ligand into an unconstrained protein was reported, but the method appears to be very promising to depict local geometry modifications of the active site upon docking of a flexible ligand.

## 9. Concluding Remarks

Up to very recently, molecular dynamics simulations of protein(DNA)–ligand complexes had not been very popular in lead finding and optimization for two major reasons. The first one was the high computing effort that was required, which precluded for the comparison of a set of related compounds in their bound states. The second one was the low number of 3D structures for macromolecular targets of interest. The situation has dramatically evolved in the last two or three years, as MD simulations are nowadays feasible for even larger and larger molecular systems, at relatively low cpu time costs, on affordable computer platforms [112]. The variety of all MD applications reviewed in this chapter is particularly well adapted to everyday drug-design problems and makes from that computing method a very powerful tool for the comparative and qualitative analysis of protein–ligand structures that may go beyond available experimental data.

However, it should be recalled that: (i) MD models will neither substitute for experimentally determined structures, (ii) MD is not the only potent conformational sampling technique [113] and (iii) the successful application of molecular dynamics simulations in a drug-discovery program absolutely needs a strong and permanent feedback to the experiment.

## Acknowledgements

I wish to thank Professor Gerd Folkers (ETH Zürich) for critical reading of the manuscript, and Arthur for being so quiet at home during the writing period. The financial support of the Schweizerischer Nationalfonds zur Förderung der wissenschaftlichen Forschung (Project No. 31-45504.95), as well as the allocation of cpu time on the CRAY-J90 and the INTEL Paragon by the calculation center of the ETHZ, is sincerely acknowledged.

## References

1. Sambrook, J., Fritsch, E.F. and Maniatis, T., *Molecular cloning: A laboratory manual*, 2nd Ed., Cold Spring Harbor Laboratory Press, 1989.
2. Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Jr., Brice, M.D., Rodgers, J.M., Kennard, O., Shimanouchi, T. and Tasumi, M., *The protein data bank: a computer-based archival life for macromolecular structures*, *J. Mol. Biol.*, 112 (1977) 535–542.

3. The current status (18 December 1996) of the PDB is: 5504 coordinate entries (5096 proteins, 396 nucleic acids, 12 carbohydrates).
4. Lybrand, T.P., *Ligand-protein docking and rational drug design*, *Curr. Opin. Struct. Biol.*, 5 (1995) 224–228.
5. Bamborough, P. and Cohen, F.E., *Modeling protein-ligand complexes*, *Curr. Opin. Struct. Biol.*, 6 (1996) 236–241
6. Brooks. C.L., III. and Case, D.A., *Simulations of peptide conformational dynamics and thermodynamics*, *Chem. Rev.*, 93 (1993)2487–2502.
7. Marshall, G.R., Barry, C.D., Bosshard, R.A., Dammkoehler, R.A. and Dunn, D.A., *The conformational parameters in drug design: The active analog approach*, In Olson, E.C. and Christoffersen, R.E. (Eds.) *Computer-assisted drug design*, ACS Symp. series, Vol. 112, American Chemical Society, Washington DC., 1979, pp. 205–226.
8. Goodford. P.J., *A computational procedure for determining energetically favorable binding sites on biologically important macromolecules*, *J. Med. Chem.*, 28 (1985) 849–857.
9. Gilson, M.K., Straatsma, T.P., McCammon, J.A., Ripoll, D.R., Faerman, C.H., Axelsen, P.H., Silman, I. and Sussmann, J.L., *Open 'back door' in a molecular dynamics simulation of acetylcholinesterase*, *Science*, 263 (1994) 1276–1278.
10. Alder. B.J. and Wainwright. T.E., *Studies in molecular dynamics: I. General methods*, *J. Chem. Phys.*, 31 (1959). 459–466.
11. Rahman, A., *Correlations in the motion of atoms in liquid argon*, *Phys. Rev.* 136 (1964) 405–411.
12. McCammon, J.A., Gelin, B.R. and Karplus, M., *Dynamics of folded proteins*, *Nature*. 267 (1977) 585–590.
13. McCammon, J.A., Wolynes, P.G. and Karplus, M., *Picosecond dynamics of tyrosine side chains in proteins*, *Biochemistry*, 18 (1979) 927–42.
14. van Gunsteren, W.F. and Karplus, M., *Effect of constraints, solvent and crystal environment on protein dynamics*, *Nature*, 293 (1981) 677–678.
15. Plimpton, S. and Hendrickson, B., *A new parallel method for molecular dynamics simulation of macromolecular systems*, *J. Comput. Chem.*, 17 (1996) 326–337.
16. Lim, K.T., Buinett, M., Iotov, M., McClurg, K.B., Vaideli, N., Dasgupta, S., Taylor, N. and Goddaard, III. W.R., *Molecular dynamics for very large systems on massively parallel computer: The MPSim program*, *J. Comput. Chem.*, 18 (1997)501–521
17. Cornell, W.D., Cieplak, P., Bayly, C.I., Gould. I.R., Merz, Jr., K.M., Ferguson, D.M., Spellmeyer, D.M., Fox, T., Caldwell, J.W. and Kollman, P.A., *A second generation force field for the simulation of proteins, nucleic acids, and organic molecules*. *J. Am. Chem. Soc.*, 117 (1995) 5179–5197.
18. Kitson, D.H., Avbelj. F., Moults, J., Nguyen, D.T., Mertz, J.E., Hadzi, D. and Hagler, A.T., *On achieving better than 1 Å accuracy in a simulation of a large protein: Steptomyces griseus protease A*, *Proc. Natl. Acad. Sci. USA.*, 90 (1993) 8920–8924.
19. Halgren, T., *Merck molecular force-field: I. Basics. the scope parameterization and performance of MMFF94*, *J. Comput. Chem.*, 17 (1996)490–511.
20. Jorgensen, W.L., Maxwell, D.S. and Tirade-Rives, J., *Development and testing of the OPLS all-atom force-field on conformational energetics and properties of organic liquids*, *J. Am. Chem. Soc.*, 118(1996) 11225–11236.
21. van Gunsteren, w., *Computer simulation of biomolecular systems: Overview of time-saving techniques*, *AIP Conf. Proc.*, 239 (1991) 131–146.
22. Brunne. R.M., Berndt. K.D., Guntert, P., Wuthrich, K. and van Gunsteren, W.F., *Structure and internal dynamics of the bovine pancreatic trypsin inhibitor in aqueous solution form long-time molecular dynamics simulations*, *Proteins: Struct. Funct. Genet.*, 23 (1995) 49–62.
23. Weerasinghe, S., Smith. P.E. and Pettitt, B.M., *Structure and stability of a model pyrimidine–purine–purine DNA triple helix with a GC.T mismatch by simulation*, *Biochemistry*. 34 (1995) 16269–78.
24. York, D.M., Vlodawer, A., Pedersen, L.G. and Darden. T.A., *Atomic-level accuracy in simulations of large protein crystals*, *Proc. Natl. Acad. Sci. USA.*, 91 (1994) 8715–8718.
25. Elber, R.E., *Novel methods for molecular dynamics simulations*, *Current Opin. Struct. Biol.*, 6 (1996) 232–235.

26. van Gunsteren, W.F. and Berendsen, H.J., *Computer simulation of molecular dynamics: methodology, applications, and perspectives in chemistry*, Angew. Chem., Int. Ed. Engl., 29 (1990) 992–1023.
27. Brunger, A.T. and Karplus, M., *Molecular dynamics simulations with experimental restraints*, Acc. Chem. Res., 24 (1991) 54–61.
28. Brunger, A.T., *X-PLOR: A system for crystallography and NMR*, Manual version 3.1. Yale University Press, New Haven, CT.
29. Kirkpatrick, S., Gelatt, C.D., Jr. and Vecchi, M.P., *Optimisation by simulated annealing*, Science, 220 (1983) 671–680.
30. Amadei, A., Linssen, A.B. and Berendsen, H.J., *Essential dynamics of proteins*, Proteins: Struct. Funct. Genet., 17 (1993) 412–425.
31. Huber, T., Torda, A.E. and van Gunsteren, W.F., *Local elevation: A method for improving the searching properties of molecular dynamics simulations*, J. Comput.-Aided. Mol. Design., 8 (1994) 694–708.
32. Simmerling, C. and Elber, R.E., *Hydrophobic collapse in a cyclic hexapeptide computer simulation of CHDLFC and CAAAAC in water*, J. Am. Chem. Soc., 116 (1994) 2534–2547.
33. Jagannadh, B., Kunwar, A.C., Thangavelu, R.P. and Osawa, E., *New technique for conformational sampling of cyclic molecules using the AMBER force field: Application to 18-crown-6*, J. Phys. Chem., 100 (1996) 14339–14342.
34. Fronczek, F.R., Hemingway, R.W., McGraw, C.W., Steyberg, J.P., Helfer, C.A. and Mattice, W.L., *Crystal structure, conformational analysis, and molecular dynamics of tetra-O-methyl-(+)-catechin*, Biopolymers, 33 (1993) 275–282.
35. Porter, L.J., Wong, R.Y., Benson, M., Chang, B.G., Wiswanadhan, V.N., Gandour, R.D. and Mattice, W.L., *conformational analysis of flavans: Proton NMR and molecular mechanical (MM2) studies of the benzopyran ring of 3',4',5',7'-tetrahydroxyflavan-3-ols: The crystal and molecular structure of the procyandin (2R, 3S, 4R)-3',4',5,7-tetramethoxy-4(-2,4,6-trimethoxyphenyl)-flavan-3-ol*, J. Chem. Res., 3 (1986) 86–87.
36. Bryant, S.D., Attila, M., Salvadori, S., Guerrini, R. and Lazarus, L.H., *Molecular dynamics conformations of deltorphin analogues advocate delta opioid binding site models*, Pept. Res. 7 (1994) 175–184.
37. Collins, S. and Hruby, V.J., *Prediction of the conformational requirements for binding to the Kappa-opioid receptor and its subtypes: I. Novel alpha-helical cyclic peptides and their role in receptor selectivity*, Biopolymers, 34 (1994) 1231–1241.
38. Folkers, G., Merz, A. and Rognan, D., *CoMFA: Scope and limitations*, In Kubinyi, H., (Ed.) 3D-QSAR: Theory, methods and applications. ESCOM Science Publishers B.V., Leiden, The Netherlands, 1993, pp.583–618.
39. Hopfinger, A.J. and Kawakami, Y., *QSAR analysis of a set of benzothioopyranindazole anti-cancer analogs based upon their DNA interaction properties as determined by molecular dynamics simulation*, Anti-Cancer Drug Design, 7 (1992) 203–17.
40. Langer, T. and Wermuth, C.G., *Inhibitors of prolyl endopeptidase Characterization of the pharmacophoric pattern using conformational analysis and 3D-QSAR*, J. Comput.-Aided Mol. Design. 7 (1993) 253–262.
41. Rognan, D., Reddehase, M.J., Koszinowski, U.H. and Folkers, G., *Molecular modeling of an antigenic complex between a viral peptide and a class I major histocompatibility glycoprotein*, Proteins: Struct. Funct. Genet., 13 (1992) 70–85.
42. Folkers, G., Merz, A. and Rognan, D., *CoMFA as a tool for active site modelling*, In Wennuth, C.G. (Ed.) Trends in QSAR and molecular modelling, ESCOM Science Publishers. Leiden. The Netherlands, 1993, pp. 233–244.
43. Thibaut, U., Folkers, G., Klehe, G., Kubinyi, H., Merz, A. and Rognan, D., *Recommendations for CoMFA studies and 3D QSAR publications*, Quant. Struct.-Act. Relat., 13 (1994) 1.
44. Ringe, D. and Petsko, G.A., *Mapping protein dynamics by X-ray diffraction*, Prog. Biophys. Mol. Biol., 45 (1985) 197–235.
45. Bjorkman, P.J.J., Saper, M.A., Samraoui, B., Bennet, W.S., Strominger, J.L. and Wiley, D.C., *Structure of the human class I histocompatibility antigen. HLA-A2*, Nature, 329 (1987) 506–512.
46. Sussman, J.L., Harel, M., Frolow, W., Oefner, C., Goldman, A., Toker, L. and Silman, I., *Atomic structure of acetylcholinesterase from Torpedo californica: A prototypic acetylcholine-binding protein*, Science, 253 (1991) 872–879.

47. Ripoll, D.R., Faerman, C.H., Axelsen, P.H., Simman, I. and Sussman, J.L., *An electrostatic mechanism for substrate guidance down the aromatic gorge of acetylcholinesterase*, Proc. Natl. Acad. Sci. USA., 90 (1993) 5128–5132.
48. Axelsen, P.H., Harel, M., Silman, I. and Sussman, J., *Structure and dynamics of the active site gorge of acetylcholinesterase Synergistic use of molecular dynamics simulation and X-ray crystallography*, Prot. Sci., 3 (1994) 188–197.
49. Faerman, C., Ripoll, D., Bon, S., Le Feuvre, Y., Morel, N., Massoulié, J., Sussman, J.L. and Silman, I., *Site-directed mutagenesis designed to test the back-door hypothesis of acetylcholinesterase function*, FEBS Lett., 3 (1996) 65–71.
50. Heemels, M.T. and Ploegh, H.L., *Generation, translocation and presentation of MHC class I-restricted peptides*, Annu. Rev. Biochem., 64 (1995) 643–691.
51. Townsend, A., Ohlen, C., Bastin, J., Ljunggren, H.G., Foster, L. and Karre, K., *Association of class I major histocompatibility heavy and light chains induced by viral peptides*, Nature, 340 (1989) 443–448.
52. Madden, D.R., Gorga, J.C., Strominger, J.L. and Wiley, D.C., *The structure of HLA-B27 reveals nonamer self-peptides bound in an extended conformation*, Nature, 353 (1991) 321–325.
53. Rognan, D., Zimmermann, N., Jung, G. and Folkers, G., *Molecular dynamics study of a complex between the human histocompatibility antigen HLA-A2 and the IMP58-66 nonapeptide from influenza protein, virus matrix* Eur. J. Biochem., 208 (1992) 101–113.
54. Madden, D.R., Garboezi, D.N. and Wiley, D.C., *The antigenic identity of peptide/MHC complexes: A comparison of the conformations of five viral peptides presented by HLA-A2*, Cell, 75 (1993) 693–708.
55. Blundell, T., Sibanda, B.L., Sternberg, M.J.E. and Thornton, J.M., *Knowledge-based prediction of protein structures and the design of novel molecules*, Nature, 326 (1987) 347–352.
56. Vedani, A., Zbinden, P. and Snyder, J.P., *Pseudo-receptor modeling: A new concept for the three-dimensional construction of receptor binding sites*, J. Recept. Res., 13 (1993) 163–177.
57. van Gunsteren, W.F. and Berendsen, H.J.C., *Groningen molecular simulation (GROMOS) library manual*, Biomos, Groningen.
58. Kern, P., Brunne, R.M., Rognan, D. and Folkers, G., *A pseudo-particle approach for studying protein-ligand models truncated to their active sites*, Biopolymers, 38 (1996) 619–637.
59. Kern, P., Rognan, D. and Folkers, G., *MD simulations in pseudo-particle fluids: Applications to active-site protein complexes*, Quant. Struct.-Act. Relat., 14 (1995) 229–241.
60. Bash, P.A., Singh, U.C., Langridge, R. and Kollman, P.A., *Free energy calculation by computer simulation*, Science, 236 (1987) 564–568.
61. Kollman, P.A., *Advances and continuing challenges in achieving realistic and predictive simulation of the properties of organic and biological molecules* Acc. Chem. Res., 29 (1996) 461–469.
62. Pealman, D.A. and Connelly, P.R., *Determination of the differential effects of hydrogen bonding and water release on the binding of FK506 to native and Tyr82 → Phe82 FKBP-12 proteins using free energy simulations*, J. Mol. Biol., 248 (1995) 696–717.
63. Miyamoto, S. and Kollman, P.A., *Absolute and relative binding free energy calculations of the interaction of biotin and its analogs with streptavidin using molecular dynamics/free energy perturbation approaches*, Proteins: Struct., Funct., Genet., 16 (1993) 226–245.
64. Bayly, C.I. and Kollman, P.A., *Molecular dynamics and free energy calculations on the peculiar bimodal alkali ion selectivity of an 8-subunit cavitand*, J. Am. Chem. Soc. 116 (1994) 697–703.
65. Branda, N., Wyler, R. and Rebek, J., Jr., *Encapsulation of methane and other small molecules in a self-assembling superstructure*, Science, 263 (1994) 1267–1268.
66. Zacharias, M., Straatsma, T.P., McCammon J.A. and Quiocho, F.A., *Inversion of receptor binding preferences by mutagenesis: Free energy thermodynamic integration studies on sugar binding to L-arabinose binding*, Biochemistry, 32 (1993) 7428–7434.
67. Orozco, M., Tirado-Rives, J. and Jorgensen, W.L., *Mechanism for the rotamase activity of FK506 binding protein from molecular dynamic simulations*, Biochemistry, 32 (1993) 12864–12874.
68. Lau, F.T. and Karplus, M., *Molecular recognition in proteins: Simulation analysis of substrate binding by a tyrosyl-tRNA synthetase mutant*, J. Mol. Biol., 236 (1994) 1049–1066.
69. Reddy, M.R., Varney, M.D., Kalish, V., Viswanadhan, V.N. and Appelt, K., *Calculation of relative differences in the binding free energies of HIV1 protease inhibitors: A thermodynamic cycle perturbation approach*, Med. Chem., 37 (1994) 1145–1152.

70. Singh, S.B., Wemmer, D.E. and Kollman, P.A., *Relative binding affinities of distamycin and its analog to d(CGCAAGTTGGC).d(GCCAACTTGCG): Comparison of simulation results with experiment*, Proc. Natl. Acad. Sci. USA., 91 (1991) 7673–7677.
71. Wlodek, S.T., Antosiewicz, J., McCammon, J.A., Straatsina, T.P., Gilson, M.K., Briggs, J.M., Humblet, C. and Sussman, J.L., *Binding of tacrine and 6-chlorotacrine by acetylcholinesterase*, Biopolymers, 38 (1996) 109–117.
72. Reddy, M.R., Viswanadhan, V.N. and Weinstein, J.N., *Relative differences in the binding free energies of human immunodeficiency virus I protease inhibitors: A thermodynamic cycle-perturbation approach*, Proc. Natl. Acad. Sci., 88 (1991) 10287–10291.
73. Jones, J.P., Trager, W.F. and Carlson, T.J., *The binding and regioselectivity of reaction of (R)- and (S)-nicotine with cytochrome P-450cam: Parallel experimental and theoretical studies*, J. Am. Chem. Soc., 115 (1993): 381–387.
74. Gillner, M., Bergman, J., Cambillau, C., Alexandersson, M. and Fernstrom, B., *Interactions of indolo[3,2-b]carbazoles and related polycyclic aromatic hydrocarbons with specific binding sites for 2,3,7,8-tetrachlorodibenzo-p-dioxin in rat liver*, Mol. Pharmacol., 44 (1993) 336–45.
75. Rao, B.G., Kim, E.E. and Murcko, M.A., *Calculation of solvation and binding free energy differences between VX-478 and its analogs by free energy perturbation and AMSOL methods*, J. Comput.-Aided Mol. Design, 10(1996) 23–30.
76. Miyamoto, S. and Kollman, P.A., *What determines the strength of noncovalent association of ligands to proteins in aqueous solution?*, Proc. Natl. Acad. Sci. USA, 90 (1993) 8402–8406.
77. Jorgensen, W.L., Buckner, J.K., Boudon, S. and Tirado-Rives, J., *Efficient computation of absolute free energies of binding by computer simulation: Application to the methane dimer in water*, J. Chem. Phys., 89 (1988) 3741–3746.
78. Mark, A. and van Gunsteren, W.F., *Decomposition of the free energy of a system in terms of specific interactions: Implication for theoretical and experimental studies*, J. Mol. Biol., 240 (1994) 167–176.
79. Boresch, S., Achonitis, S. and Karplus M., *Free energy simulations: The meaning of the individual contributions from a component analysis*. Proteins: Struct., Funct. Genet., 20 (1994) 25–33.
80. Helms, V. and Wade, R.C., *Thermodynamics of water mediating protein-ligand interactions in cytochrome P450cam: A molecular dynamics study*. Biophys. J., 69 (1995) 810–824.
81. Poornima, C.S. and Dean, P.M. Hydration in drug design: 1. *Multiple hydrogen-bonding features of water molecules in mediating protein-ligand interactions*, J. Comput.-Aided Mol. Design, 9 (1995) 500–512.
82. Appelt, K., Bacquet, R.J., Bartlett, C.A., Booth, C.L., Freer, S.T., Fuhry, MA., Gehring, M.R., Herrman, S.M., Howland, E.F., Janson, C.A., Jones, T.R., Kan, C.-C., Kathardekar, V., Lewis, K.K., Marzoni, G.P., Matthews, D.A., Mohr, C., Moomaw, E.W., Morse, C.A., Oatley, S.J., Ogden, R.C., Reddy, M.R., Reich, S.H., Schoettlin, W.S., Smith, W.W., Varney, M.D., Villafranca, J.E., Ward, R.W., Webber, S., Webber, S.E., Welsh, K.M. and White, J., *Design of enzyme inhibitors using iterative protein crystallographic analysis*. J. Med. Chem., 7 (1991) 1925–1934.
83. Lam, P.Y., Jadhav, P.K., Eyermann, C.J., Hodge, C.N., Ru, Y., Bachelier, L.T., Meek, J.L., Otto, M.J., Rayner, M.M., Wong, Y.N., Chang, C.H., Webber, P.C., Jackson, D.A., Sharpe, T.R. and Erickson-Viltanen, S., *Rational design of potent, bioavailable, nonpeptide cyclic ureas as HIV protease inhibitors*. Science, 263 (1994) 380–384.
84. Blundell, T., Hubbard, R. and Weiss, MA., *Structural biology and diabetes mellitus: Molecular pathogenesis and rational drug design*, Diabetologia, 35 (1992) 69–76.
85. Rognan, D., Scapozza, L., Folkers, G. and Daser, A., *Molecular dynamics simulation of MHC-peptide complexes as a tool for predicting potential T cell epitopes*, Biochemistry, 33 (1994) 11476–11485.
86. Rognan, D., Krebs, S., Kuonen, O., Lamas, J.R., López de Castro, J.A. and Folkers, G., *Fine specificity of antigen for two class I major histocompatibility protein Alleles (B 2705 and B 2703) differing in one amino acid*, J. Comput.-Aided Mol. Design, 11 (1997) 463–478.
87. Kognan, D., Scapozza, L., Folkers, G. and Daser, A., *Rational design of nonnatural peptides as high-affinity ligands for the HLA-B\*2705 human leukocyte antigen*, Proc. Natl. Acad. Sci. USA, 92 (1995) 753–757.



88. Scopozza, L., Rognan, D., Folkers, G. and Daser, A. *Molecular dynamics and structure-based drug design for predicting non-natural nonapeptide binding to a class I MHC protein*, Acta Cryst., D51 (1995) 541–549.
89. Rognan, D., *Molecular modeling of protein-peptide complexes: Application to major histocompatibility proteins*, Habilitationsschrift. Eidgenossische Technische Hochschule (ETH), Zurich, 1997.
90. Jardtetzky, T.S., Lane, W.S., Robinson, R.A., Madden, D.R. and Wiley, D.C., *Identification of self-peptides bound to purified HLA-B27*, Nature, 353 (1991) 326–329.
91. Daser, A., Henning, U. and Henklein, P., *HLA-B27 binding peptides derived from the 57kD heat shock protein of Chlamydia trachomatis: Novel insights into the peptide binding rules*, Mol. Immunol., 31 (1994)331-336.
92. Bouvier, M. and Wiley, DC., *Importance of peptide amino and carboxy termini to the stability of MHC class I molecules*. Science, 265 (1994) 398–402.
93. Szewezuk, Z., Gibbs, B.F., Yue, S.H., Purisima, E., Zdanov, A., Cygler, M. and Konishi, Y., *Design of a linker for trivalent thrombin inhibitors: Interaction of the main chain of the linker with thrombin*, Biochemistry, 32, (1993) 3396–3404.
94. Connolly, M.J., *Analytical molecular surface calculation*, J. Appl. Crystallogr., 16 (1983) 548-558.
95. Rognan, D. *AS and immunotherapy: Future considerations*, In, Lopez-Larrea, C. (Ed.) HLA-B27 in the development of spondyloarthropathies, R.G. Landes Co., Georgetown, MA, 1997, pp. 235–251.
96. M.D, Paulsen and Ornstein R.L. *Predicting the product specificity and coupling of cytochrome P450cam*, J. Comput.-Aided Mol. Design, 6 (1992) 449-460.
97. Bass, M.B. and Orustein, R.L., *Substrate specificity of cytochrome P450cam for L- and D- norcamphor as studied by molecular dynamics simulations*, J. Comput. Chem., 14 (1993) 541–548.
98. Bemis, G.W., Carlson-Golab, G. and Katzenellenbogen, J.A., *A molecular dynamics study of the stability of chymotrypsin acyl enzymes*, J. Ani. Chem. Soc., 114(1992) 570–578.
99. Taylor, N.R. and von Itzstein, M., *Molecular modeling studies on ligand binding to sialidase from influenza virus and thw mechanism of catalysis*, J. Med. Chem., 34 (1994) 616–624.
100. Helms, V., Deprez, E., Gill, E., Barret, C., Hui Bon Hoa, G. and Wade, R.C., *Improved binding of Cytochrome P450cam substrates analogues designed to fill extra space in the substrate binding pocket*, Biochemistry, 35 (1996) 1485–1499.
101. Boisgérault, F., Tieng, V., Stolzenberg, M.C, Dulphy, N., Khalil, I., Tamouza, R., Charron, D. and Toubert, A., *Differences in endogenous peptides presented by HLA-B\*2705 and B\*2703 allelic variants: Implications for susceptibility to spondyloarthropathies*, J. Clin. Invert., 98 (1996) 2764–2770.
102. Klebe, G., *The use of composite crystal-field environments in molecular recognition and the de novo design of protein ligands*, J. Mol. Biol., 237 (1994) 212–235.
103. Mills, J.E.J. and Dean, P.M., *Three-dimensional hydrogen-bond geometry and propability information from a crystal survey*, J. Comput.-Aided Mol. Design, 10(1996) 607–622.
104. Pearlman, D.A., Case, D.A., Caldwell, J.C., Ross, W.S., Cheatham, T.E., III, Ferguson, D.E., Seibel, G.L., Singh, C., Weiner, P.K. and Kollmann, PA., *AMBER 4. 1*. University of California, San Francisco, CA, U.S.A., 1995.
105. Lewis, R.A. and Leach, A.R., *Current methods for site-directed structure generation*, J. Comput.-Aided Mol. Design, 8 (1994) 467–475.
106. Pearlman, D.A. and Murcko, M.A., *CONCEPTS: New dynamic algorithm for de novo drug suggestion*, J. Comput. Chem., 14 (1993) 1184–1193.
107. Pearlman, D.A. and Murcko, M.A., *CONCERTS: Dynamic connection of fragments as an approach to de novo ligand design*, J. Med. Chem., 39 (1996) 1651–1663.
108. Sandak, B., Nussinov, R. and Wolfson, H.J., *An automated computer vision and robotics-based technique for 3-D flexible biomolecular docking and matching*. Comput. Appl. Biosci., 11 (1995) 87–99.
109. Rarey M., Kramer B., Lengauer T. and Klebe G., *A fast flexible docking method using an incremental construction algorithm*, J. Mol. Biol., 261 (1996) 470–489.
110. Morris, G.M., Goodsell, R., Huey, R. and Olson, A.J., *Distributed automatic docking of flexible ligands to proteins: parallel applications of AutoDock 2.4*, J. Comput.-Aided Mol, Design, 4 (1996) 293–304.
111. Di Nola, A., Roccatano, D. and Berendsen, H.J.C., *Molecular dynamics simulation of the docking of substrates to proteins*. Proteins: Struct., Funct. Genet., 19 (1994) 174–182.

112. Tirado-Rives, J. and Jorgensen, W.L., *Viability of molecular modeling with pentium-based Pcs*, J. Comput. Chem., 17 (1996) 1385–1386.
113. Jorgensen, W.L. and Tirado-Rives, J., *Monte Carlo vs molecular dynamics for conformational sampling*, J. Phys. Chem., 100 (1996) 14508–14513.

**This Page Intentionally Left Blank**

**Part III**

# **Pharmacophore Modelling and Molecular Similarity**

**This Page Intentionally Left Blank**

# Bioisosterism and Molecular Diversity

Robert D. Clark\*, Allan M. Ferguson, and Richard D. Cramer

*Tripos, Inc., 1699 S. Hanley Road, St. Louis, MO 63144, U.S.A.*

## 1. Introduction

Assessing quantitatively the similarity (or dissimilarity) between one molecule and another is central to both QSAR and molecular diversity analysis. When either endeavor is involved in pharmaceutical or agrochemical lead discovery and development, the only similarity that 'really' matters is similarity in biochemical (and/or physiological) properties. Yet, by definition, biochemical similarity cannot itself be measured before compounds are actually in hand, so some other way must be found to identify functionalities which generally 'look' similar to receptor and enzymatic ligand binding sites, even though they do not share substructural motifs — i.e. which are bioisosteric [1].

A cartographic metaphor [2] for the discovery and development of biologically active small molecules is useful in thinking about the relationship between QSAR and diversity analysis. As so ably described in chapters elsewhere in this volume, QSAR involves using compounds of known biochemical activity to find a combination of descriptors which can be used reliably to predict the activity of related compounds which have not yet been synthesized and to help better understand the underlying biochemistry. Hence, QSAR is fundamentally interpolative, a matter of surveying islands once their location is known to identify the highest peak on the island.

Molecular diversity analysis, in contrast, entails surveying an ocean to find out where the islands of activity are — and where they probably are not. Descriptor values for properties which might be useful as charting reference points generally range far more widely than would be appropriate for a QSAR study, making diversity analysis an essentially extrapolative exercise which is always looking to the horizon. The key is finding a descriptor 'lens' through which the different kinds of islands are well-separated, clearly defined and easy to distinguish.

Computational chemists have quite naturally looked to QSAR as a source of 'good' descriptors for quantitatively evaluating bioisosteric similarity. The classical whole-molecule ( $\log P$ , MR, etc.) and fragment ( $\pi$ ,  $\sigma_p$ , F, etc.) QSAR parameters developed by Hansch and Fujita [3], among others, have been employed for this purpose [4,5], but have a limited range of values and tend to be highly correlated with one another, which makes it difficult to capture satisfactorily the diversity of large structural datasets.

Steric, electrostatic and hydrophobic molecular fields have also been very successfully employed in QSAR via Comparative Molecular Field Analysis [6–8]. Extension of CoMFA to diversity analysis is particularly appealing because the information in molecular fields is localized and of high dimensionality. Binding sites in enzymes and receptors tend to present quite variegated interaction surfaces to their ligand. Therefore,

---

\* To whom correspondence should be addressed.

it is reasonable to expect that descriptors like CoMFA fields, which can differentiate between molecules in many different ways simultaneously, stand a better chance of ‘capturing’ biochemically meaningful distinctions. This is one explanation for the correlation between the dimensionality of descriptors and their suitability as diversity descriptors [9].

An added appeal of CoMFA is its ability to generalize across substitution patterns and atom types. This occurs because differences between fields are quantitative but sensitive to the structural context, whereas differences between atoms are qualitative. Hence, the field for a methylamino group is ‘between’ those of an ethyl group and a methoxy group, and the field near a ketone resembles that around a difluoromethylene. Since an enzyme or receptor binding site interacts with each ligand’s molecular orbitals, not with its atomic nuclei, it is reasonable to expect that the molecular fields involved should correlate well with biological activity. The success of CoMFA as a tool for deriving quantitative structure–activity relationships [7] is a testament to the validity of this hypothesis.

## 2. Theoretical Considerations

### 2.1. Information density

In general, the databases with which diversity analysis deals nearly always originate from molecular connectivity specifications alone (e.g. SMILES or SLN strings [10]) which are essentially two dimensional. Several excellent computer programs exist which will convert such connectivity data into 3D molecular graphs which can, in turn, be submitted to minimization routines to identify low-energy conformations. None of these operations actually adds any information to that contained in the original connectivity, and recent work indicates that this absence of added information is very real in many situations [11]. How, indeed, can any ‘derived’ 3D descriptors such as molecular fields add anything to diversity analysis over that which can be obtained from 2D fingerprints?

One way is by efficiently utilizing information about chirality which is embedded in the line notation, either explicitly or implicitly. Explicitly specified chirality encodes higher-order (‘2.5D’) information which is not captured in 2D fingerprints, but which is captured in 3D structures and, thence, in molecular fields. Certain implicit ‘chiral’ information can also be captured directly, such as the sensible disposition of equatorial and axial substituents on symmetrical, achiral molecules.

Pharmacophore multiplet fingerprints [12] attempt to do this by including the distance between analogous groups, but even quite restricted pharmacophoric classes require hundreds of thousands of bits to encode a small part of the information contained in a molecular field built from a few thousand lattice points. Moreover, in our hands, the most important information in such fingerprints is best appreciated by visual inspection: compounds with similar biochemical properties tend to form similar patterns in the pharmacophore multiplet space, a relationship which often fails to show up in quantitative bitwise comparisons.

Incorporating pharmacophore definitions into 2D fingerprints can capture some of this information effectively, but cannot capture subtle biochemical effects (distinctions between amide and carbamate nitrogens or similarities between pyrimidine analogs, for example) except as special cases. Molecular fields, on the other hand, can generalize efficiently across a range of structurally diverse but biologically equivalent functional groups — i.e. bioisosteres.

## 2.2. Encoding conformational information

CoMFA is an attractive way to compare molecules, but each conformer of any given molecule has a different molecular field. Most small molecules of interest as potential drugs are flexible and, hence, have a fairly large number of distinct conformations lying within a reasonable energy range of the ground state. Indeed, solvent effects and necessary approximations in making energetics calculations can make identification of ‘the’ ground state conformation problematic at best, and a purely theoretical exercise at worst.

One way to approach this dilemma is to average the molecular field across a large number of energetically reasonable conformations. For compounds which have only a few related but distinctive low-energy conformations, this approach will certainly enhance the information content of the molecular field as a descriptor. Unfortunately, such molecules are the exception, not the rule, and averaging across conformers is likely to blur meaningful distinctions between molecules at least as often as it illuminates significant similarities.

Alternatively, one can finesse the issue by identifying that conformation for one molecule which produces a molecular field most similar to that of another to which it is being compared [13]. Such field fitting can become computationally intractable, however, unless one molecule is fixed in a reference conformation. Given a reference molecule and conformation, this approach becomes a variation on the theme of fitting probe molecules to some enzyme or receptor binding site model. Though docking methods are indeed powerful tools for cases where a known, relatively rigid binding site structure is known, it is not very useful in designing libraries for general lead discovery — i.e. when targeted against a ‘universal receptor’. Related work using a panel of binding-sites, whether virtual [14] or realized [15], may prove more generally useful, however.

## 2.3. Characteristic conformations

A third alternative, which is the one we have found most fruitful, is to ‘project’ the four-dimensional conformational space into three dimensions by using robust, rule-based algorithms to generate *characteristic* conformations. Doing so brings the fields for similar and homologous substructures into register, which enhances the ability to detect similarity and, at the same time, enhances field differences between functionally distinct substructures.

Moreover, suitably formulated rules make it possible to encode the very potential for flexibility into molecular fields. Consider two compounds, one cyclic and one open chain, but otherwise of similar structure. Even if the nominal ground-state



conformations of the two are identical, their potentials as ligands are likely to be very different when compared across a range of receptors. A 3D building rule which sets bond torsions to all trans wherever reasonable makes the corresponding molecular fields become readily distinguishable.

Characteristic conformations also serve to emphasize similarities between molecules which may be attenuated or lost altogether when averaged across conformers. Biaryl ethers are a case in point. The aryl planes in these compounds are characteristically tilted somewhat from the perpendicular with respect to one another, so that each low-energy conformer is asymmetrical about the central oxygen. Yet energy differences between the various torsions about the ether bridge and rotations about each ether bond are negligible. If 'true' energy minima were used for CoMFA, differences between molecules would be scattered across the entire field. Choosing an arbitrary but consistent starting orientation focuses differences onto specific substructures, so that a consistent QSAR can be obtained [e.g. 16]. Characteristic conformations, such as those obtained from CONCORD [17], bring an analogous benefit as a starting point for diversity analysis.

### 3. Topomeric CoMFA

Even where knowledge of the 'true' conformational ground state is available, there are an infinite number of energetically equivalent rotations and translations that could be applied to one molecular field with respect to another when making a comparison between the two. In cases where a series of compounds include a common core, this 'alignment problem' can be dealt with by superimposing that common core.

The core atom bearing the substituent (core atom) defines the origin of the field, and the bond between the core and the fragment establishes the positive pole of the x-axis. The positive-positive quadrant of the xy plane is defined by the core atom, the fragment atom it is bonded to (the valence atom) and the first atom of the biggest substituent on the valence atom [18]. A characteristic conformation is then generated by working out from the common core, applying a series of rules to establish characteristic torsional and chirality states as one goes.

Such rules become more ambiguous as one moves away from the core, so field contributions from each atom in a substituent are attenuated by a factor of  $0.85^n$ , where n is the number of rotatable bonds separating that atom from the core.

Topomeric alignments for several trichloroethyl derivatives are shown in the top row of structures in Fig. 1. Note that dimethylamino and isopropyl groups are isosteric in this context, whereas 2-propenyl is more similar to the thiomethyl congener. In the series of trichloroacyl derivatives shown at the bottom of Fig. 1, on the other hand, dimethylamino is isosteric with 2-propenyl, which reflects conjugation with the acyl double bond. This dependence on context is a particularly useful result of using the rule-based CONCORD for initial 3D model building. In some cases, this requires a trade-off between useful distinctions between molecules and arguably artificial ones — compare the alignments of the isopentane (top) and the isopropyl ketone (bottom) in Fig. 1, for example. As discussed below with respect to validation, however, this is not a great problem for diversity analysis.

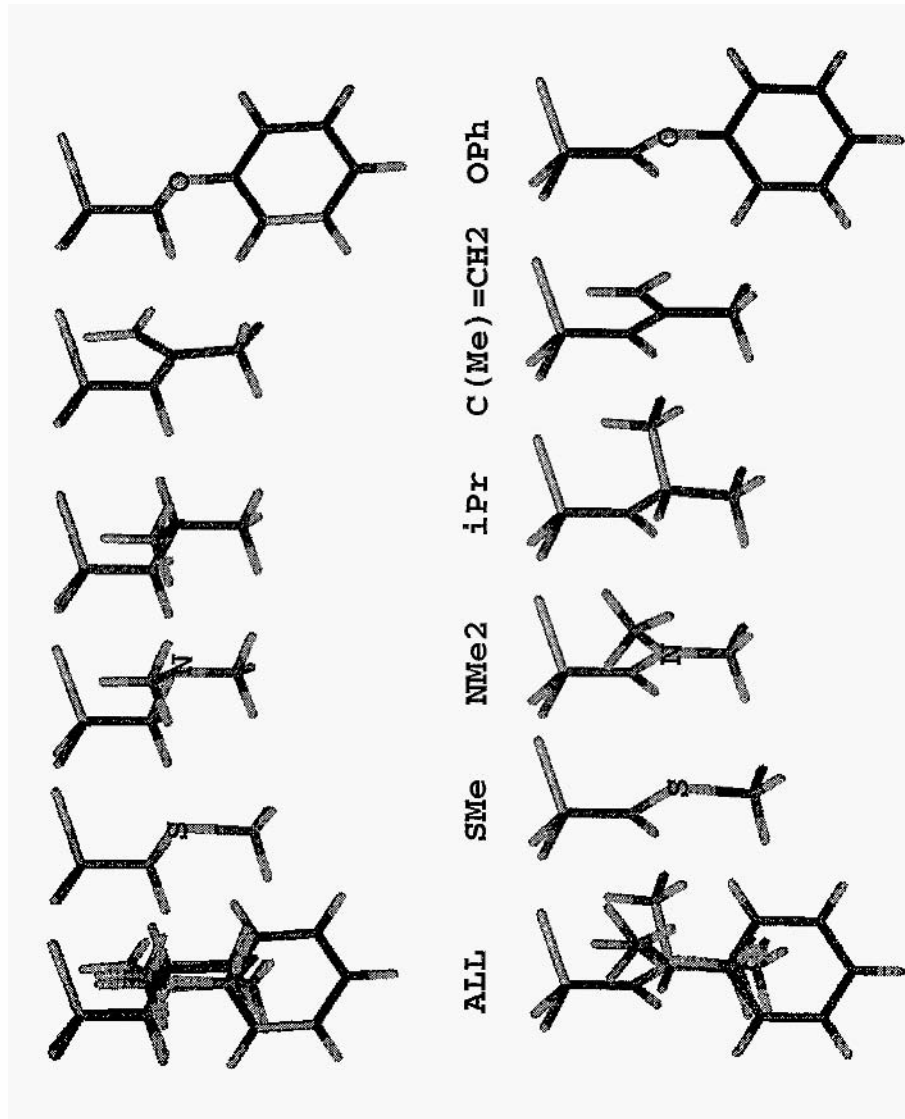


Fig. 1. Topomeric alignment of a series of trichloroethyl derivatives (top) and their trichloroacetyl counterparts (bottom). All analogs are superimposed at the left of each row. Heteroatoms are labeled in the substituents but omitted from the common core structures CONCORD [17] structures were aligned using Selector in SYBYL 6.3 [20].

A somewhat related method for optimizing the diversity of a set of side chains has been described by Chapman [19]. Here, similarity is summed across pairwise distances between each atom in the 'query' and all atoms in the target molecule. A 'soft threshold' weighting function is used, and each of a set of 'diverse' conformers is considered.

### 3.1. Applications

Topomeric CoMFA is used to cluster each class of reactant (e.g. nucleophile, electrophilic reagent, dienophile, etc.) whenever a new class of reaction is considered for inclusion in the Optiverse screening library [2,20]. Selection of a reagent from each cluster gives a representative diverse set of reagents which carry on through subsequent stages of the design. Candidate products are then screened for diversity on the basis of their 2D fingerprints. This use of both metrics together exploits their complementarity and enhances the isobiosteric diversity of the database [2].

Similarity searches can now be run in the ChemSpace virtual library [20] based on topomeric fields. This capability is proving itself very useful in identifying novel isosteric substituents for ongoing development programs. In addition, topomeric CoMFA has been used to evaluate disparate structures as potential diverse cores or scaffolds for combinatorial libraries, with quite satisfying results (R.D. Cramer, unpublished).

## 4. Inertial Field Orientation (IFO-CoMFA)

In many series of interest, compounds lack an identifiable common core. In some combinatorial libraries, a nominal common core exists, but serves only as a linker between functional moieties. In either case, an alternative to topomeric alignments is required if molecular fields are to be of use as diversity metrics. To this end, inertial field orientation (IFO-CoMFA) was introduced into Selector [20] for SYBYL 6.3 [21].

Alignments based on molecular moments of inertia have been successfully used to predict HPLC retention times of polynuclear aromatic hydrocarbons (PAHs) [22]. In this approach, the field coordinate origin is set to the molecular center of mass, and that molecular axis about which the moment of inertia is smallest defines the x axis for the field. The y axis is defined by that molecular axis perpendicular to the first which has the smallest moment of inertia.

That implementation of the technique is not directly applicable to diversity analysis, however. For one thing, there is no good way to differentiate one end of the x axis from the other, or one end of the y axis from the other. As a result, each molecule has four equivalent inertial alignments. This did not interfere with the analyses by Welsh et al. because they were working with flat, symmetrical molecules and highly isotropic types of interactions distributed across the entire molecular surface [22,23].

A second problem is that inertial field orientations which incorporate atomic masses are overly sensitive to replacement of hydrogen by fluorine, chlorine or bromine. Halogenation, or exchange of one halogen for another, has a tremendous effect on molecular moments of inertia. Hence, such homologous substitutions will introduce unreasonably large CoMFA dissimilarities if weighted moments are used for orientation.

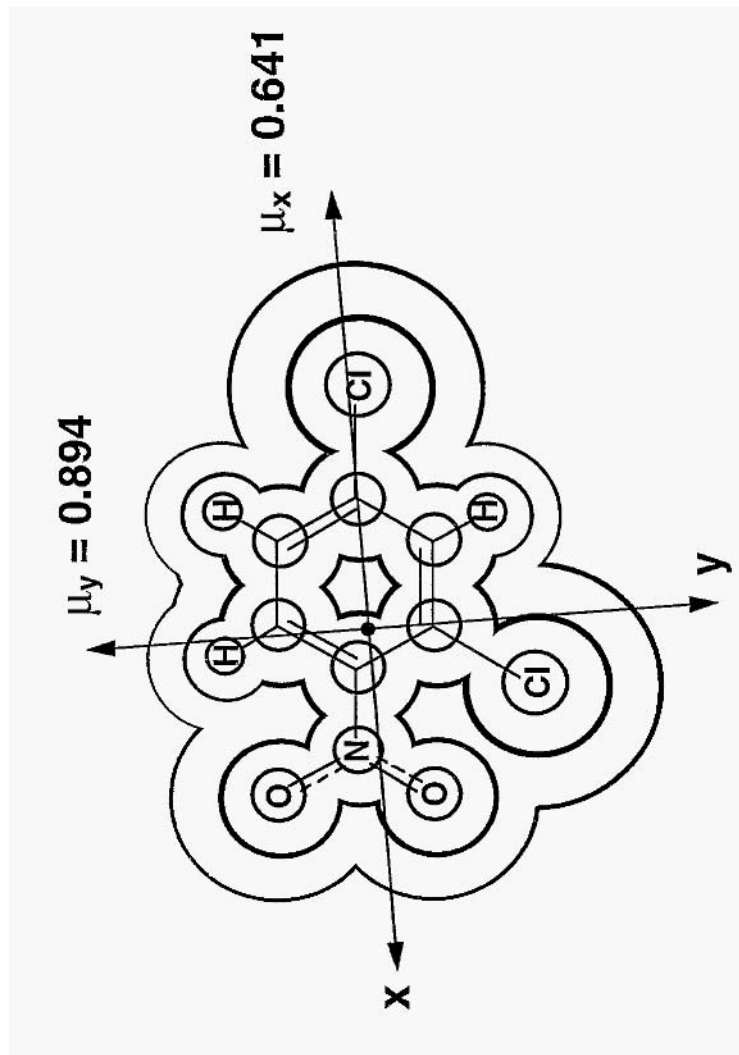


Fig. 2. Schematic representation of the isopotential shells constituting the molecular steric field 2,4-dichlorobenzene. Arrows indicate the common principal axes for the molecular graph and the isopotential shells; dipole moments along each axis are based on changes calculated using the GAST-HUCK option in SYBYL 6.3 [20].

To avoid these problems, we have turned to ‘inertial’ orientation of the steric field, as opposed to inertial orientation of the molecule itself. This can be most simply accomplished by using the principal axes of the *unweighted* molecular graph [21] as defining coordinate axes for the molecular fields. That the principal axes of the (unweighted) molecular graph are valid indicators of the principal axes of the steric field is not immediately obvious, but may be appreciated from the schematic in Fig. 2. If the steric field is thought of as a series of shells built up from summed spherically symmetrical functions about each atom, each shell will have the same principal axes. Hence, the principal axes of the 0-diameter shell —i.e. the unweighted graph—coincide with the principal axes of the steric field as a whole.

The principal axes define four (generally distinct) possible orientations for the field. We opt to use the one for which the components of the dipole moment along the x and y axes are most positive. Logically, this can be thought of as aligning each molecule unambiguously with itself, and is akin to the rationale behind characteristic conformations outlined above. Incorporating the dipole moments along each principal axis allows the overall electrostatics of the molecule to contribute to the orientation of the field, so that nonlinear molecules with distinct distributions of polar functionalities can be more readily distinguished.

Other, related approaches are also in the literature. One can, for example, align randomly sampled isosteric surfaces by their principal axes, then pick from among the four degenerate solutions for each of 20–30 conformers the one which maximizes the intersection volume [24]. In this case, similarity is measured in terms of a Tanimoto coefficient calculated from the intersection and union volumes between the (fixed) query and the target molecules.

#### 4.1. Example

Evolutionary forces can be expected to favor substrates and ligands which are geometrically and electrostatically asymmetrical. To the extent that this is so, two sterically similar molecules which ‘look’ electrostatically similar with respect to the principal axes of each are expected to be qualitatively similar in biological activity. This is illustrated in Fig. 3: steric and electrostatic IFO-CoMFA fields are shown for the anti-ulcer, histamine H<sub>2</sub> receptor antagonists ranitidine (Zantac®) and cimetidine (Tagamet®). The characteristic conformations used were generated using CONCORD [17] and oriented using the inertial field orientation option in Selector [21].

Steric similarity between molecules can be expressed as a volumetric Tanimoto coefficient [24], in terms of the correlation coefficient between steric fields [25], or as simple energy differences [8]. The steric fields in Fig. 3 differ in energy by 186 kcal/mol when summed across a 2 Å lattice, which is somewhat above the characteristic neighborhood radius of 165 kcal/mol for IFO-CoMFA (R.D. Clark, unpublished). The difference in the electrostatic fields, however, is considerably larger, on the order of 230 kcal/mol. In our hands, electrostatic fields are sometimes too dependent on conformation to discriminate reliably between molecules. It is still the case that molecules with similar enough fields will, indeed, have similar biological activities, but the

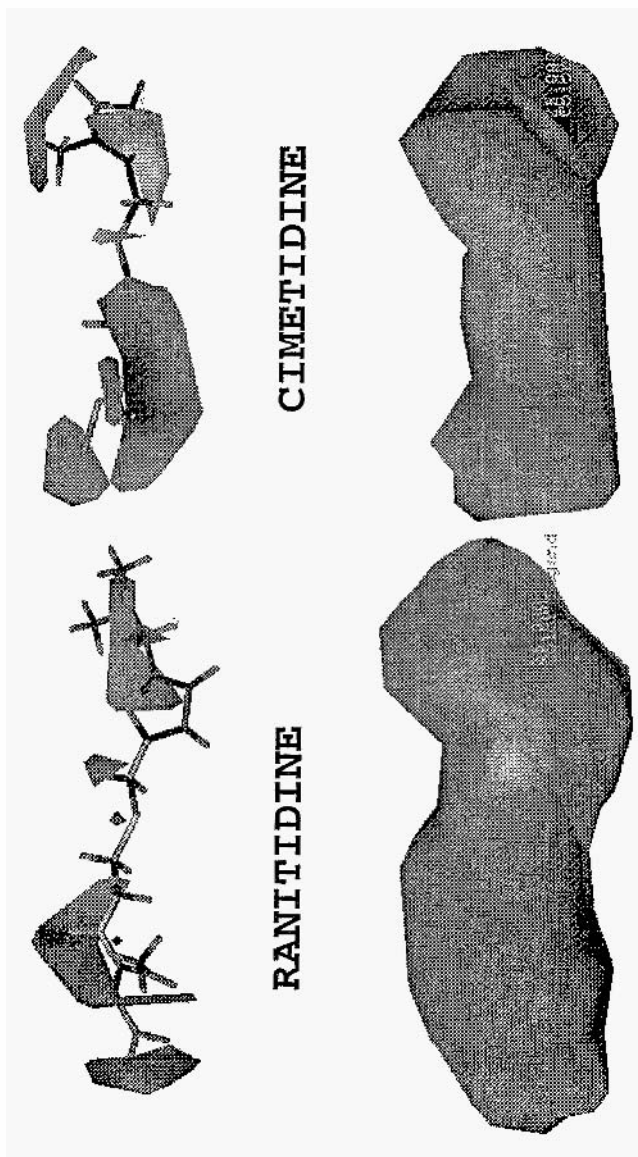


Fig. 3. Comparison of inertially oriented electrostatic (top) and steric (bottom) molecular fields for ranitidine (Zantac®) and cimetidine (Tagamet®). In the top plots, the lighter areas are more electronegative regions. The electrostatic were contoured at  $\pm 20$  kcal/mol (top) and steric fields were contoured at 25 kcal/mol (bottom).

'neighborhoods' defined by that radius may be too sparsely populated to be generally useful. H-bond fields have proven more useful in CoMFA diversity applications [9].

## **5. Validation**

As noted above, QSAR is an essentially interpolative endeavor. But assessing the differences among thousands to millions of disparate compounds, as is routinely contemplated today, is an heroically extrapolative endeavor. The essential complementarity of the two disciplines means that some good QSAR descriptors will, indeed, be good diversity metrics and vice versa, but utility in the one arena is no guarantee of usefulness in the other.

The usefulness of a descriptor in a particular QSAR application is assessed by how much it enhances the ability to interpolatively predict the biochemical activity of compounds whose activity is known but which was not included in the model. The vast majority of compounds in virtual combinatorial libraries will never be made, let alone assayed for biological activity. How, then, is the validity of a new diversity metric to be evaluated? The ultimate judgment will be based upon the track record of each metric and how it has been applied, but taking a 'wait and see' approach can be costly in terms of wasted resources or missed opportunities, or both.

A more pragmatic approach is to use the descriptor in question to cluster a range of compounds into subgroups: 'good' descriptors will tend to put sets of similar compounds into the same group. The assessment of biochemical similarity in such 'categorical validation' studies may rely on activity or inactivity in some single assay [26], or on the pharmacological class of each compound [25]. Given a set of (usually well-separated) activity islands, categorical validation tests the ability of a metric to assign compounds to the correct island [27].

IFO-CoMFA has been validated in the latter context [27], whereas the categorical validation of topomeric CoMFA was by examination of clusters in light of medicinal chemical experience [18]. Such direct assessment of bioisosterism is 'abstract' to some degree, but is altogether appropriate — the development process has relied upon it historically — and can be very powerful. Bioisosteric assessment allows experience with 'real' compounds to be extended to artificial datasets which lack the clumpiness imparted to commercial databases by directed-walk development strategies and patent considerations, and allows meaningful generalizations across receptors and active sites [18].

A complementary approach is to show that proximity in the descriptor space is a sufficient condition for proximity in bioactivity space — i.e. that the descriptor of interest exhibits good 'neighborhood behavior [9]'. This entails plotting differences in bioactivity against dissimilarity in the descriptor space. If the upper-left triangle in such a plot is sparsely populated, similar compounds rarely show large differences in bioactivity. Hence, one can rely on each compound to reliably 'report' on the bioactivity of its neighbors as defined by that descriptor [9]. Such neighborhood plots also provide estimates of the neighborhood 'radius' for a descriptor — i.e. of degree of separation in the descriptor space which corresponds to a difference of 100-fold or less in biological activity [9].

In terms of our island-charting metaphor: islands will be exclusive and locally dense in a good descriptor space, so that if one compound in the neighborhood is active, most of those similar to it will also be active. Each activity may well be spread across several islands, however. Hence, exaggeration of what might well be incidental differences, such as that between the isopentane and the isopropyl ketone in Fig. 1, do not directly compromise the reliability of a metric for diversity analysis. Such fragmentation of 'islands' is, however, undesirable, in so far as it reduces the efficiency of sampling by increasing the number of compounds required to 'cover' a particular range of structural variation.

Neighborhood behavior has been demonstrated for topomeric CoMFA [9] and for whole-molecule CoMFA using a rule-based alignment of ACE inhibitors [25]. IFO-CoMFA also exhibits neighborhood behavior with a similarity radius of 150–200 kcal/mol when summed across a 2 Å field lattice (R.D. Clark, unpublished).

## 6. Conclusion

Molecular fields are important tools for assessing molecular diversity, particularly when one wishes to go beyond 2D substructural similarity and include bioisosteric similarity. Hence, when common cores or scaffolds exist, topomeric CoMFA is a powerful complement to analyses based on similarity of 2D fingerprints. In the absence of common cores, topomeric CoMFA can be used to evaluate the 'diversity potential' of different core scaffolds. Alternatively, IFO-CoMFA is available to provide useful insight into molecular similarities.

## References

1. Fujita, T., *QSAR and database-aided bioisosteric structural transformation procedures as methodologies of agrochemical design*. In Hansch, C. and Fujita, T. (Eds.) Classical and three-dimensional QSAR in agrochemistry, ACS Symposium Series 606, ACS, Washington D.C., 1995, pp. 13–34.
2. Ferguson, A.M., Patterson, D.E., Garr, C.D. and Underiner, T.L., *Designing chemical libraries for lead discovery* J. Biomol. Screening, 1 (1996) 65–73.
3. Hansch, C. and Fujita, T., *A method for the correlation of biological activity and chemical structure*, J. Am. Chem. Soc., 86 (1964), 1616–1626.
4. Shemetulskis, N.E., Dunbar, J.B., Jr., Dunbar, B.W., Moreland, D.W. and Humblet, C., *Enhancing the diversity of a corporate database using chemical database clustering and analysis*, J. Comput.-Aided Molec. Design, 9 (1995) 407–416.
5. Martin, E.J., Blaney, J.M., Siani, M.A., Spellmeyer, D.C., Wong, A.K and Moos, W.H., *Measuring diversity: Experimental design of combinatorial libraries for drug discovery*, J. Med. Chem., 38 (1995), 1431–1436.
6. Cramer, R.D. III, Bunce, J.D., Pallerson, DE. and Frank I.E., *Crossvalidation bootstrapping, and partial least squares Compared with multiple regression in conventional QSAR studies*, Quant. Str.-Act. Relat., 7 (1988) 18–25.
7. Martin, Y.C., Kim, K.-H. and Lin, C.T., *Comparative molecular field analysis: CoMFA*. In Charton, M. (Ed.) Advances in quantitative structure–property relationships, Vol. 1, JAI Press, inc., Greenwich, CT, 1966, pp. 1–52.
8. Norinder, U., *Single and domain mode variable selection in 3D QSAR applications*, J. Chemometrics, 10 (1996) 95–105.
9. Patterson, D.E., Cramer, R.D., Ferguson, A.M., Clark, R.D. and Weinberger, L.E., *Neighborhood behavior: A useful concept for validation of molecular diversity descriptors*, J. Med. Chem., 39 (1996) 3049–3059.



- Ash, S., Cline, M.A., Homer, R.W., Hurst, T. and Smith, G.B., *SYBYL line notation (SLN): A versatile language for chemical structure representation*. J. Chem. Inf. Comput. Sci., 37 (1997) 7 1-79; SLN is part of the SYBYL and UNITY products available from Tripos, Inc., 1699 S. Hanley Road. St. Louis. MO 63411, U.S.A.
- Brown, R.D. and Martin, Y. C., *The information content of 2D and 3D structural descriptors relevant to ligand-receptor binding*, J. Chem. Inf. Comput. Sci., 37 (1997) 1-9.
- Good, A.C. and Kuntz, I.D., *Investigating the extension of pairwise distance pharmacophore measures to triplet-based descriptors*, J. Comput.-Aided Molec. Design, 9 (1995) 373-379.
- Willett, P., *Calculation of molecular similarity of the alignment of molecular fields*. In this volume.
- Briem, H. and Kuntz, I.D., *Molecular similarity based on DOCK-generated fingerprints*, J. Med. Chem., 39 (1996) 3401-3408.
- Kauvar, L.M., Higgins, D.L., Villar, H.O., Sportsman, J.R., Engqvist-Goldstein, A., Bukar, R., Bauer, K.E., Dilley, H. and Rocke, D.M., *Predicting ligand binding to proteins by affinity fingerprinting*, Chem. Biol., 2 (1995) 107-118.
- Clark, R.D., *Synthesis and QSAR of herbicidal 3-pyrazolyl  $\alpha,\alpha$ -trifluoro-tolyl ethers*, J. Agric. Food Chem., 44 (1996) 3643-3652.
- Balducci, R., McGarity, C., Rusinko, A. III, Skell, J. and Pearlman, R.S. and Pearlman, R.S. CONCORD was developed at the University of Texas at Austin and is available exclusively from Tripos, Inc., 1699 S. Hanley Road, St. Louis, MO 63144, U.S.A.
- Cramer, R.D., Clark, R.D., Patterson, D.E. and Ferguson, A.M. *Bioisosterism as a molecular diversity descriptor: Steric fields of single topomeric conformers*, J. Med. Chem., 39 (1996) 3060-3069.
- Chapman, D., *The measurement of molecular diversity: A three-dimensional approach*, J. Comput.-Aided Molec. Design. 10 (1996) 501-512.
- Tripos, Inc., 1699 S. Hanley Road, St. Louis, MO 63144, U.S.A.
- (a) Clark, R.D. and Cramer, R.D., *The many ways to skin a combinatorial centipede*, 211th ACS National Meeting (1966). Abstract CINF 067. 1966: (b) molecular Diversity Manager Manual. SYBYL Version 6.3, Tripos, Inc., 1699 S. Hanley Road. St. Louis. MO 63144, U.S.A., 1996; pp. 194-195.
- Collantes, C.R., Tong, W. and Welsh, W.J., *Use of moment of inertia in comparative molecular field analysis to model chromatographic retention of nonpolar solutes*, Anal. Chem., 68 (1966) 2038-2043.
- Welsch, W.J., Tong, W., Collantes, E.R., Chickos, J.S. and Cagarin, S.G., *Enthalpies of sublimation and formation of polycyclic aromatic hydrocarbons (PAHs) derived from comparative molecular field analysis (CoMFA): Application of moment of inertia for molecular alignment*, Thermochemica Acta, 290 (1966) 55-64.
- Hahn, M., *Three-dimensional shape-based searching of conformationally flexible compounds*, J. Chem. Inf. Comput. Sci., 37 (1997) 80-86.
- Matter, H. and Lassen, D., *Compound libraries for lead discovery*, Chimica Oggi, 6 (1996) 9-15.
- Brown, R.D. and Martin, Y.C., *Use of structure-activity data to compare structure-based clustering methods and descriptors for use in compound selection*, J. Chem. Inf. Comput. Sci., 36 (1996) 572-584.
- Clark, R.D. and Cramer, R.D., *Taming the combinatorial centipede*, CHEMTECH, 27 (1997) 26-30.

# Similarity and Dissimilarity: A Medicinal Chemist's View

Hugo Kubinyi

*Drug Design, BASF AG, 0-67056 Ludwigshafen, Germany*

Several 3D approaches discussed in this volume describe methods for the analysis and quantitative description of chemical similarity. The underlying concept is that chemical similarity is reflected by similar biological activities — i.e. chemically closely related analogs should be related in their mode of action, as well as in their relative potencies. This fundamental assumption has, indeed, been used in medicinal chemistry research, and has led to many valuable drugs.

However, chemical similarity may have different facets if a computer chemist or a medicinal chemist look at the compounds. There is no argument that for maximal affinity a ligand of a biological macromolecule has to fit the binding pocket geometrically and that hydrophobic surfaces of the ligand and the binding site have to be complementary. The functional groups of the ligand need a separate consideration. For lipophilicity, there is no significant difference between, e.g.  $-O-$  and  $-NH-$  in an organic molecule; for ionization, there is a big difference whether the nitrogen atom is part of a basic group (an amine) or a neutral group (e.g. in an amide); and for binding, potency differences of several orders of magnitude may result from the exchange of the hydrogen bond acceptor  $-O-$  against a donor function  $-NH-$ .

## 1. Similarity as a Design Principle in Lead Optimization

Nearly all drugs result from the optimization of a lead structure. Sources of such leads are natural products from plants or microorganisms, synthetic chemicals or their intermediates, hits from (high-throughput) screening of in-house and combinatorial libraries, rational concepts from a biochemical pathway or the unexpected observation of a therapeutically useful side-effect of a drug. Most often, the biological activity of a lead structure is neither optimal, with respect to its efficacy, nor with respect to specificity, bioavailability, pharmacokinetics, toxic and other side-effects. Chemists perform more or less systematic variations of lead structures, using the experience of about 100 years of medicinal chemistry and the results of (quantitative) structure–activity relationships.

The principle of bioisosteric replacement of functional groups serves as a successful optimization strategy [1–3]. Its systematic application has resulted in a broad variety of therapeutically used drugs, many of them finally having the desired combination of favorable properties. A few examples of typical but different consequences of isosteric replacement of atoms or groups are illustrated by compounds **1–3** (Fig. 1), some others with unexpected effects on biological activities are discussed in later sections of this chapter.

In their attempts to optimize lead structures, medicinal chemists intuitively follow the principles of evolution. In genetic and evolutionary algorithms, randomly generated starting models (the lead structures) are reproduced involving random mutations and crossover (the chemical variation of the structures). Better models (compounds) are kept for further modification; worse ones are discarded. The biological activity, in later

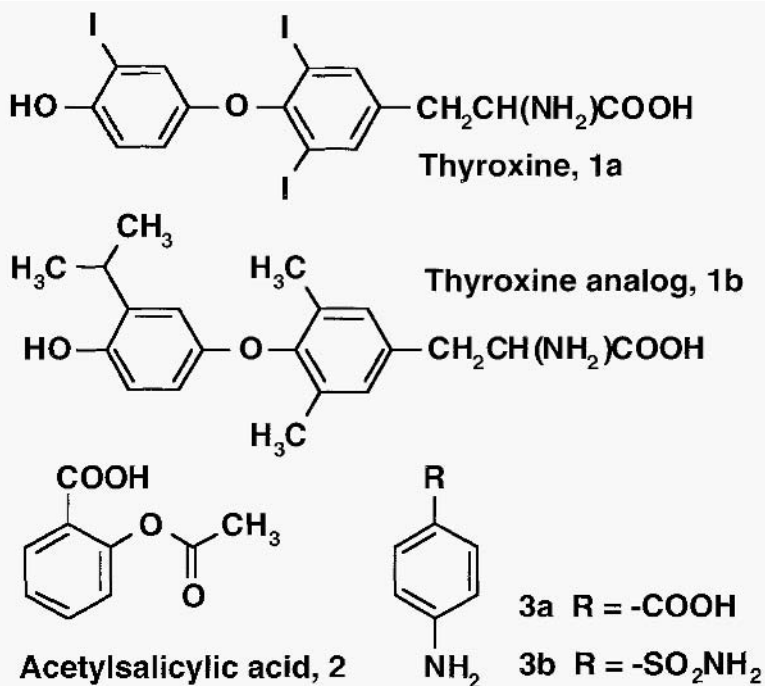


Fig. 1. Different effects of isosteric replacement. (a) The substitution of all three iodine atoms of thyroxine **1a** leads to an analog **1b** which still acts as a thyromimetic agent. (b) Replacement of an ester function,  $-O-CO-$ , by an amide function  $-NH-CO-$ , most often produces analogs with higher metabolic stability: if this replacement is done in acetylsalicylic acid **2**, an inactive analog results because the amide is no longer able to transfer an acetyl group to a certain serine residue of cyclooxygenase, (c) *p*-Aminibenzoic acid **3a** is an essential metabolite of microorganism: if the acid group,  $-COOH$ , is replaced by a sulfonamide group  $-SO_2NH_2$ , the antimetabolite sulfanilamide **3b** results (active metabolite of the antibacterial sulfonamide sulfamidochrysoidine)

stages a selectivity index or some other biological property, serves as the 'fitness function' for the 'survival' of certain structural entities. That genetic and evolutionary algorithms, indeed, reduce the effort in searching the most active analogs has recently been confirmed by dedicated investigations (e.g. [4,5]).

## 2. The Biological Activity of a Ligand Depends on its Flexibility

Other common structural modifications in the optimization of a lead structure are the dissection of rings or the rigidification of flexible molecules. Molecules with several rotatable bonds may adopt many different geometries — some of them being favorable because of low internal energies, others being less favorable because of van der Waals or electrostatic repulsions between non-bonded atoms or groups. If different conformations of such molecules are 'frozen' by closing rings between certain atoms, either one of two very different consequences results. If the frozen conformation differs from the

bioactive conformation of the flexible lead or if the added atoms interfere with the binding, biological activity will be more or less destroyed. If the ring closure stabilizes the bioactive conformation, usually a significant increase in biological activity results. This comes from the fact that the binding of a flexible analog is entropically unfavorable, due to a loss of rotational degrees of freedom, whereas the rigid analog has already lost its flexibility prior to binding. Two examples of highly similar, flexible and rigid analogs are the pairs **4** and **5** [6] and **6** and **7** [7], respectively, where the rigid analogs are significantly more active than their flexible counterparts (Fig. 2).

