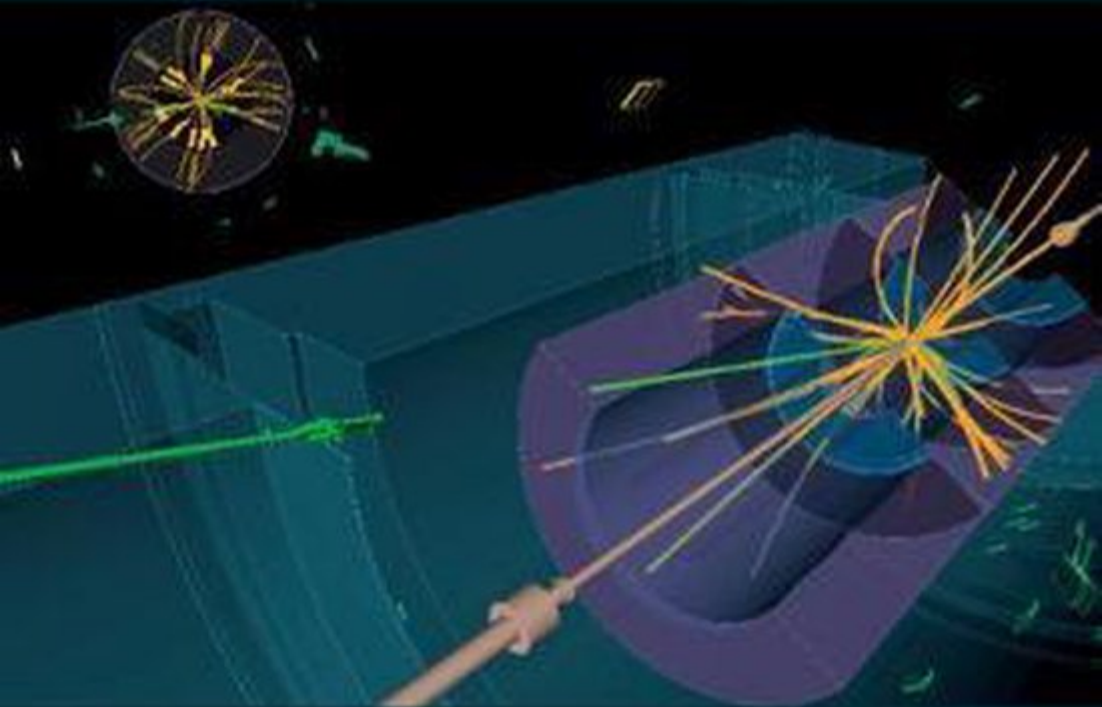


Fourth Edition

The Ideas of

# Particle Physics



James Dodd  
Ben Gripaios

# The Ideas of Particle Physics

This book is a comprehensive introduction to particle physics, bridging the gap between traditional textbooks on the subject and popular accounts that assume little background knowledge. This fourth edition is fully revised, including the most recent ideas and discoveries, and the latest avenues of research. The development of the subject is traced from the foundations of quantum mechanics and relativity, through the formulation of quantum field theories, to the Standard Model. Research now continues with the first signs of physics beyond the Standard Model and with the formulation of modern string theory which aims to include a quantum theory of gravity for the first time. This book is intended for anyone with a background in physical sciences who wishes to learn about particle physics. It is also valuable to students of physics wishing to gain an introductory overview of the subject.

**James Dodd** is co-founder of the St Cross College Centre for the History and Philosophy of Physics. He has studied Physics at the Universities of London, Oxford and Cambridge. His research focused on the production of charmed particles in high energy hadron collisions.

**Ben Gripaios** is Professor of Theoretical Physics at the Cavendish Laboratory, University of Cambridge and is a Fellow of King's College, Cambridge. He has previously held research positions at CERN, Lausanne and Oxford. His research focuses on the search for physics beyond the Standard Model.



# *The Ideas of Particle Physics*

***James Dodd***

St Cross College, University of Oxford

***Ben Gripaios***

Cavendish Laboratory, University of Cambridge

FOURTH EDITION



**CAMBRIDGE**  
UNIVERSITY PRESS



**CAMBRIDGE**  
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre, New Delhi – 110025, India

79 Anson Road, #06–04/06, Singapore 079906

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning, and research at the highest international levels of excellence.

[www.cambridge.org](http://www.cambridge.org)

Information on this title: [www.cambridge.org/9781108727402](http://www.cambridge.org/9781108727402)

DOI: [10.1017/9781108616270](https://doi.org/10.1017/9781108616270)

© Cambridge University Press 1984, 1991, 2006, 2020

This publication is in copyright. Subject to statutory exception and to the provisions of relevant collective licensing agreements, no reproduction of any part may take place without the written permission of Cambridge University Press.

First published 1984

Second edition published 1991

Third edition published 2006

Reprinted 2009

Fourth edition published 2020

Printed in the United Kingdom by TJ International Ltd, Padstow Cornwall

*A catalogue record for this publication is available from the British Library.*

*Library of Congress Cataloging-in-Publication Data*

Names: Dodd, J. E. (James Edmund), 1952– author. | Gripaios, B. M., author.

Title: The ideas of particle physics / James Dodd and Ben Gripaios.

Description: 4th edition. | New York : Cambridge University Press, 2020. |

Includes bibliographical references and index.

Identifiers: LCCN 2019042615 (print) | LCCN 2019042616 (ebook) |

ISBN 9781108727402 (paperback) | ISBN 9781108616270 (epub)

Subjects: LCSH: Particles (Nuclear physics)

Classification: LCC QC793.2 .D6 2020 (print) | LCC QC793.2 (ebook) |

DDC 539.7/2–dc23

LC record available at <https://lcn.loc.gov/2019042615>

LC ebook record available at <https://lcn.loc.gov/2019042616>

ISBN 978-1-108-72740-2 Paperback

Cambridge University Press has no responsibility for the persistence or accuracy of URLs for external or third-party internet websites referred to in this publication and does not guarantee that any content on such websites is, or will remain, accurate or appropriate.

*To our families and to  
our tutor, Jack Paton.*



# Contents

*Preface* *page* xiii

<b>Part I Introduction</b>	1
1 Matter and Light	3
1.1 Introduction	3
1.2 The Nature of Matter	3
1.3 Atomic Radiations	4
1.4 Rutherford's Atom	6
1.5 Two Problems	7
2 Special Relativity	8
2.1 Introduction	8
2.2 Galilean Relativity	8
2.3 The Origins of Special Relativity	10
2.4 The Lorentz–Fitzgerald Contraction	10
2.5 The Special Theory of Relativity	11
2.6 Mass Momentum and Energy	12
2.7 The Physical Effects of Special Relativity	13
2.8 Using Relativity	14
3 Quantum Mechanics	16
3.1 Introduction	16
3.2 Planck's Hypothesis	16
3.3 Einstein's Explanation of the Photoelectric Effect	17
3.4 Bohr's Atom	17
3.5 De Broglie's Electron Waves	19
3.6 Schrödinger's Wavefunction	20
3.7 Heisenberg's Mechanics and the Uncertainty Principle	21
3.8 The Interpretation of the Wavefunction $\psi$	22

3.9	Electron Spin	23	10.2	Internal Symmetry	65
3.10	The Pauli Exclusion Principle	24	10.3	Quarks	66
4	Relativistic Quantum Theory	26		<b>Part IV Weak Interaction Physics I</b>	69
4.1	Introduction	26	11	The Violation of Parity	71
4.2	The Dirac Equation	26	11.1	Introduction	71
4.3	Antiparticles	27	11.2	$\beta$ Decay of Cobalt	71
4.4	Quantum Field Theory (QFT)	29	11.3	Absolute-handedness and CP Invariance	73
4.5	Interacting Fields	30	12	Fermi's Theory of the Weak Interactions	75
4.6	Perturbation Theory	30	12.1	Introduction	75
4.7	Virtual Processes	32	12.2	Fermi's Theory of $\beta$ Decay	75
4.8	Renormalisation	32	12.3	Spin, Helicity and Chirality	76
4.9	The Quantum Vacuum	33	12.4	The Polarisation of $\beta$ -decay Electrons	76
4.10	Quantum Electrodynamics	34	12.5	Neutrino Helicity	77
4.11	Postscript	35	12.6	In Conclusion	78
	<b>Part II Basic Particle Physics</b>	37	13	Two Neutrinos	79
5	The Fundamental Forces	39	13.1	Introduction	79
5.1	Introduction	39	13.2	A Problem in the Weak Interactions	79
5.2	Gravity	39	13.3	The Two-neutrino Experiment	80
5.3	Electromagnetism	41	14	Neutral Kaons and CP Violation	82
5.4	The Strong Nuclear Force	43	14.1	Introduction	82
5.5	The Weak Nuclear Force	45	14.2	What is a Neutral Kaon?	82
6	Symmetry in the Microworld	47	14.3	Violation of CP Symmetry	83
6.1	Introduction	47		<b>Part V Weak Interaction Physics II</b>	85
6.2	Space–Time Symmetries	47	15	The Current–Current Theory of the Weak Interactions	87
6.3	Discrete Symmetries	48	15.1	Introduction	87
6.4	The CPT Theorem	50	15.2	The Lepton Current	87
6.5	Dynamical Symmetries	50	15.3	Higher-order Interactions	88
6.6	Internal Symmetries	51	16	An Example Leptonic Process: Electron-neutrino Scattering	90
6.7	Broken Symmetries	51	16.1	Introduction	90
7	Mesons	52	16.2	The Role of the Weak Force in Astrophysics	91
7.1	Introduction	52	17	The Weak Interactions of Hadrons	92
7.2	Yukawa's Proposal	52	17.1	Introduction	92
7.3	The Muon	52	17.2	The Hadronic Current	92
7.4	The Real Pion	53	17.3	The Hadron Current and Quarks	93
7.5	Terminology	54	18	The W Boson	94
7.6	Isotopic Spin	55	18.1	Introduction	94
8	Strange Particles	56	18.2	The W Boson	94
8.1	Introduction	56	18.3	Observing the W Boson	95
8.2	Associated Production	56		<b>Part VI Gauge Theory of the Weak Interactions</b>	97
8.3	The Kaons	57	19	Motivation for the Theory	99
8.4	The Hyperons	59	19.1	Introduction	99
8.5	Summary	59	19.2	Problems with the W Bosons	99
	<b>Part III Strong Interaction Physics</b>	61			
9	Resonance Particles	63			
9.1	Introduction	63			
9.2	Resonance Particle Experiments	63			
10	SU(3) and Quarks	65			
10.1	Introduction	65			

20	Gauge Theory	101	29.2	Electromagnetic Structure Functions	138
20.1	Introduction	101	29.3	Weak Interaction Structure Functions	139
20.2	The Formulation of QED	101	29.4	Electron and Neutrino Structure Functions Compared	140
20.3	Generalised Gauge Invariance	102	29.5	Sum Rules	140
20.4	Gauge Invariance and the Weak Interactions	103	29.6	Summary	141
21	Spontaneous Symmetry Breaking	105	<b>Part VIII Quantum Chromodynamics – the Theory of Quarks</b>		
21.1	Introduction	105	30	Coloured Quarks	145
21.2	Spontaneous Breaking of Global Symmetry	105	30.1	Introduction	145
21.3	Spontaneous Breaking of Local Symmetry – the Higgs Mechanism	106	30.2	Colour	145
22	The Glashow–Weinberg–Salam Model	108	30.3	Invisible Colour	147
22.1	Introduction	108	31	Colour Gauge Theory	150
22.2	Formulation	108	31.1	Introduction	150
22.3	Reprise	111	31.2	The Formulation of QCD	150
22.4	An Academic Postscript – Renormalisability	111	32	Asymptotic Freedom	154
23	Consequences of the Model	112	32.1	Introduction	154
23.1	Introduction	112	32.2	Violations of Scaling	157
23.2	Neutral Currents	112	33	Quark Confinement	160
23.3	The Incorporation of Hadrons – Charm	113	33.1	Introduction	160
23.4	Parity-violating Tests of the Glashow–Weinberg–Salam Model	115	33.2	Quark Forces – Hadron Forces	162
24	The Hunt for the $W^\pm, Z^0$ Bosons	116	<b>Part IX Electron–Positron Collisions</b>		
24.1	Introduction	116	34	Probing the Vacuum	167
24.2	The CERN $p\bar{p}$ Collider Experiment	116	34.1	Introduction	167
24.3	Detecting the Bosons	118	34.2	The Experiments	167
24.4	Epilogue	121	34.3	The Basic Reactions	168
<b>Part VII Deep Inelastic Scattering</b>		123	35	Quarks and Charm	171
25	Deep Inelastic Processes	125	35.1	Introduction	171
25.1	Introduction	125	35.2	The Quark Picture	171
25.2	Two Key Ideas	126	35.3	The Advent of Charm	172
26	Electron–Nucleon Scattering	127	35.4	Psychology	174
26.1	Introduction	127	35.5	Charmed Particles	176
26.2	The Scaling Hypothesis	127	36	Another Generation	178
26.3	Exploring the Structure Functions	129	36.1	Introduction	178
27	The Deep Inelastic Microscope	131	36.2	The Upsilon	178
27.1	Introduction	131	36.3	The Tau Heavy Lepton	179
27.2	Free Quarks and Strong Forces	131	36.4	Completing the Third Generation	181
28	Neutrino–Nucleon	134	<b>Part X The Standard Model</b>		
28.1	Introduction	134	37	The Model in Summary	185
28.2	Neutrino Experiments	134	37.1	Introduction	185
28.3	The Cross-section	135	37.2	Summary of the Standard Model	185
28.4	The Scaling Hypothesis	135	37.3	Consistency of the Standard Model	186
29	The Quark Model of the Structure Functions	138	38	Precision Tests of the Model	189
29.1	Introduction	138	38.1	Introduction	189
			38.2	Precision Tests of the Gauge Interactions	191



52.6	The Anthropic Principle	251	59.1	The Miracle of Duality	278
52.7	Summary	251	59.2	The String Theory Side	278
	<b>Part XIII Gravity and Gravitational Waves</b>	253	59.3	The Quantum Field Theory Side	279
53	From General Relativity to Gravitational Waves	255	59.4	The AdS–CFT Dictionary	279
53.1	Introduction	255	59.5	Applications of AdS–CFT	279
53.2	Hulse–Taylor Variation in Binary Pulsar Periodicity	256	60	Consequences of the Theory	280
53.3	Modern Astrophysics	256	60.1	The Richness of String and M-theory	280
54	The Discovery of Gravitational Waves	259	60.2	Back to the Anthropic Principle	281
54.1	Introduction	259	60.3	A Theory in Search of Experiment	281
54.2	LIGO	259	60.4	Conclusion	282
54.3	The Detection of GW150914	259		<b>Part XV The Future: To Boldly Go!</b>	283
54.4	Subsequent Events	262	61	Accelerators, Observatories and Other Experiments	285
55	Gravitational-wave and Multi-messenger Astronomy	263	61.1	Accelerators	285
55.1	Introduction	263	61.2	Observatories and Other Experiments	286
55.2	Gravitational-wave Astronomy	263	62	Known Unknowns	289
55.3	GW170817 – a Binary Neutron Star Merger	264	62.1	The Current In-tray	289
56	The Future: Super LIGO and LISA	266	63	Glittering Prizes	291
	<b>Part XIV String Theory</b>	269	63.1	The Class of 1984	291
57	Origins – the Hadronic String	271	64	Unknown Unknowns: It Must Be Beautiful	292
57.1	The Success of QFT	271	64.1	The Challenges of Quantum Gravity	292
57.2	The Problem of Gravity	271	64.2	The Beautiful Equations	293
57.3	Strings versus Particles	272		<b>Appendices</b>	295
57.4	The Hadronic String	273	A	Units and Constants	297
58	String Theory to M-theory	275	B	Glossary	298
58.1	The Search for a Consistent String Theory	275	C	List of Symbols	306
58.2	String Theories Contain More Than String	276	D	Bibliography	308
59	The AdS–CFT Correspondence	278	E	Elementary Particle Data	312
				<i>Name Index</i>	313
				<i>Subject Index</i>	314





# *Preface*

Forty years is a suitable period in which to judge the progress of a subject such as modern particle physics. After all, there have only been eight such periods since the publication of Newton's *Principia* in 1687, in nearly all of which the advances of physics have led to the modern world we know today. It is just such a period which has seen the gestation period of this book and its publication through its three previous editions.

We now have a convincing picture of the fundamental structure of observable matter in terms of certain point-like elementary particles. We also have a comprehensive theory describing their behaviour and the forces which act between them. This we believe provides a complete and correct description of nearly all non-gravitational physics.

Matter, so it seems, consists of just two types of elementary particles: fermions (such as quarks and leptons) and bosons (such as photons and others). These are the fundamental building blocks of our material world. The theory describing the microscopic behaviour of these particles has become known as the 'Standard Model' which provides an accurate account of the force of electromagnetism, the weak nuclear force (responsible for radioactive decay), and the strong nuclear force (which holds atomic nuclei together). The Standard Model has been remarkably successful and, as we shall see, has achieved exceptionally high agreement between theoretical predictions and experimental measurements.

The Standard Model is based on the principle of ‘gauge symmetry’, which asserts that the properties and interactions of elementary particles are governed by certain fundamental symmetries related to familiar conservation laws. Thus, the strong, weak and electromagnetic forces are all ‘gauge’ forces. They are mediated by the exchange of certain particles, called gauge bosons, which are, for example, responsible for the interaction between two electric charges, and for the nuclear processes taking place within the Sun. Unsuccessful attempts have been made to fit the only other known force – gravity – into this gauge framework. However, despite our clear understanding of certain macroscopic aspects of gravity, a microscopic theory of gravity has so far proved elusive. Moreover, recent experiments in neutrino physics cannot be explained within the Standard Model, showing beyond doubt that there must be a theory beyond the Standard Model, and that the Standard Model itself is only an approximation (albeit a very good one) to the true theory.

The above picture of the microworld has emerged slowly since the late 1960s, at which time only the electromagnetic force was well understood. It is the story of the discoveries which have been made since that time to which this book is devoted. The telling of the story is broadly in chronological order, but where appropriate this gives way to a more logical exposition in which complete topics are presented in largely self-contained units. The advances described in Parts VI–IX, for example, were made more or less simultaneously, but no attempt is made here to relate an accurate history. Instead, we focus on the logical development of the individual topics and give only the main historical interconnections.

Our main concern in writing this book has been to communicate the central ideas and concepts of elementary particle physics. We have attempted to present a comprehensive overview of the subject at a level which carries the reader beyond the simplifications and generalisations necessary in popular science books. It is aimed principally at graduates in the physical sciences, mathematics, engineering, or other numerate subjects. But we must stress that this is not a textbook. It makes no claim to the precision and rigour that a textbook requires. It contains no mathematical derivations, and no complicated formulae are written down (other than for the purpose of illustration). Nevertheless, simple mathematical

equations are frequently employed to aid in the explanation of a particular idea, and the book does assume a familiarity with basic physical concepts (such as mass, momentum, energy, etc.).

This book is organised in 15 parts each consisting of four or five short chapters. Parts I–IX deal with the evolution of the subject up until the first edition in the early 1980s. Parts X onwards cover the ideas and experiments of the last 40 years. Dealing with the most exciting of current research topics, it contains chapters which are rather longer than average and which will require more time and concentration on the part of the reader. Part XV looks ahead to what might emerge in the next 40! We draw the reader’s attention to the Glossary (Appendix B), which gives concise definitions of the most important of particle physics nomenclature. It should prove useful as a memory prompt, as well as a source of supplementary information.

The story begins in Part I at the turn of the last century when physicists were first beginning to glimpse the remarkable nature of ordinary matter. Out of this period came the two elements essential for the understanding of the microworld: the theories of special relativity and quantum mechanics. These are the unshakeable foundations upon which the rest of the story is based.

Part II introduces the four known fundamental forces, and is followed by a more detailed discussion of the physics of the strong and weak (nuclear) forces in Parts III–V. It was the desire to understand the weak force, in particular, which led eventually to recognition of the role of gauge symmetry as a vital ingredient in theories of the microworld. Gauge theory is the subject of Part VI, which introduces the Glashow–Weinberg–Salam theory of the electromagnetic and weak forces. This theory, often called the ‘electroweak model’, has been spectacularly verified in many experiments over the past four decades. The most impressive of these was the discovery at CERN in 1983 of the massive W and Z gauge bosons which mediate the weak force and the subsequent discovery in 2012 of the Higgs boson, the final particle of the model to be confirmed.

At about the same time as the electroweak model was being developed, physicists were using ‘deep inelastic scattering’ experiments to probe the interior of the proton. These experiments, which are described in Part VII, provided the first indication that the proton

was not truly elementary, but composed of point-like objects (called quarks). As the physical reality of quarks gained wider acceptance, a new gauge theory was formulated in an attempt to explain the strong forces between them. This theory is called ‘quantum chromodynamics’ and attributes the strong force to the exchange of certain gauge bosons called gluons. It is described in Part VIII. Together, quantum chromodynamics and the Glashow–Weinberg–Salam electroweak theory constitute the ‘Standard Model’ of elementary particle physics. Part IX describes early experiments involving collisions between electrons and positrons. These experiments were instrumental in confirming the physical reality of quarks and in testing many of the predictions of quantum chromodynamics and the electroweak theory.

Part X begins by summarising the Standard Model and describes the many tests of the model performed first in the previous generation of electron–positron colliders of the 1990s and early 2000s and more recently in the Large Hadron Collider (LHC) at CERN, which culminated in 2012 with the discovery of the Higgs. Part XI, however, goes on to explain that, despite its success, the Standard Model cannot account for all observed phenomena, indicating the need for ideas and theories beyond, including candidates such as grand unification, supersymmetry, composite Higgs

models and the possible existence of new particles such as axions.

Part XII changes perspective entirely and explains the exchange of ideas between particle physics and astrophysics and cosmology, particularly ideas under consideration to explain the current major unexplained mysteries of dark matter and dark energy. Part XIII explores further the cosmological realm explaining the origin and the recent, astonishing discovery of gravitational waves and their implications for the very largest objects in the universe (e.g. super-massive black holes) and how their behaviour may give valuable clues to the microscopic theories beyond the Standard Model.

The famous, but famously sophisticated, field of string theory attempts to unite the quantum theories of the microworld with a quantum theory of gravitation and this is explained in Part XIV. Despite its mathematical elegance and the intense work devoted to it, the theory has yet to reveal a definitively testable prediction. But as described in the final Part XV, mathematical constructs can often presage as yet undiscovered physical reality. The final part goes on to describe in brief plans for the current and next generation of accelerators and experiments and a checklist of possible discoveries which may guide us on the path of the next 40 years.



**Part I**  
**Introduction**



# 1

## *Matter and Light*

### 1.1 Introduction

The physical world we see around us has two main components, matter and light, and it is the modern explanation of these things which is the purpose of this book. During the course of the story, these concerns will be restated in terms of material particles and the forces which act between them, and we will most assuredly encounter new and exotic forms of both particles and forces. But in case we become distracted and confused by the elaborate and almost wholly alien contents of the microworld, let us remember that the origin of the story, and the motivation for all that follows, is the explanation of everyday matter and visible light.

Beginning as it does, with a laudable sense of history, at the turn of the last century, we have only to appreciate the level of understanding of matter and light around 1900, and some of the problems in this understanding, to prepare ourselves for the story of progress which follows.

### 1.2 The Nature of Matter

By 1900 most scientists were convinced that all matter is made up of a number of different sorts of atoms, as had been conjectured by the ancient Greeks millennia before and as had been indicated by chemistry experiments over the preceding two centuries. In the atomic picture, the different types of substance can be seen as arising from different arrangements of the atoms. In solids, the atoms are relatively immobile and

in the case of crystals are arranged in set patterns of impressive precision. In liquids they roll loosely over one another and in gases they are widely separated and fly about at a velocity depending on the temperature of the gas; see Figure 1.1. The application of heat to a substance can cause phase transitions in which the atoms change their mode of behaviour as the heat energy is transferred into the kinetic energy of the atoms' motions.

Many familiar substances consist not of single atoms, but of definite combinations of certain atoms called molecules. In such cases it is these molecules which behave in the manner appropriate to the type of substance concerned. For instance, water consists of molecules, each made up of two hydrogen atoms and one oxygen atom. It is the molecules which are subject to a specific static arrangement in solid ice, the molecules which roll over each other in water and the molecules which fly about in steam.

The laws of chemistry, most of which were discovered empirically between 1700 and 1900, contain many deductions concerning the behaviour of atoms and molecules. At the risk of brutal over-simplification the most important of these can be summarised as follows:

- (1) Atoms can combine to form molecules, as indicated by chemical elements combining only in certain proportions (Richter and Dalton).



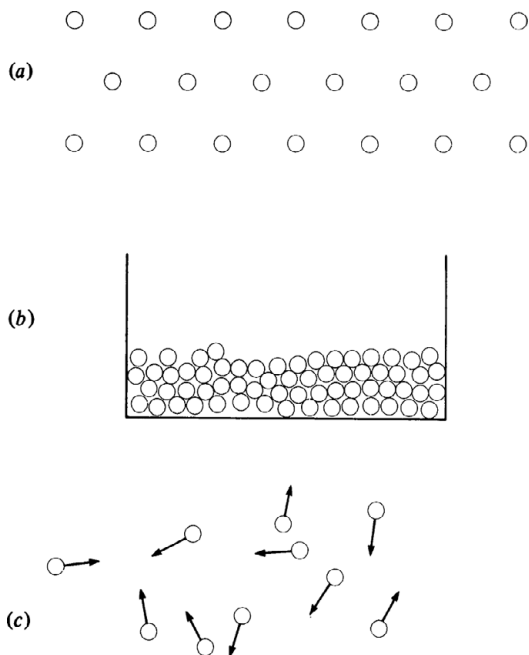


Figure 1.1. (a) Static atoms arranged in a crystal. (b) Atoms rolling around in a liquid. (c) Atoms flying about in a gas.

- (2) At a given temperature and pressure, equal volumes of gas contain equal numbers of molecules (Avogadro).
- (3) The relative weights of the atoms are approximately multiples of the weight of the hydrogen atom (Prout).
- (4) The mass of each atom is associated with a specific quantity of electrical charge (Faraday and Webber).
- (5) The elements can be arranged in families having common chemical properties but different atomic weights (Mendeleeff's periodic table).
- (6) An atom is approximately  $10^{-10}$  m across, as implied by the internal friction of a gas (Loschmidt).

One of the philosophical motivations behind the atomic theory (a motivation we shall see repeated later) was the desire to explain the diversity of matter by assuming the existence of just a few fundamental and indivisible atoms. But by 1900 over 90 varieties of atoms were known, an uncomfortably large number for a supposedly fundamental entity. Also, there was evidence for the disintegration (divisibility) of atoms.

At this breakdown of the 'ancient' atomic theory, modern physics begins.

### 1.3 Atomic Radiations

#### 1.3.1 Electrons

In the late 1890s, J. J. Thomson of the Cavendish Laboratory at Cambridge was conducting experiments to examine the behaviour of gas in a glass tube when an electric field was applied across it. He came to the conclusion that the tube contained a cloud of minute particles with negative electrical charge – the electrons. As the tube had been filled only with ordinary gas atoms, Thomson was forced to conclude that the electrons had originated within the supposedly indivisible atoms. As the atom as a whole is electrically neutral, on the release of a negatively charged electron the remaining part, the ion, must carry the equal and opposite positive charge. This was entirely in accord with the long-known results of Faraday's electrolysis experiments, which required a specific electrical charge to be associated with the atomic mass.

By 1897, Thomson had measured the ratio of the charge to the mass of the electron (denoted  $e/m$ ) by observing its behaviour in magnetic fields. By comparing this number with that of the ion, he was able to conclude that the electron is thousands of times less massive than the atom (and some 1837 times lighter than the lightest atom, hydrogen). This led Thomson to propose his 'plum-pudding' picture of the atom, in which the small negatively charged electrons were thought to be dotted in the massive, positively charged body of the atom (see Figure 1.2).

#### 1.3.2 X-rays

Two years earlier in 1895, the German Wilhelm Röntgen had discovered a new form of penetrating radiation, which he called X-rays. This radiation was emitted when a stream of fast electrons (which had not yet been identified as such) struck solid matter and were thus rapidly decelerated. This was achieved by boiling the electrons out of a metallic electrode in a vacuum tube and accelerating them into another electrode by applying an electric field across the two, as in Figure 1.3. Very soon the X-rays were identified as another form of electromagnetic radiation, i.e. radiation that is basically the same as visible light, but with a much higher frequency and shorter wavelength. An impressive demonstration of the wave nature of X-rays

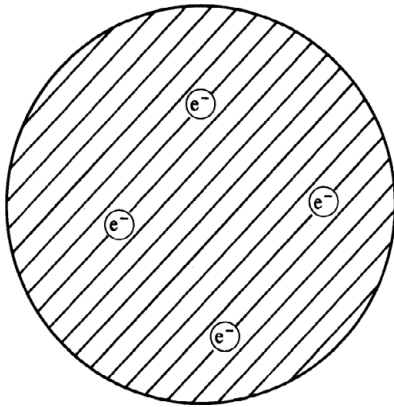


Figure 1.2. Thomson's 'plum-pudding' picture of the atom.

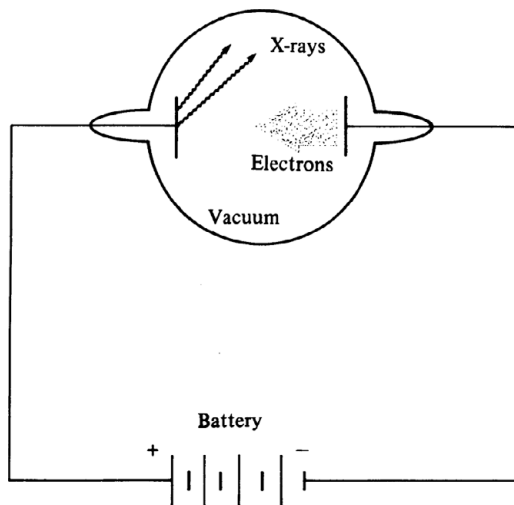


Figure 1.3. The production of X-rays by colliding fast electrons with matter.

was provided in 1912 when the German physicist Max von Laue shone them through a crystal structure. In doing so, he noticed the regular geometrical patterns characteristic of the diffraction which occurs when a wave passes through a regular structure whose characteristic size is comparable to the wavelength of the wave. In this case, the regular spacing of atoms within the crystal is about the same as the wavelength of the X-rays. Although these X-rays do not originate from within the structure of matter, we shall see next how they are the close relatives of radiations which do.

### 1.3.3 Radioactivity

At about the same time as the work taking place on electrons and X-rays, the French physicist Becquerel was conducting experiments on the heavy elements. During his study of uranium salts in 1896, Becquerel noticed the emission of radiation rather like that which Röntgen had discovered. But Becquerel was doing nothing to his uranium: the radiation was emerging spontaneously. Inspired by this discovery, Pierre and Marie Curie began investigating the new radiation. By 1898, the Curies had discovered that the element radium also emits copious amounts of radiation.

These early experimenters first discovered the radiation through its darkening effect on photographic plates. However, other methods for detecting radiation were soon developed, including scintillation techniques, electroscopes and a primitive version of the Geiger counter. Then a great breakthrough came in 1912 when C. T. R. Wilson of the Cavendish Laboratory invented the cloud chamber. This device encourages easily visible water droplets to form around the atoms, which have been ionised (i.e. have had an electron removed) by the passage of the radiation through air. This provides a plan view of the path of the radiation and so gives us a clear picture of what is happening.

If a radioactive source such as radium is brought close to the cloud chamber, the emitted radiation will trace paths in the chamber. When a magnetic field is placed across the chamber, then the radiation paths will separate into three components which are characteristic of the type of radiation (see Figure 1.4). The first component of radiation (denoted  $\alpha$ ) is bent slightly by the magnetic field, which indicates that the radiation carries electric charge. Measuring the radius of curvature of the path in a given magnetic field can tell us that it is made up of massive particles with two positive electric charges. These particles can be identified as the nuclei of helium atoms, often referred to as  $\alpha$  particles. Furthermore, these  $\alpha$  particles always seem to travel a fixed distance before being stopped by collisions with the air molecules. This suggests that they are liberated from the source with a constant amount of energy and that the same internal reactions within the source atoms are responsible for all  $\alpha$  particles.

The second component of the radiation (denoted  $\gamma$ ) is not at all affected by the magnetic field,

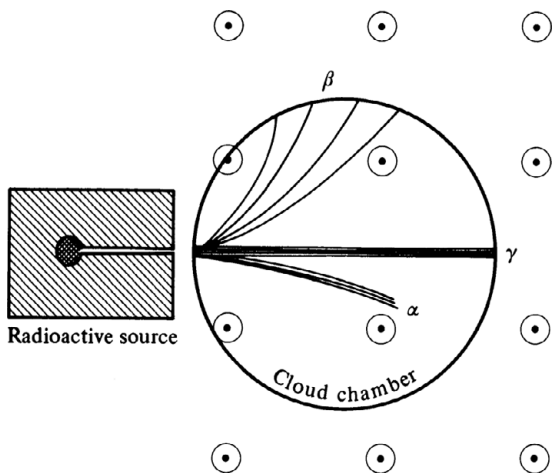


Figure 1.4. Three components of radioactivity displayed in a cloud chamber.  $\odot$  signifies that the direction of the applied magnetic field is perpendicular to, and out of the plane of, the paper.

showing that it carries no electric charge, and it is not stopped by collisions with the air molecules. These  $\gamma$ -rays were soon identified as the close relatives of Röntgen's X-rays but with even higher frequencies and even shorter wavelengths. The  $\gamma$ -rays can penetrate many centimetres of lead before being absorbed. They are the products of reactions occurring spontaneously within the source atoms, which liberate large amounts of electromagnetic energy but no material particles, indicating a different sort of reaction to that responsible for  $\alpha$ -rays.

The third component (denoted  $\beta$  radiation) is bent significantly in the magnetic field in the opposite direction to the  $\alpha$ -rays. This is interpreted as single, negative electrical charges with much lesser mass than the  $\alpha$ -rays. They were soon identified as the same electrons as those discovered by J. J. Thomson, being emitted from the source atoms with a range of different energies. The reactions responsible form a third class distinct from the origins of  $\alpha$ - or  $\gamma$ -rays.

The three varieties of radioactivity have a double importance in our story. Firstly, they result from the three main fundamental forces of nature effective within atoms. Thus the phenomenon of radioactivity may be seen as the cradle for all of what follows. Secondly, and more practically, it was the products of radioactivity which first allowed physicists to explore the interior of atoms and which later indicated

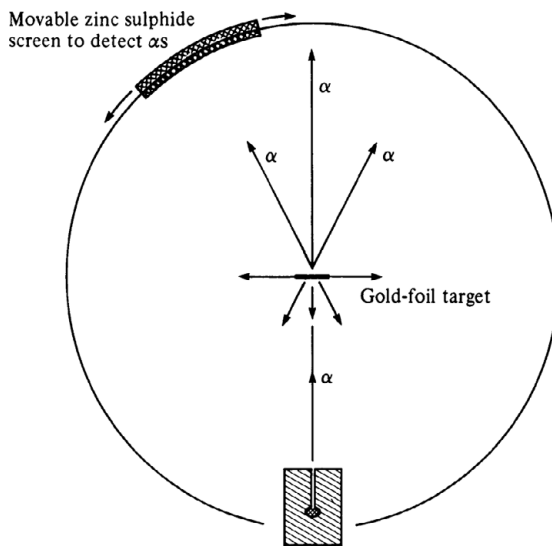


Figure 1.5. The Geiger and Marsden experiment. According to Rutherford's scattering formula, the number of  $\alpha$  particles scattered through a given angle decreases as the angle increases away from the forward direction.

totally novel forms of matter, as we shall see in due course.

#### 1.4 Rutherford's Atom

In the first decade of the twentieth century, Rutherford had pioneered the use of naturally occurring atomic radiations as probes of the internal structure of atoms. In 1909, at Manchester University, he suggested to his colleagues, Geiger and Marsden, that they allow the  $\alpha$  particles emitted from a radioactive element to pass through a thin gold foil and observe the deflection of the outgoing  $\alpha$  particles from their original paths (see Figure 1.5). On the basis of Thomson's 'plum-pudding' model of the atom, they should experience only slight deflections, as nowhere in the uniformly occupied body of the atom would the electric field be enormously high. But the experimenters were surprised to find that the heavy  $\alpha$  particles were sometimes drastically deflected, occasionally bouncing right back towards the source. In a dramatic analogy attributed (somewhat dubiously) to Rutherford: 'It was almost as incredible as if you fired a 15-inch shell at a piece of tissue paper and it came back and hit you!'

The implication of this observation is that a very strong repulsive force must be at work within the atom. This force cannot be due to the electrons as they are

over 7000 times lighter than the  $\alpha$  particles and so can exert only minute effects on the  $\alpha$ -particle trajectories. The only satisfactory explanation of the experiment is that all the positive electric charge in the atom is concentrated in a small nucleus at the middle, with the electrons orbiting the nucleus at some distance. By assuming that the entire positive charge of the atom is concentrated with the atomic mass in a small central nucleus, Rutherford was able to derive his famous scattering formula which describes the relative numbers of  $\alpha$  particles scattered through given angles on colliding with an atom (see Figure 1.5).

Rutherford's picture of the orbital atom is in contrast with our perception of apparently 'solid' matter. From the experiments he was able to deduce that the atomic nucleus, which contains 99.9% of the mass of the atom, has a diameter of about  $10^{-15}$  m compared to an atomic diameter of about  $10^{-10}$  m. For illustration, if we took a cricket ball to act as the nucleus, the atomic electrons would be 5 km distant! Such an analogy brings home forcibly just how sparse apparently solid matter is and just how dense is the nucleus itself. But despite this clear picture of the atom, indicated from the experiment, explaining how it works is fraught with difficulties, as we shall see in Chapter 3.

## 1.5 Two Problems

Just as these early atomic experiments revealed an unexpected richness in the structure of matter, so too, theoretical problems forced upon physicists more-sophisticated descriptions of the natural world. The theories of special relativity and quantum mechanics arose as physicists realised that the classical physics of mechanics, thermodynamics and electromagnetism were inadequate to account for apparent mysteries in the behaviour of matter and light. Historically, the mysteries were contained in two problems, both under active investigation at the turn of the century.

### 1.5.1 *The Constancy of the Speed of Light*

Despite many attempts to detect an effect,

no variation was discovered in the speed of light. Light emerging from a torch at rest seems to travel forward at the same speed as light from a torch travelling at arbitrarily high speeds. This is very different from the way we perceive the behaviour of velocities in the everyday world. But, of course, we humans never perceive the velocity of light, it is just too fast! This unexpected behaviour is not contrary to common experience, it is beyond it! Explanation for the behaviour forms the starting point for the theory of special relativity, which is the necessary description of anything moving very fast (i.e. nearly all elementary particles); see Chapter 2.

### 1.5.2 *The Interaction of Light with Matter*

All light, for instance sunlight, is a form of heat and so the description of the emission and absorption of radiation by matter was approached as a thermodynamical problem. In 1900 the German physicist Max Planck concluded that the classical thermodynamical theory was inadequate to describe the process correctly. The classical theory seemed to imply that if light of any one colour (any one wavelength) could be emitted from matter in a continuous range of energy down to zero, then the total amount of energy radiated by the matter would be infinite. Much against his inclination, Planck was forced to conclude that light of any given colour cannot be emitted in a continuous band of energy down to zero, but only in multiples of a fundamental quantum of energy, representing the minimum negotiable bundle of energy at any particular wavelength. This is the starting point of quantum mechanics, which is the necessary description of anything very small (i.e. all atoms and elementary particles); see Chapter 3.

As the elementary particles are both fast moving and small, it follows that their description must incorporate the rules of both special relativity and quantum mechanics. The synthesis of the two is known as relativistic quantum theory and this is described briefly in Chapter 4.

## 2

# *Special Relativity*

### 2.1 Introduction

A principle of relativity is simply a statement reconciling the points of view of observers who may be in different physical situations. Classical physics relies on the Galilean principle of relativity, which is perfectly adequate to reconcile the points of view of human observers in everyday situations. But modern physics requires the adoption of Einstein's special theory of relativity, as it is this theory which is known to account for the behaviour of physical laws when very high velocities are involved (typically those at or near the speed of light, denoted by  $c$ ).

It is an astonishing tribute to Einstein's genius that he was able to infer the special theory of relativity in the almost total absence of the experimental evidence which is now commonplace. He was able to construct the theory from the most tenuous scraps of evidence.

To us lesser mortals, it is challenge enough to force ourselves to think in terms of special relativity when envisaging the behaviour of the elementary particles, especially as all our direct experience is of 'normal' Galilean relativity. What follows is of course only a thumbnail sketch of relativity. Many excellent accounts have been written on the subject, not least of which is that written by Einstein himself.

### 2.2 Galilean Relativity

Any theory of 'relativity' is about the relationships between different sets of coordinates against

which physical events can be measured. Coordinates are numbers which specify the position of a point in space (and in time). However, for these numbers to have any meaning, we must also specify the particular coordinate system (or frame of reference) they refer to. For example, we might choose the origin of our coordinates to be the Royal Greenwich Observatory, and choose to specify coordinates in terms of the distance east of the observatory, the distance north and the height. Hence, the choice of a coordinate system involves specifying (1) an origin from which to measure coordinates (e.g. the observatory), and (2) three independent directions (e.g. east, north and up). So, relative to any chosen coordinate system, the position of a point in space is specified in terms of three independent coordinates, which we may write as  $(x, y, z)$ . These three coordinates can be denoted collectively as a vector,  $\mathbf{x} = (x, y, z)$ . A further coordinate,  $t$ , is required to specify time.

Galileo's simple example is still one of the clearest descriptions of what relativity is all about. If a man drops a stone from the mast of a ship, he will see it fall in a straight line and hit the deck below, having experienced a constant acceleration due to the force of gravity. Another man standing on the shore and watching the ship sail past will see the stone trace out a parabolic path, because, at the moment of release, it is already moving with the horizontal velocity of the ship. Both the sailor and the shoreman can write down their views of the stone's motion using

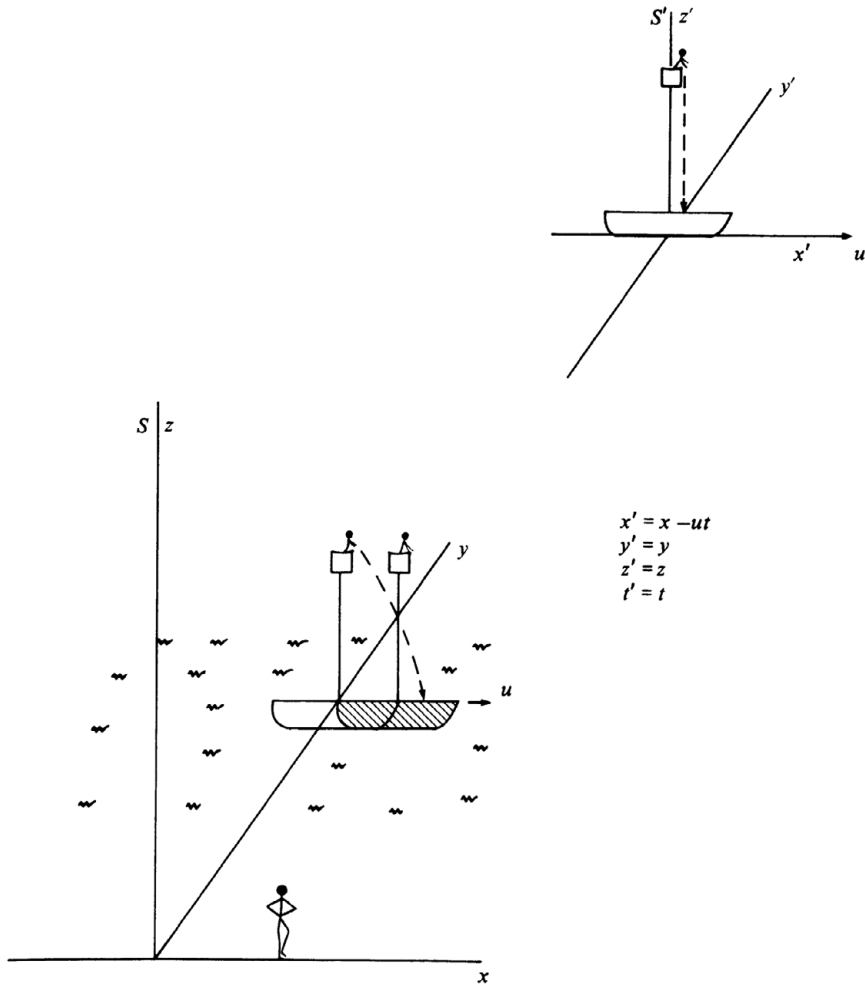


Figure 2.1. The transformations of Galilean relativity.

the mathematical equations for a straight line and a parabola respectively. As both sets of equations are describing the same event (the same force acting on the same stone), they are related by transformations between the two observers. These transformations relate the measurements of position ( $x'$ ), time ( $t'$ ), and velocity ( $v'$ ) in the sailor's coordinate system  $S'$ , with the corresponding measurements ( $x, t, v$ ) made by the shoreman in his coordinate system  $S$ . This situation, assuming that the ship is sailing along the  $x$ -axis with velocity  $u$ , is shown in Figure 2.1.

Important features of the Galilean transformations are that velocity transformations are additive and that time is invariant between the two coordinate

frames. Thus if a sailor throws the stone forward at 10 m per second in a ship travelling forward at 10 m per second, the speed of the stone to a stationary observer on shore will be 20 m per second. And if the sailor on a round trip measures the voyage as one hour long, this will be the same duration as observed by the stationary shoreman.

Lest the reader be surprised by the triviality of such remarks, let him or her be warned that this is not the case in special relativity. At the high velocities, such as are common in the microworld, velocities do not simply add to give the relative velocity, and time is not an invariant quantity. But before we address these sophistications, let us see how the idea came about.

### 2.3 The Origins of Special Relativity

The fact that Galilean transformations allow us to relate observations made in different coordinate frames implies that any one inertial frame (a frame at rest or moving at constant velocity) is as good as another for describing the laws of physics. Nineteenth-century physicists were happy that this should apply to mechanical phenomena, but were less happy to allow the same freedom to apply to electromagnetic phenomena, and especially to the propagation of light.

The manifestation of light as a wave phenomenon (as demonstrated in the diffraction and interference experiments of optics) encouraged physicists to believe in the existence of a medium called the ether through which the waves might propagate (believing that any wave was necessarily due to the perturbation of some medium from its equilibrium state). The existence of such an ether would imply a preferred inertial frame, namely, the one at rest relative to the ether. In all other inertial frames moving with constant velocity relative to the ether, measurement and formulation of physical laws (say the force of gravitation) would mix both the effect under study and the effect of motion relative to the ether (say some sort of viscous drag). The laws of physics would appear different in different inertial frames, due to the different effects of the interaction with the ether. Only the preferred frame would reveal the true nature of the physical law.

The existence of the ether and the law of the addition of velocities suggested that it should be possible to detect some variation of the speed of light as emitted by some terrestrial source. As the Earth travels through space at 30 km per second in an approximately circular orbit, it is bound to have some relative velocity with respect to the ether. Consequently, if this relative velocity is simply added to that of the light emitted from the source (as in the Galilean transformations), then light emitted simultaneously in two perpendicular directions should be travelling at different speeds, corresponding to the two relative velocities of the light with respect to the ether (see Figure 2.2).

In one of the most famous experiments in physics, the American physicists Michelson and Morley set out in 1887 to detect this variation in the velocity of propagation of light. The anticipated variation was well within the sensitivity of their measuring apparatus, but absolutely none was found.

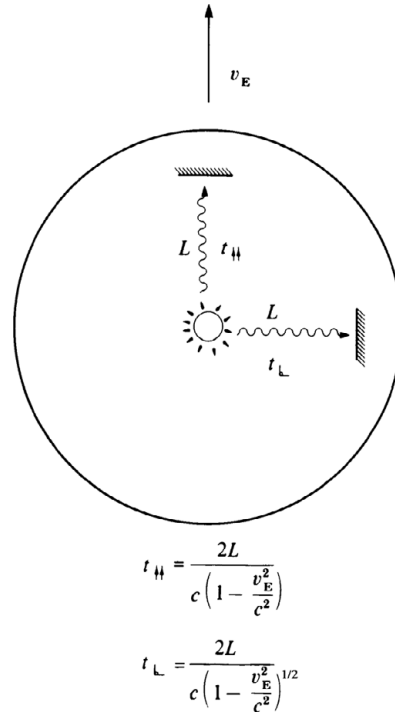


Figure 2.2. Anticipated variation in the propagation of light reflected to and fro along a distance  $L$  due to the Earth's motion through space  $v_E$ .

This experiment provided clear proof that no such ether exists and that the speed of light is a constant regardless of the motion of the source.

### 2.4 The Lorentz–Fitzgerald Contraction

Around the turn of the century, many physicists were attempting to explain the null result of the Michelson and Morley experiment. The Dutch physicist Lorentz and the Irish physicist Fitzgerald realised that it could be explained by assuming that intervals of length and time, when measured in a given frame, appear contracted when compared with the same measurements taken in another frame by a factor dependent on the relative velocity between the two. Their arguments were simply that the anticipated variations in the speed of light were cancelled by compensating changes in the distance and time which the light travelled, thus giving rise to the apparent constancy observed. It is possible to calculate geometrically that an interval of length  $x$  measured in one frame is found to be  $x'$  when measured in a

second frame travelling at velocity  $v$  relative to the first where:

$$x = \frac{x'}{\left(1 - \frac{v^2}{c^2}\right)^{1/2}}. \quad (2.1)$$

Here,  $c$  is the speed of light, which is approximately equal to  $2.998 \times 10^8$  metres per second. And, similarly, the intervals of time observed in the two frames are related by:

$$t = \frac{t'}{\left(1 - \frac{v^2}{c^2}\right)^{1/2}}. \quad (2.2)$$

These empirical relationships, proposed on an ad hoc basis by Lorentz and Fitzgerald, suggest that because the ‘common-sense’ Galilean law of velocity addition fails at speeds at or near that of light, our common-sense perceptions of the behaviour of space and time must also fail in that regime. It was Einstein who, quite independently, raised these conclusions and relationships to the status of a theory.

## 2.5 The Special Theory of Relativity

The special theory of relativity is founded on Einstein’s perception of two fundamental physical truths which he put forward as the basis of his theory:

- (1) All inertial frames (i.e. those moving at a constant velocity relative to one another) are equivalent for the observation and formulation of physical laws.
- (2) The speed of light in a vacuum is constant.

The first of these is simply the extension of the ideas of Galilean relativity to include the propagation of light, and the denial of the existence of the speculated ether. With our privileged hindsight, the amazing fact of history must be that the nineteenth-century physicist preferred to cling to the idea of relativity for mechanical phenomena while rejecting it in favour of the concept of a preferred frame (the ether) for the propagation of light. Einstein’s contribution here was to extend the idea of relativity to include electromagnetic phenomena, given that all attempts to detect the ether had failed.

The second principle is the statement of the far-from-obvious physical reality that the speed of light is truly independent of the motion of the source and

so is totally alien to our everyday conceptions. Einstein’s achievement here was to embrace this apparently ludicrous result with no qualms. Thus the theory of relativity, which has had such a revolutionary effect on modern thought is, in fact, based on the most conservative assumptions compatible with experimental results.

Given the equivalence of all inertial frames for the formulation of physical laws and this bewildering constancy of the speed of light in all frames, it is understandable intuitively that measurements of space and time must vary between frames to maintain this absolute value for the speed of light. The relationships between measurements of space, time and velocity in different frames are related by mathematical transformations, just as were measurements in Galilean relativity, but the transformations of special relativity also contain the Lorentz–Fitzgerald contraction factors to account for the constancy of the speed of light (see Figure 2.3).

The first feature of the transformations to note is that when the relative velocity between frames is small compared with that of light (i.e. all velocities commonly experienced by humans), then  $v/c \approx 0$ , and the transformations reduce to the common-sense relations of Galilean relativity.

The unfamiliar effects of special relativity contained in the transformations can be illustrated by a futuristic example of Galileo’s mariner: an astronaut in a starship travelling close to the speed of light  $c$ .

Because of the transformations, velocities no longer simply add. If, say, the astronaut fires photon torpedoes forward at speed  $1c$  from the starship, which itself may be travelling at  $0.95c$ , the total velocity of the photon torpedoes as observed by a stationary planetary observer is not the sum,  $1.95c$ , but is still  $c$ , the constant speed of light. Also, time is dilated. So a voyage which to the stationary observer is measured as a given length of time will appear less to the kinetic astronaut.

Another intriguing feature of the transformations is that continued combinations of arbitrary velocities less than  $c$  can never be made to exceed  $c$ . Thus the transformations imply that continued attempts to add to a particle’s velocity (by successive accelerations) can never break the light barrier. Indeed, the transformations themselves do not make sense for velocities greater than  $c$ , as when  $v > c$  the equations become imaginary, indicating a departure from the



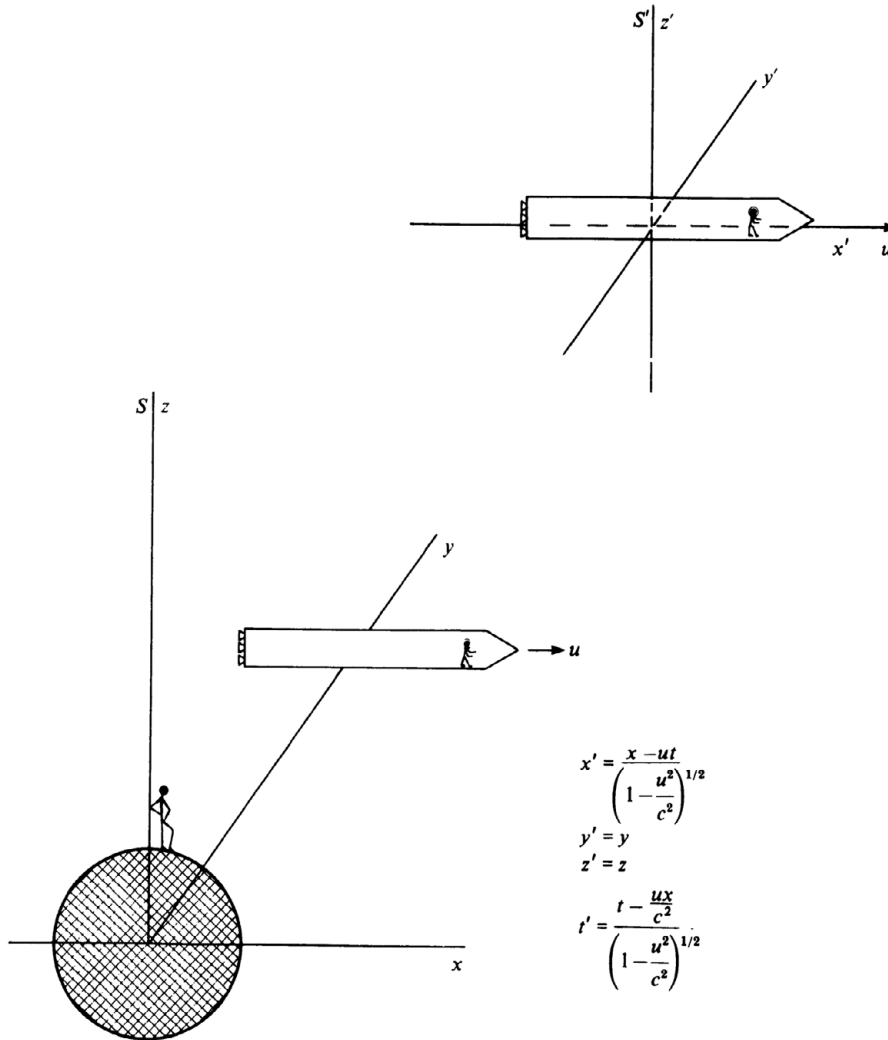


Figure 2.3. The Lorentz transformations of special relativity.

physical world. Special relativity therefore implies the existence of an ultimate limiting velocity beyond which nothing can be accelerated.

### 2.6 Mass Momentum and Energy

If the transformation laws of special relativity show diminishing returns on any attempts to accelerate a particle (by application of some force), it is reasonable to expect some compensating factor to bring returns in some other way, and so maintain energy conservation. This compensating factor is the famous increase in the mass of a particle as it is accelerated to speeds approaching  $c$ .

By requiring the laws of conservation of mass and conservation of momentum to be invariant under the Lorentz transformations, it is possible to derive the relationship between the mass of a body  $m$  and its speed  $v$ ,

$$m = \frac{m_0}{\left(1 - \frac{v^2}{c^2}\right)^{1/2}}, \tag{2.3}$$

where  $m_0$  is the mass of the body in a frame in which it is at rest. Multiplying the equation by  $c^2$  and expanding the bracket we obtain:

$$mc^2 = m_0c^2 + \frac{m_0v^2}{2} + \dots \tag{2.4}$$

We can identify the second term on the right-hand side of the equation as the classical kinetic energy of the particle. The subsequent terms are the relativistic corrections to the energy while the first is describing a quantity of energy arising only from the mass itself.

This is the origin of the mass–energy equivalence of special relativity expressed in the most famous formula of all time:

$$E = mc^2. \quad (2.5)$$

From this formula several others follow immediately. One can be obtained by substituting an expression for the momentum ( $\mathbf{p}$ ) into the expansion for  $m$  in the above:

$$\mathbf{p} = m\mathbf{v},$$

so

$$E^2 = m_0^2 c^4 + p^2 c^2. \quad (2.6)$$

For a particle with no rest mass, such as the photon, this gives:

$$\frac{E}{p} = c. \quad (2.7)$$

## 2.7 The Physical Effects of Special Relativity

The effects which we have just introduced are all wholly unfamiliar to human experience and this is perhaps one reason why, even today, the reality of special relativity is repeatedly challenged by sceptical disbelievers (see Figure 2.4). But all the effects are real and they can all be measured.

A roll call of the effects of special relativity provides a useful checklist which we should remember when envisaging the behaviour of elementary particles.

### 2.7.1 The Ultimate Speed $c$

It is possible to measure directly the velocity of electrons travelling between two electrodes by measuring the time of flight taken. It is observed that the speed does not increase with the energy which the electrons have been given as it would under classical Newtonian theory, but instead tends to a constant value given by  $c$ .

### 2.7.2 Addition of Velocities

Under special relativity, only when individual velocities are much smaller than  $c$  can they be simply added to give the relative velocities. At speeds

## DELTA PUBLICATIONS

7305, Aram Nagar, New Delhi-110055, INDIA

# A BIG HOWLER

EINSTEIN'S THEORY OF  
SPECIAL RELATIVITY

Dr S. P. Gulati & Dr (Mrs) S. Gulati, Associate Professors, Cuttington University College, Liberia; January 1982; 106 pages; Price US\$12.50, STG.6.25 (Air Parcel Postage free).

This book is an open challenging invitation to 'Einsteinians'—particularly so to persons like Professor A. I. Miller of USA who in his recent book 'Albert Einstein's Theory of Special Relativity' (Addison-Wesley, 1981) has undertaken to apotheosize Einstein whose work if not an act of straight plagiarism is definitely 'A BIG HOWLER': infested with infidelities. The 'Transformation Maze' is another interesting feature of the book. Besides, it also contains outlines of the authors' 'SIMILARITY THEORY', perhaps, the only valid alternative.

The book is obtainable either directly from the publisher or through your book-seller or the authors.

Figure 2.4. Special relativity in trouble? An advertisement from *New Scientist* magazine, 27 May 1982.

approaching  $c$ , velocities do not add, but combine in a more complicated way so that the total of any combination is always less than  $c$ . This can be tested directly by an elementary particle reaction. One kind of elementary particle we shall encounter is the neutral pion  $\pi^0$  which often decays into a pair of photons. If the pion is travelling say at  $0.99c$  when it emits a photon, we would expect the photon to have a total velocity of  $1.99c$  under the laws of Galilean relativity. This is not observed. The photon velocity is measured to be  $c$ , showing that very high velocities do not add, but combine according to the formula:

$$v_{\text{total}} = \frac{v_1 + v_2}{1 + v_1 v_2 / c^2}.$$

### 2.7.3 Time Dilation

This is the effect which causes moving clocks to run slowly and it has been measured directly in an experiment involving another type of elementary particle. The experiment looks at a species

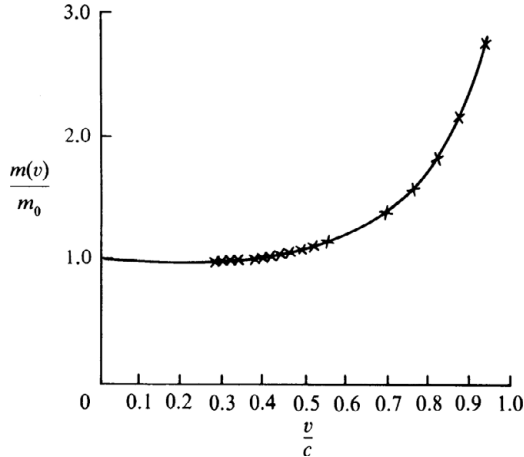


Figure 2.5. Relativistic mass increase as a function of velocity.

of elementary particle called the muon, which is produced in the upper atmosphere by the interactions of cosmic rays from outer space. The muon decays into other particles with a distribution of lifetimes around the mean value of  $2.2 \times 10^{-6}$  s when measured at rest in the laboratory. By measuring the number of muons incident on a mountain top, it is possible to predict the number which should penetrate to sea level before decaying. In fact, many times the naive prediction are found at sea level, indicating that the moving particles have experienced less time than if they were stationary. Muons moving at, say,  $0.99c$  keep time at only one-seventh the rate when stationary with respect to us.

#### 2.7.4 Relativistic Mass Increase

The last effect we shall illustrate is the well-known increase in the apparent mass of a body as its velocity increases. This has been measured directly by observing the electric and magnetic deflections of electrons of varying energies (see Figure 2.5).

### 2.8 Using Relativity

As we have seen, relativity tells us how to relate the formulations of physical laws in different frames of reference, but it does not tell us how to formulate them in the first place. This is the rest of physics! In this pursuit, special relativity is introduced by adopting kinematical prescriptions which the dynamical variables must obey.

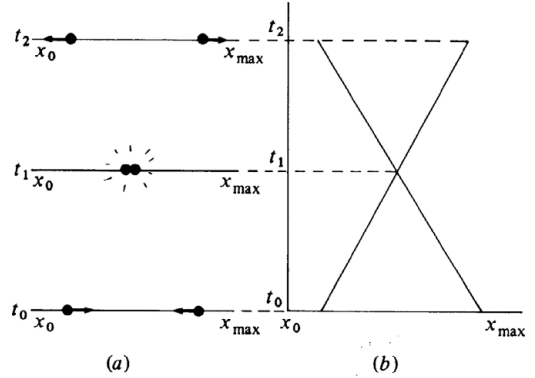


Figure 2.6. A space-time diagram particle collision (b) shown sequentially in (a).

#### 2.8.1 Space-Time Diagrams

In classical relativity, space and time are entirely separate, but in special relativity they are mixed together by the Lorentz transformations. Thus it makes little sense to visualise events as occurring only in space. A better context in which to visualise them is space-time. Space-time diagrams can be used to display events at the expense, for the purposes of visualisation, of making do with only one, or possibly two, spatial dimensions (see Figure 2.6). A point in space-time is frequently called an *event*.

#### 2.8.2 Four-vectors

Just as ordinary vectors  $\mathbf{x} = (x, y, z)$  (or three-vectors) define the components of a position or velocity in ordinary space, we can define four-vectors  $(\mathbf{x}, ct) = (x, y, z, ct)$  to define an event in space-time. The fourth coordinate is the time coordinate multiplied by  $c$  to give an equivalent distance, so matching the other three distance components.

The benefit of writing equations in three-vector form is to ensure their covariance under spatial rotations. (Covariance is not quite the same as invariance, which means that absolutely nothing changes. Covariance means that both sides of an equation change in the same way, preserving the validity of the equation.) This permits freedom in the orientation of the coordinate system employed and also ensures that conservation of angular momentum is manifest (see Chapter 6). If we can write the laws of physics in four-vector form, then the benefit is that the laws will be covariant under ‘rotations’ in space-time (which are equivalent to the Lorentz transformations of special relativity).

In addition to the position three-vector  $\mathbf{x}$ , the momentum of a particle is also a vector quantity ( $\mathbf{p}$ ). By examining the effects of the Lorentz transformations on the momentum and the energy of a particle, it is possible to form a four-vector from these quantities, namely  $(\mathbf{p}, E/c)$ . This four-vector is used to specify the dynamic state of a particle. It does not specify an event in space–time.

### 2.8.3 Relativistic Invariants

Although special relativity illustrates how perceptions of space and time may vary according to the observer’s frame, it also accommodates absolutely invariant quantities, which we might expect to vary under Galilean relativity. The speed of light in a vacuum is the obvious invariant upon which the theory is founded. Another quantity is the square of the space–time interval between an event and the origin of the coordinate system,

$$s^2 = x^2 - (ct)^2 = x'^2 - (ct')^2. \quad (2.8)$$

This is just a special case of the square of the space–time interval between two events, which is the difference between their four-vectors,

$$\Delta s^2 = (\Delta x)^2 - (c\Delta t)^2.$$

Another invariant quantity is the rest mass of a given material particle. All observers will agree on the mass of the same particle at rest in their respective frames:

$$m_0^2 = \frac{E^2}{c^4} - \frac{p^2}{c^2}. \quad (2.9)$$

Relativistic invariants are useful in high-energy physics because, once measured, their value will be known in all other frames of reference. Here it is worth appreciating that high-energy experiments regularly exercise the idea of Lorentz transformations. One experiment may arrange for two protons travelling with equal energies in opposite directions to collide head-on, while another experiment may collide moving protons and a stationary target. The centres of mass of the two experiments will be moving relative to one another with some velocity which is likely to be an appreciable fraction of the speed of light. This will require the Lorentz transformations to relate measurements in the two experiments.

This concludes our brief sketch of special relativity and now we pass on to the second of the two great pillars of twentieth-century physics: quantum mechanics.

# 3

## *Quantum Mechanics*

### 3.1 Introduction

It is fascinating to reflect on the fact that both quantum mechanics and special relativity were conjured into being in the first five years of the previous century, and interesting to compare the development of the two. Whereas special relativity sprang as a complete theory (1905) from Einstein's genius, quantum mechanics emerged in a series of steps over a quarter of a century (1900–25). One explanation of this is that, whereas in special relativity the behaviour of space and time follows uniquely from the two principles, in quantum theory there were no such simple principles which, known at the beginning, allowed the derivation of all quantum phenomena. Rather, each of the steps was a fresh hypothesis based on, or predicting, some new experimental facts and these do not necessarily follow logically one from another, still less from just one or two fundamental principles. So quantum mechanics emerged, hypothesis hand-in-hand with experiment, over 25 years or so. As indicated by the subheadings of this chapter, most of the steps in the progression can be associated closely with just one man, and we will use the examination of each of these in turn as our introduction to quantum mechanics.

### 3.2 Planck's Hypothesis

As mentioned in Section 1.5.2, quantum theory came into being when Max Planck attempted to explain the interaction of light with matter. That is, for

instance, how hot metal emits light and how light is absorbed by matter.

Using the well-known and highly trusted classical theories of thermodynamics and electromagnetism, Planck derived a formula describing the power emitted by a body, in the form of radiation, when the body is heated. To find the total power radiated, it is necessary to integrate over all the possible frequencies of the emitted radiation. But when Planck tried to do this using his classical formula, he found that the total radiated power was predicted to be infinite – an obviously nonsensical prediction!

Planck was able to avoid this conclusion only by introducing the concept of a minimum amount of energy which can exist for any one frequency of the radiation – a *quantum*. By assuming that light can be emitted or absorbed by matter only in multiples of the quantum, Planck derived a formula which gives the correct prediction for the total amount of power emitted by a hot body. A convenient analogy here may be the economic wealth of an individual, which is normally thought of as a continuously variable quantity. Yet when the individual is in economic interaction (i.e. goes shopping), his or her wealth is quantised in multiples of the smallest denomination coin available. The minimum quantum of energy  $E$ , allowed at a given frequency  $\nu$ , is given by Planck's formula

$$E = h\nu, \tag{3.1}$$

where  $h$  is Planck's quantum constant with dimensions of energy per frequency and the minute value of  $6.625 \times 10^{-34}$  joule seconds. The appearance of Planck's constant in the equations of physics is a valuable diagnostic device. When we set  $h = 0$ , then we are ignoring the existence of the quantum and so should recover the results of classical physics. However, when we examine formulae (or parts of formulae) which are proportional to  $h$ , then we are looking at wholly quantum effects which would not be predicted by classical physics.

### 3.3 Einstein's Explanation of the Photoelectric Effect

The next major step in quantum theory was taken by Einstein in the same year as his formulation of special relativity. This was his explanation of the photoelectric effect, or how metal can be made to emit electrons by shining a light on it. Planck had suggested that only light in interaction with matter would reveal its quantum behaviour at low energies. Again it was left to Einstein to generalise the idea (as he had generalised the idea of relativity to include electromagnetism). He proposed that all light exists in quanta and set out to show how this might explain the photoelectric effect.

He assumed that the electrons need a definite amount of energy to escape from the metal. If the light of a given colour which is shone on the metal consists of a large number of quanta, each of energy  $h\nu$ , then quanta which collide with the electrons provide them with the energy they need to escape. The electrons will pop out of the metal with an energy which is the difference between that of the quanta and that needed to escape the surface. If the light is below a certain frequency, then no matter how much of it is used, no single quantum will be able to give an electron enough energy to escape. Ignoring multiple quanta-electron collisions, no electrons will emerge. But if the frequency of the light is increased, scanning up the spectrum from red to blue, the electrons will suddenly appear when the quanta have just enough energy to liberate them. As the frequency is increased further still, the electrons will be ejected with higher and higher energies.

This picture exactly fits the experimental facts of the photoelectric effect discovered in 1902 by Lenard. These are that the energy of the electrons emitted depends only on the frequency of the light and not

on the intensity (the number of quanta), and that the number of electrons emitted depends only on the intensity but not the frequency.

Einstein's explanation of the photoelectric effect confirmed the quantum theory of light (and won him the Nobel Prize). This resurrection of a corpuscular theory of light causes immediate conceptual problems because light is quite demonstrably also a continuous wave phenomenon (as demonstrated by diffraction and other interference experiments). It appears to be both a discrete particle (a *photon*) and an extended wave! How can this be?

Resolution of this apparent paradox requires the introduction of a new entity which reduces to both particle and wave in different circumstances. This entity turns out to be a *field*, which we shall discuss further in Chapter 4. But before going on to this we will come to appreciate that not only light is subject to such schizophrenic behaviour.

### 3.4 Bohr's Atom

We saw in Chapter 1 how Rutherford's scattering experiments led to a picture of the atom in which the light, negatively charged electrons orbit the small, massive, positively charged nucleus located in the centre, the vast majority of the volume of the atom being empty space. This appealing picture has fundamental difficulties. Firstly, in the classical theory of electrodynamics, all electric charges which experience an acceleration should emit electromagnetic radiation. Any body constrained to an orbit is subject to an acceleration by the force which gives rise to the orbit in the first place. Thus the electrons in Rutherford's atom should be emitting radiation constantly. This represents a loss of energy from the electrons which, as a result, should spiral down into lower orbits and eventually into the nucleus itself. This 'radiation collapse' of atoms is an inescapable consequence of classical physics and represents the failure of the theory in the atomic domain. Another problem of the Rutherford atom is to explain why all the atoms of any one element are identical. In classical physics, no particular configuration of electronic orbits is predicted other than on the grounds of minimising the total energy of the system. This does not explain the identity of the atoms of any element. A fundamentally new approach is needed to describe the Rutherford atom.

It was the Danish physicist Niels Bohr who in 1913 suggested a new quantum theory of the atom,

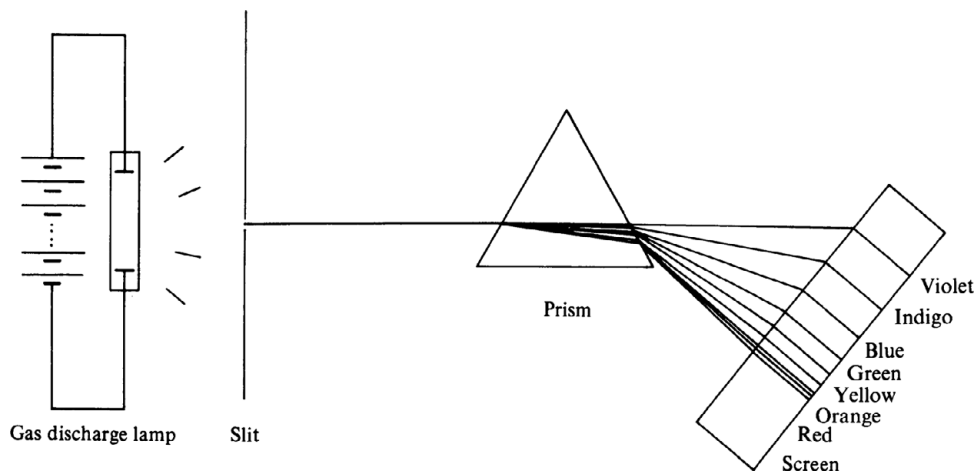


Figure 3.1. The characteristic spectrum of a gas discharge lamp.

which, at a stroke, dismissed the problem of the radiation collapse of atoms, explained the way in which light is emitted from atoms and incorporated the new quantum ideas of Planck and Einstein.

Bohr's basic hypothesis was the simplest possible application of the quantum idea to the atom. Just as Planck had hypothesised that light exists only in discrete quanta, so Bohr proposed that atoms can exist only in discrete quantum states, separated from each other by finite energy differences, and that *when in these quantum states the atoms do not radiate*. A simple way to think of these quantum states is as a set of allowed orbits for the electron around the nucleus, the space between the orbits being forbidden to the electrons.

The allowed orbits are specified as those in which the orbital angular momentum of the electron is quantised in integral units of Planck's quantum constant divided by  $2\pi$  and denoted  $\hbar$ . It may seem odd that angular momentum should be one of the few quantities to be quantised (like energy and electric charge but unlike mass, linear momentum and time). But we may have suspected as much on first meeting Planck's constant. Its rather unusual units of energy per frequency are in fact identical to the dimensions of angular momentum.

Although the atom is assumed not to radiate light when all its electrons are safely tucked into their quantum orbits, it will do so when an electron makes a transition from one of the allowed orbits to another. This process of emission should explain

the behaviour of light observed in the real world. The light from a gas discharge lamp, say a neon or mercury vapour tube, has a distinctive appearance. The atoms in a gas or vapour are widely separated and interact with each other relatively seldom. This means that the light they emit will be characteristic of the particular atoms involved. It is a mixture of just a few separate frequencies which can be split up by a prism. The resulting spectrum of frequency lines is a unique property of the element which is emitting the light (see Figure 3.1). Late in the nineteenth century, researchers such as Balmer, Lyman and Paschen looked at the spectra of many different elements and noted that they all fall into simple mathematical patterns – with several discrete patterns per element. These patterns had long defied explanation, essentially because they defy the smooth way in which quantities vary in classical physics. But with the quantum theory, Bohr was able to put forward a convincing explanation of the origin of these lines. Each pattern of frequency lines represents the energy difference between a particular quantum state and all the others in the atom from which the electron can reach that state by emitting light (see Figure 3.2).

With Bohr's model of the atom, physicists were able to calculate, in great detail, many of the spectroscopic results obtained by the experimenters of previous decades. On the basis of this understanding of atoms, Bohr himself was able to propose a tentative explanation of Mendeleev's Periodic Table of elements. The periodic table, which classifies

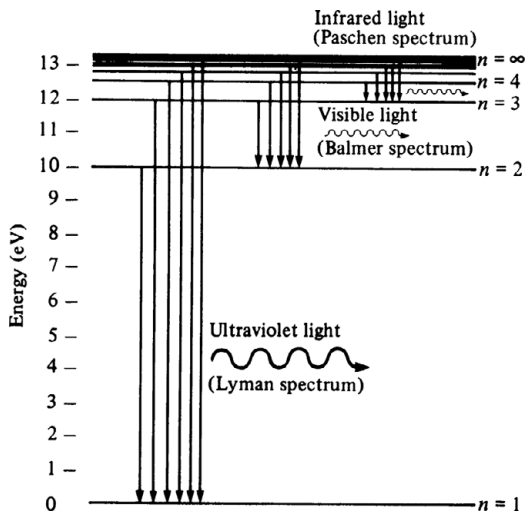


Figure 3.2. The discrete patterns of frequency lines in a given element (such as hydrogen) arise from transitions into each available state from all the others. Each state is labelled by its Bohr orbital quantum number  $n$ .

the elements into groups reflecting their chemical behaviour, is explained by the way the electronic orbits are filled in the different elements. The chemical properties of an element are determined predominantly by the number of electrons in its outermost orbit, and so by proposing the electronic orbital structures of the elements it is possible to reproduce the pattern of the table (see Figure 3.3).

While the Bohr atom was an enormous step forward, and the concept of electronic orbits is a mental crutch for our imaginations operating so far beyond their normal domain, it is important to realise that it is only the simplest quantum model of the atom and that more-sophisticated portrayals of electronic behaviour are necessary, as we are about to see.

### 3.5 De Broglie's Electron Waves

The next major conceptual advance in quantum theory came much later, in 1924. The young French physicist Louis de Broglie suggested in his doctoral thesis that just as light waves could act like particles in certain circumstances, so too could particles manifest a wavelike behaviour. In particular, he proposed that the electrons, which had previously been regarded as hard, impenetrable, charged spheres could in fact behave like extended waves undergoing diffraction and interference phenomena just like light or water waves.

According to de Broglie, the wavelength of a particle wave is inversely proportional to its momentum, the constant of proportionality being Planck's quantum constant:

$$\lambda = \frac{h}{p}. \quad (3.2)$$

So the higher the momentum of a particle, the smaller its wavelength. It is worth appreciating that de Broglie's hypothesis applies to all particles, not just to electrons and the other elementary particles. For instance, a billiard ball rolling across the table top will have a wavelength, but because Planck's constant is so minute and the ball's momentum is so comparatively large, the billiard ball's wavelength is about  $10^{-34}$  m. This, of course, is many orders of magnitude different from the typical dimensions of billiards, and so the wave character of the ball never reveals itself. But for electrons, their typical momenta can give rise to wavelengths of  $10^{-10}$  m, which are typical of atomic distance scales. So electrons may be expected to exhibit a wavelike character during interaction with atomic structures.

This wavelike character was observed in 1927 by the US physicists Clinton Davisson and Lester Germer, and independently by G. P. Thomson (J. J.'s son) who was at the time Professor of Natural Philosophy at the University of Aberdeen in Scotland. They demonstrated that electrons undergo diffraction through the lattice structure of a crystal in a fashion similar to the diffraction of light through a grating. Davisson and Thomson were jointly awarded the 1937 Nobel Prize for Physics.

De Broglie's hypothesis also provided the first rationale for Bohr's model of the atom. The existence of only certain specific electronic orbits can be explained by allowing only those orbits which contain an integral number of de Broglie wavelengths. This reflects the momentum of the electron involved (and so the energy of the orbit); see Figure 3.4.

Adoption of de Broglie's idea requires the comprehensive assimilation of particle-wave duality. For any entity in the microworld, there will be situations in which it is best thought of as a wave and situations in which it is best thought of as a particle. Neither is a truer representation of reality than the other, as both are the coarse product of our human macroscopic imaginings.



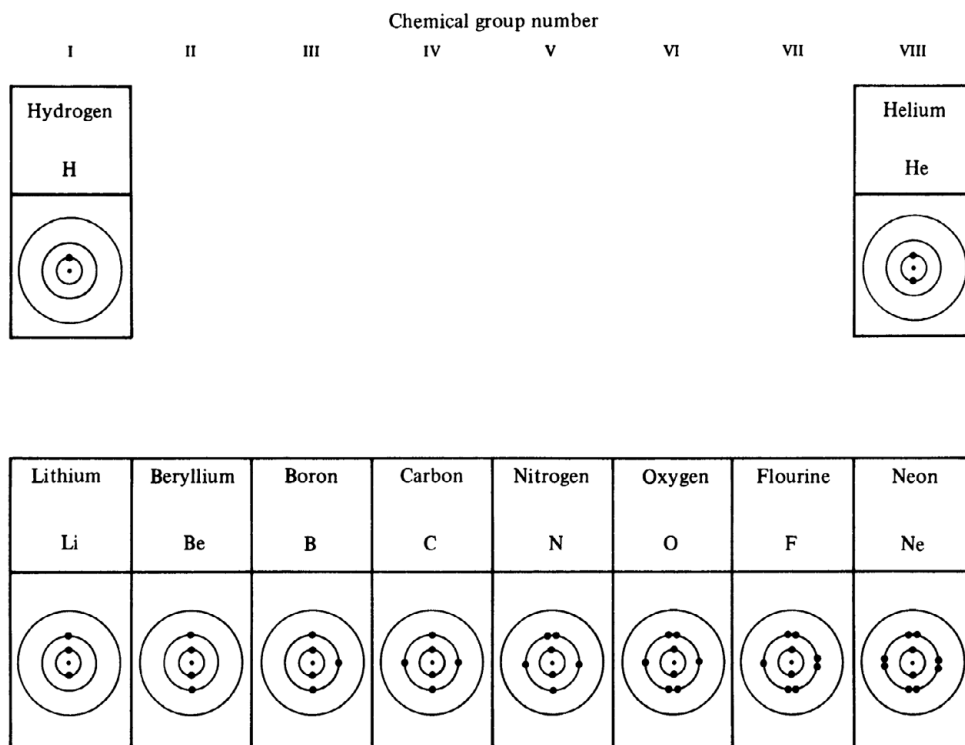


Figure 3.3. A fragment of the periodic table and the associated electronic orbital structure.

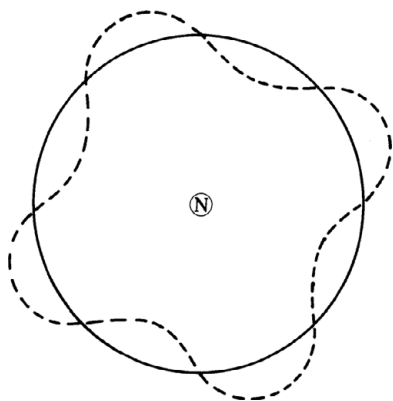


Figure 3.4. Allowed orbits are explained as containing an integral number of de Broglie wavelengths.

The advent of de Broglie's ideas marks the spark which started the intellectual bush fire of quantum theory proper. Up to the early 1920s, quantum theory was a series of prescriptions (albeit revolutionary ones) but not a dynamical theory of mechanics to

transcend that of Newton. The second wave of the quantum revolution (1924–27) was to provide just such a theory.

### 3.6 Schrödinger's Wavefunction

Following on directly from de Broglie's ideas, the Austrian physicist Erwin Schrödinger developed the idea of particle waves into a wave mechanics proper. Schrödinger's starting point was essentially the wave equation describing the behaviour of light waves in space and time. Just as this is the accurate representation of optical phenomena (which can be described approximately by the light *rays* of geometrical optics), Schrödinger formulated a matter wave equation which he put forward as the accurate representation of the behaviour of matter (which is described approximately by the particle dynamics). Schrödinger's equation (Figure 3.5) describes a particle by its wavefunction, denoted  $\psi$ , and goes on to show how the particle wavefunction evolves in space and time under a specific set of circumstances.

$$\boxed{-\frac{\hbar^2}{2m} \frac{\partial^2 \psi}{\partial x^2} + V\psi = i\hbar \frac{\partial \psi}{\partial t}}$$

Figure 3.5. Schrödinger's wave equation.

One such circumstance of very great interest is that of a single electron moving in the electric field of a proton. Using his wave equation, Schrödinger was able to show that the electron wavefunction can assume only certain discrete energy levels, and that those energy levels are precisely the same as the energies of the electronic orbits of the hydrogen atom, postulated earlier by Bohr.

The particle wavefunction is an extremely significant concept which we shall use frequently in the coming chapters. It is a mathematical expression describing all the observable features of a particle. Collisions between particles are no longer necessarily viewed as some variant of billiard-ball behaviour but, instead, as the interference of wavefunctions giving rise to effects rather like interference phenomena in optics.

But now that we have introduced the particle wavefunction, and claimed that an equation governing it can predict the behaviour of particles, what exactly is its significance? Should we think of an electron as a localised ball of stuff, or as some extended wave? And if a wave, what is doing the waving? After all, there is no such thing as a light wave; it is a handy paraphrase for time- and space-varying electric and magnetic fields. What, then, is a matter wave?

Before we answer these intriguing questions, we need one more principle of quantum theory. This is the 'uncertainty principle' which the German physicist Werner Heisenberg derived from his alternative formulation of a quantum mechanics, developed simultaneously with Schrödinger's wave formulation, but from a very different starting point.

### 3.7 Heisenberg's Mechanics and the Uncertainty Principle

Heisenberg took as his starting point the quantum state of the system under consideration (e.g. a single electron, an atom, a molecule, etc.), and argued that the only sensible way to formulate a mechanics of the system was by modelling the act of observation on it. Here, by the word 'observation' we mean any

interaction experienced by the system, such as the scattering off it of light or of an electron. In the absence of any interaction, the system would be totally isolated from the outside world and so totally irrelevant. Only by some form of interaction or observation does the system exist in a definite state.

Heisenberg's approach is the literal manifestation of Wittgenstein's parting philosophical rejoinder, 'concerning that of which we cannot speak, we must pass over in silence'. We can speak (or write equations) only of what we observe, and so observation is to have pride of place in quantum theory.

Heisenberg represented observations on a system as mathematical operations on its quantum state. This allowed him to write equations governing the behaviour of a quantum system and so led to results which were identical to the somewhat more accessible wave mechanics of Schrödinger (say in predicting the energy levels of the hydrogen atom). The equivalence of the two approaches can be appreciated by realising that the expressions Heisenberg used to represent the observations are differential operators and that they act on the quantum state, which is represented by the wavefunction of the system. So this approach will result in a differential equation in the wavefunction  $\psi$ , identical to the wave equation which Schrödinger obtained by analogy with the wave equation for light.

Heisenberg's uncertainty principle results from the realisation that any act of observation on the quantum system will disturb it, thus denying perfect knowledge of the system to the observer. This is best illustrated by analysis of what would happen if we were to attempt to observe the position of an electron in an atomic orbit by scattering a photon off it (Figure 3.6). The photon's wavelength is related to its momentum by the same equation as for any other particle:

$$\lambda = \frac{h}{p}$$

So the greater the photon's momentum, the shorter its wavelength and vice versa. If then we wish to determine the position of the electron as accurately as possible, we should use the photon with the highest possible momentum (shortest wavelength), as it is not possible to resolve distances shorter than the wavelength of the light used. However, by using a high-momentum photon, although we will gain a good estimate of the electron's position at the instant of measurement, the electron will have been violently disturbed by the high

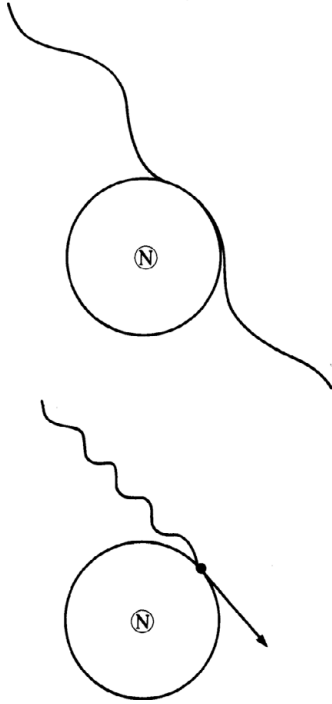


Figure 3.6. A long-wavelength (low-momentum) photon can give only a rough estimate of the position of the electron, but does not disturb the atom very much. A short-wavelength (high-momentum) photon localises the electron more accurately, but causes great disturbance.

momentum of the photon and so its momentum will be very uncertain. This is the essence of Heisenberg's uncertainty principle. Knowledge of any one parameter implies uncertainty of some other so-called 'conjugate' parameter. This is expressed mathematically by requiring that the product of the uncertainties in the two conjugate parameters must always be greater than or equal to some small measure of the effect of measurement. Not surprisingly, this measure turns out to be none other than Planck's ubiquitous constant:

$$\Delta p \Delta x \geq \frac{\hbar}{2} \quad \text{with} \quad \hbar = \frac{h}{2\pi}.$$

A similar trade-off occurs when attempting to measure the energy of a quantum system at a given time. An instantaneous measurement implies a high-frequency probe (one wavelength over in a short time), but this means a high-energy probe which will mask the energy of the quantum state itself. Conversely, a very low-energy probe, which will not unduly affect the

energy of the quantum state, implies a low-frequency probe, which means the time to be associated with the measurement is uncertain, thus:

$$\Delta E \Delta t \geq \hbar.$$

Heisenberg's uncertainty principle is an enormously powerful result when we realise that the uncertainty in a quantity provides a good guide to its minimum value. For instance, if we know that the uncertainty in a particle's lifetime is 1 s, then the lifetime is unlikely to be less than  $\frac{1}{2}$  s as the uncertainty could not otherwise be accommodated. Similarly, if we know that a particle is confined to a small volume (say the nucleus  $\Delta x \approx 10^{-15}$  m), then we can conclude that the momentum of the particle must be greater than

$$p_{\min} \approx \frac{\Delta p}{2} \approx \frac{\hbar}{2\Delta x} \approx 100 \text{ MeV}/c.$$

If the particle is confined to the nucleus, then this is a reasonable guide to the strength (energy) of the force which is keeping it there.

Armed with these ideas, we can turn to the thorny problem of just what a matter wave is.

### 3.8 The Interpretation of the Wavefunction $\psi$

Firstly, let us address the question of whether an electron is to be regarded as a localised ball or an extended wave. Which of these two descriptions applies is very much a matter of the circumstances the electron finds itself in (see Figure 3.7).

For an electron which is travelling through space with a definite momentum ( $\Delta p = 0$ ) and so is isolated from all interactions, the uncertainty in its position is infinite. Thus its wavefunction is a sine wave of definite wavelength extended throughout space. The electron is in no sense a localised particle. If an electron is vaguely localised, say we know it has disturbed an atom, then with  $\Delta x$  as the dimension of the atom, we know that there will be an uncertainty  $\Delta p$  in the electron momentum (due to its interaction with the atom) and so a spread in the wavelength of the wavefunction,  $\Delta \lambda = h/(\Delta p)$ . This spread in wavelengths (frequencies) causes the formation of a localised wavepacket in the wavefunction reflecting the rough localisation of the electron.

When the electron is very specifically localised, say in a quasi-point-like, high-energy collision with another particle, then the uncertainty in its momentum (and so the spread in the wavelength components

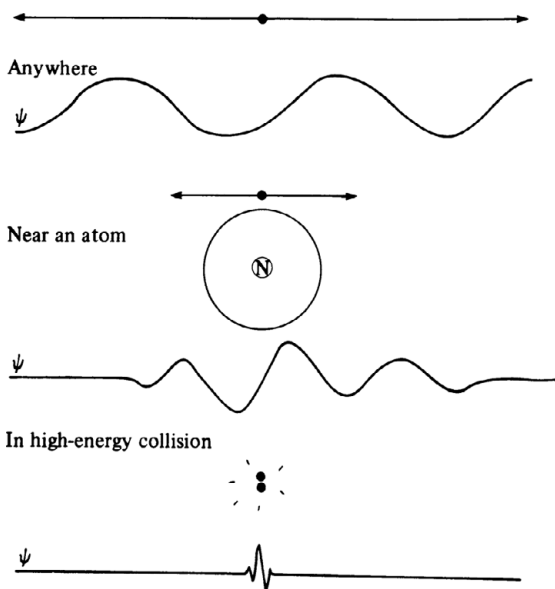


Figure 3.7. A particle's wavefunction reflects its localisation (see text).

of the wavefunction) is large, and the wavepacket becomes very localised, in which case it is sensible to regard the electron as a particle.