The computer program CAVEAT was developed for the design of rigid analogs which bear a pharmacophore in a certain geometry [8]. CAVEAT starts from a struc-

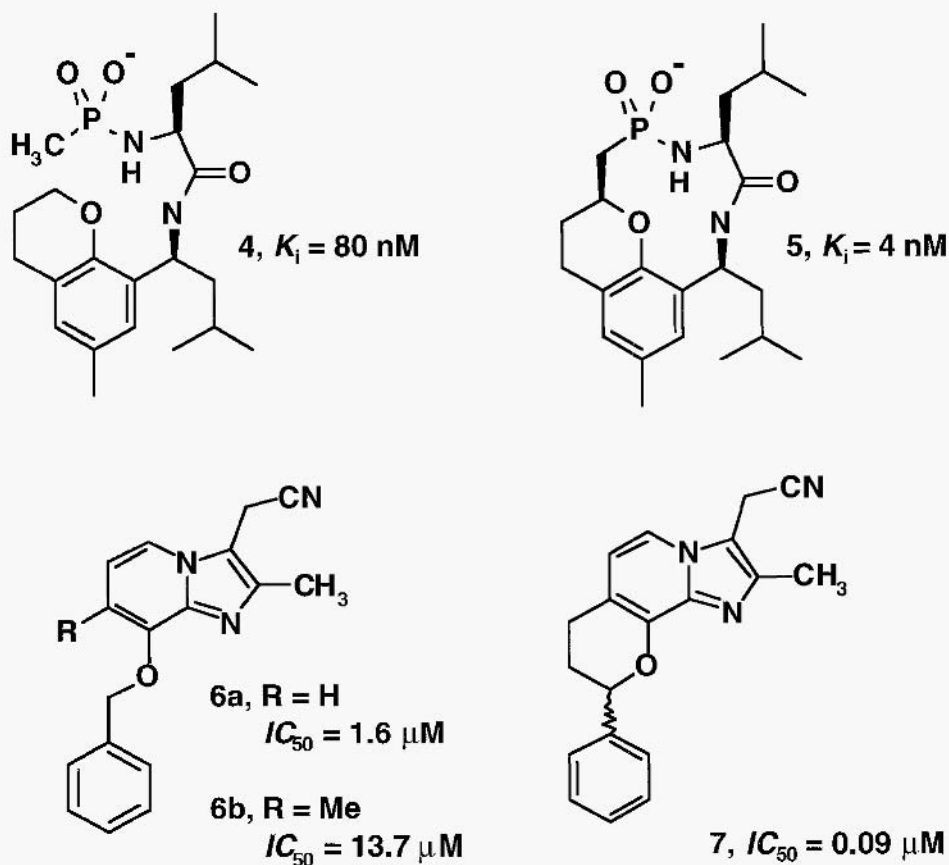


Fig. 2. Rigidification of a bioactive conformation significantly increases biological activities (a) If an additional ring is introduced into the thermolysin inhibitor **4** to produce antilog **5**, a 20-fold increase in affinity is observed (b) Introduction of a methyl group into the  $\text{H}^+/\text{K}^+$ -ATPASE inhibitor **6a** reduces biological activity by a factor of 8, most probably due to a destabilization of the bioactive conformation; if the substituent  $R$  in **6b** is extended to a new ring (compound **7**), which fixes the bioa conformation, the biological activity of **6b** is enhanced by a factor of 150.

tural hypothesis or from the known 3D structure of a ligand, e.g. a polypeptide. and extracts vectors of residues that participate in binding. In a peptide, these vectors are e.g. the  $C_\alpha$ – $C_\beta$  bonds of the amino acid side chains. Then the program identifies ring systems that are suited to accommodate these residues in exactly the same relative geometries.

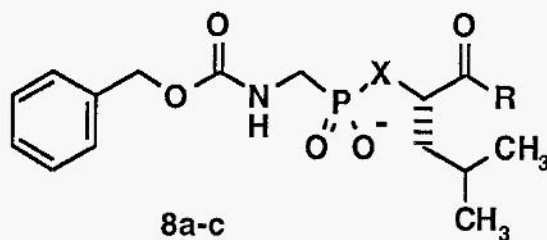
A lead structure, where a significant increase in binding affinity is achieved by steric constraints, not by ring closure, has been described by Kaplan and Bartlett [9]. In the flexible phosphonate tripeptide  $\text{Cbz-Phe-Ala}[\text{PO}_2^-]\text{-O-CH}(\text{CH}_3)\text{COOH}$  (=  $Z\text{-Phe-Ala}^p\text{-(O)-Ala}$ ;  $p$  indicates a  $\text{PO}_2^-$  residue instead of a  $\text{C=O}$  group). inhibitory activities against the zinc protease carboxypeptidase A increase significantly with the stepwise introduction of bulky residues. In the series  $Z\text{-Phe-Ala}^p\text{-(O)-Ala}$ ,  $Z\text{-Phe-Ala}^p\text{-(O)-Phe}$  and  $Z\text{-Phe-val}^p\text{-(O)Phe}$ , affinities increase from 56 nM to 0.001 nM and 0.000010–0.000027 nMol (i.e. 10–27 fMol). The more active analogs have a higher lipophilicity, due to the additional phenyl group in phenylalanine, as compared to alanine, and due to the isopropyl group instead of the methyl group of the C-terminal lactic acid residue. But in this case. the affinity difference of about six orders of magnitude cannot be attributed only to the increase in lipophilicity which changes by less than three orders of magnitude. It must also result from the conformational constraint imposed on both phenylalanines by the central valine and which, by fortune, stabilizes the bioactive conformation. In terms of ‘chemical similarity’, all three analogs are closely related: with respect to their internal flexibility and their preferred conformations, they are not.

### 3. Biological Potencies of Similar Molecules

For compounds with comparable flexibility, most often similar analogs have also similar biological potencies. That this need not always be the case can be seen by a comparison of three thermolysin inhibitors **8a–c** (Fig. 3) [10,11]. All analogs **8a** and **c**, with  $X = \text{-NH-}$  and  $\text{-CH}_2\text{-}$ , are about 1000 times more potent than the  $X = \text{-O-}$  analogs **8b** ( $R = \text{OH}$  or different amino acids). The explanation for this effect can be easily derived from the 3D structure of the complex of thermolysin with the inhibitor **8a** ( $R = \text{-Leu-OH}$ ). If  $X$  is an  $\text{-NH-}$  group (series **8a**), a hydrogen bond is formed between this group and the oxygen atom of an alanine carbonyl group. In the  $\text{-O-}$  analog **8b**, this hydrogen bond cannot be formed; in addition. an electrostatic repulsion between the two oxygen atoms results. The biological activity of the  $\text{-CH}_2\text{-}$  analog **8c** has been predicted [12] to be comparable to the  $\text{-NH-}$  analogs and much higher than for the  $\text{-O-}$  analogs. This was later confirmed by the synthesis of the inhibitors **8c** [11].

A similar but less pronounced effect is observed for the thrombin inhibitors **9a–c** (Fig. 4) [13]; in this case, the non-bonded contact between the  $\text{-X-}$  group of the ligand and the carbonyl group of Gly 216 in the binding site of thrombin is responsible for the structure–activity relationship. If, on the other hand, the carbonyl group of **10a**, which forms a hydrogen bond with the  $\text{-NH-}$  group of Gly 216, is replaced by  $\text{-CH}_2\text{-}$  (**10b**), the affinity is reduced by more than three orders of magnitude (Fig. 4) [13].

Neutral endopeptidase 24.11 (NEP 24.11; enkephalinase) is a zinc protease with some homology to thermolysin. The NEP inhibitors **11a–d** (Fig. 5) show a different



Binding constants  $K_i$  in  $\mu\text{M}$

R	X =	8a -NH-	8b -O-	8c -CH <sub>2</sub> -
-OH		0.76	660	1.4
-Gly-OH		0.27	230	0.3
-Leu-OH		0.01	9	0.01

Fig. 3. The  $X = -\text{NH}-$  group of the thermolysin inhibitors **8a** forms a hydrogen-bond with the carbonyl oxygen atom of Ala 113: this affinity-enhancing effect is only moderated by desolvation of the ligand in going from the free to the bound state. All oxygen analog **8b** are much less active, because they cannot form such a hydrogen-bond; in addition, there is the unfavorable desolvation effect and an electrostatic repulsion between the oxygen atoms of the ligand and the binding site. Like the  $-\text{O}-$  analogs **8b**, the  $-\text{CH}_2-$  analogs **8c** cannot form the hydrogen-bond to Ala 113, but this disadvantage is countbalanced by the effect that no desolvation of the  $\text{CH}_2$  group of the ligand and no electrostatic repulsion take place. The differences between  $X = -\text{NH}-$ ,  $-\text{O}-$  and  $-\text{CH}_2-$  are identically observed whether  $R = -\text{OH}$ ,  $-\text{Gly}-\text{OH}$  or  $-\text{Leu}-\text{OH}$ .

structure-activity relationship, as compared to the preceding examples. Here the lipophilic  $-\text{CH}_2-$  and  $-\text{S}-$  analogs are much less active than the  $-\text{NH}-$  and  $-\text{O}-$  analogs, which both have comparable biological activities [14]. As long as the 3D structure of NEP 24.11 remains unknown, no explanation can be given for this effect.

An unexpected structure-activity relationship has been observed for the macrocyclic renin inhibitors **12a** and **b** (Fig. 5) [15]. Modelling of the complex of **12a** with the aspartyl protease renin indicated that  $X = -\text{NH}-$  should form a hydrogen bond with one of the Asp 226 oxygen atoms. Thus, the replacement of  $-\text{NH}-$  by  $-\text{O}-$  should reduce the affinity. The opposite was observed; the affinity of analog **12b** is increased by more than two orders of magnitude. The authors discuss several possible explanations for this observation but finally conclude: 'we have attempted to find a plausible basis for this surprising reversal in potency, but without much success ... the comparison [of both analogs] is therefore intrinsically more difficult, and will constitute a more demanding test for thermodynamic simulation techniques [as compared to the thermolysin inhibitors. Fig. 3]'.<sup>2</sup>

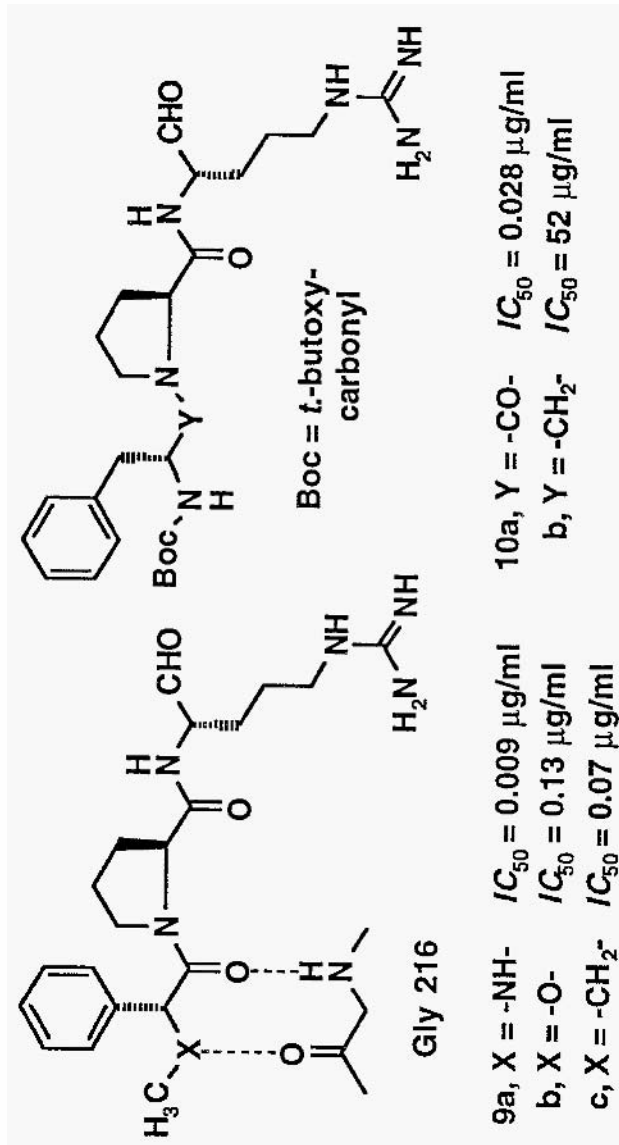


Fig. 4. The inhibitor **9a** forms two hydrogen-bonds with glycine 216 of the serineprotease thrombin. As compared to the thrombolytic inhibitors **8a-c** (Fig. 3), **a** similar but much less pronounced effect is observed by structural variation of the group X in the thrombin inhibitors **9a-c**. If, on the other hand, the carbonyl group of **10a** is replaced by a methylene group, the resulting pseudopeptide **10b** shows a 2000-fold weaker inhibitor activity.

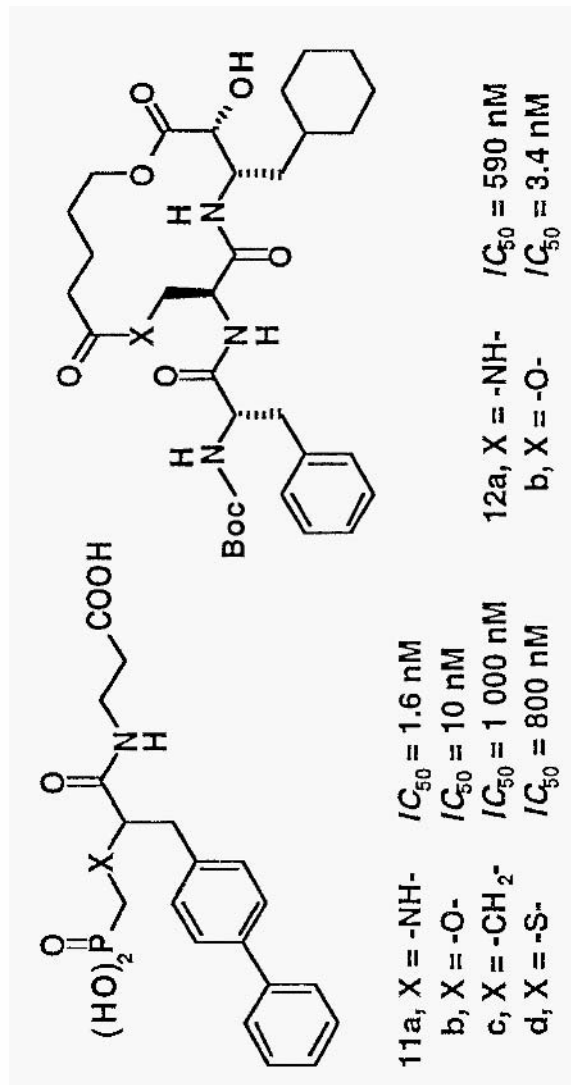


Fig. 5. The NEP 24. 11 inhibitors **11a** and **b** are much more potent than the more lipophilic analogs **11c** and **d**. The macrocyclic renin inhibitor **12a** is less potent than its oxygen cinalq **12b**, although its NH group should, form a favorable hydrogen-bond with the carboxylate of Asp 226 of the enzyme.



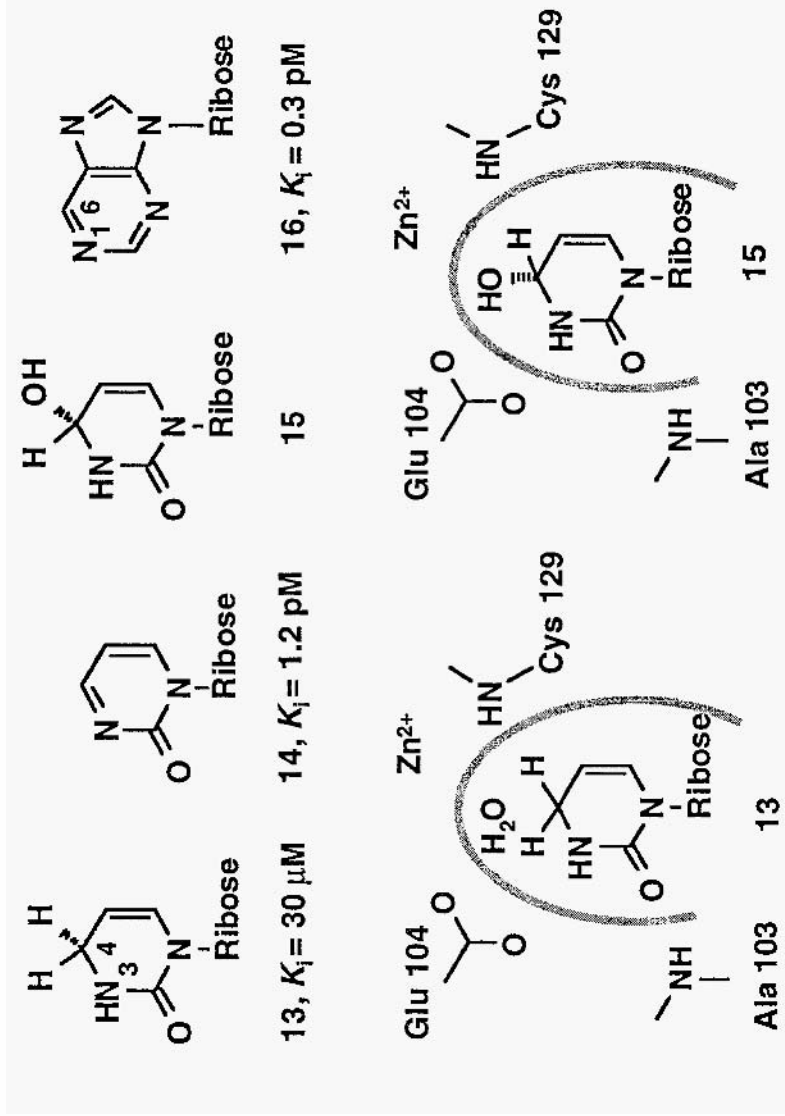


Fig. 6. 3,4-Dihydrozebularine 13 is a weak inhibitor of cytidine deaminase. Zebularine 14 interacts with the enzyme as the 3,4-hydrate 15, as confirmed by X-ray structure analysis. The favorable interactions of the hydroxy group enhance the biological activity by more than seven orders of magnitude. Nebularine 16, a strong adenosine deaminase inhibitor, is also proposed to interact as the 1,6-hydrate.

There are many examples in literature where the introduction of an –OH group into a ligand either causes an increase or a decrease of biological affinities. From a theoretical point of view, this is not surprising. If the new –OH group forms hydrogen bonds with polar groups at the binding site (either as a hydrogen bond donor or as an acceptor), the net free energy depends on the balance between the desolvation energies of the water shells at the surfaces of the ligand and the binding site, as compared to the energy of the formed hydrogen bond/s and the entropy gain by the release of some water molecules. Certain tightly bound water molecules in the binding cavity of a protein (usually seen in the X-ray structures), e.g. those which form more than two hydrogen bonds to the protein, are not easily removed. The attempt to introduce an –OH group into the ligand, to replace such a water molecule, must necessarily fail.

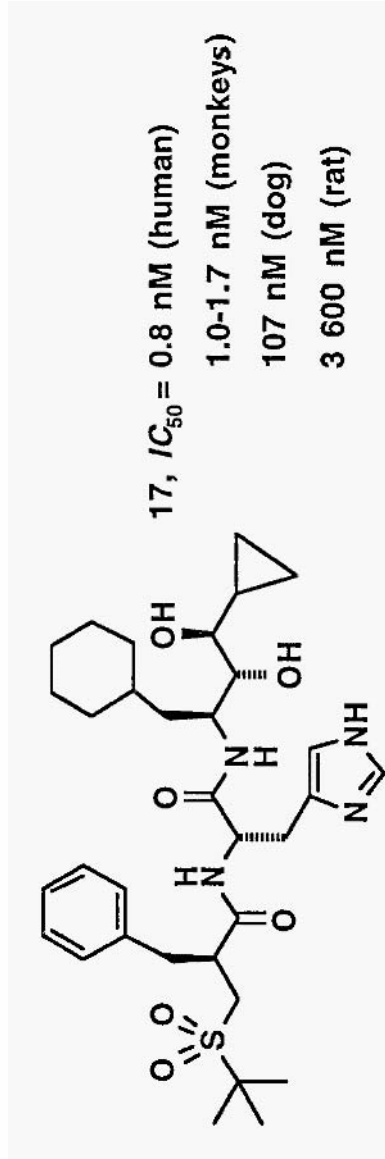
An example where this is not the case, and where significantly enhanced binding affinities result after the introduction of such a hydroxy group, are the cytidine and adenosine deaminase inhibitors **13–16** (Fig. 6) [16,17]. In this special case, the resulting affinity differences are 7 to 8 orders of magnitude!

#### 4. Biological Potencies versus Similar Biological Targets

Functionally corresponding proteins from different species have identical or very similar 3D structures, but they normally differ in their amino acid composition. Although they always show, dependent on the evolutionary relationship between the two species, a certain degree of homology, structure–activity relationships may significantly differ, even after the replacement of just one single amino acid by another one. This has been shown, for example, by a comparison of rat and human 5-HT receptors [18]. Although both have about 90% identity and more than 95% homology in their amino acid sequences, the binding affinities of a series of 5-HT receptor ligands and  $\beta$ -adrenergic blocking agents are not related ( $n = 10$ ;  $r = 0.27$ ). If one (!) amino acid of the human receptor, i.e. threonine 355, is replaced by the corresponding amino acid of the rat receptor, i.e. asparagine, the binding affinities of the ligands change significantly. Now the human mutant 5-HT receptor behaves like a rat receptor; all affinities to both receptors are more or less identical ( $r = 0.98$ ) [18].

In the past, new compounds have always been tested in animals before they could be applied to humans. Human proteins were not available, except in rare cases. Only for those proteins that could be extracted from human material, e.g. hemoglobin or thrombin from blood, could one predict the human biological activity of a new drug from *in vitro* studies, prior to animal studies. With the progress in gene technology, it is now possible to produce human proteins in sufficient quantities to establish test models. Thus, their biological activity in humans can be forecasted from investigations at the molecular level. How important this is, can be illustrated, e.g. by the activities of the renin inhibitor remikiren **17** against the renins of different species (Fig. 7) [19].

Different QSAR models for dihydrofolate reductase (DHFR) inhibitors, chemically related to the antibacterial drug trimethoprim **18** (Fig. 8), describe the inhibitory activities versus *E. coli* and *L. casei* DHFR [20]. Whereas in the case of *E. coli* DHFR, the 3-, 4- and 5-substituents at the benzyl group contribute to biological activities, only the



*Fig. 7. Remikiren 17 is a much stronger inhibitor versus human and monkey renins, as compared to dog and rat renins. If this compound had only been tested in small laboratory animals as was the case before gene technology could be applied in drug research, it would not have been considered for further development, due to its weak Biological activity.*

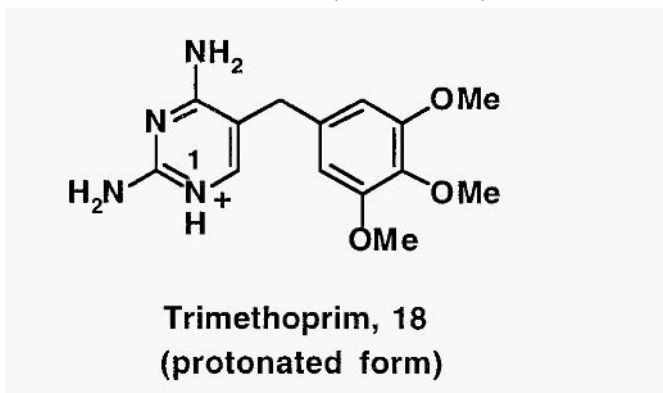
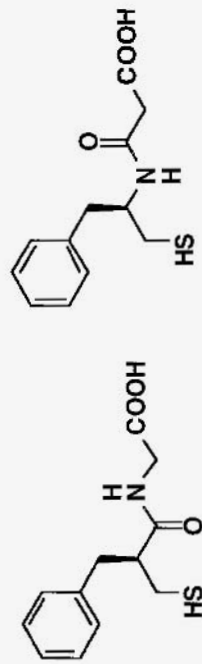


Fig. 8. Trimethoprim **18** is proposed to interact with dihydrofolate reductase in its protonated form, where N1 changes its character from a hydrogen-bond acceptor to a hydrogen-bond donor. Due to the positive charge of the ligand, the exchanged of negatively charged amino acids of the protein to neutral amino acids and of neutral to positively charged amino acids reduce the affinity of the ligand.

3- and 4-substituents are relevant for the activities against *L. casei* DHFR. Several years later, the X-ray analyses of both enzymes explained these differences: in *L. casei* DHFR, there is a much narrower pocket, due to a bulky leucine in the binding site, as compared to a flexible methionine side chain in *E. coli* DHFR [20].

Whereas many differences in the binding affinities to related biological targets can be attributed to different non-covalent interactions with the amino acids of the binding sites, such effects may also result from more distant changes in a protein. Trimethoprim **18** is thought to bind to DHFR in its positively charged form. An *E. coli* mutant, where the negatively charged glutamate residue 118 is replaced by a neutral glutamine, has a 4.5-fold lower affinity for trimethoprim, despite the fact that the modified group is about 15 Å apart from the charged N1 of trimethoprim. An even more pronounced effect is observed in a double mutant (Glu 118 Gln, Leu 28 Arg): the affinity of trimethoprim is reduced by a factor of 190, although the positive charge of the arginine residue is about 8 Å apart from the positive charge of the ligand [21]. This reduction of binding affinities has been used to explain the selectivity of 5 to 6 orders of magnitude of trimethoprim against bacterial DHFRs, as compared to avian and mammalian DHFRs. In chicken DHFR, seven amino acids in the environment of the binding site (not in the binding site itself) have changed charge from negative to neutral or from neutral to positive, as compared to *E. coli* DHFR. If one compares the effects observed in the *E. coli* double mutant with these figures, the changes seem to be sufficient to explain the observed effect [21].

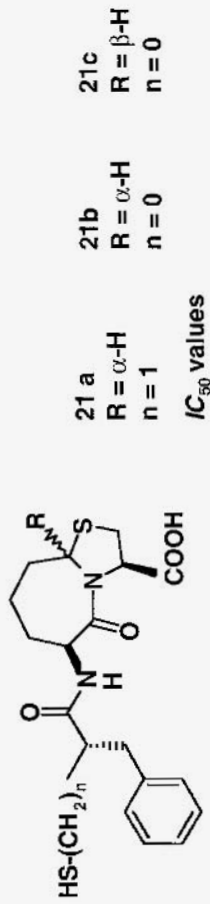
Thiorphan **19** and its *retro-inverso* peptide, *retro*-thiorphan **20** (Fig. 9), are inhibitors of the structurally related zinc proteases thermolysin and NEP 24.11. Although the affinities of the ligand pair differ, from enzyme to enzyme, by three orders of magnitude, there are no significant differences between them [22]. Thus, they may be considered to be 'similar', which was also confirmed by the X-ray structure analyses of their complexes with thermolysin; both compounds bind in an analogous manner and have corresponding interactions with the protein [23]. On the other hand, their activities



Thiorphan, 19

retro-Thiorphan, 20

Thermolysin	$K_i = 1.8 \mu\text{M}$	$K_i = 2.3 \mu\text{M}$
NEP 24.11	$K_i = 0.0019 \mu\text{M}$	$K_i = 0.0023 \mu\text{M}$
ACE	$K_i = 0.14 \mu\text{M}$	$K_i > 10 \mu\text{M}$

21 a  
R =  $\alpha$ -H  
n = 121 b  
R =  $\alpha$ -H  
n = 021 c  
R =  $\beta$ -H  
n = 0 $IC_{50}$  values

NEP 24.11	1.1 nM	11.5 nM	2 820 nM
ACE	5.5 nM	16 nM	11.5 nM

Fig. 9. Thiorphan 19 and retro-thiorphan 20 are 'similar' with respect to thermolysin and NEP 24.11, but they are highly 'dissimilar' with respect to angiotensin converting enzyme (ACE). Changing the configuration of 21a and b at the ring junction, from  $\alpha$ -H to  $\beta$ -H in 21c, does not influence the ACE-inhibitory activity, but reduces the NEP-inhibitory activity to a large extent; thus, the selectivity against ACE is increased by a factor of 300.

against yet another related zinc protease, angiotensin converting enzyme (ACE), are significantly different [22]; with respect to this enzyme, the analogs are 'dissimilar'. Whether the differences result from different binding modes or from an unfavorable binding geometry of *retro*-thiorphan in ACE will only be known if the ACE 3D structure becomes available. Corresponding differences in the structure–activity relationships are observed in some other dual zinc protease inhibitors **21a–c** (Fig. 9) [24]. Again, the degree of similarity and dissimilarity of the ligands depends on the biological target.

Corresponding problems are also observed in toxicity predictions for humans, as illustrated by the toxicities of lysergic acid diethylamide (LSD) and 2,3,7,8-tetrachlorodioxin ('dioxin') in different species. LSD is only weakly toxic to mice ( $LD_{50} = 50\text{--}60$  mg/kg) and rats ( $LD_{50} = 16.5$  mg/kg) but significantly more toxic to rabbits ( $LD_{50} = 0.3$  mg/kg). An elephant, which was (cautiously?) treated with an LSD dose of 0.3 g (corresponding to about 0.06 mg/kg), died within several minutes [25]! Humans seem to tolerate LSD quite well. Cases of death have been observed as the result of drug-induced suicides, but not by any toxicity of the drug itself.

2,3,7,8-Tetrachlorodioxin is highly toxic to several species — e.g. guinea-pigs ( $LD_{50} = 0.6\text{--}2.5$   $\mu$ g/kg) and mink ( $LD_{50} = 4$   $\mu$ g/kg). It is much less toxic for mice ( $LD_{50} = 114\text{--}280$   $\mu$ g/kg), rats ( $LD_{50} = 22\text{--}320$   $\mu$ g/kg), hamsters ( $LD_{50} = 1150\text{--}5000$   $\mu$ g/kg), rabbits ( $LD_{50} = 115\text{--}275$   $\mu$ g/kg), dogs ( $LD_{50} > 100$  and  $< 3000$   $\mu$ g/kg) and monkeys ( $LD_{50} < 70$   $\mu$ g/kg) [26]. If one extrapolates from the monkey, dioxin should be relatively 'harmless' for humans, at least if only acute toxicity is considered. However, the significantly different toxicity versus the closely related species guinea pig and hamster puts a caveat on too simple and straightforward extrapolations. Of course, for humans the  $LD_{50}$  is not the relevant endpoint, but rather the no-effect  $LD_0$  level!

## 5. Similar Structures and Activities, but Different Binding Modes

Comparable biological activities of chemically similar structures do not necessarily result from analogous binding modes of all analogs. Since at least in 3D QSAR, if not in all QSARs, the correct superposition of all analogs within a series is a precondition for reliable results, a better knowledge of the binding modes is most important. Unexpected differences in the binding modes of some analogs certainly produce problems in 3D QSAR studies that are based on such a mutual alignment of the structures. But it is to be expected that also 3D QSAR modifications, which do not require a structural alignment, should have problems in such cases.

Differences in binding modes have been extensively reviewed [1,27–29]. Thus, only some examples will be summarized here, without detailed discussion. Purine nucleoside phosphorylase (PNP) inhibitors **22** and **23** show surprising structure–activity relationships, which can be explained only by inspecting the X-ray structures of the inhibitor complexes (Fig. 10). The exchange of an acceptor nitrogen to a donor nitrogen atom changes not only the interactions with the directly involved asparagine side chain, but also of an 8-amino group with the threonine hydroxyl and methyl groups [30,31].

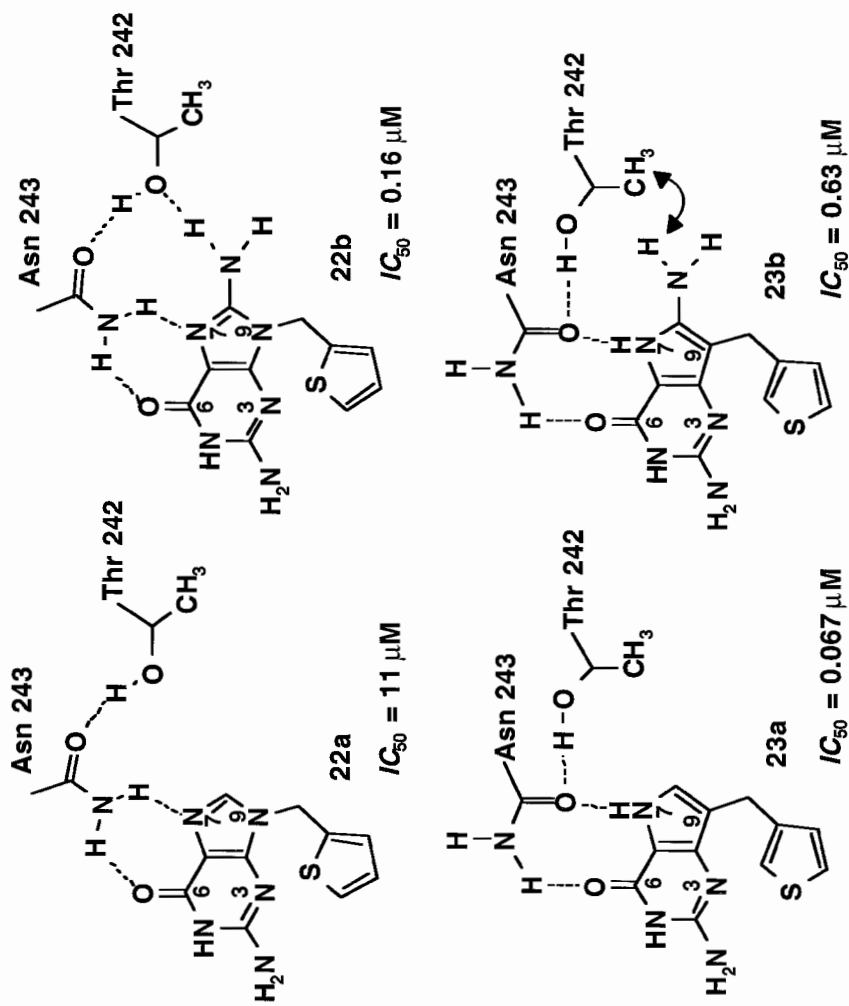


Fig. 10. The X-ray structures of the purine nucleoside phosphorylase (PNP) complexes with the inhibitors **22** and **23** show that **22a** forms less favorable hydrogen-bonds with Asp 243 than the desazapurine **23a**. On the other hand, the introduction of an 8-amino group has different consequences in **23a** and **23a**. Whereas the new amino group of **22b** forms an additional hydrogen-bond to Thr 242, it causes unfavorable polar/nonpolar interactions in **23b**.

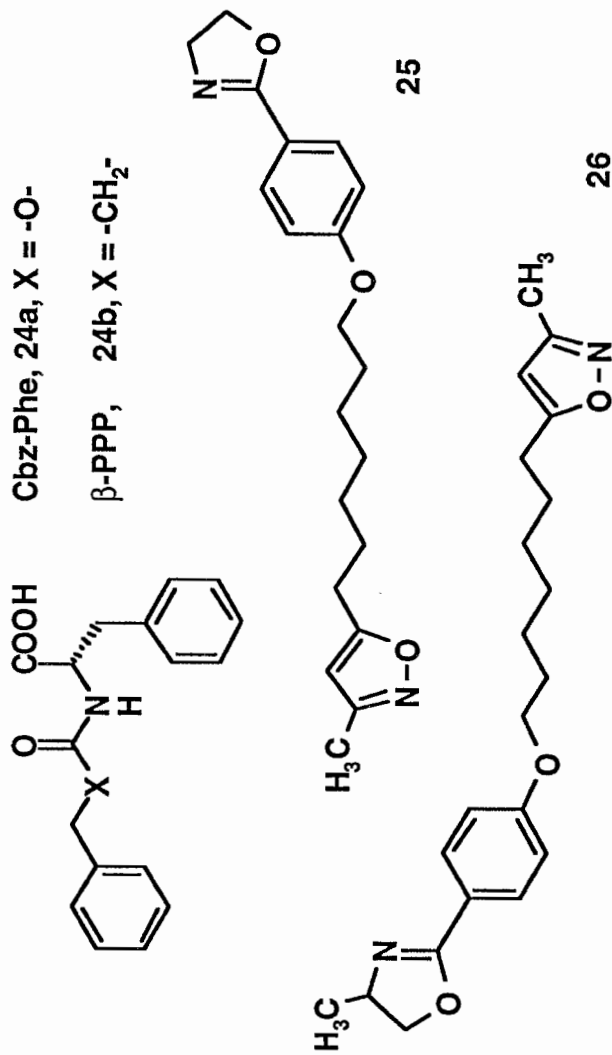


Fig. 11. Changing an oxygen atom of the thermolysin inhibitor **24a** to CH<sub>2</sub> (**24b**) leads to a reversed binding mode. Also the viral coat protein ligands **25** and **26** bind in a reverse manner, as indicated. There are only hydrophobic groups and one hydrogen-bond donor in the binding pocket of the viral protein; this donor can form a hydrogen-bond with either the oxazoline or the isoxazole ring.



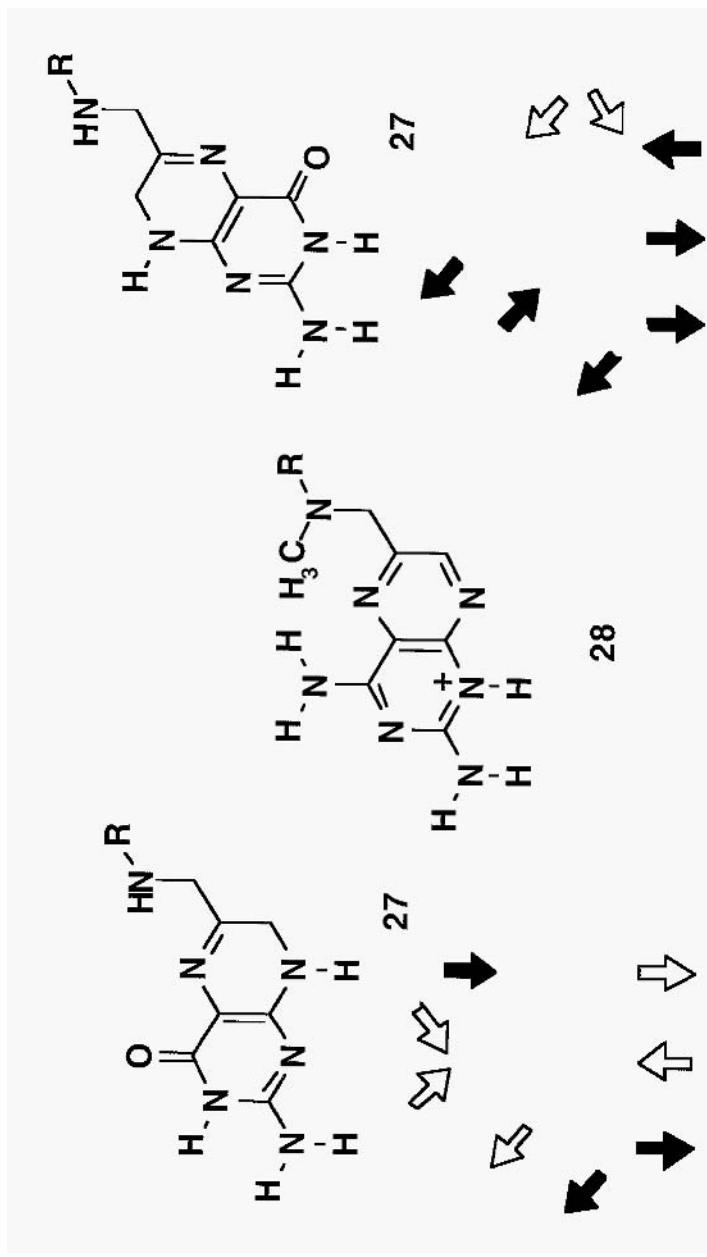


Fig. 12. Dihydrofolate 27 and the positively charged form of methotrexate 28 ( $R = p\text{-C}_6\text{H}_4\text{-CONHCH(COOH)CH}_2\text{CH}_2\text{COOH}$ ) have different patterns of hydrogen-bond donors and acceptors, if they are superimposed according to their ring systems (left side). If the ring system of 27 is rotated around the bond which connects the ring system with the rest of the molecule, a much better similarity of the hydrogen-bond patterns results (right side). Filled arrows indicate identical directions of the hydrogen-bond in the analogs 27 and 28.

Carbobenzoxy-phenylalanine (Cbz-Phe) **24a** and  $\beta$ -phenylpropionyl-phenylalanine ( $\beta$ -PPP) **24b** (Fig. 11) differ only by their X group. Nevertheless, they bind to thermolysin in an opposite direction, as uncovered by the determination of the 3D structures of the complexes [32]. Reverse binding modes are also observed for some structurally related viral coat protein ligands – e.g. compounds **25** and **26** (Fig. 11) [33,34].

Dihydrofolate **27**, the substrate of dihydrofolate reductase (DHFR), and the DHFR inhibitor methotrexate **28** seemingly differ only in minor chemical details (Fig. 12). However, this small difference has significant consequences on the hydrogen-bond donor and acceptor patterns of the compounds. Accordingly, dihydrofolate and methotrexate accommodate the binding site of DHFR in different modes [35], an effect which has been predicted from the X-ray structure of the dihydrofolate/DHFR complex [36].

The isomeric 1-, 2- and 4-phenylimidazoles are inhibitors of the oxidation of camphor by cytochrome P450<sub>cam</sub>. Nevertheless, only 1- and 4-phenylimidazole bind as expected, with one nitrogen atom of the imidazole ring coordinated to the iron atom of the porphyrin system; 2-phenylimidazole cannot be accommodated in such a position and binds in a completely different manner [37]. Structurally related dipeptide inhibitors of the serine protease elastase bind with their corresponding residues in different pockets of the enzyme [1,27,38]. The trypsin inhibitor *p*-guanidinumbenzoate seems to be one of the rare cases where X-ray crystallography provides clear evidence that one and the same ligand binds in distinctly different orientations [39]. Many other examples of different binding modes of closely related analogs have been described in the literature [1,27–29]. From the observation of such differences, Dagmar Ringe proposed to use this information to design ‘hydra-headed’ inhibitors, which fill all possible pockets of a binding site.

## 6. Different Mechanisms of Action of Similar Molecules

Several well-known examples of different modes of action of closely related analogs can be found in medicinal chemistry and pharmacology textbooks. Norepinephrine **29a**, epinephrine **29b** and isoproterenol **29c** (Fig. 13) are adrenergic agonists. However, in going from R = H to R = CH<sub>3</sub> and R = isopropyl, the mechanism of action gradually changes from a more or less specific  $\alpha$ -adrenergic agonism to a pure  $\beta$ -adrenergic agonism. If the two hydroxyl groups of isoproterenol are changed into chloro substituents, the  $\beta$ -adrenergic antagonist dichloroisoproterenol (DCI) **30** results; in fact, **30** was the first  $\beta$ -blocker.

An unexpected effect resulted after the introduction of an isobutyl group into the AT<sub>1</sub>-specific angiotensin II receptor antagonist **31a** (Fig. 13). The original AT<sub>1</sub> selectivity is completely destroyed; in addition, the new analog **31b** is no longer an antagonist, it is a strong agonist at the AT<sub>1</sub> and AT<sub>2</sub> subtypes of this receptor [40]. A related effect has been observed for some structurally diverse cholecystokinin (CCK) receptor ligands; the introduction of an isopropyl group at a certain nitrogen atom leads from peripheral CCK antagonists to agonists [41,42].

The integrin family of receptors are structurally and functionally related membrane-embedded glycoproteins that ‘integrate’ the extracellular matrix with the cytoskeleton.

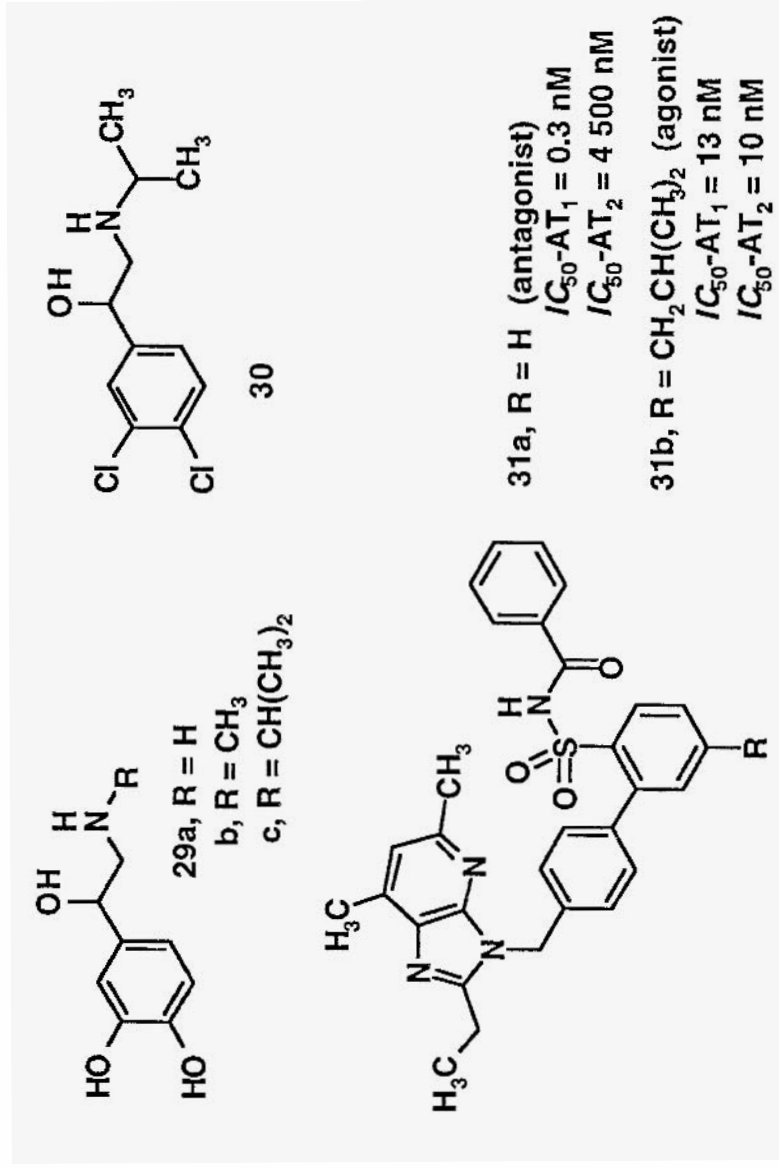


Fig. 13. The change from norepinephrine **29a** to epinephrine **29b** and isoproterenol **29c** gradually changes the mode of action from  $\alpha$ -agonism to  $\beta$ -agonism. Dichloroisoproterenol (DCI) **30** was the first  $\beta$ -adrenergic antagonist. Introduction of an isobutyl group into the  $AT_1$ -selective antagonist **31a** produces the potent, unselective  $AT_1/AT_2$  agonist **31b**.

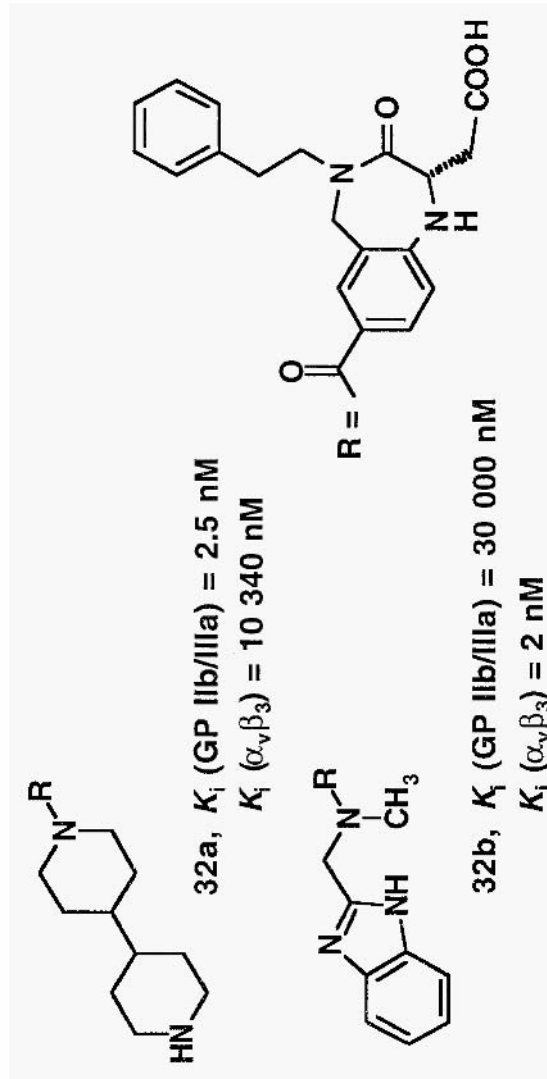


Fig. 14. The chemically closely related integrin antagonist **32a** and **32b** differ in their selectivity towards the fibrinogen (GP IIb/IIIa) and the vitronectin ( $\alpha_v\beta_3$ ) receptors by nearly eight orders of magnitude.

Important members are the GPIIb/IIIa receptor (GP for glycoprotein, I Ib/IIIa for the  $\alpha_{IIb}\beta_3$  integrin) and the vitronectin ( $\alpha_v\beta_3$ ) receptor [43,44]. Whereas the GPIIb/IIIa receptor binds, *inter alia*, fibrinogen and mediates blood platelet aggregation, the  $\alpha_v\beta_3$  receptor binds vitronectin and is responsible for angiogenesis, vascular smooth muscle migration and adhesion of osteoclasts to the bone matrix. Thus, both receptors are important for drug design, the GP I Ib/IIIa receptor for the development of antithrombotic agents and the vitronectin receptor for the development of drugs for the treatment of cancer, as well as against restenosis and osteoporosis. However, both fibrinogen and vitronectin interact with their receptors using an identical structural domain, the so-called RGD motif (RGD = Arg-Gly-Asp), but obviously in different conformations. This was first confirmed by the investigation of different cyclic pentapeptides, containing D-amino acids in different positions [44,45]. How far structure-based design (in this case, based only on the structures of ligands) can be advanced is illustrated by compounds **32a** and **32b**, which are highly selective GPIIb/IIIa and vitronectin receptor antagonists, respectively (Fig. 14) [46]. The selectivity of these analogs against the two receptors differs by nearly eight orders of magnitude, despite their close chemical similarity! The only major difference is the distance between the carboxylic group, which mimics the Asp, and the basic nitrogen atom, which mimics the Arg of the RGD motif.

A well-known example of different modes of action of closely related analogs are the anti-allergic agent promethazine **33**, the neuroleptic chlorpromazine **34**, and the anti-depressants imipramine **35a** and desipramine **35b** (Fig. 15). Despite their very similar structures, **33** acts mainly as a histamine H1 antagonist, **34** is a dopamine antagonist, **35a** is an unspecific norepinephrine- and serotonin-uptake inhibitor and **35b** is a norepinephrine-specific uptake inhibitor. Steroid hormones – e.g. **36–39** (Fig. 16) – give another example of strikingly different biological effects of chemically closely related analogs.

Several therapeutically used drugs bind to more than one receptor and are correspondingly termed ‘promiscuous’ ligands. However, whether a certain (balanced) unspecificity of their mode of action is advantageous for therapy or not still remains uncertain.

## 7. Chirality and Biological Activities

Due to the chiral nature of amino acids (except glycine), drug binding sites of proteins are asymmetric. In the past, the different actions of enantiomers of chiral molecules on enzymes and receptors were often neglected [47]. For economic reasons, racemates of synthetic drugs were used in therapy. Today, researchers and drug companies are more aware of the different effects of enantiomers and diastereomers [1,2] in their biological activities, as well as in their pharmacokinetics. Enantiomers can even be differentiated by their odor — e.g. the monoterpenes (*R*)- and (*S*)-limonene **40** and (*R*)- and (*S*)-carvone **41** (Fig. 17) [48]. Butaclamol **42** is mentioned as just one example of significantly different eudismic ratios (i.e. ratios of affinities of the different enantiomers) *versus* different receptors (Fig. 17) [47].

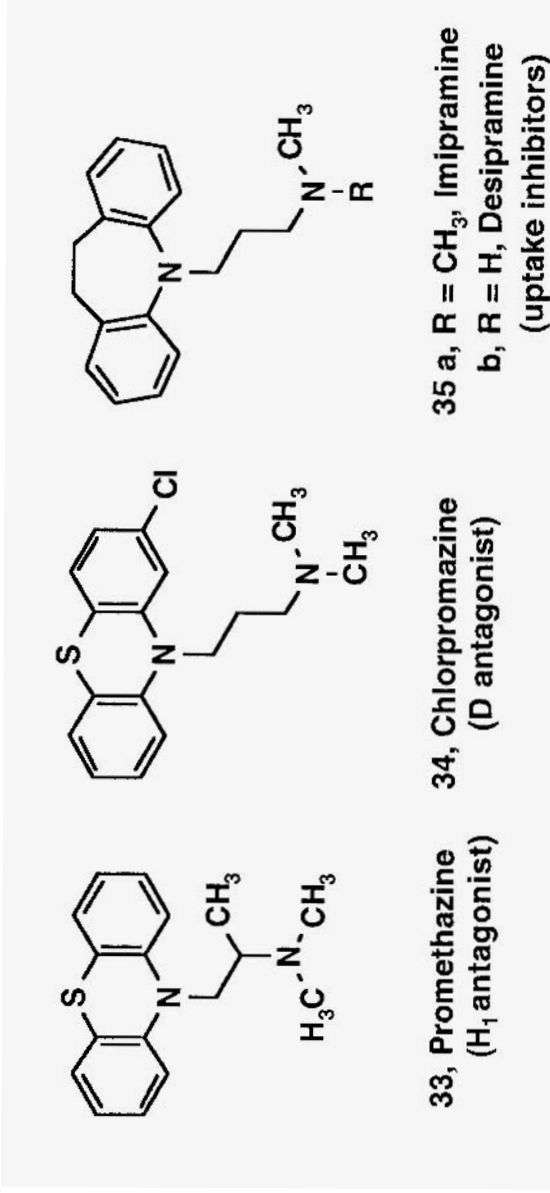


Fig. 15. Despite their chemical analogy, the tricyclic compounds **33**, **34** and **35** have different mechanisms of action and different therapeutic effects: **33** is an antiallergic agent, **34** is used for the treatment of schizophrenia and **35a** and **b** for the treatment of depression.

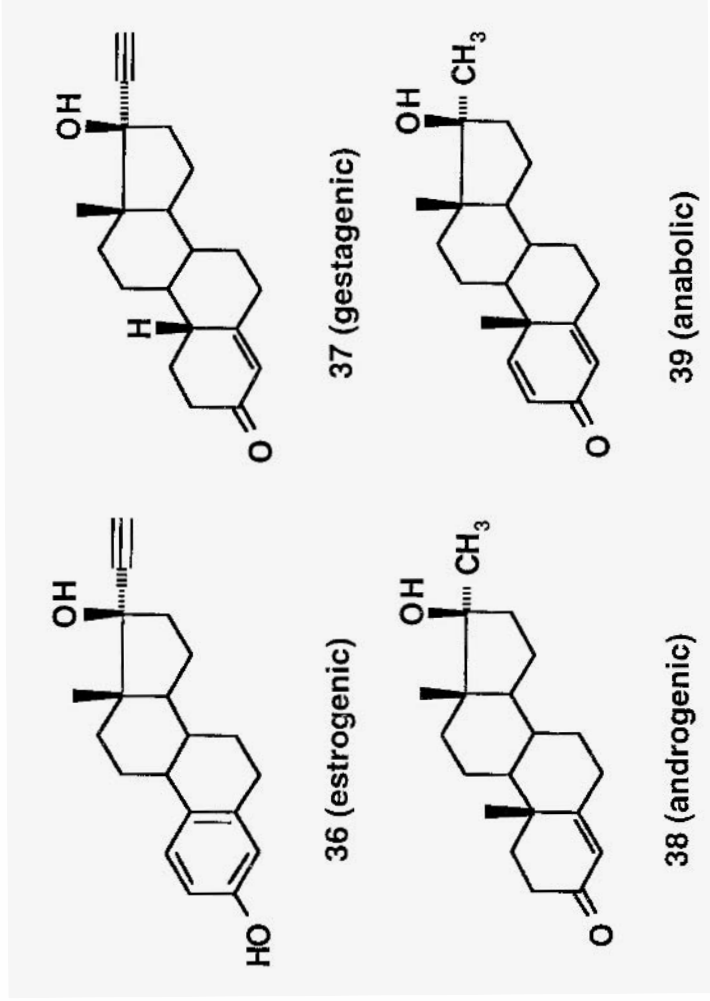


Fig. 16. In the steroid series, minor chemical differences produce very different effects on the biological properties of the compounds. Compounds 36–38 correspond in their biological actions to the different female and male sex hormones, compound 39 is a male sex hormone-related anabolic.

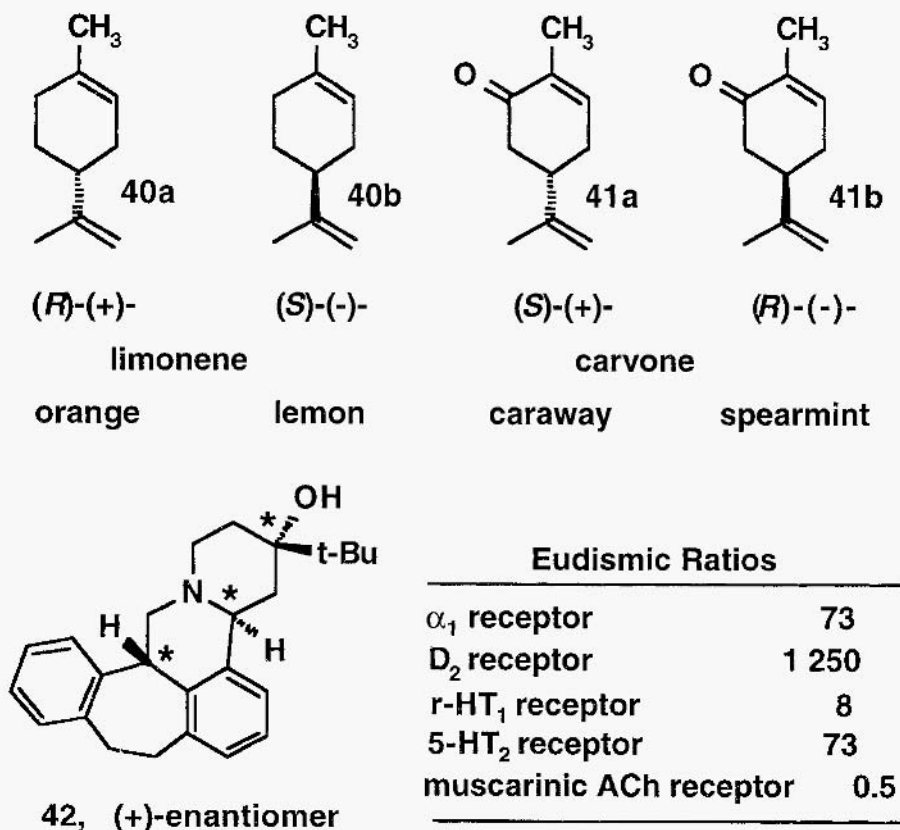


Fig. 17. Enantiomers are different: the olfactory receptors of our nose can even distinguish such minor structural differences as between **40a** and **b**, as well as between **41a** and **b** (the typical odors of the compounds are indicated below the structures). Activity differences between enantiomers are defined by their eudismic indices, the ratios of biological activities of the individual enantiomers. They significantly depend on the biological target. Whereas both enantiomers of butaclamol **42** look very 'similar' to the muscarinic acetylcholine receptor, they differ by more than three orders of magnitude in their affinities to the D<sub>2</sub> receptor.

Enantiomers may even have opposite biological effects. Certain chiral barbiturates show sedative activities in the one enantiomer and convulsive activities in the other one. The nifedipine analog Bay K 8644, **43** (Fig. 18) was originally synthesized as a racemate. *In vitro* tests with isolated smooth muscle strips showed it to be more or less inactive. However, several years later its high affinity to calcium channels was discovered: this demanded a separation of the racemate into the two enantiomers. One is a calcium channel agonist, the other a weak antagonist. Höltje explained these differences by the asymmetry of the molecular electrostatic potentials of the enantiomers, if the molecules are superimposed by their ring systems [49].



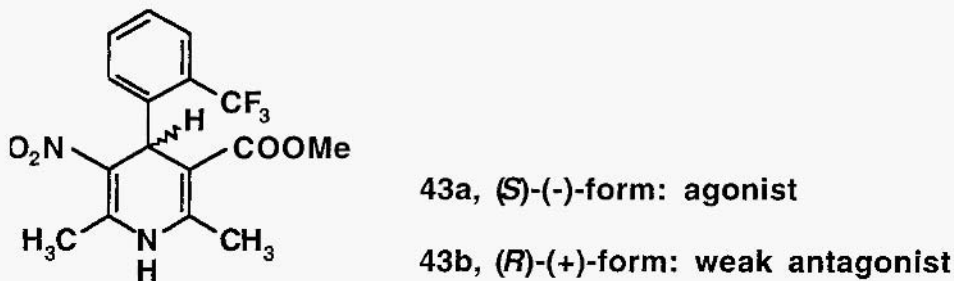


Fig. 18. The enantiomers of Bay K 8644 have opposite biological effects. Whereas **43a** stabilizes the calcium channel in its open form, **43b** stabilizes its closed form

## 8. The Similarity Principle in QSAR and 3D QSAR

Quantitative descriptions of the specific binding of ligands to a protein surface are based on the additivity principle of activity group contributions to the overall affinities. This has been extensively proven by dedicated investigations performed by Andrews, Goodford and Bohacek [50], but also by Böhm's scoring function of the *de novo* design program LUDI [51]. Some other approaches are reviewed in this book. However, despite the differences between enantiomers, discussed above, classical QSAR has no good tools to distinguish different conformers or enantiomers. 3D QSAR methods are much more efficient in this respect.

The very first approach to describe structure–activity relationships by  $N \times N$  similarity matrices ( $N$  = number of compounds) was presented by Rum and Herndon in 1991 [52]. More systematic investigations, using 3D similarity indices, were performed by Richards et al. in 1993 [53,54].  $N \times N$  Distance matrices are even appropriate for the quantitative description of nonlinear structure–activity relationships [50,55,56].

A very promising method, based on similarity fields, is the Comparative Molecular Similarity Indices Analysis (CoMSIA) approach [57,58]. As in CoMFA, all molecules are aligned according to a common pharmacophore: similarity indices of certain probes to the different molecules are then calculated in the positions of a regular grid. These fields, not just pairwise similarity indices of the molecules, are correlated with the biological activities. Due to the 'soft' character of the Gaussian potentials that are used to derive the similarity fields, smoother and more contiguous contour maps result from CoMSIA analyses [57].

As compared to the CoMFA and CoMSIA methods, the use of  $N \times N$  3D similarity matrices in QSAR has several advantages. Instead of a mutual alignment of all active and inactive molecules to a common reference frame, only pairwise alignments need to be performed [56,58]. This leads to a matrix with individual 'similarity vectors' for each analog. Inactive analogs do no longer distort the alignment of the active analogs. If SEAL similarity indices [59,60] are used in the generation of the  $N \times N$  matrix, not even a grid is needed. Thus, the frequently observed disturbance of the CoMFA results, after translation or rotation of the box around the molecules (e.g. [61]), are not ob-

served. A disadvantage of the calculation of pairwise similarity coefficients is the fact that no contour maps can be generated from the resulting QSAR model.

## 9. Conclusion

There are many lessons to be learned for 3D QSAR applications. Firstly, a correct alignment is the most important basis of each structure–activity analysis. Errors that are produced there can never lead to correct models.

An overreliance in target-independent similarity indices has to be questioned, because of the dependence of ‘similarity’ on the biological macromolecule to which the analogs bind. Sophisticated investigations on the ‘dissimilarity’ of chemical databases are most often futile. Similar compounds may have very different actions and different molecules can be very similar in their biological activities. Considering the examples presented in this chapter and the many more cases in the literature, one has to admit that we are far from a deeper understanding of the details which underline the observed structure–activity relationships. Applying the results from one series of analogs to another one may lead to completely wrong conclusions.

In the past, the use of molecular connectivity parameters in QSAR studies has led to highly controversial discussions. The same may happen to the BCUT, WHIM and EVA parameters, all described in other chapters in this volume. But one fact can be taken for sure: all these methods describe the similarity between molecules to a different extent. Thus, one should not be surprised that these approaches work, in some cases, even as well as other, more ‘rational’ methods.

Besides a correct alignment, the selection of the training and test sets is critically important to the results of a 3D QSAR study [58]. Only if both sets cover approximately the same range of structural space, can one be sure that the analysis and the results are meaningful. Cross-validation methods do not help in this respect. If the data set is highly redundant, even cross-validation in groups will produce ‘good’ results, whereas in clearly non-redundant datasets even a simple leave-one-out cross-validation must fail. One has to come back to the original goal of QSAR studies: to derive a working hypothesis and to design new analogs, based on one or several alternative working hypotheses. In this sense, QSAR is a theoretical tool which cannot solve all our problems but which definitely should accompany the experiments of the medicinal chemists and biologists.

## References

1. Böhm, H.-J., Klebe, G. and Kubinyi, H., *Wirkstoffdesign. Der Weg zum Arzneimittel*, Spektrum Akademischer Verlag, Heidelberg, 1996.
2. Wermuth, C.G. (Ed.), *The Practice of Medicinal Chemistry*, Academic Press, London, 1996.
3. Wolff, M.E. (Ed.), *Burger's Medicinal Chemistry*, 5th Ed., Vol. 1, John Wiley, New York, 1995.
4. Weber, L., Wallbaum, S., Broger, C. and Gubernator, K., *Optimization of the biological activity of combinatorial compound libraries by a genetic algorithm*, *Angew. Chem.*, 107 (1995) 2452–2154; *Angew. Chem. Intern. Ed.*, 34 (1995) 2280–2282.
5. Singh, J., Ator, MA., Jaeger, E.P., Allen, M.P., Whipple, D.A., Solowej, J.E., Chowdhary, S. and Treasurywala, A.M., *Application of genetic algorithms to combinatorial synthesis: a computational approach to lead identification and lead optimization*, *J. Am. Chem. Soc.*, 118 (1996), 1669–1676.

- Morgan, B.P., Holland D.R., Matthews, B.W, and Barlett, P.A., *Structure-based design of an inhibitor of the zinc peptidase thermolysin*, J. Am. Chem. Soc., 116 (1994) 3251–3260.
- Kaminski, J.J., Wallmark, B., Briving, C. and Anderson, B.-M., *Antiulcer agents: 5. inhibition of gastric H<sup>+</sup>/K<sup>+</sup>-ATPase by substituted Imidazo[1,2-a]pyridines and related analogs and its implications in modeling the high affinity potassium ion binding site of the gastric proton pump enzyme*, J. Med. Chem., 34(1991)513–541.
- Lauri, G. and Barlett, P.A., *CAVEAT: A program to facilitate the design of organic molecules* J. Comput.-Aided Mol. Design, 8 (1994) 51–66.
- Kaplan, A.P, and Barlett, P.A., *Synthesis and evaluation of tin inhibitor of carboxypeptidase A with a K<sub>i</sub> value in the femtomolar range*, Biochemistry, 30 (1991) 8165–8170.
- Barlett, P.A. and Marlowe, C.K., *Evaluation of intrinsic binding energy from a hydrogen bonding group in an enzyme inhibitor*, Science, 235 (1987)569–571.
- Morgan, B.P., Scholtz, J.M., Ballinger, M.D., Zipkin, I.D. and Barlett, P.A., *Differential binding energy: A detailed evaluation of the influence of hydrogen-bonding and hydrophobic groups on the inhibition of thermolysin by phosphorous inhibitors*, J. Am. Chem. Soc., 113 (1991) 297–307.
- Merz, K.M. and Kollman, P.A., *Free energy perturbation of the inhibition of thermolysin: Prediction of the free energy of binding of a new inhibitor*, J. Am. Chem. Soc., 111 (1989) 5649–5655.
- Shuman, R.T., Rothenberger, R.B., Campbell, C.S., Smith, G.F., Gifford-Moore, D.S. and Gesellchen, P.D., *A series of highly selective thrombin inhibitors*, In Smith, J.A, and Rivier, J.E. (Eds) Peptides — chemistry and biology proceedings of the 12th American Peptide Symposium. Cambridge, MA, U.S.A., 1991, ESCOM Science Publishers R.V., Leiden, 1992. pp. 801–802.
- Stanton, J.L., Ksander, G.M., de Jesus, R. and Sperbeck, D.M., *The effect of heteroatom substitution on a series of phosphonate inhibitors of neutral endopeptidase 24.11*, Bioorg. Med. Chem. Lett., 4 (1994) 539–542.
- Weber, A.E., Steiner, M.G., Krieter, P.A., Colletti, A.E., Tata, J.R., Halgren, T.A., Ball, R.G., Doyle, J.J., Schorn, T.W., Stearns, R.A., Miller, R.R., Siegl, P.K.S., Greenlee, W.J. and Patchett, A.A., *Highly potent. orally active diester macrocyclic human renin inhibitors* J. Med. Chem. 35 (1 992) 3755–3773.
- Wolfenden, R. and Kati, W.M., *Testing the limits of protein-ligand binding discrimination with transition-state analogue inhibitors*, Acc. Chem. Res., 24 (1991) 209–215.
- Xiang, S., Short, S.A., Wolfenden, R. and Carter, C.W., *Transition-state selectivity a single hydroxyl group during catalysis by cytidine deaminase*, Biochemistry, 34 (1995 ) 4516–4523.
- Parker, E.M., Grisel, D.A., Iben, L.G. and Shapiro, R.S., *A single amino acid difference accounts for the pharmacological distinctions between the rat and human 5-Hydroxytryptamine<sub>1B</sub> receptors*, J. Neurochem., 60 (1993) 380–383.
- Clozel, J.-P. and Fischli, W., *Discovery of reremikiren as the first orally active renin inhibitor*. Arzneim.-Forsch. (Drug Research), 43 (1993) 260–262.
- Li, R.-L., Hansch, C., Matthews, D., Blaney, J.M., Langridge, R., Delcamp, T.J., Susten, S.S. and Freisheim, J.H., *A comparason by QSAR, crystallography, and computer graphics of the inhibition of various dihydrofolate reductases by 5-(X-Benzyl)-2,4-diaminopyrimidines*, Quant. Struct. -Act. Relat., 1(1982) 1–7.
- Li, Z., Nguyen, D.T., Kitson, D.H., Bajorath, J., Kraut, J and Hagler, A.T., *Origin of trimethoprim's pharmacologic activity and differential binding to E. coli and chicken liver dihydrofolate reductases: Long-range electrostatic non-'lock and key' specificity*, Abstract of Presentations, Scientific Seminar Tour 1993, BIOSYM, San Diego, CA, U.S.A., 1993, pp. 14–19
- Roques, B.P., Nobel, F., Daugé, V., Fournié-Zaluski, M. and Beaumont, A., *Neutral endopeptidase 24.11: Structure, inhibition and experimental and clinical pharmacology*, Pharmacol. Rev., 45 (1993) 87–146.
- Roderick, S.L., Fournié-Zaluski, M.C., Roques, B.P and Mathews, B.W., *Thiorphan and retro-thiorphan display equivalent interactions when bound to crystalline thermolysin*. Biochemistry, 28 (1989)1493–1497.
- Slusarchyk, W.A., Robl, J.A., Taunk, P.C., Asaad, M.M., Bird, J.E., DiMarco, J. and Pan, Y., *Dual metalloprotease inhibitors: V. Utilization of bicyclic azepinothiazolidines and azepinonetetrahydrothiazenes in constrained peptidomimetics of mercaptoacyl dipeptides*, Bioorg. Med. Chem. Lett., 7 (1995) 753–758.

25. Hofmann, A., LSD — mein Sorgenkind. dtv/Klett-Cotta, Munich, 1993.
26. Hanson, D.J., *Dioxin toxicity: New studies prompt debate, regulatory action*, Chem. Eng. News, 12 August 1991. 7–14.
27. Mattos, C. and Ringe, D., *Multiple binding modes*, In Kubinyi, H. (Ed.) 3D QSAR in drug design: Theory methods and applications, ESCOM Science Publishers B.V., Leiden, 1993 pp. 226–254.
28. Meyer, E.F., Botos, I., Scapozza, L. and Zhang, D., *Backward binding and other structural surprises*, Persp. Drug Discov. Design, 3 (1993) 168–195.
29. Böhm, H.-J. and Klebe, G., *What can we learn from molecular recognition in protein-ligand complexes for the design of new drug?*, Angew. Chem., 108(1996) 2750- 2778; Angew. Chem. Intern. Edit.. 35 (1996), 2588–2614.
30. Montgomery, J.A. and Niwas, S., *Structure-based drug design*, Chemtech. 23 (1993) 30–37.
31. Montgomery, J.A., and Secrist III, J.A., *PNP Inhibitors*, Persp. Drug Discov. Design, 2 (1994) 205–220,
32. Kester, W.R., and Matthew., B.W., *Crystallographic study of the binding of dipeptide inhibitors to thermolysin: Implications for the mechanism of catalysis*, Biochemistry, 16 (1977) 2506–2516 .
33. Badger, J., Minor, I., Kremer, M.J., Oliveira, MA., Smith, T.J., Griffith, J.P., Guerin, D.M.A., Krishnaswamy, S., Luo, M., Rossmann, M.G., McKinlay, M.A., Diana, G.D., Dutko, F.J., Fancher, M., Ruechert, R.R. and Heinz, B.A., *Structural analysis of a series of antiviral agents complexed with human rhinovirus 14*, Proc. Natl. Acad. Sci. USA, 85 (1988) 3304–3308.
34. Diana, G.D., Treasurywala, A.M., Bailey, T.R., Oglesby, R.C., Pevear, D.C. and Dutko, F.J., *A model for compounds active against human rhinovirus-14 based on X-ray crystallography data*, J. Med. Chem., 33 (1990) 1306–1311.
35. Bystroff, C., Oatley, S.J. and Kraut, J., *Crystal structures of Escherichia coli dihydrofolate reductase The NADP<sup>+</sup> holoenzyme and the folate NADP<sup>+</sup> ternary complex: Substrate binding and a model for the transition state* Biochemistry. 29 (1990) 3263–3277
36. Bolin, J.T., Filman, D.J., Matthews, D.A., Hamlin R.C. and Kraut, J., *Crystal structure of Escherichia coli and Lactobacillus casei dihydrofolate reductase refined at 1.7 Å resolution*, J. Biol. Chem., 257 (1982) 13650–13662.
37. Poulos, T.L. and Howard, A.J., *Crystal structures of metyrapone- and phenylimidazole-inhibited complexes of cytochrome P-450<sub>cam</sub>*, Biochemistry, 26 (1987) 8165–8174.
38. Mattos, C., Rasmussen, B., Ding, X., Petsko, G.A. and Ringe, D., *Analogous inhibitors of elastase do not always bind analogously*, Nature. Struct. Biol., 1 (1994) 55–58,
39. Massumoto, O., Taga, T., Matsushima, M., Higashi, T. and Machida, K., *Multiple binding of inhibitors in the complex formed by bovine tryosin and fragments of a synthetic inhibitor*, Chem. Pharm. Bull., 38 (1990) 2253– 2255.
40. Underwood, D.J., Strader, C.D., Rivero, R., Patchett, A.A., Greenlee, W. and Predergast, K., *Structural model of antagonist and agonist binding to the angiotensin II, AT<sub>1</sub> subtype G protein coupled receptor*, Chem. Biol., 1 (1994) 211– 221.
41. Aquino, C.J., Armour, D.R., Bermann, J.M., Birkemo, L.S., Carr, R.A.E., Croom, D.K., Dezube, M., Dougherty, Jr., R.W. Ervin, G.N., Grizzle, M.K., Head, J.E., Hirst, G.C., James, M.K., Johnson, M.F., Miller, L.J., Queen, K.L., Rimele, T.J., Smith, D.N. and Sugg, E.E., *Discovery of 1,5-Benzodiazepines with peripheral cholecystokinin (CCK-A) receptor agonist activity: I. optimization of the Agonist 'Tigger'*, J. Riled. Chem., 39 (1996) 562–569.
42. Hirst, G.C., Queen, K.L., Sugg, E.E. and Willson, T.M., *Conversion acyclic nonpeptide CCK antagonist into CCK agonists*. Bioorg. Med. Chem. Lett., 7 (1997) 511–514.
43. Samanen, J., *GPV1Ib/IIIa antagonist*, Ann. Rep. Med. Chem., 31 (1996) 91–100.
44. Engleman, V.W., Kellogg, M.S. and Rogers, T.E., *Cell adhesion integrins us pharmaceutical targets*, Ann. Rep. Med. Chem., 31 (1996) 191– 200.
45. Aumailley, M., Gurrath, M., Müller, G., Calvete, J., Timpl, R. and Kessler, II., *Arg-Glv-Asp constrained within cyclic pentapeptides: Strong and selective inhibitors of cell adhesion to vitronectin and laminin fragment P1*. FEBS Lett. 291 (1991) 50–54.
46. Keenan, R., Miller, W., Ali, F., Barton, L., Bondinell, J., Burgess, J., Callahan, J., Calvo, R., Cousins, R., Gowen, M., Huffman, W., Hwang, S., Jakas, D., Ku, T., Kwon, C., Lago, A., Mombouyran, V., Nguyen, T., Ross, S., Samanen, J., Takata, D., Uzinskas, I., Venslavsky, J., Wong, A., Yellin, T. and

- Yuan, C., *Nonpeptide vitronectin receptor antagonists*, Abstract MEDI 236, 211th ACS National Meeting, 1996.
47. Ariëns, E.J., Wuis, E.W. and Vetinga, E.J., *Stereoselectivity of bioactive xenobiotics: A pre-Pasteur attitude in medicinal chemistry, pharmacokinetics and clinical pharmacology*, *Biochem, Pharmacol.*, 37 (1998) 9–18.
  48. Friedman, L. and Miller, J.G., *Odor incongruity and chirality*, *Science*, 172 (1971) 1044–1046.
  49. Hölftje, H.-D. and Marrer, S., *A molecular graphics study on structure-action relationships of calcium-antagonistic and agonistic 1,4-dihydropyridines*, *J. Comput.-Aided Mol. Design*, 1(1987) 23–30.
  50. Kubinyi, H., *QSAR: Hansch analysis and related approaches*. VCH, Weinheim, 1993.
  51. Böhm, H.-J., *The development of a simple empirical scoring function to estimate the binding constant for (a protein–ligand c complex of known three-dimensional structure*, *J. Comput.-Aided Mol. Design*, 8 (1994) 243–256.
  52. Rum, G. and Herndon, W.C., *Molecular similarity concepts: 5. Analysis of steroid-protein binding constants*, *J. Am. Chem. Soc.*, 113 (1991) 9055–9060.
  53. Good, A.C., Peterson, S.J. and Richards, W.G., *QSAR's, from similarity matrices: Technique validation and application in the comparison of different similarity evaluation methods*, *J. Med. Chem.*, 36 (1993) 2929–2937.
  54. Good, A.C., *3D molecular similarity indices and their application in QSAR studies*, In: Dean, P. (Ed.) *Molecular similarity in drug design*, Chapman and Hall, New York, 1995, pp. 23–56.
  55. Martin, Y.C., Lin, C.T., Hetti, C. and DeLazzer, J., *PLS analysis to detect nonlinear relationships between biological potency and molecular properties*. *J. Med. Chem.*, 38 (1995) 3009–3015.
  56. Kubinyi, H., *A General View on Similarity and QSAR Studies*, In *Computer-assisted lead finding and optimization*, Proceedings of the 11th European Symposium on Quantitative Structure–Activity Relationships, Lausanne, Switzerland, 1996: van der Waterbeemd, H., Testa, B. and Folkers, G. (Eds.): Verlag Helvetica Chimica Acta and VCH: Basel, Weinheim, 1997, pp. 7–28.
  57. Klebe, G., Abraham, U and Mietzner, T , *Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological potency*, *J. Med. Chem.*, 37 (1994) 4130–4146
  58. Kubinyi, H., Hamprecht, F.A. and Mietzner, T., *Three-dimensional quantitative similarity-activity relationships (3D QSiAR), from SEAL similarity matrices*, manuscript submitted for publication.
  59. Kearsley, S.K. and Smith, G.M., *An alternative method for the alignment of molecular structures: Maximizing electrostatic and steric overlap*, *Tetrahedron Comp. Methodol.*, 3 (1990) 615–633.
  60. Klebe, G., Mietzner, T. and Weber, F. *Different approaches toward an automatic alignment of drug molecules: application to sterol mimics, thrombin and thermolysin inhibitors*, *J. Comput.-Aided Mol. Design*, 8 (1994) 751–778.
  61. Cho, S.J. and Tropsha, A., *Cross-validated R<sup>2</sup> guided region selection for comparative molecular field analysis (CoMFA): A simple method to achieve consistent results*, *J. Med. Chem.*, 38 (1995) 1060–1066.

# Pharmacophore Modelling: Methods, Experimental Verification and Applications

Arup K. Ghose and John J. Wendoloski

*Amgen Inc., 1840 DeHavilland Dr., Thousand Oaks, CA 91320, U.S.A.*

## 1. The Concept of the Pharmacophore and its Validity

The essential functionalities of a molecule necessary for its pharmacological activity are called *pharmacophores*. Although the idea of pharmacophores existed for a long time, Ehrlich [1] first introduced this terminology following the term *chromophore* which was used to describe the groups responsible for the color of a compound. The interest in the idea of pharmacophores has grown tremendously in the last few decades due to the availability of computer graphics [2], a number of computational methods to determine the pharmacophoric geometry [3–7] and various software for 3D database mining using the concept of a pharmacophore pattern match [8]. However, the validity of a pharmacophore hypothesis came more from direct medicinal chemistry structure–activity relationships (SAR) studies rather than from any theoretical calculations. Such calculations may be possible in the near future with the availability of an ever increasing number of ligand–protein complex structures, better molecular mechanics force fields and better understanding of solvation factors. In such an approach, we have to show that the pharmacophoric groups provide the major contribution in binding affinity to the protein compared to the rest of the molecule. The main objective here is to discuss the various computational methods for pharmacophore identification and its geometry determination, a number of ways to validate the pharmacophore models and their applications in medicinal chemistry to identify novel active compounds. There are several excellent reviews and papers on pharmacophore modelling [3–7].

## 2. The Medicinal Chemistry Approach for Pharmacophore Determination and its use in Computational Chemistry

Pharmacophoric groups have been traditionally identified from the SAR data. When a compound with a novel biological activity is identified, medicinal chemists try slowly to modify this structure to optimize its potency, as well as to identify the important structural moieties responsible for the biological activity. The usual pharmacophoric groups include: (i) hydrogen (bond) donors, (ii) hydrogen acceptors, (iii) hydrophobic groups, (iv) electron acceptors, (v) electron donors and (vi) polar bonds, etc. This process of pharmacophore identification is illustrated with an example based on the hydroxamate inhibitors of the matrix metalloproteinases (MMPs) [5] collagenase and stromelysin. Collagenase cleaves collagen, the major constituent of cartilage, so blocking the activation of these MMPs is considered to be a potential strategy to prevent a tissue damage. The general structure of the peptidomimetic hydroxamate inhibitors of the MMPs are shown in Fig. 1. The pharmacophoric group determination and pharmacophoric atom selection are interrelated but may not be exactly the same. The atom selection varies

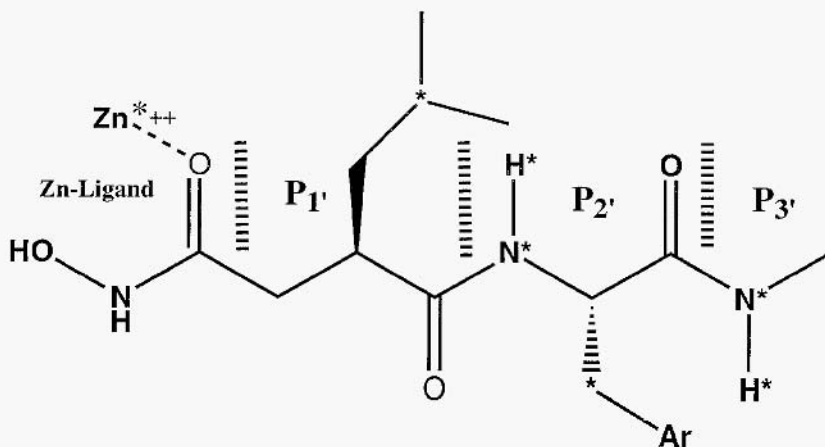


Fig. 1. The general structure and pharmacophoric groups of the peptidomimetic hydroxamate inhibitors. Pharmacophoric groups are indicated with an \*.

with our objective: when we are interested in the determination of the geometry of the pharmacophore atoms, we should select a minimum number of atoms from the pharmacophoric group. However, when we are interested in database searching for a pharmacophore, we should use a sufficient number of atoms which will match compounds from the correct chemical class. If the N–H bond of an amide is found to be a pharmacophoric group, one may select the N and H atoms as the pharmacophoric atoms for geometry determination. However, one should use the whole –CONH– group for database searching to avoid hits from amines. Keeping these differences in mind, we will analyze the SAR work done around this molecule and pharmacophoric atom selection of the hydroxamates described below.

### 2.1. A Zn ligand at the scissile peptide bond

Replacement of the scissile peptide bond by groups able to act as a Zn ligand is the basic approach that has been used in the design of collagenase inhibitors. The hydroxamate functionality is a very effective scissile bond surrogate in collagenase inhibitors [9,10]. Other Zn ligands like thiol, carboxylate and phosphonates also serve this function reasonably effectively. Removal of the Zn liganding group leads to a dramatic decrease in binding affinity, suggesting that it is a pharmacophore. When we are interested in the geometrical aspects of the pharmacophoric groups, it may be difficult to define the pharmacophore if we have different Zn liganding groups because the various ligands will have different coordination geometries around Zn. In the case where three fairly rigid Zn coordinating groups come from the protein, however, the complete coordination sphere is usually either a distorted square pyramid or a trigonal bipyramid. In this type of modelling, a dummy atom for Zn satisfying the geometry of the Zn–ligand coordination sphere [II] may be the best way to represent this pharmacophoric group. In this representation, the criterion of a valid superposition of different types of liganding

groups is that the Zn–ligand bond(s) will approach approximately from the same direction.

### *2.2. A hydrophobic group on the P1' amino acid*

It was known that the complete removal of the *iso*-butyl group led to a strong decrease in binding affinity. Decreasing the size to *n*-propyl led to a very small reduction in affinity [10]. This SAR indicated the importance of one of the terminal methyl groups for the activity. Since the two methyl groups are stereochemically different and there was no experimental evidence for selecting the correct methyl group, the second carbon was selected as one of the pharmacophoric atoms.

### *2.3. A hydrogen-bond donor at the P1'–P2'*

Replacement of the P1'–P2' amide linkage (–CONH–) by an ester (–COO–) or a retroamide (–NHCO–) led to inactive compounds, suggesting that –CO– may not be directly involved in the binding process with the MMPs. Methylation of the peptide N decreased the binding potency. These observations led to the selection of the N–H bond as a pharmacophoric group.

### *2.4. A hydrogen bond donor at the P2'–P3'*

Replacement of the P2'–P3' amide (–CONH–), as in (2.3), led to a considerable decrease in binding affinity. Two pharmacophoric atoms were used to represent each of these hydrogen donors to maintain the directionality of hydrogen-bond donation.

### *2.5. A hydrophobic alkyl or aralkyl group at the P2' side chain*

This increased the binding affinity considerably. Ala at P2', for example, showed considerably less potency [10] than Phe or Trp. This structure activity data led us to select the seven atoms as critical pharmacophoric atoms (Fig. 1).

## **3. Computational Methods for Pharmacophore Modelling**

The computational methods for pharmacophore modelling may have two steps: (i) determination of equivalent pharmacophoric atoms in a diverse set of active compounds; and (ii) pharmacophore geometry determination from a defined set of equivalent atoms in a diverse set of compounds.

### *3.1 Determination of the equivalent pharmacophoric groups in a diverse set of active compounds*

When the pharmacophore hypothesis originates from the SAR study of a single lead compound, definition or identification of the equivalent pharmacophoric groups may not be difficult. However, if we have several active compounds for the same biochemical or



biological system with a diverse set of structures. Identification of the equivalent pharmacophoric groups may not be as straightforward. Any automation along this line should have the strategies: (i) to identify similar atoms or groups in terms of their chemical or physico-chemical properties; (ii) to estimate the relative position of the possible pharmacophoric groups in the allowed low-energy conformations of the molecules for 3D structural comparison; and (iii) to weigh similar pharmacophoric groups when there are multiple choices. Although, from time to time, some success along this line has been claimed, [12–16], none of these methods have been well tested for diverse types of problem. In one of the earliest attempts, Crippen [12] abstracted the ligands by a set of representative points of specific types. The conformational flexibility of the molecules was taken into account by a distance range matrix of upper and lower limits of the distances between these points in the energetically allowed conformations of the molecules. He then combinatorially determined the maximum number of matching points of the same type. The geometrical requirement for the matching of the points is that any two matching points should have either overlapped distance ranges or the distance ranges should be within a specified distance limit. Once an acceptable match has been found, the common distance range of the matched atoms can be used in the distance geometry embedding technique [17] to find the active conformation of the individual molecules. Ghose et al. [13,14] used a set of important atoms from each molecule for a combinatorial exploration of a plausible superposition. The conformational flexibility was accounted for by taking a set of low-energy conformations separately. Superpositions of all other atoms were determined from the initial superposition of the important atoms. Each superposition was rated by a *fit function* utilizing the three most important physico-chemical properties of the atoms, namely charge, refractivity and hydrophobicity. Atomic charge density is related to the electrostatic interaction. Atomic refractivity is related to the van der Waals interaction, and atomic logP is related to hydrophobic interactions. These are the three most important forces in the binding of the ligands with the biological molecules. Martin et al. [15], in their DISCO program, automated and improved most of Crippen's approach using a selected number of low-energy conformations. Jones et al. [16] used a very similar approach, except that their fit function was optimized using the genetic algorithm (GA) where the *gene* was represented by atom match, as well as the conformational information. The general utility of these methods may be determined in the future when independent workers will apply these methods to a diverse set of problems

### 3.2 Pharmacophore geometry determination with a defined superposition hypothesis

Most of the generally explored methods for pharmacophore modelling belong to this category, since quite often the pharmacophore hypothesis can be formulated from traditional SAR studies. The objective of these methods is straightforward: to find all possible low-energy conformations of the molecules where the pharmacophoric groups can be superimposed within a predefined distance limit. If a method tries to find any one acceptable solution, its success may be limited, since it may not be the real solution. On the other hand, the method which tries to get all possible solutions may sometimes be

computationally intractable. Some of the methods reported to solve this problem are shown below.

### *3.2.1. Ensemble distance geometry*

Sheridan et al. [18] formulated an elegant way of using Crippen's distance geometry program for superimposing molecules. In this approach, they created an ensemble distance matrix of all atoms of all the molecules that are to be superimposed. The intramolecular atomic distance ranges are given the usual distance ranges for that molecule. Many of these distances are simply the bond distances, bond angle distances and fixed torsion distances. The intermolecular atomic distance ranges are zero to a small tolerance distance for the atoms to be superimposed. The rest of the distance ranges are evaluated by the embedding program using the triangular inequality rule. The method has some advantages over Crippen's original approach where the common distance range of the superimposed atoms were used to evaluate their geometry. The constraints of the non-superimposed atoms further decreased the acceptable solution space.

However, it is not completely free from all the disadvantages of regular distance geometry, namely when there are multiple solutions of conformations, it does not give any information about the other acceptable solutions, nor does it guarantee low-energy conformations. This method is computationally very inefficient too. One has to diagonalize an  $(n*m) \times (n*m)$  matrix, where  $n$  is the number of molecules and  $m$  is the number of atoms in each molecule. One can get a comparable result by diagonalizing the  $m \times m$  atomic distance matrix of each molecular separately with the common distance range for the overlapping atoms. However, the simplicity of the data preparation justifies the extra computation for this method.

### *3.2.2. Distance mapping in torsional space*

Earlier, Crippen introduced [19] the method of using a distance range matrix of the atoms for a fast evaluation of the superimposability of the molecules. Distance range was determined from the minimum and maximum distances between these atoms in the acceptable low-energy conformations. In this method, the criterion for superimposition is that the superposed atom should have an overlapping distance range. This criterion is a negative test, rather than a positive test; in other words, if this test fails, the molecules cannot be superimposed. However, qualifying this test does not guarantee a superposition. The conclusions drawn from the distance range matrix are valid only in an  $n - 1$  dimensional space, where  $n$  is the number of atoms superimposed [20]. Our physical space has only three-dimensions. One can get a structure in a three dimensional space which can best represent the distance range using distance geometry embedding. However, distance geometry embedding is not computationally fast enough to be applied in a large number of test cases.

Another problem of embedding is that the structures obtained from simple embedding may not be energetically low. To overcome these problems, Ghose and Crippen [21] proposed the construction of a distance map in torsional space. In this approach, an ensemble of distance matrices, each one representing an energetically allowed grid of torsional space, is created (Fig. 2). The grid is determined by the rotational increment of



### 3.2.3. Orientation map in distance space

Marshall et al. used an alternative approach where they mapped the torsion angles in distance space [22]. The generation of a torsion angle map is computationally more efficient since once a grid size has been defined, one can very easily assign a grid to each of the energetically allowed conformations during the conformational search according to its distance property. A schematic representation of torsional (orientation) mapping in the distance space is represented in Fig. 3. The simplest approach to test the feasibility of superimposition of a set of equivalent atoms in two molecules is to compare their orientation map. If a grid is occupied by both the molecules, those conformations are superimposable if their chirality is the same [2]

The weakness of this approach is that the outcome is dependent on the distance grid size. If one arbitrarily uses a very large grid size, most conformations will be placed in a few grids and will suggest many poorly superimposable conformations as superimposable [21] Too small a grid size will create many unoccupied grids and may suggest many fairly superimposable conformations as non-superimposable simply because they occupied a close neighboring grid. The ideal grid size that can avoid

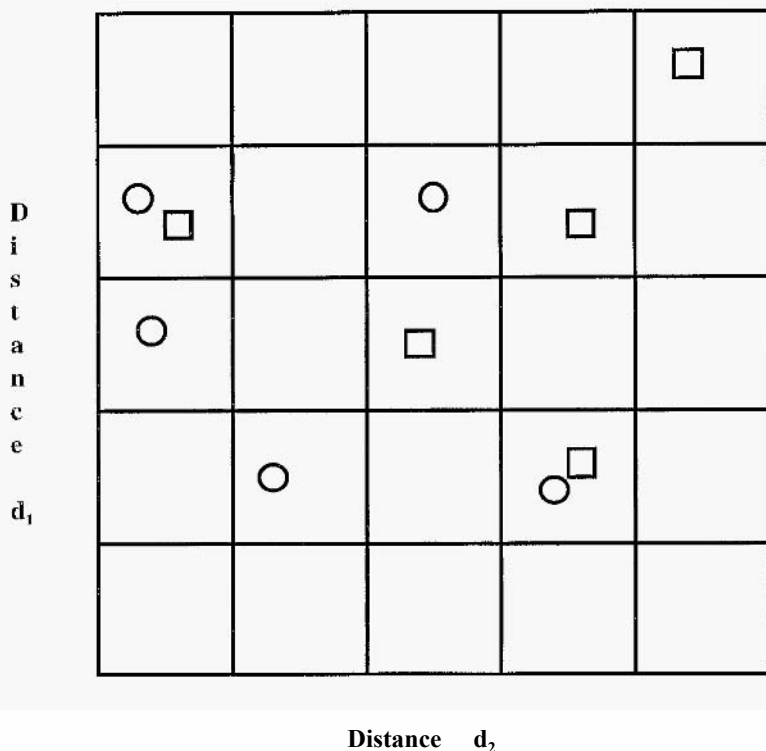


Fig. 3. A schematic two dimensional representation of an orientation map in the distance space. The circles and the squares represent the conformations of two different molecules satisfying the distance range of the occupied grids, and  $d_1$  and  $d_2$  are the two pharmacophoric atom distances.

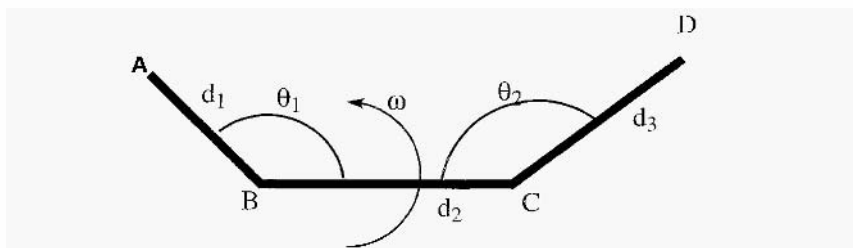


Fig. 4. Description of the various internal coordinates affecting the AD distance, as used in equations 1 and 2.

during a torsional angle increment in the conformational search process. The distance between atoms A and D during the rotation around bond B-C (Fig. 4) is given by:

$$r^2 = d_1^2 + d_2^2 + d_3^2 - 2d_1d_2\cos\theta_1 - 2d_2d_3\cos\theta_2 + 2d_1d_3(\cos\theta_1\cos\theta_2 - \sin\theta_1\sin\theta_2\cos\omega) \quad (1)$$

where  $d$ 's,  $\theta$ 's and  $\omega$  are the various bond distances, bond angle (Fig. 4).

The differentiation of  $r^2$  with respect to  $\omega$  gives:

$$(\delta r^2 / \delta \omega) = 2d_1d_3\sin\theta_1\sin\theta_2\sin\omega \quad (2)$$

Equation 2 shows that the rate of change in distance between A and D depends on  $d_1$  and  $d_3$ . The higher the values, the greater is the rate of change. In addition, the change is maximal when the bond angles as well as the torsion angle are  $90^\circ$ . Unfortunately, the pharmacophoric atoms A and D are often not directly bonded to B and C and, therefore, all of these variables may have a wide range. In other words, the appropriate grid spacing is a function of several variables and will differ from one atom pair to another and even from one conformation to another.

Creating a distance map from torsional space [21] was in a better position from these difficulties, because the distance range tends to change according to the increment step size of the torsion angles and the other variables, as shown in Eq. 1. The precision of the superimposed structure was determined by the coarseness of the torsional increment during the conformational analysis.

### 3.2.4. Disco program

In Disco, Martin et al. [15] automated the identification of possible pharmacophoric groups in a molecule and applied Kuhl et al.'s [23] clique finding algorithm to determine their superimposability. They, however, used a set of distance matrices representing a few low-energy conformations of the molecules as used by Marshall et al. [3] and Ghose et al. [13,14] instead of a distance range matrix. Since it used a distance tolerance for accepting a distance match in a discrete distance representation of a flexible molecule, it often had to redo the matching with increased tolerance when no matching was

found with at least three pharmacophoric groups. The method was successful in identifying the traditionally accepted pharmacophores of the benzodiazepines, as well as in fitting some newer 2-aminothiazoles in the traditional pharmacophores. Disco is an attractive tool for generating alternate possible pharmacophore hypotheses. It is now commercially available [24] and it is expected to be extensively explored by the researchers of this field in the near future.

### 3.2.5. Genetic algorithm (GA) based approach

Payne et al. probably was the first to apply [25] GA for pharmacophoric geometry determination. Genetic algorithm mimics the process of evolution where new generations are born by *crossover* and *mutations* of genes and the *fittest members* survive over the time. Any optimization process can be used by the GA programs by encoding the independent variables in a string of bits called *chromosomes*. The fitness of each chromosome can be measured by a function. An initial population is randomly created and the fitness of its members are evaluated. A set of reproduction operators (*crossover*, *mutation*, etc.) is used to create the *children*. The fitness of the children is evaluated. The children replace the least-fit members of the population. The process continues as long as it finds better-fitted children. Here the conformational variables were encoded in a bit string and a fitness function was represented by the distance of the superimposed atoms. It is definitely more efficient than a Monte Carlo-type search. However, when there are multiple superposable conformations, it may fail to detect all of them. There is no guarantee that the method will find the global minimum. Sometimes convergence may be a problem also.

### 3.2.6. DHYDM method

The DHYDM (*Distance Hyperspace Distance Measurement*) [S] method was developed by Ghose et al. to avoid the grid size problem of the orientation-mapping approach developed by Marshall et al. [22]. This method involves the following steps.

- (i) Do a conformational search and generate an orientation map (Fig. 3) for each molecule independently. The conformational analysis and orientation map generation can be accomplished by Sybyl software [24].
- (ii) Determine the distance of the centroid of the occupied grids of molecule 1 from those of the  $i$ 'th (initially 2nd) molecule in the distance space and store the minimum distance for each grid using the following expression:

$$d_{ij} = \sum (dc_{i,k} - dc_{j,k})^2$$

where  $d_{ij}$  represents the diatomic between the  $i$ 'th occupied grid of the molecule 1 and the  $j$ 'th occupied grid of molecule  $n$  in distance hyperspace, and  $dc_{i,k}$  represents the distance of the center of the  $i$ 'th grid along the  $k$ 'th dimension. It is simply the mean of the minimum and maximum distances of the grid along the  $k$ 'th dimension.

- (iii) Repeat step (ii) for the rest of the molecules. If the minimum distance is greater than the stored value, replace it by the current minimum.

- (iv) Rank-order the minimum distance values and accept those grids which found occupied grids within an acceptable limit.
- (v) Accept the conformation in the grid if a rigid fit gives a good rms deviation of the pharmacophoric atoms from a conformation of the reference molecule in that grid. Since distances do not contain the chirality information (enantiomeric conformations have the same distance properties), this step may be very important.

Unlike most other methods, the advantages of this method are:

1. The outcome of this experiment can be monitored to set an upper limit on the molecular mechanics energy of an acceptable superimposable conformation. This may be important if the best solution for a low-energy cutoff is not geometrically acceptable.
2. When the molecules are not sufficiently diverse in their conformational behavior and, therefore, are superimposable in many different low-energy conformations, the method will detect such a situation and will inform the investigator about multiple possibilities. In the absence of other criteria for the acceptance of these possible conformations, one can prioritize on the basis of conformational energy.
3. This method also accounts for the complete low-energy conformational space and not just the grid points explored in the search process. The computational burden of this method, however, precludes its use for a fast but approximate 3D database search [8].

The main disadvantage of this method is the computational speed. It needs a thorough conformational search for all molecules. When there are many torsion angles, the method may still be used by increasing the angle increment and setting the unimportant torsion angles to a local minimum.

#### **4. Experimental Verification of a Pharmacophore Geometry**

One obvious major problem with computer modelling is that it is not an experimental result, even though the input data was supplied from experimental findings! To improve confidence in the result, appropriate steps should be taken to validate the pharmacophore model. Several approaches can be used to validate the pharmacophore model.

##### *4.1. Analyzing binding affinities of comparable compounds*

Analysis of the binding affinities of the compounds which can/cannot attain the pharmacophore geometry in the various allowed low-energy conformations may give an indirect support to the pharmacophore model. In general, if there is a low-energy conformation of the molecule satisfying the pharmacophore geometry, it should be active, otherwise not. However, this experiment may not be very conclusive since a molecule having a perfect match for the pharmacophore may suffer from steric repulsion with the binding protein from the other parts of the molecule. Some pharmacophores may interact less strongly with the protein and give detectable binding without proper alignment or even in its absence. Some molecules may bind in a totally different binding mode.

#### 4.2. Analyzing constrained compounds

Making constrained compounds to validate a pharmacophore theory is one approach often used by medicinal chemists. The constrained compounds with appropriate pharmacophoric groups, in theory, should have a higher binding affinity due to smaller entropy loss upon binding with the protein. In reality, it may not happen the expected way because the accepted or computed pharmacophore geometry may not be the ideal geometry or for the reasons already discussed. In an alternate view, if the constrained compound is not totally free from torsional rotation, positive binding may not always confirm the pharmacophore geometry.

#### 4.3. Biophysical determination of the binding conformation

Is it possible to determine the binding conformation of the ligand without the tedious process of crystallizing and solving the ligand-protein complex structure? High-resolution NMR spectroscopy [5] may often be a good complementary tool for this purpose. This method needs appropriate isotopically labeled inhibitors for generating NOEs. The NOE constraints are determined in presence and absence of the binding protein. Wang et al. [5], for example, used an  $^{15}\text{N}$ -edited and decoupled NOESY spectrum of compound Phe analog of **I** (Fig. 5) in the presence and absence of human fibroblast collagenase (HFC). Analysis of NOE peaks derived from the labeled NH of the bound inhibitor, they inferred that there was a methyl group within a short distance from the labeled NH, possibly less than 3 Å. Since there is only one label in this compound, it was not possible to distinguish intermolecular and intramolecular NOEs. The NOESY showed that the two prochiral  $\beta$  protons have strong and comparable NOE intensities. On the other hand, the cross-peak with the phenylalanine *ortho* protons is very weak. These results suggest that the two  $\beta$  protons are at similar distances from the NH and the phenyl ring is probably in the *trans* position to it. This observation was consistent with the suggested active conformation of **I**.

#### 4.4. Biophysical determination of the ligand-receptor complex structure

The most realistic test of the pharmacophore hypothesis involves the solution of the protein–ligand crystal structure either by X-ray crystallography or NMR [5], although we do not very often get the chance to perform such an elaborate experiment. This approach needs a sufficient quantity of purified protein. The NMR study needs isotopically labeled proteins. The X-ray method needs good-quality crystals of the ligand–protein complex. In the current industrial drug-research setting, we often start with a target protein without a known 3D structure, and a long lead time before the structure is solved. In absence of the protein structure, pharmacophore identification is a common approach in drug research. As a consequence, in the future we will get more examples where ligand–receptor complex structures will support or refute pharmacophore hypothesis. For the collagenase (HFC) inhibitors, Ghose et al. used their DHYDM method and a set of partially constrained and unconstrained inhibitors (Fig. 5)



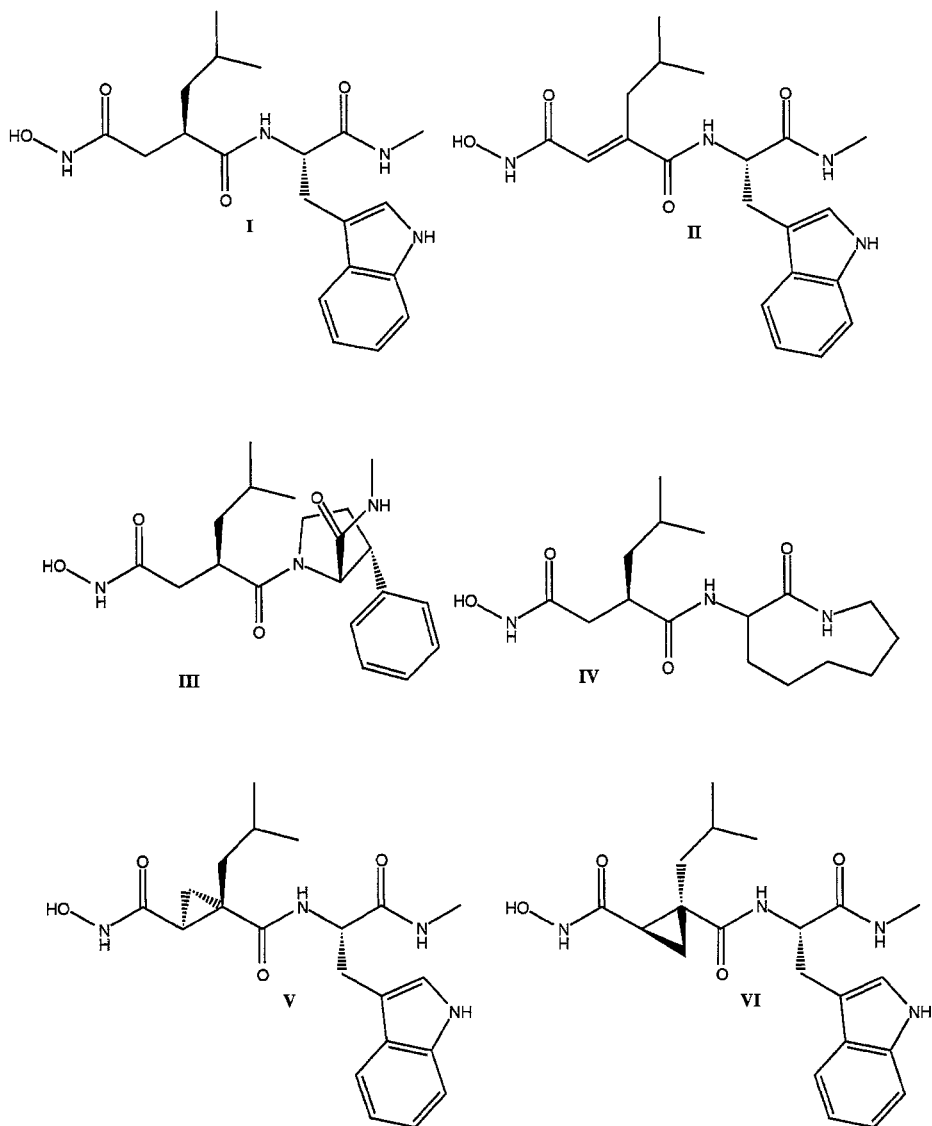


Fig. 5. The structures of the various constrained and unconstrained collagenase inhibitors used for the pharmacophore geometry determination.

to propose a pharmacophore geometry and, finally, solved a collagenase–ligand complex structure using X-ray crystallography. Ten of the eleven torsion angles of the bound conformation of the ligand were within an acceptable error limit (Table 1). A stereoview of the computed pharmacophore model and the X-ray structure of a related bound ligand are shown in Fig. 7.

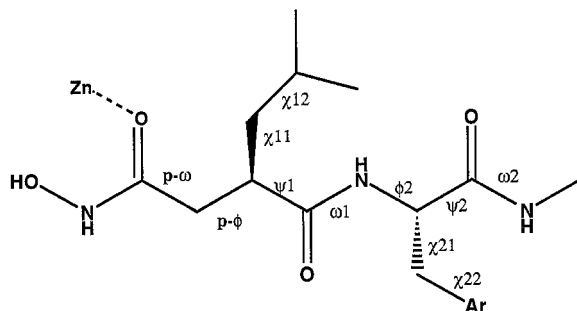


Fig. 6. Description of the torsion angles as used in Table 1.

Table 1 Comparison of the computed pharmacophore model with the X-ray structure of the protein–ligand complex

Angle name <sup>a</sup>	Pharmacophore model <sup>b</sup> (angles in degrees)	X-ray	$\Delta(P - X)^d$
p- $\omega$	91–129 (110)	253	153
p- $\phi$ 1	153–178 (166)	183	17
$\psi$ 1	89–148 (120)	110	10
$\omega$ 1	177–185 (180)	191	11
$\phi$ 2	276–302 (293)	271	22
$\psi$ 2	299–332 (307)(146) <sup>e</sup>	132	14
$\omega$ 2	179–192 (183)	178	5
$\chi$ 11	169–185 (177)	190	13
$\chi$ 12	146–202 (176)	172	4
$\chi$ 21	180 <sup>e</sup>	183	7
$\chi$ 22	90	73	17

<sup>a</sup>See Fig. 6 for the description of the torsion angles.

<sup>b</sup>The angle range is given from the first experiment; the values within parenthesis are the average value in different inhibitors.

<sup>c</sup>The value suggested from the alternate conformation of trans nine membered lactam in the Cambridge Structural Database.

<sup>d</sup>The difference from the mid-point of pharmacophore model range.

<sup>e</sup>The value from the unconstrained acyclic compound I only.

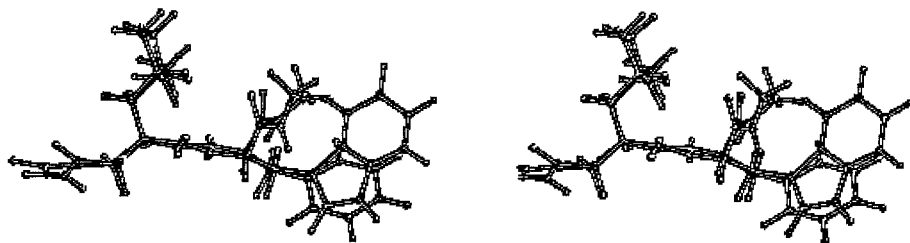


Fig. 7. A stereoview of the computed pharmacophore model superimposed on the X-ray structure of a related bound ligand.

#### 4.5. Biophysical determination of the conformation of the free ligand

In the recent literature, there were a considerable number of papers which studied the conformations of a set of bioactive compounds either by X-ray [26–30] or by NMR [31–35] to propose a pharmacophore model. Whether such methods are justified is arguable, since it is the protein or biological molecules which dictate the binding conformation, sometimes with minor adaptation on its own. A ligand will bind to the biomolecule provided it can take a complementary conformation with a low expense of energy. In other words, the binding conformation may not be the most populated conformation in the free state. Nicklaus et al. [36] recently studied the various ligands available in the protein databank. Many of these compounds were also available in the Cambridge structural database in their unbound state. None of these structures resembled the bound conformations. Only a careful analysis of the structures of multiple compounds in their unbound conformation may shed light on their binding conformation.

### 5. Applications of Pharmacophore Modelling

Pharmacophore modelling is an excellent research tool in drug research. It is useful at the very early stage when no lead has been generated, but we know a natural substrate as well in the later stage when a ligand–receptor complex structure has been solved. Being a qualitative approach, one does not need to have very good-quality binding data. As the organic synthetic approaches become more automated, pharmacophore modelling is likely to be used more extensively. We will discuss here the most common applications of pharmacophore modelling.

#### 5.1. Pharmacophore model and database searching

The most common and useful application of pharmacophore modelling is in database mining. In recent years, a large number of very useful databases have been supplied by the chemical software companies, such as ACD (*Available Chemical Directory*), MDDR (*Molecular Drug Directory Report*) and CMC (*Comprehensive Medicinal Chemistry*) [37]. The development of the reliable methods of conversion of 2D to 3D structures enabled the creation of these and proprietary 3D databases. These 3D databases can not only be searched for traditional substructures, but also for a three-dimensional pharmacophoric geometry. One can do such searches under various conditions. When several lead compounds are available, one can try to develop a uniquely defined 3D pharmacophore hypothesis using one of the approaches discussed above, and subsequently, use that information for searching novel leads. When only one lead structure is available, the possible pharmacophoric groups may be searched for a few of its low-energy conformations. When such a lead is unavailable, natural substrates may be used. However, a 3D database search may be complicated by the fact that most organic molecules are flexible. The choice of strategy used to cover the flexibility may change with the problem. For high-throughput screening, one may simply use a distance range matrix for this purpose. A comparison of the distance range matrix

(being a *negative test*, see section 3.2.2) will pick a certain number of compounds that will not be able to satisfy the pharmacophore geometry. But absorbing those false hits may not be difficult in a high-throughput screening.

More difficult screening may need more careful selection of the compounds. One such approach was used in the Galaxy 3D database [38]. The 3D database here was created after a rapid conformational search capable of identifying a very low-energy conformation, if not the global minimum energy conformation. The advantages of keeping a very low-energy conformation are two-fold. If a rigid matching is done, the hit will be valid since the conformation used is an accessible conformation. When a flexible search is desired, the conformation can be computed directly from the distance geometry embedding program by setting the required pharmacophoric distances along with the bond distance and bond angle distances. Comparison of the energy of the computed conformation (if available) with the energy of the conformation in the database will immediately prompt its energetic acceptance.

### 5.2. *Pharmacophore model and 3D QSAR*

Most 3D QSAR techniques [13,39–41] need to superimpose the 3D structures of the inhibitors as the first step. The pharmacophore modelling may be used as the basis for such superposition. In general, there are two types of 3D QSAR approaches. The physico-chemical property-based approaches like REMOTEDISC [13,40] where the local properties are clustered spatially from the superimposed 3D structures of the ligands and the clustered properties are correlated with the biological activity. Such methods give a physical interpretation of the nature of the hypothetical binding pocket of the protein. The field-based approaches like COMFA [41] calculate the interaction of the ligand molecules with atoms representative of the protein atoms at an arbitrarily defined set of grid points. The interaction energies are correlated with the binding affinity to develop a comparable interpretation of the nature of the protein-binding pocket. What comes out of these approaches depends on the conformation used, as well as the mode of superimposition of the ligands. Although COMFA does not supply any definite algorithm for the superimposition of the ligands, it is definitely a good idea to use the pharmacophore modelling to initiate the COMFA. Hopfinger's molecular shape analysis [42] and APEX-3D [43] often used pharmacophore hypothesis to a 3D QSAR analysis.

### 5.3. *Pharmacophore model and de novo design*

Pharmacophore models can be used for a *de novo* design of novel compounds. The structural moieties to connect the pharmacophore may be selected by using the attachment bond vectors of the pharmacophore and searching a 3D database of molecular structures [44] or using some standard spacers (molecular fragments) from a library [45]. These approaches definitely help medicinal chemists to get ideas about novel structural class of compounds. However, since most often such compounds have to be synthesized, the utility of these programs will be less compared to database searching programs where one can get 'ready-made' compounds.

#### 5.4. Pharmacophore modelling and combinatorial library design

The most useful combinatorial library design programs can analyze the reactants in terms of its diversity. It may be debated whether the diversity will be in terms of key physico-chemical properties like hydrophobicity, polarizability, volume, etc. or in terms of the nature of the pharmacophoric groups or both. Diversification, in terms of the nature of the pharmacophoric groups alone, is also an excellent way to probe the binding site [46]. In general, while designing a 'universal library', one should diversify the nature of the possible pharmacophoric groups of the whole molecule. The *focused libraries* should be diversified for each and every combinatorial substituents.

### 6. Critical Aspects and Comments

One should be critical while using a pharmacophore hypothesis, Constraining the conformation often leads to a loss of binding affinity. although entropy gains arising from the reduced conformational distribution should facilitate binding. Problems in such molecules may be due to: (i) constraining the conformation at an angle somewhat distant from the ideal angle; and (ii) bad interactions of the constraining structural moiety with the enzyme binding site. It is not certain how much drop in activity of a molecule should be acceptable while developing or validating the model.

The idea of pharmacophoric modelling does not hold if the inhibitors show a considerable activity, even when one or more pharmacophoric groups do not reach the same region of the active site. Forcing the molecules to attain a conformation where equivalent groups occupy the same location in such a situation may give a distorted pharmacophoric model. However, success in the drug-design process is so rare that the researchers in this area are often eager to take these risks. The only suggestion to be offered here may be to analyze the suggested (computed) conformation with the existing knowledge of conformation of similar molecules.

The pharmacophore modelling is based on the idea that similar inhibitors bind in the same way at the active site. The X-ray crystallographic data of most ligand–protein complexes usually confirm this hypothesis. However, there are several exceptions to this basic idea [47–49]. Multiple-binding mode may often result in a binding pocket where non-directional forces, like van der Waals or hydrophobic, are dominant. The application of pharmacophore modelling in such a system may be risky.

Staying close to a lead compound maximizes the chance of finding an active compound. Unless it is necessary to get a very different compound (maybe for patentability), one should try to make the smallest change from the lead compound while searching a database.

In general, pharmacophore modelling is more useful than QSAR approaches at the initial phase of drug discovery. Also it is more often applied since it does not need a very good-quality biological activity or binding data. Unfortunately, such a relaxed condition may easily lead to a misuse of various methods in this area.

## Acknowledgements

The authors want to thank Dr. Tim Hervey for a careful reading of the manuscript and for the various suggestions.

## References

1. Ehrlich, P., *Über den jetzigen Stand der Chemotherapie*, Chem. Ber., 42 (1909) 17.
2. Marshall, G.R. and Naylor, C.B., *Use of molecular graphics for structural analysis of small molecules*, In Hansch, C., Sammes, P.G., Taylor, J.B. and Ramsden, C.A. (Eds.) *Comprehensive medicinal chemistry*: Vol. 4, Pergamon Press, Oxford, 1990, pp. 431–458.
3. Marshall, G.R., Barry, C.D., Bosshard, H.E., Dammkoehler, R.A. and Dunn, D.A., *The conformational parameter in drug design*, Computer Assisted Drug Design, ACS Symp. Ser. 112, 1979, pp. 205–226.
4. Gund, P., *Pharmacophoric pattern searching and receptor mapping*, Ann. Rep. Med. Chem., 14 (1979) 299–308.
5. Chose, A.K., Logan, M.E., Treasurywala, A.M., Wang, H., Wahl, R.C., Tomezuk, B.E., Gowravaram, M.R., Jaeger, E.P. and Wendoloski, J.J., *Determination of pharmacophoric geometry for collagenase inhibitors using a novel computational method and its verification using molecular dynamics, NMR and X-ray crystallography*, J. Am. Chem. Soc., 17 (1995) 4671–82.
6. Humblet, C. and Marshall, G.R. *Pharmacophore identification and receptor mapping*, Ann. Rep. Med. Chem., 15 (1980) 267–276.
7. Klebe, G., *structural alignment of molecules* In Kubinyi, H. (Ed.) 3D QSAR in drug design: theory, methods and applications, ESCOM, Leiden, 1993, pp. 173–199.
8. Clark, D.E., Willet, P. and Kenny, P.W., *Pharmacophoric pattern matching in file of three-dimensional chemical structures: Use of smoothed bounded distances for incompletely specified query patterns*, J. Mol. Graph., 9 (1991) 157–160.
9. Johnson, W.H., Roberts, N.A. and Borkakoti, N., *Collagenase inhibitors: Their design and potential therapeutic use*, J. Enzyme Inhib., 2 (1987) 1–22.
10. Schwartz, M.A. and Van Wart, H.E., *Synthetic inhibitors of bacterial and mammalian interstitial collagenase*, Prog. Med. Chem. 29 (1992) 271–334.
11. Matthews, B.W., *Structural basis of the action of thermolysin and related Zn peptidases*, Acc. Chem. Res., 21(1988) 333–340.
12. Crippen, G.M., *Quantitative structure–activity relationships by distance geometry: systematic analysis of dihydrofolate reductase inhibitors* J. Med. Chem. 23 (1980) 599–606.
13. Ghose, A.K., Crippen, G.M., Revankar, G.R., McKernan, P.A., Srnicek, D.F. and Robins, R.K., *Analysis of the in vitro antiviral activity of certain ribonucleosides against parainfluenza virus using a novel computer aided receptor modeling procedure*, J. Med. Chem., 32 (1989) 746–756.
14. Viswanadhan, V.N., Ghose, A.K., Revankar, G.R. and Robins, R.K., *Atomic physicochemical parameters of three-dimensional structure directed quantitative structure–activity relationships: 4. Additional parameters for hydrophobic and dispersive interactions and their application for an automated superposition of certain naturally occurring nucleoside antibiotics* J. Chem. Inf. Comput. Sci., 29 (1989) 163–172.
15. Martin, Y.C., Burea, M.G., Danaher, E.A., DeLazzer, J., Lico, I. and Pavlik, P.A. *A fast new approach to pharmacophore mapping and its application to dopaminergic and benzodiazepine agonists*, J. Comput.-Aided Mol. Design, 7(1993) 83–102.
16. Jones, C., Willet, P. and Glen, R.C., *A genetic algorithm for flexible molecular overlay and pharmacophore elucidation*, J. Comput.-Aided Mol. Design, 9 (1995) 532–549.
17. Crippen, G.M. and Havel, T.F., *Stable calculation of coordinates from distance information*, Acta Crystallogr., Sect. A. 34 (1978) 282.
18. Sheridan, R.P., Nilakantan, R., Dixon, J.S. and Venkataraghavan, R. *The ensemble distance geometry: Application to the nicotinic pharmacophore*, J. Med. Chem., 29 (1986) 899–906.

19. Crippen, G.M., *Distance geometry approach to rationalizing binding data*, J. Med. Chem., 22 (1979) 988–997.
20. Chose, A.K. and Crippen, G.M., *Distance geometry approach to modeling receptor sites* In Hansch, C., Sammes, P.G., Taylor, J.B. and Ramsden, C.A. (Eds) Comprehensive medicinal chemistry, Vol. 4, Pergamon Press, 1990, pp. 715–734.
21. Ghose, A.K. and Crippen, G.M., *Geometrically feasible binding models of the flexible ligand molecule at the receptor site*, J. Comp. Chem., 6 (1985) 350–359.
22. Motoc, I., Dammkoehler, R.A. and Marshall, G.R. *Three-dimensional structure-activity relationships and biological receptor mapping*, In Trinajstić, N. (Ed.) Mathematical and computational concepts in chemistry, Ellis Horwood Ltd., Chichester, 1986, pp. 222–251.
23. Kuhl, F.S., Crippen, G.M. and Friesen, D.K., *A combinatorial algorithm calculating ligand binding* J. Comput. Chem., 5 (1984) 24–34.
24. Tripos Associates, Inc., 1699 S. Hanley Road, St. Louis, MO 63144, U.S.A.
25. Payne, A.W. and Glen, R.C., *Molecular recognition using a binary genetic search algorithm*, J.Mol. Graph., 11 (1993) 74–91.
26. Bohacek, R., Delombaert, S., McMartin, C., Priestle, J., and Grutter, M., *3-Dimensional models of ACE and NEP inhibitors and their use in the design of potent dual ACE/NEP inhibitors*, J. Am. Chem. Soc., 118 (1996) 8231–8249.
27. Dalpaiz, A., Bertolasi, V., Borea, P.A., Nacci, V., Fiorini, I., Campiani, G., Mennini, T., Manzoni, C., Novellino, E. and Greco, G., *A concerted study using binding measurements, X-ray structural data and molecular modeling on the stereochemical features responsible for the affinity of 6-arylpyrrolo [2,1-D] [1, 5] benzothiazepines toward mitochondrial benzodiazepine receptors*, J. Med. Chem., 38 (1995) 4730–4738.
28. Froimowitz, M., Patrick, K.S. and Cody, V., *Conformational analysis of methylphenidate and its structural relationship to other dopamine reuptake blockers such as CFT*, Pharmaceutic. res., 12 (1995) 1430–1434.
29. Bandoli, G., Dolmella, X., Gatto, S. and Nicolmi, M., *X-ray studies, empirical, semiempirical and statistical calculations on a series of thyrotropin-releasing-hormone derivatives*, J. Mol. Struct., 345 (1995) 213–225.
30. Morita, H., Yun, Y.S., Takeya, K., Itokawa, H. and Shiro, M., *Conformational analysis of a cyclic hexapeptide, segetalin-A from Vaccaria segetalis*, Tetrahedron, 51 (1995) 5987–6002.
31. Brandt, W., Drosihin, S., Haurand, M., Holzgrabe, U. and Nachtsheim, C., *Search for the pharmacophore in k-agonistic diazabicyclo[3.3.1]nonan-9-one-1,5-diesters and arylacetamides*, Arch. der Pharmazie, 329 (1996) 311–323.
32. Hennig, P., Raimbaud, E., Thurieau, C., Volland, J.P., Michel, A. and Fauchere, J.L., *Solution conformation by NMR and molecular modeling of 3 sulfide-free somatostatin octapeptide analogs compared to angiopeptin*, J. Comput. -Aided Mol. Design, 10 (1996) 83–86.
33. Boger, D.L. and Zhou, J.C., *N-Desmethyl derivatives of Deoxybouvardin and RA-VII: synthesis and evaluation*, J. Am. Chem. soc., 117 (1995) 7364–7378.
34. Morita, H., Yun, Y.s., Takeya, K., Itokawa, H. and Shiro, M., *Conformational analysis of a cyclic hexapeptide, segetalin-A from Vaccaria segetalis*, Tetrahedron, 51 (1995) 5987–6002.
35. Sefler, A.M., He, J.X., Sawyer, T.K., Holub, K.E., Omecinsky, D.O., Reily, M.D., Thanball, V., Akunne, H.C. and Cody, W.L., *Design and structure-activity relationships of C-terminal cyclic neurotensin fragment analogs*, J. Med. Chem., 38 (1995) 249–257.
36. Nicklaus, M.C., Wang, S., Driscoll, J.S. and Milne, G.W., *Conformational changes of small molecules binding to proteins*, Bioorg. Med. Chem., 3 (1995) 411–428.
37. MDL Information Systems Inc., 14600 Cataline Street, San Leandro, CA 94577, U.S.A.
38. AM Technologies Inc., 14785 Omicron Dr., Texas Research Park, San Antonio, TX 78245, U.S.A.
39. Ghose, A.K. and Crippen, G.M., *A general distance geometry three-dimensional receptor model for dihydrofolate reductase inhibitors*, J. Med. Chem., 27 (1984) 901–914.
40. Ghose, A.K. and Crippen, G.M., *Use of physicochemical parameters in distance geometry and related three-dimensional quantitative structure-activity relationships: A demonstration using Escheria coli dihydrofolate reductase inhibitors*, J. Med. Chem., 28 (1985) 333–346.

41. Cramer, R.D., Patterson, D.E. and Bunce, J.D., *Comparative molecular field analysis (CoMFA): I. Effect of shape on binding of steroids to carrier proteins*, J. Am. Chem. Soc., 110 (1988) 5959–5967.
42. Holzgrabe, U. and Hopfinger, A.J., *Conformational-analysis, molecular shape comparison, and pharmacophore identification of different allosteric modulators of muscarinic receptors*, J. Chem. Inf. Comp. Sci., 36 (1996) 1018–1024.
43. Hariprasad, V. and Kulkarni, V.M., *A proposed common spatial pharmacophore and the corresponding active conformations of some peptide leukotriene receptor antagonists*, J. Comput.-Aided Mol. Design, 10 (1996) 284–292.
44. Barlett, P.A., Shea, J.T. Telfer, S.J. and Waterman, S. *CAVEAT: S program to facilitate the structure-derived design of biologically active molecules* (Special publ. Molecular recognition in chemical and biological problems), 78 (1989) 182–196.
45. Tschinke, V. and Cohen, N.C., *The NEWLEAD program: A net method for the design of candidate structures from pharmacophore hypothesis*. J. Med. Chem., 36 (1993) 3863–3870.
46. Pickett, S.D., Mason, J.S. and Melay, I.M., *Diversity profiling and design using 3D pharmacophores — pharmacophore-derived queries (pdq)* J. Chem. Inf. Comput. Sci. 36 (1996) 1214–1223.
47. Mattos, C. and Ringe, D., *Multiple Binding Modes*, In H. Kubinyi (Ed.) 3D QSAR in drug design: Theory, methods and applications. ESCOM, Leiden, 1993, pp. 226–254.
48. Diana, G.D., Jaeger, E.P., Peterson, M.L. and Treasurywala, A.M., *The use of an algorithmic method of small molecule superpositions in the design of antiviral agents*, J. Comput.-Aided Mol. Design, 7 (1993), 325–335.
49. Martin, Y.C., *Pharmacophore mapping*, In Martin, Y.C. and Willet, P. (Eds) Design of bioactive molecules using 3D structure information, ACS Washington D.C., 1997.



**This Page Intentionally Left Blank**

# The Use of Self-organizing Neural Networks in Drug Design

Soheila Anzali<sup>a</sup>, Johann Gasteiger<sup>b</sup>, Ulrike Holzgrabe<sup>c</sup>, Jaroslaw Polanski<sup>d</sup>,  
Jens Sadowski<sup>e</sup>, Andreas Teckentrup<sup>b</sup> and Markus Wagener<sup>f</sup>

<sup>a</sup>E. Merck KGaA, Department of Medicinal Chemistry/Bio- and Chemoinformatics  
D-64293 Darmstadt, Germany

<sup>b</sup>Computer-Chemie-Centrum, Institut für Organische Chemie, Universität Erlangen-Nürnberg,  
D-91052 El-larigen, Germany

<sup>c</sup>Pharmazeutisches Institut der Universität Bonn, D-53 115 Bonn, Germany

<sup>d</sup>Institute of Chemistry, University of Silesia, PL-40006 Kutowice, Poland

<sup>e</sup>BASF AG, Drug Design, ZHV/W-A30, D-67056 Ludwigshafen, Germany

<sup>f</sup>Smith Kline Beecham, UW2940, King of Prussia, PA 19406-0939, U.S.A.

## 1. Introduction

The development of a new drug is an extremely laborious and time-consuming process. Thus, quite early on, computer methods have been used to further an understanding of the interactions of a drug with its receptor. Molecular modelling and rational drug design have become indispensable tools for the development of a new drug [1]. Recently, combinatorial chemistry and high-throughput screening have been introduced in order to speed up the drug development process. These methods produce massive amounts of data that have to be analyzed in an efficient manner in order to make best use of these novel methods. We will show here that self-organizing neural networks, such as the one introduced by Kohonen [2], can be used both in rational drug design and in combinatorial chemistry.

The application of neural networks in chemistry has increased dramatically in recent years [3–5]. In a Kohonen neural network (KNN), the artificial neurons self-organize in an unsupervised learning process and, thus, can be used to generate topological feature maps. It will be shown here that this potential can be utilized to analyze the shape and surface properties of those three-dimensional objects responsible for biological activity, molecules.

In these applications, there is a one-to-one mapping of a *single molecule* into a single Kohonen network. However, a Kohonen network can also be used for the analysis of *datasets of molecules*, where several molecules are simultaneously mapped into one Kohonen network. In order to make full use of the potential of self-organizing networks, novel representations of molecular structures have been developed. These methods can be put into a clear hierarchy, starting from molecular topology going all the way to molecular surfaces. They do not only encode structural information, but also information on the properties of atoms or of molecular surfaces.

## 2. Self-organizing Neural Networks

A Kohonen network can be used to study data of high-dimensional spaces by projection into a two-dimensional plane. The projection will be such that points that are adjacent in

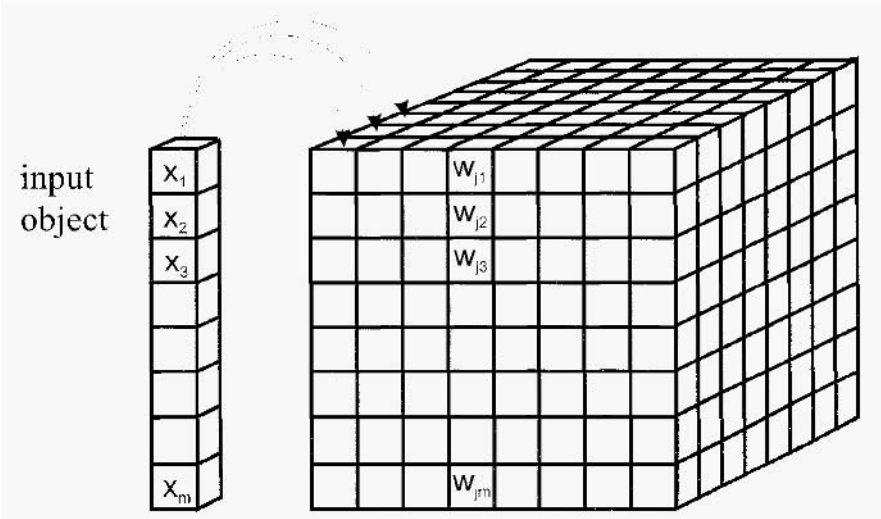


Fig. 1. Architecture of a Kohonen neural network The input object  $X = (x_1, x_2, \dots, x_m)$  is mapped into an  $n \times n$  arrangement of neurons,  $j$ , each having a weight vector  $W_j = (w_{j1}, w_{j2}, \dots, w_{jm})$ .

the high-dimensional space will also be adjacent in the Kohonen map: this explains the full name of the method. self-organizing topological feature maps.

Figure 1 shows the architecture of a Kohonen network: each column in this two-dimensional arrangement represents a neuron, each box in such a column represents a weight of a neuron [5]. Each neuron has as many ( $m$ ) weights,  $w_{ji}$ , as there are input data,  $x_i$ , for the object that is being mapped into the network.

An object, a sample,  $s$ , will be mapped into that neuron,  $sc$ , that has weights most similar to the input data (Eq. 1):

$$\text{out}_{sc} \leftarrow \min \left[ \sum_{i=1}^m (x_{si} - w_{ji})^2 \right] \quad (1)$$

The weights of this winning neuron,  $sc$ , will be adjusted such as to make them even more similar to the input data. In fact, the weights of each neuron will be adjusted but to a degree that decreases with increasing distance to the winning neuron. There are various ways for utilizing the two-dimensional maps obtained by a Kohonen network:

1. Representation: the two-dimensional map can be taken as a representation, an encoding, of the higher-dimensional information.
2. Similarity perception: objects that are mapped into the same or closely adjacent neurons can be considered as similar.
3. Cluster analysis: points that form a group in such a map, clearly distinguished from other points, can be taken as a class or category of objects having certain features in common.