This picture of the electron wave makes rather a nonsense of the simple Bohr picture of the orbiting electrons. The dimensions of the electronic wavefunction are comparable to that of the atom itself. Until some act of measurement localises the electron more closely, there is no meaning to ascribing any more detailed a position for the electron. However, this explanation is not altogether satisfactory as it stands, as we have left the electron with a rather poorly defined role in the atom. Progress in understanding this aspect is related to our other outstanding question about the wavefunction: what is it?

In 1926 the German physicist Max Born ventured the suggestion that the square of the amplitude of the wavefunction at any point is related to the probability of finding the particle at that point. The wavefunction itself is proposed to have no direct physical interpretation other than that of a 'probability wave'. When squared, it gives the chance of finding the particle at a particular point on the act of measurement. Hence, the probability density for finding the particle at the position  $\mathbf{x}$  at time  $t$  is

$$\text{probability density} = |\psi(\mathbf{x}, t)|^2.$$

So the location of the electron in the atom is not wholly indeterminate. The solution to Schrödinger's equation for an electron in the electrical field of the proton will give an amplitude for the wavefunction as a function of distance from the proton (as well as the energy levels mentioned earlier). When squared, the amplitude gives the probability of finding the electron at any particular point. Thus we can give only a probability for finding an electron in its Bohr orbit, a probability for determining its position within the orbit and probabilities for finding it in the space between orbits. There is even a small probability of this so-called orbital electron existing actually inside the nucleus!

Schrödinger's wavefunction associates with every point in space (and time) two real numbers: the amplitude (or size) of the wavefunction, and its *phase*. In general, the phase of a wave corresponds to the position in its cycle, with respect to an arbitrary reference point. In other words, it is a measure of how far away one is from a wave crest or trough. The phase is usually expressed as an angle. In contrast to the wavefunction's amplitude (which is related to the probability), its phase can never be directly observed – it is unobservable. Only differences in phase are observable (e.g. as interference patterns in optics).

### 3.9 Electron Spin

Having just developed a rather sophisticated picture of the electronic wavefunction, we shall immediately retreat to the comfortingly familiar picture of Bohr's orbital atom to explain the next important development in quantum theory!

By 1925, physicists attempting to explain the nature of atomic spectra had realised that not all was correct. Where, according to Bohr's model, just one spectral line should have existed, two were sometimes found very close together. To explain this and other similar puzzles, the Dutch physicists Sam Goudsmit and George Uhlenbeck proposed that the electron spins on its axis as it orbits around the nucleus (just as the Earth spins around the north-south axis as it orbits around the Sun; see Figure 3.8).

The splitting of the spectral lines is explained by the existence of magnetic effects inside the atom. The electron orbit around the nucleus forms a small loop of electric current and so sets up a magnetic field; the orbiting electron behaves like a small magnet. The

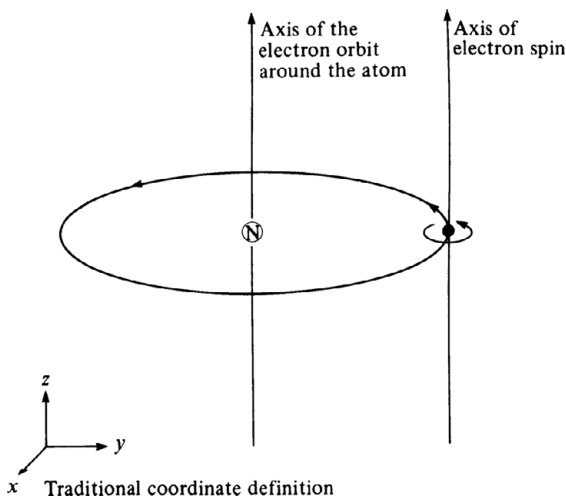


Figure 3.8. In the orbital picture of a particle electron, the electron spins on its own axis.

spin of the electron also has a magnet associated with it, which is referred to as the ‘magnetic moment of the electron’. This interacts with the orbital magnetic moment, adding to or reducing the energy, depending on the way in which the electron is spinning. This will lead to a slight difference in the energy for the different spins of the electron, and will result in the splitting of the spectral line associated with the Bohr orbit.

The above is a nice classical picture, but it has its limitations. The fact that the spectral line splits into just two components indicates that the electron cannot be spinning around at any arbitrary angular momentum but must be such that it has just two values along the line of the atom’s magnetic field (or, in the case of a free electron, along the line of any applied magnetic field). The components of the spin in this direction are referred to as the ‘z components’ (see Figure 3.8) or the ‘third components’ of spin and are measured to be quantised in half-integral units of Planck’s quantum constant (divided by  $2\pi$ ),

$$s_z = \pm \frac{\hbar}{2}.$$

Although the picture of the electron as a spinning ball is attractive, it is important to remember that it is simply a model. In fact, electron spin is purely a quantum concept (it is directly proportional to  $\hbar$ , so if  $\hbar = 0$ , there is no spin!). We must be prepared to think also of the electron as an extended wave which carries a quantum of intrinsic angular momentum, just like its quantum of electric charge.

Other particles also carry spin. The proton and the neutron carry spin quanta which are half-integral multiples of Planck’s constant, just like the electron. The photon also has something like spin, but the quantum is a whole unit of  $\hbar$ . As the photon is simply a packet of electric and magnetic fields this shows that intrinsic angular momentum can be a feature of purely non-material fields. As we shall soon see, the difference in particle spins is very important. On a fundamental level, it gives a method of categorising the behaviour of the wavefunctions of particles under the Lorentz transformations of special relativity (a connection we shall discuss further in Chapter 6). On a practical level, it implies very different behaviours of ensembles of particles (see next section).

### 3.10 The Pauli Exclusion Principle

A straightforward look at the Bohr model of the atom tells us that some fundamental principle must be missing. For there is seemingly nothing to prevent all the electrons of any atom from performing the same orbit. Yet we know that a typical atom will have its electrons spread over several different orbits. Otherwise, transitions between them would be rare, in contradiction to the observations of atomic spectra. So some rule must keep the electrons spread out across the orbits of the atom.

In 1925 the Austrian physicist Wolfgang Pauli derived the principle that no two electrons can simultaneously occupy precisely the same quantum state (i.e. have identical values of momentum, charge and spin in the same region of space). He reached this conclusion after examining carefully the atomic spectra of helium. He found that transitions to certain states were always missing, implying that the quantum states themselves were forbidden. For instance, the lowest orbit (or ground state) of helium in which the two electrons have the same value of spin is not present. But the state in which the two electron spins are opposite is observed.

The power of this principle in atomic physics can hardly be overstated. Because no two electrons can exist in the same state, the addition of extra orbital electrons will successively fill up the outer-lying electron orbits and will avoid over-crowding in the lowest one. Just two electrons are allowed in the ground state because the only difference can be the two values of spin available. More electrons are allowed in the higher orbits because their quantum states can differ by a wide range of orbital angular momenta around

the nucleus (which also turns out to be quantised). It is the Pauli exclusion principle which is responsible for the chemical identities of all atoms of the same element, as it is this principle which determines the allowed arrangements of the atomic electrons.

Although we have focused on the atom, the exclusion principle applies to any quantum system, the extent of which is defined principally by the wavefunctions of the component particles. In the case of totally isolated electrons of definite momentum whose wavefunctions extend over all space, the exclusion principle means that only two electrons with opposite spins can have the same momentum. In the case of electrons confined to a crystal (i.e. electrons whose wavefunctions extend over the dimensions of the crystal), the rule will apply to all electrons in the crystal.

Pauli's exclusion principle can be expressed alternatively in terms of the behaviour of the wavefunction of a quantum system. Although we have talked so far only of the wavefunctions of individual particles, these can be aggregated for any quantum system to give a wavefunction describing the whole system. For example, the total wavefunction of the helium atom can describe the behaviour of two electrons at the same time. Just as the wavefunction of a single electron is a wavepacket reflecting the localisation of the electron, a double-electron wavefunction will contain two wavepacket humps reflecting the localisations of the two electrons. The exclusion principle is a consequence of the fact that a multiple-electron wavefunction must change sign under the interchange of any two electrons. Wherever the wavefunction is positive it must become negative and vice versa. The wavefunction is said to be antisymmetric under the interchange of two electrons. This effect can be understood by considering the two-electron helium atom. Consider the wavefunction for one electron at position  $x_1$  and the other at position  $x_2$ . The wavefunction will be a function of the separation  $x_1 - x_2$ , and by the antisymmetry property,

$$\psi(x_1 - x_2) = -\psi(x_2 - x_1).$$

Then the probability for the two electrons to be at the same point ( $x_1 = x_2$ ) is related to the amplitude of the wavefunction at  $x_1 = x_2$ . But at  $x_1 = x_2$ , the above equation reads

$$\psi(0) = -\psi(0) = 0.$$

Thus the probability for the two electrons to be in the same place is zero and the exclusion principle follows. Note that since any two electrons are indistinguishable, all we are doing in interchanging them is relabelling the electrons and this should make no difference to the physical results (e.g. energy levels and probability densities). The antisymmetry of the wavefunction allows just this. As all physical quantities are proportional to its square, changing only its sign will make no difference.

Particles such as the electron and the proton with spin  $\frac{1}{2}\hbar$  (and other more exotic particles that we shall meet with other half-integral spins  $\frac{3}{2}\hbar, \frac{5}{2}\hbar, \dots$ ) obey the exclusion principle, have antisymmetric wavefunctions under the interchange of two identical such particles and are referred to as *fermions*. This is because ensembles of fermions obey statistics governing their dynamics, which were first formulated by the Italian physicist Enrico Fermi, and the Englishman Paul Dirac. Fermi–Dirac statistics show how momentum is distributed amongst the particles of the ensemble. Because of the exclusion principle in any quantum system, there is a limit to the number of particles which can adopt any particular value of momentum and so this leads to a wide range of momentum carried by the particles. Particles such as the photon with spin  $\hbar$  (and other particles we shall meet with integral spins  $0, \hbar, 2\hbar, 3\hbar, \dots$ ) do not obey the exclusion principle and are called *bosons*. Their wavefunction does not alter under the interchange of two particles. An assembly of Bosons obeys dynamical statistics first formulated by the Indian physicist Satiendranath Bose and Albert Einstein. In Bose–Einstein statistics there is no limit to the number of particles which can have the same value of momentum, and this allows the assembly of bosons to act coherently, as in the case of laser light.

This last principle concludes our whistle-stop tour of quantum mechanics. Although brief, the tour has included most of the new concepts introduced by the theory. For the purposes of the rest of the book, the most important of these is the wavefunction interpretation of a particle, although we will use the uncertainty and exclusion principles from time to time. As in the case of relativity, it is a constant challenge to shrug off our everyday imaginings in the microworld and learn to think in terms of these unfamiliar ideas. But before we are quite ready to approach the subject we must look at what happens when relativity and quantum mechanics are put together.

# 4

## *Relativistic Quantum Theory*

### 4.1 Introduction

Quantum mechanics, just like ordinary mechanics and electrodynamics, must be made to obey the principles of special relativity. Because the entities (particles, atoms, etc.) described by quantum theory quite often travel at speeds at or near  $c$ , this becomes an essential requirement. Special relativity will not just give corrections to conventional Newtonian mechanics, but will dictate dominant, unconventional relativistic effects.

We will see that the synthesis of relativity with quantum theory predicts wholly new and unfamiliar physical consequences (e.g. antimatter). This requires us to develop a new way of looking at matter via quantum fields. If we can then go on to develop the mechanics of interacting quantum fields, this will provide us with the most satisfactory description of the behaviour of matter (both the conventional matter we have discussed so far, and the unconventional antimatter we will introduce along the way).

### 4.2 The Dirac Equation

At the same time as Schrödinger and Heisenberg were formulating their respective versions of the quantum theory, Paul Dirac was attempting the same task. But, in addition, he was concerned that the quantum theory should manifestly respect Einstein's special relativity. This implies two distinct requirements:

firstly, that the theory must predict the correct energy–momentum relation for relativistic particles,

$$E^2 = m_0^2 c^4 + p^2 c^2,$$

and, secondly, that the theory must incorporate the phenomenon of electron spin in a Lorentz covariant fashion.

In one of the most celebrated brainstorming sessions of theoretical physics, Dirac simply wrote down the correct equation! He was guided in this task by realising that Schrödinger's equation for the electronic wavefunction cannot possibly satisfy the requirements of special relativity because time and space enter the equation in different ways (as first- and second-order derivatives respectively). Schrödinger's equation is perfectly adequate for particles moving with velocities much less than  $c$ , and it predicts the correct Newtonian energy–momentum relationship for particles,

$$E = \frac{p^2}{2m} = \frac{mv^2}{2}.$$

But because space and time are not treated correctly, it does not predict the correct relativistic relationships or incorporate energy–mass equivalence.

In the spirit of special relativity, Dirac sought an equation treating space and time on an equal basis. In this he succeeded, but found that in doing so the electron wavefunction  $\psi$  could no longer be a simple number. Incorporating time and space on an equal

basis requires the electron wavefunction  $\psi$  to contain two separate components which in the non-relativistic limit correspond to the probabilities that the electron is spin up (with spin quantum  $+\hbar/2$ ) or spin down (with spin quantum  $-\hbar/2$ ). Thus  $\psi$  is written as a two-component *spinor*,  $\psi = \begin{pmatrix} a \\ b \end{pmatrix}$ . In fact, in the full theory it is a four-component object, for reasons which will become clear in the next section.

So in attempting to incorporate special relativity into quantum mechanics it was necessary to invent electron spin! It is fascinating to wonder whether, if electron spin had not been proposed and discovered experimentally, it would have been proposed theoretically on this basis.

Dirac's equation can be used for exactly the same purposes as Schrödinger's, but with much greater effect. In Section 3.9 we saw that the spin of the electron gives rise to a splitting in the energy levels of the hydrogen atom. This is because the magnetic moment of the electron may either be aligned with, or against, the magnetic field set up by the electron's orbital angular momentum. It was noticed in experiments that the half-integral unit of spin angular momentum  $\hbar/2$  produced as big a magnetic moment as a whole integral unit of orbital angular momentum (i.e. spin is twice as effective in producing a magnetic moment as is orbital angular momentum). This is quantified by ascribing the value of 2 to the gyromagnetic ratio (the *g*-factor) of the electron. This is effectively the constant of proportionality between the electron spin and the magnetic moment resulting. In non-relativistic quantum mechanics,  $g = 2$  is an empirical fact. With the Dirac equation, it is an exact prediction.

The Dirac equation can also explain the fine splitting and hyperfine splitting of energy levels within the hydrogen atom. These result from the magnetic interactions between the electron's orbital angular momentum, the electron spin and the proton spin.

### 4.3 Antiparticles

One immediate consequence of predicting the relativistic relationship between energy and momentum for the electron wavefunction is that the Dirac equation seems to allow the existence of both positive- and negative-energy particles:

$$E = \pm \left( m_0^2 c^4 + p^2 c^2 \right)^{1/2}.$$

In an amazing feat of intellectual bravado, Dirac suggested that this prediction of negative-energy particles was not rubbish but, instead, the first glimpse of a hidden universe of antimatter.

The concept of negative-energy entities is wholly alien to our knowledge of the Universe. All things of physical significance are associated with varying amounts of positive energy. So Dirac did not ascribe a straightforward physical existence to these negative-energy electrons. Instead, he proposed an energy spectrum containing all electrons in the Universe (see Figure 4.1). This spectrum consists of all positive-energy electrons which inhabit a band of energies stretching from  $m_0 c^2$ , the rest mass, up to arbitrarily high energies. These are the normal electrons which we observe in the laboratory and whose distribution over the energy spectrum is determined by the Pauli exclusion principle. Dirac then went on to suggest that the spectrum also contains the negative-energy electrons which span the spectrum from  $-m_0 c^2$  down to arbitrarily large negative energies. He proposed that these negative-energy electrons are unobservable in the real world. To prevent the real, positive-energy electrons simply collapsing down into negative-energy states, it is necessary to assume that the entire negative-energy spectrum is full and that double occupancy of any energy state in the continuum is prevented by the Pauli exclusion principle. No electrons inhabit the energy gap between  $-m_0 c^2$  and  $m_0 c^2$ .

Viewed picturesquely, it is as if the world of physical reality conducts itself while hovering over an unseen sea of negative-energy electrons.

But if this sea of negative-energy electrons is to remain unseen, what is its effect on the everyday world? The answer to this is that elementary particle interactions of various sorts can occasionally transfer enough energy to a negative-energy electron to boost it across the energy gap into the real world. For instance, a photon with energy  $E \geq 2m_0 c^2$  may collide with the negative-energy electron and so promote it to reality. But this cannot be the end of the story, as we seem to have created a unit of electrical charge, whereas we are convinced that this is a quantity which is conserved absolutely. Also, we started out with a photon of energy  $E \geq 2m_0 c^2$  and have created an electron with an energy just over  $m_0 c^2$ . Where has the energy difference of  $m_0 c^2$  gone? We believe that positive energy is also conserved absolutely; it does not disappear into some negative-energy sea.



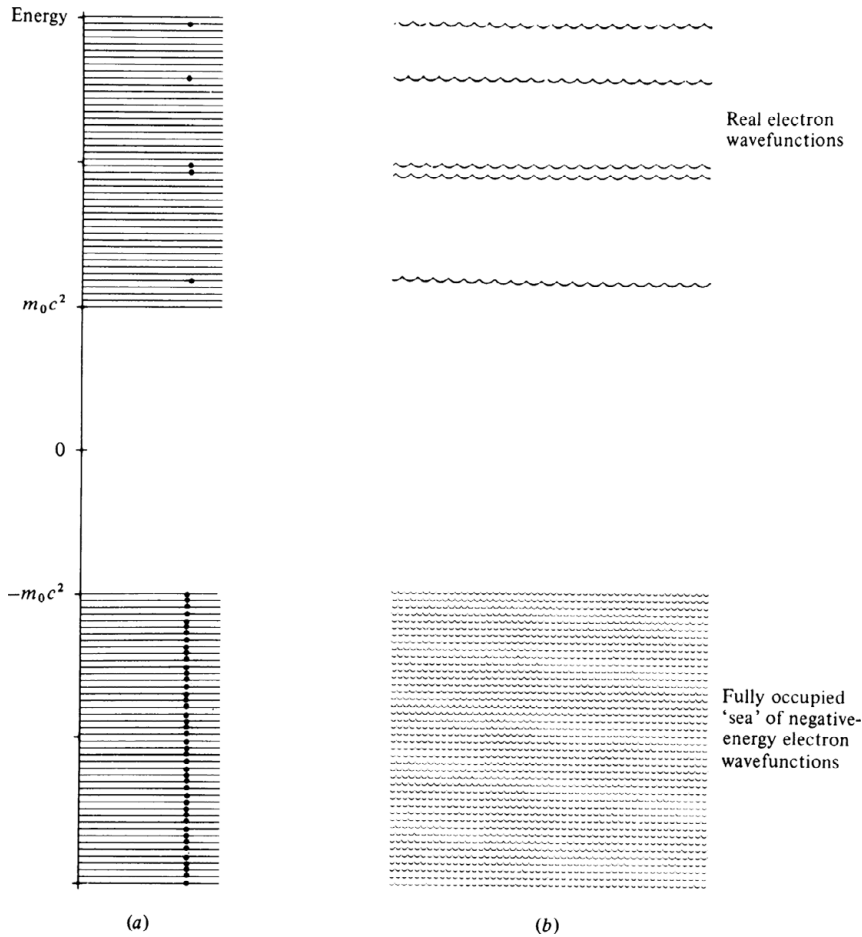


Figure 4.1. Dirac's energy spectrum of electronic states (a) and its interpretation (b).

These problems of interpretation are resolved by proposing that the hole left in the negative-energy sea represents a perceptible, positive-energy particle with an electrical charge opposite to that on the electron. (The absence of a negative-energy particle represents the presence of a positive-energy particle.) This particle is referred to as the *antiparticle* of the electron, is called the *positron*, and is denoted by  $e^+$ .

The positron was first discovered in 1931 by the American physicist Carl Anderson in a cloud chamber photograph of cosmic rays.

Although the arguments given here have concentrated specifically on the electron and the positron, it is important to appreciate that the Dirac equation applies to any relativistic spin- $\frac{1}{2}$  particle, and so too do the ideas of a negative-energy sea and antiparticles.

Both the proton  $p$  and neutron  $n$  can be described by the Dirac equation and seas of negative-energy protons and neutrons may be proposed as coexisting with those of the electrons. The holes in those seas, the antiprotons denoted  $\bar{p}$ , and antineutrons denoted  $\bar{n}$ , took somewhat longer to discover than the positron as, in their case,  $2m_0c^2$  is large. It requires high-energy accelerators to provide probes which are energetic enough to boost the antiprotons into existence. These were not available until the mid-1950s.

The electron wavefunction which is described in the Dirac equation can now be appreciated in its full four-component form. In the Newtonian limit, these components describe, respectively, the spin-up and spin-down states of both the electron and the positron.

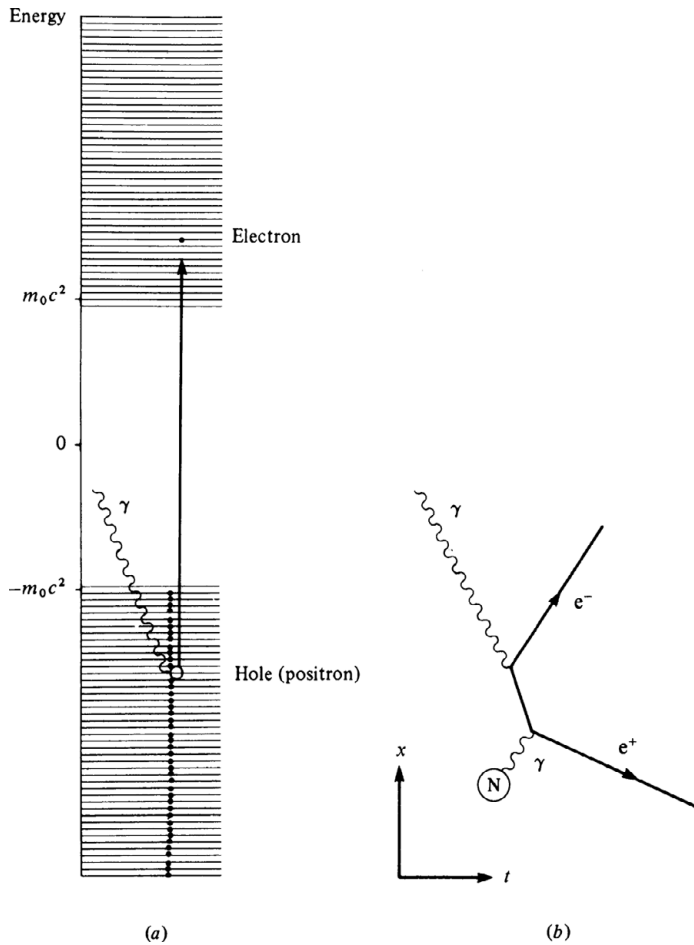


Figure 4.2. Pair creation by a photon  $\gamma$  in the Dirac picture in (a), and in a space–time diagram in (b). Energy and momentum conservation require the subsequent involvement of a nearby nucleus.

The development of the next concept in the microworld is contained in the behaviour of particles and antiparticles. We suggested that an energetic photon can promote a negative-energy electron from the sea, thus leaving a hole. So the photon can create an electron–positron pair from the vacuum. (In fact, this must take place in the presence of another particle to ensure conservation of energy and momentum; see Figure 4.2.) Similarly, an electron and a positron can annihilate each other and give rise to energetic photons. The upshot of this is that particles such as the electron can no longer be regarded as immutable, fundamental entities. They can be created and destroyed just like photons, the quanta of the electromagnetic field.

#### 4.4 Quantum Field Theory (QFT)

In the most sophisticated form of quantum theory, all entities are described by fields. Just as the photon is most obviously a manifestation of the electromagnetic field, so too is an electron taken to be a manifestation of an electron field and a proton of a proton field. Once we have learnt to accept the idea of an electron wavefunction extending throughout space and time (by virtue of Heisenberg’s uncertainty principle for a particle of definite momentum), it is not too great a leap to the idea of an electron field extending throughout space–time. Any one individual electron wavefunction may be thought of as a particular frequency excitation of the field and may be localised to a greater or lesser extent dependent on its interactions.

The electron field variable is, then, the Fourier sum over the individual wavefunctions, where coefficients multiplying each of the individual wavefunctions represent the probability of the creation or destruction of a quantum of that particular wavelength (momentum). The representation of a field as the summation over its quanta, with coefficients specifying the probabilities of the creation and destruction of those quanta, is referred to as *second quantisation*.

First quantisation is the recognition of the particle nature of a wave or of the wave nature of a particle (the Planck–Einstein and de Broglie hypotheses respectively). Second quantisation is the incorporation of the ability to create and destroy the quanta in various reactions.

There is a relatively simple picture which should help us to appreciate the nature of a quantum field and its connection with the notion of a particle. A quantum field is equivalent, at least mathematically, to an infinite collection of harmonic oscillators. These oscillators can be thought of as a series of springs with masses attached. When some of the oscillators become excited, they oscillate (or vibrate) at particular frequencies. These oscillations correspond to a particular excitation of the quantum field and hence to the presence of particles, i.e. field quanta.

We are familiar with the electromagnetic and gravitational fields because, their quanta being bosons, there are no restrictions on the number of quanta in any one energy state and so large assemblies of quanta may act together coherently to produce macroscopic effects. Electron and proton fields are not at all evident because, being fermions, the quanta must obey Pauli's exclusion principle and this prevents them from acting together in a macroscopically observable fashion. So although we can have concentrated beams of coherent photons (laser beams), we cannot produce similar beams of electrons. These instead must resemble ordinary incoherent lights (e.g. torchlights) with a wide spread of energies in the beam.

#### 4.5 Interacting Fields

Having introduced this new, rather nebulous, concept of a field representation of matter, we must now set about using it. Our ultimate objective must be to predict the values of physical quantities which can be measured in the laboratory such as particle reaction cross-sections, particle lifetimes, energy levels in bound systems, etc. We hope to achieve this

by using the idea of quantum fields to tell us the probabilities of the creation and destruction of their quanta in various reactions, and to provide us with descriptions of the behaviour of the quanta between creation and destruction (the wavefunctions). This will then allow us to calculate the probabilities associated with physical processes.

Now the probabilities follow somehow from the dynamics, and the dynamics of any system, whether it be governed by Newtonian mechanics, quantum mechanics or quantum field theory, can be derived from a single quantity describing the system, called its *Lagrangian*. The Lagrangian  $L$  for any system is the difference between its kinetic energy ( $KE$ ) and its potential energy ( $PE$ ),

$$L = KE - PE.$$

For a classical particle, say a cricket ball, moving through the gravitational field of the Earth, the potential energy is due to its height  $x$  above the Earth ( $PE = mgx$ ), and its kinetic energy is due to its velocity ( $KE = \frac{1}{2}mv^2$ ).

In quantum mechanics (or QFT), we are dealing with wavefunctions (or fields) which extend throughout space–time. Here, we do not deal with the total Lagrangian  $L$  directly, but with the Lagrangian density  $\mathcal{L}$ . The total Lagrangian can then be found by integrating the Lagrangian density over all space. Although in future discussions we shall be talking about the properties of the Lagrangian, the comments will properly apply to the Lagrangian density, a fact which we will acknowledge by using the symbol  $\mathcal{L}$ .

It is straightforward to write down the expression for the Lagrangian density of a free electron in terms of the electron wavefunction (or field). For both the cricket ball and the free electron, it is a trivial exercise to go from the Lagrangian to the equations of motion ( $F = ma$  for the cricket ball and the Dirac equation for the electron). But in the case of elementary particles *in interaction* we do not know in general the equations of motion and, where we do, we cannot solve them. We cannot therefore proceed immediately to calculate the quantities of physical interest resulting from the motions of particles, and a more subtle approach is required.

#### 4.6 Perturbation Theory

To describe elementary particle reactions in which quanta can be created and destroyed, it is

necessary to propose an expression for the Lagrangian of the interacting quantum fields. Let us concentrate on interacting electron and photon fields only. The Lagrangian will contain parts which represent free electrons  $\mathcal{L}_0(\psi_e)$  and free photons  $\mathcal{L}_0(A)$ , where  $A$  denotes a four-vector representing the electromagnetic field. It will also contain parts which represent the interactions between electrons and photons,  $\mathcal{L}_{\text{INT}}(\psi_e, A)$ , whose form will be dictated by general principles. These will include, for instance, Lorentz invariance and various conservation laws which the interactions are observed to respect (such as the conservation of electrical charge). In Chapter 20 we shall see how these principles can be expressed in terms of the symmetry of the Lagrangian under various groups of transformations.

The total Lagrangian is then the sum of all these parts:

$$\mathcal{L} = \mathcal{L}_0(\psi_e) + \mathcal{L}_0(A) + \mathcal{L}_{\text{INT}}(\psi_e, A).$$

This is the top-level specification of the fields being described and the way in which they interact. We can proceed to predict the values of physical quantities by following a method developed in the late 1940s by the American physicist Richard Feynman. Feynman derived a set of rules which specifies the propagation of the interacting field quanta as the sum of a set of increasingly complicated sub-processes involving the propagation of the free field quanta (governed by  $\mathcal{L}_0(\psi_e)$  and  $\mathcal{L}_0(A)$ ), with interactions between them (coming from  $\mathcal{L}_{\text{INT}}(\psi_e, A)$ ). Each sub-process in the sum can be represented in a convenient diagram referred to as a Feynman diagram. The rules associate with each diagram a mathematical expression. To calculate the probability of occurrence  $P$  of any physical event involving the quanta of the fields, it is first necessary to specify the initial and final states being observed, denoted  $|i\rangle$  and  $\langle f|$  respectively, and then to select all the Feynman diagrams which can connect the two. The mathematical expression for each diagram is then worked out to give the quantum-mechanical amplitude  $m$  for the sub-process. The amplitude for a number of the individual sub-processes may then be added to give the total amplitude  $M$  which is then squared to give the required probability of occurrence:

$$P = |\langle f|M|i\rangle|^2$$

$$M = m_1^{(1)} + m_2^{(1)}$$

$$m_1^{(2)} + m_2^{(2)} + m_3^{(2)} + \dots$$

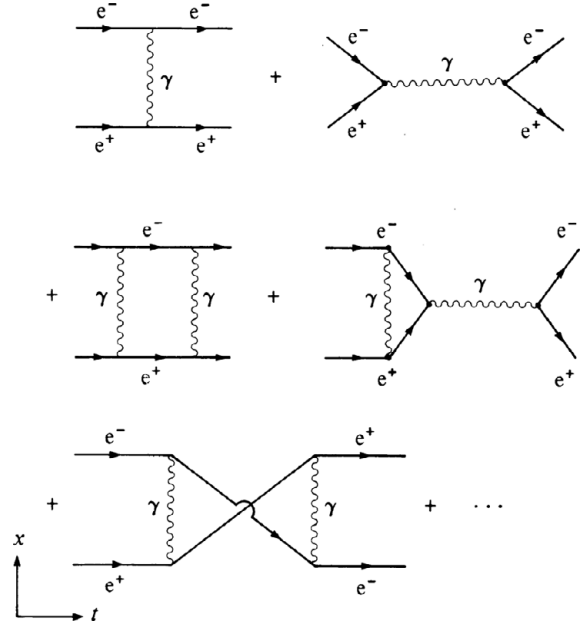


Figure 4.3. The perturbation series containing the various sub-processes possible in electron–positron scattering.

In this notation  $m_i^{(1)}$  denotes the ‘first-order’ diagrams with just two photon–electron vertices involved,  $m_i^{(2)}$  denotes ‘second-order’ diagrams with four photon–electron vertices,  $m_i^{(3)}$  denotes ‘third-order’ diagrams and so on.

For example, in the case of electron–positron elastic scattering, the initial and final states are  $|e^+e^-\rangle$  and  $\langle e^+e^-\rangle$  respectively. A few of the simplest Feynman diagrams connecting the two are shown in Figure 4.3. The first sub-process, amplitude  $m_1^{(1)}$ , is the exchange of a photon between the electron and the positron; the second,  $m_2^{(1)}$ , is the annihilation of the electron and the positron into a photon and its subsequent reconversion; the third,  $m_1^{(2)}$ , is the exchange of two photons and so on.

The probability of occurrence (i.e. of the transformation between initial and final states) may then be restated as the cross-sectional area of two colliding particles, as the mean lifetime for a particle to decay, or as some other appropriate measurable parameter. This is achieved by adopting the kinematical prescriptions which take into account factors like the initial flux of colliding particles, the density of targets available in a stationary target and so on.

The reason why this approach can be adopted is that only the first few of the simplest Feynman diagrams from the infinite series need be considered. This is because the strength of the interaction between electrons and photons (the strength of the electromagnetic force) is small. It can be regarded as a perturbation of free-particle-type behaviour. Another way of stating this is that the probability of the electron or positron interacting with a photon is small. In fact, each photon–electron vertex multiplies the probability of occurrence of the diagram by  $e/\sqrt{(\hbar c)}$ . As each new order of diagram contains a new photon line with two vertices, the relative magnitude of successive orders is reduced by  $e^2/(\hbar c) = \frac{1}{137}$ . So only the first few sub-processes need be calculated to achieve an acceptable approximation to the exact answer.

Summary	
<i>The Lagrangian (L)</i>	specifies the form of the interaction between the fields.
<i>The perturbation principle</i>	approximates the equations of motion by a series of . . .
<i>Feynman diagrams</i>	which show sub-processes between initial and final states involving quanta which may be calculated to give . . .
<i>Probabilities of physical events</i>	which may be stated as cross-sections, lifetimes, etc.

#### 4.7 Virtual Processes

It is important to understand that the dynamics of the individual field quanta within any sub-process of the perturbation expansion are *not* constrained by energy or momentum conservation, provided that the sub-process as a whole does conserve both. This microscopic anarchy is permitted by Heisenberg’s uncertainty principle which states that energy can be uncertain to within  $\Delta E$  for a time  $\Delta t$ , such that

$$\Delta E \Delta t \geq \hbar.$$

So an electron may emit an energetic photon, or a photon may convert into an electron–positron pair over microscopic timescales, provided that energy conservation is preserved in the long run.

These illicit processes are known as ‘virtual processes’. They form the intermediate states of elementary particle reactions. So although we do not

see them, we must calculate the probabilities of their occurrence and add them all up to find the number of different ways for a particle reaction to get from its initial to its final state. A good example of a virtual process is the annihilation of an  $e^+e^-$  pair into a photon. The energy of the  $e^+e^-$  pair must be

$$E_{e^+e^-} = \left(m_{e^+}^2 c^4 + p_{e^+}^2 c^2\right)^{1/2} + \left(m_{e^-}^2 c^4 + p_{e^-}^2 c^2\right)^{1/2},$$

whereas the energy momentum relation of the photon is

$$E_\gamma = p_\gamma c.$$

So it is not possible to have both

$$E_{e^+e^-} = E_\gamma \quad \text{and} \quad p_\gamma = p_{e^+} + p_{e^-}$$

because of the rest mass of the  $e^+e^-$  pair. This means that the virtual photon can exist only as an unobservable intermediate state before dissolving into a collection of material particles which do conserve energy and momentum. Virtual particles are said to be ‘off mass-shell’, because they do not satisfy the relationship  $E^2 = p^2 c^2 + m^2 c^4$ . Massless particles, such as photons, are ‘off mass-shell’ if  $E \neq pc$ .

#### 4.8 Renormalisation

In writing down all the Feynman diagrams of the sub-processes we find some whose amplitude appears to be infinite. These diagrams are generally those with bubbles on either electron or photon wavefunctions or surrounding electron–photon vertices; see Figure 4.4. These diagrams give infinite contributions owing to ambiguities in defining the electron and the photon.

An ordinary electron propagating through space is constantly emitting and absorbing virtual photons. It is enjoying self-interaction with its own electromagnetic field (of which its own charge is the source). So the wavefunction of the electron is already dressed up with these virtual photons; see Figure 4.5(a). Similarly, a photon propagating through space is free to exist as a virtual  $e^+e^-$  pair, and the full photon wavefunction already contains the probabilities of this occurring (Figure 4.5(b)). Also, the electric charge, which we denote  $e$ , already contains the quantum corrections implied by the diagram of Figure 4.4(c).

In 1949, Feynman, Schwinger, Dyson and Tomonaga showed how the infinite contributions to the perturbation series can be removed by redefining

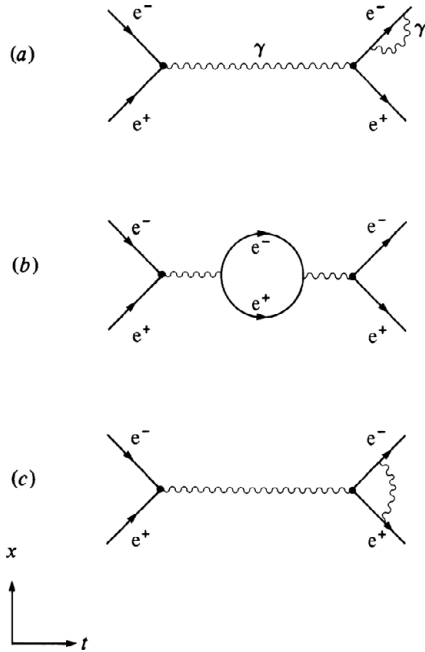


Figure 4.4. (a), (b) and (c). Diagrams with ‘bubbles’ which give infinite contributions to the perturbation series.

the electron, photon and electric charge to include the quantum corrections. When the real electrons, photons and charges appear, the infinite diagrams are included implicitly and should not be recounted. The mathematical proof of this demonstration is known as ‘renormalisation’.

Renormalisation is a necessary formal process which shows that the particles in the theory and their interactions are consistent with the principles of quantum theory. These may seem like hollow words for the familiar interactions of electrons with photons. But in the more esoteric quantum field theories we are going to encounter, both the particle content of the theories and the form of their interactions are largely unknown. In these cases, the ability to renormalise the perturbation expansion of the Lagrangian is a good guide to the acceptability of the theory.

#### 4.9 The Quantum Vacuum

In classical (non-quantum) physics, empty space–time is called the *vacuum*. The classical vacuum is utterly featureless. However, in quantum mechanics, the vacuum is a much more complex entity: it is far from featureless and far from empty. Actually,

the quantum vacuum is just one particular state of a quantum field. It is the quantum-mechanical state in which no field quanta are excited, that is, no particles are present. Hence, it is the ‘ground state’ of the quantum field, the state of minimum energy.

Let us recall the analogy, introduced above in Section 4.4, between a quantum field and an infinite collection of harmonic oscillators (masses connected to springs). In the vacuum, every oscillator is in its ground state. For a classical oscillator, this means that it is motionless: the spring holds the mass in a fixed position. However, for a quantum oscillator, the uncertainty principle means that neither position nor momentum is precisely fixed, and both are subject to random quantum fluctuations. These fluctuations are called zero-point oscillations, or zero-point vibrations. So, the quantum vacuum is full of fluctuating quantum fields. There are no real particles involved, only virtual ones. Virtual particle–antiparticle pairs continually materialise out of the vacuum, propagate for a short time (allowed by the uncertainty principle) and then annihilate.

These zero-point vibrations mean that in the vacuum – the state of minimum energy – there is a zero-point energy associated with any quantum field. Since there is an infinite number of harmonic oscillators per unit volume, the total zero-point energy density is, in fact, infinite. We have already seen that some sense can be made of infinite quantities through the process of renormalisation. As it is usually implemented, this yields a zero energy density for the standard quantum vacuum.

It is extremely difficult to observe these vacuum fluctuations, since there is no state of lower energy with which the vacuum can be compared. However, there is one situation in which its effects can be seen indirectly. In 1948, Hendrik Casimir predicted that two clean, neutral, parallel, microscopically flat metal plates attract each other by a very weak force that varies inversely as the fourth power of the distance between them. The ‘Casimir effect’ was experimentally verified in 1958. It can be understood in the following way. The zero-point energy filling the vacuum exerts pressure on everything. In most circumstances, this pressure is not noticeable, since it acts in all directions and the effect cancels. However, the quantum vacuum has different properties between the two metal plates. Some of the zero-point vibrations of the electromagnetic field are suppressed, namely, those with wavelengths too long to fit between the plates.

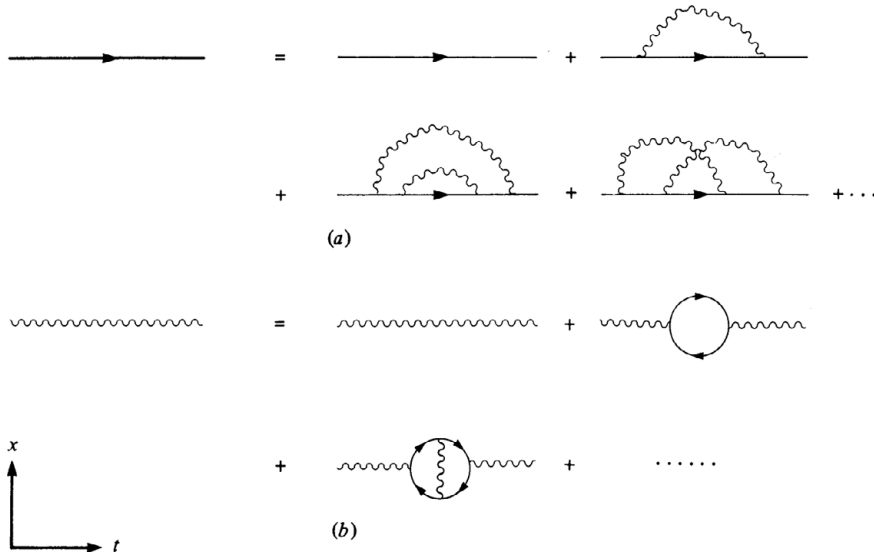


Figure 4.5. (a) The completed (‘dressed’) electron wavefunction already contains its quantum corrections (interactions with virtual photons). (b) The photon propagator likewise.

So, the zero-point energy density between the plates is *less* than that of the standard vacuum, i.e. it is negative. From this it follows that the pressure outside is greater and hence the plates feel an attractive force.

### 4.10 Quantum Electrodynamics

This is the name (often abbreviated to QED) given to the relativistic quantum field theory describing the interactions of electrically charged particles via photons. The discovery of the perturbation expansion revealed the existence of an infinite number of ever-decreasing quantum corrections to any electromagnetic process. The renormalisability of QED means that we can avoid apparently infinite contributions to the perturbation expansion by careful definition of the electron and photon. Therefore we can calculate the value of observable parameters of electromagnetic processes to any desired degree of accuracy, being limited only by the computational effort required to evaluate the many hundreds of Feynman diagrams which are generated within the first few orders (first few powers of  $e^2/(\hbar c)$ ) of the perturbation expansion. This has led to some spectacular agreements between theoretical calculations and very accurate experimental measurements.

The  $g$ -factor of the electron is not, in fact, exactly equal to 2 (as predicted by the Dirac equation).

Its value is affected by the quantum corrections to the electron propagator illustrated in Figure 4.5(a). Essentially, the virtual photons of the quantum corrections carry off some of the mass of the electron while leaving its charge unaltered. This can then affect the magnetic moment generated by the electron during interactions. The measure of agreement between QED and experimental measurement is given by the figure for the modified  $g$ -factor:

$\frac{g}{2} =$	1.001 159 652 41	experimental
	$\pm 0.000\ 000\ 000\ 20$	measurement
$\frac{g}{2} =$	1.001 159 652 38	theoretical
	$\pm 0.000\ 000\ 000\ 26$	prediction.

There are several other such amazing testaments to the success of QED, including numbers similar to the above for the  $g$ -factor of the muon (a heavy brother of the electron which we will meet soon), and yet more subtle shifting of the exact values of the energy levels within the hydrogen atom, the so-called Lamb shift.

This success makes QED the most precise picture we have of the physical world (or at least the electromagnetic phenomena in it). For this reason we shall look at QED again in Part VI in an attempt to discover the fundamental principles behind it (i.e. behind the form of the interaction between the fields). This is so

that we can attempt to repeat the theory's success for the other forces in nature.

#### **4.11 Postscript**

We have now looked at the frontiers of physics as they appeared at the turn of the last century and have seen that relativity and quantum mechanics emerged in turn from the vacuum of knowledge beyond those frontiers. The realisation that relativity and quantum mechanics must be made mutually consistent led to the discovery of antiparticles, which led in turn to the concept of quantum fields. The theory of interacting quantum fields is the most satisfactory description of elementary particle behaviour. All calculations in quantum field theory follow from the specification of the correct interaction Lagrangian, which is

determined by the conservation laws obeyed by the force under study.

We have developed this picture of the world almost exclusively in terms of the particles interacting by the electromagnetic force. It is now time to turn our attention to the other particles and forces in nature to see if they are amenable to a similar treatment.

In what follows, we shall often use the language of particle wavefunctions rather than that of quantum fields. Although somewhat imprecise, a particle wavefunction is a slightly more convenient and intuitive concept in most situations. However, there will be occasions in later chapters in which a proper understanding of certain phenomena demands that we consider the quantum fields themselves rather than wavefunctions.





**Part II**  
**Basic Particle Physics**



# 5

## *The Fundamental Forces*

### 5.1 Introduction

It is an impressive demonstration of the unifying power of physics to realise that all the phenomena observed in the natural world can be attributed to the effects of just four fundamental forces. These are the familiar forces of gravity and electromagnetism, and the not-so-familiar weak and strong nuclear forces (generally referred to as the ‘weak’ and ‘strong’ forces). Still more impressive is the fact that the phenomena occurring in the everyday world can be attributed to just two: gravity and electromagnetism. This is because only these forces have significant effects at observable ranges. The effects of the weak and strong nuclear forces are confined to within, at most,  $10^{-15}$  m of their sources.