### 3. Computational Details

The 3D structures of all molecules studied in the investigations reported here were generated by the 3D structure generator CORINA [6–9] and used without further optimization. Partial atomic charges are calculated by the PEOE (*Partial Equalization of Orbital Electronegativity*) method [10] and its extension to conjugated systems [11]. Residual electronegativity values were also obtained by the above two procedures [10,11] by considering the charge dependence of electronegativities [12]. The method for the calculation of atom polarizabilities has been published in reference [13]. These methods are collected in the program package PETRA (*Parameter Estimation for the Treatment of Reactivity Applications*) [14].

The molecular electrostatic potential was obtained in a classical manner by moving a point charge on the molecular surface and calculating the potential according to Coulomb's law from the partial atomic charges [15]. Any molecular surface can be taken into account; in most cases, we used the van der Waals surface.

The Kohonen neural networks were generated and analyzed by the Kohonen map simulator KMAP [16]. The study of the benzodiazepine and dopamine dataset was performed with an implementation of a Kohonen network on a massively parallel computer, MasPar [17,18]. In many studies reported here; the same dataset was used: 31 steroids binding to the corticosteroid binding globulin (CBG) receptor [19,20] and for which affinity data were available in the literature [21–23]. This dataset was chosen because it had been studied with other methods [20,24,25]. Quite intentionally the same dataset of CBG steroids is used again and again, in order to render the various methods comparable and show which features they emphasize.

### 4. One Molecule into One Network

#### 4.1. Maps of molecular surfaces

##### 4.1.1. Methods and results

A Kohonen network can be used to map a molecular surface into a two-dimensional plane. For this mapping of a molecular surface, points on this surface are chosen at random and their three Cartesian coordinates are taken as input into a KNN, with each neuron having three weights [26,27]. As the molecular surface is without beginning and without end, it was decided also to choose for projection a two-dimensional plane without beginning and without end, the surface of a torus. For visualization, the torus is cut along two perpendicular lines and the surface spread into a plane.

With a toroidal network, the maps can be shifted, mirrored and rotated against each other to achieve a similar position of their patterns. The surface of a molecule and the surface of a torus have a different topology and, therefore, this mapping process must lead to topological distortions that result in empty neurons. This feature of the mapping process in a Kohonen network has been analyzed and explained in detail [28].

Once the network has been trained, the entire dataset is sent again through the network and each neuron is colored with a property on the molecular surface that exists

at that point(s) that is (are) mapped into the neuron considered [27]. In this coloring process, any molecular surface property can be chosen such as molecular electrostatic potential, hydrogen-bonding potential, or just a color identifying the surfaces of different types of atoms.

In order to give an idea of the correspondence between a 3D space and its 2D map, we show in Fig. 2 (see p. 279) as an example the projected Kohonen maps of the van der Waals surface of corticosterone. The values of the electrostatic potential on the molecular surface (MEP) determine the colors of the map.

Corticosterone has two sites with a large negative value of the MEP, the carbonyl group at positions 3 (4) and the side chain COCH<sub>2</sub>OH at position 17 (1) (Fig. 2). Consistent with this, the Kohonen map (Fig. 2) shows two spaces with a red–yellow color of these sites. The spatial distance of these groups is reflected by two different shapes of the projection of the MEP into the Kohonen network. The third site with a negative value of the MEP stems from the hydroxyl group at position 11 (3). A space with a yellow color is reserved in this map for this group. Furthermore, the large positive MEP area of corticosterone is below the D-ring and the side chain at position 17 (2). The projection of the MEP into the Kohonen map indicates the location of this space (violet color) close to the space of the negative MEP area of COCH<sub>2</sub>OH at position 17.

#### 4.1.2. Visual comparison of Kohonen maps

The comparison of Kohonen maps of molecular surface properties offers a technique for the perception of similarities in ligands binding to the same receptor. Kohonen maps of the molecular electrostatic potential have been generated for four ligands that bind to the muscarinic receptors and four ligands that bind to the nicotinic receptors and are shown in Fig. 3.

Visual inspection of these maps clearly shows characteristics that are common to the four molecules binding to the muscarinic receptors and are not contained in the maps of the four ligands binding to the nicotinic receptors [20]. The nicotinic compounds, for their side, show common features different from those of the muscarinic ligands. Thus, inspection of these eight maps allowed a clear separation of molecules that either bind to the muscarinic or the nicotinic receptors.

#### 4.1.3. Averaged maps

In the previous example, a visual comparison of the Kohonen maps of the molecular electrostatic potential was made, thus allowing one to differentiate ligands that bind to two different types of receptors. The question is now whether such a comparison can be put onto a more objective basis. This will be explored with 31 steroids for which the CBG affinities are known [21–23]. The distribution of the compounds into high-, intermediate- and low-affinity classes are defined in reference [24].

For each of the 31 steroids, a Kohonen network was trained, using the three Cartesian coordinates of points on the molecular surface as input into the network. The values of the MEP determine the colors of the map. For a more objective analysis, the averaged maps for the sets of high-, medium- and low-active compounds were generated (Fig. 4). For this purpose, each neuron *n* of the Kohonen maps of the single compounds was assigned a

color index in a range of ten values representing the MEP of the potential the neuron obtained during the training process. Then the colors of the neurons in the averaged maps were obtained by averaging the color indices of the neurons in the single maps.

The MEP pattern of the most polar area in the averaged map of the highly active compounds is the most pronounced one. In the three averaged maps, the pronunciation of the polar spaces decreases according to decreasing activity of the compounds. Therefore, a comparison of the maps of steroids with the averaged map allows one to establish whether a molecule belongs to the active or inactive CBG compounds. The averaged map of the highly active compounds can be used to build a pharmacophore model.

#### *4.1.4. Maps as a two-dimensional representation of molecules*

The investigation of the previous section can be taken one step further. If, indeed, the maps of the molecular electrostatic potential allow one to distinguish high-active compounds from low-active compounds, why not take those maps as representations of molecules? In other words, we first train a, say,  $20 \times 20$ , Kohonen network with the three Cartesian coordinates of points of a molecular surface. Then, the entire dataset is sent again through the network and an extra layer of the network used for labeling the network is colored with the electrostatic potential of the points that are mapped into each neuron. (This is the procedure as outlined in section 4.1.1.) The values of the MEP (or the color) of the 20 rows of this label layer, each consisting of 20 values, are then concatenated to give a 400-dimensional vector. This vector is a two-dimensional representation of a molecule as it has been obtained by projecting a molecular surface into two dimensions and is used to train a second Kohonen network of size  $5 \times 5$ . The result is shown in Fig. 5.

It can be seen that the steroids quite nicely separate into groups of compounds of high, medium and low activity. Only one compound of medium activity shows a collision with highly active compounds by being mapped into the same neuron.

#### *4.1.5. Bioisosteric design*

The bioisostere database by Istvan Ujvary (BIOISOSTER version 1.3), a database of analog design, including 1515 bioisosteric groups was analyzed. The question was if some coherency between the physico-chemical properties and the bioisosteric effect can be deduced by looking at the calculated Kohonen maps of these groups. Figure 6 shows an example of such a structure-pair in this database. The squares show those parts of the structures which are defined as bioisosteric groups.

Several hundred pairs from this database were selected. The selection of these compounds was based on diverse structural fragment pairs as far as possible. From the selected database the bioisosteric groups are then cut out. The 3D structures of the selected fragments are calculated using the program CORINA [6-9]. The MEP were calculated on the van der Waals surface. Then, the Kohonen maps were calculated for each bioisosteric groups with a unique color plate. Figure 7 shows some examples of the calculated fragments. As shown here, the fragment pairs are structurally quite different, but their maps show a high similarity in the electrostatic potential patterns.

This is an interesting result because it offers the possibility for selecting fragment pairs of this database, which can have a general validity, *true bioisosteric groups*. Therefore, we constructed a 3D database of several hundred fragments and functional groups including their corresponding Kohonen maps. The comparison of the electrostatic potential patterns of the maps of such a library can be used to cluster bioisosteric groups and, consequently, improve the efficiency of the 3D design of bioactive molecules.

## 4.2. Comparative maps

### 4.2.1. The template approach

A Kohonen network stores the information on an object that is used for training. This fact inspired the use of a Kohonen network trained with the molecular data of a given molecule as a reference molecule, a template, for the comparison with other molecules [30,31]. The general idea of such a comparative mapping is shown in Fig. 8. The Cartesian coordinates of the points taken from the molecular surface of a butane molecule (a) are used to train a Kohonen network (c). In the example shown in Fig. 8, neurons of the map (d) are colored by giving the surface belonging to carbon atoms 1 to 4 of butane, and the hydrogen atoms bonded to these carbon atoms different shades of gray. The network (c) can now be used for the comparison of the surface of molecules other than butane. In our example, it was used for the simulation of a map of the propane molecule (b). Such a map (e) can be seen as a superposition of the compared molecule onto the template molecule, propane and butane in our case. A point from the compared molecule will find a neuron in the template network having weights quite similar to coordinates of the point from the surface of the compared molecule (cf. Eq. 1). However, neurons corresponding to those parts of the surface of the template molecule that have no counterpart on the surface of the compared molecule will not become excited and, thus, stay empty. In the comparative map that is obtained by filtering the compared molecule through the reference network of the template molecule, the empty neurons show up as white areas indicating where the surface of the reference molecule differs from the surface of the compared molecule.

Different settings of the parameters for training and testing of the Kohonen network program allow one to emphasize certain aspects of the molecular surface to a different extent. In particular, the value for the threshold that determines whether the input data of an object match with the weights of a neuron (cf. Eq. 1) can render the number of non-matching (empty) neurons and thus the white area in the comparative map larger or smaller. This is indicated in Fig. 8 where one setting (top map, Fig. 8e) somehow indicates the entire methyl group of butane to be lost in propane, whereas the second setting (bottom map, Fig. 8c) shows the major difference to reside in the three hydrogen atoms of the methyl group of butane being lost in propane.

The basis of the template approach is an analysis of the shapes of molecules and the quantification of a shape similarity or dissimilarity within a series of compounds using a reference molecule. This is of particular merit for the comparison of a series of biologically active compounds. The larger the difference in shape between the reference

L		H	M	
L, L			M	M
L, L, L, L, L		H, H, H	H	
		H	M	
L, L, L		H, H, H, M, H, H	M, M	M, M

Fig. 5. Mapping of a dataset of 31 steroids binding to the corticosteroid binding globulin (CBG) receptor into a toroidal 5 x 5 Kohonen network. Each steroid is labeled by its activity range: H = high, M = medium and L = low activity.

Table 1 Number of empty neurons for the maps of CBG compounds (total number of neurons = 2500)

Name of molecule	No. of empty neurons	Name of molecule	No. of empty neurons
Corticosterone	0	16 $\alpha$ , 17-dihydroxyprogesterone	69
Cortisol	15	19-nortestosterone	357
11-deoxycortisol	52	Dihydrotestosterone	324
17 $\alpha$ -hydroxyprogesterone	61	2 $\alpha$ -methyl-9 $\alpha$ -fluorocortisol	28
2 $\alpha$ -methylcortisol	6	4-androstenedione	350
11-deoxycorticosterone	50	Androsterone	417
Cortisolacetat	13	Eticholanolone	710
Prednisolone	140	Pregnenolone	108
Progesterone	58	17 $\alpha$ -hydroxypregnenolone	130
Epicorticosterone	46	Estriol	636
17 $\alpha$ -methylprogesterone	113	Estrone	700
Cortisone	75	Estradiol	644
19-norprogesterone	152	Dehydroepiandrosterone	378
4-pregnene-3,11,20-trione	79	Androstenediol	341
Testosterone	296	5-androstenediol	332
Aldosterone	225		



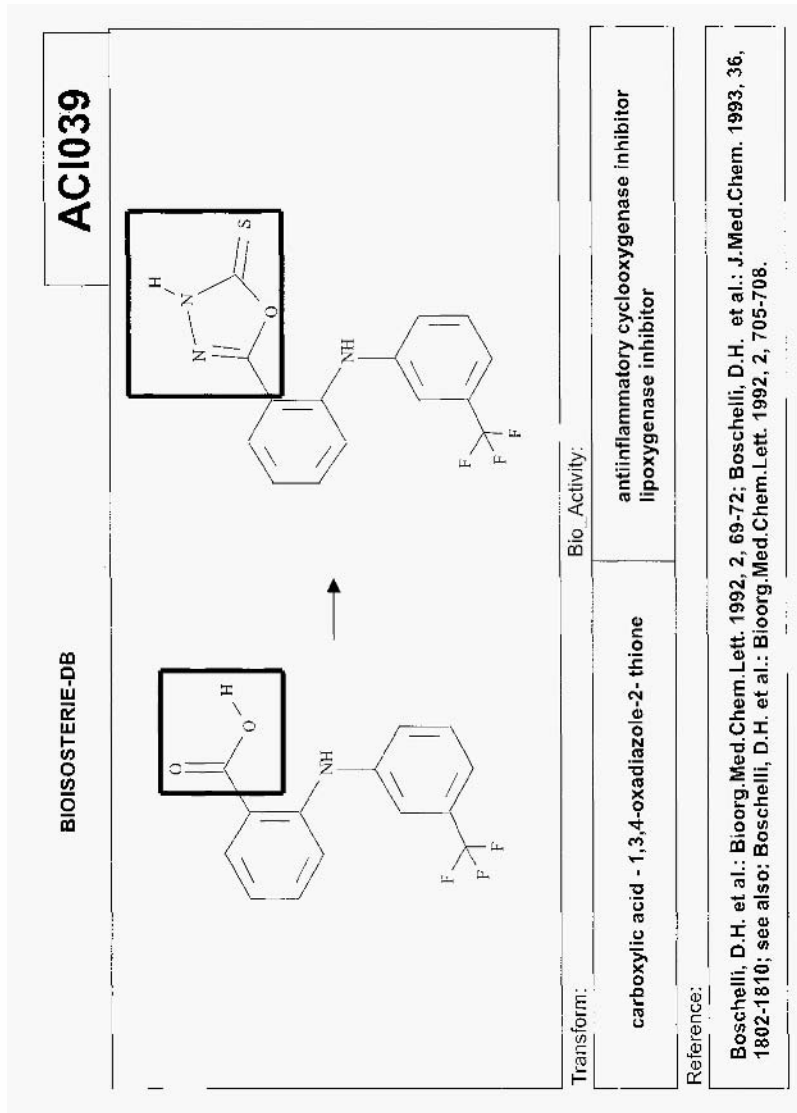


Fig. 6. A structure-pair of the Bioisostere Database. The squares show the bioisosteric groups.

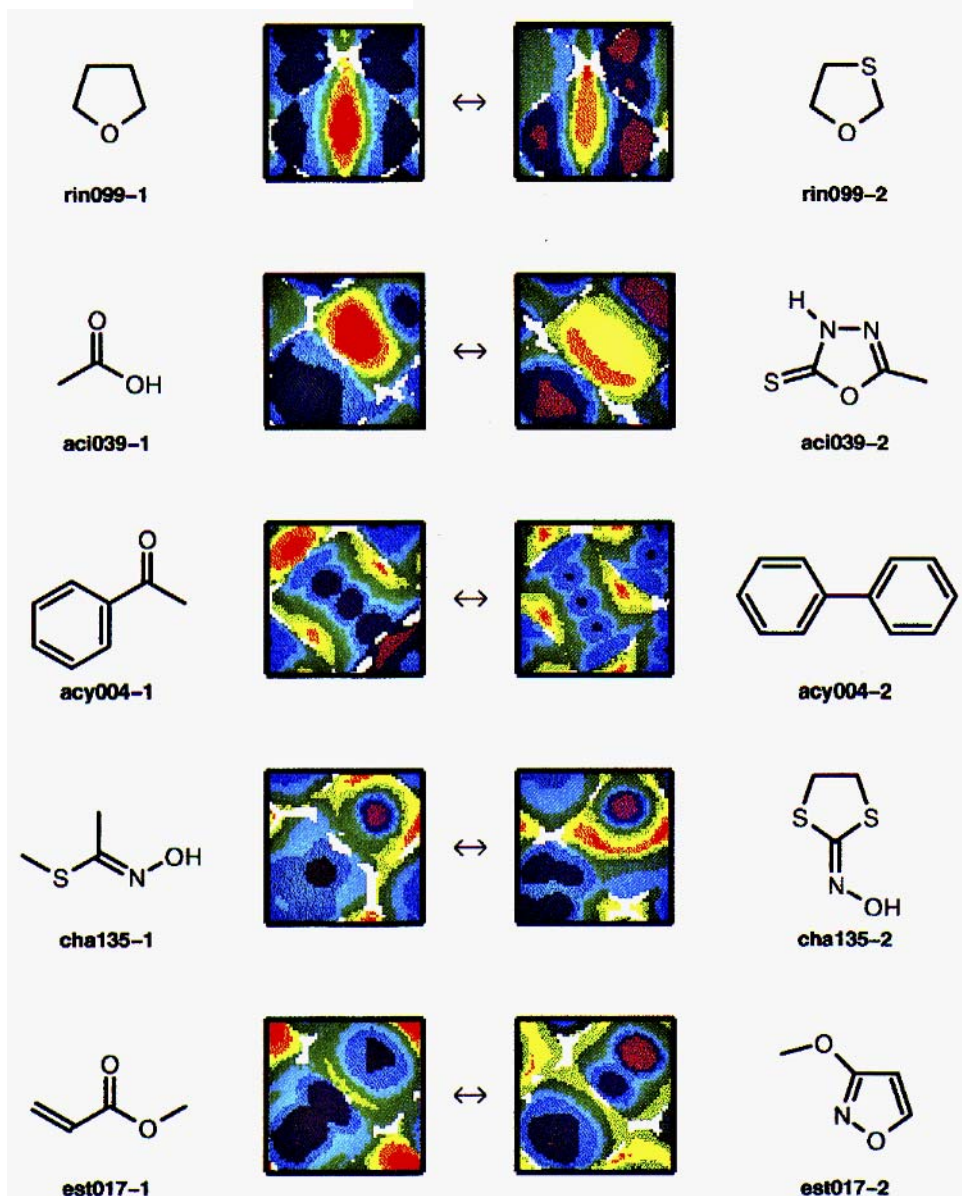


Fig. 7. Some examples of the bioisosteric groups and their calculated Kohonen maps.

#### 4.2.3. Backprojection of maps onto molecular shapes

The identification of the pharmacophore in an assembly of structurally diverse ligands is quite a difficult task, especially when the structure of the target molecule, the receptor, is not known. It will be demonstrated how Kohonen networks can be utilized for this

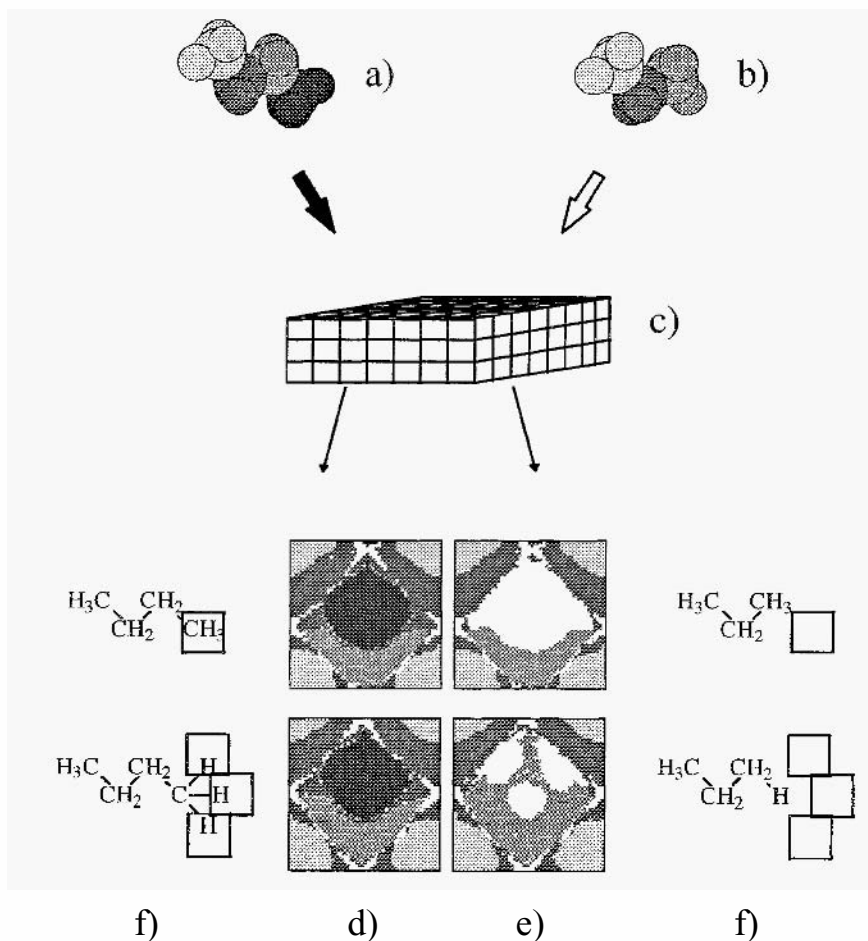


Fig. 8. The idea of comparative Kohonen mapping of two molecules. A template butane molecule (a) trains (black arrow) a Kohonen network (c) allowing for 2D visualization of the surface of two methyl ( $\text{CH}_3$ ) and two methylene ( $\text{CH}_2$ ) groups (d). Different settings of the parameters (two examples) during training result in only slightly different template patterns (d). The same network (c), if used for processing the molecular data coming from propane (b), gives a comparative pattern of the propane molecule (e). Different settings of the training parameters allow for showing such aspects of similarity that comply with the simple analysis coming from a chemist (f) — i.e. propane lacks the entire terminal methyl group of butane that is taken in the square within upper formulas; if, however, the carbon atom of this methyl group is superposed on one of hydrogen atoms of the propane molecule the difference will consist in three hydrogen atoms as indicated with squares in the bottom formulas.

problem by elucidating the pharmacophore of allosteric modulators of muscarinic receptors, a group of drugs under development. which can enhance the activity of an antagonist in a very specific manner [32]. In a previous publication [33], it was shown how the pharmacophore can be mapped out with a small number of allosteric modulators.

Here, we limit the discussion even further by only exploring alcuronium as the most potent compound, characterized by an almost rigid structure, and the flexible W84 as a representative of a wider range of hexamethonium and bispyridinium compounds [32–35].

Two different conformations of W84 were explored, the extended form **4e**, obtained from CORINA, and a slightly distorted sandwich form, **4d**, found by molecular dynamic calculations and subsequent optimization by the semiempirical AM1 method. For alignment, the rigid alcuronium was chosen as a template. Both conformations of W84 were superimposed onto alcuronium first using the positively charged nitrogens, because they were assumed to be the most important feature for the first step of ligand receptor recognition. In addition, both aromatic rings were matched onto each other. The superposition of the extended, linear conformation of W84 (**4e**) and alcuronium (**1**) with the fixpoints of two positively charged nitrogens resulted in protruding phthalimido rings at both ends of alcuronium (Fig. 9a). The alignment of alcuronium, **1**, and the distorted sandwich conformation of W84 (**4d**), revealed a much better fit (Fig. 9b).

In order to find out the similarities in the molecular surfaces and, thus, properties of these molecules, the surfaces were sent into Kohonen neural networks and colored by labeling the neurons according to the kind of atom the corresponding points belonged to (atomic surface assignment, ASA). The ASA maps showed that, firstly, the maps of alcuronium and that of the distorted sandwich conformation of W84 are similar, whereas the one with the extended form of W84 is quite different.

For a more quantitative comparison of the 3D shape of the molecules, a template approach was made sending both conformations of W84 through the Kohonen network of alcuronium as reference compound. The maps obtained for the surface of the molecules show a rather large number of empty neurons reflecting that the shape of the molecules is fairly different from the shape of the large reference molecule alcuronium.

There is an even more illustrative method for showing the correspondence of two molecular surfaces. The map of the second molecule obtained by sending it through the Kohonen network of the first, the reference molecule, can be projected back onto the three-dimensional surface of the reference structure alcuronium. Fig. 10 shows the 3D models of the surface of alcuronium **1** with a backprojection of the Kohonen map of the extended conformation and that of the distorted sandwich-like conformation of W84, **4e** and **4d**, respectively. Those places that have empty neurons in the template maps are indicated by a black open mesh on the surface of the alcuronium.

This way of representation impressively exhibits the similarities between alcuronium and the distorted sandwich conformation of W84 (**4d**). The extended conformation of W84 tills only the center of the alcuronium surface. In contrast, the distorted sandwich-like conformation of W84 covers much more area of the surface of alcuronium. Moreover, the essential features, both aromatic skeletons and both positive charges, color the surface exactly at the same places for this sandwich conformation and alcuronium.

Taken together, the following conclusions can be drawn [30]: firstly, the pharmacophore consists of two positively charged nitrogens in a distinct distance from each other and two heterocyclic, aromatic rings, both closely located in the hydrophobic central chain; secondly, the distorted sandwich geometry appears to be the conformation W84 takes up upon binding to the allosteric binding site; and thirdly, electrostatic inter-

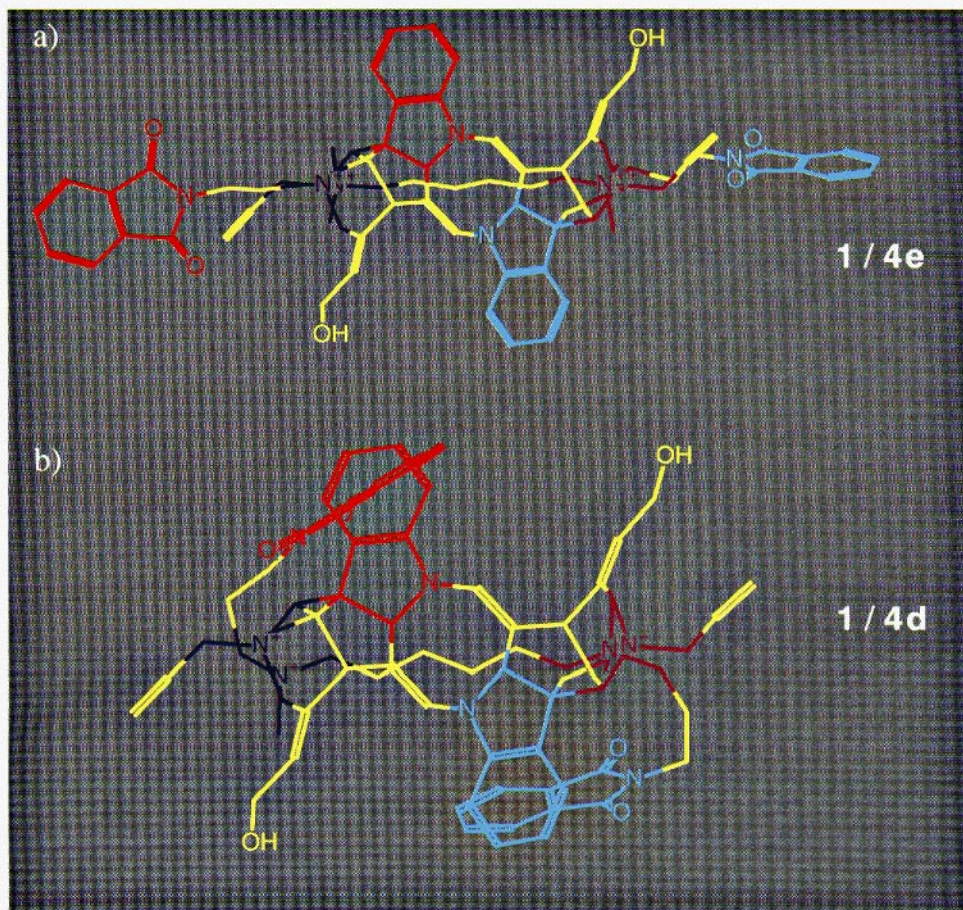


Fig. 9. Superposition of the skeleton of the extended form of W84 (4e) and the distorted sandwich form of W84 (4d) onto alcuronium (1).

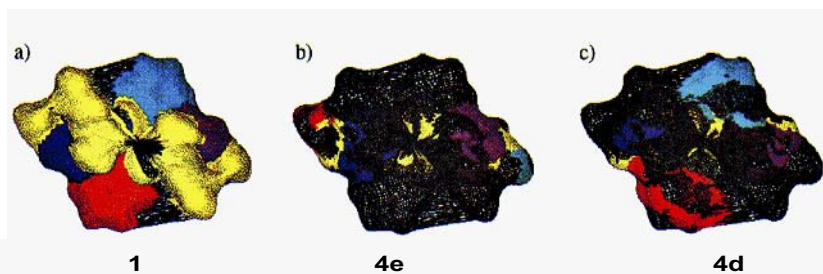


Fig. 10. Backprojection of the Kohonen maps shown in Fig. 28a–c onto the molecular surface of alcuronium 1.

actions have been found to be primarily responsible for the molecular recognition between receptor and ligand.

#### 4.2.4. Descriptors from comparative maps

The more the surface of a template (reference) and a compared molecule differ from each other, the higher is the number of empty neurons. Therefore, this value can be taken as a measure describing the difference in the *geometry* of the molecules input. This approach can be taken one step further also to describe differences in molecular surface *properties*. In principle, any molecular surface property can be taken but we limit the discussion to the molecular electrostatic potential (MEP). The entire spectrum of the electrostatic potential is divided into ranges (e.g. 10 ranges) each indicated by a specific color. Therefore, the map is coded by a matrix with elements that take discrete values from 1 to 10, while 0 codes empty neurons. Clearly, also the real values of the electrostatic potential could be used, but we wish to keep the discussion here simple.

Figure 11b shows an example of the comparative MEP maps of two similar molecules shown in Fig. 11a. Figure 11c compares the histogram of the occurrence of the neurons colored with the respective colours 0-10. We can define descriptors for the differences in color profiles of the maps. By comparing the occurrence of neurons having the same color (range of MEP):

$$K_i = \sum_j^n \sum_k^n (k_i)_{jk} \quad (2)$$

with  $k_i = 1$  for neurons colored with the respective color coded by  $i$  and  $k_i = 0$  for all other colors, while the matrix coding a feature map is of size  $n \times n$ .

The difference between the map of the template and the compared molecule is given by:

$$S_i = (K_i)_T - (K_i)_M \quad (3)$$

with the index,  $T$ , denoting the template, and  $M$  the molecule being compared. We can also include the range of the electrostatic potential coded by a certain color and obtain a modified Eq. 4:

$$EP_i = c_i S_i = \left( \sum_j^n \sum_k^n (c_i)_{jk} \right)_T - \left( \sum_j^n \sum_k^n (c_i)_{jk} \right)_M \quad (4)$$

where  $c_i$  is a value (1–10) defining the range of the electrostatic potential coded by the respective color  $i$  and  $(c_i)_{jk}$  denotes the components of the matrix, of size  $n \times n$ , coding this respective color within the feature map.

These descriptors can be calculated for a single color or for a group of different ranges of colors — e.g.  $EP_{1-3}$  will give a sum of the  $EP_1 + EP_2 + EP_3$ .

A related global  $EP$  descriptor was used by Barlow [36]. In contrast, the  $EP$  parameter calculated for a narrow range of  $EP$  will bear only the information on the polar character neglecting the shape of the molecules.

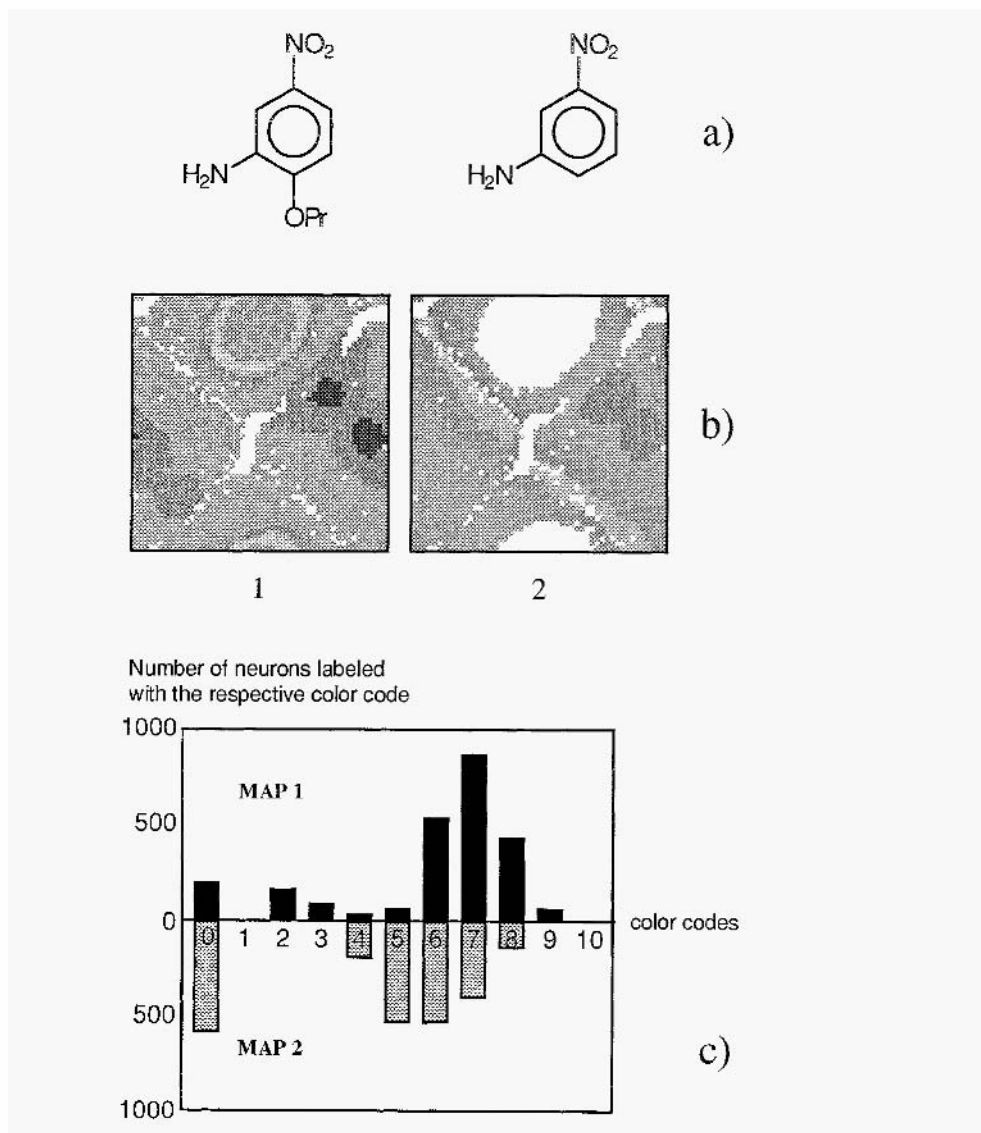


Fig. 11. comparative patterns (b) of two molecules (a) and a histogram comparing the frequency of the occurrence of different colors within the maps (c).

#### 4.3. Quantitative structure–activity studies

One of the basic approaches toward modelling SAR and QSAR relationships involves the comparison of a series of bioactive molecules. Table 2 gives an overview of a series of analogs analyzed in previous publications by quantitative structure–activity studies using descriptors developed above (section 4.2.4).

Table 2 The SAR and QSAR models obtained using the template Kohonen maps of the molecular electrostatic potential

Study	Compounds/activity	Model/statistical characterization	Descriptor	Reference
1.	Steroids/CBG	Qualitative	NE	[30]
2.	Histamine analogs/H2 activities	Qualitative	EP <sup>a</sup> <sub>all</sub>	[36]
3.	Ryanodine derivatives/binding to membrane proteins	Qualitative	NE <sup>b</sup>	[37]
4.	Nitroanilines/sweetness activity	$r = 0.99, n = 9$	EP <sub>7-10</sub>	[31]
5.	Nitro- and cyanoanilines/sweetness activity	$r = 0.96, n = 18$	EP <sub>1-10</sub>	[31]
6.	Ethylcarboxylates/Taft's E <sub>s</sub> constant	$r = 0.94, n = 22$	NE	[31]
7.	Steroids/CBG	$r = 0.89, n = 31$	- <sup>c</sup>	[38]
8.	Steroids/TBG	$r = 0.92, n = 31$	- <sup>c</sup>	[39]
9.	Arylsulfonylalkanoic acids/sweetness activity	$r = 0.94, n = 9$ <sup>e</sup>	- <sup>c</sup>	[40]

<sup>a</sup> The comparison is performed by Subtraction of the MEP matrices, a number of ranges of EP is not given.

<sup>b</sup> NE is the number of empty neurons.

<sup>c</sup> The comparison is performed by a single instar neuron.

<sup>d</sup> The induced fit between the TBG receptor and molecules stimulating are simulated using the hypermolecule technique.

<sup>e</sup> A method allows for the determination of active pharmacophoric conformation, the statistical characterization considers this conformation,



For more details, the reader is referred to the original publications. The major conclusion to be drawn is that the template approach provides a basis for the quantification of the shape and electrostatic effects of molecular surfaces, allowing the calculation of descriptors that are useful for the development of quantitative structure–activity relationships.

## 5. Several Molecules into One Network

Previously, each Kohonen network was trained with information from only *one molecule*, in most cases, one molecular surface. The maps thus generated were replications of one molecular surface, the objects mapped into the individual neuron were points from the molecular surface. Now, work will be reported where entire *datasets of molecules* are sent into a single Kohonen network. Each object mapped into a neuron will consist of an entire molecule. Various representations of molecules have been employed, as detailed in the following sections.

### 5.1. Representation of molecules

The analysis of a dataset of objects by learning methods, be it by statistical or pattern recognition methods or neural networks, asks for the objects to be represented by the same number of variables. If the objects are molecules, one has to come up with the same number of descriptors irrespective of the size of the molecule and the number of atoms in the molecule. In the following, the transformation of a molecular structure by autocorrelation is used to obtain a representation by a fixed number of variables. It will be shown that autocorrelation allows molecules to be considered with different degrees of sophistication, starting with the constitution (the topology) of a molecule, through the 3D structure all the way to representations of molecular surfaces. In addition, a variety of physico-chemical properties of the atoms or of the molecular surfaces can be considered. Such a hierarchy of representations is mainly dictated by the size of the datasets to be studied: large datasets of molecules ask for rapid encoding schemes, while smaller ones allow a more detailed consideration of molecules.

### 5.2. Topological autocorrelation and the location of biologically active compounds

The idea of using autocorrelation for the transformation of the constitution of a molecule into a fixed length representation was introduced by Moreau and Broto [41]. A certain property,  $p_k$ , of an atom  $i$  is correlated with the same property of atom  $j$  and these products are summed over all atom pairs having a certain topological distance  $d$ . This gives one element of a topological autocorrelation function  $A(p_k, d)$  of this property  $p_k$ :

$$A(p_k, d) = \sum_{j=i+d}^N \sum_{i=1}^{N-1} \delta_{ji} p_k(i) p_k(j) \quad (5)$$

with  $\delta_{ji} = 1$  if  $d_{ji} = d$ , otherwise  $\delta_{ji} = 0$ .

The following properties were calculated by previously published empirical methods for all atoms of a molecule: sigma charge,  $q_{\sigma}$  [10], total charge,  $q_{tot}$ , sigma electronegativity,  $\chi_{\sigma}$  [12], pi-electronegativity,  $\chi_{\pi}$  [11], lone-pair electronegativity,  $\chi_{LP}$ , and atom polarizability,  $\alpha$ , [13]. In addition to those six electronic variables, the identity function — i.e. each atom being represented by the number 1 — was used in Eq. 5 to just account for the connectivity of the atoms in the molecule.

The autocorrelation of these seven variables was calculated for seven topological distances (number of intervening bonds) from two to eight. The basic assumption, thus, was that interactions of atoms beyond eight bonds can be neglected. With seven variables and seven distances, an autocorrelation vector of dimension 49 was obtained for each molecule, irrespective of its size or number of atoms. The hydrogen atoms were not considered in the calculation of the autocorrelation vector.

In order to investigate the potential of topological autocorrelation functions for the distinction of biological activity, a dataset of 112 dopamine agonists (DPA) and 60 benzodiazepine agonists (BDA) was studied [18]. A Kohonen network of size 10 X 7 was used to project these 172 compounds from the 49 dimensional space spanned by these autocorrelation vectors into two dimensions.

The two types of compounds, DPA, and BDA, were nearly completely separated in the Kohonen map, underscoring the potential of this molecular representation to model biological activity. To put this capability to a more severe test, this dataset of 112 DPA and 60 BDA compounds was mixed with the entire catalog of a chemical supplier (Janssen Chimica catalog, version 1989) consisting of 8323 commercially available compounds comprising a wide range of structures from alkanes to triphenylmethane dyestuffs.

The map of Fig. 12 shows that both DPA and BDA occupy only Limited areas in the overall map. Furthermore, the areas of DPA and BDA are quite well separated from each other, only one neuron with BDA intrudes into the domain of DPA and only two neurons with conflicts, obtaining both DPA and BDA, occur. With the results obtained here, the search for new active compounds or new lead structures can be restricted to a smaller area of the entire chemical space. This opens the way for searching for compounds with a desired biological activity and for discovering new lead structures in large databases of compounds. Closer analysis of the mapping shows interesting insights that are further discussed in the original publication [18].

### 5.3. Autocorrelation of molecular surfaces properties

Ligands and proteins interact through molecular surfaces and, therefore, clearly, representations of molecular surfaces have to be sought in the endeavor to understand biological activity. Again, we are under the restriction of having to represent molecular surfaces of different size; and again, nutocomelation was employed to achieve this goal [20]. Firstly, a set of randomly distributed points on the molecular surface has to be generated. Then, all distances between the surface points are calculated and sorted into preset intervals

$$A(d) = \frac{1}{L} \sum_{ij} p(i)p(j) \quad d_1 \leq d_{ij} < d_u \quad (6)$$

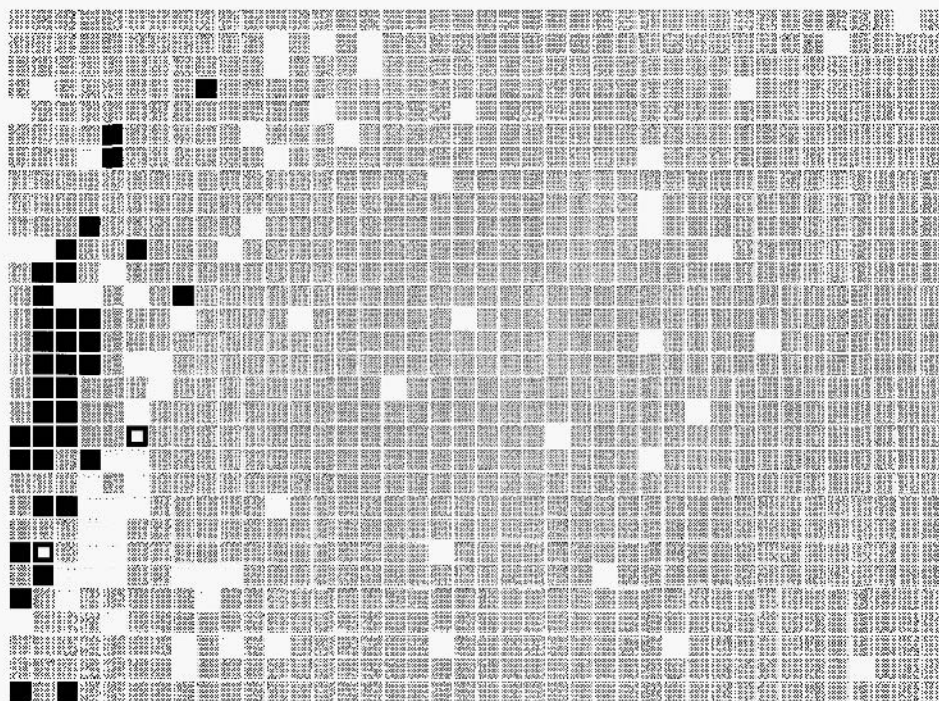


Fig. 12. Kohonen map of 40 X 30 neurons obtained by training with 112 dopamine (DPA), 60 benzodiazepine agonists (BDA) and 8323 commercially available compounds. Only the types of compounds mapped into the individual neurons are indicated. Black identifies DPA, light gray BDA and dark gray the compounds of unknown activity. Empty neurons are stored in white; the two neurons marked by a black frame indicate conflicts where both DPA and BDA are mapped into the same neuron.

where  $p(i)$  and  $p(j)$  are property values at points  $i$  and  $j$ , respectively;  $d_{ij}$  is the distance between the points  $i$  and  $j$ ; and  $L$  is the total number of distances in the interval  $[d_1, d_u]$  represented by  $d$ . For a series of distance intervals with different upper and lower bounds,  $d_1$  and  $d_u$ , a vector of autocorrelation coefficients is obtained. It is a condensed representation of the distribution of property  $p$  on the molecular surface.

#### 5.4. Modelling CBG affinity by a combination of two different neural networks

The affinity of 31 steroids binding to the corticosteroid binding globulin (CBG) receptor was modelled based on spatial autocorrelation coefficients of the molecular electrostatic potential as descriptor [20]. A vector of twelve autocorrelation coefficients corresponding to twelve distance intervals of 1 Å width between 1 and 13 Å was determined for each steroid using Eq. 5. Then, this set of descriptors was investigated using two different methods: firstly, the 12-dimensional descriptor space was projected into a plane using a Kohonen neural network in order to visualize the high-dimensional descriptor space. Then, these descriptors were used to quantitatively model CBG activity.

### 5.5. Modelling of chemical libraries

The methods introduced in previous sections have the advantage that they allow for a rapid visualization of high-dimensional descriptor spaces. The importance of this feature has increased with the advent of the large compound collections that can be generated by combinatorial chemistry and related techniques: small datasets comprising tens or hundreds of compounds can be analyzed using almost any method without reaching the limits of currently available computer hardware, whereas special techniques are needed for the handling of datasets of hundreds of thousands of compounds. To demonstrate the merits of Kohonen networks and spatial autocorrelation descriptors in handling large datasets, we analyzed three combinatorial libraries that together comprise more than 87 000 compounds [42].

Rebek et al. published the synthesis of two combinatorial libraries of semi-rigid compounds that were prepared by condensing a rigid central molecule functionalized by four acid chloride groups with a set of 19 different L-amino acids [43]. This process is summarized in Fig. 14. In addition to the two published libraries we included a third, hypothetical library with adamantane as central molecule into our study.

#### 5.5.1. Comparison of the xanthene, the cubane and the adamantane libraries

A Kohonen network with  $50 \times 50$  neurons was trained with the combined descriptors of the xanthene and the cubane libraries, each molecule represented by 12 autocorrelation values calculated from the electrostatic potential on the molecular surface. The resulting map is shown in Fig. 15a. The neurons are colored according to the most frequent central molecule that is mapped into them. All 2500 neurons of the map are occupied.

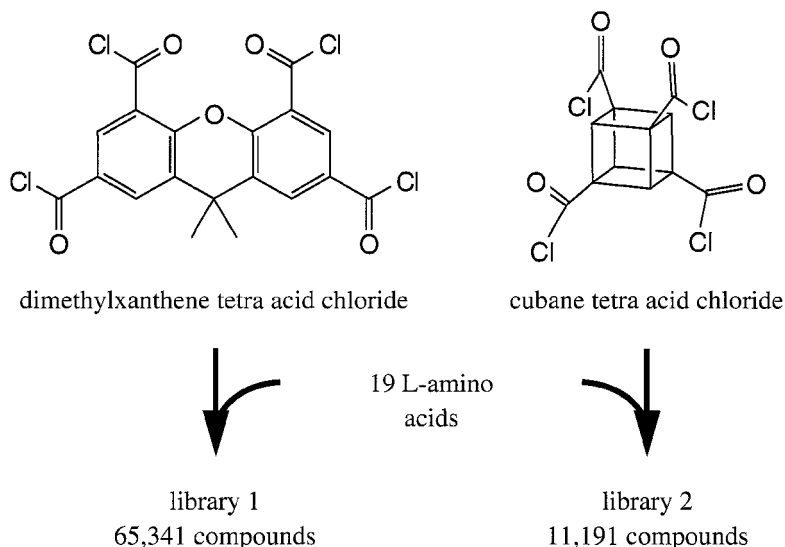


Fig. 14. Preparation of the xanthene and the cubane libraries.

The compounds of the cubane library form a cluster in the center of the map that is separated from the compounds of the xanthene library. The neural network can clearly separate the two libraries quite well — they both cover different parts of Kohonen maps and, thus, it can be concluded that they are from different parts of the chemical space. Consequently, they are remarkably different and, thus, both worthwhile to be considered in a screening program.

In a second experiment, we trained the same network with the combined data set of all three libraries. This resulted in the Kohonen map shown in Fig. 15b. Again, a distinct cluster that is clearly separated from the xanthene derivatives can be seen in the center of the map. The cubane and adamantane derivatives, on the other hand, cannot be distinguished by the neural network. They are tightly mixed in the central cluster, even more than can be concluded from Fig. 15b, as 92% of the cubane and adamantane compounds are mapped into common neurons. The cubane and adamantane libraries, thus, cover the same part of the chemical space — they are so similar to each other that considering both of them in a screening program is both a waste of resources and time. The xanthene library is evidently different from the other two libraries. Therefore, the xanthene and either one of the cubane or adamantane libraries should be used for screening.

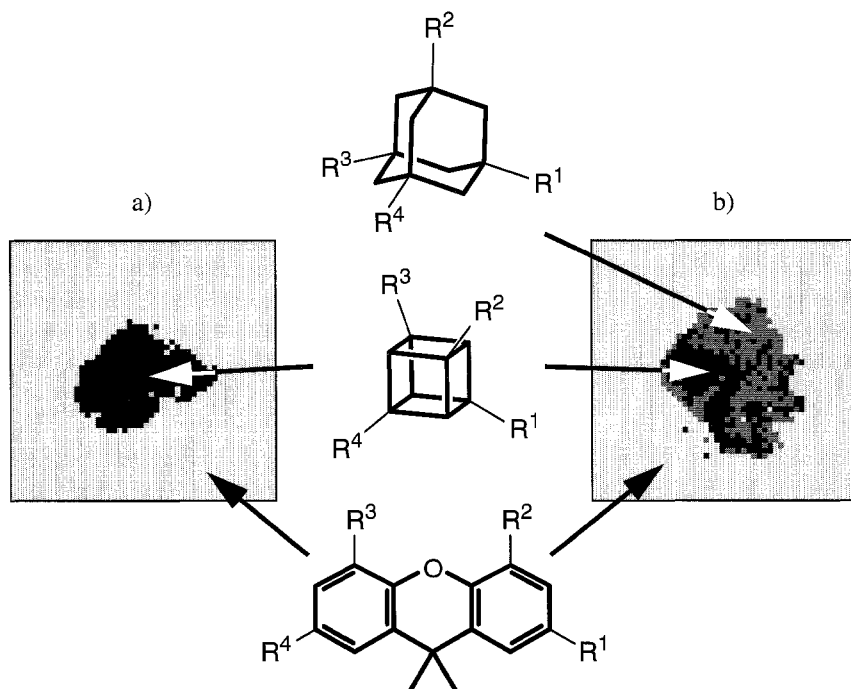


Fig. 15. Kohonen map of (a) the combined xanthene and cubane libraries and (b) the combined xanthene, cubane and adamantane libraries.

### 5.5.2. Deconvolution of xanthene sublibraries

Rebek et al. used their libraries to screen for novel trypsin inhibitors. Only the xanthene library showed significant trypsin inhibition, so that they concentrated further efforts on this library. In the next round of screening, they divided the xanthene library into six sublibraries by using subsets of only 15 amino acids for the generation of the libraries. These subsets were generated by omitting three amino acids in turn from a set of 18 amino acids. This process resulted in sublibraries of 25 425 coinpounds that were tested for their trypsin inhibition. To study the diversity of the six sublibraries, we first trained a network with the complete xanthene dataset resulting in a map with all neurons occupied. In this map, we then sent the compounds of the different sublibraries, obtaining altogether six different maps, one each for each sublibrary (Fig. 16).

The six maps show remarkable differences: some of them are nearly completely filled, some of them exhibit large white areas representing neurons that no compound was mapped into. The larger these white areas are, the less the corresponding sublibrary covers the chemical space of the original xanthene library. The omission of the basic or acidic amino acids, for example, has led to a decreased diversity as shown by the large number of empty neurons. On the other hand, the omission of the larger alkyl amino acids or the -OH and -S- substituted amino acids from the xanthene library does not lead to a remarkable decrease in diversity as there are only small white areas in the corresponding maps.

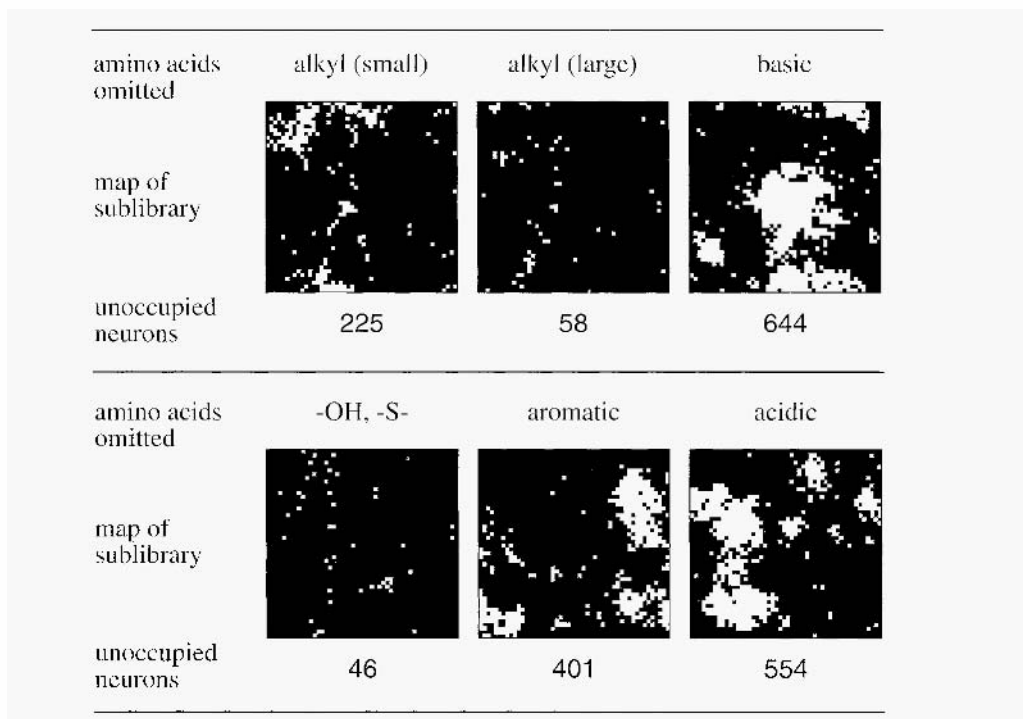


Fig. 16. Kohonen maps of a network trained with the entire xanthene library. Neurons occupied by the different sublibraries are shown in black, unoccupied neurons in white.

## 6. Conclusion and outlook

The human brain generates maps of the environment from sensory information, This capability of the human brain is modelled by self-organizing neural networks such as the one developed by Kohonen. Kohonen networks can be used for the mapping of molecular surface properties. It is shown that maps of the molecular electrostatic potential provide valuable information for understanding biological activity and searching for new lead structures. Kohonen networks can also be used for the mapping of datasets of molecules. Autocorrelation vectors derived from the topology of molecules or from molecular surface properties provide an encoding of molecular structures that can be used as input to Kohonen networks and, thus, allow a clustering of molecules that reflects biological activity. Such a mapping can be used for the assessment of the similarity and diversity of chemical libraries.

The algorithms presented here, both those for the calculation of physico-chemical effects such as the molecular electrostatic potential and that for the Kohonen network, work quite rapidly. In addition, by their very nature, neural networks are of a parallel manner allowing their implementation on parallel machines. This all taken together makes it possible to study large molecules and very large datasets.

## References

1. Kubinyi, H. (Ed.), *3D QSAR in Drug Design: Theory, Methods and Applications*, ESCOM, Leiden. 1993.
2. Kohonen. T., *Self-organized formation of topologically correct feature maps*, Biol. Cybern., 43 (1982) 59–69; *Self-Organization and Associative Memory*, 3rd Ed., Springer. Berlin. 1989; Proc. IEEE, 78 (1990) 1460–1480; *Self-Organizing Maps*, Eds. Huang, T.S., Kohonen. T. and Schröder, M.R., Springer, Berlin, 1995.
3. Zupan, J. and Gasteiger, J., *Neural networks: a new method for solving chemical problems or just a passing phase?*, Anal. Chim. Acta. 248 (1991) 130.
4. Gasteiger, J. and Zupan, J., *Neural networks in chemistry*, Angew. Chem. Int. Ed. Engl., 32 (1993) 502–527; Angew. Chem., 105 (1993) 510–536.
5. Zupan, J. and Gasteiger, J., *Neural Networks for Chemists: An Introduction*, VCH Publishers. Weinheim, 1993.
6. Sadowski, J., Rudolph. C. and Gasteiger, J., *The generation of 3D-models of host-guest complexes*. Anal. Chim. Acta, 265 (1992) 233–241.
7. Gasteiger, J., Rudolph, C. and Sadowski, J., *Automatic generation of 3D-atomic coordinates for organic molecules*, Tetrahedron Comput. Method.. 3 (1990) 537–547.
8. Sadowski, J. and Gasteiger, J., *From atoms and bonds to three-dimensional atomic coordinates: Automatic model builders*, Chem. Reviews. 93 (1993) 2567–2581.
9. Sadowski, J., Gasteiger, J. and Klebe, G., *Comparison of automatic three-dimensional model builders using 639 X-ray structures*, J. Chem. Inf. Comput. Sci., 34 (1994) 1000–1008.
10. Gasteiger, J. and Marsili, M., *Iterative partial equalization of orbital electronegativity — a rapid access to atomic charges*, Tetrahedron, 36 (1980) 3219–3228.
11. Gasteiger, J. and Saller, H., *Calculation of the charge distribution in conjugated systems by a quantification of the resonance concept*, Angrew, Chem. Int. Ed. Engl., 24 (1985) 687–680; Angew. Chem., 97 (1985) 699–701.
12. Hutchings, M.G. and Gasteiger, J., *Residual Electronegativity — an empirical quantification of polar influences and its application to the proton affinity of amines*, Tetrahedron Lett.. 24 (1983) 2541–2544.

13. Gasteiger, J. and Hutchings, M.G., *Quantification of effective polarizability: Applications to studies of x-ray photoelectron spectroscopy and alkylamine protonation*, J. Chem. Soc. Perkin, 2 (1084) 559–564.
14. PETRA program package. Gasteiger, J., Marsili, M., Saller, H., Hutchings, M.G. and Fröhlich, 4., 1995.
15. SURFACE program, Version 1.0, Sadowski, J. and Gasteiger, J., 1994.
16. KMAP, Version 2.1. Li, X., Wagener, M. and Gasteiger, J., 1996.
17. Bauknecht, H. and Zell, A., Universität Stuttgart, 1995.
18. Bauknecht, H., Zell, A., Bayer, H., Levi, P., Wagener, M., Sadowski, J. and Gasteiger, J., *Locating biologically active compounds in medium-sized heterogeneous datasets by topologic al autocorrelation vectors: Dopamine and benzodiazepine agonists*, J. Chem. Inf. Comput. Sci., 36 (1996) 1205–1213.
19. Cramer, R.D., III, Patterson, D.E. and Bunce, J.D., *Comparative molecular field analysis (CoMFA): I. Effect of shape on binding of steroids to carrier proteins*, J. and Chem. Soc., 1 10 (1988) 5959–5967
20. Wagener, M., Sadowski, J. and Gasteiger, J., *Autocorrelation of molecular surface properties for modeling corticosteriod binding globulin and cytosolic ah receptor ac tivity by neural networks*, J. Am. Chem. Soc., 117 (1995) 7769–7775.
21. Dunn, J.F., Nisula, B.C. and Rodbard, D., *Transport of steroid hormones: Binding of 21 endogenous steroids to both testosterone-binding globulin and corticosteriod-binding globulin in human plasma*. J. Clin. Endocrinol. Metab., 53 (1981) 58–68.
22. Mickelson, K.E., Forsthoefel, J. and Westphal, U., *Steroid protein interactions, human corticosteriod binding globulin: Some physicochemical properties and binding specificity*, Biochemistry, 20 (1981) 6211–6218.
23. Westphal, U. (Ed.). *Steroid–Protein Interaction II*, Springer, Berlin. Germany, 1986.
24. Good, A.C., So, S.S. and Richards, W.G., *Structure-activity relationships from molecular similarity matrices*, J. Med. Chem., 36 (1993) 433–438.
25. Jain, A.N., Koile, K. and Chapman, D., *Compass: Predicting biological activities from molecular surface properties: Performance comparison on a steroid benchmark*, J. Med. Chem., 37 (1994) 2315–2327.
26. Gasteiger, J., Li, X., Rudolph, C., J. Sadowski, J. and Zupan, J., *Representation of molecular electrostatic potentials by topological feature maps*, J. Am. Chem. Soc., 1 16 (1994) 4608–4620.
27. Gasteiger, J., Li, X. and Uschold, A., *The beauty of molecular surfaces as revealed by self-organizing neural networks*, J. Mol. Graphics, 12 (1994) 90–97.
28. Li, X., Gasteiger, J. and Zupan, J., *On the topology distortion in self-organizing feature maps*, Biol. Cybern., 70 (1993) 189–198.
29. Gasteiger, J. and Li, X., *Mapping the electrostatic potential of muscarinic and nicotinic agonists with artificial neural networks*, Angew. Chem. Int. Ed. Engl., 33 (1994) 643–646; Angew. Chem., 106 (1994) 671–674.
30. Anzali, S., Barnickel, G., Krug, M., Sadowski, J., Wagenrr, M., Gasteiger, J. and Polanski, J., *The comparison of geometric and electronic properties of molecular surfaces by neural networks: Application to the analysis of corticosteriod binding globulin activity of steroids*, J. Comput.-Aided Mol. Design, 10(1996)521–534.
31. Polanski, J., Gasteiger, J., Wagener, M. and Sadowski, J., *The comparison of molecular surfaces by neural networks and its application to quantitative structure activity studies*, Quant. Struct. Act. Relat. (in print).
32. Holzgrabe, U. and Mohr, K., *Allosteric modulators of lagand binding to muscarinic acetylcholine receptors*, Drug Discovery (in print).
33. Gasteiger, J., Holzgrabe, U., Kostenis, E., Mohr, K., Sürig, U. and Wagener, M., *Variation of the oxime function of bispyridinium-type allosteric modulators of M<sub>2</sub>-cholinoceptors*, Pharmazie, 50 (1995) 99–105.
34. Bejeuhr, G., Holzgrabe, U., Mohr, K., Sürig, U. and von Petersenn, A., *Molecular modelling and synthesis of potent stabilizers of antagonist binding to M<sub>2</sub>-cholinoceptors*, Pharm. Pharmacol. Lett., 2 (1992) 100–103.
35. Holzgrabe, U., Wagener, M. and Gasteiger, J., *Comparison of structurally different allosteric modulators of muscarinic receptors by self-organizing neural networks*, J. Mol. Graph. 14 (1996) 185–195.
36. Barlow, T.W.J., *Self-organizing maps and molecular similarity*, J. Mol. Graph., 13 (1995) 24–27.



37. Anzali, S., Barnickel, G., Krug, M., Sadowski, J., Wagener, M. and Gasteiger, J., *Evaluation of molecular surface properties using a Kohonen neural network: Studies on structure-activity relationships*, In Devillers, J. (Ed.) *Neural networks in QSAR and drug design*, Academic Press, London, 1996, pp. 209–222.
38. Polanski, J., *Neural nets for the simulation of the molecular recognition of MS-WINDOWS environment*, *J. Chem. Inf. Comput. Sci.*, 36 (1996) 694–705.
39. Polanski, J., *The receptor-like neural network for modelling corticosteroid and testosterone binding globulins*, *J. Chem. Inf. Comput. Sci.*, 37 (1997) 553–561.
40. Polanski, J., Ratajczak, A., Gasteiger, J., Galdecki, Z. and Galdecka, E., *Molecular modelling and X-ray analysis for a structure-taste study of  $\alpha$ -Arylsulfonylalkanoic acids*, *J. Mol. Struct.*
41. Moreau, G. and Broto, P., *Autocorrelation of molecular structures: Application to SAR studies*, *Nouv. J. Chim.*, 4 (1980) 757–764.
42. Sadowski, J., Wagener, M. and Gasteiger, J., *Assessing similarity and diversity of combinatorial libraries by spatial autocorrelation functions and neural networks*, *Angew. Chem. Int. Ed. Engl.*, 34 (1995) 2674–2677; *Angew. Chem.*, 107 (1995) 2892–2895.
43. Carell, T., Wintner, E.A., Bashir-Hashemi, A. and Rebek, J., Jr., *A Novel Procedure for synthesis of libraries containing small organic molecules*, *Angew. Chem. Int. Ed. Engl.*, 33 (1994) 2059–2061; *Angew. Chem.*, 106 (1994) 2159–2162; Carell, T., Wintner, E.A., Bashir-Hashemi, A. and Rebek, J., Jr., *A solution-phase screening procedure for the isolation of active compounds from a library of molecules*, *Angew. Chem. Int. Ed. Engl.*, 33 (1994) 2061–2064; *Angew. Chem.*, 106 (1994) 2162–2165.

**This Page Intentionally Left Blank**

# Calculation of Structural Similarity by the Alignment of Molecular Electrostatic Potentials

David A. Thorner, David J. Wild, Peter Willett\* and P. Matthew Wright

Krebs Institute for Biomolecular Research and Department of Information Studies, University of Sheffield, Sheffield S10 2 TN, U.K.

## 1. Introduction

Database searching plays an increasingly important role in drug-discovery programs [1]. Facilities for *substructure searching*, the identification of all of those molecules in a database that contain a user-defined query substructure, have been available in chemical information systems for many years [2]. The last few years have seen the introduction of complementary facilities for *similarity searching* [3]. This involves matching a *target* molecule of interest, such as a weak lead from a high-throughput screening program, against all of the molecules in a database to find the *nearest neighbors* — i.e. those molecules that are most similar to the target using some quantitative measure of intermolecular similarity.

Early database searching systems were designed for the storage and retrieval of two-dimensional (2D) chemical structures but the development of structure generation programs [4] has focused interest on techniques for the processing of three-dimensional (3D) structural information [5], and there have already been several reports of systems for 3D similarity searching that are sufficiently fast in operation to allow them to be used with databases of non-trivial size [6–12]. However, few of these approaches take explicit account of the electrostatic, steric and hydrophobic fields that form the basis of modern approaches to 3D QSAR (as illustrated by the many other papers in this volume), and an ongoing project at the University of Sheffield is hence developing methods for field-based similarity searching. Like many previous workers [13–21], our experiments have focused on the Molecular Electrostatic Potential (MEP), but the techniques that we have developed are applicable, in principle at least, to any field-like attribute that can be represented by real values in a 3D grid surrounding a molecule.

The electrostatic potential,  $P_r$ , at a point  $r$  for a molecule of  $n$  atoms is calculated from the point charges  $q_i$  on each atom  $i$  in the molecule, so that

$$P_r = \sum_{i=1}^n \frac{q_i}{|r - R_i|}$$

where  $R_i$  denotes the position of the  $i$ -th atom. A molecule is positioned at the center of a 3D grid and the potential is calculated at each point in the grid. The similarity between a pair of molecules is estimated by aligning them such that similar features are superimposed, taking the product of the two molecular potentials at each point and then

---

\* To whom correspondence should be addressed.

summing over the entire grid, with a suitable normalizing factor being used to bring the similarities into the range  $-1.0$  to  $+1.0$ . This numerical approach necessarily involves the matching of very large numbers of grid points and is, thus, extremely demanding of computational resources; however, Good et al. [16] have reported an alternative approach in which the potential distribution is approximated by a series of Gaussian functions that can be processed analytically, with a substantial increase in the speed of the similarity calculation and with only a minimal effect on its accuracy. This elegant idea removes one of the main limitations of field-based approaches to 3D similarity searching, but still requires the alignment of the two molecules that are being compared prior to the calculation of the similarity, which is why field-based similarity methods have, thus far, only been applied in the context of small datasets.

The principal objective of the work reported here has been to develop methods for the generation of alignments that are sufficiently rapid in execution to permit the use of MEP-based similarity measures for database searching. This chapter presents the two methods we have developed for this purpose: one method is based on the application of a *maximal common subgraph* (MCS) isomorphism algorithm to a structure representation that we shall refer to subsequently as a *field-graph*; and the other method is based on a *genetic algorithm*, hereafter a GA. Full details of the work reported here are presented by Thorne et al. and by Wild and Willett [22-24].

## 2. Calculation of Similarities Using Field-Graphs

### 2.1. Generation of field-graphs

Chemical database systems use methods of representation and search that are based on *graph theory*. The methods were first developed for 2D substructure searching, where the atoms and bonds of a chemical compound are denoted by the nodes and edges of a labelled graph and where searching is effected by the application of a subgraph isomorphism algorithm to such graph representations [2]. Analogous techniques are used for 3D pharmacophore searching, where the nodes of a graph again denote the atoms of a chemical compound but where the edges denote the interatomic distances (or distance ranges) in a rigid (or flexible) 3D molecule [25]. Graph-based similarity searching methods have also been described, in which an MCS algorithm is used to find molecules that are similar to a target molecule [3,26,27].

It is simple to develop analogous graph methods for the processing of grid-based representations of molecular fields, since a grid can be represented by a graph in which the nodes correspond to each of the grid points and in which the edges correspond to the inter-point distances. The overlap between two sets of molecular fields, and hence the similarity of the two corresponding molecules, will then be given by the overlap between the two graphs, this being estimated most obviously by their MCS. Unfortunately, existing MCS algorithms are capable of handling graphs containing only a few tens of nodes at most, whereas even a coarse molecular grid will involve many hundreds or thousands of grid-points, and hence graph nodes. Accordingly, the use of an MCS algorithm for the generation of molecular alignments requires a substantial

reduction in the sizes of the grid-based graphs that are to be matched, while ensuring that the resulting field-graphs still encompass the main features of the underlying fields, so that the alignments are chemically (and hence biologically) meaningful. Once the MCS procedure has been used to generate an alignment, the similarity between the aligned molecules can be calculated using the fast Gaussian similarity procedure described previously. The similarity measure used is the so-called Carbo index [28], which is actually a form of the long-established cosine coefficient [3] and which is defined to be

$$\frac{\int P_A P_B dv}{\left(\int P_A^2 dv\right)^{1/2} \left(\int P_B^2 dv\right)^{1/2}}$$

where  $P_A$  and  $P_B$  are the properties (such as the MEPS) of the two molecules that are being compared.

A field-graph is generated from a set of grid points by identifying a subset of them that have potential values meeting some criterion (to be discussed below) and then grouping points that meet the chosen criterion and that are close to each other (in some sense). There are many ways in which a field-graph can be generated from a set of potential values, depending on the criterion that is used and the definition of 'close' that is used to determine whether two grid-points are to be considered as belonging to the same group. There are no obvious guidelines, *a priori*, as to how field-graphs should be created, and we have accordingly evaluated a range of different procedures for creating nodes. In fact, we have simplified the problem by considering only a single measure of closeness, which is that two grid-points,  $P$  and  $Q$ , are considered as being contained within the same graph node if  $P$  and  $Q$  are vertically, horizontally or diagonally adjacent to each other. This may be illustrated by considering the set of grid-points in Fig. 1 (which is in just two dimensions for simplicity). If the criterion was that a grid-point value was to be at least 10 kcal/mol, then the four boxed elements would be selected as forming a node in the field-graph.

The basic approach involves thresholding the grid-point values using a user-defined threshold potential: the application of this threshold identifies all of the positive (negative) grid-points with values greater (less) than or equal to the threshold value. The

7	8	9	5
8	11	13	9
4	10	9	7
1	6	10	5
3	2	1	0

Fig. 1. Application of a threshold of +10 to a set of grid-point values to generate a field-graph. In this simple, planar grid, the four values surrounded by bold lines are taken as forming a single field-graph node since they all meet the threshold criterion and each is adjacent to at least one other grid-point that meets the threshold criterion. The resulting four-point node is shown shaded.

selected subset of the original points is then considered for inclusion in one of the nodes in the field-graph, using the adjacency-based grouping criterion described above. Thorner et al. describe several different methods for the selection of the most appropriate threshold potential: the most cost-effective of these methods was found to be one that seeks to maximize the number or nodes in the field-graph that contain at least some minimal number of grid-points [24].

Each node in a field-graph has an associated label defining the threshold level (either positive or negative) at which it was formed and the size of the node — i.e. the number of grid-points that it represents. The location of a node is defined by the geometric coordinates of the center of the corresponding cluster of points, so that the distance between two nodes (which represents the graph-edge that connects those two nodes) is the distance between the two centers. Thus, a field-graph containing  $N$  nodes (i.e. one that contains  $N$  clusters of grid-points after the application of the procedures described above) will contain  $N(N - 1)/2$  distinct distances (since the inter-node distance matrix is completely symmetric), and a database of 3D structures can be represented for search by the corresponding set of field-graphs.

Pairs of these graphs, one representing a target structure and one representing a database structure, are matched using an MCS procedure based on the Bron-Kerbosch clique-detection algorithm [29,30]. Two nodes are regarded as being equivalent if their labels have the same potential sign (either positive or negative); it is possible to adopt a more rigorous matching criterion that takes account of the magnitude of the local potential, as well as its sign or of the size of the node (i.e. the number of constituent grid-points), or of some combination of the two, but all our experiments have employed just the simple, sign-based matching criterion. The Bron-Kerbosch algorithm identifies the MCS based on such node-to-node matches, and the constituent pairs of matching nodes are input to a least-squares fitting routine. This routine aligns the database structure with the target structure so as to minimize the squared distances between the centroids of the field-graph nodes from the database structure with the centroids or the field-graph nodes from the target structure to which they have been mapped. The resulting alignment is then used for the final grid-based similarity calculation.

The Bron-Kerbosch algorithm is designed, and normally used, to identify the largest subgraph common to a pair of graphs. In the present context, however, we have used it to generate all of the subgraphs common to a pair of field-graphs, and not just the largest such subgraph(s). This has been done to increase the number of possible alignments that are considered in the precise similarity calculation, and thus to ensure that the minimum possible number or close neighbors to the target structure are overlooked because of the approximations involved in the generation of the field-graphs. The similarity between an individual database structure and the target structure is taken to be the largest of the calculated similarities over all of the possible alignments resulting from the first-level, clique-detection search.

## 2.2 *An operational implementation*

The field-graph generation procedure described above was applied to the structures in the Zeneca Agrochemicals corporate database. The 3D structures of these molecules

were generated using CONCORD [31] and then the atomic point charges calculated using the MNDO routines in MOPAC Version 5.00 [32]. A grid was constructed around each molecule, extending for 5.0 Å beyond the maximum and minimum extents of the molecule in each plane, as determined by the centers of the atoms on the molecular surface. The grid had a user-defined step-size (0.5 Å in all of the experiments reported here), with grid-points being ignored if they fell within the van der Waals radius of any of the atoms in the molecule. The grids were then converted to field-graphs: the mean number of nodes in the field-graphs for the 173 197 molecules in the database was 7.13 (standard deviation of 3.22), with the single largest graph containing 41 nodes. The generation of the field-graphs took about 20 CPU days on an R4000 Silicon Graphics workstation (excluding the very extensive computation associated with the calculation of the atomic partial charges).

Preliminary testing showed that the search times for a scan of the entire file were likely to be excessive unless a fair amount of program optimization was to be carried out.

The relative computational requirements of the MCS-detection and Gaussian-calculation stages for the matching of the target structure with a database structure depend upon the number of possible alignments resulting from the application of the MCS algorithm. Specifically, if only a few possible alignments are identified, then the MCS stage takes more time, while the Gaussian calculation takes more time if there are many alignments (since each one of them needs the similarity calculation to be carried out). Most of the time requirement for the Gaussian calculation is occasioned by the need to calculate square roots and exponentials, and the routine was hence modified to minimize the number of square roots and to calculate most of the exponentials by means of a precalculated look-up table. This optimization was also used in all of the GA experiments described later in this chapter.

The time requirement for the MCS algorithm rises rapidly with an increase in the number of nodes in the target structure, given a fixed set of database structures, and this requirement will hence be minimized if the target structure contains just a few nodes. A field-graph with only a few nodes will also generate only a small number of alignments, and hence reduce the time requirement of the final similarity calculations. A series of searches was hence carried out in which a threshold was applied to the sizes of the nodes in the field-graph representing the target structure, and a cutoff applied, so that only those nodes containing more than  $n$  grid-points, where  $n$  is defined by the user, were considered in the matching of a target structure with a database structure. This procedure will certainly increase the speed of searching, but may also mean that molecules that are, in fact, strongly similar to the target structure in the final similarity calculation will be overlooked since the appropriate alignments are not forthcoming from the MCS stage of the search. Following a detailed series of experiments [24], the final implementation was such that the alignment stage considered only those target field-graph nodes for which  $n \geq 3$ . The chosen value for  $n$  could, of course, be changed by a user if this was felt to be desirable in the context of a particular search.

With the incorporation of these and other modifications, a search of a typical target structure against the entire database can be accomplished in about 16 hours of elapsed time on an R4000 Silicon Graphics workstation (i.e. in a single overnight run), although

large target structures can mean that a search will take a day or even more. It should be noted that some of the molecules in the database can lead to very large numbers of alignments, each of which must then be checked in the subsequent Gaussian calculation: searches with several target structures demonstrated that the search time could be reduced by approx. 40% simply by ignoring the top-ranked 5% of the database when the structures were ranked in decreasing order of the number of alignments that needed to be processed.

Searches with the system at Zeneca Agrochemicals demonstrate clearly that it often leads to the retrieval of structures with a high degree of novelty that would not be retrieved by conventional similarity searching techniques that are based on patterns of atoms (in either 2D or 3D) [3]. The system hence provides an effective way of suggesting novel bioisosteres for known active compounds. That said, the focus on just a single type of field means that while the top-ranked structures generally provide a reasonable level of MEP similarity, many of them are of little interest since their steric and/or lipophilic characteristics render them inappropriate for the biological system under investigation. The principal value of the system is hence as an ideas-generator that can suggest previously unexplored chemical classes to the chemist requesting the search, albeit at the cost of (in some cases) a low level of search precision.

There are two other obvious limitations to the current system. The first is that the graph-creation routines have an inherent failure rate of around 6% since the threshold criteria that are used to create a field-graph can result in the identification of less nodes than are necessary for the generation of a unique alignment [24]. Secondly, inspection shows that suboptimal alignments are generated in some cases, with the result that molecules that are similar to the target structure can be ranked less highly than they should be. This is almost inevitable, given the simplicity of the representation that is used.

### 3. Calculation of Similarities Using a Genetic Algorithm

#### 3.1. The algorithm

Genetic algorithms are computational problem-solving methods that mimic some of the principal characteristics of biological evolution and genetic reproduction [33,34]. A GA creates a randomly chosen set, known as a *population*, of individuals, each of which contains a representation of a possible problem solution. This solution is encoded in a linear string, called a *chromosome*. The effectiveness of the solution encoded by each of the chromosomes in a population is measured by the *fitness function*, and the GA manipulates the chromosomes so as to maximize the value of the fitness function. This it does by the creation of subsequent populations that include features from the fitter strings in the previous population, in an iterative procedure that can be thought of as an algorithmic representation of biological reproduction. Parents are selected from the population, and information is taken from their chromosomes to produce one or more child individuals that are inserted into the population. Chromosomes are manipulated by *mutation* (where the chromosomal material may be altered slightly in a random fashion) and *crossover* (where new child chromosomes are created by taking some chromosomal



material from one parent, and some from the other) operators. A GA may be considered to have succeeded when *convergence* occurs — i.e. when the members of the population all lie in the same region of search space.

There have been several reports of the application of GAs to problems in chemical structure handling, demonstrating that they provide both an effective and an efficient mechanism for the investigation of a range of complex matching problems [35,36]. Following earlier work by Payne and Glen [37], the GA we have developed here seeks to identify a combination of translations and rotations that will align one MEP with another, fixed, MEP so as to give the highest possible similarity. Each chromosome contains six components, three to encode rotations and three to encode translations, with each component being allocated one byte. The chromosomes are initially set to random values, and then decoded by applying the indicated rotations and translations to the 3D coordinates of the atoms in one of the two molecules that are being aligned. The resulting coordinates are passed to the fitness function for the evaluation of the alignment defined by that particular set of rotations and translations: this function is the Gaussian similarity.

While GAs are simple in concept, there are generally many different ways in which they can be parameterized. Wild and Willett [22] describe the very extensive comparative experiments that were carried out to ensure an appropriate combination of *effectiveness* — i.e. the ability to identify good MEP alignments — and *efficiency* — i.e. the CPU time required (since the GA must be sufficiently fast in operation to search databases of non-trivial size). The final GA that was used for the experiments reported in the next section had a population size of ten, steady-state-without-duplicates reproduction, binary encoding, a static uniform crossover rate of 20%, linear normalization and a diversely initialized population.

### *3.2. Comparison of alignment methods*

Having described two different ways of generating alignments, the question arises as to which is the more effective — this, in turn, requiring some quantitative means of evaluating the effectiveness of the similarity measures that are being tested. Previous work in our laboratory on the comparison of different similarity or clustering methods [38] has made use of the *similar-property principle* of Johnson and Maggiora [39] for this purpose. Here, simulated property-prediction experiments are carried out using datasets for which both structural and property data are available, so as to ascertain which methods (e.g. which similarity coefficients) result in measures of structural similarity that are most closely correlated with measures of property similarity. We have adopted a rather different approach in the work reported here. The extensive studies of Richards and co-workers [13,16,19,20] have shown clearly that there is a strong correlation between biological activity and the similarities that result from grid-based MEP calculations (and further evidence of such a correlation is presented later in this chapter). Accordingly, we can compare the effectiveness of the field-graph and GA methods for matching MEPs by the magnitudes of the Gaussian similarities since a high similarity will be achieved if, and only if, an appropriate alignment has been achieved. A

comparable approach had been taken previously in identifying the best way of generating field-graphs [24].

Our experiments used a test database of 1000 molecules taken from the Fine Chemicals Database (FCD), from which 100 molecules were chosen at random to act as the target molecules for which the nearest-neighbor molecules were required. The dataset was processed using CONCORD and the MNDO routines in MOPAC, as described previously, and then searches were carried out using the field-graph and GA methods for generating alignments, with the methods parameterized such that they took about the same amount of time (1–2 CPU seconds on a medium-level Unix workstation) to match a pair of MEPs.

Let  $S_i$  be the Gaussian similarity for the  $i$ -th most similar molecule to the  $m$ -th target molecule; then we define the performance measure  $E_{mn}$  by

$$E_{mn} = \frac{1}{n} \sum_{i=1}^n S_i$$

where typical values for  $n$  are 5, 10 or 20 — i.e.  $E_{mn}$  is the mean MEP similarity for the  $n$  nearest-neighbor structures of the  $m$ -th target structure. The overall effectiveness of the set of searches for the  $n$  nearest neighbors of each target structure,  $E_{mn}$ , is then obtained by taking the mean of the 100 individual  $E_{mn}$  values. When this was done, there was found to be no obvious difference between the  $E_n$  values obtained using the field-graph and GA methods for generating the alignments; for example, both gave  $E_{20}$  values of 0.70 when averaged over the 100 searches of the FCD dataset, and a Wilcoxon signed-rank test [40] showed that the two sets of searches were not significantly different in performance at the 0.05 level of statistical significance. Wild and Willett report additional experiments in which the alignments resulting from the GA were also shown to be at least as good as those resulting from a bit climber and from a simplex optimizer [22].

#### 4. Inclusion of Conformation Flexibility

The experiments that have been described thus far have assumed that the molecules that are being processed are completely rigid in nature. However, most organic molecules contain one or more rotatable bonds, and it is to be expected that improved intermolecular similarity relationships will be identified in a database search if the MEP-alignment procedure is able to take account of flexibility in a target structure and/or in a database structure.

The MEP at any point in 3D space is a function of the partial atomic charges and the distances of each atom from that point, and a full treatment of the effect of conformational flexibility on MEP-based similarity searching should thus take account of the changes in both the atom-to-point distances and the partial atomic charges that can occur as a molecule flexes [41]. However, the calculation of the partial atomic charges required for the generation of an MEP is very time-consuming. Our work on flexible field-based searching hence involves making the assumptions that the partial charges do

not vary with the conformation adopted by a flexible molecule, and that the MEP is affected only by the changes in the atom-to-point distances resulting from torsional rotation.

In seeking to find an appropriate alignment procedure, we have been guided by the extensive studies that have been carried out into techniques for flexible 3D substructure searching, where two main approaches have been described [1,5,25]. In the first approach, a flexible molecule is characterized by a small number of conformations that are checked to ascertain whether any of them contain a query pharmacophore [42–44]. The alternative approach involves a torsional optimization approach that permits an exploration of the full conformational space of a flexible molecule at search time, seeking to determine whether it can adopt a conformation that contains the pharmacophore [45–47]. These two approaches are considered further below.

#### 4.1. Searching flexible molecules using field-graphs

Our initial experiments involved the application of field-graphs to the multi-conformation method of flexible searching [23]. Specifically, we sought to determine the numbers of field-graphs that are required to delineate fully the variations in MEP that result from variations in conformation, since the field-graph approach will only be applicable to searching databases of non-trivial size if these numbers are not large.

The experiments involved a dataset of eleven compounds, taken from the Cambridge Structural Database (CSD) [48], that had been used previously by Ghose et al. to evaluate conformational searching methods [49]. The SYBYL SEARCH module was used to generate a number of conformations (between 48 and 1589) for each of the eleven molecules by systematic increments of their rotatable bonds, and a field-graph was then generated from each of the resulting conformations. The set of field-graphs for the set of conformations for each molecule was next converted into a searchable database, with the target structure in each case being the field-graph that was generated from that molecule's CSD crystal structure.

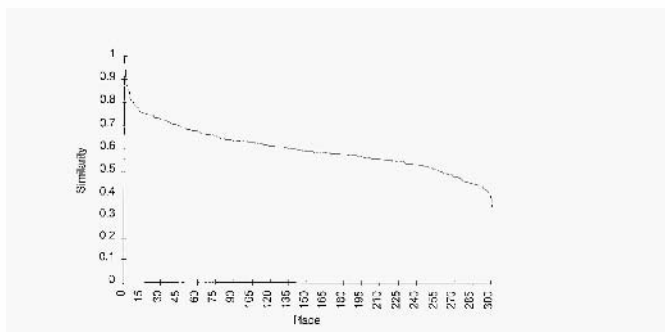
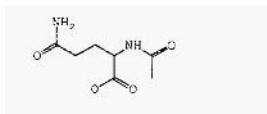


Fig. 2. Similarity values for the molecule AGLUAM10 after sorting into descending order. Each value is the MEP similarity for the matching of the CSD structure of this molecule against one of its conformers (see text for details).

The results of the searches were summarized as shown in Fig. 2. which is a sorted list of the similarities calculated for the 301 conformations that were generated for one of the eleven molecules, AGLUAM10.



AGLUAM10

It will be seen that there are a few conformations with similarities in excess of 0.80 but that the great majority of the conformations have much smaller similarities: for this dataset, the mean similarity is 0.601 with a standard deviation of 0.096. Similar results were obtained with all of the other molecules that were considered, with the mean (standard deviation) similarity varying from 0.308 (0.071) to 0.887 (0.058). If all of the conformations for a particular molecule gave comparable MEPs, and hence comparable field-graphs, then most of the similarities with the target structure would be near to 1.0; in fact, most of them are very much smaller than this (with one of the conformations for one of the molecules yielding a similarity as low as 0.202).

It is, thus, clear that torsional rotations can bring about substantial changes in the MEP similarity with, in extreme cases, just a small change in a single torsion angle resulting in changes of up to 40% in the similarity between the target structure and a conformation [23]. There are two reasons for such drastic changes in the similarity. The first, and most obvious, is a change in the MEP itself arising from the rotation of a bond near to the center of a molecule. This can result in large-scale changes in the overall geometry of the molecule, and hence in the atom-to-point distances that are used in the calculation of the potential at each point in the 3D grid surrounding a molecule. A second problem is the lack of robustness in the routine that is used to generate a field-graph from the set of grid-point potentials, once they have been calculated. We have found that the form of the field-graph produced by this routine can be overly sensitive to the precise values of these potentials, with only small changes in the values sometimes leading to changes in the number or nodes in the resulting field-graph: these changes can affect the alignments generated by the MCS procedure and, hence, the final intermolecular similarities.

Substructure searches make use of extensive screening strategies to minimize the number of molecules for which a detailed search needs to be carried out. One strategy that has been found to be particularly effective in the context of flexible 3D substructure searching is to associate a *distance range* with each pair of atoms in a flexible molecule [45]. Here, the lower and upper bounds of the range correspond to the minimum and maximum separations of the two atoms as the molecule flexes, so that the set of distance ranges for a molecule contains all of the geometrically feasible conformations which that molecule can adopt. This representation allows the use of graph-based screening procedures, which ensure that only those molecules that match the query at the graph level proceed to the final, detailed conformational search [46,47].

Some preliminary experiments were thus carried out to investigate the extent to which it might be possible to associate an analogous *potential range* with each point in the 3D MEP grid surrounding a molecule. Specifically, each point in a grid was charac-

terized by not one but two values, MAX and MIN, which are the maximum and the minimum values of the potential that are observed at that point as the molecule flexes. However, it was found that the potential ranges,  $|\text{MAX}-\text{MIN}|$ , were generally large and, thus, unable to provide any useful screening capability [23].

We have already noted some inherent limitations of the field-graph approach when discussing its application to the searching of rigid 3D molecules. The results obtained here suggest that it will be difficult for a field-graph representation derived from a single conformation to provide an adequate description of the variations in MEP that can occur as a result of torsional rotations: instead, many conformations, and thus field-graphs, will be required for such a description. Since the matching of the field-graphs representing a single conformation of a target structure and a single conformation of a database structure requires 1–2 CPU seconds, it is clear that the use of multiple field-graphs for flexible field-based searching will be extremely time-consuming. In addition, the parameter-driven nature of a GA makes it easier to bias a search towards efficiency (i.e. a search that runs quickly but that may miss some good hits) or towards effectiveness (i.e. a search using a larger population size and/or a greater number of generations). Accordingly, the remainder of this chapter focuses upon the use of the GA method for flexible similarity searching.

#### *4.2. Searching flexible molecules using a genetic algorithm*

The GA that we have developed can be used in two ways. In the first, which is described below and which formed the basis for most of the initial experiments, a rigid target structure is assumed but each of the database structures is allowed to be flexible; alternatively, the target structure can also be allowed to be flexible. We shall refer to these two types of search as ‘Flex’ and ‘FlexFlex’, respectively; searches when both molecules are rigid will be referred to as ‘Fixed’.

The Flex GA is a straightforward extension of that described previously. The chromosome again encodes the translations and rigid-body rotations, but augments these by an extra component for each rotatable bond in the database structure (and in the target structure in the more general, FlexFlex case) such that not only the position, but also the conformation, of a molecule is encoded, and a further component to encode corner flipping in each flexible ring. The fitness function is augmented by a simple van der Waals radius bumpcheck procedure, which ensures that the torsion angles encoded in a chromosome do not represent a high-energy conformation. If the bumpcheck is successful, then the alignment defined by the chromosome acts as the input to the Gaussian similarity calculation, with the resulting similarity being the raw fitness value for that chromosome. The chromosomes are ranked in decreasing order of raw fitness from the size of the population down to one (which is the least-fit chromosome in the current population), with their position in the ranked list being taken as the fitness. A starting value was also added to each chromosome to provide a fitness window. This last approach was found to give a value of 1.1 (for Flex) and 1.5 (for FlexFlex) for the *selection pressure*, that is, the ratio of the fitness of the fittest chromosome to the mean fitness of the population.

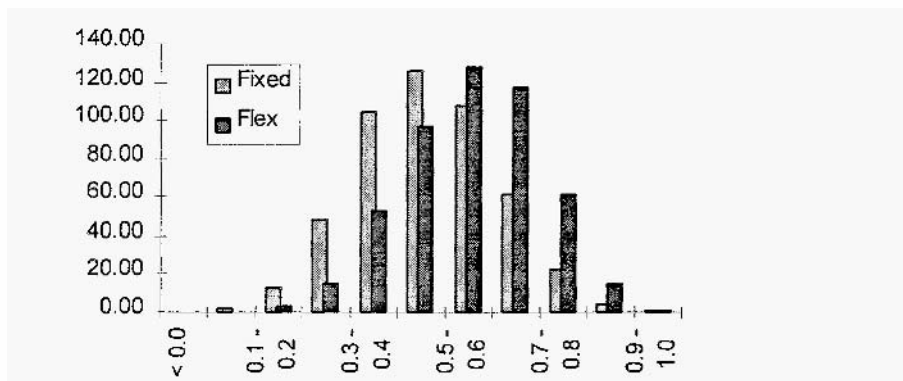


Fig. 3. Distribution of observed MEP similarities for matching a rigid target structure against 491 rigid ('Fixed') or flexible ('Flex') database structures. Each column denotes the mean number of similarities in the indicated range, when averaged over the 39 target molecules and the 10 runs for each target molecule that were used in these experiments

As with the previous GA, a large number of parameterization experiments were carried out to maximize the performance of the algorithm. These experiments, which are detailed by Thorner et al. [23], led to the use of two-point crossover and a bit-flip mutation operator, with other parameters having the following values (with the range of values that were tested included in brackets): a population of 150 chromosomes (varied in the range 5–500); a selection pressure of 1.5 (varied in the range 1.05–1.90); 1250 generations (varied in the range 50–5000); a crossover rate of 35% (varied in the range 2–90%); and a mutation rate of 7% (varied in the range 0.1–90%). The effectiveness of the alignments produced by the GA, and hence the values of the final similarity coefficient, increases in line with the number of generations; the 1250 value used here was felt to provide the best trade-off between effectiveness and efficiency, with run times being about 3.5 CPU seconds for calculating the similarity between a rigid target structure and a flexible database structure, using an implementation of the algorithm in the C programming language on a Silicon Graphics R4000 workstation.

The GA was tested using a set of 491 structures from the Fine Chemicals Database, 39 of which were used in turn as the rigid target structure for two types of search. In the first set of Flex searches, the database structures were allowed to be completely flexible, with the search being effected using the GA parameter values given previously. In the second set of Fixed searches, the database structures were kept entirely rigid so as to determine the increase in performance, if any, resulting from the inclusion of torsional flexibility in the matching process. The GAs that were used in these two sets of runs were parameterized so as to take about the same amount of CPU time, and both sets of searches were repeated ten times to encompass the variations in performance that result from the non-deterministic nature of GAs. The mean similarity of each target structure to each database structure, when averaged over all 39 target structures, all 491 database structures in each case and all ten sets of runs was 0.465 for the Fixed searches and 0.556 for the Flex searches. The distribution of the calculated similarities is shown in Fig. 3, which

illustrates the marked shift to higher similarities that results from the inclusion of database-structure flexibility in the matching algorithm.

Thus far, we have considered only the flexibility of the database structures, while keeping the target structure rigid. It is, however, simple to extend the algorithm to include any inherent flexibility in the target structure. All that is required is to extend the chromosome by a further byte for each rotatable bond in the target structure, and to treat these bytes in just the same way as is currently done for the corresponding bytes in a description of a database structure. It might be expected that allowing both the target structure and the database structures to flex would further increase the similarities that were obtained, when compared with the Flex searches. This was, however, not found to be the case, since while the FlexFlex searches gave mean similarities that were again greater than those of the Fixed searches, the largest mean similarities resulted from the Flex searches. Specifically, the Fixed, Flex and FlexFlex searches gave mean similarities of 0.398, 0.522 and 0.515, respectively, when averaged over these sets of searches and when parameterized to take similar amounts of CPU time. We believe that this seemingly counter-intuitive result arises from the frequent non-convergence of the FlexFlex searches during the limited time available for the execution of the GA, especially when the target structure had a large number of rotatable bonds.

We, thus, concluded that while the inclusion of conformational flexibility in a field-based similarity search enables the identification of better MEP overlaps than if only rigid molecules are considered, the cost-effectiveness of allowing both the molecules in a comparison to flex required further study, as detailed below.

## **5. Prediction of Biological Activity**

Searching a chemical database to find molecules that are similar to a bioactive target structure, and that might thus also be expected to exhibit the activity of interest, is one of the principal applications of similarity searching [1,3]. The work reported thus far in this chapter has not considered the effectiveness of our matching algorithms for this purpose and we, hence, now report some initial experiments to ascertain the extent to which field-based similarity searching might be of use in lead discovery programs. In these sets of experiments, the 3D CONCORD structures were minimized using SYBYL MAXMIN, and the atomic charges calculated using the PM3 routines in MOPAC.

The first tests used six small datasets that we have employed previously in a study of methods for distance-based 3D similarity searching [50]. Each of these datasets contains molecules for which qualitative (active/inactive) data are available, as follows:

1. 209 9-anilinoacridines of which 150 showed anti-tumor activity [51];
2. 147 barbiturates of which 37 had sufficient durations of activity to be classed as active [52];
3. 112 nitrobenzenes of which 53 were musk odorants [53];
4. 141 aromatic amines of which 98 were carcinogenic [54];
5. 113 steroids of which 69 showed potent anti-inflammatory activity [55];
6. 145 nitrosamines of which 112 were carcinogenic [56].

For each dataset, an active compound was selected and used as the target structure for Fixed, Flex and FlexFlex searches of that dataset. This was repeated for a total of ten different target structures, and the mean similarities that were obtained are listed in Table 1. It will be seen that, on average, the Flex searches again give the largest similarities and hence, one might assume, the best alignments. Different results are obtained, however, when the number of active nearest neighbors that were retrieved is considered, as illustrated in Table 2. The columns in the main body of the table list the mean number of actives that were retrieved when the 20 nearest neighbors were considered for each of the ten active target structures mentioned previously. The columns headed Fixed, Flex and FlexFlex have their normal meanings, while those headed AM and MCS represent the results that were obtained using the Atom Mapping (AM) and Maximal Common Substructure (MCS) similarity measures discussed by Pepperrell and Willett in their review of distance-based measures for 3D similarity searching [50]. The results suggest that FlexFlex is the best of the three MEP-based similarity measures and that it is at least as effective as the two distance-based similarity measures.

The QSAR datasets are limited both in size and in structural heterogeneity, and also have a large proportion of actives; the second set of experiments hence used a file

Table 1 Mean similarities when averaged over 10 target structures for Fixed, Flex and FlexFlex MEP similarities of each of six small QSAR datasets

Dataset	Fixed	Flex	FlexFlex
A	0.493	0.512	0.592
B	0.642	0.805	0.736
C	0.608	0.679	0.667
D	0.739	0.764	0.694
E	0.684	0.774	0.821
F	0.729	0.811	0.749
Average	0.670	0.724	0.709

Table 2 Mean number of actives in the top 20 nearest neighbors when averaged over 10 target structures for Fixed, Flex, FlexFlex, AM (atom mapping) and MCS (maximal common substructure) similarity measures in each of 6 small QSAR datasets

Dataset	Fixed	Flex	FlexFlex	AM	MCS
A	13.3	13.6	15.4	16.8	16.0
B	7.1	8.9	9.7	6.8	10.6
C	11.5	11.9	12.3	12.3	11.3
D	16.5	14.9	15.6	14.9	15.5
E	15.6	14.7	14.7	16.2	14.7
F	18.2	18.2	19.0	16.3	17.4
Average	13.7	13.7	14.4	13.9	14.2



of 3500 structures drawn from the *World Drugs Index* (WDI), which contains 2D structures and broad-class activity indicants for some tens of thousands of drugs.

The target structures we adopted were those used recently by Kearsley et al. in a study of properly-based similarity searching [57]. These molecules are listed in Table 3, together with their associated WDI activity classes and the number of molecules belonging to that particular class. (Captopril is described in the WDI as an angiotensin antagonist, rather than its true nature as an inhibitor of angiotensin-converting enzyme, and we have thus used this description in the searches reported here.) For the purposes of these experiments, drugs that lie within the same activity class or classes as the target structure are considered to be actives. For each of the target structures, all of the actives were added to the WDI subset mentioned previously, and then Fixed, Flex and FlexFlex similarity searches carried out to retrieve the nearest neighbors for the target structure in each case.

The results that were obtained are summarized in Table 4, which contains mean values averaged over the ten target structures that were used. Here, the first row in the main body of the table gives the mean similarity between the target structure and the database structures, while the following four rows list the mean numbers of actives retrieved when the top 10, top 20, top 50 and top 100 nearest neighbors were retrieved. It will be seen that the relative performance of the three types of search depends on the number of nearest neighbors that are retrieved. with the Fixed searches giving the best results for a precision-oriented search in which just a few, highly similar, structures are required, and with the FlexFlex searches giving a relatively better level of performance as more recall-oriented searches are required. As with the earlier experiments. the Flex searches yield the highest mean similarities, and we would thus hypothesize that the FlexFlex would perform even better at retrieving active molecules if it were allowed to run for long enough to give similarities that were comparable with those obtained in the Flex searches.

It must be emphasized that the results presented here are only preliminary, but they do suggest that the inclusion of flexibility information can increase the effectiveness of field-based searching at little computational cost, when compared with the corresponding non-flexible searches.

## **6. Conclusion**

Several previous workers have described methods for calculating the similarities between pairs of molecules characterized by their MEPs. In this chapter, we have presented algorithms and data structures that enable such similarities to be calculated sufficiently rapidly to enable MEPs to be used for field-based similarity searching in chemical databases.

Our first approach involves the use of an MCS algorithm to align the field-graphs representing the target structure and each database structure. The alignment(s) resulting from this algorithm are then used for the calculation of the final MEP-based similarity measure. The second approach involves a GA that encodes the translations and rotations needed to maximize the overlap of the MEPs of a target structure and a database structure. This is found to be comparable in effectiveness with the field-graph approach

Table 3 Target structures used in the searches of the World Drugs Index

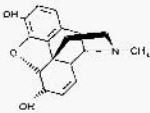
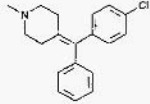
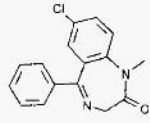
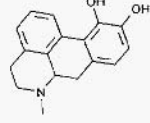
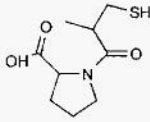
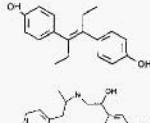
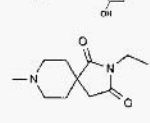
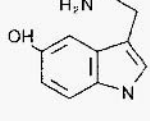
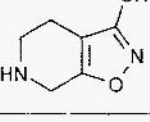
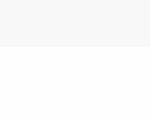
Target	Structure	WDI classification	No. of Actives
Morphine		Narcotic or opioid	128
Cycliramine		Antihistaminic-III	261
Diazepam		Tranquilizer or benzodiazepine-agonist	113
Apomorphine		Dopaminergics	117
Captopril		Angiotensin antagonists	86
Diethylstilbesterol		Estrogens	189
Fenoterol		Sympatho-mimetics-beta	65
RS86		Parasympatho-mimetics	93
Serotonia		Serotoninergics	38
Gaboxadol		Gabaminergics	34

Table 4 Results of searching the World Drug Index; the Similarity row gives the mean similarity when averaged over the 10 target structures while each of the remaining rows lists the total number of actives retrieved for a given number of nearest neighbors (NNs)

Effectiveness Measure	Fixed	Flex	FlexFlex
Similarity	0.481	0.592	0.556
Actives (NN = 10)	5.4	4.8	4.0
Actives (NN = 20)	8.8	7.8	7.9
Actives (NN = 50)	12.5	10.9	13.9
Actives (NN = 100)	18.8	17.8	19.1

when rigid structures are considered, but is noticeably more efficient if flexible similarity searching is to be carried out.

The studies reported here are now being extended in two ways. Firstly, we are carrying out further searches on the *World Drugs Index* dataset using not just the MEP similarity measure, but also measures based on 2D fingerprints and on 3D interatomic distances (specifically the atom-mapping measure [6]). The aim of this work is to investigate the extent to which the various similarity measures retrieve different sets of bioactive molecules as the output from a similarity search. Our initial experiments suggest that the search outputs do, indeed, differ, with those resulting from the MEP measure being noticeably more heterogeneous than those resulting from the 2D and 3D measures, thus supporting the hypothesis that field-based searching provides a way of identifying novel bioisosteres that would not be retrieved by more conventional, atom-based searching procedures. Secondly, we have already noted, when discussing the system at Zeneca Agrochemicals, that the focus on just a single type of field means that while the top-ranked structures in a search generally provide a reasonable level of MEP similarity, many of them are of little interest since their steric and/or lipophilic characteristics render them inappropriate for the biological system under investigation. We are, accordingly, extending the GA to enable all three types of field information to be taken into account when deciding which molecules should be retrieved in a database search, with the expectation that this will further improve the retrieval of bioactive molecules from database searches. Full details of both of these sets of experiments will be reported shortly.

## Acknowledgements

We thank the following: the Biotechnology and Biological Sciences Research Council, the James Black Foundation, the Science and Engineering Research Council and Zeneca Agrochemicals for funding; Harold Cox, Jonathan Davies, Bobby Glen, Gareth Jones, Caroline Low, Anne Mullaley, Robin Taylor and Andy Vinter for helpful discussions on genetic algorithms and molecular electrostatic potentials; Derwent Information for providing the *World Drugs Index*; and Tripos Inc. for hardware and software support. The Krebs Institute for Bionolecular Research is a designated Bioinolecular Sciences Centre of the Biotechnology and Biological Sciences Research Council.

## References

1. Good, A.C. and Mason, J.S.. *Three-dimensional structure database searches*, Rev. Comput. Chem., 7(1995)67-117.
2. Barnard, J.M., *Substructure searching methods: Old and new*, J. Chem. Inf. Comput. Sci., 33 (1993) 532-538.
3. Downs, G.M. and Willett, P., *Similarity searching in databases of chemical structures*, Rev. Comput. Chem., 7 (1995) 1-66.
4. Sadowski, J. and Gasteiger, J., *From atoms and bonds to three-dimensional atomic coordinates: Automatic model builders*, Chem. Rev., 93 (1 993) 2567-2581.
5. Martin, Y.C. and Willett, P. (Eds.). *Designing bioactive molecules: Three-dimensional techniques and applications* (in press).
6. Pepperrell, C.A., Willett, P. and Taylor, R.. *Implementation and use of an atom-mapping procedure for similarity searching in databases of 3D chemical structures*, Tetrahedron Comput. Methodol., 3 (1990) 575-593.
7. Perry, N.C. and van Geerestein, V.J., *Database searching on the basis of three-dimensional molecular similarity*, using the SPERM program. J. Chem. Inf. Comput. Sci., 32 (1992) 607-616.
8. Fisanick, W., Cross, K.P. and Rusinko, A., *Similarity searching of CAS Registry substances: I. Global molecular property and generic atom triangle geometric searching*, J. Chem. Inf. Comput. Sci., 32 (1992) 664-674.
9. Bemis, G.W. and Kuntz, I.D.. *A fast and efficient method for 2D and 3D molecular shape description*, J. Comput.-Aided Mol. Design, 6 ( 1992) 607-628.
10. Nilakantan, R., Bauman, N. and Venkataraghavan, R., *A new method for rapid characterisation of molecular shape: Applications in drug design*. J. Chem. Inf. Comput. Sci., 33 (1993) 79-85.
11. Bath, P.A., Poirrette, A.R., Willett, P. and Allen, F.H., *Similarity searching in files of three-dimensional chemical structures: Comparison of fragment-based measures of shape similarity*, J. Chem. Inf. Comput. Sci., 34 (1994) 141-147.
12. Good, A.C., Ewing, T.J.A., Gschwend, D.A. and Kuntz, I.D.. *New molecular shape descriptors: Application in database screening*, J. Comput.-Aided Mol. Design, 9 (1995) 1-12.
13. Burt, C., Richards, W.G. and Huxley, P.. *The application of molecular similarity calculations*, J. Comput. Chem., 11 (1990) 1139-1146.
14. Manaut, F., Sanz, F., Jose, J. and Milesi, M., *Automatic search for maximum similarity between molecular electrostatic potential distributions*, J. Comput.-Aided Mol. Design, 5 (1991) 371-380.
15. Richard, A.M., *Quantitative comparison of molecular electrostatic potentials for structure-activity studies*. J. Comput. Chem., 12 (1991) 959-969.
16. Good, A.C., Hodgkin, E.E. and Richards, W.G.. *The utilisation of Gaussian functions for the rapid evaluation of molecular similarity*; J. Chem. Inf. Comput. Sci., 32 (1992) 188-191.
17. Sanz, F., Manaut, F., Rodriguez, J., Lozoya, E. and Lopez-de-Brinas, E., *MEPSIM: A computational package for analysis and comparison of molecular electrostatic potentials*, J. Comput.-Aided Mol. Design, 7 (1993) 337-347.
18. Petke, J.D., *Cumulative and discrete similarity analysis of electrostatic potentials and fields*, J. Comput. Chem., 14 (1993) 928-933.
19. Good, A.C., Peterson, S.J. and Richards, W.G.. *QSARS, from similarity matrices: Technique validation and application in the comparison of different similarity evaluation methods*, J. Med. Chem., 36 (1993) 2929-2937.
20. Burt, C., *Molecular similarity calculations for the rational design of bioactive molecules*, In Vinter, J.G. and Gardner, M. (Eds.) *Molecular modelling and drug design*. Macmillan, London, 1994, pp. 305-332.
21. Johnson, M.A., Maggiora, G.M., Lajiness, M.S., Moon, J.B., Petke, J.D. and Rohrer, D.C., *Rational use of chemical and sequence databases*. In van de Waterbeemd, H. (Ed.) *Advanced computer-assisted techniques in drug discovery*, VCH, Weinheim, 1995, pp. 89-110.
22. Wild, D.J. and Willett, P., *Similarity searching in files of three-dimensional chemical structures: Alignment of molecular electrostatic potentials with a genetic algorithm*, J. Chem. Inf. Comput. Sci., 36 (1996) 159-167.

23. Thorner, D.A., Wild, D.J., Willett, P. and Wright, P.M., *Similarity searching in files of three-dimensional chemical structures: Flexible field-based searching of molecular electrostatic potentials*, J. Chem. Inf. Comput. Sci., 36 (1996) 900–908.
24. Thorner, D.A., Willett, P., Wright, P.M. and Taylor, R., *Similarity searching in files of three-dimensional chemical structures: Representation and searching of molecular electrostatic potentials using field-graphs*, J. Comput.-Aid. Mol. Design, 11 (1997) 163–174.
25. Bures, M.G., Martin, Y.C. and Willett, P., *Searching techniques for databases of three-dimensional chemical structures*, Topics Stereochem., 21 (1994) 467–511.
26. Ho, C.M.W. and Marshall, G.R., *FOUNDATION: A program to retrieve all possible structures containing a user-defined number of matching query elements from three-dimensional databases*, J. Comput.-Aid. Mol. Design, 7 (1993) 3–22.
27. Moon, J.B. and Howe, W.J., *3D database searching and de novo construction models in molecular design*, Tetrahedron comput. Methodol., 3 (1990) 697–711.
28. Carbo, R., Leyda, L. and Arnau, M., *How similar is a molecular to another? An electron density measure of similarity between two molecular structures*, Int. J. Quant. Chem., 17 (1980) 1185–1189.
29. Bron, C. and Kerbosch, J., *Algorithm 457: Finding all cliques of an undirected graph*, Comm. ACM, 16(1973)575–577.
30. Brint, A.T. and Willett, P., *Algorithms for the identification of three-dimensional maximal common substructures*, J. Chem. Inf. Comput. Sci., 27 (1987) 152–158.
31. CONCORD is distributed by the University of Texas at Austin and Tripos Inc., St Louis, MO, U.S.A.
32. Stewart, J.J.M., *MOPAC; a semiempirical molecular orbital program*, J. Comput.-Aided Mol. Design, 4 (1990) 1–105.
33. Goldberg, D.E., *Genetic Algorithms in Search, Optimisation and Machine Learning*. Addison-Wesley, New York, 1989.
34. Michalewicz, Z., *Genetic Algorithms + Data Structures = Evolution Programs*, 2nd Ed., Springer Verlag, New York, 1994.
35. Willett, P., *Genetic algorithms in molecular recognition and design*, Trends Biotech., 13 (1995) 516–521.
36. Clark, D.E. and Westhead, D.R., *Evolutionary algorithms in computer-aided molecular design*, J. Comput.-Aided Mol. Design, 10 (1996) 337–358.
37. Payne, A.W.R. and Glen, R.C., *Molecular recognition using a binary genetic search algorithms*, J. Mol. Graph., 11 (1993) 74–91.
38. Willett, P., *Similarity searching and clustering algorithms for processing database of two-dimensional and three-dimensional chemical structures*, In Dean, P.M. (Ed.). *Molecular similarity in drug design*, Chapman and Hall, Glasgow, 1994, pp. 110–137.
39. Johnson, M.A. and Maggiora, G.M. (Eds.), *Concepts and Applications of Molecular Similarity*. John Wiley, New York, 1990.
40. Siegal, S., *Nonparametric Statistics for the Behavioural Sciences*. McGraw-Hill Kogakusha, Tokyo, 1956.
41. Reynolds, C.A., Essex, J.W. and Richards, W.G., *Atomic charges for variable molecular conformations*, J. Am. Chem. Soc., 114 (1992) 9075–9079.
42. Clark, D.E., Willett, P. and Kenny, P.W., *Pharmacophoric pattern matching in files of three-dimensional chemical structures: Use of smoothed-bounded distance matrices for the representation and searching of conformationally-flexible molecules*, J. Mol. Graph., 10 (1992) 194–204.
43. Moock, T.E., Henry, D.R., Ozkarak, A.G. and Alamgir, M., *Conformational searching in ISIS/3D databases*, J. Chem. Inf. Comput. Sci., 34 (1991) 184–189.
44. Hurst, T., *Flexible 3D searching The directed tweak technique*, J. Chem. Inf. Comput. Sci., 34 (1994) 190–196.
45. Milne, G.W.A., Nicklaus, M.C., Driscoll, J.S. and Wang, S., *National cancer Institute Drugs Information System 3D Database*, J. Chem. Inf. Comput. Sci., 34 (1994) 1219–1224.
46. Smellie, A., Kahn, S.D. and Teig, S.L., *Analysis of conformational coverages: 1. Validation and estimation of coverage*, J. Chem. Inf. Comput. Sci., 35 (1995) 285–294.
47. Sinellie, A., Kahn, S.D. and Teig, S.L., *Analysis of conformational coverage: 2. Applications of conformational models*, J. Chem. Inf. Comput. Sci., 35 (1995) 295–304.

48. Allen, F.H., Davies, J.E., Galloy, J.J., Johnson, O., Kennard, O., Macrae, C.F., Mitchell, E.M., Mitchell, G.F., Smith, J.M. and Watson, D.G., *The development of Versions 3 and 4 of the Cambridge Structural Database System*, J. Chem. Inf. Comput. Sci., 31 (1991) 187–204.
49. Ghose, A.K., Jaeger, E.P., Kowalczyk, P.J., Peterson, M.L. and Treasurywala, A.M., *Conformational searching methods for small molecules: 1. Study of the SYBYL search method*, J. Comput. Chem., 14 (1993) 1050–1065.
50. Pepperrell, C.A. and Willett, P., *Techniques for the calculation of three-dimensional structural similarity using inter-atomic distances*, J. Comput.-Aid. Mol. Design, 5 (1991) 455–474.
51. Henry, D.R., Jurs, P.C. and Denny, W.A., *Structure-antitumour activity relationships of 9-anilino-acridines*, J. Med. Chem., 25 (1982) 899–908.
52. Stuper, A.J. and Jurs, P.C., *Structure-activity studies of barbiturates using pattern-recognition techniques*, J. Pharm. Sci., 67 (1978) 745–751.
53. Chastrette, M., Zakarya, D. and Elmouaffek, A., *Structure-odor relations of the nitrobenzene musk family*, Eur. J. Med. Chem., 21 (1996) 505–510.
54. Yuta, K. and Jurs, P.C., *Computer-assisted structure-activity studies of chemical carcinogens: Aromatic amines*, J. Med. Chem., 24 (1981) 241–251.
55. Stouch, T.R. and Jus, P.C., *Computer-aided studies of the structure-activity relationships between some steroids and their anti-inflammatory activity*, J. Med. Chem., 29 (1986) 2125–2136.
56. Rose, S.L. and Jurs, P.C., *Computer-assisted studies of structure-activity relationships of N-nitroso compounds using pattern recognition*, J. Med. Chem., 25 (1982) 769–776.
57. Kearsley, S.K., Sallamack, S., Fluder, E.M., Andose, J.D., Mosley, R.T. and Sheridan, R.P., *Chemical similarity using physicochemical property descriptors*, J. Chem. Inf. Comput. Sci., 36 (1996) 118–127.

# Explicit Calculation of 3D Molecular Similarity

Andrew C. Good<sup>a</sup> and W. Graham Richards<sup>b</sup>

<sup>a</sup> Glaxo Wellcome Medicines Research Center, Gunnels Wood Road Stevenage, Herts, SG1 2NY, U.K.

<sup>b</sup> Physical Chemistry Laboratory, South Parks Road, Oxford OX1 3QZ, U.K.

## 1. Introduction

The measurement of (dis)similarity between molecules in 3D is central to many aspects of Computer-Aided Molecular Design (CAMD). From evaluating possible molecular superpositions to applying descriptors to QSAR model construction, it is the degree of molecular similarity between molecules which is being evaluated. In general, 3D modelling techniques such as CoMFA [1] and DISCO [2] apply an implicit measurement of the similarity between structures during molecular comparison. There is a branch of CAMD, however, which attempts the explicit calculation of molecular similarity in order to elucidate SAR data. In this chapter, we summarize a number of such techniques, detailing their application, merits and pitfalls.

## 2. A Brief History of 3D Molecular Similarity Coefficients

A multitude of indices have been suggested for the calculation of explicit molecular similarity, using a variety of 3D molecular descriptors. These indices can be separated into two basic classes. The first is best described as the set of ‘cumulative’ similarity indices. For these formulae, molecular similarity is evaluated via the accumulation of overlap or difference values over all descriptor space. The second can be characterized as ‘discrete’ in nature. Such indices evaluate similarity at discrete points in descriptor space, with overall molecular similarity determined from the average of these point values.

### 2.1. Cumulative formulae

Hopfinger [3,4] proposed a number of indices including equations for measuring the common volume of steric overlap between molecular pairs. One example is shown here:

$$Vo_{ij} = \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} (V_{A_i} \cap V_{B_j}) \quad (1)$$

Each atom is described by a sphere of van der Waals (vdW) radius. The atomic overlap volume for all intermolecular atom-pair combinations is calculated. Summation of these molecular overlap data provides a measure of shape similarity. Hermann and Herron [5] and Masek [6] et al. have both proposed variants of this approach.

Perhaps the most widely applied ‘cumulative’ molecular similarity index used in 3D SAR studies was pioneered by Carbo et al. [7]:

$$C_{AB} = \frac{\int P_A P_B dV}{\left(\int P_A^2 dV\right)^{1/2} \left(\int P_B^2 dV\right)^{1/2}} \quad (2)$$

Molecular similarity  $C_{AB}$  is determined from the 3D descriptor properties  $P_A$  and  $P_B$  of molecules  $A$  and  $B$  being compared over all space. As in Eq. 1, the numerator term measures property overlap, while the denominator normalizes the similarity result. Its structure is identical in construction to the cosine coefficient which is also utilized in 2D database similarity searching [8,9]. The formula known as the Carbo index measures the deviation of two molecular properties from proportionality and is, thus, sensitive to the shape of property distribution rather than to its magnitude. This is highlighted by the fact that when the measured properties of two molecules correlate ( $P_A = nP_B$ ), the similarity index tends towards unity. Variants of this index have been proposed by Hodgkin and Richards, and also by Petke, in an attempt to increase the sensitivity of the formula to property magnitude [10–12].

The indices of Carbo and Hodgkin have been used extensively in molecular similarity investigations. As originally applied, the Carbo index was utilized to measure the similarity of quantum-mechanically derived electron density between molecules [7,13–15]. Other quantum-mechanically derived measures of electron distribution have also been used in conjunction with variants of the index [16–18]. Electron density has the quality of being an analytical property firmly grounded in quantum chemistry. Its calculation is, in general, CPU intensive, however, and it is not a particularly discriminating property. For 3D QSAR calculations, the Molecular Electrostatic Potential (MEP) and molecular shape derived measures tend to be preferred, due to their ease of calculation and improved discrimination. To this end, similarity evaluations using the variants of the Carbo index have been extended to include the comparison of such MEP [10,19–23] and shape [24–29] descriptors.

A close relative of the Carbo index, the Tanimoto coefficient, which is widely used in 2D similarity calculations [8], has also been applied in modified form (Eq. 3) to the measurement of shape similarity [29]:

$$T_{AB} = \frac{\int P_A P_B dV}{\int w_1 (P_A - P_B) dV + \int w_2 (P_B - P_A) dV + \int P_A P_B dV} \quad (3)$$

$P_A P_B$  equals the volume overlap between structure  $A$  and shape query  $B$  ( $B$  is the shape derived from a single molecule, group or molecules or active site),  $P_A - P_B$  is the structure volume not in the query,  $P_B - P_A$  the query volume not in the structure and  $w_1$  and  $w_2$  are user-defined weights. A more distant relative of the Carbo index, the Spearman rank-correlation coefficient, has also been used to calculate 3D molecular similarity:

$$S_{AB} = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n^3 - n} \quad (4)$$

This index has been used for molecular similarity calculation in a number of studies involving MEP and accessible surface [30–35]. Rather than applying the actual numeri-



cal values of the properties being evaluated, the formula calculates the Pearson correlation coefficient of the relative ranks of the data. The numerator term  $d_i$  is the difference in the property ranking at point  $i$  of two structures, and  $n$  is the total number of points over which the property is measured.

Another widely used family of indices analyzes molecular dissimilarity through the measurement of the Euclidean distance between molecular properties. An example of this is shown in Eq. 5, which calculates the root mean squared difference between properties:

$$RMSD_{AB} = \left[ \sum_{i=1, n} \left\{ (P_{A,i} - P_{B,i})^2 \right\} / n \right]^{1/2} \quad (5)$$

Perry and van Geerstein [36] used Eq. 5 for optimizing molecular surface shape similarity. Other variants of this group of indices include the measurement of sum squared errors and mean squared difference [32,37–39].

## 2.2. Discrete formulae

More recently, indices which measure similarity at discrete points in space have been employed to allow greater control over molecular similarity calculations [12,22,40–42]. The application of such formulae permits both graphical and statistical analysis to be executed on the resultant similarity grid. For such discrete indices, overall molecular complementarity is obtained by summing the similarities over all points in space and then dividing the result by the total number of points present. The resulting value is the average similarity over measured property space. Reynolds et al. [40] proposed the first discrete similarity formula for 3D calculations, in the form of the ‘linear’ similarity index:

$$I_{AB} = 1 - \frac{|P_A - P_B|}{\max(|P_A|, |P_B|)} \quad (6)$$

$\max(|P_A|, |P_B|)$  equals the larger molecular property magnitude between  $P_A$  and  $P_B$  at the grid-point where the similarity is being calculated. An exponential variant of this index was also proposed by Good [22], while Petke [12] created two further discrete indices, including one based on the Hodgkin index. These discrete indices have been applied to the measurement of MEP and Molecular Electrostatic Field (MEF) similarity [12,22],

Recently, Klebe et al. [41] proposed a discrete formula for the evaluation of molecular similarity in 3D QSAR calculations. This methodology is interesting, in that rather than calculating a pair-wise molecular similarity between molecules, similarity is determined between a given molecule and a probe atom:

$$A_j^{q_k} = \sum_{i=1}^n W_{probe,k} W_{ik} e^{-\alpha r_{iq}^2} \quad (7)$$

where the similarity  $A$  for physico-chemical property  $k$  at grid-point  $q$  is calculated through Summation across all  $n$  atoms of molecule  $j$ ,  $w_{ik}$  is the physico-chemical property value of atom  $i$ ,  $w_{probe,k}$  the probe atom property (+1 atomic charge, 1 Å radius and +1 hydrophobicity),  $\sigma$  the attenuation factor and  $r_{iq}$  the mutual distance between probe atom at grid-point  $q$  and atom  $i$  of molecule  $j$ . The resulting molecular similarity grids have been applied successfully to 3D QSAR calculations [43].

### 2.3. Comparing coefficients

Choosing the equation which best suits one's need for 3D molecular similarity calculations is no simple task. While many studies have been undertaken to compare the relative merits of similarity indices in 2D database searches (detailed elsewhere, see reference [8]), equivalent comparisons involving 3D descriptors are less common. Petke [12] and Good [22] explored issues of sensitivity in some detail, comparing similarity index values for MEPs and MEFs across a number of datasets. Both these studies suggest that discrete indices tend to be more sensitive than their cumulative counterparts. The test systems used (e.g. the comparison of index behavior at a single discrete point in space) were, however, rather artificial in nature. Calculations involving actual ligand series yielded more complicated relationships [22]. Indeed, extensive QSAR studies by Good et al. [43], involving multiple datasets and indices, suggest that despite its perceived lack of sensitivity, the Carbo index produces results comparable to many of its more complex contemporaries.

These results highlight the problems in attempting to define which is the best index. Each coefficient has its own strengths and weaknesses. The Spearman rank-correlation coefficient (Eq. 3) has been shown to perform poorly in QSAR studies [43]. Its structure is such that it is scale invariant, however, making it potentially useful with systems containing a net charge, as it prevents highly charged regions dominating all other system features. Discrete indices, such as Eq. 6, allow improved control and graphical/statistical analysis of similarity evaluations [22], while formulae such as the Carbo index allow rapid analytical evaluation using Gaussian functions [23,27]. Many of the indices have been applied with success to a variety of problems. The choice of index and its method of evaluation should, therefore, be made with careful consideration to the molecular design problem at hand. The information provided in section 3 should make the nature of such considerations clearer. More extensive comparisons of similarity coefficients in general, and those applied to 3D problems in particular, are detailed elsewhere [44–46].

## 3. 3D Descriptors and their Application in Explicit Molecular Similarity Calculations

In many cases, it is not so much the similarity index, but the methodology applied to property calculation, that determines the success of a molecular similarity evaluation. A host of 3D descriptors have been used in conjunction with the molecular similarity coefficients described above. Many of the measures have been selected or tailored to

meet particular applications, from *ab initio* quantum mechanical calculations for thorough molecular comparison to specially devised atomic triplet shape descriptors for rapid database searches. A number of these approaches are detailed here, together with their associated applications.

### 3.1. Calculations based on explicit molecular shape and MEP evaluation

A variety of techniques have been exploited in comparing the molecular fields between molecules, including numerical evaluation on rectilinear grids, analytical comparisons employing Gaussian function approximations, gnomonic projection and molecular volume measurement. These and other approaches are detailed below.

#### 3.1.1. Descriptor calculations in grid space

In CAMD, continuous molecular properties are often approximated through the application of rectilinear grids, and this is no less the case in molecular similarity calculations. When originally applied to the problem of measuring molecular similarity, many indices were (and still are) used to measure molecular similarity based on quantum-mechanically derived properties [7, 13–18,47–48]. CPU hungry calculations and lack of results discrimination led Hodgkin and Richards [10,11] to apply similarity indices in numerical evaluations of MEP and MEF (Fig. 1). The increased speed of MEP and MEF similarity calculations allows systems of greater biological significance to be chosen for study. Burt et al. [19] further developed the methodology by undertaking MEP and MEF similarity calculations for a series of nitromethylene insecticides. Multiple grid increments, extents and charge schemes were employed in order to determine the optimum compromise between speed and accuracy of calculation. Burt and Richards [20] then added the ability to include flexible fitting during similarity optimization. They found that by allowing torsional flexibility, molecular superpositions with significantly higher similarity could be realized. In order to lower the chance of

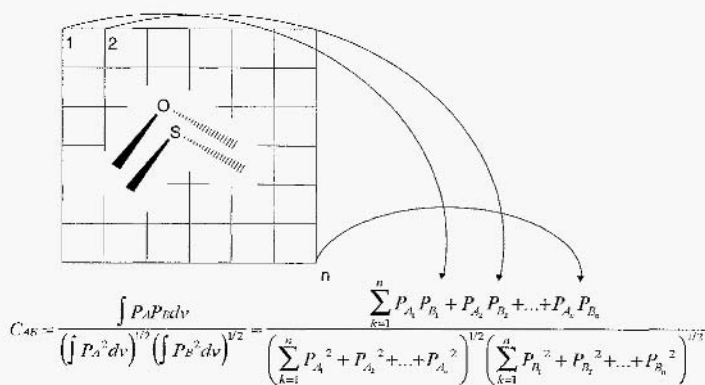


Fig. 1. Numerical approximation for the Carbo index using molecular properties defined on a rectilinear grid. adapted from reference [43].

obtaining energetically unfavorable conformations during such calculations, the similarity index calculation was weighted using a Boltzmann factor to penalize significant increases in internal energy:

$$C_{AB} = \left( \frac{\int P_A P_B dv}{\left( \int P_A^2 dv \right)^{1/2} \left( \int P_B^2 dv \right)^{1/2}} \right) e^{-c\Delta E / RT} \quad (8)$$

where  $c$  is the weighting factor,  $\Delta E$  the energy of rotated conformation minus energy of initial conformation,  $R$  the gas constant and  $T$  the temperature. All the software functions described above have been integrated into a single software package known as ASP [49]. Similar procedures have also been applied to measurement of hydrophobic similarity, essentially replacing point charges with atom-based hydrophobic potentials [50].

Comparable functionality to that described above has also been developed by Manaut et al. [34–35] in the form of the MEPCOMP and MEPSIM software. These programs employ the Spearman rank-correlation coefficient (Eq. 4) to determine and optimize MEP similarity from *ab initio* derived MEP grids, as well as those calculated from point charges. They also use a novel method for determining the grid-points to be included in the optimization, setting inner and outer exclusion volumes proportional to the vdW radii of the constituent atoms. The internal shell is usually set to a small value, so that the high MEP values can attempt to simulate steric interactions.

One of the problems of a rectilinear grid with uniform point density is that the system gives equal importance to points close to and far from the molecules being compared. Richard [21] proposed an alternative in the form of a molecular atom-centered radial (MACRA) grid. As its name suggests, a MACRA grid produces grid-points emanating from the atom centers of a molecule. A template sphere employing a fixed or vdW radius with approximately uniformly distributed surface points is centered on each atom. This forms the first layer of the grid. The second layer is created by scaling the sphere radius through the addition of the average distance between grid-points on the lower layer. Points clashing with lower shells of other atoms are removed. The primary advantage of such an approach is that, as the layers are constructed, the resulting radial grid requires around 2% of the points of a 1.0 Å increment rectilinear grid to encompass a ligand of drug-like proportions. The technique contains certain drawbacks however, not the least of which is the inability of such a grid to deal adequately with concave molecular surfaces.

The calculation of molecular shape similarity has also been undertaken employing rectilinear grids. Meyer and Richards [24] utilized a point counting function version of the Carbo index through modifications of the ASP program. Every grid-point is tested to see whether it falls inside the vdW surface of each molecule. The results are then applied to the following index:

$$C_{AB} = \frac{B}{(T_A T_B)^{1/2}} \quad (9)$$

$B$  is the number of grid-points falling inside both molecules, while  $T_A$  and  $T_B$  are the total number of grid-points falling inside each individual molecule. The use of

extremely fine grids (0.2 Å separation) in conjunction with this technique make for prolonged calculation times, restricting its utility to SAR calculations rather than molecular alignment. More recently, Hahn [29] has developed a more rapid grid-based method for shape evaluation. The technique uses rapid volume and principal axes indices to pre-screen for molecules similar in shape to a given query. Molecules passing this screen are aligned by their principal axes (and symmetrically equivalent superpositions) and their shape similarity evaluated through volumetric comparison. Additional optimization, flexible fitting and electrostatic similarity comparison functions are also included. The use of such rapid pre-screens and coarser grids allows this approach to be used for full-scale 3D database searching.

### 3.1.2. Gaussian function evaluations

While grid-based similarity evaluation techniques are common, their numerical foundations impart inherent drawbacks. The largest of these problems is that, to gain computation speed, the grids employed are normally coarse, with the consequence that resulting evaluations of spatial properties are somewhat rough. In particular, the similarity optimization through the modification of relative molecular position is coarse and crude. It is, for example, very difficult for a grid-based similarity optimization to superimpose a molecule on top of itself, since the program tends to converge prematurely at some discrete point.

The mathematical structures of certain similarity coefficients are such that analytical Gaussian functions may be exploited in their evaluation — for example, MEP calculations employing the standard point charge approach, where the charges ( $q_i$ ) assigned to each atom ( $i$ ) create an electrostatic potential at a point  $r$  for a molecule of  $n$  atoms according to the following equation:

$$P_r = \sum_{i=1}^n \frac{q_i}{(r - R_i)} \quad (10)$$

where  $R_i$  is the nuclear coordinate position of atom  $i$ .

The inverse distance dependence term of Eq. 10 was substituted with a Gaussian function approximation by Good et al. [23] (Fig. 2):

$$P_r = \sum_{i=1}^k q_i (\gamma_1 e^{-\alpha_1 r^2} + \gamma_2 e^{-\alpha_2 r^2} + \gamma_k e^{-\alpha_k r^2}) \quad (11)$$

When this potential function is substituted into the Carbo index (Eq. 2), the resulting integral terms expand into a series of two-center Gaussian overlap integrals. A two-center Gaussian overlap integral has a simple solution based on the exponent values and distances between atom centers shown in Eq. 12 [51]:

$$\int e^{-\alpha_1(r-R_i)^2} e^{-\alpha_2(r-R_j)^2} d\mathbf{v} = \left( \frac{\pi}{\alpha_1 + \alpha_2} \right)^{3/2} \exp\left( \frac{\alpha_1 \alpha_2}{\alpha_1 + \alpha_2} |R_i - R_j|^2 \right) \quad (12)$$

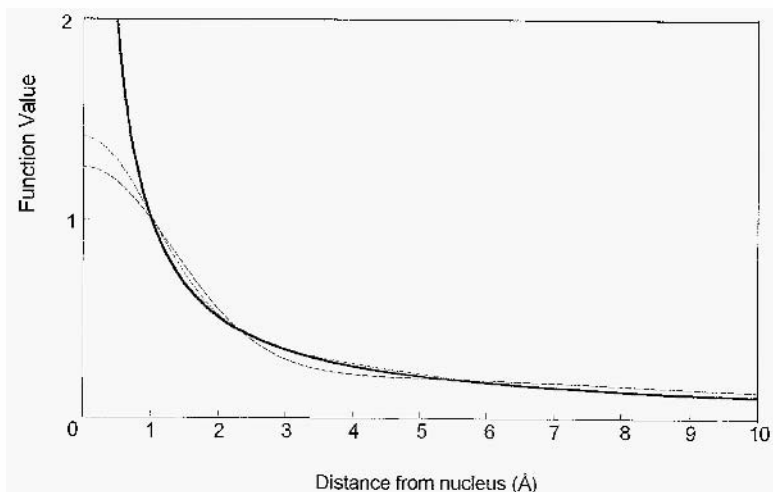


Fig. 2. Two and three Gaussian function approximations to the  $1/r$  distance dependence curve. Adapted from reference [22]: —  $1/r$ ; ..... 2 Gaussians; —, —, 3 Gaussians.

The similarity calculation can thus be broken down into a succession of readily calculable exponent terms. As a result of this, it is possible to evaluate MEP similarity rapidly and analytically, and as no singularity exists when the potential approaches an atomic nucleus (Fig. 2), the calculation need not be restricted to regions outside the atomic vdW radii.

These Gaussian functions were incorporated into the ASP program by Good et al. [23], and MEP similarity optimization calculations were undertaken on identical molecules in different spatial orientations. The results confirmed that the analytical Gaussian functions produced similarity values comparable with grid-based calculations, with a two orders of magnitude increase in evaluation speed. They were also found to be capable of superimposing two identical molecules back on top of each other during similarity optimizations, calculations for which equivalent numerical evaluations always converged prematurely. Further speed increases have been obtained by exploiting the analytical nature of the Gaussian functions to replace simplex optimization routines with gradient-based methods [52]. In this way, the time required for similarity optimization can be reduced by up to an order of magnitude. Willett and co-workers [53–54] have exploited the speed of Gaussian-based evaluations, coupled with efficient methods for undertaking molecular alignment (genetic algorithms, MEP field graphs and bit climbers), to allow the rapid comparison of MEPs between molecular pairs (< 1 CPU second on a good workstation). Using such an approach, it has been possible to compare whole databases of molecules against given lead structures, searching for similar MEP distributions. This is of particular interest as the molecules found to be similar often show little structural resemblance to each other, unlike those located using traditional substructure search techniques.