With this in mind, it is worthwhile summarising a few key facts about each of the four forces before going on to look at the variety of phenomena they display in our laboratories. In each case we are interested in the sources of the force and the intrinsic strength of the interactions to which they give rise. We are interested also in the space–time properties of the force: how it propagates through space and how it affects the motions of particles under its influence. Finally, we must consider both the macroscopic (or classical) description of the forces (where appropriate) and the microscopic (or quantum-mechanical) picture (where possible).

### 5.2 Gravity

Gravity is by far the most familiar of the forces in human experience, governing phenomena as diverse as falling apples and collapsing galaxies. At the non-relativistic level, the source of the gravitational force is mass and, because there is no such thing as negative mass, this force is always attractive. Furthermore, it is independent of all other attributes of the bodies upon which it acts, such as electric charge, spin, direction of motion, etc.

The gravitational force is described classically by Newton’s famous inverse square law, which states that the magnitude of the force between two particles is proportional to the product of their masses and inversely proportional to the square of the distance between them:

$$F = G \frac{m_1 m_2}{r^2}.$$

The strength of the force is governed by Newton’s constant,  $G$ , and is extremely feeble compared with the other forces (see Table 5.1). We notice the effects of gravity only because it is the *only* long-range force acting between electrically neutral matter. In the microworld, the effects of gravity are mainly negligible. Only in exotic situations, such as on the boundary of a black hole and at the beginning of the Universe, do the effects of gravity on the elementary particles become important.

Table 5.1. Relative strengths of forces as expressed in natural units.\*

Force	Range	Strength	Acts on
Gravity	$\infty$	$G_{\text{Newton}} \approx 6 \times 10^{-39}$	All particles
Weak nuclear force	$< 10^{-18}$ m	$G_{\text{Fermi}} \approx 1 \times 10^{-5}$	Leptons, hadrons
Electromagnetism	$\infty$	$\alpha = \frac{1}{137}$	All charged particles
Strong nuclear force	$\approx 10^{-15}$ m	$g^2 \approx 1$	Hadrons

\* The dimensionality of the forces is removed by dividing out the appropriate powers of  $\hbar$  and  $c$ , to leave a dimensionless measure of the forces' intrinsic strengths. Note that for gravity and the weak force, a mass must also be introduced to give a dimensionless quantity. In the table we have used the proton mass.

The mechanism which gives rise to this force in the classical picture is that of the gravitational field, which spreads out from each mass-source to infinity. A test mass will interact with the gravitational field. At each point in space, the interaction between the test mass and the gravitational field reproduces the gravitational force. However, according to Newton's theory, when a mass-source moves, the gravitational field it sets up changes instantaneously to accommodate its new position. This instantaneous change is fundamentally incompatible with the theory of special relativity, which requires that disturbances cannot propagate faster than light. This motivated Einstein to formulate a new theory of gravity and relativity, called general relativity, which he completed in 1915.

A further feature of Newton's formula is that the quantity characterising a source of gravity – its *gravitational* mass – is identical to the quantity – its *inertial* mass – which characterises its acceleration in response to an applied force, as given by another of Newton's famous equations:

$$F = ma.$$

This equivalence between gravitational and inertial mass, which was known to many generations of physicists before him, led Einstein to speculate on the connection between gravity and acceleration. The principle of equivalence is the apotheosis of this connection, and formed the basis of his conceptual leap from the theory of special relativity to the theory of general relativity. Put simply, the equivalence principle declares that, at any given point, gravitation and acceleration are indistinguishable phenomena.

### 5.2.1 General Relativity

We saw how, in special relativity, observers' perceptions of time and space are modified by factors

depending on their relative velocities. From this it follows that during acceleration (changing velocity) an observer's scales of time and space must become distorted. By the principle of equivalence, an acceleration is identical to the effects of a gravitational field, and so this too must give rise to a distortion of space–time. Einstein's general relativity goes on to explain the somewhat tenuous reality of the gravitational field as the warping of space–time around a mass-source. Thus a mass will distort space–time rather like a bowling ball laid on a rubber sheet. And the effect of gravity on the trajectory of a passing particle will be analogous to rolling a marble across the curved rubber sheet (see Figure 5.1).

So, general relativity suggests that instead of thinking of bodies as moving under the influence of a gravitational force, we should think of them as moving freely through a warped, or curved, space–time. Hence, the force of gravity is reduced to the curved geometry of space–time. Geometry has, as we know, different rules on a curved surface. For example, on the curved surface of the Earth, two north-pointing lines which are parallel at the equator (lines of longitude) actually meet at the north pole; whereas on a flat surface, two parallel lines never meet. In fact, in curved space–time, straight lines must be replaced by *geodesics* as the shortest path between two points; free particles move along geodesics. (On the surface of the Earth, geodesics correspond to great circles.) Einstein embodied this interpretation of gravity as geometry in his field equations of general relativity:

$$G_{\mu\nu} = 8\pi GT_{\mu\nu},$$

which loosely translates as

$$\left( \begin{array}{c} \text{geometry of} \\ \text{space–time} \end{array} \right) = 8\pi G \times \left( \begin{array}{c} \text{mass and} \\ \text{energy} \end{array} \right).$$

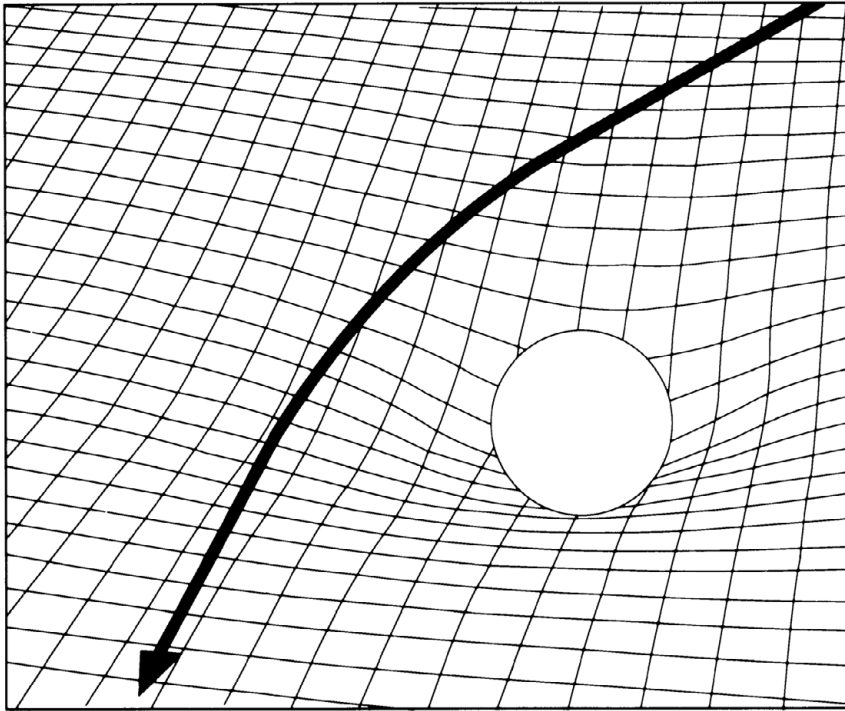


Figure 5.1. According to general relativity, mass bends space–time, which gives rise to the trajectories associated with gravity.

So, mass and energy determine the curvature and geometry of space–time; and the curvature and geometry of space–time determine the motion of matter. In other words, ‘matter tells space–time how to bend’, and ‘space–time tells matter how to move’.

The theory goes on to predict the existence of *gravitational waves* propagating through space as the result of some changes in a mass-source (such as the collapse of a star into a neutron star or black hole). In this event, distortions in space–time will spread out spherically in space at the speed of light, rather like ripples spread out circularly across the surface of a still pond into which a stone is dropped. In Part XIII, we will review the spectacular 2015 discovery of these waves at the LIGO observatories in the US and the resulting possibilities for the emerging fields of gravitational-wave and multi-messenger astronomy.

### 5.2.2 Quantum Gravity

It is important to remember that Einstein’s general relativity is still a classical theory; it does not account for gravity in the quantum-mechanical

regime. A successful quantum theory of gravity has not yet been formulated, and the reconciliation of general relativity with quantum mechanics is one of the major outstanding problems in theoretical physics. It is straightforward enough to take the first few steps towards such a theory, following an analogy with quantum electrodynamics. We can propose that the gravitational field consists of microscopic quanta called *gravitons* which must be massless (to accommodate the infinite range of gravity) and of spin 2 (for consistency with general relativity). The gravitational force between any two masses can then be described as an exchange of gravitons between them. Problems arise, however, because, unlike quantum electrodynamics, certain graviton sub-processes always seem to occur with an infinite probability – quantum gravity is not renormalisable. We shall return to quantum gravity in Part XIV.

### 5.3 Electromagnetism

This is the force of which we have the fullest understanding. This is possibly a reflection of its

physical characteristics: it is of infinite range, allowing macroscopic phenomena to guide our understanding of classical electromagnetism, and it is a reasonably weak force, allowing its microscopic quantum phenomena to be understood using perturbation theory. The strength of the electromagnetic force is characterised by the *fine structure constant*:  $\alpha = e^2/\hbar c = \frac{1}{137}$ .

The source of this force is, of course, electric charge which can be either positive or negative, leading to an attractive force between unlike charges and a repulsive force between like charges. When two charges are at rest, the electrostatic force between them is given by Coulomb's law, which is very similar to Newton's law of gravity, namely that the magnitude of the force is proportional to the product of the magnitudes of the charges involved (empirically observed to exist only as multiples of the charge on an electron) and inversely proportional to the square of the separation between them:

$$F = K \frac{N_1 e \cdot N_2 e}{r^2},$$

where  $N_1$  and  $N_2$  are integral multiples of the charge on the electron  $e$  and  $K$  is a constant depending on the electrical permittivity of free space. New mysteries are introduced by the concept of electric charge. What is it, other than a label for the source of a force we observe to act? Why does it exist only in quanta? Why is the charge quantum on the electron exactly opposite to that on the proton? These are largely taken for granted in classical electromagnetism and are only now being addressed in the modern theories described in Part IX.

Unlike gravity, when electric charges start to move, qualitatively new phenomena are introduced. A moving charge has associated with it not only an electric field, but also a magnetic field. A test charge will always be attracted (or repelled) along the direction of the electrical field, i.e. along a line joining the centres of the two charges. But the effect of the magnetic field is that a test charge will be subjected to an additional force along a direction which is mutually perpendicular to the relative motion of the source charge and to the direction of the magnetic field (Figure 5.2). These properties imply that the combined electromagnetic force on a particle cannot be described simply by a number representing the magnitude of the force but, instead, by a vector quantity describing the magnitude of the forces acting in each of the three directions.

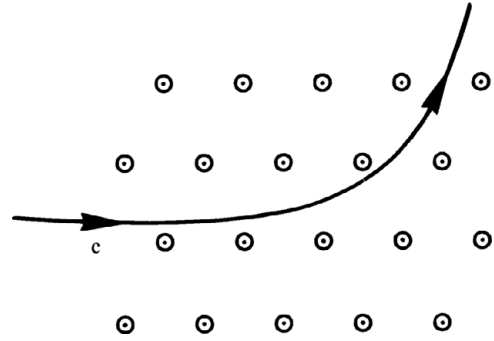


Figure 5.2. The motion of a charged particle in a magnetic field directed out of the plane of the paper.

When a charge is subject to an acceleration, then a variation in electric and magnetic fields is propagated out through space to signal the event. If it is subject to regular accelerations, as may occur when an alternating voltage is applied to a radio aerial, then the charge emits an electromagnetic wave which consists of variations in the electric and magnetic fields perpendicular to the direction of propagation of the wave, see Figure 5.3. Such an electromagnetic wave is part of the electromagnetic spectrum which contains, according to the frequency of oscillation of the fields, radio waves, infrared waves, visible light, ultraviolet light, X-rays and gamma rays, see Figure 5.4.

Electromagnetic phenomena are all described in the classical regime by Maxwell's equations, which allow us to calculate, say, the electric field resulting from a particular configuration of charges, or the wave equation describing the propagation of electric and magnetic fields through space.

One interesting feature of these equations is that they are asymmetric owing to the absence of a fundamental quantum of magnetic charge. It is possible to conceive of a source of a magnetic field which would give rise to an elementary magnetostatic force. Such a magnetic charge would appear as a single magnetic pole, in contrast to all examples of terrestrial magnets which consist invariably of north-pole–south-pole pairs. These conventional magnets are magnetic dipoles which are the result of the motions of the atomic electrons. The possibility of the existence of truly fundamental magnetic monopoles has been a popular topic of research in recent decades following

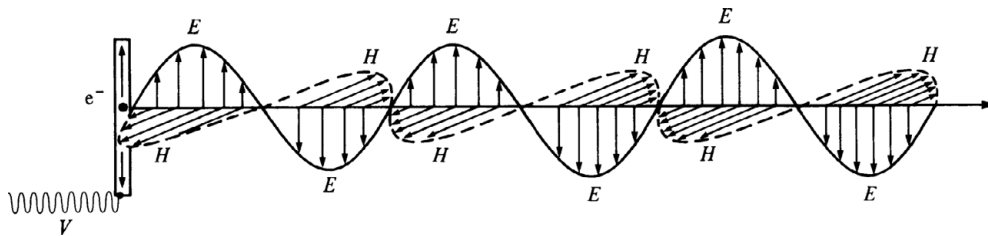


Figure 5.3. The propagation of an electromagnetic wave resulting from the regular accelerations of a charge.

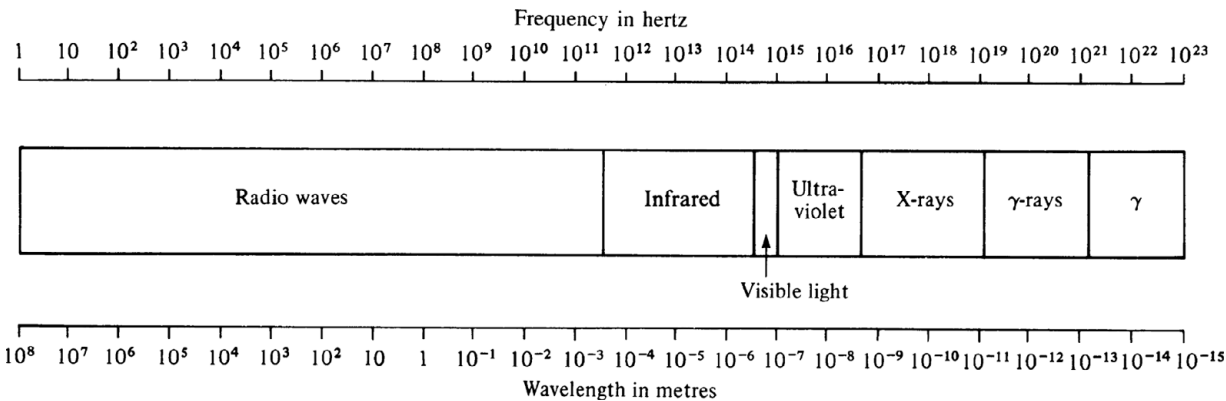


Figure 5.4. The electromagnetic spectrum.

their emergence from the most modern theories (see Parts X onwards).

We have already seen, in Part I, how we can formulate the quantum theory of electrodynamics by describing the interactions of charged particles via the electromagnetic field as the exchange of the quanta of the field, the photons, between the particles involved. QED is the paradigm quantum theory towards which our descriptions of the other forces all aspire.

### 5.4 The Strong Nuclear Force

When the neutron was discovered by James Chadwick in 1932, it became obvious that a new force of nature must exist to bind together the neutrons and protons (referred to generically as *nucleons*) within the nucleus. (Prior to this discovery physicists seriously entertained the idea that the nucleus might have consisted of protons and electrons bound together by the electromagnetic force.) Several features of the new force are readily apparent.

Firstly, as the nucleus was realised to consist only of positively charged protons and neutral neutrons confined within a very small volume (typically of diameter  $10^{-15}$  m), the strong force must

be very strongly attractive to overcome the intense mutual electrostatic repulsions felt by the protons. The binding energy of the strong force between two protons is measured in millions of electronvolts, MeV, as opposed to typical atomic binding energies which are measured in electronvolts (see Appendix A for definitions of these units of energy).

The second fact concerning the strong nuclear force is that it is of extremely short range. We know this because Rutherford's early scattering experiments of  $\alpha$  particles by atomic nuclei could be described by the electromagnetic force alone. Only at higher energies, when the  $\alpha$  particles are able to approach the nuclei more closely, are any effects of the strong force found. In fact, the force may be thought of as acting between two protons only when they are actually touching, implying a range of the strong force similar to that of a nuclear diameter of about  $10^{-15}$  m.

Finally, the last fact we shall mention is that the strong nuclear force is independent of electric charge in that it binds both protons and neutrons in a similar fashion within the nucleus.

One consequence of the solely microscopic nature of the strong force is that we should expect it to



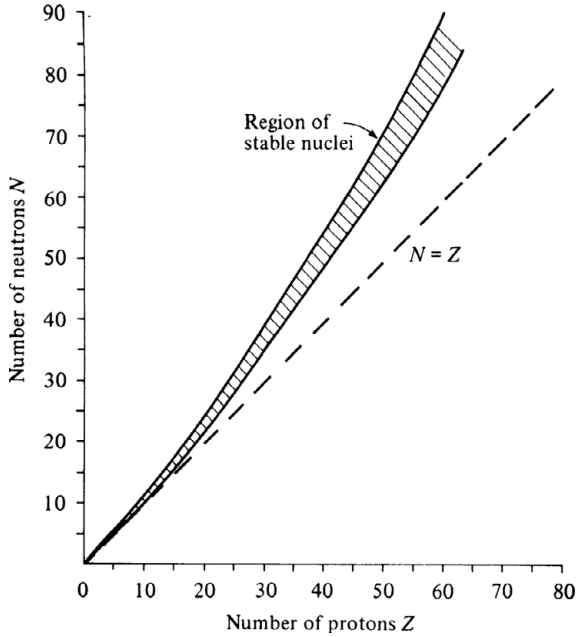


Figure 5.5. Nuclear stability against radioactive decay is governed by the ratio of protons to neutrons.

be a uniquely quantum phenomenon. We can expect no accurate interpretation in terms of classical physics but only in the probabilistic laws of quantum theory.

One of the prime sources of early information on the strong force was the phenomenon of radioactivity and the question of nuclear stability. This involves the explanation of the neutron/proton ratios of the stable or nearly stable nuclei. These can be displayed as a band of stability on the plane defined by the neutron number,  $N$ , and the proton number,  $Z$ , of the nucleus, as in Figure 5.5.

The fact that it is predominantly the heavier nuclei which decay confirms our picture of the short-range nature of the strong force. If we naively think of the nucleus as a bag full of touching spheres, then if the force due to any one nucleon source were able to act on all other nucleons present, we would expect nuclei with more nucleons to enjoy proportionately stronger binding and thus greater stability. (Adding the  $n$ th nucleon to a nucleus would give rise to an extra  $(n - 1)$  nuclear bonds and so a binding energy which increases with  $n$ .) This is observed not to be the case. It is the heavier nuclei which suffer radioactive decay, indicating an insufficient binding together of

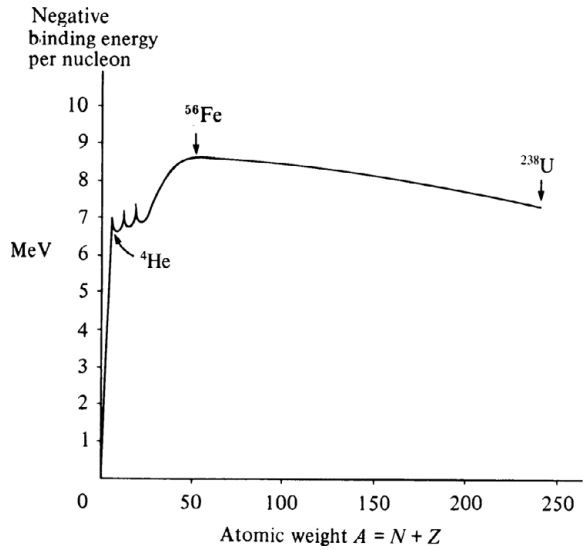


Figure 5.6. Nuclear stability can be expressed in terms of the binding energy available per nucleon in the nucleus.

the nucleons. This is because the nuclear force acts only between touching, or ‘nearest-neighbour’ nucleons. The addition of any extra nucleon will then give rise only to a constant extra binding energy whereas the electric repulsion of the protons is a long-range force and does grow with the number of protons present.

Thus the question of nuclear stability may be described in part by the balance of the repulsive electrical forces and the attractive strong forces affecting the nucleons. It is possible to calculate the sum of these two forces for each nucleus and so to calculate the average binding energy per nucleon in each case, the more negative the binding energy indicating that the more strongly bound are the nucleons within the nuclei. This can be shown graphically as in Figure 5.6. The relatively small negative binding energy of the light atoms results from them not having enough nucleons to saturate fully the nearest-neighbour strong interactions available. The most strongly bound nuclei are those in the mid-mass range, like iron, which more efficiently use the strong force without incurring undue electric repulsion. The heavier nuclei suffer because the electric repulsion grows by an amount proportional to the number of protons present.

As nature attempts to accommodate heavier and heavier nuclei, a point is reached where it becomes

energetically more favourable for the large nuclei to split into two more tightly bound, mid-mass-range nuclei. This gives rise to an upper limit on the weights of atoms found in nature, occupied by uranium-238 with 92 protons and 146 neutrons. By bombarding uranium with neutrons, it is possible to exceed nature's stability limit causing the uranium + neutron nucleus to split into two. This is nuclear fission.

Radioactive  $\alpha$  decay occurs when an element is not big enough to split into two, but would still like to shed some weight to move up the binding energy curve to a region of greater stability. The  $\alpha$  particle (which is a helium nucleus consisting of two protons and two neutrons) will have existed as a 'nucleus within a nucleus' prior to the decay. By borrowing energy for a short time according to Heisenberg's uncertainty principle, it will be able to travel beyond the range of the strong attractive forces of the remaining nucleons to a region where it is subject only to the electrical repulsion due to the protons. Thus the nucleus is seen to expel an  $\alpha$  particle, see Figure 5.7. Because the energy is borrowed according to the probabilistic laws of quantum theory, it is not possible to specify a particular time for  $\alpha$  decay, but only to specify the time by which there will be a, say, 50% probability of a given nucleus having undergone the decay (corresponding to the average time needed for 50% of a sample to decay). This is called the *half-life* of the element, denoted by  $\tau_{1/2}$ .

Another feature of nuclear stability can be explained by the action of another quantum principle. Although we have explained why too many protons in a massive nucleus may cause it to break up, we have not explained why this cannot be countered by simply adding an arbitrary number of neutrons to gain extra attractive strong forces. The reason is due to Pauli's exclusion principle. Because both protons and neutrons are fermions, no two protons and no two neutrons can occupy the same quantum state. We cannot simply add an arbitrary number of neutrons to dilute the repulsive effects of the electric charges on the protons, as the exclusion principle forces the neutrons to stack up in increasingly energetic configurations leading to a reduction in the negative binding energy per nucleon and so to decreased stability.

Although we have now reviewed some facts about the strong force (its short range, charge independence and spin dependence via the exclusion principle, etc.), we have done nothing to explain the mechanism

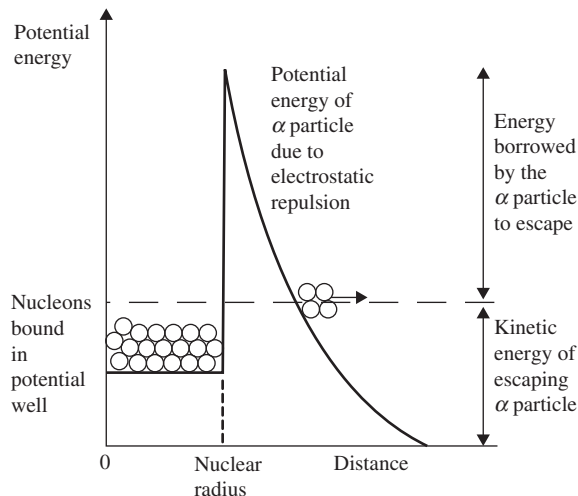


Figure 5.7. Radioactive  $\alpha$  decay. An  $\alpha$  particle within the nucleus borrows enough energy via Heisenberg's uncertainty principle to overcome the potential binding energy of the strong force.

of its action apart from noting that only a quantum picture will be suitable for such a microscopic phenomenon. Yukawa's formulation of his meson theory of the strong force is the point of departure into particle physics proper from the inferences of nuclear physics just discussed, and is described in Chapter 7.

### 5.5 The Weak Nuclear Force

One of the most obvious features of the neutron is that it decays spontaneously into a proton and an electron with a half-life of about 10 minutes. This period is much longer than any of the phenomena associated with the strong force, and it is difficult to imagine how the electromagnetic force could be responsible for this comparative stability. So, we are led to the conclusion that neutron decay is due to some qualitatively new force of nature.

It is this 'weak' force causing neutron decay which lies behind the phenomenon of the radioactive  $\beta$  decay of nuclei described in Chapter 1. The decay of the neutron into a proton allows a nucleus to relieve a crucial neutron surplus which, because of the action of the Pauli exclusion principle, may be incurring a substantial energy penalty and eroding the binding energy of the nucleus.

The same interaction may also allow the reverse reaction to occur in which a nuclear proton transfers into a neutron by absorbing an electron. (This may

occur because of the very small but finite chance that the electron may find itself actually inside the nucleus, according to the positional uncertainty represented by the electron wavefunction.) This reaction will allow a proton-rich nucleus suffering from undue electric repulsion to dilute its proton content slightly, thereby strengthening its binding.

One problem soon encountered in attempts to explain radioactive  $\beta$  decay is that the electrons which are emitted from decaying nuclei are seen to emerge with a range of energies up to some maximum which is equal to the difference in the masses of the initial and final nuclei involved. When the electrons emerge with less than this maximum figure we seem to have lost some energy. To avoid this apparent violation of energy conservation (and also an accompanying apparent violation of angular momentum conservation), Pauli postulated in 1930 that another, invisible, particle was also emitted during the decay, which carries off the missing energy and angular momentum. As the original reactions do conserve electric charge, then this new particle must be neutral. On the strength of this, Fermi called it the *neutrino*.

Several properties of the neutrino are apparent from the facts of  $\beta$  decay. Because of conservation of energy, it is necessary that the neutrino be very

light or indeed massless (because some electrons do emerge with the maximum energy allowed by the mass difference). Similarly, because of angular momentum conservation the neutrino must be spin  $\frac{1}{2}$ . Another interesting feature is that the neutrino interacts with other particles only by the weak force and gravity (because the strong interaction is obviously not present in neutron decay, and because the uncharged neutrino experiences no electromagnetic effects).

The apparent invisibility of the neutrino is due to the very feebleness of the weak force, as indicated in Table 5.1. This reluctance to interact allows it to pass through the entire mass of the Earth with only a minimal chance of interaction en route. Because of this, the neutrino was not observed (i.e. collisions attributable to its path were not identified) until large neutrino fluxes emerging from nuclear reactors became available. This was achieved in 1956 by Reines, some 26 years after Pauli's proposal.

The weak force, like its strong counterpart, acts over microscopic distances only. In fact, to all intents and purposes it makes itself felt only when particles come together at a point (i.e. below any resolving power available to physics, say less than  $10^{-18}$  m). We shall discuss it further in Parts IV, V and VI.

# 6

## *Symmetry in the Microworld*

### 6.1 Introduction

In the everyday world, symmetries in both space and time have a universal fascination for the human observer. In nature, the symmetry exhibited in a snowflake crystal or on a butterfly's wings might be taken to indicate some divine guiding hand, while in art the pleasures of design or of a fugue may be seen as its imitation. Pleasing as symmetry may be, however, its significance generally remains unappreciated.

In the world of physics, and especially in the microworld, symmetries are linked closely to the actual dynamics of the systems under study. They are not just interesting patterns or an artistic disguise for science's passion for classification. Indeed, it is no exaggeration to say that symmetries are the most fundamental explanation for the way things behave (the laws of physics).

Historically this has not always been appreciated. It is, of course, the case that physicists notice natural phenomena and write down equations of motion to describe them (notably Newton, Einstein and Dirac, to name but an illustrious few). But in describing the microworld it is generally far too difficult to write down the equations of motion straight away; the forces are unfamiliar and our experiments provide only ground-floor windows into the skyscraper of the high-energy domain. So we are forced to consider first the symmetries governing the phenomena under study, generally indicated by the action of conservation rules of one sort or another (e.g. energy, momentum and

electric charge). The symmetries may then guide our investigation of the nature of the forces to which they give rise.

Symmetry is described by a branch of mathematics called 'group theory'. A group is simply a set of symmetry transformations, changes under which a system stays the same. The action of these transformations on a specific set of objects is called a *representation*. The notion is made non-trivial by the demand that repeated transformations are equivalent to another transformation. When a physical system has such a symmetry, the Lagrangian governing the system does not change under the group transformations. This then implies the existence of a conserved quantity. This connection is due to a remarkable mathematical theorem by Emmy Noether which states that, for every continuous symmetry of a Lagrangian, there is a quantity which is conserved by its dynamics.

We can proceed to put some flesh on this theoretical skeleton by giving examples of four kinds of symmetry used extensively in particle physics: continuous space-time symmetries, discrete symmetries, dynamical symmetries and internal symmetries. After a look at each of these, we will mention in closing how even broken symmetries can provide a useful guide to the formulation of physical laws.

### 6.2 Space-Time Symmetries

Foremost amongst these are the operations of translation through space, translation through time and

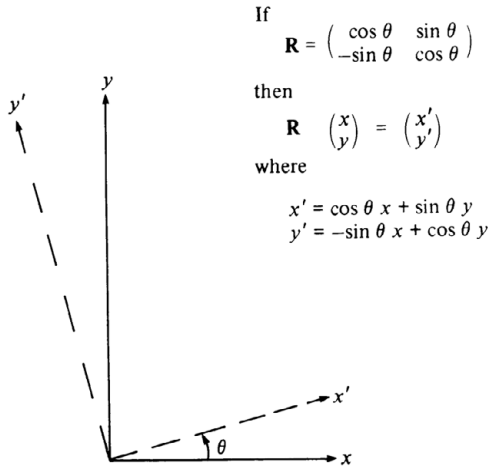


Figure 6.1. Rotations redefine a coordinate system. Invariance of the laws of physics with respect to such a rotation implies the conservation of angular momentum.

rotation about an axis. The physical laws governing any process are formulated with a particular origin and a particular coordinate system in mind; for instance, laws of terrestrial gravity might use the centre of the Earth as their origin, while the laws of planetary motion might use the centre of the Sun.

However, the physical laws should remain the same whatever the choice and so any mathematical expression of the laws should remain the same under any of these transformations. Application of Noether's theorem, then, reveals the conserved quantity corresponding to each particular invariance. Invariance under a translation in time (i.e. that the laws of physics this year are the same as last year) implies that conservation of energy is built into the laws describing the process. Invariance under a translation in space (e.g. that physics is the same in London as in New York) implies conservation of momentum. And invariance under spatial rotations implies conservation of angular momentum; see Figure 6.1.

The laws of physics are also invariant under the Lorentz transformations of special relativity (Figure 2.3 in Section 2.5). More generally, physical laws are unchanged under any combination of a Lorentz transformation and a space-time translation. These are called Poincaré transformations after the French mathematician Henri Poincaré. Invariance under the complete group of Poincaré transformations incorporates all of the above space-time symmetries.

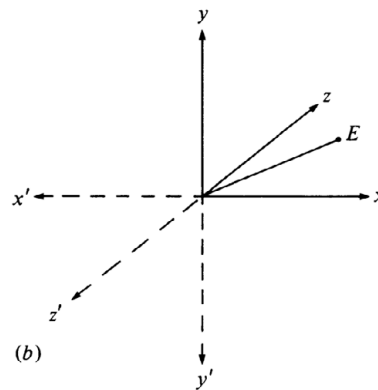
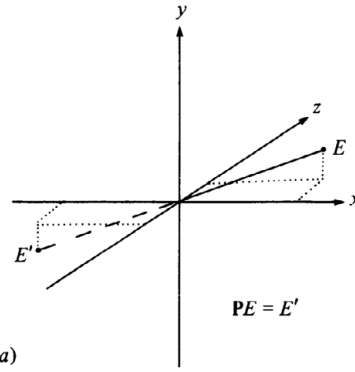


Figure 6.2. A parity transformation reverses the spatial coordinates of an event  $E$  (a) or, equivalently, converts a left-handed coordinate system into a right-handed one (b).

### 6.3 Discrete Symmetries

The continuous space-time symmetries are called *proper* Lorentz transformations because they can be built up from a succession of infinitesimally small ones. However, there are also *improper* symmetries which cannot be so built up. These improper, or discrete, symmetries do not have corresponding conservation laws as important as those of the continuous symmetries. However, they have proved very useful in telling us which particle reactions are possible with a given force and which are not. We shall deal with the three most important improper symmetries in turn.

#### 6.3.1 Parity or Space Inversion

In this operation, denoted  $\mathbf{P}$ , the system under consideration (e.g. a particle wavefunction) is reflected through the origin of the coordinate system, as in Figure 6.2(a). An alternative way of thinking of this is as the reversal of a left-handed coordinate system

into a right-handed one, as shown in Figure 6.2(b). The parity operation is equivalent to a mirror reflection followed by a rotation through  $180^\circ$ .

If a system (a particle, or collection of particles) is described by a wavefunction  $\psi(\mathbf{x}, t)$ , then the parity operation will reverse the sign of the *spatial* coordinates:

$$\mathbf{P}\psi(\mathbf{x}, t) = \psi(-\mathbf{x}, t).$$

If the system is to remain invariant under the parity operation, then the observable quantity which must not change is the probability density, which is just the absolute square of the wavefunction:

$$|\psi(\mathbf{x}, t)|^2 = |\psi(-\mathbf{x}, t)|^2.$$

So, if  $\psi$  describes a state of definite parity, such that

$$\mathbf{P}\psi(\mathbf{x}, t) = \lambda\psi(\mathbf{x}, t) = \psi(-\mathbf{x}, t),$$

then the absolute square of  $\lambda$  should be one. Furthermore, acting twice with  $\mathbf{P}$  (two inversions) gets us back to where we started, so that

$$\mathbf{P}^2\psi(\mathbf{x}, t) = \lambda^2\psi(\mathbf{x}, t) = \psi(\mathbf{x}, t).$$

Therefore

$$\lambda = \pm 1.$$

So if the system is to remain invariant under the parity operation, the system's wavefunction may either remain unchanged,  $\mathbf{P}\psi = +\psi$ , in which case we say the system is in an *even* parity state; or the system's wavefunction may change sign,  $\mathbf{P}\psi = -\psi$ , in which case we say the system is in an *odd* parity state.

If the forces governing the system respect parity, an even parity state cannot change into an odd one or vice versa. This helps us define the ways in which a system may evolve.

An example of this is the way in which light is emitted from atoms. Each state an electron can occupy has a definitive parity assignment, even or odd, which is determined by the magnitude of the orbital angular momentum of the electron about the nucleus and by the orientation of electron spin. As the electromagnetic force respects parity, and the photon has odd intrinsic parity (see below), transitions can only occur between atomic states of opposite parity. This limits the transitions possible and so prescribes the energies of the photons emitted. By observing the spectral lines emitted from atoms we can thus check the conservation of parity.

It is also necessary to consider the *intrinsic parity* of a single particle for which the operation of space inversion is not so obvious. This is illustrated by the decay of one particle into two. The final state of two particles with some well-defined motion with respect to one another can be examined under parity transformations and either even or odd parity assigned. If then the interaction responsible for causing the decay conserves parity, the initial one-particle state must also be a state of well-defined parity. Thus a particle can be assigned some intrinsic parity, even (+1) or odd (-1), which is multiplied together with the spatial parity to obtain the overall parity of the state.

Intrinsic parity has meaning only because particles can be created or destroyed. If the particles in a system were always the same, then the product of their intrinsic parities in any initial or final states would always be the same and so would be a meaningless quantity. In this hypothetical immutable world we should be free to assign any particle either even or odd parity with no reason. In the real world this arbitrariness allows us to define the intrinsic parity of certain particles (normally the nucleons are given even parity) and then the intrinsic parities of all other particles are established from experiment.

### 6.3.2 Charge Conjugation Symmetry

Another useful symmetry in particle physics is that of the interchange of particles with their antiparticles, denoted  $\mathbf{C}$ . This symmetry means that if physical laws predict the behaviour of a set of particles, then they will predict exactly the same behaviour for the corresponding set of antiparticles. For example, a collision between an electron and a proton will look precisely the same as a collision between a positron and an antiproton (see Figure 6.3).

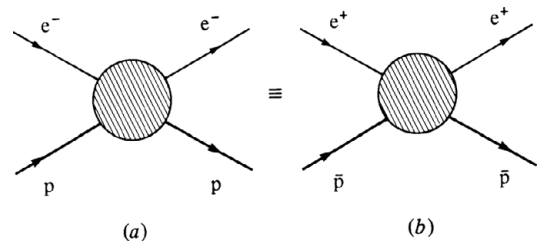


Figure 6.3. Symmetry under the charge conjugation operation implies the equivalence of (a) particle and (b) antiparticle reactions.

The symmetry applies also to the antiparticles of particles with no electric charge, such as the neutron. The interaction of a proton and a neutron is the same as that of an antiproton and an antineutron.

As with the parity operation, the wavefunction of a system may be even or odd under the action of the charge conjugation operation:

$$C\psi = \pm\psi.$$

An example of the use of this symmetry is provided by particle decay into photons by the electromagnetic force. A single photon is odd under  $C$  symmetry.

Observing the decay of a particle into two photons, then, determines the particle to be even under charge conjugation symmetry, as this is given simply by the product of the two photons' symmetries  $(-1)^2$ . We then know that the particle cannot decay into an odd-charge conjugate state, such as three photons, if  $C$  symmetry is to be preserved.

### 6.3.3 Time Reversal

The last of the three discrete symmetries, denoted  $T$ , connects a process with that obtained by running backwards in time. Despite the rather intriguing name, the operation refers simply to that process obtained by reversing the directions of motion within the system. Symmetry under 'time reversal' implies that if any system can evolve from a given initial state to some final state, then it is possible to start from the final state and re-enter the initial state by reversing the directions of motion of all the components of the system.

## 6.4 The CPT Theorem

It is possible also to define product symmetries which can be obtained by operating two or more of these discrete symmetries simultaneously. For instance, a system of particles in a coordinate system can be subject to the operations of parity and charge conjugation simultaneously to reveal a system of antiparticles in the reverse-handed coordinate system. If the laws governing the system are invariant under  $CP$  operation, then the two systems will behave in exactly the same way. Also, it is possible to assign an even or odd symmetry under the combined  $CP$  symmetry to any state and so require that the system evolves to a state of the same symmetry.

There are no utterly fundamental reasons for supposing that the individual symmetries should be

preserved by the various forces of nature (that there should be symmetry between a process and its mirror image, between a process and its anti-process and so on). But it seems a reasonable assumption and was taken for granted for many years. In fact, as we will soon see, the symmetries are *not* exact and there do exist phenomena which display slight asymmetries between process and mirror process, and process and anti-process (see Chapters 11, 14).

But there are very good reasons for supposing that the combined CPT symmetry is absolutely exact. Then, for any process, its mirror image, antiparticle and time-reversed process will look exactly as the original. This is the so-called CPT theorem, which can be derived from only the most fundamental of assumptions, such as the causality of physical events (cause must precede effect), the locality of interactions (instantaneous action at a distance is not possible) and the connection between the spin of particles and the statistics governing their collective behaviour.

The consequences of the CPT theorem are that particles and their antiparticles should have exactly the same masses and lifetimes, and this has always been observed to be the case. Another consequence is that if any one individual (or pair) of the symmetries is broken, as mentioned, there must be a compensating asymmetry in the remaining operation(s) to cancel it and so ensure exact symmetry under CPT.

## 6.5 Dynamical Symmetries

The symmetries of space and time give rise to universal conservation laws such as those of energy, momentum and angular momentum. As these laws must be respected by all processes, the Lagrangian of any system must be invariant under the groups of transformations through time, space and angular rotations, respectively.

But other conservation laws are also known to exist, such as the conservation of electrical charge. This can be represented by requiring the Lagrangian to remain invariant under arbitrary shifts in the phases of the charged particle wavefunctions appearing in the Lagrangian (see Figure 6.4).

We will learn that there are many other quantities which are conserved in interactions arising from the various forces of nature. This implies that the Lagrangians describing these interactions must be invariant under appropriate symmetry operations. We will see that demanding such invariances gives rise to

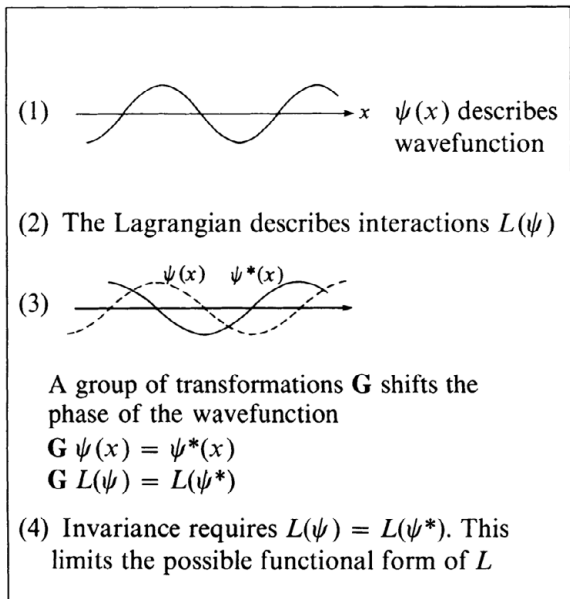


Figure 6.4. A symbolic representation of the action of a dynamical symmetry.

physically significant predictions such as the existence of new particles and values for their electric charges, spins and other quantum numbers yet to be introduced.

## 6.6 Internal Symmetries

The symmetry operations we have introduced so far are the fundamental ways of describing the conservation laws we observe to obtain in particle interactions. But symmetry can also help us categorise particles according to their intrinsic properties.

In addition to the familiar particles carrying only electrical charge, we will soon meet particles with wholly new quantum numbers such as 'strangeness', 'charm' and so on. The values of these quantum numbers carried by the particles allow them to be

classified into fixed patterns or multiplets; we shall see this at work later.

Suffice at this stage to say that, in the microworld, symmetry does fulfil its traditional role of arranging disparate elements into regular patterns (just like the periodic table of elements).

## 6.7 Broken Symmetries

Symmetries are sufficiently valuable that even broken ones can be useful. For many purposes a broken mirror is as good as a whole one! We have mentioned already how the individual reflection symmetries  $\mathbf{P}$ ,  $\mathbf{C}$  and  $\mathbf{T}$  might be broken in some classes of reaction (which turn out to be those governed by the weak nuclear force). But for other forces which do respect them, they are still a valuable guide for indicating which reactions are possible and which are forbidden.

Similarly, conservation laws and the internal symmetries on which they are based may not be exact. The first successful internal symmetry scheme for classifying the reactions of the strongly interacting particles was known from the start to be badly broken but, nevertheless, it provided a valuable ordering effect on the variety of reactions observed.

Of particular interest is the case when the Lagrangian governing the dynamics due to some force or forces is not quite invariant under some group of transformations, but only under a restricted group or when additional particles have been introduced. This indicates that the relatively more complicated forces arising from the imperfect or restricted symmetries have their origin in a truly general symmetry (and its simpler forces) which may have obtained under different circumstances. This is the gist of the modern approach to the unified theory of the forces of nature in which approximate symmetries are a guide to the nature of forces in unfamiliar circumstances (e.g. just after the Big Bang).



# 7

## *Mesons*

### 7.1 Introduction

Modern particle physics can be thought of as starting with the advent of mesons. For these are not constituents of everyday matter, as are the protons and the electrons, but were first proposed to provide a description of nuclear forces. The subsequent discoveries of a bewildering variety of mesons heralded an unexpected richness in the structure of matter, which took many decades to understand.

### 7.2 Yukawa's Proposal

In attempting to describe the features of the strong nuclear force, physicists in the 1930s had to satisfy two basic requirements. Firstly, as the force acts in the same way on both protons and neutrons, it must be independent of electric charges and, secondly, as the force is felt only within the atomic nucleus, it must be of very short range. In 1935 the Japanese physicist H. Yukawa suggested that the nuclear force between protons is mediated by a massive particle, now called the pi-meson or pion, denoted by  $\pi$ , in contrast to the massless photon which mediates the infinite-range electromagnetic force. It is the mass of the mediating particle which ensures that the force it carries extends over only a finite range. This is indicated by Heisenberg's uncertainty principle which allows the violation of energy conservation for a brief period. If the proton emits a pion of finite mass, then energy conservation is violated by an amount equal to this mass energy. The time for which this situation

can obtain places an upper limit on the distance which the pion can travel, and this distance is a guide to the maximum effective range of the force.

From the  $\alpha$ -particle scattering experiments, we know that the effective range of the strong force is about  $10^{-15}$  m, which gives a pion mass of about 300 times that of the electron, or about 150 MeV. To account for all the possible interactions between nucleons, the pions must come in three charge states. For instance, the proton may transform into a neutron by the emission of a positively charged pion or, equivalently, by the absorption of a negatively charged pion. But the proton may also remain unchanged during a nuclear reaction, which can be explained only by the existence of an uncharged pion. So the pion must exist in three charge states: positive, neutral and negative ( $\pi^+$ ,  $\pi^0$ ,  $\pi^-$ ).