The use of Gaussian functions need not be limited to MEP calculations. In principle, any property which can be approximated to a set of Gaussians could be compared in this way. Hodgkin and Richards [14] parameterized Gaussian in functions centered at electron density maxima, so to reproduce the *ab initio* electron density distributions of molecular fragments. The technique was found to work well for the small fragments investigated, but was not pursued because of the difficulty of breaking a larger drug molecule down into a suitable set of fragments. More recently, Good and Richards [27] proposed a more elementary approach, with electron density simply approximated to the square of the STO-3G atomic orbital wave functions [55]. The results produced using this technique were found more than two orders of magnitude faster than equivalent calculations using the numerical shape analysis of Meyer and Richards [24], with little loss in calculation accuracy. Grant and Pickup [56–57] applied Gaussian approximations to the hard sphere model of molecular shape, further improving on the approach by exploiting the fact that the product of multiple Gaussians is a single Gaussian centered at a coalescence point. This allows molecular shapes based on Gaussians to be collapsed into increasingly simple elliptical representations, further increasing calculation speed. Gaussian functions have also been applied by Chapman [58] in the field of combinatorial library profiling, using the approach to quantify steric dissimilarity between potential reagents.

It is clear from the above studies that calculations employing Gaussian functions offer a versatile technique for molecular similarity evaluation. Their analytical nature and ease of calculation make for robust and rapid similarity optimizations, while their empirical nature appears to have little consequence on their effectiveness. (Good et al. [43] showed that when used as parameters in QSAR calculations, analytically derived MEP similarity matrices were more predictive than their numerical counterparts.)

### 3.1.3. Gnomonic projection

The technique of gnomonic projection extends the molecular properties of a structure onto the surface of a sphere. The sphere is approximated by a tessellated icosahedron, or an icosahedron and dodecahedron oriented such that the vertices of the dodecahedron lie on the vectors from the center of the sphere through the mid-points of the icosahedral faces [26,36,59]. Gnomonic projection displays many useful properties, including the dramatic simplification of how to map two irregular surfaces to each other. Calculation times are rapid, especially when symmetry elements inherent to the system are exploited to reorient the projections. As a result of these features, gnomonic projection techniques have been widely applied to quantitative similarity measurement. Applications include exhaustive similarity comparisons of MEP and surface shape between molecular pairs [31], full-scale shape similarity database searching [26,36] and interactive graphical and numerical similarity analysis [38-39,60]. The primary drawback of gnomonic projections is that the technique is essentially restricted to comparisons in rotational space, rendering the final results of any similarity calculation dependent on the chosen centers of projection for each molecule.

#### 3.1.4. Volume measurements

The measurement of molecular shape through the use of volume overlap calculations was one of the first ways in which 3D quantitative similarity measures were made. Hopfinger [3] first employed such measures as QSAR parameters through the overlap calculation between all pairs of atoms in the molecules being compared (Eq. 1), using the hard sphere approximation for each atom. Hermann and Herron [5] developed the OVID program, which optimizes molecular alignment by maximizing the overlap volume of a prespecified subset of ligand atoms. OVID is fast but requires the user input to define the atoms considered important in the binding mode, essentially measuring fit quality to a postulated pharmacophore. Masek et al. [6] employ the more accurate analytical volume calculations presented by Connolly [61] to determine optimum molecular shape comparisons. It is also possible to weight for overlap between atoms with matching chemistry. These searches are more exhaustive and, consequently, have a higher CPU requirement.

#### 3.1.5. Dot Surface evaluation

As well as the OVID program, Hermann and Herron [5] developed an alternative, SUPER, for measuring surface shape complementarity. This software is used to undertake more exhaustive molecular alignment calculations. The program basically works by comparing molecular dot surfaces of overlaid molecules with surface points considered as matching when they lie within a predefined distance of one another. MEP data can also be used to weight for matches with matching electrostatics. Badel et al. [37] use what they term bi-dimensional surface profiles to measure surface complementarity. A 2D slice is made through the molecule Connolly surface [62], with the angular profile of the resulting surface contour determined by moving along the cross-section, calculating the angles formed by three successive surface points. Sections of this resulting profile are then compared in order to measure shape complementarity. The technique attempts to provide a condensed 2D measure of shape, but the results are ultimately dependent on the orientation of the molecule relative to the cross-sections taken. Masek et al. [63] have developed a method for molecular surface similarity evaluation based on the overlap of regions between the vdW and solvent accessible surface. The technique was pioneered because it is not possible to use volume overlap comparisons to measure the similarity of ligands with vastly different size (e.g. comparison of small molecule with exposed loop of protein ligand). Overlap is calculated using the analytical volume overlap calculations of Connolly [62]. Such evaluations are extremely CPU intensive, however, so to speed up the calculations, Perkins et al. [63] developed an alternative where molecular surface representations are precomputed and stored on grids for comparison. Such evaluations are thus rapid, but this does lead to the problem that molecular flexibility cannot be dealt with explicitly, as a new grid would need to be calculated in response to any change in conformation.

#### 3.1.6. QSAR methodology exploiting field-based similarity data

A number of different approaches have been taken in the application of molecular similarity calculations to the creation of 3D QSAR models. The easiest way to correlate molecular similarity results with biological data is via simple regression analysis. Many



examples of such studies can be found in the literature. Seri-Levy and Richards [65–66] exploited molecular similarity data to construct QSARs for ligand enantiomer eudismic ratios. Each enantiomeric pair was superimposed through a least-squares fit to maximize the overlap of their stereogenic centers, and the ASP program was used to measure shape and MEP dissimilarity (defined as 1-similarity). The eudismic ratios were then correlated against these dissimilarities. Similar calculations have also been undertaken to quantify the complementarity of the peptide bond to a series of isosteres [67]. Burt et al. [19] correlated activity for a set of nitromethylene insecticides with their MEP similarity to the most active molecule in the series. A similar approach has also been employed by Montanari et al. [68]. Hopfinger et al. [3,69-73] have successfully applied their molecular shape calculations in the derivation of 3D QSAR models across a wide variety of systems. Correlations were determined through the calculation of molecular similarity, using each dataset molecule in turn, or even the shape of an ensemble of aligned active molecules. as the template for regression analyses.

It is clear from the above studies that molecular similarity-based regression equations can produce good QSARs. Nevertheless, it is unlikely that a single molecule will contain all or most of the structural information inherent to a given ligand dataset. In an effort to capture this information, full  $N \times N$  (similarity of each ligand in dataset calculated against all other ligands) molecular similarity results matrices originally employed by Rum and Hernden in QSAR calculations [74] were extended to incorporate 3D property data [43,75]. This would make all the SAR data embedded within the molecular similarity values available for model generation. The information contained within such matrices is similar to that present within a CoMFA data matrix. the primary difference being that. while CoMFA will describe a region around a group of molecules using a large number of grid-points, the similarity matrix will attempt to describe the same region using just a few numbers. Such matrices thus provide an efficient 3D QSAR descriptor set, and the predictivity of resulting 3D QSARs has been found to compare well with CoMFA calculations undertaken on identical datasets [43,76]. The primary problem of this methodology is that, while CoMFA is able to display the coefficients of its QSAR equations as maps of favorable and unfavorable structural interactions. no such method exists for extracting chemical meaning from similarity matrix equations. To address this, Klebe et al. [41] have proposed an approach to 3D QSAR which applies a discrete rather than cumulative 3D molecular similarity index (Eq. 7). Similarities are evaluated between molecules in the QSAR dataset and a probe atom. The resulting molecular similarity grids are then analyzed in a CoMFA-like manner using PLS. While, as in CoMFA, this leads to much larger data matrices than the  $N \times N$  cumulative similarity matrix approach, the resulting QSAR models can be analyzed graphically. These graphical QSAR model depictions were shown to be of superior quality when compared to CoMFA graphical representations of the same datasets [41], while the QSAR models showed similar predictivity.

### *3.2. Molecular similarity evaluations based on atom distribution*

While a number of the techniques described above have utilized clever techniques to speed up MEP and shape similarity comparisons, it is generally the case that such

calculations are somewhat time-consuming. As a consequence, they do not lend themselves particularly well to database searching and library profiling protocols. To lower the CPU requirement of 3D molecular similarity evaluation, a variety of alternative methods has been devised which attempts to determine similarity using simplified structural descriptors based on 3D atom distributions. Such techniques, while less accurate, can be calculated rapidly and are, therefore, well suited to the large dataset comparisons required in database searching and library profiling. Many of these techniques are described below.

### 3.2.1. *Shape measures*

A host of approaches have been developed to allow shape similarity measurements based on matching atom distributions. Nilakantan et al. [28] have developed a 3D database search system, based on the distribution of all atom triplet distance combinations found within a molecule. Simplified binary signatures of template and database molecule triplets are compared during the first stage of a database search, with the triplets of those molecules deemed similar enough regenerated on the fly for final comparison. The number of triplet matches found is used as the molecular shape property to quantify similarity. Good et al. [77] extended the use of simple ligand triplet descriptors to encompass molecular surface comparisons, and altered the descriptors to allow storage on hard disk, thus facilitating faster searches. Bemis and Kuntz [78] applied a simplified version of triplet matching to permit rapid 3D database clustering. The perimeter of each triplet in a molecule is measured, and the resultant distance is used to augment the appropriate bin of a molecular shape histogram. These histograms are then compared to quantify shape similarity. A variety of different triangle geometric properties has also been exploited by Fisanick et al. [79] to aid similarity searching of molecules found in the Chemical Abstracts Service Database. Another interesting method introduced by Norel et al. [80] employs, again, triangle descriptors to calculate molecular shape. For each molecule in a database, every pair of atoms within a defined distance range of each other are extracted, and triangles are constructed by adding a third vertex in the form of a molecular surface point. Triangle data for all the atom pair-surface triangle descriptions of a molecular database are stored in a hash table. This table is then used rapidly to screen for potentially complementary ligand-receptor interactions, through complementarity calculations with equivalent receptor triangle data.

### 3.2.2. *Measures incorporating pharmacophoric information*

It is possible to extend the descriptors described in section 3.2.1. through the incorporation of chemical information with the atom distribution data. Moon and Howe [25] have developed a 3D database search system where queries are built up from single or multiple overlapped active ligands. Database molecules are fitted to the query in multiple orientations using the clique detection algorithms of Kuntz et al. [81]. Atoms match when their centers are found within a predetermined distance of each other. The score applied then depends on the degree of chemical complementarity between matched atoms. Query atoms are defined as required or optional: database molecules must have an atom match with required atoms in order to be retrieved as a hit, while optional

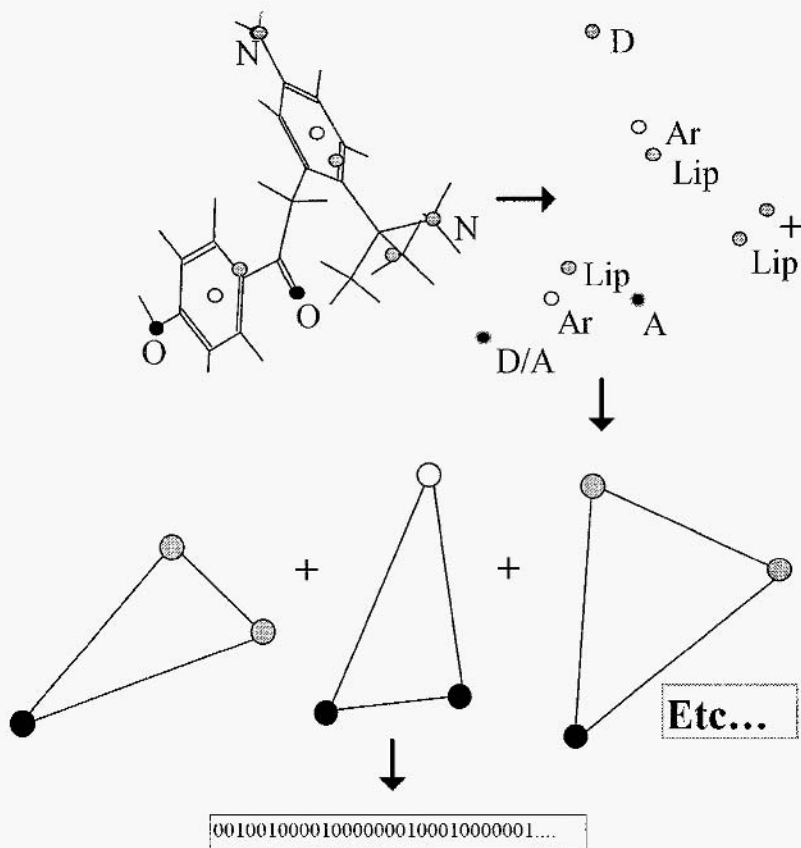


Fig. 3. Pharmacophore centre abstraction and triplet description creation paradigm for pharmacophore based similarity measures (references [79,81-83]). Key to pharmacophore center types: A= hydrogen-bond acceptor; D= hydrogen-bond donor; Ar= aromatic ring centroid; Lip= lipophilic centroid (center of group of atoms with -0 charge); += charged positive. Adapted from reference [82].

atoms matches are used only to augment the similarity score after matching has occurred.

All of the techniques described so far in section 3.2 are based on descriptors derived from explicit molecular conformations. Thus, while fast, they do not account for the inherent flexibility which might be present in a given system. The primary problem with attempting to incorporate molecular flexibility in conjunction with triplet similarity measures is the  $n^3$  dependence ( $n$  = number of centers per molecule) of the descriptor keying times. To overcome this problem, Good and Kuntz [82] developed a triplet-based measure grounded on a simplified description of each molecule. Rather than using all the heavy atoms in a structure, triplet calculation is restricted to pharmacophoric centers (e.g. donor, acceptor, lipophilic; see Fig. 3). In this way, the CPU requirement for descriptor calculation is dramatically reduced, allowing the

incorporation of data from multiple conformations. Pickett et al. [83] extended this idea using their PDQ (pharmacophore derived query) approach, where triplet descriptors are derived from multiple 3D database searches against a battery of predetermined theoretical pharmacophores. Such a technique has the advantage of building up a full profile of the number and type of pharmacophores present in a given dataset. This can be useful in library profiling where dataset diversity measures are extremely important. Quality of fit to any given pharmacophore can also be determined easily. The major drawback with the PDQ approach is its keying time which is orders of magnitude slower than internal pharmacophore measurement on the fly [82,84]. This type of methodology has been further advanced by the program *Chem-Diverse* [84–85], which exploits pharmacophore triplet information in combinatorial library profiling. The *Chem-Diverse* protocol for molecular similarity calculation is based on trying to obtain the maximum coverage of pharmacophore space by potential combinatorial chemistry products. The overlap of pharmacophore triplets between entire datasets is compared, allowing the assessment of their similarity and hence diversity. Such methodology is currently an area of much interest [86].

#### 4. Conclusion

This chapter has summarized many of the techniques available for quantifying explicit 3D molecular similarity. A multitude of molecular properties, indices and protocols have been presented, from the detailed comparison of MEP and shape for QSAR construction, through to the rapid analysis of pharmacophore triplet descriptors for diversity analysis. The sheer variety of such approaches permits their application in many key areas of molecular design. Such flexibility makes quantitative molecular similarity evaluation an important tool in the hands of the computational chemist. With the ever increasing demand for the quantification of diversity in combinatorial library profiling, this importance is likely to grow.

#### References

1. Cramer, R.D. III, DePriest, S.A., Patterson DE. and Hecht, P., *The developing practice of comparative Molecular Field Analysis*, In 3D QSAR in drug design. Kubinyi, H. (Ed.) ESCOM, Leiden, 1993, pp. 443–485.
2. Martin, Y.C., Bures, M.G., Danaher, E.A., DeLazzer, J., Lico, I. and Pavlik, P.A., *A few new approaches to pharmacophore mapping and its application to dopaminergic and benzodiazepine agonists*, J. Comput.-Aided Mol. Design, 7 (1993) 83–102.
3. Hopfinger, A.J., *A QSAR investigation of DHFR inhibition by Bakers Triazines based upon molecular shape analysis*, J. Am. Chem. Soc., 102 (1980) 7196–7206.
4. Hopfinger, A.J., *Theory and analysis of molecular potential energy fields in molecular shape analysis: A QSAR study of 2,4-diamino-5-benzylpyrimidines as DHFR inhibitors*, J. Med. Chem., 26 (1983) 990–996.
5. Hermann, R.R. and Herron, D.K., *OVID and SUPER: Two overlappograms for drug design*, J. Comput.-Aided Mol. Design, 5 (1991) 511–524.
6. Masek, B. B., Merchant, A. and Matthew, J. B., *Molecular shape comparisons of angiotensin II receptor antagonist*, J. Med. Chem., 36 (1993) 1230–1238.
7. Carbo, R., Leyda, L. and Arnau, M., *An electrostatic density measure of the similarity between two compounds*, Int. J. Quantum Chem., 17 (1980) 1185–1189.

8. Downs, G.M. and Willett, P., *Similarity searching in databases of chemical structures*. In Lipkowitz, K.B. and Boyd, D.B. (Eds.) *Reviews in computational chemistry*, Vol. 7, VCH, New York, 1995, pp. 1–66.
9. Holland, J.D., Ranade, S.S. and Willett, P., *A fast algorithm for selecting sets of dissimilar molecules from large chemical databases*, *Quant. Struct.-Act. Relat.*, 14 (1995) 501–506.
10. Hodgkin, E.E. and Richards, W.G., *Molecular similarity based on electrostatic potentials and electric field*, *Inf. J. Quantum Chem. Quantum Biol. Symp.*, 14 (1987) 105–110.
11. Hodgkin, E.E. and Richards, W.G., *Molecular similarity*, *Chem. Br.*, (1988) 1141–1144.
12. Petke, J.D., *Cumulative and discrete similarity analysis of electrostatic potentials and fields*, *J. Comp. Chem.*, 14 (1993) 928–933.
13. Bowen-Jenkins, P.E. and Richards, W.G., *Molecular similarity in terms of valence electron density*, *J. Chem. Soc. Chem. Commun.* (1986) 133–135.
14. Hodgkin, E.E. and Richards, W.G., *A semi-empirical method for calculating molecules similarity*, *J. Chem. Soc. Chem. Commun.*, (1986) 1342–1344.
15. Carbo, R. and Domingo, L., *LCAO-MO similarity measures and taxonomy*, *Int. J. Quantum Chem.*, 32 (1987) 517–545.
16. Amovilli, C. and McWeeny, R., *Shape and similarity: Two aspects of molecular recognition*, *J. Mol. Struct.*, 227 (1991) 1–9.
17. Cioslowski, J. and Fleischmann, E.D., *Assessing molecular similarity from results of ab initio electronic structure calculations*, *J. Am. Chem. Soc.*, 113 (1991) 64–67.
18. Cooper, D.L. and Allen, N.L., *A novel approach to molecular similarity*, *J. Comput.-Aided Mol. Design*, 3 (1991) 253–259.
19. Burt, C., Huxley, P. and Richards, W.G., *The application of molecular similarity calculations*, *J. Comp. Chem.*, 11 (1990) 1139–1146.
20. Burt, C. and Richards, W.G., *Molecular similarity: The introduction of flexible fitting*, *J. Comput.-Aided Mol. Design* 4 (1990) 231–238.
21. Richard, A.M., *Quantitative comparison of MEPs for structure activity studies*, *J. Comp. Chem.*, 12 (1991) 959–969.
22. Good, A.C., *The calculation of molecular similarity: Alternative formulas, data manipulation and graphical display*, *Mol. Graph.*, 10 (1992) 144–151
23. Good, A.C., Hodgkin, E.E. and Richards, W.G., *The utilisation of Gaussian functions for the rapid evaluation of molecular similarity*, *J. Chem. Inf. Comput. Sci.*, 32 (1992) 188–191.
24. Meyer, A.M. and Richards, W.G., *Similarity of molecular shape*, *J. Comput.-Aided Mol. Design*, 5 (1991) 426–439.
25. Moon, J.B. and Howe, W.J., *3D database searching and de novo design construction methods in molecular design*, *Tetrahedron Comput. Methodol.*, 3 (1992) 697–711.
26. van Geerestein, V.J., Perry, N.J., Grootenhuys, P.D.J. and Haasnoot, C.A.G., *3D database searching on the basis of shape using the SPERM prototype method*, *Tetrahedron Comput. Methodol.*, 3 (1992) 595–613.
27. Good, A.C. and Richards, W.G., *Rapid evaluation of shape similarity using Gaussian functions*, *J. Chem. Inf. Comput. Sci.*, 33 (1993) 112–116.
28. Nilakantan, R., Bauman, N. and Venkataraghavan, R., *New method for rapid characterization of molecular shapes: Applications in drug design*, *J. Chem. Inf. Comput. Sci.*, 33 (1993) 79–85.
29. Hahn, M., *Three-dimensional shape-based searching of conformationally flexible compounds*, *J. Chem. Inf. Comput. Sci.*, 37 (1997) 80–86.
30. Namasivaynam, S. and Dean, P.M., *Statistical method for surface pattern matching between dissimilar molecules: Electrostatic potentials and accessible surfaces*, *J. Mol. Graph.*, 4 (1986) 46–50.
31. Chau, P.-L. and Dean, P.M., *Molecular recognition: 3D surface structure comparison by gnomonic projection*, *J. Mol. Graph.*, 5 (1987) 97–100.
32. Dean, P.M. and Chau, P.-L., *Molecular recognition Optimised searching through molecular 3-space for pattern matches on molecular surfaces*, *J. Mol. Graph.*, 5 (1987) 152–158.
33. Dean, P.M., Callow, P. and Callow, P.-L., *Molecular recognition: Blind searching for regions of strong structural match on the surface of two dissimilar molecules*, *J. Mol. Graph.*, 6 (1988) 28–34.

34. Manaut, M., Sanz, F., Jose. J. and Milesi, M., *Automatic search for maximum similarity between MEP distributions*, J. Comput.-Aided Mol. Design, 5 (1991) 371–380.
35. Sanz, F., Manaut, F., Rodriguez, J., Lozoya, E. and Lopez-de-Brinao, E., *MEPSIM: A computational package for analysis and comparison of Molecular Electrostatic Potentials*, J. Comput.-Aided Mol. Design, 7 (1993) 337–347.
36. Perry, N.C. and van Geerestein, V.J., *Database searching on the basis of 3D molecular similarity using the SPERM program*, J. Chem. Inf. Comput. Sci., 32 (1992) 607–616.
37. Badel, A., Mornon, J.P. and Hazout, S., *Searching for geometric molecular shape complementarity using bi-dimensional surface profiles*, J. Mol. Graph., 10 (1992) 205–211.
38. Blaney, F.E., Finn, P., Phippen, R.W. and Wyatt, M., *Molecular surface comparison: Application to molecular design*, J. Mol. Graph., 11 (1993) 98–105.
39. Blaney, F.E., Naylor, D. and Woods, J., *MAMBAS: A real time graphics environment for QSAR*, J. Mol. Graph., 11 (1993) 157–165.
40. Reynolds, C.A., Burt, C. and Richards, W.G., *A linear molecular similarity index*, Quant. Struct.-Act. Relat., 11 (1992) 34–35.
41. Klebe, G., Abraham, U. and Mietzner, T., *Molecular similarity indices in a comparative analysis (ComSIA) of drug molecules to correlate and predict their biological activity*, J. Med. Chem., 37 (1994) 4130–4146.
42. Kearsley, S.K. and Smith, G.M., *An alternative method for the alignment of molecular structure: Maximizing electrostatic and steric overlap*, Tetrahedron Comput. Methodol., 3 (1990) 615–633.
43. Good, A.C., Peterson, S.J. and Richards, W.G., *QSARs from similarity matrices: Technique validation and application in the comparison of different similarity evaluation methods*, J. Med. Chem., 36 (1993) 2929–2937.
44. Sneath, P.H.A. and Sokal, R.R., *Numerical Taxonomy*, W.H. Freeman, San Francisco, CA, 1973.
45. Good, A.C., *30 molecular similarity indices and their application in QSAR studies*, In Dean, P.M. (Ed.) *Molecular similarity in drug design*, Blackie Academic and Professional, Glasgow, 1995, pp. 138–162.
46. Brown, R.D. and Martin, Y.C., *The information content of 2D and 3D structural descriptor relevant to ligand-receptor binding*, J. Chem. Inf. Comput. Sci., 37 (1997) 1–9.
47. Burgess, E.M., Ruell, J.A., Zalkow, L.H. and Haugwitz, R.D., *Molecular similarity from atomic electrostatic multipole comparisons: Application to anti-HIV drugs*, J. Med. Chem., 38 (1995) 1635–1640.
48. Measures, P.T., Mort, K.A., Allan, N.L. and Cooper, D.L., *Applications of momentum space similarity*, J. Comput.-Aided Mol. Design, 9 (1995) 331–340.
49. Automated Similarity Package, developed and distributed by Oxford Molecular, the Medewar Centre, Oxford Science Park, Oxford OX4 4GA, U.K.
50. Bone, R.G. and Villar, H.O., *Discriminating D1 and D2 agonists with a hydrophobic similarity index*, J. Mol. Graph., 13 (1995) 165–74.
51. Szabo, A. and Ostland, N.S., *Modern quantum chemistry*, Macmillan, New York, 1982, pp. 410–412.
52. McMahon, A.J. and King, P.M., *Optimization of the Carbo molecular similarity index using gradient methods*, J. Comp. Chem., 18 (1997) 151–158.
53. Wild, D.J. and Willett, P., *Similarity searching in files of three-dimensional chemical structures: Alignment of molecular electrostatic potential fields with a genetic algorithm*, J. Chem. Inf. Comput. Sci., 36 (1996) 159–167.
54. Thorner, D.A., Wild, D.J., Willett, P. and Wright, P.M., *Similarity searching in files of three-dimensional chemical structures: Flexible field-based searching of molecular electrostatic potentials*, J. Chem. Inf. Comput. Sci., 36 (1996) 900–908.
55. Frisch, M.J., Head, G.M., Schlegel, H.B., Raghavachari, K., Binkley, J.S., Gonzalez, C., Defrees, D.J., Fox, D.J., Whiteside, R.A., Seeger, R., Melius, C.F., Baker, J., Martin, R., Kahn, L.R., Stewart, J.J.P., Fluder, E.M., Topiol, S. and Pople, J.A., *Gaussian 88*, Gaussian, Inc., Pittsburgh, PA, U.S.A.
56. Grant, J.A. and Pickup, B.T., *A Gaussian description of molecular shape*, J. Phys. Chem., 99 (1995) 3503–3510.
57. Grant, J.A., Gallardo, M.A. and Pickup, B.T., *A fast method of molecular shape comparison: A simple application of a Gaussian description of molecular shape*, J. Compt. Chem., 17 (1996) 1653–1666.
58. Chapman, D., *The measurement of molecular diversity: A three-dimensional approach*, J. Comput.-Aided Mol. Design, 10 (1996) 501–512.

59. Bladen, P., *A rapid method for comparing and matching the spherical parameter surfaces of molecules and other irregular objects*, J. Mol. Graph., 7 (1989) 130–137.
60. Blaney, F.E., Edge, C., Phippen, R.W., *Molecular surface comparison: 2. Similarity of electrostatic surface vectors in drug design*, J. Mol. Graph., 13 (1995) 165–74.
61. Connolly, M.L., *Computation of molecular volume*, J. Am. Chem. Soc., 107 (1985) 1118–1124.
62. Connolly, M.L., *Analytical molecular surface calculation*, J. Appl. Cryst., 16 (1983) 548–558.
63. Masek, B.B., Merchant, A. and Matthew, J.B., *Molecular skins: A new concept for quantitative shape matching of a protein with its small molecular mimics*, Protein, 17 (1993) 193–202.
64. Perkins, T.D.J., Mills, J.E.J. and Dean, P.M., *Molecular surface-volume and property matching to superpose flexible dissimilar molecules*, J. Comput.-Aided Mol. Design, 9 (1995) 479–490.
65. Seri-Levy, A. and Richards, W.G., *Chiral drug potency: Pfeiffer's rule and computed chirality coefficients*, Tetrahedron Asymmetry, 4 (1993) 1917–1921.
66. Seri-Levy, A., West, S. and Richards, W.G., *Molecular similarity, quantitative chirality and QSAR for chiral drugs*, J. Med. Chem., 37 (1994) 1727–1732.
67. Dughan, L., Burt, C. and Richards, W.G., *The study of peptide bond isosteres using molecular similarity*, J. Mol. Struct., 235 (1991) 481–488.
68. Montanari, C.A., Tute, M.S., Beezer, A.E. and Mitchell, J.C., *Determination of receptor-bound drug conformations by QSAR using flexible fitting to derive a molecular similarity index*, J. Comput.-Aided Mol. Design, 10 (1996) 67–73.
69. Cardozo, M.G., Kawai, T., Iimura, Y., Sugimoto, H., Yamanishi, Y. and Hopfinger, A.J., *Conformational analyses and molecular shape comparisons of a series of inandone-benzylpiperidine inhibitors of acetylcholinesterase*, J. Med. Chem., 35 (1992) 590–601.
70. Tokarski, J.S. and Hopfinger, A.J., *Three-dimensional molecular shape and analysis-quantitative structure-activity relationship of a series of cholecystokinin-A receptor antagonists*, J. Med. Chem., 37 (1994) 3639–3654.
71. Burke, B.J., Dunn, W.J. III and Hopfinger, A.J., *Construction of a molecular shape analysis-three-dimensional quantitative structure-analysis relationship for an analog series of pyridobenzodiazepinone inhibitor of muscarinic 2 and 3 receptor*, J. Med. Chem., 37 (1994) 3775–3788.
72. Rhyu, K.B., Patel, H.C. and Hopfinger, A.J., *A 3D-QSAR study of anticoccoidal triazines using molecular shape analysis*, J. Chem. Inf. Comput. Sci., 35 (1995) 771–778.
73. Holzgrabe, U. and Hopfinger, A.J., *Conformational analysis, molecular shape comparison, and pharmacophore identification of different allosteric modulators of muscarinic receptors*, J. Chem. Inf. Comput. Sci., 36 (1996) 1018–1024.
74. Rum, G. and Herndon, W.C., *Molecular similarity concepts: 5. Analysis of steroid protein binding constants*, J. Am. Chem. Soc., 113 (1991) 9055–9060.
75. Good, A.C., So, S. and Richards, W.G., *Structure Activity Relationships from Similarity Matrices*, J. Med. Chem., 36 (1993) 433–438.
76. Horwell, D.C., Howson, W., Higginbottom, M., Naylor, D., Ratcliffe, G.S. and Williams, S., *Quantitative structure-activity relationships (QSARs) of N-terminus fragments of NK1 tachykinin antagonists: A comparison of classical QSARs and three-dimensional QSARs from similarity matrices*, J. Med. Chem., 38 (1995) 4454–4462.
77. Good, A.C., Ewing, T.J.A., Gschwend, D.A. and Kuntz, I.D., *New molecular shape descriptors: Application in database screening*, J. Comput.-Aided Mol. Design, 9 (1995) 1–12.
78. Bemis, G.W. and Kuntz, I.D., *A fast efficient method for 2D and 3D molecular shape description*, J. Comput.-Aided Mol. Design, 6 (1992) 607–628.
79. Fisanick, W., Cross, K.P. and Rusinko, A. III, *Similarity searching on CAS registry substances: 1. Global molecular property and generic atom triangle geometric searching*, J. Chem. Inf. Comput. Sci., 32 (1992) 664–674.
80. Norel, R., Fischer, D., Wolfson, H.J. and Nussinov, R., *Molecular Surface Recognition by a Computer Vision Based Technique*, Protein Eng., 7 (1994) 39–46.
81. Kuntz, I.D., Blaney, J.M., Oatley, S.J., Langridge, R. and Ferrin, T.E., *A geometric approach to macromolecule-ligand interactions*, J. Mol. Biol., 161 (1982) 269–288.
82. Good, A.C. and Kuntz, I.D., *Investigating the extension of pair-wise distance pharmacophore measures to triplet based descriptors*, J. Comput.-Aided Mol. Design, 9 (1995) 373–379.

83. Pickett, S.D., Mason, J.S. and Melay, I.M., *Diversity profiling and design using 3D pharmacophores — pharmacophore derived queries (PDQ)*, J. Chem. Inf. Comput. Sci., 36 (1996) 1214–1233.
84. Davies, E.K. and Briant, C., *Combinatorial chemistry library design using pharmacophore diversity*, URL <http://www.awod.com/netsci/Science/Combichem/feature05.html>.
85. Chem-Diverse, developed and distributed by Chemical Deisng Ltd., Roundway House, Cromwell Park, Chipping Norton, Oxon OX7 5SR, U.K.
86. Lewis, R.A., Good, A.C. and Pickett, S.D., *Quantification of molecular similarity and its application to combinatorial chemistry*, In Computer-assisted lead finding and optimization, Proceedings of the 11th European Symposium on QSAR, Lausanne, Switzerland, 1996, van de Waterbeemd, H., Testa, E. and Folkers, G. (Eds.) Wiley-VCH, Easel, 1997, 135–156.



# Novel Software Tools for Chemical Diversity

Robert S. Pearlman\* and K.M. Smith

Laboratory for Molecular Graphics and Theoretical Modelling College of Pharmacy, University of Texas, Austin TX 78712. U.S.A.

## 1. Introduction: Diversity-related Tasks

Although the concept of chemical diversity has been intuitively considered by chemists for many years, the advent of combinatorial chemistry and high-throughput screening have focused attention on the need for efficient software tools to address a variety of diversity-related tasks. Perhaps the most fundamental task related to chemical diversity is that of selecting a diverse subset of compounds from a much larger population of compounds. The obvious objective of that task is to identify a subset which best represents the full range of chemical diversity present in the larger population, either to avoid the time and expense of synthesizing 'redundant' compounds or to avoid the time and expense of screening 'redundant' compounds. However, in addition to *simple subset selection*, recent practical experience has revealed other equally (possibly more) important diversity-related tasks which must also be addressed in the pharmaceutical and agrochemical industry.

High-throughput screening (HTS) can be an effective approach to lead discovery but is obviously limited by the structural diversity of compounds being screened. What if that population does not include representatives of one or more chemical classes or pharmacophores? Identifying *diversity-voids* or *missing diversity* is an important task and, obviously, *filling in diversity voids* with compounds from other sources is equally important. It is also important to be able to recognize and choose among the many compounds which might fill a particular diversity-void. These tasks become increasingly important as the number of combinatorially synthesizable compounds increases with advances in combinatorial chemical methods. and as the number of commercially available compounds increases. Similarly, *comparing diversities* of alternative compound libraries is another important diversity-related task.

In addition to simple diverse subset selection, it is often desirable to select a subset chosen not only to provide structural diversity, but also to satisfy one or more non-structural criteria or 'biases'. For example, compound availability and/or physical properties may be important when selecting a subset for HTS purposes. Reagent cost and/or reagent usage frequency may be important when deciding which compounds actually to synthesize out of a very large range of compounds synthetically accessible through combinatorial chemical methods. Clearly, non-structurally biased subset selection will yield subsets with somewhat less structural diversity than a subset chosen simply to maximize structural diversity, but practical considerations often make *biased subset selection* a very important diversity-related task.

---

\*To whom correspondence should be addressed.

The following section will discuss how chemical structures can be described for chemical diversity purposes. We shall refer to such descriptors as ‘metrics’ of a ‘chemistry-space’. An extremely important but often overlooked diversity-related task is that of *choosing the chemistry-space* metrics which *best represent the structural diversity of a given population of compounds*. For example, combinatorially generated populations, or populations chosen to be similar to a particular active (‘lead’) coinpound, are inherently less diverse than other more randomly assembled populations. Thus, it is quite reasonable to expect that metrics specifically tailored to focus on the limited diversity of such ‘focused populations’ will provide some advantages over metrics which were developed to best represent the broad range of diversity found in ‘non-focused populations’. Last but not least, the notion of considering alternative chemistry-space metrics reminds us of the need for a rational approach for *validating chemistry-space metrics* — an important and often misunderstood diversity-related task.

## 2. Chemistry-space Concepts and Diversity-related Algorithms

Before discussing software tools for addressing the aforementioned diversity-related tasks, it is useful to review a few fundamental concepts. The notions of chemical similarity, dissimilarity and, consequently, ‘diversity’ are all related to the distance between chemical compounds positioned in some multi-dimensional ‘chemistry-space’, the axes of which are the structure-related ‘chemistry-space metrics’ mentioned above. In order to be a well-defined vector-space, our chemistry-space axes must be orthonormal (mutually orthogonal, uncorrelated and normalized). We must also define a method for computing a true distance (one which satisfies the triangle inequality) within that space. The ‘diversity’ of compounds positioned in chemistry-space is intuitively related to the inter-compound distance as measured in that space.

Whereas the dimensionality of our physical world is predefined as 3, the dimensionality of a chemistry-space (as well as the definition of axes) can be chosen to best represent the diversity of a given population of compounds. Most software for addressing chemical diversity uses some form of ‘molecular fingerprint’ to describe each compound in a population. Fingerprints are bit-strings (sequences of 1s and 0s) representing the answers to yes/no questions about the presence or absence of various substructural features within the molecular structure of a given compound. Although not often discussed in such terms, each bit represents an axis in a multi-dimensional chemistry-space. Each axis could have either of two values: 0 or 1. Fingerprints represent very high-dimensional chemistry-spaces: typically between 150 and 200 bits for MDL applications [1], a few thousand for Tripos [2] and Daylight [3] applications or millions for the pharmacophore fingerprints introduced by Chemical Design Ltd. [4]. With the exception of the latter, fingerprints were developed to enable similarity and substructure searching.

Comparing fingerprints using the well-known Tanimoto similarity index,  $T$ , has proven useful for finding similar coinpounds within very large databases of chemical structures. Thus, it is typically assumed that the ‘Tanimoto dissimilarity’,  $(1-T)$ , represents a useful measure of distance within such high-dimensional spaces. Distance-based

diversity algorithms consider only the distances between compounds in chemistry-space and are used to select diverse subsets by choosing compounds guaranteed to be distant from other compounds in the selected subset. Practical experience and various validation studies (e.g. Brown and Martin [5]) indicate that such high-dimensional, distance-based diversity-related algorithms are, indeed, useful for simple diverse subset selection. However, the following criticisms should be considered:

1. The ‘questions’ corresponding to the bits of fingerprints were specifically developed to identify compounds *similar* to one another. They were not developed to focus on *differences* which would constitute a structurally diverse subset.
2. Although some software permits users to redefine the ‘questions’ corresponding to the bits of a fingerprint, this is rarely (if ever) done in actual practice. Thus, we use fingerprints developed to find similar compounds in diverse populations to select dissimilar (diverse) compounds not only from diverse populations, but also from focused populations.
3. The Tanimoto similarity index was developed to gauge similarity, not dissimilarity. That is, if  $T(A,B)$  and  $T(A,C)$  (the Tanimoto similarities between compounds  $A$ ,  $B$  and  $C$ ) are 0.9 and 0.8, compounds  $A$  and  $B$  are probably more structurally similar than compounds  $A$  and  $C$ . However, it is not at all certain that if  $(1 - T(X,Y))$  and  $(1 - T(X,Z))$  are 0.9 and 0.8, that compounds  $X$  and  $Y$  are more dissimilar (diverse) than compounds  $X$  and  $Z$ .
4. Although the Tanimoto similarity index appears useful for the similarity purposes for which it was designed,  $(1 - T)$  is *not* a valid measure of distance since it does not obey the triangle inequality. Thus, it appears that distance-based diversity (and, possibly, similarity) algorithms might be improved by using either the Euclidean distance or the Hamming distance as suggested by Pearlman [6,7].

Despite these criticisms, distance-based diversity algorithms appear to work satisfactorily for simple diverse subset selection, the most fundamental of the diverse-related tasks. However, as the term implies, distance-based algorithms consider only inter-compound distances, not the absolute positions of compounds in chemistry-space. As a result, distance-based algorithms are inherently limited and are ill-suited for many of the other diversity-related tasks. For example, locating diversity voids in chemistry-space is essentially impossible since distance-based algorithms do not reference location.

In contrast, by dividing each axis of a multi-dimensional space into ‘bins’, *cell-based diversity algorithms* partition chemistry-space into a lattice of multi-dimensional hypercubes and, thereby, consider not only inter-compound distance, but also absolute position of compounds in chemistry-space. As will be illustrated below, this additional information makes cell-based diversity algorithms much more powerful and easily applicable to *all* of the diversity-related tasks mentioned in the preceding section. However, whereas distance-based algorithms can be applied in either a high-dimensional or low-dimensional representation of chemistry-space, cell-based algorithms can only be applied in low-dimensional chemistry-spaces. (For example, a 1000-bit fingerprint corresponds to a 1000-dimensional chemistry-space which would

be partitioned into  $2^{1000}$  cells — an astronomically large number of cells almost all of which would contain no compounds.)

In order to take advantage of the power and utility of cell-based diversity algorithms, we must identify *low-dimensional matrices* from which to construct a less than 10-dimensional chemistry-space satisfactory for diversity purposes. There have been numerous attempts to use 'traditional' molecular descriptors (e.g. molecular weight, shape-factors, estimated logP, surface area, dipole moment, HOMO-LUMO gap. etc.) as the axes of a low-dimensional chemistry-space. There are three basic reasons for which these efforts have not proven particularly useful:

1. Many of the 'traditional' descriptors are highly correlated; the axes of a vector-space should be orthogonal (uncorrelated).
2. Some traditional descriptors (e.g. logP and pK<sub>a</sub>) are strongly related to drug transport or pharmacokinetics but are only weakly related to receptor affinity or activity as measured in most screening-based drug discovery efforts.
3. The traditional descriptors are whole-molecule descriptors which convey very little information about the details of molecular substructural differences which are the basis of structural diversity.

The first problem above could be addressed to a limited extent by using principal components of the 'traditional' descriptors as the axes, but the second and third more fundamental problems would remain. The advantages of cell-based methods cannot be realized unless some non-traditional chemistry-metrics can be found which enable the definition of a meaningful low-dimensional chemistry-space.

### 3. BCUT Values: Novel Low-dimensional Chemistry-space Metrics

In 1989, Burden [8] suggested that a 'molecular ID number' could be defined in terms of the two lowest eigenvalues of a matrix representing the hydrogen-suppressed connection table of the molecule. More specifically, Burden suggested putting the atomic numbers on the diagonal of the matrix. Off-diagonal matrix elements were assigned values of 0.1 times the nominal bond-type if the two atoms are bonded and 0.001 if the two atoms are not bonded. He also added 0.01 to the off-diagonal elements representing 'leaf edges' in the molecular graph (i.e. terminal bonds to the last atom in a chain). In suggesting that structurally similar compounds would be near each other in an ID-ordered list, Burden was actually proposing a one-dimensional chemistry-space. Since fingerprint-based similarity searching method? were just becoming available for modestly sized databases (under 0.5 million compounds), Burden's seemingly far-fetched suggestion was generally ignored.

In 1993, eager to find some sort of 'similarity searching method' applicable to the Chemical Abstracts Service (CAS) Registry File of approximately 12 million structures, Rusinko and Lipkus [9] applied Burden's suggestion to a test database of 60 000 compounds. The results were relatively poor compared to fingerprint-based similarity searching methods, but much better than expected. They also experimented with the notion of assigning a constant value to all diagonal matrix elements or a constant value

for all bonded off-diagonal elements but, in each case, were using the lowest eigenvalue of a single matrix to define a one-dimensional chemistry-space.

Based on Burden's (B) original suggestion and CAS's (C) 'validation' of that suggestion, Pearlman at the University of Texas (UT) added the following significant extensions which resulted in what we now refer to as the BCUT approach [6,7,10]:

1. Given that a one-dimensional chemistry-space showed some signs of promise, a similarly defined multi-dimensional chemistry-space should be even more promising. This is easily accomplished by using more than one matrix to represent each compound.
2. Mathematical analysis reveals that all eigenvalues of such matrices contain information related to molecular structure. The lowest and highest eigenvalues reflect the most different information (are least correlated). Considering both the lowest and highest eigenvalues provides another mechanism to extend Burden's original suggestion to a multi-dimensional space.
3. Pharmaceutical and agrochemical researchers are interested in structural diversity with respect to the way in which compounds might interact with a bioreceptor. Since atomic number has almost no bearing on the strength of intermolecular interactions, much more relevant metrics can be defined by putting more relevant atomic properties on the diagonals of four 'classes' of BCUT matrices: atomic charges, polarizabilities, H-bond donor- and acceptor-abilities corresponding to the electrostatic, dispersion and H-bonding modes of bimolecular interaction.
4. Burden's suggestion of using nominal bond-type information for the off-diagonal elements of the matrices was very good and should be retained. However, using CONCORD [11] to generate 3D structures opens the possibility of putting various functions of interatomic distance on the off-diagonals and, thereby, defining metrics which encode information about the 3D structure.
5. Another approach to incorporating aspects of 3D structure is to use atomic surface areas to weight the atomic properties placed on the diagonals.
6. Noting that the matrices contain atomic properties on the diagonals and connectivity information on the off-diagonals, there is clearly a need for a scaling-factor to provide the proper balance of the two types of information.

Given the large number of possible combinations of diagonal, off-diagonal and scaling-factor choices, it is clear that some method must be developed for rationally deciding which combination of BCUT values (eigenvalues) would form the chemistry-space which best represents the structural diversity of a given population of compounds. Many combinations can be quickly eliminated by requiring that the axes of the chemistry-space be mutually orthogonal. For example, different charge-related values (e.g. Gasteiger-Marsili charges, AM1 charges, AM1 densities, etc.) all convey the same fundamental information and, therefore, will be intercorrelated. On the other hand, with the exception of the H-bond-ability matrices 1, both the highest and lowest eigenvalues should be relevant and turn out to be relatively uncorrelated. (The lowest eigenvalues of these matrices contain 'information' about atoms in the molecule which are neither H-bond donors or acceptors. Since all non-H-bonding atoms have the same zero value

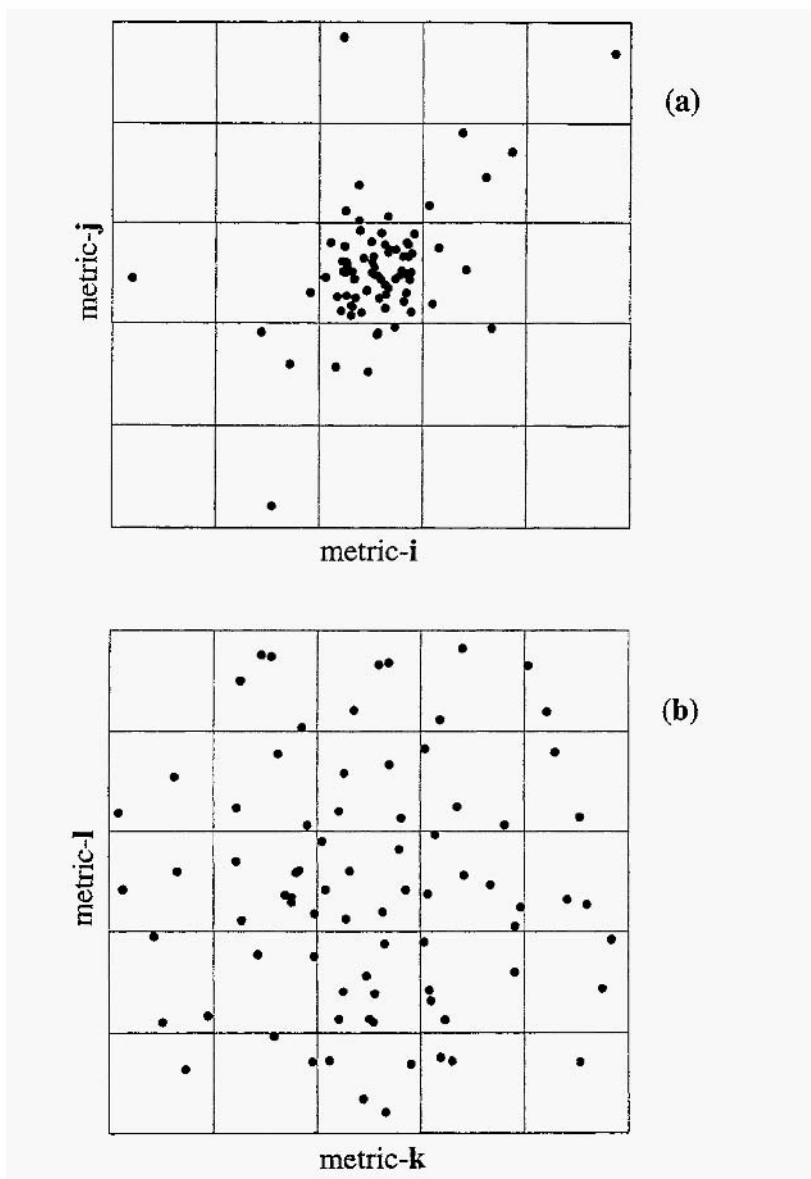


Fig. 1. (a) Cartoon representation of non-optimal two-dimensional chemistry space showing poorly distributed compounds. (b) Representation of a better two-dimensional chemistry-space showing more evenly distributed compounds.

of H-bond-ability, the lowest eigenvalues of these matrices are information-poor.) Sometimes, a 6-dimensional chemistry-space (two charge-BCUTs, two polarizability BCUTs and two H-bond-BCUTS) yields the best chemistry-space for a given population. Often, the H-bond-acceptor- and charge-BCUTs are correlated, yielding a

5-dimensional chemistry-space. Pearlman and Smith [6,7,10] developed a powerful 'auto-choose' algorithm which automatically determines both the best dimensionality of the chemistry-space and best choice of exactly which metrics best represent the structural diversity of a given population of compounds.

The rationale for the auto-choose algorithm can most easily be explained by reference to Figs 1a and 1b which depicts two 'cartoon' representations of the same population of compounds distributed in two different two-dimensional chemistry spaces. Recall that the most fundamental of diversity-related tasks is that of rational, structure-based diverse subset selection. Suppose that one were asked to select a subset of 25 diverse compounds. The chemistry-space defined by axes *i* and *j* in Fig. 1 a does a poor job of distinguishing one compound from another and positions most of the compounds in the central 'cell', thereby forcing us to choose several compounds from that cell *at random*. Clearly, this would be a very poor choice of chemistry-space axes for this population of compounds since it would provide little or no advantage over a random choice made without reference to any chemistry-space considerations. In contrast, the chemistry-space defined by axes *k* and *l* in Fig. 1b does a much better job of distinguishing compounds based on structural diversity and enables the selection of a subset of 24 diverse compounds (satisfactorily close to the desired size of 25) by simply choosing one compound from each of the 24 occupied cells. Note that by choosing compounds as close to the center of each occupied cell as possible, we have described a very natural cell-based subset selection algorithm which yields a set of compounds and 'covers' the range of diversity represented by the population and includes compounds which are mutually distant from one another. Clearly, the chemistry-space yielding the most uniform distribution of compounds would best suit our purposes. However, real-world populations will not be distributed in a perfectly uniform fashion. Valence and steric considerations limit the continuity of structures which can be achieved and, more obviously, the history of discovery efforts at a given company will result in corporate databases reflecting those focused efforts. However, the  $\chi$ -squared statistic provides a measure of how well one distribution matches another. Thus, minimizing the  $\chi$ -squared statistic can reveal which combination of metrics yields a distribution of compounds closest to the hypothetical uniform distribution. Simultaneously, by considering a range of dimensionalities, this  $\chi$ -squared approach also reveals the dimensionality of the chemistry-space which best represents the diversity of the given population of compounds. Clearly, in order to include as much structure-distinguishing information as possible, the algorithm will choose the highest-possible dimensionality which does not result in correlated (non-orthogonal) axes. (Correlated axes would result in some highly populated cells along a diagonal of the chemistry-space and corresponding empty cells surrounding that diagonal. This would yield a high  $\chi$ -squared and the chemistry-space would be rejected).

Note that the  $\chi$ -squared approach yields the combination of metrics which best represents the diversity of a given population of compounds. For 'truly diverse' populations of compounds, we are not surprised to find the same (or similar) 'universal chemistry-space' definition being reported by different users. In contrast, populations resulting from generating different combinatorial libraries should be expected to occupy different and individually less diverse regions of 'universal chemistry-space'. Clearly, the

$\chi$ -squared approach enables us to tailor a chemistry-space to best represent a focused population — an important diversity-related task listed in the Introduction section of this chapter.

It should also be noted that this auto-choose,  $\chi$ -squared approach can be applied not only to combinations of BCUT values, but to combinations involving any other low-dimensional chemistry-space metrics. Although experience to date strongly supports the use of BCUT values as metrics, software developed by Pearlman and Smith [12] encourages the user to consider his own metrics in addition to BCUT values. However, this must be done with extreme caution for the following somewhat ironic reason. Imagine assigning random numbers to each of compounds of a large populations and then considering those numbers as a potential axis of a chemistry-space. Since the random numbers would be uniformly distributed over the population of compounds, the  $\chi$ -squared approach would perceive this ‘metric’ (and other similarly random ‘metrics’) as good choices as axes of a chemistry-space. This brings us, rather dramatically, to the need to validate the choice of metrics used to define a chemistry-space.

#### 4. Validation of Chemistry-space Metrics

Obviously, chemistry-space ‘metrics’ which are merely random numbers with no relation to structure would be of no use for diversity-related tasks or any other purpose. How can we demonstrate that a given set of metrics is actually reflecting differences in molecular structure and, thereby, validate those metrics for use in addressing chemical diversity-related tasks?

Perhaps the most intuitive approach to metric validation is to use the metrics as QSAR descriptors; i.e. establish a linear regression equation relating the metrics to the experimentally measured ‘activities’ of a set of compounds. Let us imagine that we can put aside the differences between receptor-affinity and actual activity (due to transport issues or secondary processes in the cascade of events between initial receptor binding and eventual pharmacological effect). Since it would be impossible to establish a statistically significant regression based on meaningless, random numbers, demonstrating such a regression would be proof that the metrics are not random numbers, but true indicators of chemical structure. After auto-choosing a six-dimensional BCUT chemistry-space to best represent the diversity of their entire corporate database, Weintraub and Demeter [13] used those six BCUT values to regress the  $\log IC_{50}$  values measured for 800 ligands at the benzodiazepine site of the GABA<sub>A</sub> receptor and obtained a PLS model essentially as good as one they obtained previously based upon 70 classical QSAR descriptors.

While the results of Weintraub and Demeter certainly confirmed the validity of BCUT values as chemistry-space metrics, those results are, unfortunately but not unexpectedly, quite rare! As will be illustrated below, chemistry-space metrics are *not* QSAR descriptors and rarely yield regressions as good as that reported by Weintraub and Demeter, Chemistry-space metrics are intended to position compounds in a structure-based chemistry-space. QSAR descriptors are intended to provide quantitative estimates of bioactivity. Chemistry-space metrics are intended to reflect (in a necessarily



crude manner) *all* features of molecular structure. In contrast, QSAR descriptors are specifically chosen to reflect (as accurately as possible) only those features of molecular structure which have been found relevant for a *particular* bioactivity. It is well known that QSARs will give unreliable estimates of activity when applied to compounds containing structural features not present in the training set used to identify the ‘relevant’ features characterized by the QSAR descriptors. We certainly should *not* expect QSARs based on chemistry-space metrics to do any better since (i) the metrics were not intended for this purpose and (ii), as will be explained below, position in chemistry-space is not *quantitatively* related to activity.

How, then, should chemistry-space metrics be validated? Pearlman and Smith [10] have presented a simple yet novel approach to metric validation which they refer to as *activity-seeded, structure-based clustering*. Unlike typical clustering algorithms (based on structure alone) which can be used for a variety of tasks, this algorithm requires activity data (preferably, quantitative data) for a set of compounds and is intended *only* for the diversity-related task of *validating chemistry-space metrics*. Given a set of active compounds which all bind to a given receptor in the same way, it is certainly reasonable to expect that those active compounds should be positioned near each other in a small region of chemistry-space *if* the chemistry-space metrics are valid. The activity-seeded, structure-based clustering algorithm provides a method for directly testing that expectation in the typical case in which the chemistry-space dimensionality is greater than 3 and, thus, simple visual inspection of the distribution of active compounds is difficult or impossible. The algorithm consists of the following procedure:

1. Choose a unit-cluster radius: a small distance in the chemistry-space to be validated.
2. Center a sphere of that radius on the most active compound in the validation set.
3. Assign other active compounds located within that sphere to that ‘unit-cluster’.
4. Center another sphere on the next most active compound not already assigned to some unit-cluster.
5. Repeat steps 3 and 4 until all active compounds have been assigned to some unit-cluster.
6. ‘Coalesce’ adjoining (overlapping) unit-clusters and record the number of unit-cluster spheres per coalesced-cluster.

The algorithm can be implemented as an  $O(N)$  process and, thus, is extremely fast. More significantly, the algorithm can be used to validate all types of chemistry-space definitions including other (non-BCUT) low-dimensional chemistry-spaces and those based on high-dimensional fingerprints as well. When used in a cell-based context, the unit-cluster radius is typically chosen to yield a tiny hypersphere of volume equal to that of a single hypercubic cell reflecting the ‘resolution’ corresponding to a user-specified number of bins/axis (see below). In any case, the total number of unit-cluster spheres contained in all coalesced-clusters provides an *upper bound* on the volume of chemistry-space required to contain all the active compounds.

Using the activity-seeded, structure-based clustering algorithm, Pearlman and Deanda [14] have performed a number of validation studies. For example, after auto-choosing

the BCUT chemistry-space which best represents the diversity of compounds in MDL's MDDR database [15], they computed the positions of 197 relatively diverse ACE inhibitors in that chemistry-space. The 197 inhibitors were culled from the primary literature [16–23]. Measured activities ( $-\log IC_{50}$ ) were reported for all compounds and spanned the range 5.24 to 9.64. The 78 most active compounds (top 40%) had activities in the range 7.85 to 9.64 and were identified as 'highly active' compounds. If the BCUT values used as chemistry-space metrics were random numbers or quantities unrelated to structure and intermolecular interaction, the active compounds would be randomly distributed throughout chemistry-space. However, using the activity-seeded, structure-based clustering algorithm, they found that the 78 'highly active' compounds are all contained by just 3 coalesced clusters occupying less than 0.02% of the entire chemistry-space and less than 0.19% of occupied chemistriespace. Significantly, the 3 clusters were close to each other; the largest inter-cluster distance being just  $3.2R$  where  $R$  is the unit-cluster radius.

It is instructive to consider the analogous results obtained using all 197 compounds (including the 119 'poorly active' compounds). Once again, the active compounds were all clustered relatively near each other but they occupied a much larger volume of chemistry-space than that occupied by just the 78 'highly active' compounds. This result is entirely consistent with expectations. There can be many different structures which exhibit poor to modest activities. In contrast, there are relatively fewer structures which exhibit high activities. This fact may be easier to appreciate by considering the notion of making structural modifications of a very highly active compound. There may be a few modifications which preserve high activity but there are far more modifications which reduce or even completely destroy activity.

The fact that poorly to modestly active compounds are spread over larger regions of chemistry-space than highly active compounds illustrates one reason for which chemistry-space metrics cannot (and should not) be used as QSAR descriptors. Compounds with significantly different structures would be positioned at widely distant points in chemistry-space but could exhibit low or moderate bioactivities (or affinities) of exactly equal magnitudes. The converse illustrates a second reason for which chemistry-space metrics cannot (and should not) be used as QSAR descriptors. For example, adding just a single methylene unit to the middle of the  $-\text{CH}_2\text{CH}_2\text{OH}$  side chain of some highly active compound could completely destroy the activity if the propyl-hydroxy derivative no longer fits into the receptor. Thus, two highly similar compounds (which any valid metrics would place very near each other in chemistry-space) could have entirely dissimilar activities. Clearly, neither QSAR nor any other approach based on the assumption of a quantitative relationship between activity and precise position in chemistry-space will be a valid approach to metric validation. On the other hand, the activity-seeded, structure-based clustering approach clearly indicates whether a given set of metrics places compounds active against the same receptor in the same small region of chemistry-space and, thus, provides a rational basis for metric validation.

## 5. Using Low-dimensional Metrics for Diversity-related Tasks

Once one has determined which metrics define the chemistry-space which best represents the diversity within a given population of compounds, one is then able to use

various cell-based algorithms to address all of the other diversity-related tasks mentioned in the introductory section of this chapter. Recall that such algorithms can exploit not only the knowledge of inter-compound distances, but also the knowledge of absolute positions of compounds in chemistry-space. Structurally similar compounds are positioned near each other in chemistry-space and, thus, are found 'clustered' in the same or neighboring cells.

A chemistry-space, like any other vector-space, must be comprised of normalized axes (so that a distance of, say, 4 units in one direction is equivalent to a distance of 4 units in any other direction). Thus, the 'cells' are hypercubes resulting from dividing each of the normalized axes of a chemistry-space into equal numbers of evenly spaced 'bins'. The number of bins/axis is directly related to the 'resolution' with which one examines the distribution of compounds across chemistry-space and is inversely related to the apparent 'occupancy' of that chemistryspace. For example, if 250 000 compounds are distributed in some 6-dimensional chemistry-space and each axis is 'divided' into just one single bin, all 250 000 compounds would be contained in just one single cell. The occupancy (number of occupied cells divided by total number of cells) would be 100% but the resolution would, obviously, be uselessly low. If each axis were divided into 20 bins, there would be  $20^6 = 64\,000\,000$  tiny cells. In this case, the occupancy would be extremely low and the resolution would be uselessly high: most cells would be empty and even very similar compounds could be in different, non-neighboring cells. Cell-based algorithms for some tasks automatically choose the number of bins/axis most appropriate for that task and population. Other tasks require that the user decide on the resolution (see below). Recalling that typical populations of compounds are not uniformly distributed, experience has shown that choosing the number of bins/axis which yields roughly 12% to 16% occupancy provides an appropriate level of resolution for most purposes.

## 6. Simple and Biased Diverse Subset Selection

Our explanation of the  $\chi$ -squared approach to auto-choosing the metrics of a chemistry-space also illustrated the essence of the natural, cell-based approach to diverse subset selection. As implied in that illustration, we recommend selecting one compound from each occupied cell, although our software also allows one to sample each cell in proportion to its occupancy or by selecting up to some fixed number of compounds per cell. Once the user has specified the sampling protocol (number per cell) and the size of the desired subset, the software automatically finds the number of bins/axis which yields the number of occupied cells required to provide a subset closest in size to that requested. Cell-based algorithms are *extremely* fast and especially well-suited to handling very large populations of compounds. Even if the software must make three of four guesses before finding the best number of bins/axis, selecting a subset of 50 000 structurally diverse compounds from a population of 0.5 million would take approximately 20 cpu seconds on a modest workstation.

By selecting compounds nearest the center of each cell, we can avoid choosing compounds near each other but just barely on opposite sides of a plane separating two cells. In other words, by selecting compounds nearest the center of each cell, we are selecting a subset of maximal structural diversity, namely *simple diverse subset selection*.

*Biased subset selection* can easily be accomplished by allowing the user to construct a modified selection rule. In other words, rather than choosing the compound nearest the center of a given cell, one can arrange to choose the compound which provides the best (user-specified) compromise between distance from center and some non-structural property. For example, given a choice between two compounds from the same small region (cell) of chemistry-space, availability (price, quantity on hand, etc.) might certainly be important considerations for assembling a subset for general screening purposes. Recalling that logP in a poor chemistry-space metric but, nevertheless, quite important for bioactivity, choosing compounds from each cell closest to some particular 'ideal' logP could be advantageous.

Biased subset selection can also be used to improve the efficiency and economy of combinatorial library synthesis. Imagine that 1000 A-type and 1000 B-type reactants could be used to make 1000 000 AB-type products but that just 10 000 products are desired for screening purposes. Selecting 100 diverse A's and B's offers obvious practical advantages but, clearly, does not yield as diverse a set of 10 000 as could have been selected from the complete set of 1 000 000 products. Simple diverse subset selection from all the products would undoubtedly result in the need to use many more than 100 of each type of reactant. By keeping track of the frequency with which each reactant is used in the products being selected, and by specifying the format of the plates used for the syntheses (e.g. typical  $8 \times 12 = 96$ -well plate), DiverseSolutions [12] used in conjunction with CombinDBMaker (for combinatorial database generation [24]) enables the user to specify a selection rule which chooses compounds providing the best (user-specified) compromise between distance from center of cell and economy.

## 7. Identifying and Filling in Diversity Voids

In order to address the possibility of finding leads to bioactive compounds in regions of chemistry-space not covered by their current collection of compounds, pharmaceutical and agrochemical companies allocate a certain fraction of their resources to compound acquisition programs: purchasing, trading for or synthesizing additional compounds for screening. Practical considerations (cost, screening capacity, etc.) limit the number of compounds companies choose to acquire.

Identifying and filling in diversity voids is trivially simple using cell-based algorithms. Obviously, 'empty' cells represent regions of missing diversity. 'Empty' can be defined to mean either that the cell contains no compounds or that it contains less than some user-specified number. When identifying the diversity voids in a given population of compounds, the number of empty cells will depend not only on how those compounds are distributed, but also upon the 'resolution' at which the 'search' for empty cells is performed. Whereas the number of bins/axis can be chosen by the software during subset selection, the user must choose the number of bins/axis which will yield a number of diversity voids consistent with those practical considerations which limit the number of compounds his company chooses to acquire (see below).

Since cell-based algorithms (unlike distance-based algorithms) reference absolute compound coordinates in chemistry-space, tilling in diversity-voids is also trivially

simple using cell-based methods: compounds (from some secondary population) are acquired if they would occupy a previously ‘empty’ cell in the chemistry-space containing the primary population. Of course, this entails precomputing the coordinates (metrics) of the secondary population in the same chemistry-space as that used to contain the primary population. Since we know exactly which cell would be filled by each candidate compound, we can easily bias our choice of fill-in compounds using the same sort of lion-structural criteria as discussed for biased subset selection. The user can also specify how many compounds lie wants to add to ‘empty’ cells. *DiverseSolutions* then presents the list of compounds to acquire in various formats to facilitate purchase decisions (e.g. compounds which would ensure at least 1 compound in each ‘empty’ cell, compounds which would ensure at least 2 compounds in each ‘empty’ cell, etc.).

Finding the diversity voids in a population of 0.5 million compounds typically takes less than 5 cpu seconds on a modest workstation. Filling in those voids (to the extent possible) from a library of 50 000 compounds typically takes less than 4 cpu seconds. Thus, the user can easily experiment with several values of *bind/axis* and several filling-in protocols.

It is worth noting that the filling-in process does not require any information about the compounds contained in the primary population. All that is required is the definition of the chemistry-space (i.e. name and range of the metric corresponding to each axis), the number of bins/axis, and the cell-numbers of the ‘empty’ cells. Thus, without revealing the compounds in its proprietary database, company X can, in essence, enable company Y to identify compounds which would fill in company X’s missing diversity.

## **8. Comparing the Diversities of Two or More Populations**

Occasionally, it may be useful to compare the diversities of two (or more) populations of compounds — perhaps alternative third-party libraries one could purchase or alternative combinatorial libraries one could synthesize to augment the diversity of a corporate database. Distance-based approaches merely allow the comparison of statistics related to nearest-neighbor distances within the two populations. Such statistics provide no information regarding the redundancy of compounds contained in both populations, or even the extent to which the regions covered by the populations overlap in chemistry-space.

In contrast, a cell-based approach provides an extremely rapid answer to the fundamental, pragmatic questions at the heart of the population comparison issue: if population A and population B are alternative libraries and population X is a corporate database,

1. how many population-A compounds fill voids in population X?
2. how many population-B compounds fill voids in population X?
3. how many population-A compounds fill voids in population B?
4. how many population-B compounds fill voids in population A?

Questions 1 and 2 are easily answered by identifying the voids in population X and hypothetically using both populations A and B to fill those voids. Similarly, questions 3

and 4 can be answered simultaneously by a 'compare diversities' algorithm which, essentially, performs two find-voids and fill-in tasks at the same time. In addressing questions 1 and 2, it is natural to use the chemistry-space which best represents the diversity of population X. In addressing questions 3 and 4, one might use a chemistry-space previously defined for some related population or use a chemistry-space defined to best represent the union of the A and B populations.