### 7.3 The Muon

In 1937, five years after his discovery of the positron, Anderson observed in his cloud chamber yet another new particle originating from cosmic rays. The particle was found to exist in both positive- and negative-charge states with a mass some 200 times that of the electron, about 106 MeV. At first, the particle was thought to be Yukawa's pion and only gradually was this proved not to be the case. Most importantly, the new 'mesons' seemed very reluctant to interact with atomic nuclei, as indicated by the fact that they are able to penetrate the Earth's atmosphere to reach

the cloud chamber at ground level. For particles which were expected to be carrying the strong nuclear force such behaviour was unlikely. Also, there was no sign of the neutral meson. Theorists eventually accepted that this new particle was not the pion; instead it was named the *muon*, and is denoted by  $\mu$ .

The muon was a baffling discovery as it seemed to have no purpose in the scheme of things. It behaves exactly like a heavy electron and it decays into an electron in  $2 \times 10^{-6}$  s; and so is not found in ordinary matter. Although we shall see later how the muon can fit into a second generation of heavy elementary particles, the reason for this repetition is still by no means obvious. So, the muon is not a meson at all, but a *lepton* like the electron.

#### 7.4 The Real Pion

If Yukawa's pion is to interact strongly with atomic nuclei, it is unreasonable to expect it to penetrate the entire atmosphere without being absorbed. So experiments at ground level are unlikely to detect it. In 1947 C. Powell, C. Lattes and G. Occhialini from Bristol University took photographic plates to a mountain top to reduce the decay distance which pions created at the top of the atmosphere had to traverse before being detected. They found Yukawa's pion, which quickly decays into a muon, which itself then decays (Figure 7.1). The mass of these charged pions  $\pi^\pm$  was determined to be 273 times the mass of the electron (140 MeV), very close to Yukawa's original estimate. Since the initial discovery of the charged pions, it has been established that decay into a muon and a neutrino is their main decay mode with a lifetime of about  $2.6 \times 10^{-8}$  s. Other decay modes do exist but are thousands of times less likely.

The uncharged pion  $\pi^0$  was eventually discovered in accelerator experiments in 1950. The delay was due to the fact that uncharged particles leave no obvious trace in most particle detectors and so cannot be observed directly. The most likely decay mode of the  $\pi^0$  is into two photons which also leave no tracks. Only by observing the electron–positron pairs created by the photons can the existence of the  $\pi^0$  be inferred (Figure 7.2). The mass of the  $\pi^0$  was found to be slightly less than that of its charged counterparts at 264 times the mass of the electron, but its lifetime is much shorter at  $0.8 \times 10^{-16}$  s. The reason for this large difference in lifetimes is that the  $\pi^0$  decays by the action of the electromagnetic force, as indicated by

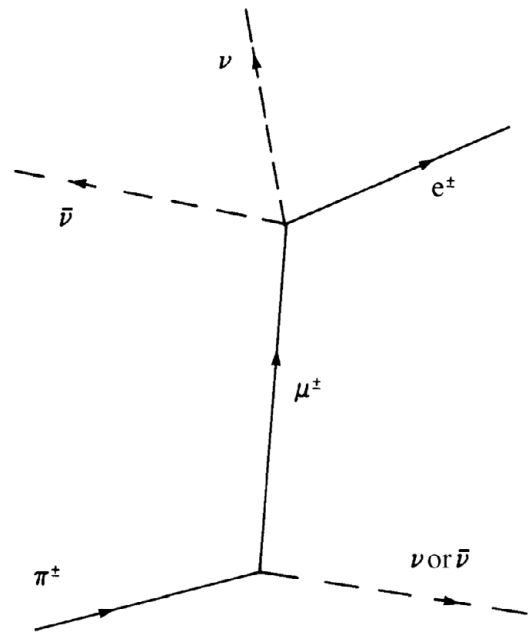


Figure 7.1. The pion decays to a muon, which then decays to an electron. Neutrinos are emitted to ensure conservation of energy and angular momentum.

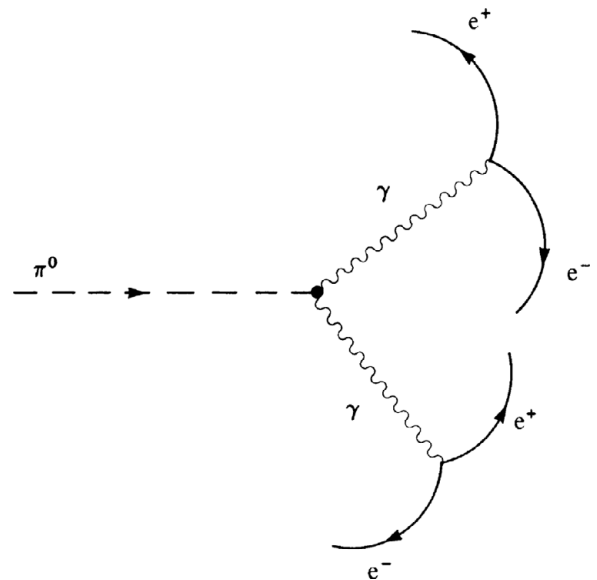


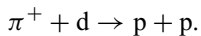
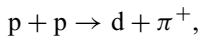
Figure 7.2. The decay of the neutral pion.

the presence of the two photons, whereas the charged pions decay by the weak force, as indicated by the presence of the neutrino.

Spin assignment	Particle	Orientation of components of spin in space in the presence of a magnetic field
$s = \frac{1}{2}$	$e^-$	
Isospin assignment	Particle	Orientation of components of isospin in charge space in the presence of electromagnetism
$I = \frac{1}{2}$	N	
$I = 1$	$\pi$	

Figure 7.3. The analogy between particle spin in real space and particle isospin in abstract charge space.

In 1953 the pions were established as spin 0 by comparing the relative magnitudes of the cross-sections of the reactions:



The relative magnitude of the two can depend only on the spins of the particles in the collision, and

knowing the spins of the protons (p) and the deuteron (d) determines that of the pion ( $\pi^+$ ). Such reactions also establish the intrinsic parity of the pion (relative to the nucleons). This is found to be odd (-1).

### 7.5 Terminology

At this point it is worth both introducing some of the generic names which are used for these particles

and defining their essential features. A few of the most often used are:

- nucleons: neutrons and protons;
- hadrons: all particles affected by the strong nuclear force;
- baryons: hadrons which are fermions (half-integral spin particles) such as the nucleons;
- mesons: hadrons which are bosons (integral spin particles) such as the pion;
- leptons: all particles not affected by the strong nuclear force, such as the electron and the muon.

Particles which are baryons are assigned a baryon number  $B$  which takes the value  $B = 1$  for the nucleons,  $B = -1$  for the antinucleons and  $B = 0$  for all mesons and leptons. In all particle reactions, the total baryon number is found to be conserved (i.e. the total of the baryon numbers of all the ingoing particles must equal that of all outgoing particles). Similarly, particles which are leptons are assigned a lepton number which is also conserved in particle reactions. This is explained further in Chapter 15.

## 7.6 Isotopic Spin

We have met, so far, two sorts of particles which differ only slightly in their masses but which have different electric charges: the nucleon (the proton and the neutron) and the pions. The strong nuclear force seems to ignore totally the effects of electric charge and influences all nucleons in the same way and all pions in the same way. As far as the strong force is concerned, there is only one nucleon and only one pion. In 1932 Heisenberg described this mathematically by introducing the concept of *isotopic*

*spin* or *isospin*. This concept is the prototype of both the elementary particle classification schemes and the modern dynamical theories of the fundamental forces, so it merits some attention.

Recall that the two different orientations in (real) space of the ‘third components’ of the spin of the electron (see Chapter 3) provide two distinct states in which the electron can exist (in the presence of a magnetic field). Analogously, Heisenberg proposed that different orientations in an abstract charge space of the third components of an imaginary isotopic spin would be a mathematically convenient way of representing the charge states within a family of particles (in the presence of electromagnetism), see Figure 7.3. Similarly, just as the different components of electron spin are separated in energy by a magnetic field (causing the fine structure in spectral lines), so too the different components of isotopic spin in a particle family are separated in mass by the effects of the electromagnetic force (causing the slight mass differences between the proton and the neutron, and between charged and neutral pions).

The electric charges of the hadrons,  $Q$ , are related to their isospin assignments by the simple formula,

$$Q = e \left( I_3 + \frac{B}{2} \right).$$

So for the pions which have zero baryon number ( $B = 0$ ), the charges are simply the units of the electronic charge corresponding to the three ‘third’ components of spin  $I_3(1, 0, -1)$ . For the nucleon which has unit baryon number ( $B = 1$ ), the two isospin states with third component  $+\frac{1}{2}$  and  $-\frac{1}{2}$  become the positive and neutral charge states respectively.

# 8

## *Strange Particles*

### 8.1 Introduction

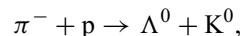
In 1947 the British physicists G. D. Rochester and C. C. Butler observed more new particles, about a thousand times more massive than the electron, in cloud chamber photographs of cosmic rays. As these particles were often associated with V-shaped tracks, they were at first called V particles (see Figure 8.1). Their origin and purpose were an entire mystery. If we remember that this same year saw the discovery of the real pion and the subsequent redundancy of the muon, it is fair to think of it as the beginning of the baroque era of particle physics, in which an increasing number of particles were discovered, seemingly with no other purpose than to decorate cloud chambers. For the following six years, the V particles were observed in cosmic ray experiments and two kinds became apparent. There are those whose decay products always include a proton and are called *hyperons*, and there are those whose decay products consist only of mesons and are called *K mesons*, or *kaons*.

The hyperons and kaons soon became known as the *strange* particles because of their anomalous behaviour. They were observed frequently enough to indicate production by the strong nuclear force, say, between two protons, or a pion and a proton, and so we would expect a decay time typical of a strong nuclear process (i.e. about  $10^{-23}$  s). But, from the length of their tracks in the photographs, it was possible to estimate their average lifetimes at about  $10^{-10}$  s, the timescale typical of weak interaction processes.

This behaviour seemed to contradict the microscopic reversibility of reactions and required explanation.

### 8.2 Associated Production

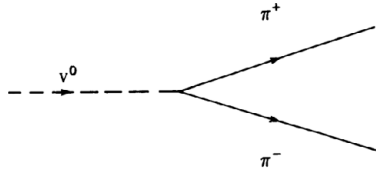
The first step in search of this explanation was provided in 1952 by the American physicist A. Pais. He suggested that the strange particles could not be produced singly by the strong interaction, but only in pairs. This was confirmed in experiments at the Brookhaven accelerator in 1953, when strange particles were man-made for the first time. The strange particles always emerged in pairs in reactions such as



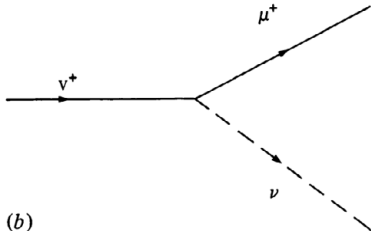
where  $\Lambda^0$  denotes the hyperon and  $K^0$  denotes a neutral kaon.

In the same year Gell-Mann and Nishijima explained this mechanism of associated production by proposing the introduction of a new conservation law, that of *strangeness*, which applies only to the strong interaction. Each particle is assigned a quantum number of strangeness, in addition to its quantum numbers of spin, intrinsic parity and isospin. Then, in any strong interaction, the total strangeness of all the particles before and after the reaction must be the same.

Associated production can now be explained by assigning a positive strangeness to one of the strange particles produced and a negative strangeness to the



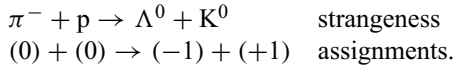
(a)



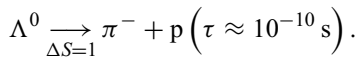
(b)

Figure 8.1. (a) A neutral  $V^0$  particle decays into pions. (b) A charged  $V^+$  decays into a muon and a neutrino. These  $V$  particles are now called *kaons*, and denoted  $K^0$  and  $K^+$ .

other, so that the total strangeness of the final state is zero, the same as that of the non-strange initial state:



The decay of strange particles into non-strange particles cannot proceed by the strong interaction, which must, by definition, conserve strangeness. Instead, such decays proceed by the weak interaction, which need not, and which allows the strange particles a comparatively long life:



The strangeness of the strongly interacting hadrons is defined by

$$Q = e \left( I_3 + \frac{B + S}{2} \right).$$

When  $S = 0$ , we recover the equation of Section 7.6 which relates the charge to the third component of isospin for pions and nucleons and other non-strange hadrons.

### 8.3 The Kaons

There are two charged strange mesons  $K^+$  and  $K^-$  which each have a mass of 494 MeV, and a neutral

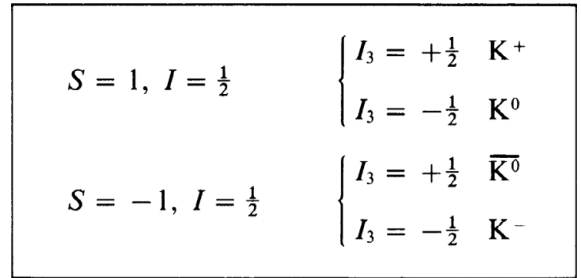


Figure 8.2. Isospin doublets of the K mesons.

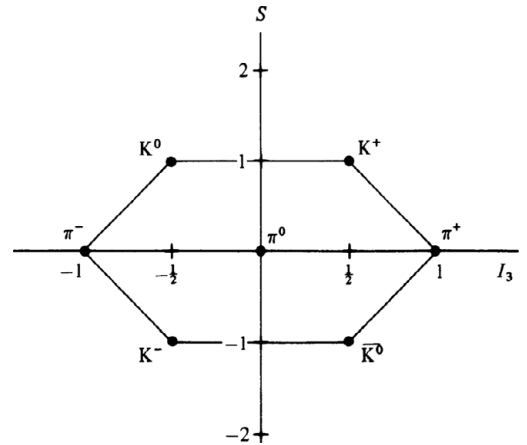


Figure 8.3. A multiplet of pions and K mesons arranged according to value of strangeness and third component of isospin.

one  $K^0$  of mass 498 MeV. This makes the K mesons about three times more massive than the pions. But, like the pions, the kaons were found to be spin 0 and to have odd intrinsic parity. They are thus in some sense close relations of the pions. However, they have a very different multiplet structure. Let us recall that the three charge states of the pion ( $\pi^+$ ,  $\pi^0$ ,  $\pi^-$ ) are the different  $I_3$  states of the same  $I = 1$  pion, and that the uncharged pion is its own antiparticle. This is not the case with the kaons because of complications due to the strangeness quantum number. If we assign to the neutral kaon  $K^0$  a value of strangeness  $S = 1$ , then from the formula in Section 8.2, the value of total isospin  $I = 1$  is ruled out and the kaons cannot form any isospin triplet like the pions. Instead, the kaons are grouped into isospin doublets as shown in Figure 8.2. From this we can see that the uncharged kaon must come in two versions with opposite strangeness if the scheme is to work. So although the  $K^-$  is the

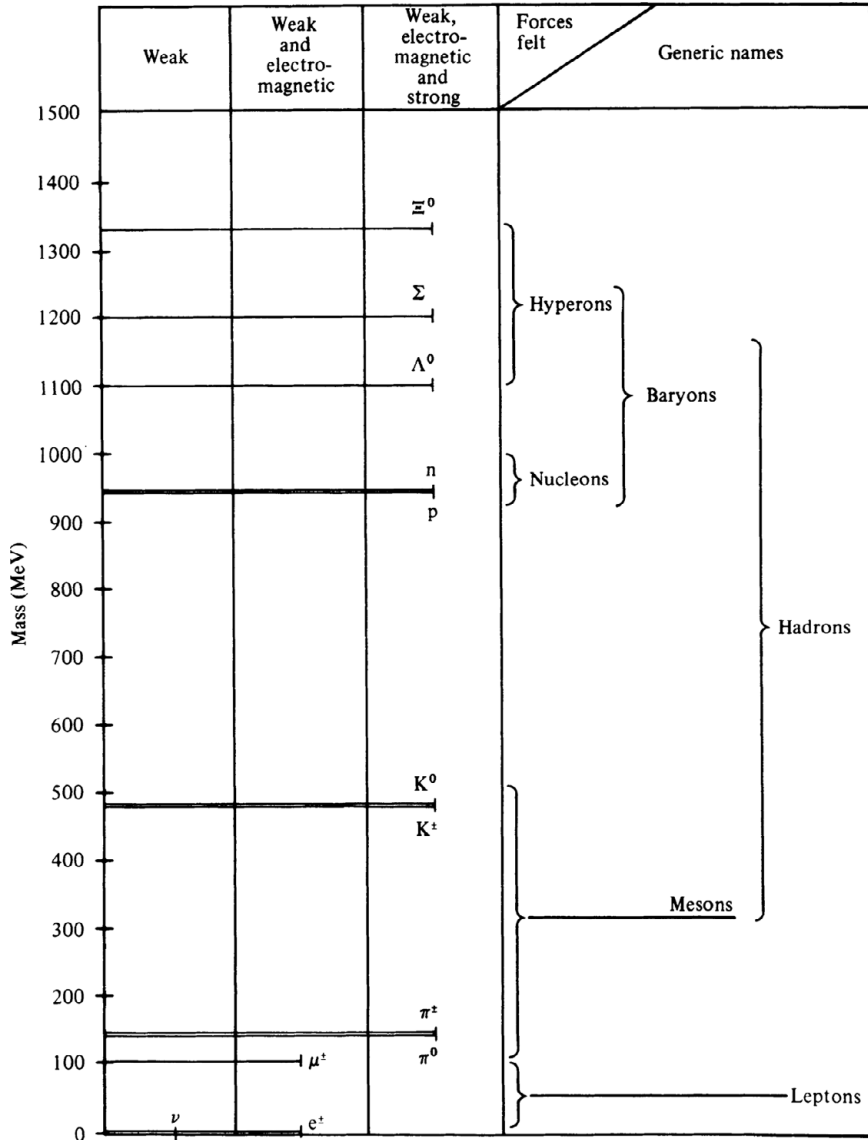


Figure 8.4. The basic set of elementary particles known by the early 1950s.

antiparticle of the  $K^+$ , the  $K^0$  is not its own antiparticle, which must have different strangeness:

$$\bar{K}^+ \equiv K^- \quad \text{but} \quad \bar{K}^0 \neq K^0.$$

Because the  $K^0$  is different from the  $\bar{K}^0$  only by the value of its strangeness, it might somehow be able to exhibit effects directly attributable to strangeness. After all, so far we have merely categorised observed particle decay patterns by awarding the particles different values of a hypothetical quantum number. If

we could observe some experimental effect due to strangeness, then we might be more convinced of its physical reality. This thought occurred at the time to Fermi, who challenged Gell-Mann to prove the worth of his strangeness by demonstrating some difference between the  $K^0$  and the  $\bar{K}^0$ . This led to some very important work, as we shall see in Chapter 14.

We can very neatly summarise our knowledge of the mesons discussed so far by plotting their assignments of isospin and strangeness (Figure 8.3). These

graphs are known as multiplets for particles of the same spin and intrinsic parity and we shall see how they form the basis of the elementary particle classification scheme in Chapter 10.

#### 8.4 The Hyperons

The hyperons are the strange particles which eventually decay into a proton and which, like the proton, have spin  $\frac{1}{2}$  and are baryons with baryon number 1. The lambda hyperon  $\Lambda^0$  is the least massive at 1115 MeV and has isospin zero (it exists only as a neutral particle). The sigma hyperon  $\Sigma$  has a mass of 1190 MeV and has isospin 1 and so exists in three different charge states ( $\Sigma^+$ ,  $\Sigma^0$ ,  $\Sigma^-$ ). Finally, the xi hyperon  $\Xi$ , known also as the cascade particle, has mass 1320 MeV and isospin  $\frac{1}{2}$  and has strangeness  $-2$ . To decay into non-strange particles, it therefore needs to undergo two weak interactions, as the weak force can only change strangeness by one unit at a time.

For the hyperons, we often prefer to use *hypercharge*  $Y$  as the distinguishing quantum number, which is the sum of baryon number and strangeness:

$$Y = B + S.$$

Those  $\Lambda$ ,  $\Sigma$  and  $\Xi$  hyperons of spin  $\frac{1}{2}$  which have been mentioned form only the basic set of those which

exist. There are very many more massive hyperons which have spins  $\frac{3}{2}$ ,  $\frac{5}{2}$  or even  $\frac{7}{2}$ . These resonances are short-lived and generally decay quickly into one of the hyperons in the basic set by the strong interaction (conserving strangeness, or hypercharge) before these eventually decay by the weak interaction back into non-strange baryons.

#### 8.5 Summary

In Figure 8.4 all the particles we have mentioned so far are plotted according to their masses and are categorised according to their generic names. The origin of the names is clear from the diagram: the leptons are the lightweights, the mesons are the middleweights and the baryons are the heavyweights. We also show the applicability of the fundamental forces to the various categories of particles. We may think it more than just coincidence that the strongly-interacting hadrons are the most massive category if we believe that the mass of the particles somehow arises from the interactions they experience.

We now know that the mass alone is not a reliable way to categorise the species. Recent experiments have found both leptons and mesons more massive than the baryons. Nowadays the classifications are taken to refer to the interactions experienced by the various classes, which is taken to be a more fundamental attribute than mass.





**Part III**  
**Strong Interaction Physics**



# 9

## *Resonance Particles*

### 9.1 Introduction

Most of the particles which we have discussed up to this point have lifetimes sufficient for them to leave observable tracks in bubble chambers or other detectors, say greater than about  $10^{-12}$  s. But there is no reason for us to demand that anything we call a particle should necessarily have this property. It may be, for instance, that some particles exist only for a much shorter time before decaying into others. In this case we should not expect to detect them directly, but to have to infer their existence from the indirect evidence of their decay products. These transient particles are called *resonance particles* and many have been discovered with widely varying properties. It was the attempts to categorise the large number of resonances which first led to an appreciation of the need for a more fundamental pattern of order, which in turn led to the idea of quarks.

### 9.2 Resonance Particle Experiments

Resonance particles can be produced in two different types of experiment: resonance formation and resonance production experiments. In the formation experiments, two colliding particles come together to form a single resonance which acts as an intermediate state between the original colliding particles and the final outgoing products of the collision. The presence of the resonance is indicated when the cross-section for the collision (i.e. the effective target area of the colliding particles) peaks dramatically over a small

range of collision energy centred on the mass of the resonance, see Figure 9.1. The value of the energy range corresponding to one-half of the height of the resonance peak is referred to as the *width* of the resonance and this is a measure of the uncertainty in the mass of the particle.

Only if a particle is perfectly stable can it be thought of as having a uniquely defined mass; for an unstable particle there will always be uncertainty in the value of its mass, given by Heisenberg's uncertainty principle:

$$\Delta E \Delta t > \hbar.$$

From this we can see that the narrower the width  $\Delta E$  of the resonance, the larger will be the uncertainty in the lifetime  $\Delta t$ , thereby implying a longer-lived particle. Conversely, if the resonance is broad, this implies a short lifetime. Typical widths for hadronic resonances, such as the  $N^*$  resonances in pion-proton scattering, are a few hundred MeV, which correspond to lifetimes of about  $10^{-23}$  s. This makes them the most transient phenomena studied in the natural world.

In resonance production experiments, the presence of a resonance is inferred when it is found that the outgoing particles prefer to emerge with a particular value of combined mass. Finding the resonances in this fashion is more difficult because it is first necessary to look at all the possible combinations of outgoing particles which might have arisen from the resonance, and then to plot the combined masses of

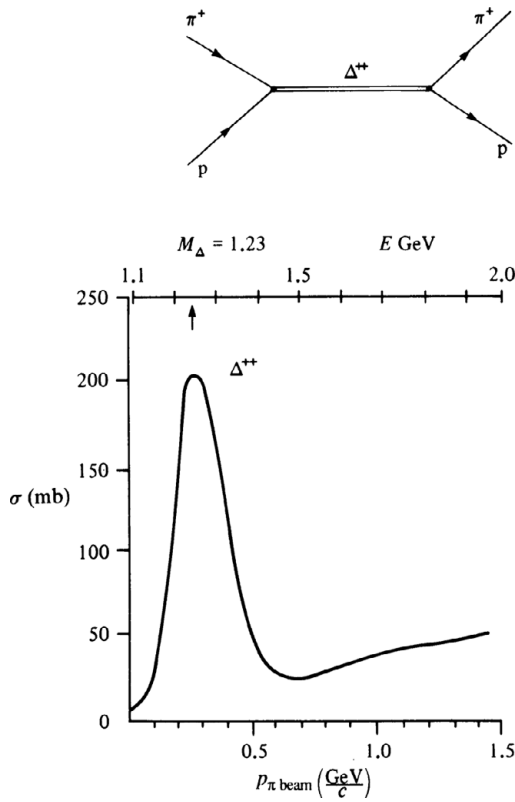


Figure 9.1. An example of resonance formation. A large increase in the pion–proton cross-section,  $\sigma$ , plotted against the pion beam momentum  $p_{\pi\text{beam}}$  signals the formation of a resonance particle.

the combinations to see if any preferred values exist, see Figure 9.2.

One advantage of the production method is that it does not require us to study only the resonances which can be formed by the ingoing particles. In high-energy collisions, any number of new and interesting particles may emerge and it is possible to see if they have originated from some previously unknown resonance. In this fashion we can study the resonances made from  $\pi$ s and Ks only, even though we cannot arrange collisions between only  $\pi$ s and Ks as the ingoing particles. These methods have allowed physicists to build up a picture of literally hundreds of resonances, all of which may be legitimately regarded as just as elementary as the neutron or the pion.

Over the years, interesting regularities in the resonance spectrum became apparent. Often a particular

Table 9.1. Two mass series of meson resonances.

Meson symbol	$I$	$S$	Mass (MeV)	Spin	Decay	Force acting
$\pi$	1	0	140	0	$\mu\nu$	weak
$\rho$	1	0	768	1	$\pi\pi$	strong
$a_2$	1	0	1320	2	$\rho\pi$	strong
$P_3$	1	0	1690	3	$4\pi$	strong
K	$\frac{1}{2}$	1	494	0	$\mu\nu$	weak
$K^*$	$\frac{1}{2}$	1	892	1	$K\pi$	strong
$K_2^*$	$\frac{1}{2}$	1	1425	2	$K\pi$	strong

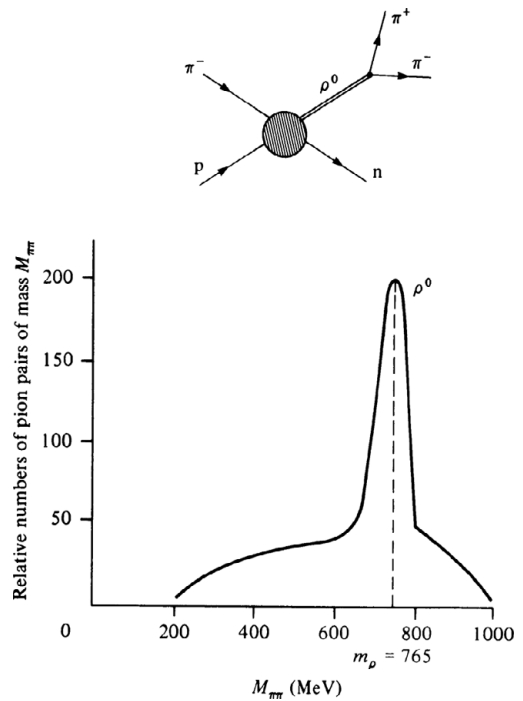


Figure 9.2. An example of resonance production.

set of quantum numbers for isospin and strangeness, such as for the pion ( $I = 1, S = 0$ ) and the kaons ( $I = \frac{1}{2}, S = 1$ ) are duplicated by particles with higher masses and spins. These higher-mass versions of the quantum numbers generally decay very quickly back down to the least-massive particle with those particular numbers, by the strong interaction. This least-massive version can then itself decay more slowly by the weak force, violating quantum-number conservation as it does so, see Table 9.1.

# 10

## *SU(3) and Quarks*

### 10.1 Introduction

By the early 1960s it became clear that many hundreds of so-called ‘elementary’ resonance particles exist, each with well-defined values of the various quantum numbers such as spin, isospin, strangeness and baryon number and with widths which are generally seen to increase (or lifetimes which are generally seen to decrease) as their masses become larger. At that time, the most urgent task for physicists was to discover the correct classification scheme for the particles, which would do for the elementary particles what Mendeleeff’s periodic table had done for the variety of chemical elements known in the nineteenth century. A closely related problem was whether or not it was sensible to regard such a plethora of different particles as truly elementary. Most of the resonance particles are very massive compared with, say, the electron and occupy a finite region of space with a radius of about  $10^{-15}$  m, while many have high values of spin and the internal quantum numbers. All these factors argue in favour of the existence of more fundamental constituents combining in a variety of ways to make up the known hadrons, just as a few fundamental atomic constituents (electrons, protons and neutrons) can combine to make up the variety of elements. But, historically, it was not possible to pass directly to the analysis of these fundamental constituents, which at the time were extremely speculative. Initially, it was necessary to classify the bewildering variety of hadrons according to some symmetry scheme from

which clues to the nature of the constituents could be derived.

### 10.2 Internal Symmetry

Such a classification scheme is provided by an internal symmetry group, as described in Chapter 6. The scheme was proposed independently by Murray Gell-Mann and Yuval Ne’eman in 1961. The starting point of this symmetry group is the charge independence of the strong nuclear forces, as expressed in Chapter 7 by the concept of isospin. By regarding the neutron and the proton as the isospin-down and isospin-up components of a single nucleon, the strong interaction’s indifference to ‘neutron-ness’ and ‘proton-ness’ can be expressed as the invariance of strong interactions to rotations in the abstract isospin space. The group of transformations which achieves these rotations is the Special Unitary group of dimension 2 called  $SU(2)$ , which acts on the 2-dimensional space defined by the proton and the neutron, redefining the proton and neutron as a mixture of the original two:

$$\mathbf{G}^{SU(2)} \begin{pmatrix} p \\ n \end{pmatrix} \rightarrow \begin{pmatrix} p^* \\ n^* \end{pmatrix}.$$

Of course, the same must also be true of the pions, which form a 3-dimensional space ( $\pi^+$ ,  $\pi^0$ ,  $\pi^-$ ), and the  $\Delta$  baryons ( $\Delta^{++}$ ,  $\Delta^+$ ,  $\Delta^0$ ,  $\Delta^-$ ), which form a 4-dimensional space. These are referred to as the 2-, 3- or 4-dimensional representations of  $SU(2)$ .

When conservation of strangeness is added to that of isospin as a property of the strong interaction, it is clear that the strongly interacting particles are governed by a bigger symmetry group. Although it seems obvious, it took a great deal of work to show that  $SU(3)$  is the appropriate group. The transformations of the  $SU(3)$  group generate many dimensional representations (multiplets), **1, 3, 6, 8, 10, 27**, etc., each of which is a well-defined quantum-number pattern. It was a triumph for the originators of the scheme to find that some of these exactly fitted the quantum-number structure of the observed hadrons, see Figure 10.1. The identification of the correct symmetry group for the strong interactions, and the assignment of hadrons to the multiplets, led to the prediction in 1962 of a new hadron necessary to complete the spin- $\frac{3}{2}$  baryon decuplet **10**. This is the famous  $\Omega^-$  particle with strangeness assignment  $-3$ . Its spectacular discovery in 1964 in bubble chamber photographs at Brookhaven convinced a previously sceptical world of the validity of  $SU(3)$ .

The correct symmetry group having been found, a major problem remained. It was necessary to explain why the mesons filled some multiplets and the baryons fitted others, but other multiplets had no particles. In particular it seemed odd that the fundamental 3-dimensional representation should remain unfilled (i.e. the most basic representation of the  $SU(3)$  group). In an unsuccessful symmetry scheme prior to that of Gell-Mann and Ne'eman, the proton, neutron and hyperon were assigned to this triplet, but the logical consequences of such an assignment were incompatible with the experimental evidence.

### 10.3 Quarks

In 1964, Gell-Mann and George Zweig pointed out that the representations of  $SU(3)$  which were occupied by particles could be chosen from amongst all those mathematically possible by assuming them to be generated by just two combinations of the fundamental representation. Gell-Mann called the entities in the fundamental representation *quarks* (a word abstracted from the novel *Finnegan's Wake* by James Joyce). The three varieties of quark, or flavours as they are now called, have since come to be known as the up, down and strange quarks, the up and down labels referring to the orientation of the quarks' isospin. The combinations of quarks which give the occupied

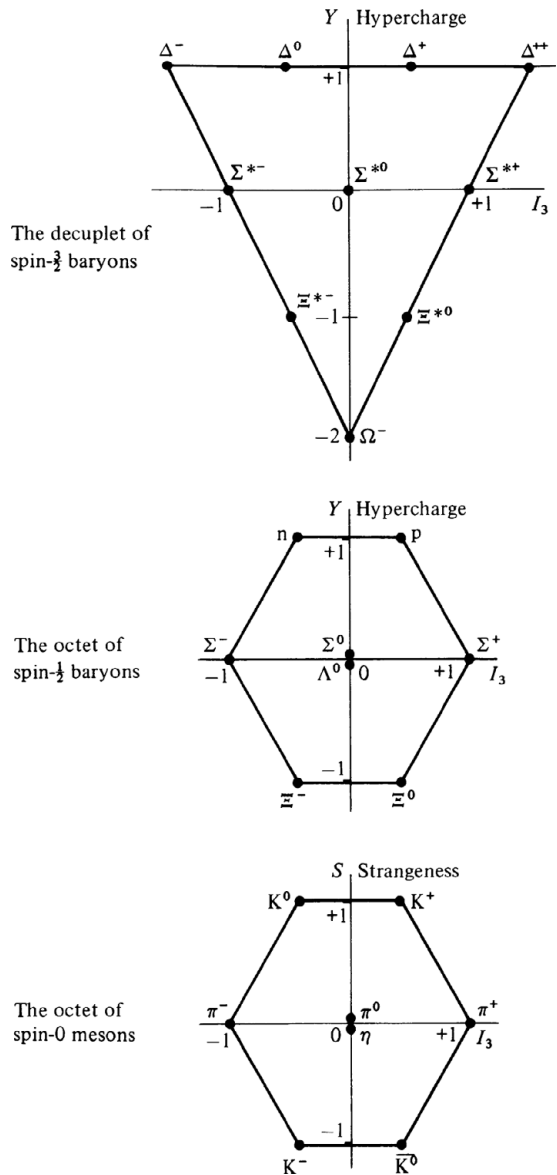


Figure 10.1.  $SU(3)$  representations provide the quantum-number patterns for the elementary particles.

representations of  $SU(3)$  are a quark-antiquark pair for the meson multiplets and three quarks for the baryon multiplets. This is expressed mathematically by combining the representations of the group:

$$\begin{aligned} \mathbf{q} \otimes \mathbf{q} \otimes \mathbf{q} &\equiv \mathbf{3} \otimes \mathbf{3} \otimes \mathbf{3} \rightarrow \mathbf{1} \otimes \mathbf{8} \otimes \mathbf{8} \otimes \mathbf{10} \\ \mathbf{q} \otimes \bar{\mathbf{q}} &\equiv \mathbf{3} \otimes \mathbf{3}^* \rightarrow \mathbf{1} \otimes \mathbf{8}. \end{aligned}$$

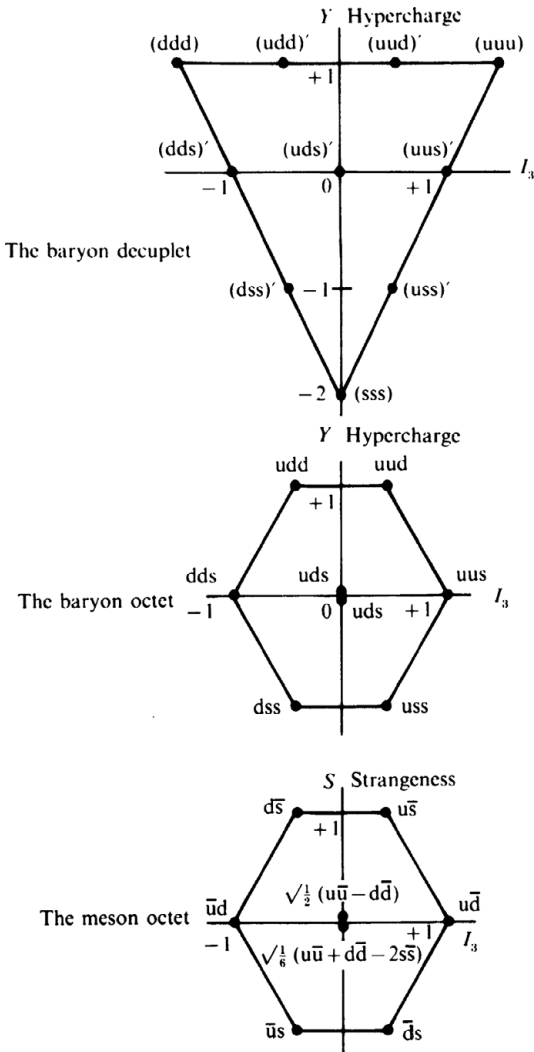


Figure 10.2. The quark content of the  $SU(3)$  representations.  $(qqq)'$  signifies the summation over the cyclic permutations of the quarks.

The quark constituents of the baryon decuplet and of the baryon and meson octets are illustrated in Figure 10.2.

One significant consequence of this scheme is that if three quarks are to make up each baryon with a baryon number 1, then the quarks themselves must have baryon number  $\frac{1}{3}$ . From the formula relating charge to isospin and baryon number, this means that they must also have fractional electronic charge. Also, to ensure that the baryons generated are fermions and that the mesons are bosons, it is necessary to assign the

Table 10.1. *The quantum number assignments of the early quarks.*

Quark	q	Spin	Charge	$I$	$I_3$	$S$	$B$
Up	u	$\frac{1}{2}$	$+\frac{2}{3}$	$\frac{1}{2}$	$+\frac{1}{2}$	0	$\frac{1}{3}$
Down	d	$\frac{1}{2}$	$-\frac{1}{3}$	$\frac{1}{2}$	$-\frac{1}{2}$	0	$\frac{1}{3}$
Strange	s	$\frac{1}{2}$	$-\frac{1}{3}$	0	0	-1	$\frac{1}{3}$

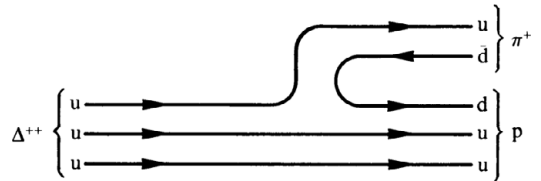


Figure 10.3. Quark line diagram for the decay  $\Delta^{++} \rightarrow p\pi^+$ . Forward-directed arrows indicate quarks and backward-directed arrows antiquarks.

quarks spin  $\frac{1}{2}$ . A summary of their properties is shown in Table 10.1.

Quarks are also useful in providing a qualitative understanding of various hadronic processes through what are called *quark line diagrams*. As an example, consider the decay of the  $\Delta^{++}$  (uuu) into a proton (uud) and a pion (ud) shown in Figure 10.3. Forward-directed arrows indicate quarks while arrows directed backwards in time indicate antiquarks. Note that these are not the same as Feynman diagrams because the quarks are confined inside hadrons and the strong interactions between them are generally not shown. Because of conservation of baryon number, quark lines cannot be broken.

The quarks were referred to earlier as entities rather than particles for good reason. It is not necessary to assume their existence as observable particles to enjoy the successes of the  $SU(3)$  flavour scheme. They may be thought of as the mathematical elements only for such a scheme, devoid of physical reality. This was a fortunate escape clause at the beginning of the quarks' career because their fractional electronic charges and the failure to observe them in experiments encouraged scepticism in the naturally conservative world of physics. As we shall see, indirect evidence for the physical reality of quarks is now very convincing – despite the fact that they have never been seen directly in isolation. But this evidence has mounted



only rather slowly since 1968 with the beginning of the ‘deep inelastic’ experiments at the Stanford Linear Accelerator Center (SLAC) in California. Prior to this, most physicists preferred to reserve judgement on the reality of quarks, content to rely on the mathematics of  $SU(3)$  only.

Because of this historical background of doubt, conclusions which rely on the mathematics of group theory are put together as the  $SU(3)$  scheme of the elementary particles, and conclusions which rely on the physical reality of quarks are referred to as the ‘quark model’. The mathematics of  $SU(3)$ , in addition to generating the multiplet structure of the observed particles, can also provide simple predictions of relationships between the masses of the particles in an  $SU(3)$  multiplet. If the  $SU(3)$  symmetry were perfect, then all the particles in the same  $SU(3)$  multiplet would have to have the same mass. This is obviously not true and so we know that  $SU(3)$  cannot be a perfect symmetry: it is broken. But by making assumptions about just how the symmetry fails, it is possible to derive mass formulae which seem to hold good:

$$\begin{aligned}\frac{1}{2}(m_N + m_\Xi) &= \frac{1}{4}(3m_{\Lambda^0} + m_\Sigma) && \text{baryons,} \\ m_k^2 &= \frac{1}{4}(3m_n^2 + m_\pi^2) && \text{mesons.}\end{aligned}$$

In the quark model, the effects of symmetry breakdown can be described by saying that the strange quark has a larger mass than either of the equal mass up and down quarks. Also, in the quark model, it is possible to assume the existence of forces holding the quarks together and then to generate the spectrum of elementary particle masses for particles of the same quantum number and different spins. The predictions are found to match the experimentally measured masses really rather well, better than the approximations of the model would seem to justify in fact. But the simple quark model is unable to explain the outstanding problems surrounding the quarks. Why are they not seen? Why do they seem to form only in certain combinations? What is the nature of the forces which they experience? These questions, as we will see, had to await the advent of a theory of quarks on a par with the QED theory of electrons.

## **Part IV**

### **Weak Interaction Physics I**



# 11

## *The Violation of Parity*

### 11.1 Introduction

The decay of the strange kaons led to a great deal of confusion in the early 1950s. Two decay modes in particular seemed so different that they were at one time thought to originate from two different parent particles, called the  $\tau$  and  $\theta$  mesons:

$$\begin{aligned}\tau^+ &\rightarrow \pi^+ + \pi^+ + \pi^-, \\ \theta^+ &\rightarrow \pi^+ + \pi^0.\end{aligned}$$

However, detailed study of the two- and three-pion final states indicated that the  $\tau$  and  $\theta$  were indeed both manifestations of the same charged kaon,  $K^+$ . In both cases the mass was the same, and so too was the lifetime – about  $10^{-8}$  s, a timescale which indicates it is the weak force that is responsible for the decays. The decays were thought to be incompatible because the parities of the two final states are different. If they originate from the same initial particle, they imply that parity is not conserved by the force responsible for the decays. This means that the force behaves differently in left-handed and right-handed coordinate systems: it can distinguish left from right, or image from mirror image.

Such a revolutionary conclusion was not seriously entertained until 1956, when T. D. Lee and C. N. Yang pointed out that, although evidence existed for the conservation of parity by the strong and electromagnetic forces, there was no evidence for its conservation by the weak force. Certainly, the  $\tau$ – $\theta$  puzzle

indicated that the weak force did not conserve parity, and Lee and Yang proposed that this was a general feature of all weak interactions.

### 11.2 $\beta$ Decay of Cobalt

Within months of Lee and Yang's suggestion, experiments were performed to test for parity violation in other weak processes. The first and most famous was conducted on the  $\beta$  decay of cobalt by C. S. Wu and E. Ambler at the National Bureau of Standards in Washington. The point of the experiment was to observe some spatial asymmetry in the emission of  $\beta$ -decay electrons from the cobalt, which could lead to a distinction between  $\beta$  decay and its mirror-image process. The process in question was the ordinary radioactive  $\beta$  decay of cobalt into nickel:



Firstly, it was necessary to establish some direction in space of which the cobalt was aware, and with respect to which the emission of  $\beta$ -decay electrons could be measured. This was done by putting a magnetic field across a specimen of cobalt, cooled to a very low temperature. In this situation, the spin of the nuclei align predominantly along the direction of the magnetic field. By measuring the emission of the  $\beta$ -decay electrons along or against the orientation of nuclear spin (the orientation of the magnetic field), any asymmetry can be detected. It is possible

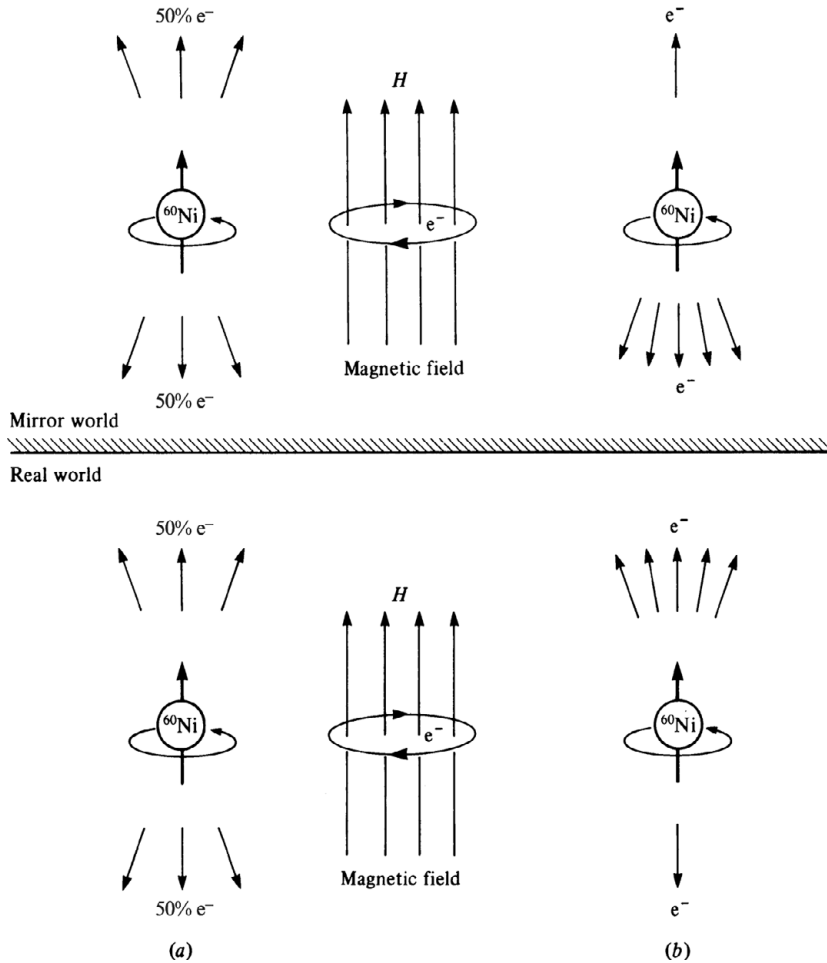


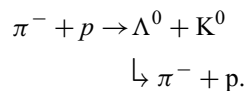
Figure 11.1. (a) If no asymmetry were detected in the emission of decay electrons, the real world and the mirror world would be indistinguishable. (b) If asymmetry were detected, this would result in a distinction between the two. This latter case is observed in experiments.

to show that the direction of spin will not change under mirror reflection, nor will the direction of the magnetic field. But the direction in which the  $\beta$ -decay electrons are emitted will change under mirror reflection and so any asymmetry of electron emission measured with respect to the magnetic field direction will appear to be reversed in the mirror-image process (see Figure 11.1). Hence the process and its mirror image are distinguishable, and the weak force responsible for nuclear  $\beta$  decay can tell its right hand from its left.