## 9. Summary

If properly constructed, high-dimensional (fingerprint) and low-dimensional metrics can provide equally valid representations of chemistry-space for chemical diversity purposes. High-dimensional metrics offer the advantage of providing substantial detail regarding the topological aspects of molecular substructure but suffer the disadvantage that they can be used only for distance-based algorithms for addressing the various diversity-related tasks encountered in pharmaceutical and agrochemical industry. Low-dimensional metrics offer the advantage of enabling the use of either distance-based or cell-based algorithms but traditional molecular descriptors are often cross-correlated, provide little or no substructural information and, thus, are poor choices for chemistry-space metrics, BCUT values constitute a novel class of molecular descriptors which not only encode substructural topological (or topographical) information, but also encode atom-based information relevant to the strength of ligand-receptor interaction.

We have presented an algorithm for choosing those low-dimensional metrics which best represent the diversity of a given population of compounds, an algorithm for validating the chosen metrics and cell-based algorithms using those metrics to address all of the diversity-related tasks. Work is currently in progress to develop additional metrics and a more efficient implementation of the algorithm for choosing the best chemistry-space.

## References

1. MACSS and Isis are distributed by Molecular Design Ltd. Information Systems. San Leandro, CA.
2. Unity is distributed by Tripos, Inc., St. Louis, MO.
3. DayMenus is distributed by Daylight Chemical Information Systems, Inc., Irvine, CA.
4. ChemDBS-3D is distributed by Chemical Design Ltd., Oxford, UK.
5. Brown, R.D. and Martin, Y.C., *Use of structure-activity data to compare structure-based clustering methods and descriptors for use in compound selection*, J. Chem. Inf. Comp. Sci., 36 (1996), 572-584.
6. Pearlman, R.S., Novel software tools for addressing chemical diversity, *NetSci*, <http://www.awod.com/netsci>, 2 (1996), 1.
7. Pearlman, R.S., *Diverse Solutions User's Manual*, University of Texas, Austin, TX, 1996.
8. Burden, F.R., *Molecular identification number for substructure searches*, J. Chem. Inf. Comp. Sci., 29 (1989), 225-7.
9. Rusinko III, A. and Lipkus, A.H., unpublished results obtained at Chemical Abstracts Service, Columbus OH.
10. Pearlman, R.S. and Smith, K.M., manuscript in preparation.
11. CONCORD was developed by R.S. Pearlman, A. Rusinko, J.M. Skell, and R. Balducci at the University of Texas, Austin TX and is distributed by Tripos, Inc., St. Louis, MO.
12. DiverseSolutions was developed by R.S. Pearlman and K.M. Smith at the University of Texas, Austin TX and is distributed by Tripos, Inc., St. Louis, MO.

13. Weintraub, H.J.R. and Demeter, D.A., personal communication.
14. Pearlman, R.S. and Deanda, F., manuscript in preparation.
15. Modern Drug Data Report distributed by Molecular Design Ltd. Information Systems, San Leandro, CA.
16. Sweet, C.S., Ulm, E.H., Gross, D.M., Vassil, T.C. and Stone, C.A., *A new class of angiotensin-converting enzyme inhibitors*, Nature, 288 (1980) 280–283.
17. Suh, J.T., Skiles, J.W., Williams, B.E., Youssefeyeh, R.D., Jones, H., Loev, B., Neiss, E.S., Schwab, A., Mann, W.S., Khandwala, A., Wolf, P.S., and Weinryb, I., *Angiotensin-Converting Enzyme Inhibitors. New Orally active antihypertensive (Mercaptoalkanoyl)- and [(Acylothio)alkanoyl]glycine derivatives*, J. Med. Chem., 28 (1985) 57–66.
18. Menard, P.R., Suh, J.T., Jones, H., Loev, B., Neiss, E.S., Wilde, J., Schwab, A., and Mann, W.S., *Angiotensin-Converting Enzyme Inhibitors. (Mercaptoaroyl)amino acids*, J. Med. Chem., 28 (1985) 328–332.
19. Wyratt, M.J. and Patchett, A.A., *Recent developments in the design of angiotensin-converting enzyme inhibitors*, Med. Res. Rev., 5 (1985) 483–531.
20. Karanewsky, D.S., Badia, M.C., Cushman, D.W., DeForrest, J.M., Dejneka, T., Loots, M.J., Perri, M.G., Petrillo, E.W., and Powell, J.R., *(Phosphinyloxy)acyl amino acid inhibitors of angiotensin-converting enzyme (ACE). 1. Discovery of (S)-1-[6-amino-2-[[hydroxy(4-phenylbutyl)phosphinyl]oxy]-l-oxohexyl-L-proline novel orally active inhibitor of ACE*, J. Med. Chem., 31 (1988) 204–212.
21. Yanagisawa, H., Ishihara, S., Ando, A., Kanazaki, T., Miyamoto, S., Koike, H., Iijima, Y., Oizumi, K., Matsushita, Y. and Hata, T., *Angiotensin-Converting Enzyme Inhibitors. 2. Perhydroazepin-2-one derivatives*. J. Med. Chem., 31 (1988) 422–428.
22. Krapcho, J., Turk, C., Cushman, D.W., Powell, J.R., DeForrest, J.M., Spitzmiller, E.R., Karanewsky, D.S., Duggan, M., Rovnyak, G., Schwartz, J., Natarajan, S., Godfrey, J.D., Ryono, D.E., Neubeck, R., Atwal, K.S., and Petrillo, E.W., *Angiotensin-Converting Enzyme Inhibitors. Mercaptan, carboxyalkyl dipeptide, and phosphinic acid inhibitors incorporating 4-substitutedprolines*, J. Med. Chem., 31 (1988) 1148–1160.
23. Karanewsky, D.S., Badia, M.C., Cushman, D.W., DeForrest, J.M., Dejneka, T., Lee, V.G., Loots, M.J., and Petrillo, E.W., *(Phosphinyloxy)acyl amino acid inhibitors of angiotensin-converting enzyme. 2. Terminal amino acid analogues of (S)-1-[6-amino-2-[[hydroxy(4-phenylbutyl) phosphinyl]oxy]-l-oxohexyl]-L-proline*, J. Med. Chem., 33 (1990) 1459–1469.
24. CombinDBMaker was developed by R.S. Pearlman and E.L. Stewart at the University of Texas, Austin TX and is available from the authors.

**This Page Intentionally Left Blank**



# New 3D Molecular Descriptors: The WHIM theory and QSAR Applications

Roberto Todeschini\* and Paola Gramatica

*Department of Environmental Sciences, via Emanuelli, 15, I-20126 Milan, Italy*

## 1. Introduction

Molecular descriptors represent the way chemical information, contained in the molecular structure, is transformed and coded to deal with chemical, pharmacological and toxicological problems in quantitative structure–activity (QSAR) and structure–property (QSPR) studies. Molecular descriptors take into account different aspects of the chemical information. The approach to obtaining this information can (a) be through experiments, theoretical calculations or simple counting operations, (b) consider the whole molecule, fragments of it or functional groups, (c) require the knowledge of the 3D structure of the molecule or its molecular graph, or simply its formula, or (d) call for information defined by scalar values, vectors or scalar fields. In recent years, several approaches have been explored and many kinds of molecular descriptors have been proposed [1–10].

Among the theoretical descriptors, the best known are molecular weight and structural descriptors (1D descriptors, i.e. counting of bonds, atoms of different kinds, presence or counting of functional groups and fragments, etc.), obtained from a simple knowledge of the formula, and topological descriptors (2D descriptors), obtained from the knowledge of the molecular topology [5–8].

The complexity of the chemical information contained in 3D molecular structure calls for descriptors able to also take into account properties related to a more sound and three-dimensional representation of the molecules, WHIM (*Weighted Holistic Invariant Molecular*) descriptors are 3D molecular indices that represent different sources of chemical information [9–13]. WHIM descriptors contain information about the whole 3D molecular structure in terms of size, shape, symmetry and atom distribution. These indices are calculated from x,y,z-coordinates of a 3D structure of the molecule, usually from a spatial conformation of minimum energy, within different weighting schemes in a straightforward manner and represent a very general approach to describe molecules in a unitary conceptual framework. These indices have already been successfully used to search for QSAR and QSPR relationships for several classes of compounds and different responses [9–16].

The WHIM descriptor approach has also been extended to treat interaction scalar fields [17]: G-WHIM (*Grid-Weighted Holistic Invariant Molecular*) descriptors are defined and calculated from the coordinates of the grid-points where an interaction energy field between the molecule and a probe has been evaluated. In both WHIM approaches, the chemical information contained in the molecular structure or in the whole

---

\*To whom correspondence should be addressed.

grid of interaction energy values is synthesized in a few parameters that are completely invariant to rotation and translation and that represent interpretable properties of the molecules. Moreover, in the G-WHIM approach, the difficulties commonly related to chemical information spread out over a great number of grid-points and to the problem of the dependence of the results upon molecule alignment are avoided.

The interpretability of the results is quite evident and defined by the same mathematical properties of the algorithm used for their calculation. In this review, the theories of both WHIM approaches are presented and some applications in the QSAR field are discussed. A simple didactical example is also used to explain properties, characteristics and chemical meaning of the WHIM descriptors. Due to the continuous development of the theory itself, there is some discrepancy in the WHIM symbols used in previous publications and those used here, the final choice is in this review.

## 2. Theory of WHIM Descriptors

WHIM descriptors are built in such a way that they capture relevant molecular 3D information regarding molecular size, shape, symmetry and atom distribution with respect to invariant reference frames. The algorithm consists in performing a Principal Components Analysis (PCA) on the centered molecular coordinates by using a weighted covariance matrix **S** obtained from different weighting schemes for the atoms. The elements of the covariance matrix are:

$$s_{jk} = \frac{\sum_{i=1}^n w_i (q_{ij} - \bar{q}_j)(q_{ik} - \bar{q}_k)}{\sum_{i=1}^n w_i} \quad (1)$$

where  $n$  is the number of atoms.  $w_i$  the weight of the  $i$ -th atom.  $q_{ij}$  represents the  $j$ -th coordinate ( $j = 1,2,3$ ) of the  $i$ -th atom and  $\bar{q}_j$  is the average of the  $j$ -th coordinates.

### 2.1. The weighting schemes

Six different weighting schemes have been proposed: (1) the unweighted case **U** ( $w_i = 1$   $i = 1, n$ , where  $n$  is the number of atoms for each compound), (2) atomic masses **M** ( $w_i = m_i$ ), (3) the van der Waals volumes **V** ( $w_i = vd w_i$ ), (4) the Mulliken atomic electronegativities **E** ( $w_i = eln_i$ ), (5) the atomic polarizabilities **P** ( $w_i = pol_i$ ) and (6) the electrotopological indices of Kier and Hall **S** ( $w_i = S_i$ ) [5].

All the weights (1)–(5) are also scaled with respect to the carbon atom and their values (original and scaled values) are shown in Table 1. As all the weights must be positive, the electrotopological indices are scaled as follows:

$$S'_i = S_i + 7 \quad S'_i > 0 \quad (2)$$

In this case, only the non-hydrogen atoms are considered and the atomic charge of each atom is dependent on its atom neighbor, following the Kier and Hall approach.

Table 1 Atomic weights and relative atomic weights used for calculation of the WHIM descriptors

ID	Atomic mass		vdW volume		Electronegativity		Polarizability	
	M	M/M(C)	v	W/V(C)	E	W/E(C)	P	W/P(C)
H	1.01	0.084	6.709	0.299	2.592	0.944	0.667	0.379
B	10.81	0.900	17.875	0.796	2.275	0.828	3.030	1.722
C	12.01	1.000	22.449	1.000	2.746	1.000	1.760	1.000
N	14.01	1.166	15.599	0.695	3.194	1.163	1.100	0.625
O	16.00	1.332	11.404	0.512	3.654	1.331	0.802	0.456
F	19.00	1.582	9.203	0.410	4.000	1.457	0.557	0.316
Al	26.98	2.246	36.511	1.626	1.714	0.624	6.800	3.864
Si	28.09	2.339	31.976	1.424	2.138	0.779	5.380	3.057
P	30.97	2.579	26.522	1.181	2.515	0.916	3.630	2.063
S	32.07	2.670	14.429	1.088	2.957	1.077	2.000	1.648
Cl	35.45	2.952	23.228	1.035	3.175	1.265	2.180	1.239
Fe	55.85	4.650	41.052	1.829	2.000	0.728	8.400	4.773
co	58.93	4.907	35.011	1.561	2.000	0.728	7.500	4.261
Ni	58.69	4.887	17.157	0.764	2.000	0.728	6.800	3.864
Cu	63.55	5.291	11.494	0.512	2.033	0.740	6.100	3.466
Zn	65.39	5.445	38.351	1.708	2.223	0.810	7.100	4.034
Br	79.90	6.653	31.059	1.384	3.319	1.172	3.050	1.733
sn	118.71	9.884	45.830	2.042	2.298	0.837	7.700	4.375
I	126.90	10.566	38.792	1.728	2.778	1.012	5.350	3.040

Depending on the kind of weighting scheme, different covariance matrices and different principal axes are obtained. For example, using atomic masses as the weighting scheme, the directions of the three principal axes from PCA are the directions of the moments of inertia axes. Thus, the WHIM approach can be viewed as a generalization searching for the principal axes with respect to a defined atomic property (the weighting scheme). For each weighting scheme, a set of statistical indices is calculated on the atoms projected onto each principal component  $\mathbf{t}_m$  ( $m = 1, 2, 3$ ), as described below. The whole procedure is shown in Fig. 1.

The invariance to translation of the calculated parameters is guaranteed from the centering of the atomic coordinates and the invariance to rotation from the uniqueness of the PCA solution. It must be noted that, in the first version of the WHIM descriptors [9–12], the molecules have been centered on their baricenter (i.e. using the atom weights for the calculation of the center) and not in the center of the coordinates (i.e. firstly, centering the atomic coordinates and then calculating the weighted covariance matrix).

## 2.2. Directional WHIM descriptors

Directional WHIM descriptors are univariate statistical indices calculated from the scores of each individual principal component (1, 2, 3). The first group of descriptors are the *eigenvalues*  $\lambda_1$ ,  $\lambda_2$  and  $\lambda_3$  that are related to molecular size, the second group is constituted by the eigenvalue proportions  $\vartheta_1$ ,  $\vartheta_2$  and  $\vartheta_3$  that are related to molecular shape:

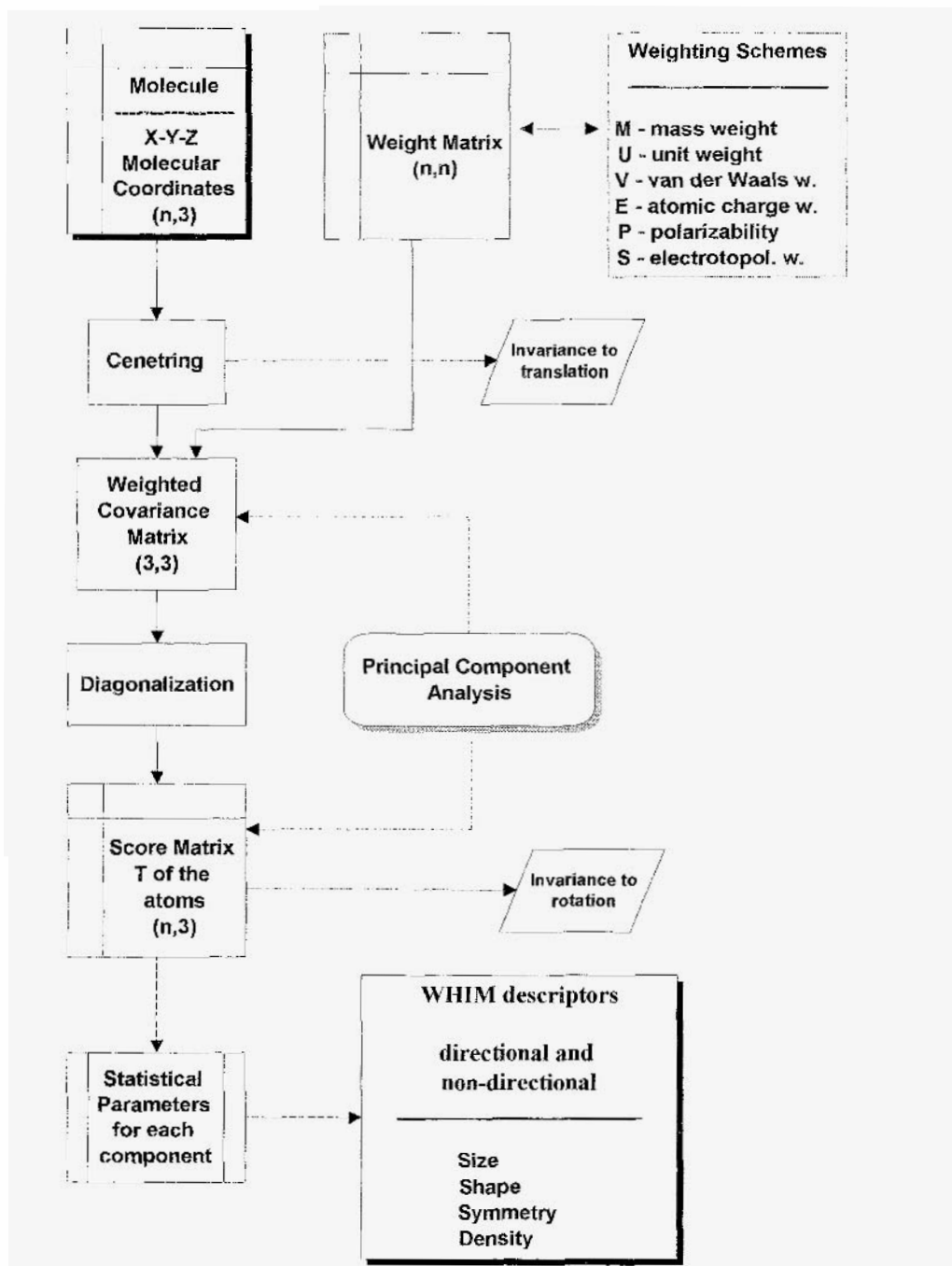


Fig. 1. Flow chart of the procedure for the calculation of the WHIM descriptors.

$$\vartheta_m = \frac{\lambda_m}{\sum_m \lambda_m} \quad m = 1, 2, 3 \quad (3)$$

Because of the closure condition ( $\vartheta_1 + \vartheta_2 + \vartheta_3 = 1$ ), only two are independent.

The third group of descriptors is constituted by the *symmetries*  $\gamma_1$ ,  $\gamma_2$  and  $\gamma_3$  calculated from an information content index on the symmetry along each principal component with respect to the center of the scores:

$$\gamma'_m = - \left[ \frac{n_s}{n} \cdot \log_2 \frac{n_s}{n} + n_a \cdot \left( \frac{1}{n} \cdot \log_2 \frac{1}{n} \right) \right] \quad \gamma_m = \frac{1}{1 + \gamma'_m} \quad 0 < \gamma_m \leq 1 \quad (4)$$

where  $n_s$ ,  $n_a$  and  $n$  are, respectively, the number of central symmetric atoms (along the  $m$ -th component), the number of non-symmetric atoms and the total number of atoms of the molecule.

Finally, the fourth group of descriptors is constituted by the inverse of the *kurtosis*  $\kappa_1$ ,  $\kappa_2$  and  $\kappa_3$ , calculated from the fourth-order moments of the scores  $t_m$ , that are related to the atom distribution and density around the origin and along the principal axes:

$$\kappa_m = \frac{\sum_i t_{im}^4}{\lambda_m^2 \cdot n} \quad \eta_m = \frac{1}{\kappa_m} \quad m = 1, 2, 3 \quad (5)$$

To avoid the problems related to the infinite (or very high) kurtosis values, obtained when along a principal axis all the atoms are projected in the center (or near the center, i.e. leptokurtic distribution), the inverse of the kurtosis is used.

Low values of the kurtosis are obtained when the data points (i.e. the atom projections) assume opposite values ( $-t$  and  $t$ ) with respect to center of the scores. When an increasing number of data values are within the extreme values  $\pm t$  along a principal axis, the kurtosis value increases (i.e.,  $\kappa = 1.8$  for a uniform distribution of points,  $\kappa = 3.0$  for a normal distribution). When the kurtosis value tends to infinity, the corresponding value tends to zero.

Thus, the group of descriptors  $\eta_m$  can be related to the *density* of the atoms distribution — i.e. to the quantity of unfilled space per projected atom — and has also been called *emptiness*: the greater the  $\eta_m$  values, the greater the projected unfilled space. The  $\eta_m$  descriptors are used in place of the kurtosis descriptors  $\kappa_m$  (previously proposed, see reference [11]).

### 2.3. Non-directional WHIM descriptors

The non-directional WHIM descriptors are directly derived from the directional WHIM descriptors. Thus, for non-directional WHIM descriptors, any information related to the principal axes disappears and the description is related only to a global — holistic — view of the molecule.

In many cases, size descriptors can, in modelling, play a significant role independently of the measured directions, allowing more simple models. Thus, in view of the

importance of this quantity, a group of descriptors of the total *molecular size* is considered in three different ways:

$$\begin{aligned} T &= \lambda_1 + \lambda_2 + \lambda_3 \\ A &= \lambda_1\lambda_2 + \lambda_1\lambda_3 + \lambda_2\lambda_3 \\ V &= \prod_{m=1}^3 (1 + \lambda_m) - 1 = T + A + \lambda_1\lambda_2\lambda_3 \end{aligned} \quad (6)$$

$T$  and  $A$  are, respectively, related to linear and quadratic contributions to the total molecular dimension;  $V$ , that contains also the third-order term, is the complete expression. Thus the molecular size is taken into account by all three size descriptors in different ways.

For each function, six total dimensions are obtained. one for each weighting scheme. The *molecular shape* is represented by the following expression:

$$K = \frac{\sum_m \left| \frac{\lambda_m}{\sum_m \lambda_m} - \frac{1}{3} \right|}{4/3} \quad 0 \leq K \leq 1 \quad (7)$$

The term  $4/3$  is the maximum value of the numerator term and is used to scale  $K$  between 0 and 1. This expression also has a more general meaning and has been proposed to evaluate the global con-elation in multivariate data [18].

For example. for an ideal straight molecule, both  $\lambda_2$  and  $\lambda_3$  are equal to zero and  $K = 1$ ; for an ideal spherical molecule, all three eigenvalues are equal to  $1/3$  and  $K = 0$ . For all planar molecules. the third eigenvalue  $\lambda_3$  is 0. there being no variance from the molecular plane. and  $K$  ranges between 0.5 and 1, depending on the linearity.

The  $K$  shape term definitely substitutes for the acentric factor. the former being more general than the previously proposed acentric factor  $\omega$  [9–12] and defined as  $\omega = \vartheta_1 - \vartheta_3$ .

The total *molecular symmetry* is defined as the following:

$$G = (\gamma_1 \cdot \gamma_2 \cdot \gamma_3)^{1/3} \quad (8)$$

where  $G$  is the harmonic mean of the directional symmetries. It equals 1 when the molecule shows a central symmetry along each axis and tends to 0 when there is a loss of symmetry along at least one axis. Different symmetry values are obtained only when unitary, mass and electrotopological weights are used: for this reason, only three kinds of symmetry parameters are retained;  $G_u$ ,  $G_m$  and  $G_s$ .

The *total molecular density* is represented by the following expression:

$$D = \eta_1 + \eta_2 + \eta_3 \quad (9)$$

The molecular descriptors defined above, due to their invariance to rotation and translation, are generalized molecular properties within each weighting scheme. Thus, for each weighting scheme, a total of 11 molecular directional descriptors ( $\vartheta_3$  was eliminated),

Weighted Holistic Invariant Molecular (WHIM) descriptors, can be obtained from each molecule geometry

$$\lambda_1 \lambda_2 \lambda_3 \vartheta_1 \vartheta_2 \gamma_1 \gamma_2 \gamma_3 \eta_1 \eta_2 \eta_3$$

The total number of directional WHIM descriptors is 66.

For planar compounds, the modelling requires only 8 descriptors for each weighting scheme:

$$\lambda_1 \lambda_2 \vartheta_1 \vartheta_2 \gamma_1 \gamma_2 \eta_1 \eta_2$$

As the molecular weight MW is the sum of the atomic weights of the molecule, similar additive quantities can be analogously defined from the other weights used for the calculation of the WHIM descriptors. Thus, the sum of the van der Waals volumes ( $S_v$ ), the sum of the electronegativities ( $S_e$ ), the sum of the polarizabilities ( $S_p$ ) and the sum of the electrotopological charges ( $S_s$ ), together with the molecular weight, can be added to the 33 non-directional WHIM. The sum of the unweighted atoms ( $S_u$ ) corresponds to the total number of atoms (NAT), but is not here considered among the WHIM descriptors. All these descriptors are independent of conformational (and configurational) geometries, but it is interesting to observe that  $S_v$  and  $S_p$  represent spatial additive properties, while MW represents an inertial additive property. As a whole, the non-directional WHIM are 5 for each weighting scheme (6):  $T$ ,  $A$ ,  $V$ ,  $K$ ,  $D$ , plus  $Gu$ ,  $Gm$ ,  $Gs$ .  $S_v$ ,  $S_e$ ,  $S_p$  and  $S_s$ , giving a total number of 37 descriptors.

#### 2.4. The meaning of the WHIM descriptors

To understand and explain the chemical meaning of WHIM descriptors, a dataset of 40 simple but heterogeneous molecules is used. The list of the 40 molecules is shown in Table 2 and the WHIM values can be obtained on request. Once calculated, the WHIM descriptors can be plotted, forming simple scatterplots that give a deeper insight into the descriptor chemical meaning.

In Fig. 2, the scatterplot of the 40 compounds is obtained from the variables  $Vm$  (size) and  $Km$  (shape) — i.e. descriptors calculated using the atomic masses as atom weights. The arrow A represents the increase in size and linearity ( $Km \approx 0.7-0.9$ ) of the five linear alkanes (1-5). Ethane (1) is more linear (higher  $km$  value than propane due to the absence of any central  $Csp_3$  carbon atom contribution along the second component. The same increase in size and linearity can be observed for cycloalkanes (11-15, arrow B), but, as expected, at a lower level of linearity ( $Km \approx 0.3-0.4$ ). The arrows C and D represent the increase in size and linearity of the halo-substituted benzenes (27-30, arrow C) and of the condensed benzenes (21, 39, 40, arrow D), respectively. The remaining mono-substituted benzenes are at intermediate positions. Arrow E highlights the large change in shape from neopentane (7, spherical shape) to 2-butyne (10, near-linear shape); note also the difference in shape between the two 2-butene isomers (8, *cis* and 9, *trans*) in this same direction. The alcohols (16-20) are

Table 2 List of the 40 compounds used in the example

ID	Compound	ID	Compound
1	Ethane	21	Benzene
2	Propane	22	Toluene
3	n-butane	23	Phenol
4	n-pentane	24	Benzoic acid
5	n-hexane	25	Aniline
6	Isobutane	26	Nitrobenzene
7	Neopentane	27	F-benzene
8	<i>cis</i> -2-butene	28	Cl-benzene
9	<i>trans</i> -2-butene	29	Rr-benzene
10	2-butyne	30	I-benzene
11	Cyclopropane	31	2-propanone
12	Cyclobutane	32	2-propanol
13	Cyclopentane	33	2-propylamine
14	Cyclohexane	34	2-fluoropropane
15	Cyclohexanone	35	2-iodopropane
16	Methanol	36	2-propanethiol
17	Ethanol	37	Methylamine
18	Trifluoroethmol	38	Dimethylamine
19	2-aminoethanol	39	Naphthalene
20	Propano	40	Anthracene

approximately aligned like the n-alkanes, the exception being the trifluoroethanol (18) because the masses of the three fluorine atoms substituting the hydrogen atoms give a tridimensional isotropic contribution. resulting in an increase in size and a significant decrease in linearity. Due to the iodine mass. increased linearity is observed for the 2-iodopropane (35), compared to the other 2-substituted propane isomers (31–36),

As shown in Figs. 3 and 4 of reference [13]. other interesting information about the molecular structure can be obtained by plotting, for example, shape ( $K$ ) versus symmetry ( $G$ ) descriptors, or densities descriptors weighted from masses and electrotopological charges ( $Dm$  versus  $Ds$ ). As shown in Figs. 6–9 of reference [13]. a principal component analysis of the 33 non-directional WHIM descriptors shows as the different sources of WHIM information are separated. The first six components explain the major part or the total variance (97%). where the first four represent individually molecular size. shape, symmetry anti density. while the last two components arc mainly related to single density descriptors.  $Dm$  and  $Ds$ . respectively.

### 3. Theory of G-WHIM descriptors

The algorithm proposed for the WHIM approach can be applied to any discretized image: in particular. this algorithm can be extended to deal with 3D grid-points. In such a case. the grid-points substitute for the atomic coordinates of a molecule and defined electronic or steric properties are used as weights. The descriptors derived from this approach are called Grid- eighed *H*olistic *I*nvariant *M*olecular descriptors (G-WHIM).



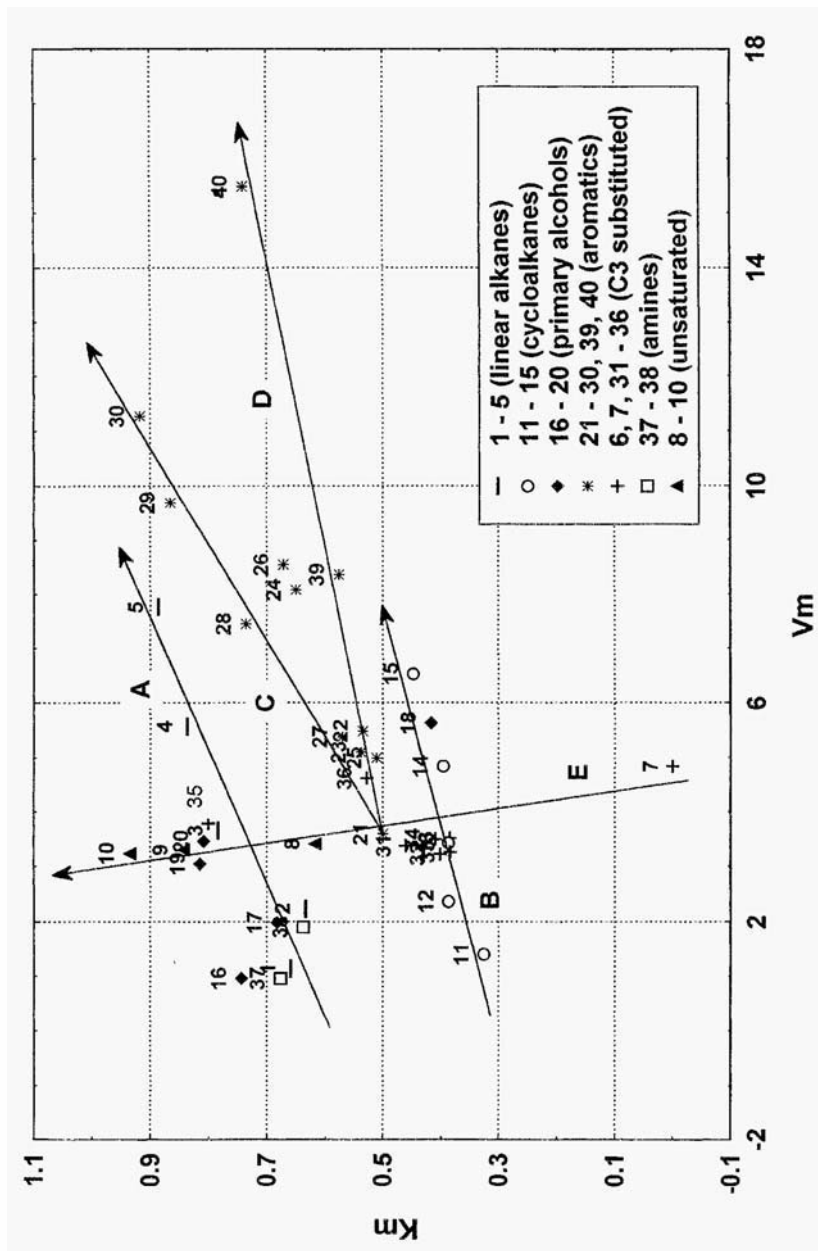


Fig. 2. Scatterplot of  $V_m$  (size) and  $K_m$  (shape) descriptors for the 40 molecules of Table 2.

The theory of the G-WHIM descriptors has been presented in reference [17]. In this review, a summary of its main elements is newly presented, updating symbols and descriptors on the base of the state-of-art of the WHIM theory presented in this chapter.

In principle, if a molecule is placed into an infinite, isotropic and evenly very dense grid, the scalar field  $F^3$  calculated at the grid-points must contain the same information, independent of the molecule's orientation and only depending on the potential energy of the selected probe and the mathematical functions representing the interaction. Thus  $F^3$  contains the whole information about the interaction properties of the molecule. In practice, this ideal situation cannot be obtained, but it can be simulated by plunging the molecule into a finite grid  $N^3$ : the aim is to represent the theoretical  $F^3$  scalar field by a finite sampling of this field.

Molecule position, grid dimensions and spacing between grid-points are crucial aspects to be defined in order to ensure that the obtained field is representative of the ideal one. In the simplest approach, the molecule must be placed in an isotropic grid — i.e. the obtained scalar field  $N^3$  is constituted of a finite number of points evenly distributed along the three dimensions. When a class of similar compounds is considered, the isotropic grid requirement can be relaxed and a grid of  $n_1 \times n_2 \times n_3$  points can be used if all the molecules are also oriented using a selected criterion. However, when the compared molecules are different in shape and size, this last approach to the grid definition cannot be considered because the scalar fields would then be evaluated with different sampling sizes in the three directions: thus, the resulting description is no longer unique and the invariance to rotation is not preserved. In any case, in order to avoid field distortions due to the truncation of the non-zero interaction values at the neighborhoods of the finite grid, the molecule must be centered in the grid.

Grid dimensions depend on the field selected as each field shows a different analytical dependence from the molecule-probe distance ( $r$ ). For example, when the non-bonding and electrostatic energy expressions are analyzed, the interaction energies decay with minus six and minus one power of  $r$ , respectively. In the first case, a grid very close to the van der Waals surface of the molecule is sufficient to contain the most informative points. In the second case, where the electrostatic energy decays very slowly with  $r$ , very large grids are required. To overcome this problem, an energy cutoff criterion can be proposed, for which only electrostatic energy values relevant for the considered interaction (e.g. long-range or chemical interaction) are taken into account. In this way, points far from the molecule and not contributing to the interaction are not included in the calculations.

In spite of the centering of the molecule, the selection of an isotropic grid and the energy cutoff definition, the invariance to rotation is, again, not preserved if the grid-points are not dense enough. This is a very important problem since a too sparse distribution of grid-points represents an inadequate sampling of the ideal scalar field and is not able to guarantee that the calculated scalar field is representative of the ideal scalar field in such a way as to preserve rotational invariance. Preliminary calculations of the grid step can be used to select the optimal step.

Once the optimal choices for the grid are selected, the G-WHIM descriptors are used to condense the whole information contained in the scalar field into a few global parameters, whose values are independent of the molecular orientation within the grid.

For each molecule, the G-WHIM descriptors are calculated by the following steps:

1. The molecule is freely and separately imbedded in the center of the grid.
2. The scalar field is calculated by using the selected probe.
3. The *scalar field values are used as weights for the grid-point coordinated*: this is the main difference between G-WHIM and WHIM descriptors. In fact, in the latter case, the coordinates are the spatial atomic coordinates, each weighted by one of the six different kinds of weights defined above: unitary weights  $U$ , atomic masses  $\mathbf{M}$ , van der Waals volumes  $\mathbf{V}$ , atomic electronegativities  $\mathbf{E}$ , polarizabilities  $\mathbf{P}$ , and electrotopological charges  $\mathbf{S}$ .
4. Finally, the G-WHIM descriptors are calculated in the same way as for the WHIM descriptors — that is, by the calculation of a weighted covariance matrix, principal component analysis and the calculation of statistical parameters on the projected points along each principal component (i.e. on the score values).

Point 3, above, deserves some further consideration. Firstly, it should be noted that only points with non-zero interaction energy are effective in the computation of the descriptors. Secondly, when the calculated interactions give both positive and negative values, the scalar field values cannot be used directly in this form as statistical weights, which must be always semi-positively defined. In this case, the scalar field values are divided in two blocks: a grid-negative (positive) block containing the grid coordinates associated with negative (positive) interaction values, getting their absolute values and setting the positive (negative) values to zero.

By this assumption, two sets of G-WHIM descriptors are obtained: the first describing the positive part of the molecular field, and the latter describing the negative one. Thus, for each weighting block (positive (+) and negative (-)), the G-WHIM descriptors consist of 8 directional plus 5 non-directional molecular descriptors (26 for a complete description of each interaction field), calculated from each molecule (point d):

$$\lambda_1 \lambda_2 \lambda_3 \vartheta_1 \vartheta_2 \eta_1 \eta_2 \eta_3, \text{ plus } T, A, V, K, D$$

The directional  $\eta$  and the non-directional  $G$  parameters, defined for WHIM descriptors and containing information about the molecular symmetry, are not considered in the frame of the G-WHIM approach because their meaning becomes doubtful, depending heavily upon the point sampling. However, the information regarding the molecular symmetry can be obtained by directly using the WHIM symmetry parameter.

The meaning of the G-WHIM descriptors is that previously defined for the WHIM descriptors, but now the descriptors are referred to the interaction field instead of the molecule. For example, the eigenvalues  $\lambda_1$ ,  $\lambda_2$  and  $\lambda_3$  will be related to the *interaction field size*; the eigenvalue proportions  $\vartheta_1$ ,  $\vartheta_2$  and  $\vartheta_3$  will be related to the *interaction field shape*; the group of descriptors constituted by the inverse function of the kurtosis ( $k$ ), i.e.,  $\eta_m = 1/\kappa_m$  will be related to the *interaction field density* along each axis. Moreover, global information about the interaction field is obtained from the non-directional WHIM ( $T, A, V, K, D$ ), with the meaning previously defined.

In the first chapter [17], the acentric factor has been used as shape descriptor — defined as  $w = \vartheta_1 - \vartheta_3$ , ranging between zero (spherical interaction field) and one

(linear interaction field): moreover, among the non-directional WHIM, only the eigenvalue sum  $T$  has been previously used.

### 3.1. Grid and points setting

The optimal grid size and point density must be evaluated by preliminary calculations on the largest molecule (or better, on the molecule with the broadest molecular field of the selected molecule set. The behavior of the G-WHIM descriptors with variation in the grid size and the point sharpness has been tested on a simple molecule such as the chlorobenzene, using the electrostatic potential to represent a very broad interaction.

To do this, a number of assumptions were made. The first assumption is to evaluate the chemical information due to the interaction outside the van der Waals surface: the field values in the inner part of the molecule have not been considered. The second assumption is to define different energy cutoffs: three values corresponding to  $-2$ ,  $-5$  and  $-10$  kcal/mol have been selected for this case. The higher the cutoff value, the smaller the considered region around the molecule (i.e. the total number of non-zero weighted grid-points).

After preliminary calculations, the grid size was fixed as a 20 Å edged cube in order to ensure the inclusion of the smallest cutoff surface. The steps defining the field points were fixed at 0.10, 0.25, 0.50, 0.75 and 1 Å, respectively. For each of the 15 cases (3 cutoffs  $\times$  5 steps), the G-WHIM descriptors were calculated: the values obtained are reported in table 1 of reference [17].

Figure 3 shows the trend of  $T$  (the eigenvalue sum) as a function of the step for each cutoff value; all the other descriptors show similar behavior. As can be seen, the  $T$  parameter for each cutoff value tends to become constant as the step decreases, converging towards the most reliable value, corresponding to the smallest step (0.1 in this case). A single calculation at step 0.05 and cutoff  $-5$  kcal/mol confirms the obtained convergence.

The step for which the parameter values show only small differences represents the optimal step for which  $N^3 \rightarrow F^3$ ; this also means that using smaller steps does not change the G-WHIM parameter values. For cutoff values very close to the minimum energy value (in this case, lower than  $-10$  kcal/mol), the sampling of the interaction field can be unreliable because too few points are considered to represent the great energy variability of the field. The parameters obtained for each cutoff value are different, representing three different chemical situations, as can be easily observed from the  $T$  values: a relatively long-range interaction ( $-2$  kcal/mol) shows a higher  $T$  value (greater total field dimension), while a relatively short-range interaction ( $-10$  kcal/mol) shows a lower  $T$  value (smaller total field dimension).

### 3.2. Invariance to rotation

The invariance to rotation is a point of utmost importance in order to avoid problems typical of other QSAR strategies, such as molecular alignment. To check the invariance to rotation, the electrostatic potential of chlorobenzene was used, as above. The follow-

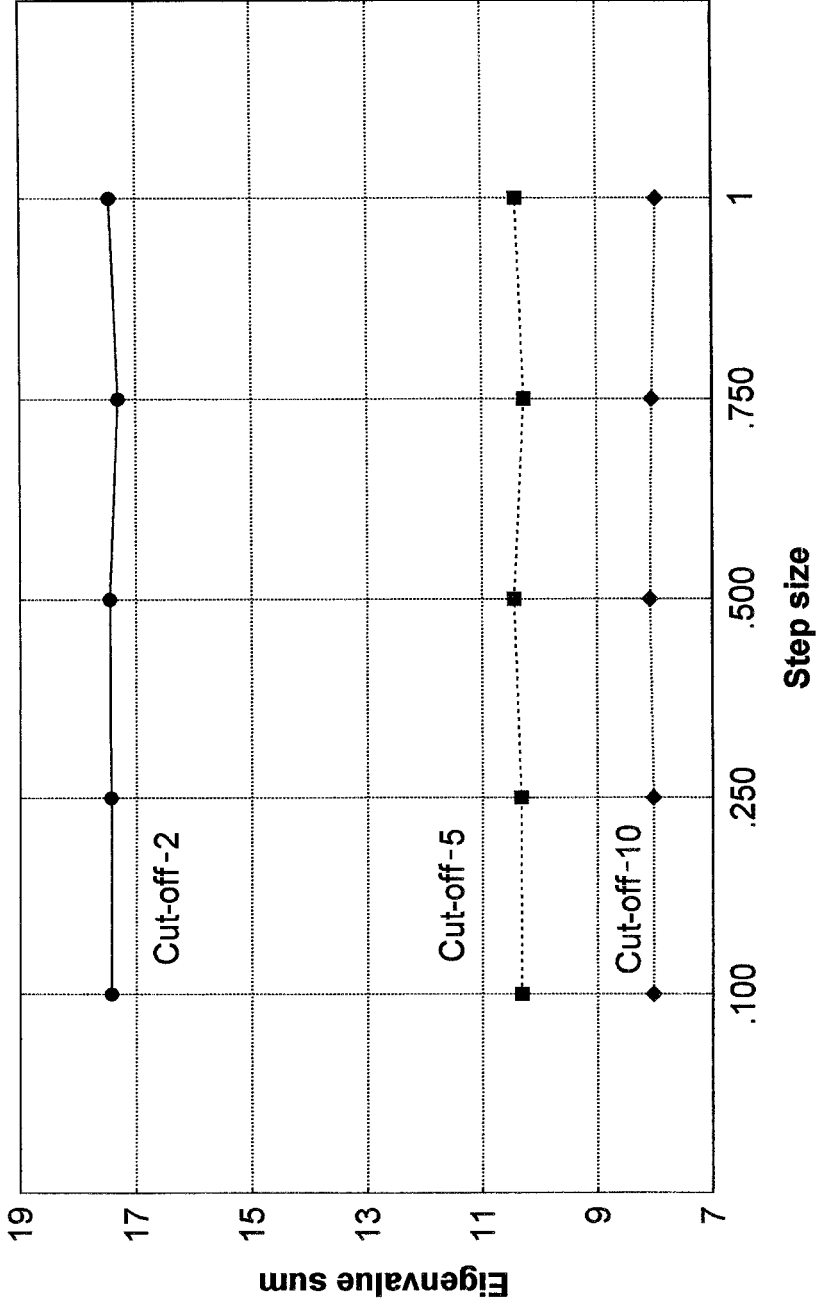


Fig. 3. Values of the eigenvalue sum ( $T$ , size) for different step sizes calculated for three cutoff values.

ing calculations were performed: the G-WHIM parameters were calculated for the same cubic isotropic grid for 5 different steps (0.1, 0.25, 0.5, 0.75 and 1.0 Å, respectively) and for 21 different orientations of a centered molecule.

Figure 4 shows the values of the eigenvalue sum ( $T$ ) for the 21 molecule rotations, for different step sizes. For this case, note the stabilization due to reduction of step size, the value of the considered descriptor clearly shows good invariance with respect to rotations at step = 0.5: for all the descriptors, the standard deviations are less than 0.075.

The values of the standard deviations of each descriptor, calculated from the descriptor values obtained for the 21 molecule rotations, decrease as the step size decreases, confirming the expected trend observed above. The 3 eigenvalues ( $\lambda$ ) and their sum ( $T$ ), represented by numbers not limited to between 0 and 1 (as for other descriptors), show standard deviations higher, on average, than those of the other descriptors.

Thus, as expected from the theory, for a representative sampling of the interaction field — i.e. when the step size is small enough — G-WHIM parameters independent of the molecular orientation within the grid space can be obtained. Similar behavior (or even better) is shown by all the other G-WHIM descriptors. Moreover, the results obtained in the case of chlorobenzene have also been confirmed in the case where phenol was used as the test molecule and 4 different cutoff values (0.5, 1.0, 1.5 and 2.0 Å) were used.

### 3.3. Perspectives of the G-WHIM approach

From the theoretical point of view, the G-WHIM approach to QSAR problems appears very promising, integrating the information contained in the WHIM descriptors and overcoming any problems due to the alignment of the different molecules and the explosion of variables arising from traditional grid approaches. In particular, the G-WHIM approach can take into account both all the points within the cutoff values, excluding only the positive interactions within the inner part of the molecule and the surface points at a cutoff value — i.e. the points on the iso-interaction-energy surfaces.

The ability to define different molecular interaction regions, taking into account the individual parameters provided from different cutoff values, is undoubtedly a fascinating possibility which may lead to a deeper chemical insight into molecular interactions and properties.

## 4. Descriptors

The WHIM descriptors defined above can be used separately in modelling: the set of 37 *non-directional WHIM* and the set of 66 *directional WHIM*. For the sake of comparison, other sets of theoretical descriptors are sometimes used, either as a single set or jointly with WHIM descriptors. In particular, the software produced by our research group (WHIM-3D/QSAR, see below) calculates two other sets of descriptors: the first is constituted by the so-called *structural descriptors* (30) — i.e. the number of different kinds of atoms (e.g. nH, nC, nF and nX are the number of hydrogens, carbons, fluorines, halogens, respectively), the number of bonds (nBO), the number of some functional groups

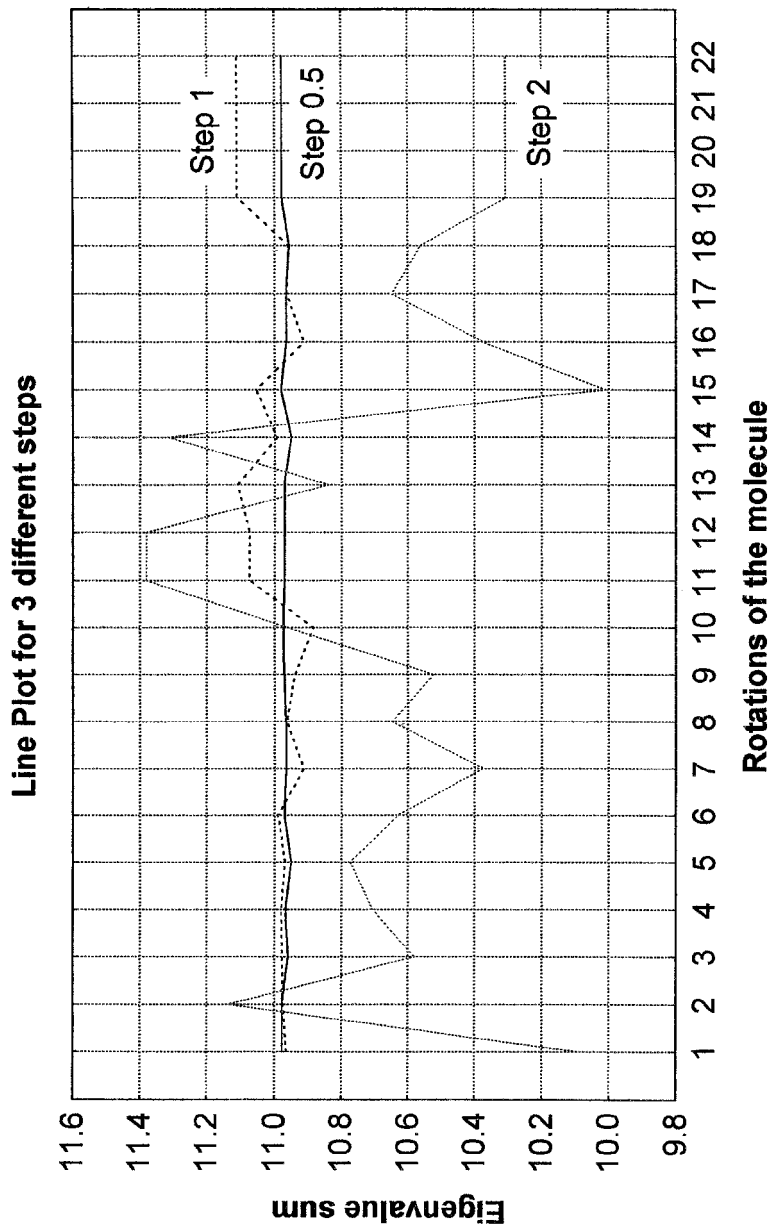


Fig. 4. Values of the eigenvalue sum ( $T$ , size) for 21 different rotations of the test molecule (chlorobenzene), for different step sizes

(e.g. nOH, nCO and nNH<sub>2</sub> are the number of OH, CO and NH<sub>2</sub> groups, respectively) and the number of rings of different size (nR03, ..., nR10 are the number of 3-membered to 10-membered rings). Moreover, the number of atom acceptors and donors of H-bonds (nHA and nHD) are also considered. The second set is constituted by the more frequently used 34 *topological descriptors* (information and connectivity indices), as defined in reference [19].

To all the sets of descriptors, the molecular weight (MW) has also been added. The G-WHIM descriptors, as defined above, have been used in some cases. The descriptors provided from theories, like topological or WHIM, are numerous and are usually internally Correlated. Chemometric strategies that use variable subsets selection procedures (e.g. genetic algorithms or simulated annealing) and model validation techniques play a key role in obtaining predictive and stable models in QSAR studies (see section 5 on Methods).

## 5. Methods

The minimum energy conformations of all the compounds were obtained by the molecular mechanics method of Allinger (MM2), using the package *HyperChem* [20]; WHIM descriptors were calculated from the obtained coordinates using our package *WHIM-3D/QSAR* for WINDOWS/PC (now available on request [21]). This package has been extended to calculate also structural and topological indices. A special version has also been extended to the calculation of the G-WHIM descriptors, reading the grid coordinates and the grid-point property values for each molecule. Principal Component Analysis (PCA) has been performed by *STATISTICA* [22]. The *HyperChem/Chemplus* package [20] has been also used to calculate some simple physico-chemical properties.

The selection of the best subset variables (variable subset selection, VSS method) for modelling the selected properties was through the genetic algorithm (GA-VSS) approach [23] where the response is obtained by ordinary least-squares regression (OLS), using our package *Moby Digs* for variable selection for WINDOWS/PC [24].

All calculations were performed by *Moby Digs* using the leave-one-out procedure of cross-validation, maximizing the cross-validated *R*-squared ( $Q^2$ ). In many cases, more demanding validation procedures were also used and always when models with highly correlated variables had been selected by GA-VSS. In these cases, when possible, an exhaustive leave-more-out procedure — i.e. leaving out *g* objects in all the possible ways — was used; otherwise, a random selection of *g* objects to be left out was repeated some thousands of ways. For the obtained models, all variables are highly significant within the 95% confidence level.

## 6. QSAR applications

In the gaining of confidence with the WHIM approach to QSAR problems, a number of datasets have been used and discussed in previous papers [9–12,14], testing the prediction capabilities for several responses, either physico-chemical properties or biological/toxicological activity responses.

Owing to the importance of the surface of a molecule for the determination of various physical, thermodynamic, toxicological, biological and transport properties, the WHIM



descriptors were successfully applied [9] to calculate the total surface area (TSA) of 101 heterogeneous compounds. On a small and homogeneous class of contaminants, 12 *chlorobenzenes* that are widely distributed throughout the environment, properties of very different kinds have been modelled, such as melting and boiling points, solubility, hydrophobicity, bioconcentration factor, toxicity on algae and Microtox. In all cases, high prediction powers resulted: as is already known, size parameters play the main modelling role for all these properties but WHIM descriptors of molecular shape, atomic distribution and entropic contribution from symmetry add useful information [14]. Similar results were obtained in the modelling of *chlorophenol* toxicity on seven different biosensors [12].

On a large dataset of environmentally relevant *polycycle aromatic hydrocarbons* (PAH, 82 compounds), some important, wide-range, physico-chemical properties such as melting point, boiling point and hydrophobicity have been successfully modelled [10], with good prediction powers. A study was made of highly heterogeneous compounds [11], belonging to the EEC priority list of dangerous chemicals (amines, chlorobenzenes, organotin and organophosphorous compounds; 49 compounds) with different toxic actions: the toxicity on *Daphnia magna* and hydrophobicity was modelled, leading also to general models and predictions for external compounds. WHIM descriptors used with topological indices were the subject of a recent work [16] on an extended class of 118 *environmental priority chemicals* dangerous to the aquatic environment, selected by the European Union according to the directive 76/464/EEC and included in the so-called 'List 1'. In particular, physico-chemical properties (melting point, boiling point, density, refraction index, solubility and hydrophobicity) and toxicological properties (algae, bacteria, *Daphnia*, fish and mammals) were modelled with mixed approaches (WHIM and other sets of descriptors). In the reported paper [16], a new hydrophilicity index (Hy), showing good modelling power, was also defined.

A comparison between WHIM and topological descriptors was recently performed [15] on a wide class of *haloaromatic compounds* (73 halobenzenes and 89 halotoluenes). In all cases, the WHIM descriptors gave better predictive results in modelling melting, boiling and flash points and density.

An extensive comparison of the latest developments of the WHIM descriptors with other different QSAR approaches has been performed on the datasets reported above. With regard to the G-WHIM descriptors, an application of a raw modification [25] of the original G-WHIM version [17] has been applied to the classical dataset of 31 *steroids* studied by Cramer et al. [26]. Despite some G-WHIM descriptors having been calculated roughly, the results are encouraging compared with the original Cramer calculations and later CoMFA refinements.

The G-WHIM approach has been improved in analogy with the WHIM approach and has been further tested for the toxicity of 14 dioxins.

### 6.1. Chlorophenols

In a previous paper [12], an extensive study was made of chlorophenol toxicities, assessed by different biosensors (Microtox, bacteria, *Daphnia*, algae, guppy, flounder

and SMP), with good results:  $Q^2$  between 95% and 97% for almost all the responses. In the same paper,  $\log K_{ow}$  was also modelled with size descriptors such as MW,  $Tu$  and  $Te$  ( $Q^2 = 94.5$  and  $R^2 = 96.1$ ): however, this model significantly decreases its predictive power ( $Q^2 = 80.8$ ) when cross-validated with a perturbation of 30%. Instead, a satisfactory and stable two-variable model ( $Vu$  and  $Ve$  size descriptors) was recalculated from the complete set of non-directional WHIM descriptors ( $Q^2 = 94.1$  and  $R^2 = 95.9$ ). This model is perfectly stable when 30% of the objects is left out ( $Q^2 = 93.3$ ), in spite of the great correlation between  $Vu$  and  $Ve$ .

A local property like pKa, difficultly modelled by global descriptors, is instead better modelled by directional WHIM descriptors with  $Q^2$  ranging from 76% and 78% for the three-, four- and five-variable best models. In these models, the most selected variables are those related to directional size, atomic density and symmetry, which in some way represent the relative position or the substituents responsible of the phenol group acidity. For example, the three-variable model, constituted by  $\lambda_{2v}$  (size),  $\eta_{1m}$  (density) and  $\gamma_{2v}$  (symmetry), is:

$$n = 20 \quad p = 3 \quad Q^2 = 75.7 \quad R^2 = 82.7 \quad (10)$$

The G-WHIM approach has also been applied to these compounds to verify, for the first time, the applicability of these new descriptors [17]. In this case, the WHIM and G-WHIM descriptors were successfully used jointly to improve on the results obtained using only WHIM descriptors, on  $\log K_{ow}$ , pKa, toxicity on flounder and melting point.

## 6.2. *N,N*-dimethyl-2-Br-phenethylamines

To predict the antagonism of 22 *N,N*-dimethyl-2-Br-phenethylamines to epinephrine in the rat ( $\log 1/ED_{50}$ ), a comparison between Hansch and Free-Wilson approaches was performed by Kubinyi [27]. A more extended comparison, using also topological and WHIM descriptors, has been performed and several models have been evaluated by using the last WHIM descriptor version [14] and comparing the results with those of the CoMFA approach [28]. A first set of models, based on substituent descriptors, has been proposed by Unger and Hansch [29]. In particular, a two-variable model (ID = 10) is constituted by the variables  $\sigma^+$  and  $\pi$  (Hammett electronic and hydrophobic constants, respectively). A three-variable extension of this model (ID = 7) corresponds to the regression model:

$$\log(1/ED_{50}) = 7.058 - 1.015\sigma^+ + 0.621r_p + 0.828\pi \quad (11)$$

where  $r_p$  is the van der Waals radius from the *para* position of the substituent. The GA-VSS approach applied to the whole set of Hansch descriptors confirms these two models.

The Free-Wilson approach (ID = 14), based on 10 binary descriptors (2 sites  $\times$  5 substituents), has also been used [30]. In spite of its good fitting performance, a check of the predictive power of this model gives a prediction power of zero! This result is not

unusual for the Free-Wilson approach when predictive performances are searched for and it is not used for sorting group contributions.

Other groups of descriptors have been applied to this dataset: structural, topological indices, directional and non-directional WHIM. used either separately or jointly. Using the topological descriptors, as defined in reference [19], the GA-VSS approach led to a three-variable regression model (ID = 9):

$$\log(1/ED_{50}) = -3.429 + 0.006 MW - 2.822 AAC + 6.773 ICEN \quad (12)$$

where *MW*, *AAC* and *ICEN* are the molecular weight, the average atomic composition index and the centric index, respectively.

Using the non-directional WHIM descriptors, a three-variable model, based on size and shape descriptors, is obtained (ID = 4):

$$\log(1/ED_{50}) = -4.133 + 0.599 Tv + 0.263 Te + 4.981 Ks \quad (13)$$

Using the directional WHIM descriptors, the results concerning the best three- and four-variable models are (ID = 5 and ID = 3):

$$\log(1/ED_{50}) = 8.781 + 4.235 \eta_1 v + 0.711 \lambda_1 e - 47.564 \gamma_2 p \quad (14)$$

$$\log(1/ED_{50}) = 6.233 + 4.316 \eta_1 v + 0.736 \lambda_1 e - 43.783 \gamma_2 p + 1.973 \vartheta_1 s \quad (15)$$

Finally, structural descriptors (only 11 are not constant for this dataset) have also been tried: the best model (ID = 13) is a three-variable model with *MW*, *NAT* (total number of atoms) and *nCl* (number of chlorine atoms), giving a  $Q^2 = 62.2$ .

For both topological and WHIM models, the dependence of the response on size (*MW*, *AAC*, *Tv*, *Te* or  $\lambda_1 e$ ) and shape (*ICEN*, *Ks* or  $\vartheta_1 s$ ) is confirmed. For the WHIM models, the presence of descriptors depending on electronegativity (*e*) and electrotopological (*s*) weights highlights the importance of electronic parameters, in accordance with  $\sigma^+$  of the Hansch model, in which the importance of molecular size is also confirmed ( $\pi$ ,  $r_p$ ). However, molecular weight alone is unable to give a predictive model ( $Q^2 = 8.0$ ,  $R^2 = 25.1$ ), as can be expected for structural isomers with different values of biological activity.

Mixed models have also been tried by using jointly structural, topological and both sets of WHIM descriptors. In this case, the sequence of the best models was calculated from one- to four-variable models (ID = 12, 6, 2, 1). Table 3 shows the more interesting models obtained from the different approaches, sorted on the  $Q^2$  values.

The descriptors 1K and 2K are the topological shape indices of first- and second-order as defined by Kier [31,32]; the remaining descriptors are directional and non-directional WHIM.

As can be observed, a good model (ID = 12) is already obtained with the global dimensional variable *Tv*; a significant improvement is obtained by adding  $\lambda_2 p$  (ID = 6),

Table 3 Comparison of the best models obtained by different QSAR approaches on the biological activity of 22 variously substituted *N,N*-dimethyl-2-Br-phenethylamines

ID	QSAR approach	Size	Q <sup>2</sup>	R <sup>2</sup>	Model descriptors
1.	Mixed (Struct./Top./WHIM)	4	97.1	98.1	2K $\vartheta_{1s}$ $\eta_{1s}$ Tc
2.	Mixed (Struct./Top./WHIM)	3	96.2	97.4	1K $\vartheta_{1s}$ $\eta_{1s}$
3.	Directional WHIM	4	95.5	97.1	$\eta_{1v}$ $\lambda_{1e}$ $\gamma_{2p}$ $\vartheta_{1s}$
4.	Non-directional WHIM	3	94.7	96.7	Tv Tc Ks
5.	Directional WHIM	3	94.0	95.8	$\eta_{1v}$ $\lambda_{1e}$ $\gamma_{2p}$
6.	Mixed (Struct./Top./WHIM)	2	91.5	93.6	Tv $\lambda_{3p}$
7.	Hansch	3	88.4	92.9	$\sigma^+$ $r_p$ $\pi$
8.	Non-directional WHIM	2	88.2	92.0	Tv Dp
9.	Topological	3	84.5	89.8	MW AAC ICEN
10.	Hansch	2	84.1	89.2	$\sigma^+$ $\pi$
11.	CoMFA	2 <sup>a</sup>	80.5	91.3	Lattice points
12.	Mixed (Struct./Top./WHIM)	1	79.4	83.2	Tv
13.	Structural	3	62.2	74.0	MW NAT nCl
14.	Fracc-Wilson	10	0	93.9	Sites $\times$ substituents

<sup>a</sup> Number of significant principal components.

which takes into account out-of-plane substituent contributions. Alternatively, the out-of-plane information can be represented by shape descriptors (*Ks* or  $\vartheta_{1s}$ ).

For these compounds, the CoMFA approach was also applied using steric and electronic fields. The common substructure features of these compounds greatly reduce alignment problems: molecular superpositions were performed by minimizing the *rms* distance between all the common heavy atoms of the considered compounds. CoMFA was carried out using the QSAR option of SYBYL 6.2 [33]. The steric and electrostatic probe–ligand interaction energies were calculated with Lennard-Jones and Coulomb potential functions within the Tripos force field, using a *Csp*<sup>3</sup> probe atom with a charge of +1. The dimension of the CoMFA lattice was determined through the provided automatic procedure in order to extend the lattice walls beyond the dimensions of each structure by 4.0 Å in all directions. The lattice spacing established was 2.0 Å. The steric energies were truncated at 30 kcal/mol and the electrostatic ones dropped to within the steric cutoff for each molecule. The regression model (ID = 11) was obtained using PLS, as implemented in SYBYL 6.2 and validated with the leave-one-out procedure. The optimal number of cross-validated PLS components is two and the corresponding prediction power is  $Q^2 = 80.5$ , a value comparable with the simplest WHIM model based only on the size descriptor Tv ( $Q^2 = 79.4$ ).

### 6.3. Dioxins and analogs

On a dataset of 73 polyhalogenated aryl derivatives [34], WHIM models were searched for 71 pRB and 69 pAHH responses. The biological response pRB is  $pRB = -log$

Table 4 Comparison of the best models obtained by different QSAR approaches on the pRB and pAHH biological responses of dioxin analogues

Response (approach)	Size	Q <sup>2</sup>	R <sup>2</sup>	Model descriptions
pRB (WHIM)	3	81.9	83.5	Tm Ts Vu
pRB (MTD)	1	n.a.	70.9	MTD
pAHH (WHIM)	4	66.8	71.8	Se Te De Vu
pAHH (WHIM)	3	64.3	68.0	Tm Vs Du
pAHH (WHIM)	2	62.7	66.6	Se Te
pAHH (MTD)	1	n.a.	66.4	MTD

n.a.; not available.

EC<sub>50</sub>(RB), where EC<sub>50</sub>(RB) values are the *in vitro* rat hepatic cytosolic Ah receptor binding affinities. The pAHH response is  $pAHH = -\log EC_{50}(AHH)$ , where EC<sub>50</sub>(AHH) values correspond to the *in vitro* induction of aryl hydrocarbon hydroxylase (AHH).

A close dependence of the responses on molecular size seems strongly suggested both by the WHIM models [14] and the models obtained with Minimal Topology Difference (MTD) approach [34], which is also related to the size of a molecule (Table 4). For both responses, models using topological descriptors, as defined in reference [19], were also tried, but the results obtained were unacceptable.

These 72 compounds were also used in a preliminary study of the relationships between WHIM descriptors and properties calculated using some well-known models largely used in commercial packages. For this purpose, the *HyperChem/Chemplus* package [20] was used to calculate some simple physico-chemical responses: total surface area (TSA, [35]), molar volume (Vm, [35]), molar refraction (MR, [36]), polarizability (Pol, [37]) and octanol–water partition coefficient (logK<sub>ow</sub> or ClogP, [36]). The compounds' properties were calculated in the *Chemplus* approaches on the same optimized molecular structures used for the calculation of the WHIM descriptors. Table 5 collects Q<sup>2</sup> values (leave-one-out procedure) for some selected models (common to the five studied properties).

The results are quite surprising, showing very high prediction powers for all the considered properties by one- and two-variable WHIM models. Moreover, all the properties

Table 5 Q<sup>2</sup> values (leave-one-out procedure) of the selected non-directional WHIM models for five calculated physico-chemical properties of 73 dioxin analogues

Size	Variables	TSA	Vm	MR	Pol	ClogP
1	Sp	97.7	99.2	99.2	97.1	84.3
1	Sv	97.0	98.9	98.8	97.5	80.3
2	Sp Ts	98.6	99.5	99.2	97.6	84.9
2	Sv Tu	97.1	99.1	99.4	99.3	95.4
2	Sv Dv	98.6	99.4	98.8	97.7	88.9

(with the exception of ClogP, in part) are ultimately modelled by a small number of similar-size WHIM descriptors, highlighting a general (and expected) correlation between these properties and size parameters.

If the quality of the obtained regression models will be confirmed for more heterogeneous compounds, at least the first four properties could be predicted by simple global models using the WHIM approach, avoiding problems due to limited parameterizations or time-consuming algorithms.

From this dioxin analogue dataset, the subset of 14 *PolyChloroDibenzoDioxins* (PCDD), for which the toxicity measured as Ah receptor binding affinity (pRB) is known, has also been studied separately. Regression models have been calculated using different QSAR approaches and different sets of descriptors: the Free-Wilson approach [30], the topological indices [19], the values of Molecular electrostatic potential (MEP) at selected points [38] and the WHIM and G-WHIM descriptors recalculated in the version proposed here. The results can be seen in Table 6.

The three topological indices (BAL, IDM and WIA) are the Balaban index, the information index on the magnitude of distances and the average Wiener index, respectively. The model based on MEP was obtained on the basis of a variable selection performed on a few MEP values chosen in the 3D distribution of this property and used as molecular descriptors.

For the two WHIM models of Table 6, a more demanding validation procedure (20% of perturbation, 3 objects left out, exhaustive calculations) gives  $Q^2 = 89.8$  and  $Q^2 = 84.2$ , respectively. For these compounds, the G-WHIM approach has also been used [39]. The molecular electrostatic potential was calculated in all the grid-points between the van der Waals surface and a predefined iso-potential surface (i.e. a threshold surface). The selected threshold values are  $-4$  and  $+10$  kcal/mol for the negative and positive parts of the potential distribution, respectively.

26 G-WHIM descriptors were used, 13 for each part of the potential distribution and the two best models found by the genetic algorithms are reported in Table 6. The prediction powers of these models are exceptionally high; moreover, an exhaustive validation (20% of perturbation, i.e. 3 objects left out at each step) was also performed. In both cases, the obtained leave-more-out  $Q^2$  values remain high:  $Q^2 = 98.2$  and

Table 6 Comparison of models obtained by different QSAR approaches on the toxicity of 14 polychlorinated dibenzodioxins (PCDD)

QSAR approach	Size	$Q^2$	$R^2$	Model descriptors
G-WHIM	5	98.6	99.5	$\eta_1(-)$ T(-) D(-) $\lambda_3(+)$ K(+)
G-WHIM	4	96.3	98.6	T(-) D(-) $\lambda_3(+)$ K(+)
WHIM (dir. and non-dir.)	4	92.5	96.6	$\eta_1u$ $\lambda_{1,m}$ Tu As
WHIM (non-dir.)	2	85.4	88.7	Dv Vs
MEP	3	76.0	87.0	MEP points
Topological indices	3	62.6	79.8	BAL IDM WIA
Free-Wilson	8	43.7	78.6	Eight substituent positions

$Q^2 = 95.7$ , respectively. From the analysis of the MEP distributions of the model variables and of their standardized regression coefficients, high binding affinity is determined by a great global extension of the negative distribution ( $T(-)$ ), with a small portion of 'empty space' between the separated negative distributions ( $D(-)$  and  $\eta_1(-)$ ) and the spherical shape of the positive distribution ( $K(+)$ ), together with a little extension of the positive distribution out of the molecular plane ( $\lambda_3(+)$ ).

The G-WHIM descriptors efficiently condense/extract the information contained in MEP distributions leading to good quantitative models of the relationships between electrostatic properties and binding affinity pRB of the PCDD. These 14 PCCDs, included in a dataset of 25 polyhalogenated dioxins, have also been studied by Waller with the CoMFA approach [40]. In this case, the result was:

$$n = 25 \quad Q^2 = 71.5 \quad R^2 = 91.9 \quad (16)$$

## 7. Conclusion

The complexity of chemical phenomena today calls for new and complementary terms to explain such phenomena. In fact, one must be aware that, for example, the biological activity of a molecule, as well as several physico-chemical properties, are, in many cases, the result of complex interactions between the molecule and its chemical, physical and biological medium. Because of this involved complexity, high prediction capabilities cannot usually be expected from very simple relationships between the response and only a few independent descriptors.

WHIM descriptors, obtained from different weighting schemes, can be viewed as an adaptive descriptor space, containing both global and directional information (spread along orthogonal axes) that, in many cases, seems able to capture this complexity. In any case, WHIM descriptors give a holistic representation of the molecule, whereas reductionistic approaches (chemical interpretation in terms of local properties, functional groups, additive schemes based on molecular fragments or on atomic types, etc.) are rather limited. These descriptors are the first 3D molecular descriptors that are invariant to rotation and translation, thus avoiding any molecule alignment problem. The algorithms for their calculations are very simple and are not time-demanding. Unlike topological descriptors, WHIM descriptors are able to distinguish different conformations of the same molecule and, obviously, different geometric isomers.

Their interpretability is discussed in this chapter and their meaning in relation to 3D structural characteristics of molecules is highlighted, although a deeper insight is still needed to better understand their meaning and correlation with several physico-chemical properties. For example, the peculiar role of the symmetry descriptors (Gw) in modelling the melting point is always confirmed for all the studied cases.

As highlighted in the studied cases, the high modelling power of WHIM descriptors to face QSAR problems has been confirmed, although the modelling power of directional WHIM descriptors needs further investigation for cases where they may be of advantage for the correlation of responses with local properties.

The G-WHIM approach appears to be a powerful tool for the future of QSAR studies — in particular, in the pharmacological field. In fact, it overcomes problems due to the

alignment of different molecules and to the explosion of variables arising from traditional grid approaches, based on interaction energy fields.

Finally, encouraging signs of the wide applicability and the modelling power of the WHIM approaches are constantly coming from the QSAR studies of our research group. In fact, studies on the gas-chromatographic relative retention time, physico-chemical properties such as Henry's constant, total surface area, melting point, solubility, aqueous activity coefficients and hydrophobicity of 209 PolyChloroBiphenyls (PCB) are in progress with good preliminary results [41]. Research on phytotoxicity of triazine derivatives, classification of aquatic pollutants according water quality objectives (WQO) [42], atmospheric reactivity (with OH\*, NO<sub>3</sub> and O<sub>3</sub>) of several organic chemicals released into the environment [43], environmental behavior of ChloroFluoroCarbons (CFC) and HydrogenChloroFluoroCarbons (HCFC) [44] and biodegradability of organic pollutants [45] have given encouraging preliminary predictive models. Moreover, the WHIM approach has also been applied to the classification of 152 organic solvents. As a result of this, a new general classification based on multivariate analysis has been proposed [46].

## Acknowledgements

We wish to thank, for their contributions to this review: Drs. Laura Bonati, Ugo Cosentino, Marina Lasagni, Giorgio Moro and Professor Demetrio Pitea (Department of Physical Chemistry, University of Milan, Italy), Drs. Alessandro Maiocchi (BRACCO spa, Milan, Italy) and Natalia Navas (University of Granada, Spain).

## References

1. Horvath, A.L., *Molecular Design*. Elsevier, Amsterdam, 1992.
2. Karcher, W. and Devillers, J. (Eds.), *Practical Applications of Quantitative Structure-Activity Relationships (QSAR) in Environmental Chemistry and Toxicology*, Kluwer, Academic Publications, Dordrecht, 1990.
3. Hansch, C. and Leo, A., *Exploring QSAR*, American Chemical Society, Washington D.C., 1995.
4. Mercy, P.G. (Ed.), *Mathematical Modeling in Chemistry*. VCH, Weinheim, 1991.
5. Kier, L.B. and Hall, L.H., *Molecular Connectivity in Structure-Activity Analysis*. Research Studies Press, Letchworth, U.K., 1986.
6. Bonchev, D., *Information Theoretic indices for Characterization of Chemical Structures*, Research Studies Press, Letchworth, U.K., 1983.
7. Sabljic, A., *Topological indices and environmental chemistry*, In Karcher, W. and Devillers, J. (Eds.) *Practical applications of quantitative structure-activity relationships (QSAR) in environmental chemistry and toxicology*, Kluwer Academic Publications, Dordrecht, 1990. pp. 61-82.
8. Basak, S.C., *A nonempirical approach to predicting molecular properties using graph-theoretic invariants*, In Karcher, W. and Devillers, J. (Eds.) *Practical applications of quantitative structure-activity relationships (QSAR) in environmental chemistry and toxicology*. Kluwer Academic Publishers, Dordrecht, 1990. pp. 83-103.
9. Todeschini, R., Lasagni, M. and Marengo, E., *New molecular descriptors for 2D and 3D structures theory: Part I*, *J. Chemometrics*, 8 (1994) 263-272.
10. Todeschini, R., Gramatica, P., Provenzani, R. and Marengo, E., *Weighted holistic invariant molecular descriptors: Part 2. Theory development and application on modeling physico-chemical properties of polyaromatic hydrocarbons*, *Chemom. Intell. Lab. System.*, 27 (1995) 221-129.



11. Todeschini, R., Vighi, M., Provenrani, R., Finizio, A. and Gramatica, P., *Modeling and prediction by using WHIM descriptors in QSAR studies: Toxicity of heterogeneous chemicals on Daphnia magna: Part 3*, Chemosphere, 32 (1996) 1527–1545.
12. Todeschini, R., Bettiol, C., Giurin, G., Gramatica, P., Miana, P. and Argese, E., *Modeling and prediction by using WHIM descriptors in QSAR studies: Submitochondrial particles (SMP) as toxicity biosensors of chlorophenols: Part 4*, Chemosphere, 33 (1996) 71–79
13. Todeschini, R. and Gramatica, P., *3D-modelling and prediction by WHIM descriptors: Part 5. Theory development and chemical meaning of WHIM descriptors*, Quant. Struct.-Act. Relat., 16 (1997) 113–119.
14. Todeschini, R. and Gramatica, P., *3D-modelling and prediction by WHIM descriptors: Part 6. Application of WHIM descriptors in QSAR studies*, Quant. Struct.-Act. Relat., 16 (1997) 120–125.
15. Chiorboli, C., Gramatica, P., Piazza, R., Pino, A. and Todeschini R., *3D-modelling and prediction by WHIM descriptors: Part 7. Physico-chemical properties of haloaromatics: comparison between WHIM and topological descriptors*, SAR QSAR Envir. Res., (1997) (in press).
16. Todeschini, R., Vighi, M., Finizio, A. and Gramatica, P., *3D-modelling and prediction by WHIM descriptors: Part 8. Toxicity and physico-chemical properties of environmental priority chemicals by 20-TI and 3D-WHIM descriptors*, SAR QSAR Envir. Res., (1997) (in press).
17. Todeschini, R., Moro, G., Boggia, R., Bonati, L., Cosentino, U., Lasagni, M., and Pitea, D., *Modelling and prediction of molecular properties: Theory of grid-weighted holistic invariant molecular (G-WHIM) descriptor*, Chemom. Intell. Lab. Systems, 36 (1997) 65–73.
18. Todeschini, R., *Data correlation, number of significant principal components and shape of molecules: The K correlation index*, Anal. Chim. Acta, 348 (1997) 419–430.
19. Todeschini, R., Cazar, R. and Collina, E., *The chemical meaning of topological indices*, Chemom. Intell. Lab. Systems, 15 (1992) 51–59.
20. HyperChem, rel. 4, and Chemplus, rel. 1.0 for WINDOWS, Autodesk, Inc., Sausalito, CA, U.S.A., 1995.
21. Todeschini, R., *WHIM-3D/QSAR — Software for the calculation of WHIM descriptors*, rel. 2.1 for WINDOWS. Talete srl, Milan, 1996.
22. STATISTICA, rel. 5.0 for WINDOWS. StatSoft, Inc., Tulsa, OK, U.S.A., 1995.
23. Leardi, R., Boggia, R. and Terrile, M., *Genetic algorithms as a strategy for feature selection*, J. Chemometrics, 6 (1992) 267–281.
24. Todeschini, R., *Moby Digs — Software for Variable Subset selection by Genetic Algorithm*, rel. 1.0 for WINDOWS. Talete srl, Milan, 1997.
25. Bravi, G., Gancia, E., Mascagni, P., Pegna, R., Todeschini, R. and Zahiani, A., *MS-WHIM, new 3D theoretical descriptors derived from molecular surface properties: A comparative 3D-QSAR study in a series of steroids*, J. Comput.-Aided Mol. Design, 11 (1997) 79–92.
26. Cramer III, R.D., Patterson, D.E., Bunce, J.D., *Comparative molecular field analysis (CoMFA): I. Effect of shape on binding of steroids to carrier proteins*, J. Am. Chem. Soc., 110 (1988) 5959–5967.
27. Kubinyi, H. (Ed.), *QSAR: Hansch Analysis and Related Approaches*, VCH, Weinheim, 1993, pp. 57–68.
28. Cramer III, R.D., DePriest, S.A., Patterson, D.E. and Hecht, P., *The developing practice of comparative molecular force field analysis*, In Kubinyi, H. (Ed.), 3D QSAR in drug design, ESCOM, Leiden 1993, pp. 443–505.
29. Unger, S.H. and Hansch, C.J., *On model building in structure-activity relationship: A reexamination of adrenergic blocking activity of  $\beta$ -halo- $\beta$ -arylalkylamines*, J. Med. Chem., 16 (1973) 745–749.
30. Free, S.M. and Wilson, J.W., *A mathematical contribution to structure-activity studies*, J. Med. Chem., 7 (1964) 395–402.
31. Kier, L.B., *Shape index from molecular graphs*, Quant. Struct.-Act. Relat., 4 (1985) 109–116.
32. Kier, L.B., *An index on molecular flexibility from Kappa shape attributes*, Quant. Struct.-Act. Relat., 8 (1989) 218–221.
33. SYBYL, Release 6.0, Tripos Assoc. Inc., St. Louis, MO, U.S.A., 1993.
33. Sulea, T., Kurunczi, L. and Simon, Z., *Dioxin-type activity for polyhalogenated aryl derivatives: A QSAR model based on MTD-method*, SAR QSAR in Environ. Rea., 3 (1995) 37–61.
35. Bodor, N., Gabanyi, Z. and Wong, C., *A new method for the estimation of partition coefficient*, J. Am. Chem. Soc., 111 (1989) 3783–3786.

36. Ghose, A.K. and Crippen, G.M., *Atomic physicochemical parameters from three-dimensional structure directed quantitative structure-activity relationships: Part 4*, J. Chem. Inf. Comp. Sci., 29 (1989) 163–172.
37. Miller, K.J., *Additivity methods in molecular polarizability*, J. Am. Chem. Soc., 112 (1990) 8533–8542.
38. Bonati, L., Fraschini, E., Lasagni, M., Palma Modoni, E. and Pitea, D., *A hypothesis on the mechanism of PCDD biological activity based on molecular electrostatic potential modelling: Part 2*, Theochem., 340 (1995) 83–95.
39. Moro, G., Bonati, L., Lasagni, M., Cosentino, U., Pitea, D., and Todeschini, R., *Molecular descriptors derived from 3D-distributions of molecular scalar fields (G-WHIM descriptors): quantitative structure-activity relationships of polychlorodibenzo-p-dioxins* (forthcoming).
40. Waller, C.L. and McKinney, J.D., *Comparative molecular field analysis of polyhalogenated dibenzo-p-dioxins, dibenzofurans, and biphenyls*, J. Med. Chem., 35 (1992) 3660–3666.
41. Todeschini, R., Gramatica, P. and Navas, N., *3D-modelling and prediction by WHIM descriptors: Part 9. Chromatographic relative retention time and physico-chemical properties of PolyChloroBiphenyls (PCB)*, Cheman. Intell. Lab. Syst. (in press).
42. Colombo, S., Thesis in Environmental Sciences, University of Milan, 1996.
43. Consonni, V., Thesis in Environmental Sciences, University of Milan, 1997.
44. Triacchini, G., Thesis in Environmental Sciences, University of Milan, 1997.
45. Merli, V., Thesis in Environmental Sciences, University of Milan, 1996.
46. Tagliabue, M., Thesis in Chemistry, University of Milan, 1996.

# EVA: A Novel Theoretical Descriptor for QSAR Studies

Trevor W. Heritage<sup>a\*</sup>, Allan M. Ferguson<sup>a</sup>, David B. Turner<sup>b</sup>) and Peter Willett<sup>b</sup>

<sup>a</sup> *Tripos Inc., 1699 South Hanley Road, St. Louis, MO 63144, U.S.A.*

<sup>b</sup> *Krebs Institute for Biomolecular Research and Department of Information Studies, University of Sheffield, Wertern Bank, Sheffield S10 2TN, U. K.*

## 1. Introduction

Since the advent of classical QSAR techniques, exemplified by Hansch [1], there has been considerable progress in the development of molecular descriptors and chemometric techniques for use in such studies. The development of 3D QSAR techniques [2] that attempt to correlate biological activity with the values of various types of molecular field, for example, steric, electrostatic or hydrophobic, has been of particular interest [3,4]. The original and most well-known of the 3D QSAR techniques is Comparative Molecular Field Analysis [3], (CoMFA) which uses steric and electrostatic field values calculated at the intersections of a three-dimensional grid that surrounds the structures in the dataset. A major limitation of CoMFA and most other 3D QSAR techniques, is the dependency upon the relative orientation of the molecules in the dataset [5,6]. Despite efforts to improve the efficiency of the alignment process [7-9], the selection of the molecular alignment is regarded as the major variable in the analysis. These problems are further exacerbated when the conformational flexibility of the molecules in the dataset is considered.

There is, therefore, considerable interest in the development of new descriptions of molecular structure that do not require the alignment of molecules, but retain the 3D and molecular property information encoded within molecular fields. Alternative descriptions of molecular fields than those used in CoMFA or molecular surface properties, for example, methods based on autocorrelation vectors [10], molecular moments [11] or 3D WHIM descriptors [12], may provide effective orientation-independent descriptions of molecular structure. In this chapter, we review a novel alignment-free descriptor of molecular structure, known as EVA (EigenValue descriptor), that is derived from calculated infrared (IR) range vibrational frequencies. As discussed later in this chapter, EVA has been found to yield statistically robust QSAR models that are comparable, in statistical terms, to those derived using CoMFA, with the advantage that EVA does not require structural alignment.

## 2. The EVA Descriptor

During the late 1980s, workers at Shell Research Ltd. [13] reasoned that a significant amount of information pertaining to molecular properties, in particular, biological

---

To whom correspondence should be addressed.

activity, might be contained within the molecular vibrational wave function, of which the vibrational spectrum is a fingerprint. The EVA descriptor is derived from normal coordinate Eigenvalues (i.e. the vibrational frequencies) that are either calculated theoretically or extracted from experimental IR spectra. Typically, a classical normal coordinate analysis (NCA) [14] is performed on an energy-minimized structure and the resulting eigenvalues represent the normal mode frequencies from which the EVA descriptor is derived. The associated normal coordinate eigenvectors (i.e. the vibrational motions) are not used within the EVA descriptor. The force constants upon which a normal coordinate analysis is dependent may be determined using a molecular mechanics, semiempirical or *ab initio* quantum mechanical method. The accuracy of the calculated vibrational eigenvalues is, therefore, determined entirely by the quality of the force constants applied or derived.

Using the standard Cartesian coordinate system as a basis for describing the displacement of an atom from its equilibrium position in a vibrating molecule requires  $3N$  coordinates for a molecule containing  $N$  atoms. Three of these coordinates describe rigid-body translational motion and a further three describe rigid-body rotations. Thus, in the general case for a molecule of  $N$  atoms, there are  $3N-6$  vibrational degrees of freedom, or  $3N-5$  for a linear molecule such as acetylene (only two coordinates are required to fix the orientation). The number of vibrational degrees of freedom is equivalent to the number of fundamental vibrational frequencies (normal modes of vibration) of the molecule. While each of these fundamental vibrations can be calculated, they may or may not appear in an experimental IR absorption spectrum due to symmetry considerations — i.e. they may have zero (or close to zero) intensity [14].

Thus, in terms of the derivation of the EVA descriptor, each structure is initially characterized by  $3N-6$  (or  $3N-5$ ) vibrational modes. In all but the special case where the molecules in the dataset contain the same number of atoms it is not possible to compare the vibrational frequencies directly. This so-called dimensionality problem does not arise during a CoMFA analysis because the molecular fields arising from each molecule are calculated across a fixed set of lattice points, but this would be an issue if, for example, one wanted to compare directly the atomic point charges from which the electrostatic fields are derived. Furthermore, even in cases in which it is desired to compare molecules that do contain the same number of atoms and hence the same number of vibrational modes, it is difficult to establish which vibrations are directly comparable between molecules: this problem arises from inherent and effectively indeterminate contributions made by individual atoms to a given vibrational mode [15].

In EVA, the dimensionality of the descriptor is unified across the entire dataset by a three-step procedure that involves transformation of the sets of vibrational frequencies onto a scale where they are directly comparable (i.e. a scale of fixed dimensionality). In the initial step of this standardization process, the frequency values are projected onto a bounded frequency scale (BFS) with individual vibrations represented by points on this axis (Fig. 1). The bounds chosen for the BFS of 0 and  $4000\text{ cm}^{-1}$  encompass the frequencies of all fundamental molecular vibrations and facilitate comparison against experimentally derived IR spectra. The second step in the standardization process involves the generation of a value at each point on the BFS whereby each calculated frequency is

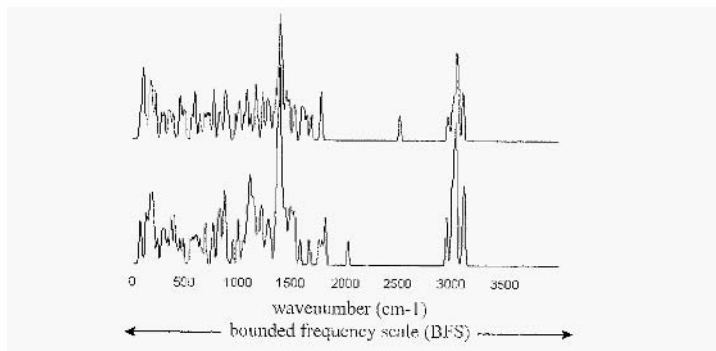


Fig. 1. Bounded frequency scale with superimposed EVA descriptors for cimetidine (lower spectrum and ranitidine (upper spectrum).

characterized in terms of 'peak' height, width and shape. Each of the calculated vibrations is weighted equally during this process. The resulting value associated with each of the calculated vibrations permits the proportion of overlap of vibrations to be determined and may be considered analogous to, but in no way representative of, peak intensity. In principle, it should be possible to use theoretical vibrational intensities derived from derivatives of calculated or measured dipole moments, but the resulting intensities are notoriously inaccurate and this has not been attempted at the time of writing.

In practice, in the second step a Gaussian function of fixed standard deviation ( $\sigma$ ) is placed over each vibrational frequency value for a given structure, resulting in a series of  $3N-6$  (or  $3N-5$ ) identical and overlapping Gaussians (Fig. 2). The value of the EVA descriptor,  $EVA_x$ , at any chosen sampling point,  $x$ , on the bounded frequency scale is

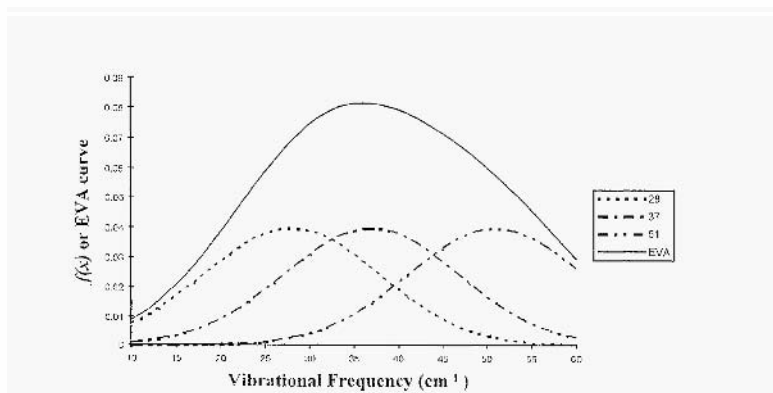


Fig. 2. Profile of the summed overlap Gaussians ('EVA curve') for three arbitrary vibrational frequencies and using a term of  $10 \text{ cm}^{-1}$ . The curve 'EVA' is determined by summing the estimated 'intensities' of the vibrations centered at  $28 \text{ cm}^{-1}$ ,  $37 \text{ cm}^{-1}$  and  $51 \text{ cm}^{-1}$ , respectively. The EVA descriptor is exacted by sampling the frequency scale at fixed intervals of  $L \text{ cm}^{-1}$ .

then determined by summing the contributions from each and every one of the  $3N-6$  (or  $3N-5$ ) overlaid Gaussians at that point

$$EVA_x = \sum_{i=1}^{3N-6} \frac{1}{\sigma \sqrt{2\pi}} e^{-(x-f_i)^2 / 2\sigma^2}$$

where  $f_i$  is the  $i$ 'th frequency for the structure.

It is important at this stage to remind the reader that the purpose of the above EVA smoothing procedure is not an attempt to simulate the infrared spectrum of the molecule of interest, since the transition dipole data is ignored, but rather to provide a basis upon which vibrations occurring at slightly different frequencies may be compared to one another. The Gaussian function applied to define peak shapes adds a probabilistic element, in that the peak maxima are centered at each of the calculated frequency values ( $f_i$ ) and, thus, these points are taken to be the most probable values of the respective frequencies. An EVA descriptor sampled at a point  $x \neq f_i$  is thus considered to be a less probable value of the  $i$ 'th frequency and the corresponding contribution of  $f_i$  to the final value of  $EVA_x$  will be less than the maximum possible contribution. To a certain extent, this behavior of the EVA descriptor reduces the dependency of the final QSAR model on the accuracy of the original calculated vibrational frequencies, which are sensitive to the molecule geometry optimization criteria and to the theoretical approximations or empirically based parameters of the NCA procedure. Furthermore and as discussed in detail below, this behavior has implications regarding the sensitivity of the descriptor to molecular conformation, in that small changes in vibrational frequencies arising from conformational changes may have insignificant effects on the resulting EVA descriptor values.

In the third and final, step of the data standardization process, the EVA function is sampled at fixed increments of  $L \text{ cm}^{-1}$  along the BFS, which results in the  $4000/L$  values that comprise the EVA descriptor. Typically, a descriptor set is derived using a Gaussian standard deviation ( $\sigma$ ) term of  $10 \text{ cm}^{-1}$  and a sampling increment ( $L$ ) of  $5 \text{ cm}^{-1}$ , resulting in 800 descriptor variables. As is the case with the CoMFA technique, the dimensionality of the EVA descriptor is much larger than the number of compounds in a typical QSAR dataset and, thus, data reduction methods such as partial least squares to latent structures (PLS) [16] or principal components regression (PCR) are applied to yield robust correlations with biological data.

### 3. Applications of the EVA Descriptor in QSAR

One of the first demonstrations in the public domain of the regressive modeling capability of the EVA descriptor was obtained in a QSPR study [19], using the BCDEF dataset of Cramer [20]. The dataset consists of measured  $\log P$  values for a highly heterogeneous set of 135 small organic chemicals, ranging from polycyclic aromatics such as the highly lipophilic phenanthrene ( $\log P = 4.46$ ) to small hydrophilic moieties including methanol ( $\log P = 0.64$ ). The EVA descriptors were derived using a Gaussian spread ( $\sigma$ ) term of  $10 \text{ cm}^{-1}$  and a sampling increment ( $L$ ) of  $5 \text{ cm}^{-1}$  based on normal coordinate fre-

quencies calculated using the AM1 [21] Hamiltonian in the MOPAC [22] semiempirical molecular orbital program. These parameters yielded an EVA descriptor consisting of 800 variables per structure, which were regressed against the  $\log P$  values using PLS. A regression equation based on only five PLS factors, that explained 96% of the variance in the  $\log P$  values, was obtained in this way. Full leave-one-out cross-validation of this dataset yielded a cross-validated- $r^2$  (i.e.  $q^2$ ) of 0.68. This model was then used to predict the  $\log P$  value for a test set of 76 'unseen' chemicals, resulting in a predictive  $r^2$  of 0.65. This study demonstrates the value of EVA as both an explanatory and a predictive tool and, in addition, highlights one of the key advantages over 3D QSAR techniques such as CoMFA. In cases such as this, where no intuitive alignment of the dataset structures exists, it is very difficult or even impossible to apply CoMFA in a meaningful way, but with EVA no such complexity exists. Furthermore, bulk properties such as  $\log P$  have no orientation dependence and, thus, any attempt to introduce such a dependency for QSAR purposes is entirely arbitrary. The diversity of structures exemplified in this dataset also suggests that EVA may be applied to the analysis of diverse sets of compounds rather than just to congeneric series, which is a limitation for most alternative descriptors.

In subsequent studies [23,24], the general applicability of the EVA descriptor in QSAR studies has been investigated in detail using datasets exhibiting a range of biological end-points (Table 1). Using EVA descriptors derived from AM1 modes, good PLS models (in terms of  $q^2$ ) can be obtained for nine of the eleven datasets. The exceptions to this are the oxadiazole [25] and biphenyl [26] datasets for which, at best, only poor models can be obtained. It is important to remind the reader that although the EVA QSAR models presented in Table 1 are satisfactory, they are based solely upon the default EVA descriptor parameters ( $\sigma = 10 \text{ cm}^{-1}$  and  $L = 5 \text{ cm}^{-1}$ ). Additional studies [23,24] have been performed in which the effect of changes to these parameters on the quality of the final QSAR models has been investigated and for nearly all of the datasets listed there exist combinations of  $\sigma$  and  $L$  that give rise to superior PLS models. A range of these parameters should, therefore, be investigated prior to settling on a final model. A protocol recommended by Turner et al. [23] suggests that a value of  $10 \text{ cm}^{-1}$  is a reasonable starting point for a QSAR study and thereafter if satisfactory results are not achieved, to supplement this with analyses based on  $\sigma$ -terms of 5.25 and  $50 \text{ cm}^{-1}$ .

A useful benchmark in determining the effectiveness of the EVA descriptor for QSAR studies is to compare the statistical performance and model characteristics based on the EVA descriptor with the analogous CoMFA model for the same datasets. A key limitation in all such comparative studies [23,24] is that the datasets have been selected because a good, published CoMFA model exists. This, therefore, leads to significant bias in favor of the CoMFA technique, but none the less, the results do provide interesting insights into the nature and scope of the EVA descriptor.

Examination of Table 1 shows that, at least in terms of the  $q^2$  scores, the EVA descriptors provide roughly equivalent correlations for the cocaine [27], dibenzofuran [26], dibenzo-*p*-dioxin [26], piperidine [25], sulphonamide [25] and steroid datasets [3]. Although not as high as CoMFA, good predictive correlations are also obtained using EVA for the  $\beta$ -carboline [28] and nitroenamine [25] datasets. The two cases where

Table 1 Summary of QSAR analyses using EYA (derived from AMBER and MOPAC AM1 normal modes) and CoMFA descriptors<sup>a</sup>

Dataset	AMBER			MOPAC AM1			CoMFA (both fields)					
	$q^2$	$r^2$	SE	$F$	$q^2$	$r^2$	SE	$F$	$q^2$	$r^2$	SE	$F$
$\beta$ -Carbolines <sup>b</sup>	0.29(6)	0.97	0.50	180.6	0.50(6)	0.97	0.57	195.5	0.68(4)	0.89	1.12	69.9
Biphenyls	0.16(1)	0.72	0.48	30.3	0.28(2)	0.90	0.30	49.0	0.49(3)	0.87	0.36	21.9
Cocaines	0.57(2)	0.91	0.26	51.9	0.49(2)	0.95	0.20	94.0	0.59(4)	0.88	0.28	63.0
Dibenzo- <i>p</i> -dioxins	0.48(2)	0.85	0.59	61.5	0.68(2)	0.88	0.53	77.2	0.66(1)	0.80	0.66	92.3
Dibenzofurans	0.61(1)	0.74	0.69	103.4	0.78(4)	0.97	0.25	273.9	0.72(6)	0.85	0.57	29.9
Muscarinics	0.42(3)	0.88	0.27	81.7	0.53(4)	0.95	0.17	171.3	0.59(4)	0.84	0.31	46.0
Nitroenamides	0.47(2)	0.86	0.59	41.3	0.49(3)	0.93	0.41	61.5	0.84(3)	0.96	0.34	92.4
Oxadiazoles	-0.36(1)	-	-	-	-0.19(1)	0.38	0.33	12.9	0.51(2)	0.85	0.07	56.2
Piperidines	0.71(5)	0.84	0.42	137.9	0.76(4)	0.84	0.43	174.7	0.73(3)	0.80	0.48	175.2
Steroids (TBC <sup>c</sup> activity)	0.42(5)	0.99	0.17	206.4	0.70(4)	0.98	0.20	175.0	0.62(3)	0.92	0.37	67.7
Steroids (CBG <sup>d</sup> activity)	0.79(2)	0.90	0.38	85.0	0.70(2)	0.87	0.45	59.1	0.75(2)	0.91	0.37	93.0
Sulphonamides <sup>e</sup>	-	-	-	-	0.54(6)	0.80	16.4	60.2	0.65(5)	0.82	15.1	87.7

<sup>a</sup> The leave-one-out  $q^2$  values are reported together with the optimal number of LVs in brackets. All  $q^2$  values of  $\leq 0$  are indicated as  $\leq \phi$  and  $LV_{opt}$  omitted as meaningless. Models are based on the selection of  $LV_{opt}$  by minimum  $SE_{cv}$  score. Full (fitted) models were derived only where  $q^2 > 0$ .

<sup>b</sup> AMBER had the required force-field parameters for only 39 structures.

<sup>c</sup> Testosterone-binding globulin affinity as the target activity.

<sup>d</sup> Corticosterone-binding globulin affinity as the target activity.

<sup>e</sup> AMBER force-field parameters not available.



EVA performs poorly, the oxadiazole [25] and biphenyl [26] datasets, also yield the poorest CoMFA results, although statistically significant correlations ( $q^2 \approx 0.5$ ) are still obtained using CoMFA.

The robustness of PLS models derived using EVA has been extensively tested by Turner [24,31], in terms of both randomization permutation testing [16] and the ability of those models to make reliable predictions for test chemicals. Using the standard steroid dataset from the original CoMFA study [3], albeit with structures corrected according to Wagener et al. [10], a predictive- $r^2$  value of 0.69 is obtained for the ten test chemicals; the biological end-point used was the affinity for corticosteroid-binding globulin (CBG) expressed as  $1/[\log K]$ . This compares to a much lower value for CoMFA combined steric and electrostatic fields of 0.35. The apparently poor CoMFA test set predictive performance is almost entirely due to an extremely poor prediction for the only structure in the test set containing a fluorine atom, omission of which raises the CoMFA predictive- $r^2$  to 0.84. In contrast, the EVA predictive performance is raised by 0.05 when this compound is excluded, a small but none the less significant improvement. Clearly, in terms of the EVA descriptor space this compound cannot be considered an extreme outlier, but in terms of CoMFA fields it is too different from the structures in the training set for a reliable prediction to be made.

The main advantage of EVA over CoMFA for QSAR purposes is the fact that orientation and alignment of the structures in the dataset is not required. In CoMFA, the alignment is the major variable, providing in some instances different modelling statistics for even quite small changes in the relative positions of the atoms in a pair of structures. However, given the nature of the field-based descriptors used in CoMFA, alignment does facilitate a powerful means of visualizing the important features of a QSAR model in the form of plots of the structural regions that are most highly correlated (either positively or negatively) with the biological property of interest. Despite the undoubted utility of these CoMFA plots, they do not indicate precisely which atoms are responsible for the modelled correlations since the electrostatic and steric fields are composed of contributions from each and every atom in the molecule, although the Predicted Activity Contributions (PAC) [17] method has been reported to overcome this problem. A further point to note is that it is not possible to predict the effects that structural changes may have on the resultant CoMFA fields. In contrast to CoMFA, there exists no obvious means of backtracking from those components of the EVA descriptor which are highly correlated with changes in biological activity to the corresponding molecular structural features; a discussion of the ways of achieving this is presented at the end of this chapter.

#### 4. EVA Descriptor Generation Parameters

The judicious selection of parameters is a prerequisite to the success of any QSAR method and EVA is no exception. The most fundamental parameters involved in the derivation of the EVA descriptor are the Gaussian standard deviation ( $\sigma$ ) and the sampling increment ( $L$ ).

#### 4.1. Gaussian standard deviation ( $\sigma$ )

The effect of varying the  $\sigma$  term of the EVA descriptor is illustrated in Fig. 3 in which, as  $\sigma$  is increased, the features of the descriptor profile are progressively smoothed. The effect of the application of a Gaussian function during the EVA descriptor standardization process is to 'smear out' a particular vibrational frequency such that vibrations occurring at similar frequencies in other structures overlap to a lesser or greater extent. It is this overlap that provides the variable variance upon which PLS modelling is dependent. By definition, each and every Gaussian must overlap, but for the most part this occurs at small (negligible) values and, consequently, the contribution to variance is very small. Only where the frequency values are sufficiently close to one another relative to the value of  $\sigma$  is it likely that interstructural overlap of Gaussians will occur at values of significant magnitude. The selection of the Gaussian standard deviation, therefore, determines the number of and extent to which, vibrations of a particular frequency in one structure can be statistically related to those in the other structures in the dataset.

In addition to interstructural overlap of Gaussians, the  $\sigma$  term also governs the extent to which vibrations within the same structure may overlap at non-negligible values. Intrastructural Gaussian overlap of this type, which is also dependent on the 'density' (i.e. proximity) of vibrations at various regions of the spectrum, causes EVA variables to consist of significant contributions from more than one vibrational frequency. The mixing of information contributed by individual normal coordinate frequencies is generally considered undesirable, but in order to provide sufficient interstructural Gaussian overlap, it is inevitable that a certain degree of intrastructural overlap occur.

Thus, small values of  $\sigma$  give rise to minimal intrastructural Gaussian overlap, while at larger values  $\sigma$  of significant overlap arises. In the former case, there will be a reduction in interstructural overlap, perhaps to such an extent that there exists no overlap of the Gaussians at significant values. In this instance, the descriptor takes on the characteristics of a binary indicator, showing only the presence or absence of specific features, thereby rendering the descriptor useless for regression analysis, but perhaps still of utility in classification analysis. In cases where larger  $\sigma$  values are used, increased mixing of the information encoded by one frequency with that encoded by other frequencies arises.

#### 4.2. Sampling increment ( $L$ )

Detailed investigation into the effect of various combinations of the  $\sigma$  and  $L$  parameters on the resulting  $q^2$  value has been carried out by Turner [24]. Turner's results indicate that, for the most part, the final  $q^2$  value is insensitive to small changes in either of these parameters —i.e. the information content of the EVA descriptor remains consistent. The most significant variations in  $q^2$  are seen as  $\sigma$  is reduced (giving a more spiky vibrational 'intensity') and the sampling increment ( $L$ ) is increased; this is analogous to lowering the spectral resolution. This result is intuitively reasonable since one would anticipate that, as  $L$  becomes very large relative to  $\sigma$  some of the Gaussian peaks (or

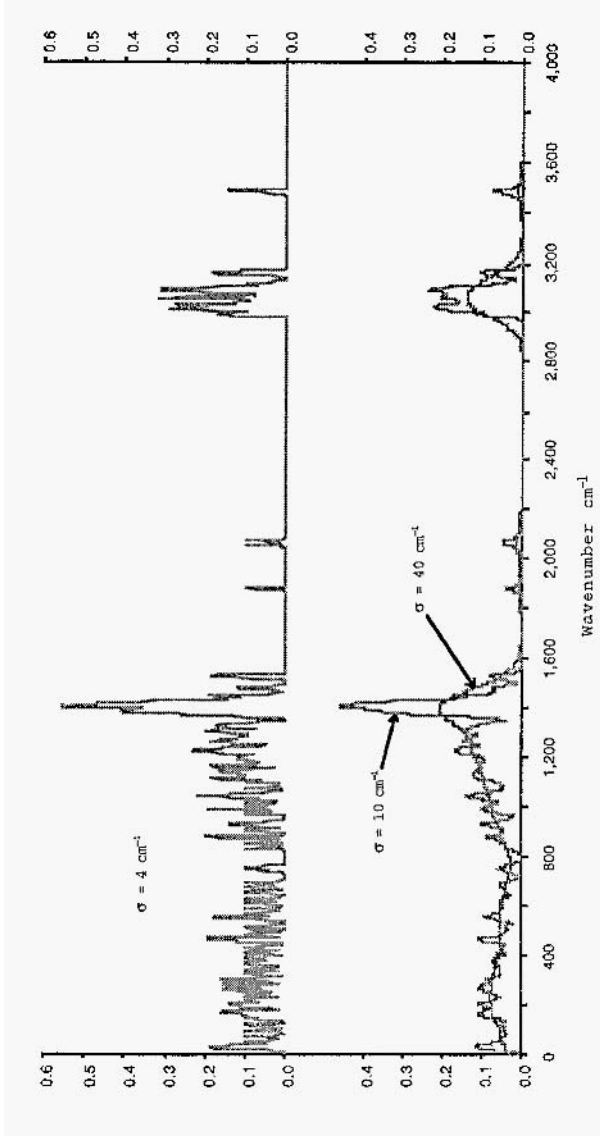


Fig. 3. Effect of the  $\sigma$  term on EVA descriptor profile.

information encoded within them) will be omitted from the descriptor. In some cases, the information omitted will be predominantly noise resulting in a superior QSAR model; but in other cases, signal may be accidentally omitted, resulting in degradation of the QSAR model. This phenomenon is known as *blind variable selection* since variables are selected or excluded from the descriptor on a completely arbitrary basis which is, of course, undesirable. The value of  $L$  at which blind variable selection begins to occur is related to the  $\sigma$  term; the larger the  $\sigma$  term the higher the permissible value of  $L$ . Thus, to avoid blind variable selection, one might wish to minimize the value of  $L$ , but this must be balanced against the additional computational requirements associated with such a practice. Conversely, therefore, the value of  $L$  should be maximized in order to reduce the computational overhead and this leads to the concept of critical  $L$  values ( $L_{crit}$ ) which are  $\sigma$  specific and which, if exceeded, result in a sampling error. Table 2 lists generally applicable  $L_{crit}$  values for various  $\sigma$  values that were derived by a systematic study of  $L$  and  $\sigma$  parameter settings for several EVA datasets [23]. Table 2 confirms that the intuitively reasonable and default, selection of an  $L$  of  $5\text{ cm}^{-1}$  with a  $\sigma$  term of  $10\text{ cm}^{-1}$  should result in no blind variable selection and that in point of fact, the value of  $L$  may be increased to  $20\text{ cm}^{-1}$  with no apparent information loss (change in  $q^2$ ).

The existence of these  $L_{crit}$  values is important not least because one of the problems with CoMFA at present is that the coarse grid-point spacing (typically  $2\text{ \AA}$ ) that is generally used is such that there is incomplete sampling of the molecular fields, resulting in information loss. The consequence of this is that reorientation of an aligned set of molecules as a rigid body within the defining CoMFA 3D region often results in substantial changes to the resulting QSAR model [30], as evidenced in the  $q^2$  values. EVA, on the other hand, does not suffer from such sampling errors, provided that the  $L_{crit}$  values given in Table 2 are not exceeded.

## 5. Characteristics of the EVA Descriptor

Although the EVA descriptor is not intended to simulate the infrared spectrum of a molecule, it is useful to visualize the EVA descriptor in the form of a ‘spectrum’. This permits the interpretation of the EVA descriptor by examination of the distribution of vibrations in a molecule or in a set of molecules. Figure 4 shows plots of the EVA descriptor for deoxycortisol (one of the most active CBG-binding compounds in the original steroid dataset used by Cramer [3]) and estradiol (one of the inactive structures) over the spectral range  $1$  to  $4000\text{ cm}^{-1}$ . Also shown in Fig. 4 is the univariate standard deviation of the descriptor over the entire dataset of 21 structures [3]. The density of

Table 2 critical values of  $L$  ( $L_{crit}$ ) for selected Gaussian terms

Gaussian standard deviation, $\sigma^{-1}$	1	2	3	4	6	8	10	14	21
Threshold increment, $L_{crit}$ ( $\text{cm}^{-1}$ ) <sup>a</sup>	2	4	5	8	10	16	20	25	32

<sup>a</sup> In order to avoid sampling errors the value of  $L$  should be chosen to be less than  $L_{crit}$  for a given  $s$ ; these values have been chosen such that effects resulting from the choice of sampling frame (determined by  $S$ ) are accounted for.

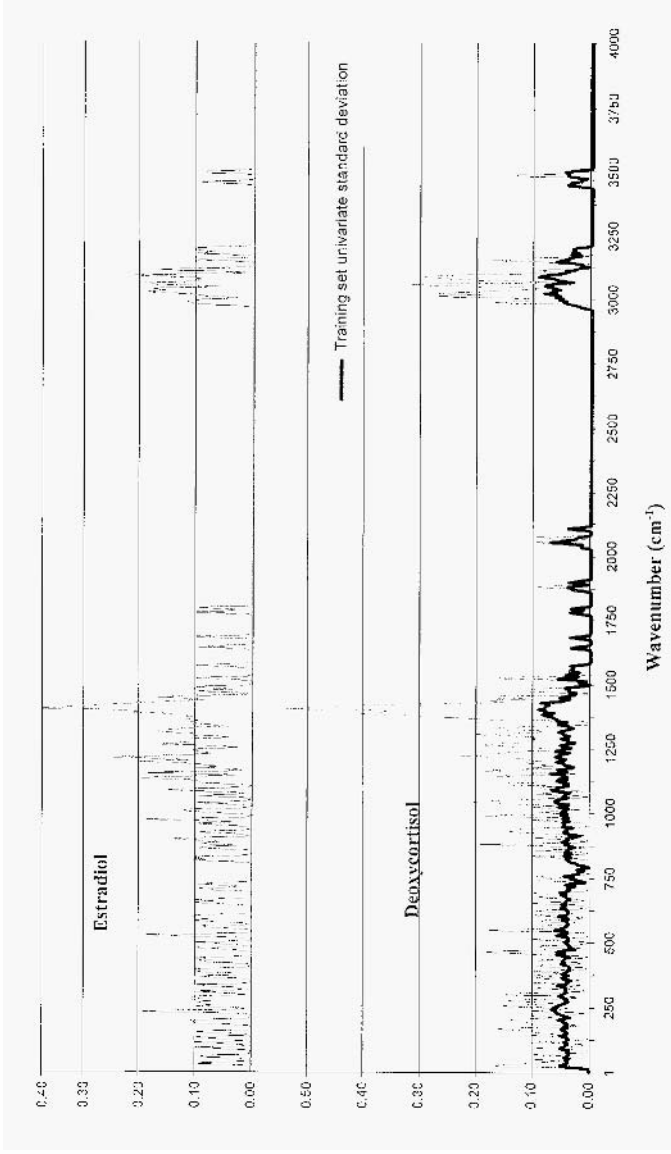


Fig. 4. EVA pseudo-spectra for estradiol (inactive for CBG-binding) and deoxycortisol (highest CBC-binding activity); a Gaussian of 4 cm<sup>-1</sup> has been used. The univariate standard deviation for the 21 training set structures is given by the heavy line.

peaks in the fingerprint region (1 to 1500  $\text{cm}^{-1}$ ) indicates that there is considerably more vibrational information in this region than in the functional group region (1500 to 4000  $\text{cm}^{-1}$ ) of the spectrum, as is typical in most infrared spectra. The EVA descriptor values and the standard deviation over the entire dataset are largest at frequencies centered around 1400 and 3100  $\text{cm}^{-1}$ , corresponding to C–H bending and stretching vibrations. Figure 4 also highlights the errors associated with the calculation of normal coordinate frequencies (in this case, using MOPAC), since a carbonyl stretching frequency is expected (from experiment) to appear at around 1700  $\text{cm}^{-1}$ , but is represented on this plot by peaks in around 2060  $\text{cm}^{-1}$ . This feature of the EVA descriptor, once again, indicates that there is no attempt to simulate an experimental IR spectrum, but does not detract from the usefulness of the descriptor for QSAR purposes, since consistency rather than accuracy across the dataset is critical. Furthermore, for QSAR purposes, relative rather than absolute differences in vibrational frequency across the dataset are important. One might expect that this would become more of an issue should heterogeneous datasets be used since the consistency with which errors associated with the reproduction of equivalent vibrational frequencies may be more erratic. In practice, however, reasonable QSAR results have been obtained using a variety of heterogeneous datasets [19,23].

## 6. Conformational Sensitivity of the EVA Descriptor

The sensitivity of CoMFA to the molecular orientation and alignment and, therefore, to the molecular conformation is well established [32,33], but while EVA is completely independent of molecular orientation and alignment, the impact of the molecular conformation on EVA QSAR performance has, thus far, not been discussed. Intuitively, it is obvious that a change in conformation will result in changes in the force constants between atoms and, therefore, in the normal coordinate frequencies and displacements. The questions are: ‘to what extent are these changes evident within the EVA descriptor?’ and ‘how much of this is accounted for by the Gaussian spread term?’. Some limited studies of these conformational effects have been performed [31,33]. In one such study [33] performed by Shell Research Ltd., five classes of chemical known to act at the same biological target, encompassing pyrazoles, thiazoles, piperidines, quinolines and thiochromans, and totalling more than 250 structures, were clustered using a nearest-neighbor algorithm, based on the EVA descriptor. The conformations of each molecule were repeatedly randomized, new EVA descriptors generated and the clustering process repeated. The conclusions from this study were that, while the nearest-neighbor relationships between compounds change, the overall cluster membership is approximately constant. This result suggests that, in the vast majority of cases, a conformational change does not lead to a sufficiently large change in the resulting EVA descriptor to cause a change in the underlying statistical model.

In a more recent study [31], EVA descriptors for test chemicals were generated for a conformation which matched that used in the training set and also for a non-matching conformation. At low  $\sigma$  values, the predictions made based on the non-matching conformation are considerably poorer than those made for the matched conformation. This

difference gradually decreases until convergence is achieved at  $\sigma = 12 \text{ cm}^{-1}$ , thereafter the predictions from the two conformations are roughly equivalent. In general, the conformational sensitivity of the EVA descriptor decreases as  $\sigma$  is increased. As would be expected, the predictions made using CoMFA for non-matching conformations are much poorer than any of those obtained using EVA, thereby highlighting the relative conformational sensitivity of the two methods.

## 7. QSAR Model Interpretation

In CoMFA, 3D isocontour plots are used to visualize those regions of space indicated by the PLS model to be most highly positively or negatively correlated with biological activity. While no such 3D visualization is possible with EVA, a variety of 2D plots have been suggested [24,31] that indicate the relative importance of regions of the spectrum in correlating biological activity. Figure 5 shows two such plots based on a two-component PLS model for the steroid dataset [3] that, in some ways, facilitate interpretation of an EVA QSAR model in analogous fashion to the interpretation of an experimental IR spectrum. The two measures shown in the figure are the magnitudes of the regression coefficients (B) and the variable influence on projection (VIP) [34]. It is pertinent to remind the reader that the peak heights depicted in Fig. 5 represent the relative importance of the EVA variables in the PLS analysis and are in no way related to vibrational intensity.

To backtrack to the important structural features indicated by the PLS model, it is necessary first to identify the variables most highly correlated with activity, decompose those variables into the contributing vibrational frequencies and then to interpret and visualize the underlying normal mode vibrations.

Two simple approaches have been proposed for identifying the most important variables in the PLS analysis [31]. The first approach suggests that important variables will have regression coefficients in excess of half of the largest coefficient. The second method, based upon the VIP score, states that important variables will have a VIP score greater than 1.0, while unimportant variables will have a VIP score less than 0.8 [34]. Analysis of the EVA descriptor ( $\sigma = 4 \text{ cm}^{-1}$ ) for the steroid dataset by Turner et al. [31] results in the selection of too many EVA variables at a threshold of  $\text{VIP} \geq 1.0$  (183 variables), but a threshold of  $\text{VIP} \geq 3.0$  yields a more manageable number (17 variables). It is reasonable to use such a high VIP threshold since these are the variables most heavily weighted by PLS and, thus, may be used to get some feel for the main structural features used to discriminate between the training set structures.

The decomposition of the selected (important) EVA variables into their contributory normal mode frequencies is most straightforward and certainly less ambiguous, if each EVA variable is composed of one and only one normal coordinate frequency. For this reason, it is important that the smallest value is used during the analysis as possible, since, as discussed earlier,  $\sigma$  directly affects the degree of intrastructural Gaussian overlap. Examination of the underlying frequencies for EVA variables with  $\text{VIP} \geq 3.0$  is not straightforward. However, for the steroid dataset, PLS appears to discriminate between high-, medium- and low-active structures based on the presence or absence of

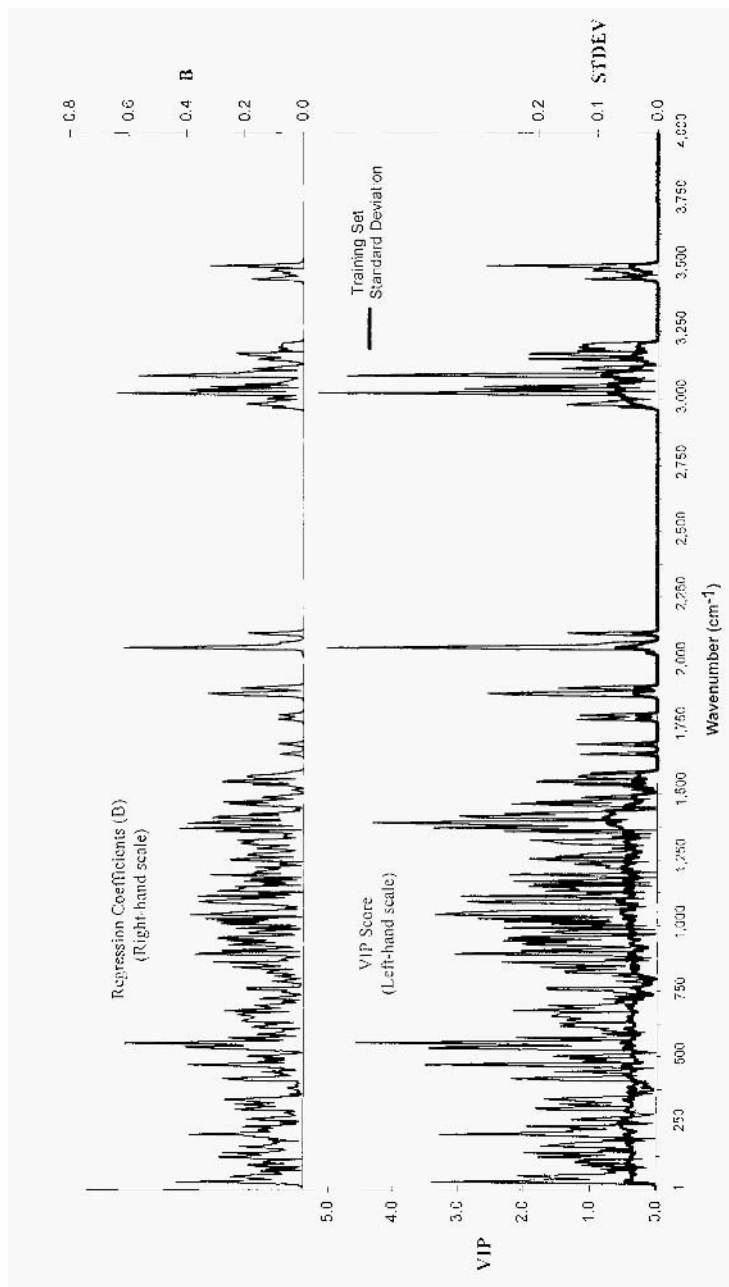


Fig. 5. Regression coefficient (B) and VIP score plots from EVA 4  $\text{cm}^{-1}$  two LV fitted model.



specific frequencies that are characteristic of the functionalities considered important for binding affinity. For example, the variable with the second-highest VIP score at 2056  $\text{cm}^{-1}$  relates to the position-3 carbonyl group stretching mode. This group is one of the features deduced by Mickelson et al. [35] to be critical for CBG-binding and is present in all of the high- and medium-activity compounds, as well as the most active of the low-activity compounds.

The attempts at interpreting QSAR models based upon the EVA descriptor, discussed herein, are encouraging, in that the classifications between structures can, to some extent, be rationalized in terms of the features postulated to be necessary for activity. None the less, EVA QSAR models cannot, to date, be interpreted to the same extent as CoMFA models in which the correlations may be related to probe interaction energies.

## **8. Summary**

One of the main problems encountered with QSAR techniques that use fields to characterize molecules, such as CoMFA, is the need to align the structures concerned. The selection of such alignments, in terms of the molecular orientation and conformation, is essentially arbitrary, but has profound effects on the quality of the derived QSAR model. For this reason, a number of groups have attempted to develop new 3D QSAR techniques that extend beyond this limitation, with varying degrees of success. This chapter has reviewed the progress made with one such methodology, that based upon molecular vibrational eigenvalues and known as EVA.

EVA provides an entirely theoretically based descriptor derived from calculated, fundamental molecular vibrations. Molecular structure and conformational characteristics are implicit in the descriptor since the vibrations depend on the masses of the atoms involved and the forces between them. The significant advantage that EVA offers relative to CoMFA and related 3D QSAR techniques is that molecular vibrational properties are orientation independent, thereby eliminating ambiguity associated with the well-known molecular alignment problem.

The discussion of the QSAR modelling performance of EVA herein illustrates that the general applicability of the descriptor and the robustness of the resultant QSAR (PLS) models in terms of cross-validation statistics. In addition, extensive randomization testing of the PLS models discussed herein [31] shows that the probability of obtaining similar correlations by chance to those actually obtained using the EVA descriptor is essentially zero. Randomization and related statistical tests [16] have played a crucial role in conclusively demonstrating that EVA can be used to correlate biological activity or other properties and generate statistically valid QSAR models. In most, but not all, cases examined EVA compares favorably with CoMFA, in terms of the ability to build statistically robust QSAR models from training set structures and in terms of the ability to use those models to predict reliably the activity of 'unseen' test chemicals. Furthermore, EVA has yielded predictively useful QSAR models for quite heterogeneous datasets, where the application of CoMFA is difficult or impossible.

The promising results presented herein may lead one to believe that development of the EVA methodology has been completed, but this is not the case. There is

considerable interest in exploring several aspects of the descriptor, including the correlation with specific types of effects (e.g. hydrophobic, steric or electrostatic) and the rational selection of localized  $\sigma$  values as a basis for establishing suitable probability density functions for particular types of vibration or regions of the infrared spectrum. In addition, despite the example provided herein of taking significance-of-variable plots coupled with techniques for selecting these variables as a means to interpreting an EVA QSAR model, there is need for more sophisticated techniques for the decomposition of EVA variables into the underlying normal mode vibration(s) and thereby to the groups of atoms that are characteristic of those vibrations. A further area that requires investigation is the sensitivity of EVA to the molecular conformation used and to what extent this governs the choice of  $\sigma$  parameter.

As the EVA methodology matures other applications, besides 3D QSAR, will begin to emerge that take advantage of the strengths of the technique. One such example [36], centers on the use of EVA for similarity searching in chemical databases, in which the overall conclusions are that EVA is equally effective for this purpose as the more traditional 2D fingerprint method. although depending on the similarity measure applied, the hits returned by EVA and 2D similarity measures may be structurally quite different. A consequence of this finding is that EVA-based similarity searching may provide an alternative source of inspiration to a chemist browsing a database. The applicability of EVA in the context of database similarity searching is in stark contrast to the complexities associated with field-based similarity searching [9] in chemical databases.

Finally, the technique described herein that yields the standardized EVA descriptor from the calculated vibrational frequencies is not limited to that purpose and may, in principle, be applied in any circumstance where the property or descriptor is non-standard. For example, the standardization procedure may be applied to interatomic distance information, either for a single conformation or as a means of summarizing conformational flexibility. Furthermore, the same procedure may be applied to other descriptions of molecular structure that are dependent on the number of atoms, such as electron populations, partial charges or vibrational properties other than normal coordinate eigenvalues (EVA), including transition dipole moments (intensity) or eigenvector data (directionality of the vibrations). The EVA standardization methodology, therefore, provides a novel means of transforming data. Furthermore, it is conceivable that descriptor strings derived from different sources, such as these, may be concatenated in a manner similar to that of the Molecular Shape Analysis method of Dunn et al. [37].

## References

1. Hansch, C. and Fujilta, T.,  $\rho$ - $\sigma$ - $\pi$  analysis: A method for the correlation of biological activity and chemical structure, *J. Am. Chem. Soc.*, 86 (1964) 1616–1626.
2. Wiese, M., In Kubinyi, H. (Ed.) 3D QSAR in drug design, ESCOM, Leiden, 1993.
3. Cramer, R.D., Patterson, D.E. and Bunce, J.D. comparative molecular field analysis (CoMFA): 1. Effect of shape on binding of steroids to carrier proteins, *J. Am. Chem. Soc.*, 110 (1988) 5959–5967.
4. Kim, K.H. and Martin, Y.C., Direct prediction of linear free-energy substituent effects from 3D structures using comparative molecular-field analysis: 1. Electronic effects of substituted benzoic-acids, *J. Org. Chem.*, 56 (1991) 2723–2729.

5. Klebe, G., Abraham, U. and Mietzner, T., *Molecular similarity indexes in a comparative-analysis (CoMSIA) of drug molecules to correlate and predict their biological activity*, J. Med. Chem., 37 (1994) 4130–4146.
6. Kellogg, G.E., Semus, S.F. and Abraham, D.J., *HINT — A new method of empirical hydrophobic field calculation for CoMFA*, J. Comput.-Aided Mol. Design, 5 (1991) 545–552.
7. Good. A.C., *The calculation of molecular similarity: Alternative formulas, data manipulation and graphical display*, J. Mol. Graph., 10 (1992) 144–151.
8. Good. A.C., Hodgkin, E.E. and Richards. W.G., *The utilisation of Gaussian functions for the rapid evaluation of molecular similarity*, J. Chem. Inf. Comput. Sci., 32 (1992) 188–191.
9. Thorner, D.A., Wild, D.J., Willett, P. and Wright, P.M., *Similarity searching in files of three-dimensional structures: Flexible field-based searching of MEP*, J. Chem. Inf. Comput. Sci., 36 (1996) 900–908.
10. Wagener, M., Sadowski, J. and Gasteiger, J., *Autocorrelation of molecular surface properties for modeling corticosteriod binding globulin and cytosolic Ah receptor activity by neural networks*, J. Am. Chem. Soc., 117 (1995) 7769–7775.
11. Silverman, B. D. and Platt. DE., *Comparative molecular moment analysis (CoMMA): 3D QSAR without molecular superposition*, J. Med. Chem., 39 (1996) 2129–2140.
12. Clementi, S., Cruciani, G., Riganelli, D. and Valigi, R., In Dean. P.M., Jolles, G. and Newton, C.G. (Eds.) *New perspectives in drug design*. Academic Press, London, 1995. pp. 285–310.
13. Ferguson, A.M. and Heritage.T.W., Shell Research Ltd. Internal Report. 1990 (not publicly available).
14. Herzberg, G., *Molecular Spectra and Molecular Structure: II, Infrared and Raman Spectra Polyatomic Molecules*, 8th Ed., D. Van Nostrand Company Inc., New York, 1945.
15. Ferguson, A.M., and Jonathan. P., Shell Research Ltd. Internal Report, 1990 (not publicly available).
16. Lindberg, W., Persson, J.-A. and Wolds, S., *Partial least-squares method for spectrofluorimetric analysis of mixtures of humic acid and ligninsulfonate*, Anal. Chem., 55 (1983) 643–648.
17. Waszkowycz, B., Clark. D.E., Frenkel, D., Li, J., Murray, C. W., Robson, B. and Westhead. D.R., *PROG-LIGAND — an approach to de Novo molecular design: 2. Design of novel molecules from molecular field analysis (MFA) models and pharmacophores*, J. Med. Chem., 37 (1994) 3994–4002.
18. Weiner, S.J., Kollman, P.A., Case, D.A., Singh, U.C., Ghio, C., Alagona, G., Profeta, Jr., S. and Weiner, P., *A novel force field for molecular mechanical simulation of nucleic acids and proteins*, J. Am. Chem. Soc., 106 (1984) 765–784.
19. Ferguson, A.M., Heritage, T.W., Jonathan. P., Pack, S.E., Phillips. L., Rogan, J. and Snaith, P.J., *EVA: A new theoretically based molecular descriptor for use in QSAR/QSPR analysis*, J. Comput. -Aided Mol. Design. 11 (1997) 143–152.
20. Cramer III, R.D., *BC (DEF) Parameters: 1. The intrinsic dimensionality of intermolecular interactions in the liquid state*, J. Am. Chem. Soc., 102 (1980) 1837–1849.
21. Stewart. J.J.P., *Optimisation of Parameters for Semiempirical Methods: 2. Applications*, J. Comp. Chem., 10 (1989) 221–264.
22. Stewart. J.J. P., *MOPAC: A semiempirical molecular orbital program*, J. Comput.-Aided Mol. Design. 4 (1990) 1–105.
23. Turner, D.B., Willett, P., Ferguson, A.M. and Heritage. T.W., *Evaluation of a novel infra-red range vibration-based descriptor (EVA) for QSAR studies: 1. General application*, J. Comput. -A ided Mol. Design, 11 (1997) 409–422.
24. Turner, D.B., *An Evaluation of a novel molecular descriptor (EVA) for QSAR studies and the similarity searching of chemical structure databases*, PhD, thesis, University of Sheffield, 1996.
25. Jonathan. P., McCarthy, W.V. and Roberts, A.M.I., *Discriminant analysis with singular covariance matrices: A method incorporating crossvalidation and efficient randomized permutation tests*, J. Chemometrica, 10 (1996) 189–214.
26. Waller, C.L. and McKinney, J.D., *Comparative molecular field analysis of polyhalogenated dibenzop-dioxins, dibenzofurans and biphenyls*, J. Med. Chem., 35 (1992) 2660–3666.
27. Carroll, F.I., Gao, Y.G., Rahman, M.A., Abraham, P., Parham, K., Lewin, A.H., Boja, J.W. and Kuhar, M.J., *Synthesis, ligand-binding, QSAR and CoMFA study of 3-β-(para-substituted phenyl)tropane-2-β-carboxylic acid methyl-esters*, J. Med. Chem., 34 (1991) 2719–2725.

28. Allen, M.S., Laloggia, A.J., Dorn, L.J., Matin, M.J., Costantin, G., Hagen, T.J., Koehkr, K.F., Skolnick, P. and Cook, J.M., *Predictive binding of  $\beta$ -carboline inverse agonists and antagonists via the CoMFA/GOLPE approach*, *J. Med. Chem.*, 35 (1992) 4001–4010.
29. Greco, G., Novellino, E., Silipo, C. and Vittoria, A., *Comparative molecular-field analysis on a set of muscarinic agonists*, *Quant. Struct.-Act. Relat.*, 10 (1991) 289–299.
30. Cho, S. and Trophsha, A., *Crossvalidated  $R^2$ -guided region selection for comparative molecular field-analysis: A simple method to achieve consistent results*, *J. Med. Chem.*, 38 (1995) 1060–1066.
31. Turner, D.B., Willett, P., Ferguson, A.M. and Heritage, T.W., *Evaluation of a novel infra-red range vibration-based descriptor (EVA) for QSAR studies: 2. Model validation*, *J. Med. Chem.* (submitted).
32. Kroemer, R.T. and Hecht, P., *Replacement of steric 6-12 potential-derived interaction energies by atom-based indicator variables in CoMFA leads to models of higher consistency*, *J. Comput.-Aided Mol. Design*, 9 (1995) 205–212.
33. Heritage, T.W., Shell Research Ltd. Internal Report, 1992 (not publicly available).
34. Wold, S., Johansson, E. and Cocchi, M., *PLS—partial least squares to latent structures*, In Kubinyi, H. (Ed.) *3D QSAR in drug design*, ESCOM, Leiden, 1993, pp. 523–550.
35. Michelson, K.E., Forsthoefel, J. and Westphal, U., *Steroid–protein interactions: Human corticosteroid binding globulin: Some physicochemical properties and binding specificity*, *Biochemistry*, 20 (1981) 6211–6218.
36. Ginn, C.M.R., Turner, D.B., Willett, P., Ferguson, A.M. and Heritage, T.W., *Similarity searching in files of three-dimensional chemical structures: Evaluation of the EVA descriptor and combination of rankings using data fusion*, *J. Chem. Inf. Comput. Sci.* 37 (1997) 23–37.
37. Dunn, W.J., Hopfinger, A.J., Catana, C. and Duraiswami, C., *Solution of the conformation and alignment tensors for the binding of trimethoprim and its analysis to dihydrofolate-reductase — 3D-quantitative structure–activity study using molecular-shape analysis 3-way partial least-squares regression, and 3-way factor analysis*, *J. Med. Chem.*, 39 (1996) 4825–4832.

## Author Index

- Alber, F. 169  
Andreoni, W. 161  
Anzali, S. 273
- Beck, B. 131
- Carloni, P. 169  
Clark, R.D. 213  
Clark, T. 131  
Cramer, R.D. 213
- Erion, M.D. 85
- Ferguson, A.M. 213, 381
- Gago, F. 19  
Gasteiger, J. 273  
Ghose, A.K. 253  
Good, A.C. 321  
Gramatica, P. 355  
Grootenhuis, P.D.J. 99
- Harrison, R.W. 115  
Heritage, T.W. 381  
Holloway, M.K. 63  
Holzgrabe, U. 273
- Knegt, R.M.A. 99  
Kubinyi, H. 225
- Liljefors, T. 3
- Marshall, G.R. 35
- Oprea, T.I. 35  
Ortiz, A.R. 19
- Pearlman, R.S. 339  
Polanski, J. 273
- Reddy, M.R. 85  
Richards, W.G. 321  
Rognan, D. 181
- Sadowski, J. 273  
Smith, K.M. 339
- Teckentrup, A. 273  
Thornier, D.A. 301  
Todeschini, R. 355  
Turner, D.B. 381
- Viswanadhan, V.N. 85
- Wade, R.C. 19  
Wagener, M. 273  
Weber, I.T. 115  
Wendoloski, J.J. 253  
Wild, D.J. 301  
Willett, P. 301, 381  
Wright, P.M. 301

**This Page Intentionally Left Blank**