The world of physics could scarcely have been more surprised when Wu and her colleagues duly observed the asymmetry which Lee and Yang's work

had implied. Not since the discovery of the quantum nature of light had nature seemed so contrary to common perception. The shock is reputed to have led one eminent physicist to accuse God of being 'weakly left-handed'.

Other experiments soon confirmed this parity-violating effect. One example is provided by the decay of the hyperon in the process.



It is possible to define a plane formed by the tracks of the incoming  $\pi^-$  and the outgoing hyperon.

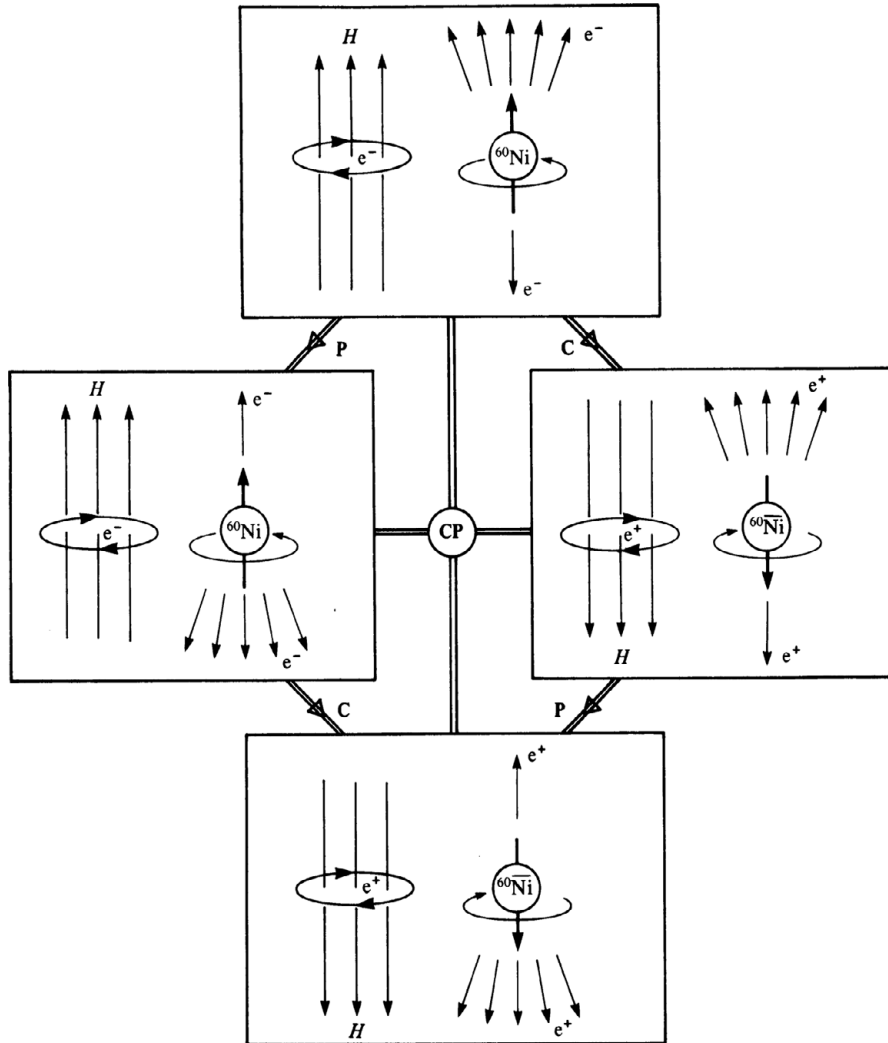


Figure 11.2. How the basic  $^{60}\text{Co}$  experiment transforms under repeated **C** and **P** transformations.

If parity were conserved, equal numbers of outgoing  $\pi^-$  would emerge on either side of this plane. In 1957 an experiment yet again detected an asymmetry indicating parity violation.

### 11.3 Absolute-handedness and CP Invariance

It is an interesting, if rather academic, question of philosophy to ask whether or not it is possible to distinguish absolutely between left and right, using the parity-violating effects of the weak force. The famous thought-experiment of such a distinction is to attempt to communicate our convention for left and right (or

clockwise and anticlockwise) to an intelligent alien in a distant galaxy.

We might think of achieving this by instructing the alien to perform the  $^{60}\text{Co}$  experiment and telling him that our definition of anticlockwise is the direction required of the electron loop that provides the magnetic field, when viewed from the direction towards which most of the  $\beta$ -decay electrons are emitted. This would certainly suffice for aliens in our galaxy, but would not necessarily for aliens in more distant parts of the Universe. This is due to the possibility that distant aliens may

be made of antimatter and may be conducting an anticobalt-60 experiment in which precisely the same procedures would lead to the opposite of our intended conclusions. The reason for this is that although the weak force violates parity, it also violates the symmetry of charge conjugation (matter–antimatter interchange) in such a way that the product symmetry of the two, denoted **CP**, is almost exactly conserved.

Starting from our original experiment in which the majority of  $\beta$ -decay electrons emerge in the direction of the field, we can see that the operation of space inversion will lead to an observable difference: the electrons are emitted against the direction of the field. However, if we then imagine the *additional* operation of charge conjugation we are led to an exact copy of the original process: the particles are emitted

along the direction of the field (see Figure 11.2). So a real-matter alien looking at our original experiment from the direction in which most decay products are emitted will see a clockwise current loop, but the antimatter alien will see an anticlockwise current loop. Thus an alien observing that the emitted particles come out preferentially in the direction of the field would know that *either* his conventions about left–right and particles–antiparticles were the same as ours *or* that they were both different.

Of course, we have not been able to test the validity of the **CP** conservation using anticobalt, although other experiments have been conducted to show that it is preserved to a high degree. But this is not the end of this particular story, as we will see in Chapter 14.

# 12

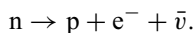
## *Fermi's Theory of the Weak Interactions*

### 12.1 Introduction

Prior to the early 1960s, just three different leptons were recognised: the electron, the muon and the neutrino (together with their antiparticles). The best place to study the weak interaction is in processes involving these leptons only. This ensures that there are no unwanted strong interaction effects spoiling the picture. Unfortunately, early opportunities to study purely leptonic reactions were limited, being restricted to the muon decay into an electron and neutrino. The most common weak interactions available for study are the radioactive  $\beta$  decay of nuclei and the decay of the pions and kaons (which are described generically as the weak decay of hadrons), and it was predominantly these reactions which formed the basis for the first description of the weak interactions formulated by Fermi in 1933.

### 12.2 Fermi's Theory of $\beta$ Decay

The simplest manifestation of  $\beta$  decay is the decay of a free neutron into a proton, an electron and an antineutrino (see Figure 12.1):



Fermi took this to be the prototype for the weak interactions, which he thus described as four fermions reacting at a single point. He expressed this mathematically by saying that, at a single point in space-time, the quantum-mechanical wavefunction of the neutron is transformed into that of the proton and that

the wavefunction of the incoming neutrino (equivalent to the outgoing antineutrino we actually see) is transformed into that of the electron. So a description of this reaction is provided by multiplying the wavefunctions by unknown factors  $\Gamma$  which effect the transformations, and by another factor  $G_F$  called the Fermi coupling constant. This is the quantity which governs the intrinsic strength of the weak interactions, and so the rate of the decay. Thus the amplitude for  $\beta$  decay is given by

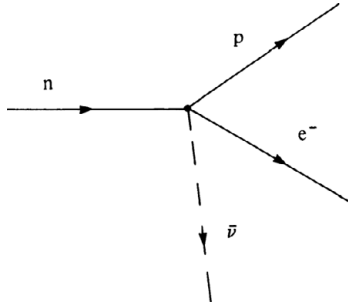
$$M = G_F (\bar{\psi}_p \Gamma \psi_n) (\bar{\psi}_e \Gamma \psi_\nu).$$

The factors  $\Gamma$  contain the essence of the weak interaction effects which give rise to the transformation of the particles. The challenge was to discover the nature of these quantities (whether they are just numbers (scalars) or vectors, tensors, etc.). By examining the angles of emission between the outgoing products of  $\beta$  decay and their various energies, it is possible to narrow down the choice. This took many years: their nature was not confirmed until the parity-violating effect of the weak force was known.

In 1956 Feynman and Gell-Mann proposed that the interaction factors  $\Gamma$  be a mixture of vector and axial-vector quantities, to account for the parity-violating effects of the interactions.

A vector quantity has well-defined properties under a Lorentz transformation. For instance, it will change sign if rotated through  $180^\circ$  and will appear identical after rotation through  $360^\circ$ . An axial-vector



Figure 12.1. The  $\beta$  decay of a free neutron.

quantity will transform just like a vector under rotations, but will transform with the opposite sign to a vector under improper transformations such as parity. Thus, if the interaction comprises vector and axial-vector components, it will look different after a parity transformation (the components might add together instead of cancelling), which is just what we need to describe the weak interactions. By inserting this form of interaction factor  $\Gamma$  into the amplitude  $M$  for  $\beta$  decay, it is possible to calculate the features of particle emission in free neutron decay.

### 12.3 Spin, Helicity and Chirality

In order to understand the weak interaction in greater depth, we need to first delve further into the properties of relativistic fermions. In Section 4.2, we learnt that relativistic fermions are described by two-component spinors (with another two-component spinor for the antiparticle). In the Newtonian limit, when fermions move slowly, these two components can be interpreted as the two spin states of the fermion: the fermion can either be spin-up or spin-down. However, when the fermions are moving close to the speed of light, the notion of spin is no longer so useful and we need a new way in which to classify the two fermion states. It turns out that there are two useful ways to do this. The first, which is closely related to spin, is to define the *helicity* as the component of the fermion's spin in the direction of motion of the fermion. The spin can either be aligned with or against the momentum, and the fermion is referred to as being in the helicity-plus or helicity-minus state respectively.

By measuring the spin and momentum of a particle, the helicity can be measured directly in experiments. However, the helicity is not invariant under proper Lorentz transformations as described

in Chapter 6. For example, consider an observer whose speed as measured by a second observer is greater than that of the fermion under observation. The two observers will disagree on the direction of the momentum of the fermion, but not on the spin, and therefore the helicities they measure will be opposite. Because of this, it is useful to make a second classification of the two possible states of a relativistic fermion, called the *chirality* or *handedness*. The chirality is defined in such a way that it is invariant under proper Lorentz transformations. A particle can either be left-handed or right-handed.

Both helicity and chirality states have the important property that they are interchanged under the parity operation, so that left-handed becomes right-handed and helicity-plus becomes helicity-minus. This suggests that they may be important for the weak interaction, which, as we saw in the last chapter, violates parity. We will see in fact that the weak interaction couples uniquely to left-handed fermions (and right-handed antifermions).

For massless fermions, the definitions of helicity and chirality coincide. A massless left-handed particle and a massless helicity-minus particle are one and the same thing.

### 12.4 The Polarisation of $\beta$ -decay Electrons

We have already seen how parity violation manifests itself in  $\beta$  decay as an asymmetry in the direction of emission of electrons. But it also affects the helicities of the emitted electrons. In the absence of parity violation, as many helicity-plus as helicity-minus electrons should be emitted (see Figure 12.2). But because of it, the electrons show a net preference to spin in one of the two possible ways. We define a polarisation  $P$  of the electrons to quantify this preference:

$$P = \frac{N^+ - N^-}{N^+ + N^-},$$

where  $N^+$  ( $N^-$ ) is the number of helicity-plus (helicity-minus) electrons in a measured sample. When  $P = 1$ , all the electrons are helicity-plus and when  $P = -1$ , all the electrons are helicity-minus. Assuming that the final-state proton does not recoil and that all electrons produced are left-handed, the polarisation can be calculated to be equal to minus the ratio of the electron's speed to that of light:

$$P = -\frac{v_e}{c}.$$

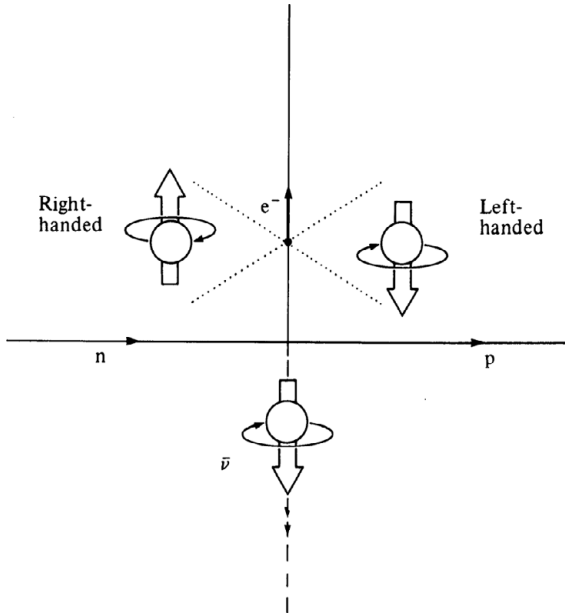


Figure 12.2. The emission of helicity-plus and helicity-minus electrons in  $\beta$  decay. Left-handed electrons are found to predominate experimentally.

So when the electrons are emitted slowly,  $v_e \approx 0$  and there is no net polarisation. But when the electrons are emitted relativistically,  $v_e \approx c$ , they are nearly all helicity-minus. In 1957, F. Frauenfelder and his colleagues observed the polarisation of the electrons from the  $\beta$  decay of  $^{60}\text{Co}$  by scattering them through a foil of heavy atoms. They found a net polarisation of  $-0.4$  for electrons travelling at  $0.49c$ , which is taken as a satisfactory agreement with the prediction.

### 12.5 Neutrino Helicity

The four-fermion interaction constrains, very tightly, the spins which can couple through the weak interactions. In fact, when we look at the fermion spins which are allowed to couple through the  $\Gamma$  we have specified, we find that only left-handed fermions and right-handed antifermions can take part in the weak interactions. As the neutrinos interact only by the weak interactions, only these two possible cases are ever seen. (Neutrinos and antineutrinos of the opposite chirality are now believed to exist, as we discuss in Chapters 41 and 42. If so, they would not take part in any of the known interactions, except for gravity.)

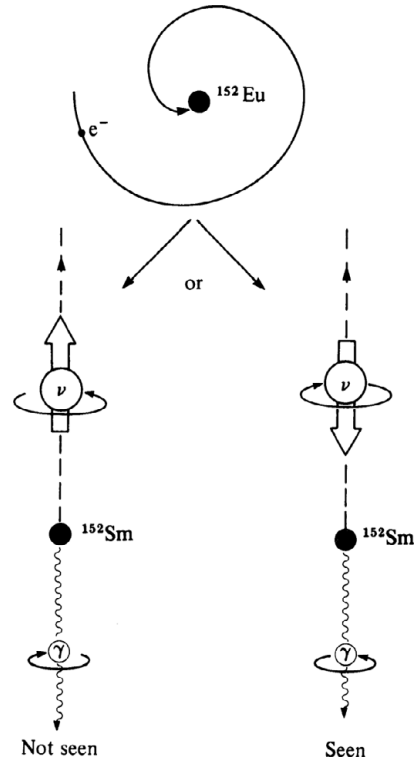


Figure 12.3. The helicity of the photon emitted in the hybrid decay of  $^{152}\text{Eu}$  shows the single-handedness of the neutrino.

To measure the neutrino helicity (which is the same as chirality for massless neutrinos) we must look for a particularly simple example of  $\beta$  decay and deduce it from the helicities of the other decay products, using the principle of angular momentum conservation.

Such an experiment was performed by M. Goldhaber and his colleagues in 1958. The spin-0 nucleus  $^{152}\text{Eu}$  is observed to undergo hybrid  $\beta$  decay in which an electron is captured and a neutrino emitted, leaving an excited state of the nucleus  $^{152}\text{Sm}$  with spin 1. This then decays to its spin-0 ground state by the emission of a photon (Figure 12.3).

The initial and final states of the nuclei are spin 0, so if there is no angular momentum between the neutrino and the photon, then their spins must be opposite. By observing the photon helicity, that of the neutrino can be inferred, and it is found to be negative (corresponding to left-handed chirality).

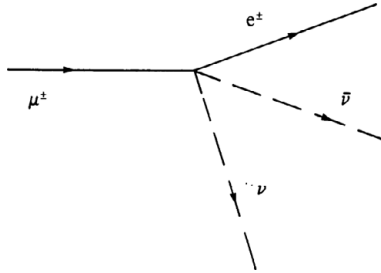


Figure 12.4. The four-fermion picture of the decay of the muon.

**12.6 In Conclusion**

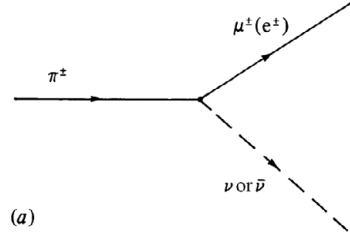
We can say that these experiments conducted in the late 1950s lend support to Fermi’s original idea of a four-fermion interaction acting at a single point, and that the phenomenon of parity violation can be incorporated by choosing the interaction factors  $\Gamma$  such that left-handed neutrinos (right-handed antineutrinos) couple to left-handed electrons.

Other reactions went on to confirm this picture. In particular, the purely leptonic decay of the muon is an obvious candidate for Fermi’s description as a four-fermion interaction (Figure 12.4):

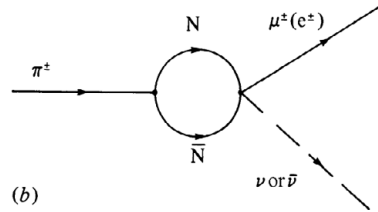
$$\mu^\pm \rightarrow e^\pm + \nu + \bar{\nu}.$$

The amplitude which describes this reaction is the same as that for the  $\beta$  decay of the free neutron, with the appropriate wavefunctions substituted. Thanks to this, it is possible to establish that the value of  $G_F$ , which is necessary to account for the observed rate of muon decay, is equal to within 2% of the value needed to account for neutron  $\beta$  decay. In this way we are sure that it is the same force that is responsible for these two very different processes.

The weak decay of the pion, see Figure 12.5(a), may at first be thought not to fit into the four-fermion description, viz.,



(a)



(b)

Figure 12.5. The weak decay of the (a) charged pion and (b) its four-fermion interpretation.

$$\pi^\pm \rightarrow \mu^\pm + \nu \text{ or } \bar{\nu},$$

or

$$\pi^\pm \rightarrow e^\pm + \nu \text{ or } \bar{\nu}.$$

But these can be accommodated by imagining the pion to dissociate into a virtual nucleon–antinucleon pair by borrowing energy for a time allowed by Heisenberg’s uncertainty principle, so the process becomes (Figure 12.5(b)):

$$\pi^\pm \rightarrow N + \bar{N} \rightarrow \mu^\pm + \nu \text{ or } \bar{\nu}.$$

However, this rather contrived way of describing the decay is avoided in the modern quark picture of the weak decays of hadrons, as we will soon see.

# 13

## *Two Neutrinos*

### 13.1 Introduction

Before nuclear  $\beta$  decay was fully understood, it was not known if the neutrinos emitted in neutron  $\beta$  decay were the same as those emitted in proton  $\beta$  decay or if they were different. (Remember that proton  $\beta$  decay can occur only within the nucleus, as a free proton is stable to all intents and purposes.) As the positron emitted in proton decay is the antiparticle of the electron emitted in neutron decay, it was suggested that a neutrino is emitted from one and an antineutrino from the other. This allows us to formulate a law of lepton-number conservation which was first put forward in 1953 by Konopinski and Mahmoud. If we assign a lepton number  $+1$  to the electron, the negatively charged muon and the neutrino, and a lepton number  $-1$  to the positron, the positively charged muon and the antineutrino, and a lepton number  $0$  to all other particles, then in any reaction the sum of lepton numbers is preserved. These assignments are summarised in Table 13.1. We can check this law in the weak interactions we have met so far:

In neutron  $\beta$  decay

$$\begin{aligned}n &\rightarrow p + e^- + \bar{\nu} \\(0) &\rightarrow (0) + (1) + (-1) \checkmark.\end{aligned}$$

In proton  $\beta$  decay

$$\begin{aligned}p &\rightarrow n + e^+ + \nu \\(0) &\rightarrow (0) + (-1) + (1) \checkmark.\end{aligned}$$

In pion decay

$$\begin{aligned}\pi^\pm &\rightarrow \mu^\pm + \bar{\nu}(\nu) \\(0) &\rightarrow (\mp 1) + (\pm 1) \checkmark.\end{aligned}$$

The assignment of these lepton numbers is shown to have a physical significance by the absence of reactions which do not conserve them, but which otherwise seem feasible. For instance

$$\bar{\nu} + p \rightarrow e^+ + n$$

is observed, whereas the similar reaction

$$\bar{\nu} + n \rightarrow e^- + p$$

is not observed.

### 13.2 A Problem in the Weak Interactions

The Fermi theory of the weak interactions and the lepton-number conservation law with the experiments of  $\beta$ , pion and muon decays formed the content of weak interaction physics until 1960. And although the theory could adequately explain the experimental observations, it could not equally well explain what was not observed. In particular, the decay of a muon into an electron and a photon was not observed despite seeming to be a perfectly valid electromagnetic transition,

$$\mu^- \rightarrow e^- + \gamma.$$

The solution to this impasse is the proposal that any neutrino belongs to either of two distinct species: one

Table 13.1. The assignment of simple lepton number.

Particle	$e^- \mu^- \nu$	$e^+ \mu^+ \bar{\nu}$	Others
Lepton number	1	-1	0

Table 13.2. The assignment of lepton-type number.

Particle	$e^- \nu_e$	$e^+ \bar{\nu}_e$	$\mu^- \nu_\mu$	$\mu^+ \bar{\nu}_\mu$	Others
Electron number	1	-1	0	0	0
Muon number	0	0	1	-1	0

associated with the electron and one associated with the muon, and that electron-type neutrinos  $\nu_e$  can never transform into muons, nor muon-type neutrinos  $\nu_\mu$  into electrons. So the  $\beta$  decay of the neutron involves only electron-type antineutrinos:

$$n \rightarrow p + e^- + \bar{\nu}_e;$$

and the muon decay of the pion involves only muon-type antineutrinos.

The muon can decay into an electron only if a muon-type neutrino carries off the ‘muon-ness’ and an electron-type antineutrino cancels out the ‘electron-ness’ of the electron:

$$\pi^- \rightarrow \mu^- + \bar{\nu}_\mu.$$

The decay of the muon into an electron and a photon is forbidden because the ‘muon-ness’ is apparently transformed into ‘electron-ness’.

All this means that the law of lepton conservation is now extended to that of lepton-type conservation and works in a similar fashion. Both electron-type number and muon-type number must be conserved separately in each reaction. Using the revised lepton-number assignments in Table 13.2, we can see how this works for muon decay:

$$\mu^- \rightarrow e^- + \bar{\nu}_e + \nu_\mu$$

$$\text{Muon number } (1) \rightarrow (0) + (0) + (1) \checkmark$$

$$\text{Electron number } (0) \rightarrow (1) + (-1) + (0) \checkmark.$$

Strictly speaking, each time we have written down the symbol for the neutrino in the preceding chapters, we should have associated with it a suffix denoting electron- or muon-type, a procedure we shall follow from now on.

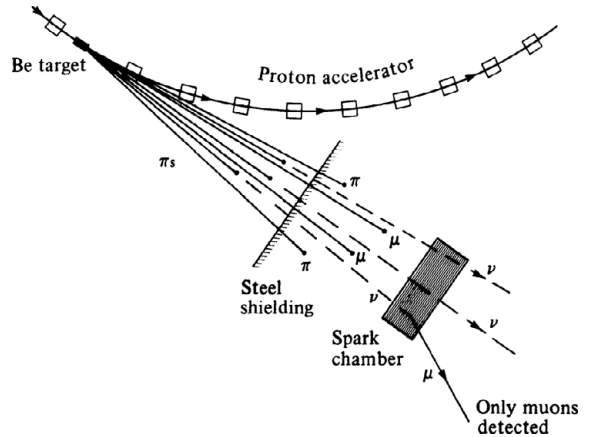


Figure 13.1. Schematic diagram of the two-neutrino experiment.

Once again, the adoption of a new conservation law has side-stepped our difficulties (the last occasion being the introduction of strangeness in Chapter 8). It will not be the last time we adopt such an approach. Our inescapable conclusion from this conservation law is that the electron-type and muon-type neutrinos should be physically different particles, and the first modern neutrino experiment was designed to illustrate just this.

### 13.3 The Two-neutrino Experiment

Neutrino experiments have peculiar difficulties due to the extremely feeble nature of the weak interactions. It is possible to use the low-energy neutrino flux from a nuclear reactor for some experiments, but others require higher-energy neutrinos which have the benefit of interacting more frequently with the targets presented. The first source of high-energy neutrinos became available in the early 1960s with the construction of the first of the big accelerators, the Alternating Gradient Synchrotron in Brookhaven in the US. With this machine, protons could be collided into a solid target such as beryllium to produce a large flux of pions. As we have seen, these pions then decay predominantly into muons and muon-type neutrinos. It is possible to separate out just the neutrinos by passing the beam through vast quantities of iron (at thicknesses of about 20 m) to filter out the muons and any other extraneous particles, see Figure 13.1.

If there is no distinction between electron- and muon-neutrino types, then we would expect the two possible reactions to occur with equal likelihood:

$$\nu_{\mu} + n \rightarrow \mu^{-} + p,$$

$$\nu_{e} + n \rightarrow e^{-} + p.$$

However, if the two types really are distinct we would expect the first reaction to occur to the almost total exclusion of the second, as the neutrino beam consists almost entirely of muon-type neutrinos.

In the first really massive accelerator experiment of modern physics, Leon Lederman, Melvin Schwartz and Jack Steinberger showed that the first (muon) reaction does indeed predominate. In 25 days

of accelerator time, some  $10^{14}$  neutrinos traversed their spark chamber, which produced just 51 reactions resulting in a final-state muon. The ratio of the electrons to muons produced was later measured at CERN to be  $0.017 \pm 0.005$ , so conclusively demonstrating the existence of two separate neutrino types. This experiment demonstrated the validity of lepton-type conservation and explained the observed absence of decays which would otherwise be allowed. Lederman, Schwartz and Steinberger shared the 1988 Nobel Prize for their discovery.

## *Neutral Kaons and CP Violation*

### 14.1 Introduction

Soon after the observation of the weak interaction's violation of parity, it was discovered that it does not preserve charge conjugation symmetry **C** either. This was demonstrated by examining the spins of the electrons and positrons emitted in the decays of positively charged and negatively charged muons respectively. But physicists hoped that these two symmetry violations cancelled each other out exactly, so that the combined **CP** symmetry would be preserved by the weak interactions. To test this it is necessary to define an elementary particle state which is either even or odd under the **CP**-symmetry operation, allow the weak interaction to act, and then check that the final result has the same **CP** symmetry. It is possible to assign even or odd parity to an elementary particle state because the nature of the state remains unchanged under the parity operation, the only effect being the possible change in sign of the state wavefunction for a state of odd parity. Unfortunately, it is not possible to assign a well-defined **CP** symmetry to the  $K^0$ . This is because the operation always transforms it into its antiparticle and so changes the identity of the wavefunction. To compare the nature of the wavefunctions before and after an intended symmetry operation we must at least be sure they represent the same particle. We phrase this technically by saying that the  $K^0$  is not an *eigenstate* of the **CP** operation.

Were **CP** to be a good symmetry (one preserved by the weak interaction), it would follow that

to describe the interaction satisfactorily, we should consider it as acting on states which have a well-defined value of **CP**, i.e. on the eigenstates of **CP**. As  $K^0$  and  $\bar{K}^0$  are not these eigenstates it means that the weak interaction does not really 'see' these particles, but some others instead which are eigenstates. The simplest of these are simple mixtures of the original particles:

$$K_1^0 = \frac{1}{\sqrt{2}}(K^0 + \bar{K}^0),$$

$$K_2^0 = \frac{1}{\sqrt{2}}(K^0 - \bar{K}^0).$$

The wavefunctions of these two states keep their identity under the **CP** operation; the  $K_1^0$  has even **CP** symmetry and the  $K_2^0$  odd **CP** symmetry (i.e. **CP**  $K_1^0 = +K_1^0$  and **CP**  $K_2^0 = -K_2^0$ ). The weak interaction acts on these states, not on the neutral K mesons produced by the strong forces.

### 14.2 What is a Neutral Kaon?

The answer to this question depends on the interaction by which the particle is observed. The kaon which is produced in the strong interaction is either  $K^0$  or  $\bar{K}^0$ , both of which are eigenstates, see Table 14.1, of the parity operation with odd intrinsic parity and which have definite assignment of strangeness. The 'particle' which decays by the weak interaction is  $K_1^0$  or  $K_2^0$ , which are eigenstates of the combined **CP** operation, but which do not have a well-defined strangeness quantum number.

Proof that the  $K_1^0$  and  $K_2^0$  are like real particles to the weak interaction as are  $K^0$  and  $\bar{K}^0$  to the strong can be found from their decays. The  $K_1^0$  (which is even under **CP** symmetry) can decay only to states which are also even, such as a state of two pions. But  $K_2^0$ , which is odd under **CP**, can decay only to **CP** odd states, such as a state of three pions. This gives rise to very different mean lifetimes of the two particles:

$$\begin{aligned} K_1^0 &\rightarrow 2\pi & \tau &= 0.9 \times 10^{-10} \text{ s,} \\ K_2^0 &\rightarrow 3\pi & \tau &= 5.2 \times 10^{-8} \text{ s.} \end{aligned}$$

Another remarkable distinction between  $K_1^0$  and  $K_2^0$  is that they have different masses, although they are both equal mixtures of  $K^0$  and  $\bar{K}^0$  which have identical masses. This seemingly paradoxical conclusion was reached in 1961 when the mass difference was measured experimentally by a method which neatly shows up the identity crisis suffered by neutral K mesons.

When a neutral K meson is first produced by the strong interaction, it is definitely either  $K^0$  or  $\bar{K}^0$ , depending on the strangeness of the reaction, e.g.

$$\pi^- + p \rightarrow \Lambda^0 + K^0. \quad (14.1)$$

Immediately, at the point of creation, the  $K^0$  is an equal mixture of  $K_1^0$  and  $K_2^0$ . However, we know that  $K_1^0$  has a much shorter mean lifetime than  $K_2^0$  and so the longer the time since its creation, the more likely it is to be a  $K_2^0$ . When the time elapsed is much greater than the mean lifetime of the  $K_1^0$  we can say that the kaon is almost entirely  $K_2^0$ . This means that, according to the equation

$$K_2^0 = \frac{1}{\sqrt{2}} (K^0 - \bar{K}^0),$$

what originally started out as a particle (i.e.  $K^0$ ), now has a 50% chance of being its antiparticle (i.e.  $\bar{K}^0$ ). We can see this transformation explicitly by using the kaon produced from the reaction above to undergo the reaction which produces hyperons,

$$\bar{K}^0 + N \rightarrow \Lambda + \pi. \quad (14.2)$$

Because of strangeness conservation, this can occur only with  $\bar{K}^0$  but not with  $K^0$ . So if we take the neutral kaon in (14.1) and wait for the  $K_1^0$  content to drop, we should start to see reaction (14.2) occur when a suitable target is placed in the beam. The frequency

Table 14.1. *The strong and weak forces 'see' different kaon eigenstates*

Interaction	Relevant eigenstates
Strong	$K^0, \bar{K}^0$
Weak:	
(1) were <b>CP</b> conserved	$K_1^0 = \frac{1}{\sqrt{2}} (K^0 + \bar{K}^0)$ $K_2^0 = \frac{1}{\sqrt{2}} (K^0 - \bar{K}^0)$
(2) with <b>CP</b> violated	$K_S^0 = K_1^0 - \varepsilon K_2^0$ $K_L^0 = K_2^0 - \varepsilon K_1^0$

of reaction (14.2) will depend on the fraction of  $\bar{K}^0$  which is generated in the beam by the  $K_1^0$  component decaying. The intensity of the  $\bar{K}^0$  component of the beam can be plotted as a function of time and it turns out that the nature of the variation depends on any mass difference which exists between  $K_1^0$  and  $K_2^0$ . Experiments indicate the existence of a mass difference of about  $3.5 \times 10^{-6}$  eV. Bearing in mind the mass of  $K^0$  of 498 MeV, the mass difference between  $K_1^0$  and  $K_2^0$  is about one part in  $10^{14}$ !

### 14.3 Violation of CP Symmetry

In 1964, Christenson, Cronin, Fitch and Turlay decided to check that the weak interaction did conserve **CP** symmetry exactly, and so to justify the use of states  $K_1^0$  and  $K_2^0$  as the particles appropriate to the weak force. They chose simply to observe a beam of  $K_2^0$  and look for any decay into just two pions. If any were observed, then this would mean that the  $K_2^0$  particle with **CP** = -1 had transformed into the two-pion state with **CP** = +1 and that the weak interaction does not conserve the symmetry. In the experiment the beam was allowed to travel about 18 m to ensure as few  $K_1^0$  present as possible. The products of the particle decays of the  $K_2^0$  beam were then observed as they left their tracks through the particle detectors, which also measured their energies, see Figure 14.1. They observed just a few of the forbidden decays of the  $K_2^0$  beam into pairs of oppositely charged pions: about 50 out of a total of 23 000 decays. This was far higher than any background event rate which may have resulted from the accidental presence of  $K_1^0$  particles still in the beam, and the team concluded



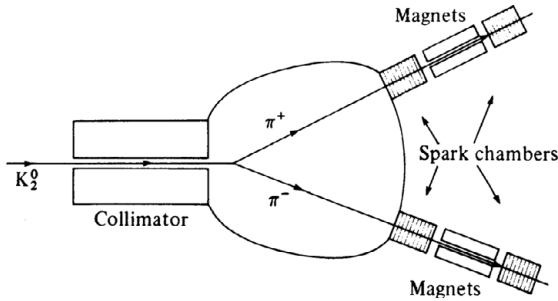


Figure 14.1. A schematic drawing of the **CP**-violation experiment of Christenson *et al.* (1964). Any decay of  $K_2^0$  into just two pions represents the violation of **CP** symmetry.

that the  $K_2^0$  could decay into just two pions and that **CP** symmetry was not preserved exactly by the weak interaction after all.

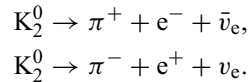
Because of this **CP**-violating effect, it follows that  $K_1^0$  and  $K_2^0$  are not quite the particles ‘seen’ by the weak interaction. Instead, the particles, as ‘seen’ by the weak interaction, are basically the **CP** eigenstates  $K_L^0$  and  $K_S^0$ , but each with a small admixture of the other. These are called the long-lived  $K_L$ , and short-lived  $K_S$ , kaons, respectively, where,

$$K_L^0 = K_2^0 + \varepsilon K_1^0, \quad K_S = K_1^0 - \varepsilon K_2^0,$$

$$\varepsilon \approx 2 \times 10^{-3}.$$

This small admixture of  $K_1^0$  (i.e. the ‘wrong’ **CP** eigenstate) in  $K_L^0$  is due to a transition between  $K^0$  and  $\bar{K}^0$ . It is a higher-order weak interaction process and hence very small.

The violation of **CP** has profound theoretical consequences which are not fully understood even now. We shall return to this intriguing story in Chapter 39. One intellectually satisfying consequence of **CP** symmetry violation is that we can at last convey to our intelligent alien the absolute distinction between left and right. Violation of **CP** symmetry gives rise to an observable difference in the probabilities of occurrence (or branching ratios) of the reactions:



We can now communicate that we define the neutrino by specifying the branching ratio of the reaction in which it is present. This established a common matter–antimatter convention which allows our alien to identify uniquely our handedness convention.

## **Part V**

# **Weak Interaction Physics II**



## *The Current–Current Theory of the Weak Interactions*

### 15.1 Introduction

In Part IV we examined some of the processes of the weak interactions and learnt some of their physical attributes (relatively long lifetimes of weak decays, parity-violating effects, etc.). What we want to do now is to introduce a framework which relates the very disparate phenomena of the weak force, ranging from nuclear  $\beta$  decay and muon decay to high-energy neutrino collisions with matter. Because some of the processes involve hadrons, it will be necessary to ensure that our framework incorporates the consequences of the internal  $SU(3)$  flavour symmetry of the hadrons, and furthermore, it is desirable that it be able to accommodate quarks as the origin of this symmetry.

This framework provides a description of the weak interactions in terms of the interaction of two ‘currents’, specifying the flow of particles. For example, in  $\beta$  decay, one current converts a neutron into a proton, and the other creates an electron and its antineutrino. We begin the construction of this framework by dividing the weak interactions into three classes reflecting the categories described above:

- (1) *leptonic reactions* involving only leptons, such as muon decay,  $\mu^- \rightarrow e^- + \nu_\mu + \bar{\nu}_e$ ;
- (2) *semi-leptonic reactions* involving both leptons and hadrons, such as neutron  $\beta$  decay,  $n \rightarrow p + e^- + \bar{\nu}_e$ ;

- (3) *hadronic weak reactions* involving only hadrons such as the pionic decay of the kaons,  $K_1^0 \rightarrow \pi^+ + \pi^-$ .

### 15.2 The Lepton Current

The ultimate aim is to achieve a common description of all three classes of weak interactions. But we start by concentrating only on one, the leptonic reaction. What we have seen of the weak force provides us with our description. In these reactions we saw that whenever an electron-neutrino is absorbed, an electron is created; and equivalently, whenever an electron-neutrino is created a positron has to be created also. This is as a result of the laws of conservation of lepton number and lepton-type number. This means that in our description, the lepton wavefunctions must always come in pairs. Also, from our knowledge of  $\beta$  decay, we know that these wavefunctions must be coupled together via an interaction factor  $\Gamma$  which combines the spins together in a correct, parity-violating way. We can now write down a ‘lepton current’  $L^W$  which describes the flow of leptons during the weak interaction:

$$L^W = \bar{\psi}_e \Gamma \psi_{\nu_e} + \bar{\psi}_\mu \Gamma \psi_{\nu_\mu},$$

$$\bar{L}^W = \bar{\psi}_{\nu_e} \Gamma \psi_e + \bar{\psi}_{\nu_\mu} \Gamma \psi_\mu.$$

The second line is essentially the antiworld equivalent of the familiar process.

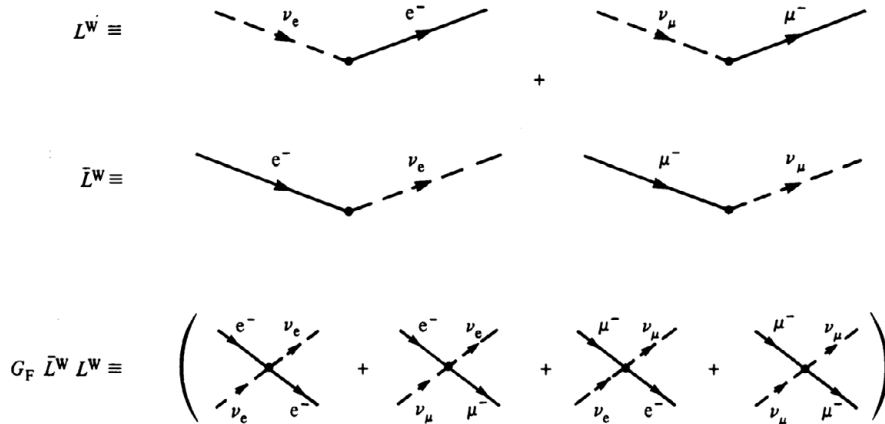


Figure 15.1. The leptonic current  $L^W$  can be multiplied with its antiworld partner  $\bar{L}^W$  to generate all the observed weak interactions of leptons.

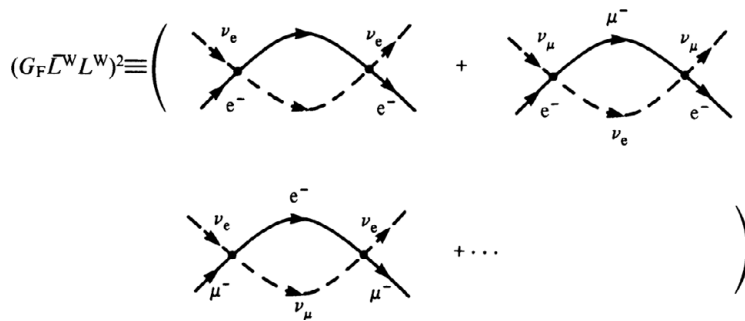


Figure 15.2. Higher-order interactions (repeated weak interactions) can be generated by multiplying the current-current interaction with itself.

We can now generate the first-order amplitudes  $m^{(1)}$  of all the leptonic processes by specifying interactions between leptonic currents. In fact, the simple product of the two lepton currents shown seems to generate all the reactions we see:

$$m^{(1)} \supset G_F \bar{L}^W L^W.$$

We can illustrate the currents and their couplings diagrammatically, as in Figure 15.1. Inspecting the interaction diagrams, we must bear in mind that the destruction of a particle is equivalent to the creation of its antiparticle, so the same diagram can describe, for instance, both electron-muon scattering and muon decay.

The weak leptonic current shown above flows between a charged lepton (e.g.  $e^-$ ) and its neutrino ( $\nu_e$ ). Because these particles have different electric charge, it is called a *charged* current. It was widely believed for many years that all weak interactions

were charged-current processes. However, in 1973 the discovery of a weak interaction *neutral* current (in which particles do not change identity) was made at CERN. We shall return to neutral currents shortly.

### 15.3 Higher-order Interactions

We can use this description to show also the higher-order interactions which can occur between the leptons. These result when the weak force acts on the leptons more than once and are described essentially by the product of two amplitudes at successive instants. For the second-order amplitudes  $m^{(2)}$  this is given by the square of the simple current-current interaction:

$$m^{(2)} \supset \left( G_F \bar{L}^W L^W \right)^2.$$

These are shown diagrammatically in Figure 15.2. In fact, these higher-order interactions are not of

practical importance for the weak interaction. This is because the second-order amplitudes are proportional to  $G_F^2$  and the  $n$ th-order amplitudes are proportional to  $G_F^n$ .

Because the weak interaction is weak, the Fermi coupling constant  $G_F$  is small and the higher powers of  $G_F$  are even smaller.

So to achieve the accurate description of a process, only the first-order term is significant. But the higher-order terms are of theoretical significance. It is desirable in principle that they should be calculable, and it was this motivation which has led us to the most recent theory of the weak force, as we shall soon see in Part VI.

# 16

## *An Example Leptonic Process: Electron-neutrino Scattering*

### 16.1 Introduction

Elastic electron-neutrino–electron scattering provides us with an example of the weak interaction at its simplest. The amplitude for the process is found in the current–current interaction:

$$m^{(1)}(v_e + e^- \rightarrow e^- + v_e) = G_F (\bar{\psi}_e \Gamma \psi_{v_e}) (\bar{\psi}_{v_e} \Gamma \psi_e).$$

By inserting the mathematical expression for the wavefunctions and interaction factors into the amplitude it is possible to calculate the cross-section for the process in the laboratory frame of reference:  $\sigma_{\text{LAB}}$ . The final answer is of a particularly simple form when the incoming neutrino energy  $E_{v_e}$  is large:

$$\sigma_{\text{LAB}}(v_e + e^- \rightarrow e^- + v_e) = \sigma_0 \frac{E_{v_e}}{m_e},$$

where  $\sigma_0$  is a constant factor arising from the calculation with a value of

$$\sigma_0 = 9 \times 10^{-45} \text{ cm}^2.$$

This is the very tiny effective area of interaction between the neutrino and an electron. So it is easy to understand why neutrino interactions are so rare. We can perform a very similar calculation to work out the cross-section for antineutrino–electron scattering. The answer, as we might expect, is very similar at high energies:

$$\sigma_{\text{LAB}}(\bar{v}_e + e^- \rightarrow e^- + \bar{v}_e) = \frac{\sigma_0 E_{v_e}}{3 m_e},$$

where we shall later see how the difference of a factor of 3 between the two results from the different-handedness of neutrinos and antineutrinos.

The fact that these cross-sections rise linearly with the energy of the incoming neutrino presents us with an almost ironic situation. When the neutrino energy is low, say  $E_v \approx 5$  MeV, which corresponds to  $E_v/m_e$  of about 10, then the cross-section remains around  $10^{-43} \text{ cm}^2$ . This is a minute cross-section, even for the microworld. The neutrinos which emerge from nuclear reactors are of about this energy, and so experiments using them to observe these neutrino–electron collisions must use a very high flux of neutrinos and must be prepared to wait a very long time to gather enough observations. However, when the neutrino energy is higher, say around 5 GeV (corresponding to  $E_v/m_e$  of about 10 000), then the cross-section rises to around  $10^{-40} \text{ cm}^2$  and the reactions are, in principle, much more accessible. Unfortunately, the only such high-energy neutrinos so far produced are from the decay of pion beams in the high-energy accelerators. Not only does this mean that the neutrino flux is limited to a rather meagre level, but also that they will nearly all be muon-type neutrinos. We can see from Figure 15.1 that elastic muon-neutrino–electron scattering is not included in the basic current–current interaction, and so such a process, if in fact it does exist, could still not provide data to test our answers. (In fact, examples of this class of process, the so-called ‘neutral current’ reactions, have been discovered and

they necessitate modifications to the current–current interaction; we shall discuss this further in Part VI.)

So we are forced back to waiting for rare events involving neutrinos from nuclear reactors to check our answers for the cross-sections. The results of these experiments can be summarised by saying that they are not inconsistent with the cross-section having the predicted levels.

## 16.2 The Role of the Weak Force in Astrophysics

Electron–neutrino–electron scattering processes may play a role in astrophysics by allowing substantial numbers of  $(\nu_e, \bar{\nu}_e)$  pairs to transfer energy from the interiors of stars to their outer layers. Normally, photons might be thought of as fulfilling this role, but in a stellar environment they are absorbed too quickly to perform an effective transfer. So it is left to the more weakly interacting neutrinos. When a heavy star has finished burning all its hydrogen (i.e. fusing hydrogen nuclei into helium and releasing energy), it moves on to a stage in which helium is burnt, after which it burns carbon and then successively heavier elements, with each stage being hotter than the previous one. At higher temperatures, the neutrinos transfer heat more efficiently throughout the star, leading to, for example, a shorter carbon-burning phase than would otherwise be the case. As a consequence, we see a larger ratio of helium-burning stars to carbon-burning stars than we would were neutrino scattering processes absent.

In fact, this is just the start of a deep and increasing role of the importance of the weak interaction in

astrophysics. Following the very discovery of neutrinos from man-made fission reactors, the US physicist Ray Davies commenced experiments in the 1960s to detect the neutrinos which would be produced by the fusion reactions occurring in the Sun. This was achieved by monitoring a mass of dry-cleaning fluid in the depths of a disused mine, an arrangement necessary to filter out the noise from neutrinos and other particles arriving in the flux of cosmic rays arriving on Earth.

After considerable efforts, a flux of neutrinos was indeed detected but at a rate of one-third what was expected. After considerable debate the reason became clear. As we will see in future chapters there are in fact three types of neutrinos, one associated with each of the electron, the muon and a heavier lepton, the tau, to be introduced in Chapter 36. The reason for the shortfall in the rate of detection was the phenomenon of oscillation between neutrino types, a very significant phenomenon we will discuss further in Chapter 43.

Another astrophysical source of neutrinos is in the giant supernovae explosions observed as massive stars collapse down into either neutron stars or black holes. The first such flux of these neutrinos was observed in 1987 when just 25 neutrinos were observed in a variety of detectors around the globe. Many more such are expected now that the first black hole and neutron star mergers have been observed from 2015 onwards at the LIGO and VIRGO observatories, as we will discuss in Part XIII.



## The Weak Interactions of Hadrons

### 17.1 Introduction

Having written down a lepton current which provides a description of the purely leptonic reactions, we must now extend the concept to include the semi-leptonic processes, such as nuclear  $\beta$  decay, which are historically more important, and also the hadronic processes. Both these categories are divided up into strangeness-conserving reactions and strangeness-changing reactions, as a first step in categorising the effects of the weak interactions. Examples of the reactions in each category are shown in Table 17.1. The most obvious and serious difficulty in writing down a hadronic current (just as we wrote down the leptonic current) is that it is impracticable to write down wavefunctions for the observed hadrons; there are simply too many of them! If we were to use a separate wavefunction for each hadron, a current describing all the possible interactions would fill a book by itself. This is too complicated to be feasible.

### 17.2 The Hadronic Current

To proceed we can avoid the problem of wavefunctions for the hadrons and simply characterise the hadronic current in terms of its effect on the quantum numbers of the participating particles. So we can write the total weak interaction current as a sum of leptonic and hadron components:

$$J^W = L^W + H^W.$$

As before, the weak interaction amplitudes are generated by the product of the total current with its antimatter conjugate multiplied by the Fermi coupling:

$$G_F \bar{J}^W J^W = G_F (\bar{L}^W L^W + \bar{L}^W H^W + \bar{H}^W L^W + \bar{H}^W H^W).$$

This contains all the leptonic reactions ( $\bar{L}^W L^W$ ), the semi-leptonic reactions ( $\bar{L}^W H^W + \bar{H}^W L^W$ ), and the purely hadronic reactions ( $\bar{H}^W H^W$ ). The form of the hadronic current (less the wavefunctions) consists of one part which conserves the strangeness of the participating hadrons,  $h^\pm$ , and of one part which changes it,  $s^\pm$ :

$$H^W = h^\pm \cos \theta_C + s^\pm \sin \theta_C.$$

The relative strength of the two components is governed by the Cabbibo angle  $\theta_C$  which is a parameter intrinsic to the weak interactions and which must be measured from experiments. A variety of measurements indicate that  $\cos \theta_C$  is around 0.97, such that  $\sin \theta_C = 0.24$ , and so strangeness-conserving weak interactions predominate over strangeness-changing ones.

At this point we must remember that just as the leptonic current  $L^W$  contains the interaction factors  $\Gamma$  (which are a mixture of vector and axial-vector quantities), so too do the separate components of the hadronic current to ensure the correct parity-violating

Table 17.1. Categories of the weak interactions of hadrons.

Reaction class	Strangeness-conserving	Strangeness-changing
Semi-leptonic	$\pi^\pm \rightarrow l + \nu(\bar{\nu})$ $n \rightarrow p + e^- + \bar{\nu}_e$ $\mu^- + p \rightarrow \nu_\mu + n$ $K^0 \rightarrow \pi^\pm + e^\mp + \nu(\bar{\nu})$ $\bar{\nu} + p \rightarrow e^+ + n$	$K^\pm \rightarrow l + \nu(\bar{\nu})$ $K^\pm \rightarrow \pi^0 + l + \nu(\bar{\nu})$ $\Lambda^0 \rightarrow p + l + \bar{\nu}$ $\Xi \rightarrow \Lambda + l + \bar{\nu}$ $\bar{\nu} + p \rightarrow e^+ + \Lambda$
Hadronic	Parity-violating effects in ordinary hadron physics, e.g. $p + p \rightarrow p + p$	$K^0 \rightarrow n\pi$ ( $n = 2, 3$ ) $\Lambda^0 \rightarrow \pi^- p$ $\Sigma \rightarrow n\pi$

$l$  denotes  $e^\pm$  or  $\mu^\pm$ .

coupling of hadronic spins in the reactions. So the hadronic weak current has four separate components:

- (1) a vector current which conserves strangeness;
- (2) a vector current which changes strangeness;
- (3) an axial-vector current which conserves strangeness;
- (4) an axial-vector current which changes strangeness.

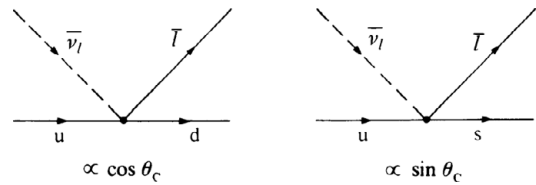


Figure 17.1. The weak interaction transforms quarks.

### 17.3 The Hadron Current and Quarks

The hadron current takes on a very simple form when written in terms of the quarks. Naively, we can think of the weak current as simply changing one flavour of quark (u, d or s) into another, and so changing the quantum numbers of the parent particle.

In fact, it is not quite as simple as this because the weak interaction does not specify a unique transformation say, from a u quark into a d quark. As we have seen, the weak current has both strangeness-conserving and strangeness-changing components, which implies that the u quark can have a certain probability of transforming into a d quark, and another of transforming into an s quark. So the current is written

$$H^W = \bar{u}\Gamma (d \cos \theta_C + s \sin \theta_C).$$

This current is shown symbolically in Figure 17.1. The decay of a strange meson is illustrated in Figure 17.2.

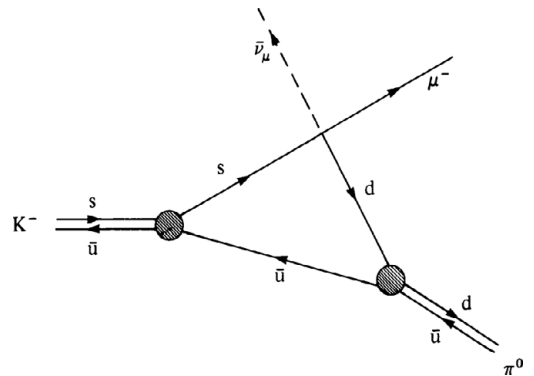


Figure 17.2. The quark picture of quantum-number flow during the semi-leptonic decay of the kaon.

This representation is very useful for envisaging the flow of quantum numbers during a reaction. Unfortunately, it is of limited use in calculating dynamical quantities because we have very little idea of how to represent the confinement of quarks mathematically. In the diagram the ignorance is contained within the shaded blobs.

# 18

## *The W Boson*

### 18.1 Introduction

It is true to say that we can explain all the data from low-energy weak interaction processes with the Fermi theory (expressed in the framework of the current–current theory of lepton processes). Unfortunately, this theory of a point-like interaction makes unacceptable predictions for high-energy weak interactions. We have just seen how the theory predicts that cross-sections for neutrino–electron scattering will rise linearly with the energy of the incoming neutrino. But we must realise that this prediction cannot be true for arbitrarily high energies. For instance, if it were true, neutrinos with exceedingly high energies (say those in cosmic rays, originating in space) would have a very high cross-section for interacting with matter, so we would expect neutrino collisions to be commonplace events in cosmic ray photographs and more common in laboratory bubble chambers. This is just not true, and we must accept that our formula for the neutrino cross-section is valid only at small energies. Also, there are extremely well-founded theorems resting only on assumptions, such as causality, which constrain the rate at which cross-sections can rise with energy.

To solve this problem, and also to put the description of the weak interaction on a common footing with the theories of electromagnetism and the strong nuclear force, it is necessary to abandon the four-fermion point-like interaction and replace it with

a particle exchange mechanism (just like, say, pion exchange between nucleons).

The particle which carries the weak interaction is called the intermediate vector boson, denoted  $W$ , see Figure 18.1. What we must do is to go back and describe all the weak interaction phenomena in terms of a particle exchange mechanism, which must approximate to the four-fermion point-like interaction at low energies to preserve its successful explanation of the data.

### 18.2 The W Boson

The essence of the  $W$ -boson mechanism is that the two currents involved in a weak interaction process no longer couple directly to each other at a single point. Instead, each current couples to the  $W$ -boson wavefunction at different space–time points and the  $W$  boson mediates the interaction between the currents. The basic weak interaction amplitude  $m^{(1)}$  is thus the coupling of the current with the  $W$ -boson wavefunction  $W$  (Figure 18.2):

$$\begin{aligned} m^{(1)} &\equiv g \left( L^W \bar{W} + W \bar{L}^W \right) \\ &\equiv g \left( \bar{\psi}_e \Gamma \psi_{\nu_e} \bar{W} + W \bar{\psi}_{\nu_e} \Gamma \psi_e \right). \end{aligned}$$

The first thing to note about the  $W$  boson is that it must come in both positively and negatively charged versions if it is to allow the transformations of positrons and electrons into antineutrinos and neutrinos respectively. Also, if we wish to describe neutral current

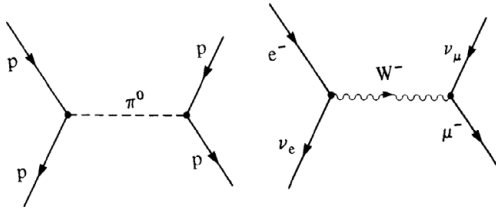


Figure 18.1. Just as the pion carries the strong force between hadrons, so the W boson carries the weak force between leptons.

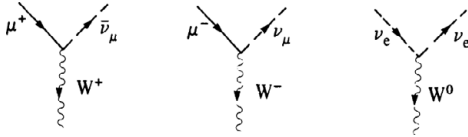


Figure 18.2. Basic lepton processes defining the charge states required of the W boson.

phenomena, we must allow the existence of a neutral intermediate boson as well which, for the moment, we will simply call W<sup>0</sup>. The role of the differing charge states is shown in Figure 18.2.

Another property of the W boson, which is easily established, is that it must be very massive. Recall our simple argument of Chapter 7, using Heisenberg's uncertainty principle: the range of the force may be thought of as typified by the maximum distance which its carrier can travel in the time element allowed by the uncertainty principle. The more massive the carrier, the shorter the range of the force. As the Fermi theory managed quite satisfactorily with a point-like assumption, it follows that the W boson must be much more massive than the pion (which allows the strong force the measurable range of 10<sup>-15</sup>m).

It is clear that the simple amplitudes of Figure 18.2 represent the creation (or destruction) of a W boson from the (into the) familiar leptons. The reactions involving only leptons as external particles will result from higher-order interactions, represented by products of these basic amplitudes:

$$m^{(2)} = g^2 \left( L^W \bar{W} W \bar{L}^W \right) \\ = g^2 \left( \bar{\psi}_e \Gamma \psi_{\nu_e} \langle \bar{W} W \rangle \bar{\psi}_{\nu_e} \Gamma \psi_e \right).$$

Figure 18.3 shows the W-exchange amplitudes for some of the basic weak interaction processes we have

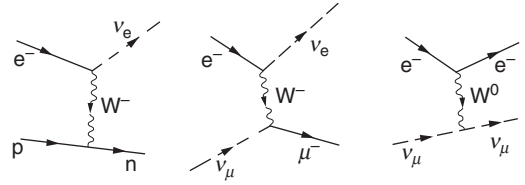


Figure 18.3. Basic weak interaction processes mediated by W exchange.

met so far. The factor in the amplitude describing the propagation of the virtual W boson ( $\langle \bar{W} W \rangle$ ) is known as the W-boson propagator, which acts as its wavefunction between its creation and its destruction. It is this new factor which improves the unacceptable high-energy behaviour of total cross-sections. The mathematical expression for the W-boson propagator describes its mass and its spin (spin 1 for a vector particle) and allows us to relate the old Fermi coupling  $G_F$  to the new lepton-W coupling  $g$ . At low energies we find:

$$G_F \propto \frac{g^2}{M_W^2}.$$

Knowing the expression for the W propagator as well as the wavefunction for the external leptons and the interaction factors allows us to recalculate the cross-sections for all processes of interest. At low energies, the answers are the same as for the Fermi theory, as desired.

### 18.3 Observing the W Boson

We have suggested that the quantum of the weak force, the W boson, has spin 1 (like the photon), comes in three charge states (W<sup>+</sup>, W<sup>0</sup>, W<sup>-</sup>), and decays into the familiar leptons. Its discovery was anticipated for many years, during which time it was generally believed that its mass was so large as to prevent it being seen in experiments. However, as we shall see in the next part, its mass can be predicted from the modern theory of the weak interactions (and the relevant experimental data). Furthermore, in experiments carried out at CERN in 1983, the W boson finally revealed itself, just as predicted with a mass around 80 GeV.



## **Part VI**

# **Gauge Theory of the Weak Interactions**



# 19

## *Motivation for the Theory*

### 19.1 Introduction

The description of the weak interactions of leptons afforded by the current–current theory of Chapter 15 provides a good account of low-energy experimental observations. This description includes the use of wavefunctions for the leptons (possible because there are just a few of them) and can also incorporate the use of W bosons in the role of ‘the photons of the weak interactions’. We might then wonder why it is not possible to write down a fully fledged theory of the weak interactions of leptons mediated by the W bosons, similar to the QED theory of electrons and photons. Indeed, there is considerable theoretical motivation for doing so. Firstly, it would be satisfying to be able to calculate answers to any desired degree of accuracy, in contrast to using only the simplest interactions to which our current understanding of the W bosons limits us. Secondly, we would like to have a predictive theory which could demonstrate its relevance by actually revealing new phenomena. And finally, if we can derive a theory similar to QED, this may allow us to formulate a unified theory of weak and electromagnetic forces.

### 19.2 Problems with the W Bosons

Given the motivation and the basic building blocks of the theory, we immediately discover several problems arising in the use of the W bosons. These problems originate in the fact that these particles have

spin 1 *and* non-zero mass. Let us investigate the consequences of this seemingly innocuous combination.

When we first talked about quantum-mechanical spin, we noted that although the naive picture of the spinning ball is helpful, it is a simplification of a more fundamental attribute. In fact, the spin of particles is the method for categorising their transformation properties under Lorentz transformations. A spin-1 particle is said to *transform* like a vector, which means that it has three components to define its orientation (i.e. its polarisation) at any point. Normally, we define the components such that two (i.e. the transverse) are perpendicular to and one (i.e. the longitudinal) is parallel to the direction of the momentum of the particle, see Figure 19.1. This is appropriate for a *massive* spin-1 particle, but for a *massless* spin-1 particle, like the photon, the transformations of special relativity show that the longitudinal degree of freedom has no physical significance. The photon can always be considered to be polarised in the plane transverse to its direction of motion. The difference becomes important when the propagators carry very high momentum. The transverse propagator for the massless photon behaves like  $1/p^2$  and becomes very small at high momentum (i.e. large  $p^2$ ). But for the massive vector particle, the presence of the extra longitudinal component in its propagator spoils this behaviour. At very high momenta, the massive propagator approaches a constant value of  $1/M^2$ .



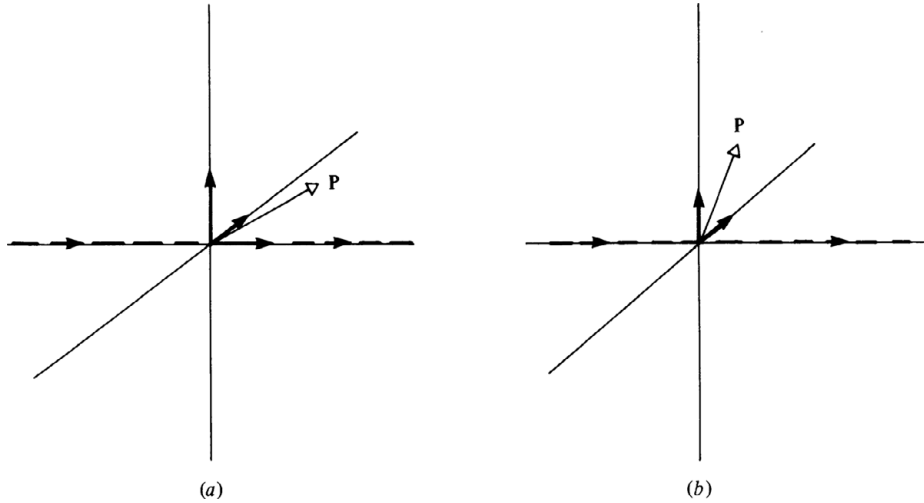


Figure 19.1. Vector propagators. A massive vector particle can have three components of polarisation (a), a massless vector particle only two (b).

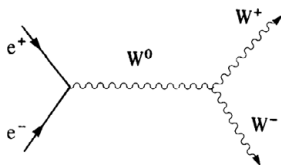


Figure 19.2. A process with bad high-energy behaviour caused by the mass of the W bosons.

We have seen that, in perturbation theory, the probability of an event occurring is given by the sum of contributions from a series of increasingly more complex Feynman diagrams. Each of these contributions should be, to all intents and purposes, a dimensionless number – probability carries no dimensions. Furthermore, as we have seen in QED, it is necessary to sum over all the possible values for the unobserved momenta of all the internal virtual particles in any diagram. However, at very high momenta a massive W-boson propagator contributes a factor of  $1/M^2$ . To compensate for this, each propagator (i.e. each wiggly line in Figure 19.2) must be multiplied by corresponding factors of the momentum in order to give a dimensionless contribution of the form  $p^2/M^2 c^2$ .

The presence of these momentum factors multiplying each propagator means that the mathematical expressions for the increasingly more complicated diagrams will be infinite when summed over all possible internal momenta. The diagrams are said to *diverge*.

In summary, it is the mass factor in the W-boson propagator which leads to an ever-increasing number of infinite contributions to the perturbation series. It is not possible for these to be reabsorbed into redefinitions of the masses and couplings. The theory is unrenormalisable and incapable of providing sensible answers to an arbitrary degree of accuracy.

Related to the problem of the theory's lack of renormalisability is the bad high-energy behaviour exhibited by some processes involving the W bosons. The presence of the mass factor in the W-boson propagator causes the cross-sections for these processes to rise with energy faster than is allowed by fundamental theorems, see Figure 19.2. This bad high-energy behaviour is precisely the problem which the W bosons were introduced to cure! As it is the mass of the W boson which seems to be the source of the trouble, the best thing for us to do is to study the origins of this particle and its mass further.

**20.1 Introduction**

The principle of gauge invariance is perhaps the most significant of the concepts used in modern particle theories, as it is the origin of the fundamental forces themselves. It appears to apply to *all* of the four known forces described in Chapter 5 (in one guise or another), and so may eventually provide us with the basis for a comprehensive unified theory. The basic method of gauge theory is to ensure that the Lagrangian describing the interaction of particle wavefunctions remains invariant under certain symmetry transformations which reflect conservation laws observed in nature. As a first step, we can see how this works in QED.

**20.2 The Formulation of QED**

QED seeks to explain the interaction of charged particles, say electrons, in such a way that total electric charge is always conserved. To represent this, the Lagrangian which describes the electron wavefunction must be invariant under a certain group of symmetry transformations  $\mathbf{G}$  (Figure 6.4.). We write this symbolically:

$$\mathbf{G}\mathcal{L}(\psi_e) \rightarrow \mathcal{L}(\psi_e^*).$$

In fact, the group in question, denoted  $U(1)$ , corresponds to a simple shift in the phase of the electron wavefunction. This is called a *global* phase transformation because it represents an identical operation at all points in space–time. From Section 3.8, we know

that the actual value of the phase of the electron wavefunction is unobservable – but we can observe differences in phase. However, before so doing, we must first establish a convention as to the starting point from which such phase differences are measured. Clearly, if the results we obtain are to make any sense, they must be independent of whichever convention we choose. Furthermore, performing a global phase transformation corresponds to changing this convention. So, the global phase symmetry is just a statement of the fact that the laws of physics are independent of the choice of phase convention.

The global phase symmetry is a relatively simple symmetry and does not place a very strong constraint on the form of the Lagrangian. The exercise becomes more interesting if we demand that the theory be invariant under *local* phase transformations which vary according to position:

$$\mathbf{G}(x)\mathcal{L}(\psi_e) \rightarrow \mathcal{L}^*(\psi_e^*),$$

where  $x = (\mathbf{x}, t)$  denotes a space–time four-vector. These are called *local gauge transformations*. A local gauge transformation corresponds to choosing a convention for defining the phase of the electron wavefunction, which is different at different space–time points. That is, the convention can be decided independently at every point in space and at every moment in time. But because of the space–time dependence, the Lagrangian representing the electron wavefunction is changed by the transformation ( $\mathcal{L} \neq \mathcal{L}^*$ ), and

the theory is not invariant under this more demanding symmetry. However, it comes as a pleasant surprise to find that by introducing another field, which compensates for the local change in the electron wavefunction, we can obtain a Lagrangian which indeed exhibits such symmetry.

The required field must have infinite range, since there is no limit to the distances over which the phase conventions must be reconciled. Hence the quantum of this new field must be massless. In fact, the field required for local gauge invariance is none other than the electromagnetic field whose quantum is, of course, the photon.

We already know that the interaction of two electrons should most correctly be described in terms of one electron interacting with a photon at one point, the propagation of the photon, and its subsequent interaction with the other electron at another space–time point. It so happens that the changes in the photon wavefunction cancel out the changes in the Lagrangian resulting from a local phase transformation. So, the introduction of the photon leads to local gauge invariance. We write this symbolically as

$$\mathbf{G}(x) \mathcal{L}(\psi_e, A) \rightarrow \mathcal{L}(\psi_e^*, A^*),$$

where the symbol  $A$  denotes a four-vector describing the electromagnetic field (i.e. the photon’s wavefunction).

So, invariance of the Lagrangian under local gauge transformations requires the existence of a massless gauge boson: the photon, which is the quantum of the long-range electromagnetic field. A physical explanation of its role is that the photon communicates the different space–time–dependent conventions, which define the phase of the electron wavefunction, between different points in space–time. Furthermore, the gauge symmetry implies a conservation law (see Section 6.1), namely, the conservation of electric charge. Hence, the gauge theory of QED successfully explains the interactions of electrons (and other charged particles).

An interesting fact is that the presence in the Lagrangian of a term attempting to describe a hypothetical mass for the photon would destroy the gauge invariance. As we have already seen, massive spin-1 particles generally give rise to non-renormalisable theories. So we may suspect local gauge invariance to be a good guide to the renormalisability of theories.

### 20.3 Generalised Gauge Invariance

Our next task is to generalise the principle of gauge invariance to other particles and other forces. The most convenient example is a historical one which, in fact, turned out to be untrue, but which follows on naturally from our previous discussions. It was originally proposed as a theory describing the strong interaction.

The charge independence of the strong nuclear force means that it acts identically on both protons and neutrons. It cannot distinguish between them but instead ‘sees’ only one basic nucleon  $N$ . This led us to categorise the proton and neutron as the two isospin components of the isospin- $\frac{1}{2}$  nucleon. The charge independence of the force may then be expressed as its invariance under rotations in isospin space, and associated with the invariance is the conservation of the total isospin of the system on which the forces act. (Of course, this isospin symmetry is broken by the electromagnetic interaction, which discriminates between protons and neutrons. But as far as the strong interaction is concerned, isospin is a good symmetry.) The Lagrangian which describes the interaction of nucleons should then be invariant under the group of global isospin rotations  $SU(2)$ :

$$\mathbf{G}^{SU(2)} \mathcal{L}(N) \rightarrow \mathcal{L}(N^*),$$

where the group  $SU(2)$  effectively rotates a proton into a neutron and vice versa. As the strong force cannot distinguish between the two, we must establish a convention for what we call a proton and what we call a neutron: any possible mixture of the two can be used to define a nucleon. The global transformation essentially redefines the nucleon convention at all points in space.

As before, it is possible to require that the theory be symmetric under the more demanding local gauge transformations and, as before, it is found necessary to introduce a massless gauge particle  $\rho$  to ensure the invariance of the Lagrangian:

$$\mathbf{G}^{SU(2)}(x) \mathcal{L}(N, \rho) \rightarrow \mathcal{L}(N^*, \rho^*).$$

The source of this new gauge particle is isospin, just as the source of the electromagnetic gauge field is electric charge. Because the nucleon may or may not change its electric charge in interaction, the new gauge particle must come in three charge states to do its job,  $(\rho^+, \rho^0, \rho^-)$ . Furthermore, because electric charge is related to isospin by  $Q = e[I_3 + (Y/2)]$ , the  $\rho$  gauge

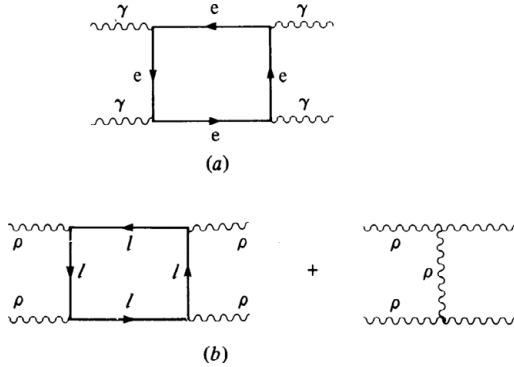


Figure 20.1. In QED, photons cannot interact directly, only with electrons (a). In  $SU(2)$  gauge theory, the gauge particles can interact both directly and indirectly (b).

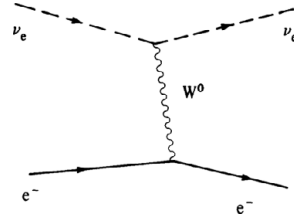


Figure 20.2. The neutral gauge particle allows weak neutral current reactions.

particle carries its own isospin and so can act as its own source. This allows the charged gauge particle to interact with itself, in contrast to the neutral photons in QED, see Figure 20.1. The difference between the two is expressed by saying that QED is an Abelian field theory while the  $SU(2)$  gauge theory is non-Abelian. The difference is a consequence of the mathematical structure of the gauge groups. The simple shift in phase in QED is said to be Abelian because a series of transformations can be performed in any order to produce the same effect as one big transformation. The group of rotations in isospin space  $SU(2)$  (which is also the group of ordinary spatial rotations in the three-dimensional space of the everyday world) is non-Abelian because a series of transformations does depend on the order of operation.

Gauge invariance was first generalised to the isospin invariance of the Lagrangian by Yang and Mills in 1954 and Shaw in 1955, but further work was deemed nugatory as such invariance apparently required the existence of a charge triplet of massless spin-1 gauge bosons (the  $\rho$  particles) which do not exist. A few years later the  $\rho$  mesons were discovered and the possibility of a gauge theory for the strong interaction was briefly entertained. Unfortunately, the  $\rho$  mesons are massive bound states of two pions and cannot be considered as candidates for this fundamental role.

### 20.4 Gauge Invariance and the Weak Interactions

The first step we must take in formulating a sensible theory of the weak interactions is to incorporate

the basic laws of leptonic physics which we have already met. These are the separate laws of electron-number and muon-number conservation. It is natural, therefore, to group the leptons' wavefunctions into doublets of the same lepton type:

$$l_e = \begin{pmatrix} \nu_e \\ e^- \end{pmatrix} \quad l_\mu = \begin{pmatrix} \nu_\mu \\ \mu^- \end{pmatrix}.$$

If gauge theory is now to be used, it is necessary to identify some conservation law which will imply the existence of some symmetry of the Lagrangian. To discover this, we note the similarity of the above doublets to the isospin doublet of the nucleon containing the proton and the neutron. We can then carry this similarity further and propose that the weak interaction also is independent of the electrical charge of the particles on which it acts. The weak interaction 'sees' only a lepton and cannot distinguish between a neutrino and an electron.

This leads us to define 'weak isospin' in a fashion exactly analogous to the isospin of nucleons, and allows us to require that the weak interaction be invariant under rotations in this weak isospin space. So the Lagrangian must be invariant under the group of weak isospin rotations, denoted by  $SU(2)^W$  to show it is acting on the leptons' wavefunctions,

$$\mathbf{G}^{SU(2)^W} \mathcal{L}(l_e, l_\mu) \rightarrow \mathcal{L}(l_e^*, l_\mu^*).$$

In exact parallel to the previous discussion, enforcing the more demanding local symmetry requires the introduction of massless gauge particles,  $W$ , to guarantee the invariance of the Lagrangian:

$$\mathbf{G}^{SU(2)^W}(x) \mathcal{L}(l_e, l_\mu, W) \rightarrow \mathcal{L}(l_e^*, l_\mu^*, W^*).$$

Here the physical reason for the existence of the  $W$  boson is to communicate between interacting leptons the locally defined convention governing the mixture of 'electron' and 'neutrino' constituting the lepton.

Again, the gauge particle must be a charge triplet ( $W^+$ ,  $W^0$ ,  $W^-$ ), and here we have something new for the weak interactions: the presence of the  $W^0$  particle allows *neutral-current* reactions in which a neutrino does not have to become an electron, see Figure 20.2. However, this theory was proposed well before the discovery of neutral currents, and the prediction of a  $W^0$  seemed a positive inconvenience. Also inconvenient was the fact that gauge invariance requires the gauge particles to be massless, yet the absence of

experimental evidence for the W boson led inevitably to the conclusion that it must be heavy.

These two factors hampered the development of gauge theory for almost a decade. What was required was a mechanism by which the W bosons may be allowed to have mass while originating from a gauge-invariant (and thus possibly renormalisable) theory. Also, the observed absence of neutral currents withheld the experimental confirmation of the theory necessary for its credibility and its development.

# 21

## *Spontaneous Symmetry Breaking*

### 21.1 Introduction

The existence of asymmetric solutions to a symmetric theory is common to many branches of physics. Consider, for example, an ordinary magnet. Its magnetic field clearly defines a preferred direction in space (i.e. rotational symmetry is broken), but the equations governing the motions of the individual atoms in the magnet are entirely rotationally symmetric. How has this come about? The answer lies in the fact that the symmetric state is *not* the state of minimum energy (i.e. the ground state), and that in the process of evolving towards the ground state, the intrinsic symmetry of the system has been broken.

A simple mechanical example is the behaviour of a marble inside the bottom of a wine bottle (see Figure 21.1). The symmetric state obviously corresponds to the marble taking the central position on top of the hump: but this is not the state of least energy, as the marble possesses potential energy due to its elevation. A small perturbation will send the marble tumbling down into the trough, where the system will possess least energy, but will also be rotationally asymmetric. When the symmetry of a physical system is broken in this way by an asymmetric ground state, we say the system exhibits ‘spontaneous symmetry breaking’.

### 21.2 Spontaneous Breaking of Global Symmetry

We now want to apply these ideas to particle physics to see if a spontaneously broken gauge

theory has anything to do with gauge boson mass. In this context, spontaneous symmetry breaking means that, although the Lagrangian is symmetric, the actual ground state is asymmetric. Recall from Section 4.4 that associated with every particle is an underlying quantum field; the particle is the quantum of the field. The ground state of the field – its vacuum state – is a state in which the field has its lowest possible energy and no particles are present (Section 4.9). For most fields, the energy is minimised when the average value of the field is zero; but for some, the energy is minimised only when it takes some uniform non-zero value.

Let us start by considering a hypothetical spinless particle consisting of two components:

$$\Phi = (\phi_1, \phi_2).$$

(This is analogous to considering the nucleon as a particle with two components: the proton and neutron, i.e.  $N = (p, n)$ .) Let us now write down a hypothetical Lagrangian which specifies the interaction between  $\phi_1$  and  $\phi_2$ . Suppose we choose the wine-bottle shape of the previous section to describe the interaction energy. Figure 21.2 shows such a choice for the interaction energy. The axes labelled  $\phi_1$  and  $\phi_2$  correspond to the average values of the associated quantum fields. For the interaction we have chosen, the energy is not a minimum at zero values of the fields, but around the circle defined by

$$\phi_1^2 + \phi_2^2 = R^2.$$

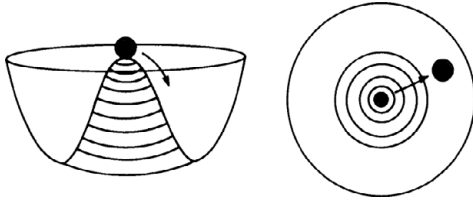


Figure 21.1. The initial position of the marble is symmetric but not minimum energy. A small perturbation will cause the rotational symmetry to be broken and the system to assume the state of minimum energy.

This equation defines the vacuum states of this theory, which are characterised by non-zero average values for  $\phi_1$  and  $\phi_2$ . Despite this unusual property, the Lagrangian is still symmetric under transformations between  $\phi_1$  and  $\phi_2$ , i.e. under rotations in the  $\phi_1$ – $\phi_2$  plane:

$$\mathbf{G}\mathcal{L}(\phi_1, \phi_2) \rightarrow \mathcal{L}(\phi_1^*, \phi_2^*),$$

where  $\mathbf{G}$  is the rotation group in the plane. Note that for the moment we are dealing with global transformations:  $\mathbf{G}$  acts in the same way at all points in space–time.

For definiteness, consider the particular vacuum state given by  $\phi_1 = 0$  and  $\phi_2 = R$ . For reasons of convenience, we might want the average values of the fields to be zero in the vacuum state. In fact, this is necessary if we wish to use the mathematics of perturbation theory. Moreover, it can be arranged by a simple redefinition of the fields:

$$\phi'_1 = \phi_1, \quad \phi'_2 = \phi_2 - R.$$

This simply corresponds to drawing new axes through the point  $R$  in Figure 21.2. If we write the Lagrangian in terms of the new fields, it should describe exactly the same physics (after all, all we have done is to make a redefinition):

$$\mathcal{L}(\phi_1, \phi_2) \equiv \mathcal{L}'(\phi'_1, \phi'_2).$$

However, some interesting features arise in this redefined system. Firstly, the vacuum is *not* invariant under the original group  $\mathbf{G}$  of rotations. Secondly, the Lagrangian now describes  $\phi'_2$  as a massive particle (mass proportional to  $R$ ), and  $\phi'_1$  as a massless particle. This is in contrast to the original Lagrangian in which the concept of mass was rather ill-defined. The net result is that the global symmetry of the original

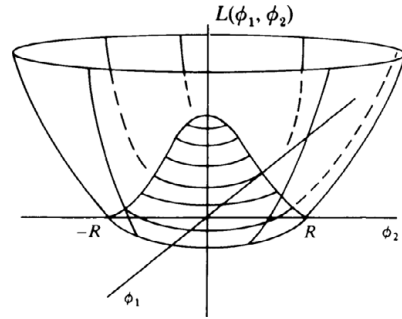


Figure 21.2. The interaction energy chosen for the two components of a hypothetical quantum field. The state of minimum energy corresponds to a non-zero value for the field.

Lagrangian is necessarily broken in the vacuum and, as a consequence, one of the particles has been given a mass while the other remains massless. The interesting question now is whether or not particle masses in the real world originate in a similar fashion from some originally gauge-invariant (and hence possibly renormalisable) interaction.

Unfortunately, things are not of much use as they stand in this simple model. The presence of the massless spin-0 particle turns out to be a general consequence of this type of mechanism (a theorem proved by Cambridge physicist Jeffrey Goldstone in the early 1960s): whenever a *global* symmetry is spontaneously broken, a massless, spin-0 particle results. This particle is called a ‘Goldstone boson’. This situation is unfortunate as no such massless, spinless particle appears to exist in the real world. Moreover, the particle which has developed a mass is nothing like a W boson (which has spin 1) and this remains massless. Our trouble seems to be increasing! Happily, however, all these difficulties can be resolved, as we shall now see.

### 21.3 Spontaneous Breaking of Local Symmetry – the Higgs Mechanism

Now take the original Lagrangian with the same wine-bottle-shaped interaction energy, and demand that it be invariant under *local* gauge transformations (i.e. under rotations in the  $\phi_1$ – $\phi_2$  plane which vary from place to place). Then we know from Chapter 20 that we must introduce a gauge particle, which we shall denote by  $A$ , in order to maintain the invariance:

$$\mathbf{G}(x)\mathcal{L}(\phi_1, \phi_2, A) \rightarrow \mathcal{L}(\phi_1^*, \phi_2^*, A^*).$$

In this instance, the gauge particle is responsible for communicating the  $\phi_1, \phi_2$  content of  $\Phi$  from place to place. As before, we now redefine the fields so as to arrange that our axes pass through the point of minimum energy:

$$\phi'_1 = \phi_1, \quad \phi'_2 = \phi_2 - R.$$

Rewriting the Lagrangian in terms of these redefined fields does not change the underlying physics, so we have

$$\mathcal{L}(\phi_1, \phi_2, A) \equiv \mathcal{L}'(\phi'_1, \phi'_2, A).$$

It is in this last step that something remarkable occurs. The redefined  $\phi'_2$  particle acquires, as before, a mass proportional to  $R$ , but astonishingly the massless Goldstone boson  $\phi'_1$  disappears. Moreover, the formerly massless gauge particle  $A$  now acquires a mass, again proportional to  $R$ .

What in fact happens is that the mathematical expressions describing the original massless gauge particle become mixed with  $R$  (the vacuum value of  $\phi_2$ ) in such a way as to create a mass term. At the same time, the Goldstone boson  $\phi'_1$  becomes absorbed into the gauge particle in such a way as to lose its physical significance. Physicists say that the gauge particle ‘eats’ the Goldstone boson and thereby becomes massive.

The physical interpretation of all this is the following. The original Lagrangian describes a two-component particle,  $\Phi = (\phi_1, \phi_2)$ , and a massless vector gauge particle  $A$ , consisting of two spin polarisation states. However, the redefined Lagrangian describes one massive spinless particle,  $\phi'_2$ , and one massive vector gauge particle,  $A'$ , which, by virtue of its mass, now contains three polarisation states. The total number of physical degrees of freedom remains the same (i.e. four), but the hapless Goldstone boson has become the third polarisation state of the massive gauge boson. This looks more encouraging.

Table 21.1. *The Higgs mechanism. Spontaneous breaking of local symmetry avoids the unwanted Goldstone boson and generates a mass for the vector boson. Both before and after symmetry breaking, the total number of physical degrees of freedom is four.*

Before		After
$A$	}	→ massive $A'$
$\phi_1$		
$\phi_2$	→	massive $\phi'_2$

The Goldstone boson has been avoided by using a *local* gauge symmetry, a step first taken by Peter Higgs of Edinburgh University and others in 1964. What is more, despite having started with a gauge-invariant theory, the gauge boson has acquired mass; this was the point of the entire exercise. The price to be paid for this success is the presence of the massive spin-0 particle,  $\phi'_2$  – the famous Higgs boson (see Table 21.1). The hunt for the Higgs has been one of the great sagas of modern physics culminating in the construction of the Large Hadron Collider at CERN in which it was finally detected in 2012, as discussed in Chapter 37.

Spontaneous breaking of *local* symmetry avoids the unwanted Goldstone boson and generates a mass for the vector boson. Both before and after symmetry breaking, the total number of physical degrees of freedom is four.

We have presented only a very simple example of how a local gauge symmetry may be spontaneously broken by the Higgs mechanism. This mechanism is quite general and can be applied straightforwardly to the gauge theory of the weak interactions, as we shall see in the next chapter.



## *The Glashow–Weinberg–Salam Model*

### 22.1 Introduction

In 1967 and 1968 respectively, Steven Weinberg of Harvard and Abdus Salam of London independently formulated a unified theory for the weak and electromagnetic interactions, based in part on work developed previously by Sheldon Glashow, also of Harvard. The theory describes the interactions of leptons by the exchange of W bosons and photons, and incorporates the Higgs mechanism to generate the masses for the W bosons. Because the Lagrangian prior to spontaneous symmetry breaking (i.e. prior to the redefinition of the fields) is gauge-invariant, Weinberg and Salam conjectured, although were not able to prove, that the theory is renormalisable. The proof was demonstrated subsequently by Gerard 't Hooft of Utrecht, in 1971.

The idea of the model is to write down a locally gauge-invariant Lagrangian describing the interactions of leptons with massless W-gauge bosons, just as described in Chapter 20. Hypothetical Higgs fields are then introduced with a suitably chosen interaction Lagrangian which is added to that for the leptons. Following the redefinition of the Higgs fields, the Lagrangian describes particles with mass. Because the theory is renormalisable, the Feynman rules can be used to calculate finite answers for any physical quantities to any desired degree of accuracy.

We must take care during the spontaneous symmetry breaking to ensure that the photon remains massless while the W bosons are made massive. This is achieved by a sufficiently clever choice of the Higgs

interactions such that, after redefinition of the fields, the vacuum is still invariant under some sub-group of the local gauge transformations. This sub-group is precisely the  $U(1)$  gauge symmetry of QED, which describes the interactions of massless photons with charged particles.

### 22.2 Formulation

We have previously grouped the electron with the electron-neutrino as two different, weak isospin states of a single lepton wavefunction,

$$l_e = \begin{pmatrix} \nu_e \\ e^- \end{pmatrix}.$$

However, a straightforward grouping like this is unsatisfactory, for while the neutrino is massless and left-handed, the massive electron is both right- and left-handed. We have already mentioned that the weak interaction prefers its electrons to be left-handed, such that if the electron were actually massless (a state well approximated by very relativistic electrons) it would act only on left-handed electrons. Therefore, let us split the electron wavefunction into separate right- and left-handed components and group them separately:

$$l_e = \begin{pmatrix} \nu_e \\ e^- \end{pmatrix}_L, \quad e_R^-.$$

In order to be able to do this consistently, the electron must be massless. So, we need to arrange for it to acquire a mass later. (This same separation is also

made for the muon and all particles that ‘feel’ the weak force. They too must be massless at this stage.)

### 22.2.1 Weak Interaction Charges and Gauge Symmetry

We now wish to write down a Lagrangian describing the interaction of these leptons, and to introduce gauge bosons by requiring the Lagrangian to be invariant under certain gauge transformations. To do this, we need to know the generators of these transformations. Or, equivalently, we need to know the quantities (i.e. the charges) which are conserved by the interactions. These weak interaction charges will differ before and after spontaneous symmetry breaking. Our hypothesis is that the present state of the world is the result of this symmetry breaking, and so we have relative freedom in choosing conservation laws prior to the breaking, provided that, after it, electric charge is conserved.

We have already identified weak isospin as a plausible candidate for a conserved charge for the weak interaction. The neutrino and the left-handed component of the electron form the weak isospin doublet,  $l_e$ , with  $I_3^W = +\frac{1}{2}$  and  $I_3^W = -\frac{1}{2}$  respectively. At this stage, prior to spontaneous symmetry breaking, the two components of the lepton doublet must be identical except for the value of the third component of their weak isospin,  $I_3^W$ . However, the electron has negative electric charge, while the neutrino is neutral. So, we must relate this electric charge difference to the difference in their  $I_3^W$  values. We can do this by introducing ‘weak hypercharge’  $Y^W$ , which is defined by the equation:

$$Q = e \left( I_3^W + Y^W / 2 \right).$$

So, by awarding both  $\nu_e$  and  $e_L^-$  a weak hypercharge of  $-1$ , the difference in their electric charges is given by the difference in their component  $I_3^W$  values. Furthermore, since  $I_3^W$  and  $Y^W$  are both conserved charges (before symmetry breaking), then electric charge will also be conserved.

The ‘weak quantum numbers’ of the leptons are summarised in Table 22.1. Note that because the right-handed component of the electron has no weakly interacting partner, it must have weak isospin zero,  $I^W = 0$  and  $I_3^W = 0$  (since it must transform into itself under weak isospin rotations). Consequently, its weak

Table 22.1. *The weak quantum numbers of the leptons.*

	$I^W$	$I_3^W$	$Y^W$	$Q$
$\nu_e$	$\frac{1}{2}$	$\frac{1}{2}$	$-1$	$0$
$e_L$	$\frac{1}{2}$	$-\frac{1}{2}$	$-1$	$-1$
$e_R$	$0$	$0$	$-2$	$-1$

hypercharge is directly related to its electric charge, and  $Y^W = -2$ .

We now demand that the interactions between leptons conserve weak isospin and weak hypercharge. We implement this by requiring the Lagrangian to be invariant under the  $SU(2)_L^W$  group of weak isospin transformations *and* under the  $U(1)^W$  group of weak hypercharge transformations (which correspond to simple shifts in the phase of the lepton wavefunction). So, the total symmetry group is  $SU(2)_L^W \times U(1)^W$ .

Antileptons have opposite values of  $I_3^W$ ,  $Y^W$  and  $Q$ .

To realise these invariances under local (space–time-dependent) transformations, we must introduce the appropriate gauge particles. Invariance under local rotations of weak isospin requires the introduction of the gauge particle  $W = (W^+, W^0, W^-)$ . Then

$$G^{[SU(2)_L^W]}(x) \mathcal{L}(l_L, W) \rightarrow \mathcal{L}(I_L^*, W^*).$$

Furthermore, to maintain invariance under shifts in the phase of the lepton wavefunction, we must introduce an additional gauge particle  $B$ , so that

$$G^{[U(1)^W]}(x) \mathcal{L}(l_L, e_R, B) \rightarrow \mathcal{L}(I_L^*, e_R^*, B^*).$$

So, the total gauge invariance can be expressed as

$$\begin{aligned} G^{[SU(2)_L^W \times U(1)^W]}(x) \mathcal{L}_1(l_L, e_R, W, B) \\ \rightarrow \mathcal{L}_1(I_L^*, e_R^*, W^*, B^*). \end{aligned}$$

### 22.2.2 Spontaneous Symmetry Breaking

At this point, we add two further terms to the Lagrangian. Each involves the Higgs fields, which, like the left-handed leptons, take the form of a doublet:

$$\Phi = \begin{pmatrix} \phi^0 \\ \phi^- \end{pmatrix},$$

where the weak quantum numbers are the same as those of the left-handed lepton doublet. Firstly, we add

the term associated with the wine-bottle-shaped interaction energy. As this term must also be locally gauge-invariant, it must also contain the gauge particles:

$$\mathcal{L}_2(\Phi, B, W).$$

Secondly, we may allow the Higgs fields to interact with the leptons as we have yet to generate their masses:

$$\mathcal{L}_3(l_L, e_R, \Phi).$$

Then the local gauge invariance of the total Lagrangian  $\mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_3$  under the  $SU(2) \times U(1)$  symmetry group is broken by the neutral Higgs component taking a non-zero vacuum value,  $\phi^0 = R$ , corresponding to the state of minimum energy. We must now redefine the Higgs field  $\Phi \rightarrow \Phi'$  so that it is zero at the state of minimum energy:

$$\phi^{0'} = \phi^0 - R, \quad \phi^{-'} = \phi^{-}.$$

After this redefinition, the Lagrangian must still describe the same physics, and so

$$\begin{aligned} \mathcal{L}_2(\Phi, B, W) &\equiv \mathcal{L}'_2(\Phi', B, W), \\ \mathcal{L}_3(l_L, e_R, \Phi) &\equiv \mathcal{L}'_3(l_L, e_R, \Phi'). \end{aligned}$$

Once the mathematical smoke has cleared following this redefinition, the following features emerge.

Weak isospin and weak hypercharge are no longer conserved charges since the  $SU(2) \times U(1)$  gauge symmetry has been broken. However, it has been broken in such a way that the combination corresponding to electric charge (i.e.  $Q = e(I_3^W + Y^W/2)$ ) is still conserved. This implies that the  $U(1)$  gauge symmetry of QED remains unbroken, and the photon remains massless.

Gauge boson masses are generated by the mixing of  $R$  (the vacuum value of  $\phi^0$ ) with  $B$  and  $W$  in  $\mathcal{L}_2$ . However, as we have noted, the photon remains massless while the other gauge bosons become massive. We can see how this happens when we notice that neither the  $W$ -gauge particle of weak isospin, nor the  $B$ -gauge particle of weak hypercharge, can be identified with the electromagnetic gauge particle, the photon. Just as the electric charge is a mixture of weak isospin and weak hypercharge, the electromagnetic gauge particle is similarly a mixture of the neutral gauge particles of weak isospin  $W^0$  and weak hypercharge  $B$ :

$$A = W^0 \sin \theta_W + B \cos \theta_W,$$

where the weak angle,  $\theta_W$ , is the parameter which adjusts the relative proportions of the two. The remaining portions of the  $W^0$  and  $B$  wavefunctions also mix together to produce another gauge particle:

$$Z^0 = W^0 \cos \theta_W - B \sin \theta_W.$$

This combination corresponds to the part of the weak interaction which has the same quantum numbers as the photon (i.e. zero electric charge). It is the neutral gauge boson corresponding to the weak neutral current mentioned previously, and is denoted  $Z^0$  to signify its origin as a combination of the two fundamental gauge fields.

Spontaneous symmetry breaking in the electroweak model leads to three massive vector bosons and one massive Higgs boson. Both before and after symmetry breaking, the total number of physical degrees of freedom is 12.

By careful choice of the form of the interaction energy, the  $Z^0$  can be given a mass, while the photon  $A$  remains massless. In obtaining a mass, the  $Z^0$  absorbs a mixture of  $\phi^0$  and its antiparticle  $\bar{\phi}^0$ , which provides the third polarisation state of the massive particle. We are left with a real massive Higgs particle,  $\phi'$ , which comes from the remaining mixture of  $\phi^0$  and  $\bar{\phi}^0$ . The charged components of the  $W$ -gauge particles (i.e.  $W^-$  and  $W^+$ ) obtain masses by absorbing  $\phi^-$  and its antiparticle,  $\phi^+$ , see Table 22.2. The values which emerge from the mathematics are:

$$M_{W^\pm} = \frac{38.5}{\sin \theta_W} \text{GeV}, \quad M_{Z^0} = \frac{M_{W^\pm}}{\cos \theta_W},$$

which clearly depend on the value of the weak angle,  $\theta_W$ . This is a free parameter in this model and must be determined by experiment. The experimental value is given by  $\sin^2 \theta_W \approx 0.23$ , and so the model predicts masses of approximately 80 GeV and 90 GeV respectively. This explains why the  $W$  bosons were

Table 22.2. *The electroweak Higgs mechanism*

Before	→	After
$\phi^-; W^-$	→	massive $W^-$
$\phi^+; W^+$	→	massive $W^+$
$\phi^0, \bar{\phi}^0; W^0, B$	→	$\left\{ \begin{array}{l} \text{massive } \phi' \\ \text{massive } Z^0 \\ \text{massless } A \end{array} \right.$

so difficult to detect. It was not until 1983 that accelerators had enough energy to produce such massive particles.

Recall that in order to consistently split the electron wavefunction into left- and right-handed components, it needed to be massless (at least before symmetry breaking). After symmetry breaking, we find that  $l_e$  and  $e_R$  are mixed with  $R$  (the vacuum value of  $\bar{\phi}^0$ ) in  $L_3$ , and the correct mass is indeed generated for the electron. However, this part of the exercise does not have the same predictive power as the mass generation for the gauge bosons. Because we have incomplete knowledge of how the Higgs field interacts with leptons, we are free to choose appropriate coefficients in  $\mathcal{L}_3$  so as to guarantee the correct electron mass.

### 22.3 Reprise

In presenting the basic structure of what is now the accepted theory of the ‘electroweak’ interactions, we have, for reasons of clarity, considered only the electron and its neutrino. However, *all* particles which ‘feel’ the weak force fit into the above scheme in a straightforward way. That is, all left-handed fermions form weak isospin doublets with  $I^W = +\frac{1}{2}$  and  $I_3^W = \pm\frac{1}{2}$ , and all right-handed fermions form weak-isospin singlets with  $I^W = 0$  and  $I_3^W = 0$ . So, for the muon we have

$$\begin{pmatrix} \nu_\mu \\ \mu^- \end{pmatrix}_L, \quad \mu_R^-.$$

The same is true even for the quarks, but we postpone discussion of this until the next chapter.

This then is the Glashow–Weinberg–Salam model of the ‘electroweak’ interactions. It incorporates the successful theory of QED and provides a description of the weak force in terms of the exchange of massive vector bosons. It has, moreover, introduced a weak neutral current in a natural fashion, and the discovery of neutral-current reactions in 1973 was a great boost to the acceptability of the model. The

model’s cleverest feature, however, is the way it ensures the masslessness of the photon, while giving mass to the weak interaction gauge bosons  $W^\pm$  and  $Z^0$ . This is achieved by the use of the Higgs mechanism and a suitable choice of Higgs fields. The gauge boson masses depend on the weak angle,  $\theta_W$ , which is a parameter which must be measured from experiments ( $\sin^2 \theta_W \approx 0.23$ ). In addition, the theory predicts the existence of a spinless Higgs particle  $\phi'$ , eventually discovered in 2012, (see chapter 41). The production of  $W^\pm$  and  $Z^0$  bosons at CERN in 1983 with precisely the predicted masses was a great triumph for the electroweak model and indeed for the concept of non-Abelian gauge theory.

### 22.4 An Academic Postscript – Renormalisability

Although the model was essentially formulated as above in 1967 and 1968, it was not enthusiastically received until its renormalisability had been demonstrated. As we have mentioned, the local gauge invariance of the Lagrangian prior to spontaneous symmetry breaking suggested that it would be, but it remained to be shown that the symmetry breaking itself did not spoil this property.

’t Hooft’s proof of renormalisability essentially consisted of showing that the Feynman rules of the theory lead to mathematical expressions for the W-boson propagators which avoid the problems associated with the use of massive spin-1 particles in perturbation theory (see Chapter 19). ’t Hooft showed that when the momentum flowing through a W-boson propagator is very large, then the mathematical expression of that propagator does not depend on the mass at all. As there is no mass dependence, there is no need for compensating momentum factors to ensure that the resulting probabilities are dimensionless numbers. It is these extra momentum factors which lead to the divergences (infinite values) when summing over all the internal momentum configurations of a complicated Feynman diagram; without them, the probabilities are finite and thus the theory is renormalisable.

## *Consequences of the Model*

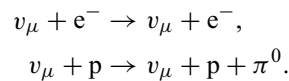
### 23.1 Introduction

The Glashow–Weinberg–Salam model is now the undisputed theory of the weak (and electromagnetic) interactions. It was established over the decade 1973–83 by a series of experiments which have confirmed the model’s predictions. Firstly, in 1973, the discovery of neutral currents revealed a qualitatively new phenomenon, just as predicted by the model. Soon after, in the mid-1970s, new mesons were discovered which support the existence of a new quark type carrying the ‘charm’ quantum number – just as had been suggested by theorists attempting to describe the weak interactions of hadrons using the Glashow–Weinberg–Salam model. Next, in the late 1970s, various parity-violating effects were found to be in close agreement with the quantitative predictions of the model. The discovery at CERN of the  $W^\pm$  and  $Z^0$  bosons in 1983 with precisely the masses predicted by the theory established the gauge theory structure of the weak interactions confirmed, eventually, by the 2012 discovery of the Higgs Boson, also at CERN.

### 23.2 Neutral Currents

In the discussion of neutrino–electron scattering in Chapter 16, we floated the possibility of the existence of neutral current processes by which, for example, an incoming neutrino may not have to turn into a charged lepton but could emerge with its identity unscathed. The emergence of the neutral gauge boson  $Z^0$  from the Glashow–Weinberg–Salam electroweak

model provides a natural explanation for such neutral current phenomena, and its detection was seen in the early 1970s as being an important test for the validity of the model. Reactions which proceed by the neutral current are, for instance, elastic muon–neutrino–electron scattering or quasi-elastic muon–neutrino–proton scattering:



It was possible to check for the existence of these reactions only when neutrino beams were intense and energetic enough to permit detailed accelerator experiments to be performed. This became reality in 1973 when, much to everyone’s surprise, neutral currents were found to be a significant effect, comparable in magnitude to the well-established charged currents. The process first seen was the second of the two mentioned above, see Figure 23.1, the magnitude of which, relative to its charged current version, was measured to be:

$$\frac{\sigma(\nu_\mu + p \rightarrow \nu_\mu + p + \pi^0)}{\sigma(\nu_\mu + p \rightarrow \mu^- + p + \pi^+)} = 0.51 \pm 0.25,$$

showing them to be effects of the same order. Since the first observation of this reaction, other neutral current processes have been observed as predicted by

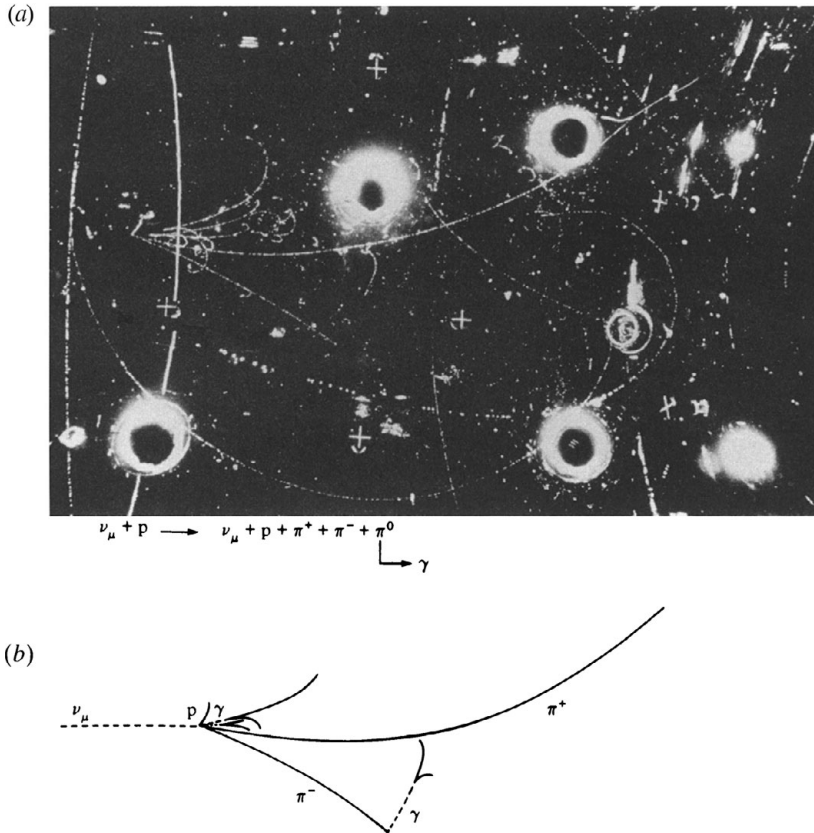


Figure 23.1. (a) A neutral current reaction photographed in the Gargamelle bubble chamber at CERN, and its interpretation (b). (Photo courtesy CERN.)

the electroweak model, including the elastic muon–neutrino–electron scattering mentioned above.

All these reactions allow us to determine the value of the weak angle  $\theta_W$ , which is the parameter which essentially fixes the relative mixing of the weak and electromagnetic interactions. The value which represents the average of the different experiments conducted to date is about

$$\sin^2 \theta_W = 0.23117 \pm 0.00016.$$

As we saw in Chapter 22, the masses of the intermediate  $W^\pm$  bosons depend directly on this quantity, which therefore predicts

$$M_{W^\pm} \approx 80 \text{ GeV} \quad M_{Z^0} \approx 90 \text{ GeV}.$$

### 23.3 The Incorporation of Hadrons – Charm

It is necessary also for the Glashow–Weinberg–Salam model to describe the weak interactions of hadrons. In Chapter 17 we saw how these could be

described using a weak interaction current written in terms of quarks. This form is the most convenient for investigating the consequences of the model. The charged current describing weak interactions where electric charge is changed is the same as before. It comprises a part which changes strangeness and a part which conserves strangeness, the relative importance of the two being regulated by the Cabbibo angle  $\theta_C$ .

However, the model also requires an electrically neutral current, which, in the absence of further specifications, will similarly consist of a part which changes the strangeness of the particles and a part which conserves it. An example of the strangeness-changing neutral current is given by the decay of the long-lived neutron kaon,

$$K_L^0 \rightarrow \mu^+ + \mu^-.$$

This decay can occur as a higher-order process involving *charged* currents, as in the diagram Figure 23.2(a). Unfortunately, these processes, apparently allowed in

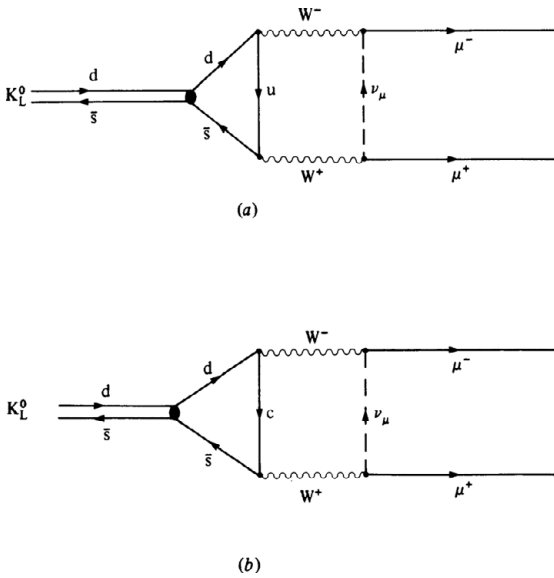


Figure 23.2. Strangeness-changing neutral currents which can proceed by the intermediate quark process in (a) can be cancelled out of the model by additional quark processes involving the charmed quark in (b).

the model, do not occur in the real world. Experimentally, the probability of this decay occurring is less than one part in one hundred million. So the model must be modified to ensure that these processes do not occur.

Several explanations were advanced in an attempt to cure this problem. The most successful, later to be rewarded by striking experimental evidence, is the ‘charm’ scheme proposed in 1970 by the international triumvirate of Glashow, Iliopoulos and Maiani, commonly known as GIM. The key assumption in the scheme is the existence of a new ‘charmed’ quark, *c*. This quark is to carry a new quantum number, charm, just as the strange quarks carry strangeness. All the other quarks, *u*, *d* and *s*, are assigned zero charm. It is then possible to arrange that the unwanted strangeness-changing neutral current of Figure 23.2(a) is cancelled out by the corresponding diagram involving the charmed quark, shown in Figure 23.2(b).

In this way, it is possible to arrange for the disappearance of all the unwanted currents in reactions such as

$$\begin{aligned} K^+ &\rightarrow \pi^+ + \mu^+ + \mu^-, \\ K^+ &\rightarrow \pi^+ + e^+ + e^-. \end{aligned}$$

Table 23.1. The weak quantum numbers of the quarks.

	$I^W$	$I_3^W$	$Y^W$	$Q$
$u_L$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{2}{3}$
$d_L$	$\frac{1}{2}$	$-\frac{1}{2}$	$\frac{1}{3}$	$-\frac{1}{3}$
$u_R$	0	0	$\frac{4}{3}$	$\frac{2}{3}$
$d_R$	0	0	$-\frac{2}{3}$	$-\frac{1}{3}$

Antiquarks have opposite values of  $I_3^W$ ,  $Y^W$  and  $Q$ .

The consequences of the GIM scheme are enormous. The existence of this fourth quark flavour implies the existence of whole new families of charmed mesons and charmed baryons, rather like a repeat showing of all the strange particles. What is more, for the charmed quark diagram to cancel out the unwanted reactions effectively requires the mass of the charmed quark (and hence charmed particles) to be relatively small at about 1.5 GeV. So we should see charmed particles copiously produced in modern accelerators.

The charm scheme seemed to many physicists to be rather an ‘expensive’ way of solving the problem, i.e. having to introduce whole families of unknown particles just to alter the reaction rate of a small obscure class of reactions. However, this scepticism soon turned to amazement as the required particles tumbled out of the accelerators in the mid-1970s (see Chapter 35).

We mentioned in the previous chapter that the Glashow–Weinberg–Salam model treats all particles that ‘feel’ the weak force in precisely the same way. That is, all left-handed fermions form weak isospin doublets with  $I^W = \frac{1}{2}$  and  $I_3^W = \pm\frac{1}{2}$ , while all right-handed fermions form weak isospin singlets with  $I^W = 0$  and  $I_3^W = 0$ . This is true even for the quarks. For instance, for the up and down quarks we have

$$\begin{pmatrix} u \\ d' \end{pmatrix}_L, u_R, d_R.$$

Their weak quantum numbers are given in Table 23.1. Similarly, for the charmed and strange quarks we have

$$\begin{pmatrix} c \\ s' \end{pmatrix}_L, c_R, s_R.$$

Note that it is not  $d$  and  $s$  which appear in the above quark doublets, but  $d'$  and  $s'$ . This is because the charged hadronic current has both strangeness-conserving and strangeness-changing components. Therefore, it is not  $d_L$  and  $s_L$  which couple to the charged  $W^\pm$  bosons but the combinations

$$\begin{aligned}d'_L &= s_L \sin \theta_C + d_L \cos \theta_C, \\s'_L &= s_L \cos \theta_C + d_L \sin \theta_C,\end{aligned}$$

where  $\theta_C$  is the Cabbibo angle. Because of the trigonometric relation  $\sin^2 \theta_C + \cos^2 \theta_C = 1$ , this rotation makes no difference to the neutral currents.

### 23.4 Parity-violating Tests of the Glashow–Weinberg–Salam Model

In addition to the qualitatively new phenomena described in the previous two sections, the Glashow–Weinberg–Salam model also tells us the magnitude of the various parity-violating effects due to the weak force. The most convincing evidence in this area is provided in polarised electron–deuteron scattering in which all the electrons can be collided with their spins pointing in a specified direction. The experiment measures the difference in the scattering cross-sections between left- and right-handedly polarised electrons off polarised deuteron targets. Not only is the magnitude predicted correctly, giving a difference between cross-sections of about 0.01%, but the way this difference changes with the energy of the collision is also explained.

Another interesting parity-violating experiment is that concerning the interaction of light with matter at low energies. This provides a test of the model in a wholly new domain, and is thus that much stronger a test of its general worth. There are many variants of the experiment, but the basic idea is to shine a beam of polarised light (light whose electric field vector is aligned in a specific direction) through a vapour of metal atoms. The light is absorbed and re-emitted by the various transitions of the atomic electrons and, because of the effects of the weak interaction between the atomic nucleus and the electrons, this leads to a very slight, but well-defined, rotation of the plane of polarisation in a given direction. Such a given rotation, of course, implies a distinction between clockwise and anticlockwise and so is indicative of parity violation. Because the effect is slight and because there are uncertainties about using the Glashow–Weinberg–Salam model in the atomic environment, the experiments and their interpretation are extremely hard work. But after initial doubts and discrepancies, there is now a convincing consensus of experimental results supporting the model. The electroweak model also predicts novel effects in electron–positron annihilations (see Chapter 34) and these also provide support. The model is now our standard working understanding of the electromagnetic and weak interactions (sometimes now referred to as the unified electroweak force) and in recognition of this, Glashow, Weinberg and Salam were awarded the Nobel Prize in 1979.



## *The Hunt for the $W^\pm$ , $Z^0$ Bosons*

### 24.1 Introduction

The intermediate vector bosons of the weak force were by far the most eagerly awaited particles of the 1980s. Their existence is crucial to the validity of the Glashow–Weinberg–Salam electroweak theory, and, by implication, to the acceptability of all the modern theories of the ‘gauge’ type. It was not surprising then that between 1976 and 1983 an enormous experimental effort was dedicated to their detection. But as we have seen, the masses of these particles are very large indeed. So just how were they detected?

The  $Z^0$  boson decays into both quark–antiquark pairs and lepton–antilepton pairs. So, by the principle of microscopic reversibility of particle reactions, we may anticipate that sufficiently energetic collisions between quarks and antiquarks, or leptons and antileptons, will produce  $Z^0$  bosons. The ‘clearer’ of the two possibilities is provided by lepton–antilepton annihilation, such as in the  $e^+e^-$  experiments described in Part IX. The LEP collider at CERN, which was commissioned in 1989, was justified largely by this prospect, despite the inherent difficulties in the handling of very high energy  $e^+e^-$  beams (see Section 34.2).

In the mid-1970s, physicists realised that an easier and less expensive means of searching for the  $W^\pm$  and  $Z^0$  bosons (although a less satisfactory environment in which to study them) was provided by quark–antiquark annihilation in collisions between protons and antiprotons. This was despite the very messy final states resulting from the spectator quarks. In such

reactions, one of the quarks inside the proton ( $uud$ ) annihilates with an antiquark inside the antiproton ( $\bar{u}\bar{u}\bar{d}$ ) to produce a  $W^\pm$  boson if the pair is dissimilar (i.e.  $ud$  or  $\bar{u}\bar{d}$ ), or a  $Z^0$  boson if the pair is similar (i.e.  $u\bar{u}$  or  $d\bar{d}$ ). This is illustrated in Figure 24.1.

In 1976, Carlo Rubbia, David Cline and their colleagues suggested converting a conventional ‘fixed-target’ proton accelerator into a proton–antiproton collider in order to provide the earliest possible opportunity for discovering these massive gauge bosons.

### 24.2 The CERN $p\bar{p}$ Collider Experiment

Following the general acceptance of the  $pp$  idea, the accelerator chosen for the job was the super proton synchrotron accelerator (SPS) at CERN which, as of 1976, was one of the highest-energy machines in the world, able to accelerate protons to 400 GeV stop. The great beauty of the  $pp$  idea is that, because the antiparticles ( $\bar{p}$ ) have the opposite charge to, but the same mass as, the particles ( $p$ ), the accelerator configuration which accelerates protons in one direction will automatically accelerate antiprotons in the opposite direction. So the one-beam ring of the original SPS was made to accommodate the two counter-rotating beams of  $p$  and  $\bar{p}$ . In the process of this conversion, each beam was designed for an energy of 270 GeV, giving a head-on collision energy of 540 GeV. This, of course, is far greater (over five times) than the thresholds for  $W^\pm$ ,  $Z^0$  production, but this is necessary

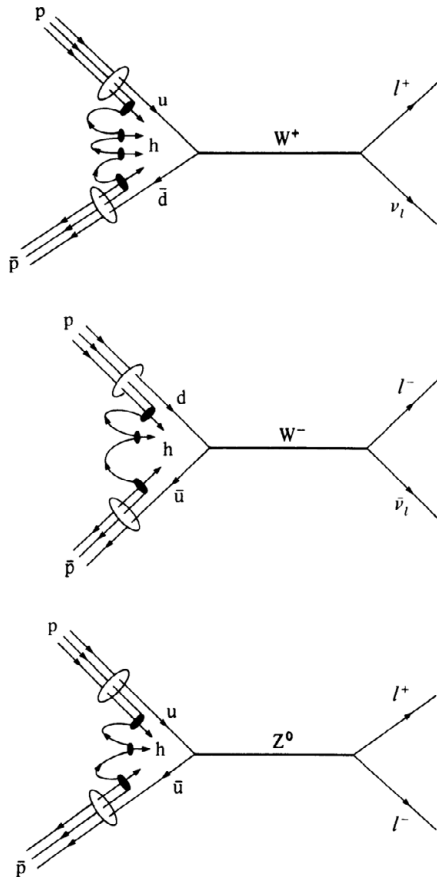


Figure 24.1. The mechanisms for  $W^\pm, Z^0$  boson production in  $p\bar{p}$  collisions. In all cases  $h$  indicates the hadronic debris resulting from the presence of the spectator quarks.

as only a fraction of the energy will go into the quark-antiquark annihilations. The rest stays with the spectator quarks and gives rise to complicated, long-range hadron production.

Although the basic idea behind the experiment is simple, the reality is complicated by the absence of naturally occurring antiprotons. These must be painstakingly manufactured in preparatory particle collisions and stored until a sufficient number have been collected to form a beam of sufficient density (referred to as luminosity) for an observable rate of reactions to be possible. Achievement of this 'antimatter factory' is one of the wonders of modern physics and demonstrates convincingly the sophistication with which the experimentalists are now able to

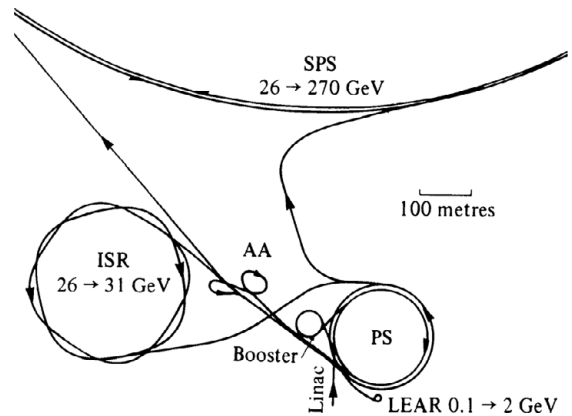


Figure 24.2. High-energy plumbing. Both the proton synchrotron (PS) and the antiproton accumulator (AA) form an integral part of the CERN SPS  $p\bar{p}$  collider experiment. The intersecting storage rings (ISR) conduct separate experiments.

control elementary particle beams, through a veritable 'sphagettiscape' of accelerators, see Figure 24.2.

Initially, protons are accelerated to 26 GeV in the 1959 proton synchrotron (PS) machine and are collided into a tungsten target. From the input of  $10^{13}$  protons, approximately 20 million antiprotons of about 4 GeV energy emerge. These are then piped to the antiproton accumulator where an increasing number of such collision bunches are stored. When a bunch first enters the accumulator it is regulated or 'cooled' to remove all random components of the antiprotons' individual movements. This process is called 'stochastic cooling' and is achieved with a sophisticated control system which detects any such deviation of the antiprotons' movements from the ideal orbit, flashes a signal across the diameter of the accumulator ring and 'kicks' the antiproton bunch back into shape by the application of a tailored magnetic pulse. All this in the time that it takes the antiproton bunch, travelling at practically the speed of light, to travel half-way around the ring! After about two seconds, the antiproton bunch is sufficiently cooled for it to be manoeuvred by magnetic fields into a 'holding' orbit in the accumulator while the next bunch is introduced and cooled, after which it too is manoeuvred to join the holding orbit. Over some two days about 60 000 injections are achieved to form a few orbiting bunches of about  $10^{12}$  antiprotons in each. Eventually, the antiproton bunches are sent back to the PS to be

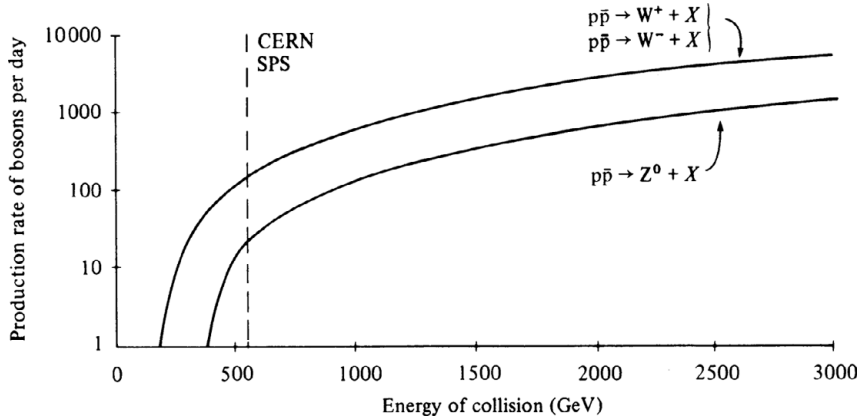


Figure 24.3. The predicted production rate of  $W^\pm$ ,  $Z^0$  bosons as a function of collision energy, assuming the design luminosity of the CERN  $p\bar{p}$  experiment.

accelerated up to 26 GeV, after which they are injected into the SPS. The PS also injects bunches of 26 GeV protons into the SPS in the opposite direction.

The SPS can then accelerate the counter-rotating  $p$  and  $\bar{p}$  bunches each to 270 GeV when they can be collided through each other at specified interaction regions around the SPS, where various experiments observe the interactions. After this, the same  $p\bar{p}$  beam bunches can go on providing interactions over many hours.

Knowing the luminosities of the beam bunches (equivalent to approximately  $10^{30}$  antiprotons per square cm per second) and using the Glashow–Weinberg–Salam theory to calculate the probabilities of occurrence of the reactions producing the  $W^\pm$ ,  $Z^0$  boson allows us to estimate the rate of production of the bosons in the CERN  $p\bar{p}$  collisions. As we can see in Figure 24.3, several hundred were expected each day. The problem then became one of finding the bosons amongst the debris of the collisions.

### 24.3 Detecting the Bosons

Once produced, the  $W^\pm$ ,  $Z^0$  bosons are far too short-lived to leave detectable tracks. As is often the case, their presence must be inferred from the behaviour of their decay products. The most significant of these for the  $W^\pm$ ,  $Z^0$  bosons are the charged leptons arising from the decays:

$$W^+ \rightarrow \mu^+ + \nu_\mu,$$

$$W^- \rightarrow \mu^- + \bar{\nu}_\mu,$$

and

$$Z^0 \rightarrow \mu^+ + \mu^-.$$

Several features of the charged lepton distributions emerging from  $p\bar{p}$  collisions can provide the telltale signs of the  $W^\pm$ ,  $Z^0$  bosons.

Firstly, the presence of the parity-violating effect indicates the action of the weak force. The effect of angular momentum conservation and the unique-handedness of the neutrino and antineutrino requires an excess of positively charged leptons emerging in the direction of the incoming antiproton beam (and, similarly, an excess of negatively charged leptons in the direction of the incoming proton beam). Although this is strong circumstantial evidence for the bosons, strictly speaking it indicates only the presence of the weak force and not specifically mediation via the  $W^\pm$ ,  $Z^0$  bosons.

This more demanding information is indicated by the momenta and energies of the emerging leptons. The production of the  $W^\pm$ ,  $Z^0$  bosons gives rise to a far higher proportion of leptons carrying a significant momentum perpendicular to the axis of the  $p\bar{p}$  collision. By measuring the distribution of this transverse momentum of the emerging leptons, the presence of the bosons is manifestly obvious, see Figure 24.4(a). Also, the decay of the  $Z^0$  boson involves no ‘invisible’ neutrinos to carry off any of the energy and so gives rise to the additional distinguishing feature of a very sharp peak in the mass distribution of emerging lepton–antilepton pairs centred on the mass of the  $Z^0$ , see Figure 24.4(b).

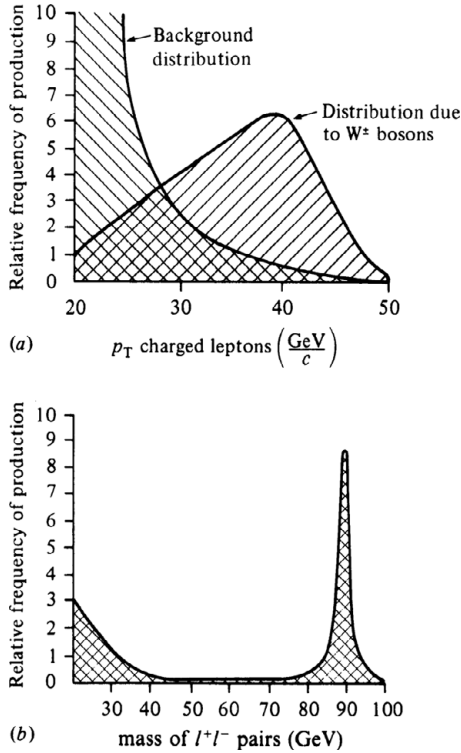


Figure 24.4. The anticipated effects of  $W^\pm, Z^0$  bosons. In (a) we see the increased distribution of leptons emerging with high transverse momentum,  $p_T$ . In (b) the mass spectrum of charged lepton–antilepton pairs shows a peak at the mass of the  $Z^0$ .

During the latter part of 1982, the SPS  $p\bar{p}$  collider achieved a sufficient luminosity to make feasible the discovery of the bosons at the rate of a few events per day. So two separate experiments using different detectors set out to find the bosons. The experiments were referred to as UA1 and UA2, respectively, denoting the different underground areas in which their detectors are located round the SPS ring. Each of the experiment's detectors were very massive assemblies of a variety of particle detection devices (2000 tonnes in the case of the UA1, 200 tonnes in UA2). The output of these detectors was fed directly into online computers, which allowed the subsequent reconstruction and analysis of the tracks of each event, see Figure 24.5(a) and (b).

The experiments observed about  $10^9 p\bar{p}$  collisions of which  $10^6$  were recorded for subsequent analysis. The two experiments then applied various criteria to the observed collisions to find examples of the process.

$$p + \bar{p} \rightarrow W^\pm + X \rightarrow e^\pm + \nu + X.$$

The UA1 experiment sought just two separate classes of event: firstly, those with an isolated electron with large transverse momentum and, secondly, those events with a large fraction of transverse energy missing (carried off by the undetected neutrinos).

Starting from its initial sample of 140 000 events, the UA1 experimental group was able to identify those events containing the isolated electron with large transverse momentum by applying a series of exclusion conditions (such as demanding that the electron track should have originated in the central detector, that other tracks should have low transverse momentum and so on). Eventually, just five isolated electron events were identified.

Then, starting a new search on a sub-set of 2000 of the original events, the group searched for those events with missing energy (i.e. those events containing energetic neutrinos). This is done essentially by adding up the measured energies of all the observed tracks and checking the total against the known collision energy of 540 GeV. After another set of exclusion conditions just seven events were found, five of which were just those events with the isolated, high-energy electrons. So the UA1 experimental group was able to claim the discovery of five  $W^\pm$  bosons. Furthermore, by adding up the measured energy of the electron track and the missing energy of the inferred neutrino track, the UA1 group was able to estimate the mass of the  $W^\pm$  bosons from which these two particles had originated. The result was in excellent agreement with the prediction of the Glashow–Weinberg–Salam model.

The UA2 experimental group was able to perform a similar analysis of its sample of  $p\bar{p}$  interactions, and was able to identify four candidate  $W^\pm$  boson events and give an estimate of the mass which was also in good agreement with the prediction.

These discoveries of the  $W^\pm$  bosons were announced in January 1983. On 1 June of the same year, the two experimental groups announced the discovery of the  $Z^0$  boson, which was first identified in the process.

$$p + \bar{p} \rightarrow Z^0 + X \rightarrow e^+ + e^- + X.$$

This identification followed from another set of exclusion criteria in which the  $e^+e^-$  pair emerged back-to-back with equal and opposite high transverse momenta.

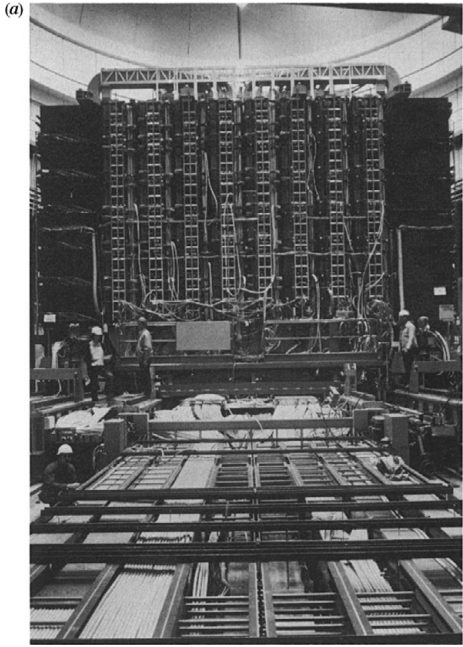
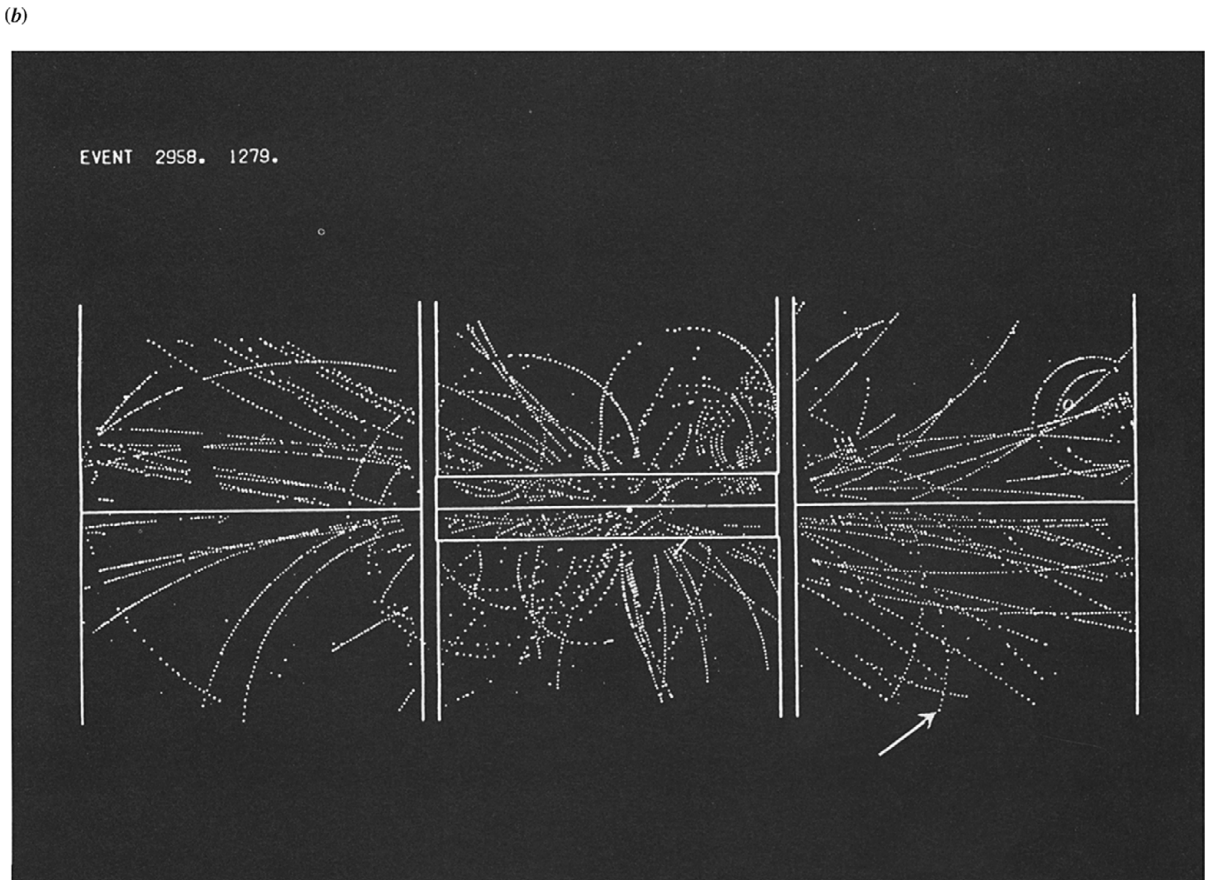


Figure 24.5. (a) The 2000 tonne UA1 detector at the CERN proton–antiproton collider. (Photo courtesy CERN.) (b) Particle tracks heralding the discovery of the W boson in the UA1 detector. An electron with high transverse momentum (arrowed) emerges from the interaction point, and missing energy betrays the escape of an invisible neutrino. (Photo courtesy CERN.)



Carlo Rubbia, the driving force behind the experimental effort, and Simon van der Meer, the inventor of stochastic cooling, were jointly awarded the 1984 Nobel Prize in recognition of the major roles they played in these discoveries, which brought to a climax a decade of successful experiments verifying the gauge theory framework of the Glashow–Weinberg–Salam electroweak model.

#### **24.4 Epilogue**

Since 1983, the massive electroweak gauge bosons have been produced in much greater numbers

both at the CERN  $p\bar{p}$  collider and also at the ‘Tevatron’  $p\bar{p}$  collider at Fermilab in Illinois. In 1989, two  $e^+e^-$  machines – LEP at CERN and SLC at Stanford – came into operation and began producing huge numbers of  $Z^0$  bosons. This has permitted an accurate mass determination, and enabled physicists to refine their measurements of the width of the  $Z^0$  peak (Figure 24.4(b)), hence establishing the particle’s lifetime. These measurements had important implications for the total number of neutrino species, which in October 1989 was finally narrowed down to precisely three.



## **Part VII**

# **Deep Inelastic Scattering**





## Deep Inelastic Processes

### 25.1 Introduction

Among the most important experiments of the last 50 years have been those which use the known interactions of the leptons to probe the structure of the nucleons. Their importance lies in the fact that they provided the first dynamical evidence for the existence of quarks, as opposed to the static evidence provided by the success of the internal symmetry scheme  $SU(3)$ .

The term *deep inelastic scattering* arises because the nucleon which is probed in the reaction nearly always disintegrates as a result. This is obvious from the momentum–wavelength relation for particle waves:

$$p\lambda = h.$$

The proton is approximately  $10^{-15}$  m in diameter and so to resolve any structure within this requires that the probing particle wave has a smaller wavelength. The formula then gives the required momentum of the probe as being greater than 1 GeV/c, under the impact of which the target nucleon is likely to disintegrate.

Deep inelastic experiments divide into two classes, depending on the nature of the probe used, which in turn dictates the force involved. In electroproduction, electrons or muons are scattered off the target nucleon and the force involved is electromagnetic. The leading process of the scattering is that of single-photon exchange, which is assumed to be a sufficiently good description of the interaction (Figure 25.1(a)) although, in principle, more-complicated

multi-photon processes may become significant as the energy of the collision becomes very large. The second class of experiment is called neutrino production and, in this, neutrinos are scattered off the target nucleon by the weak nuclear force. The leading process is that of single-W-boson exchange, other more complicated processes being insignificant. Both charged and neutral currents may contribute, see Figure 25.1(b), (c), but in practice it is the better-understood charged currents which are used in experiments. Indeed, this was necessarily the case in the early experiments, 1967–73, as the neutral currents had not then been discovered.

The main measurement of the experiments is the variation of the cross-section (the effective target area of the nucleon) with the energy lost by the lepton during the collision and with the angle through which the incident lepton is scattered. The energy lost by the lepton  $\nu$  is simply the difference between its incident and final energy:

$$\nu = E_i - E_f.$$

The angle through which the lepton is scattered is related to the square of the momentum transferred by the photon  $q^2$  from the lepton to the nucleon by the formula

$$q^2 = 2E_i E_f (1 - \cos \theta). \quad (25.1)$$

These are the two main observables in deep inelastic scattering, which connect the data from experiments with our theoretical picture of the nucleon interior.

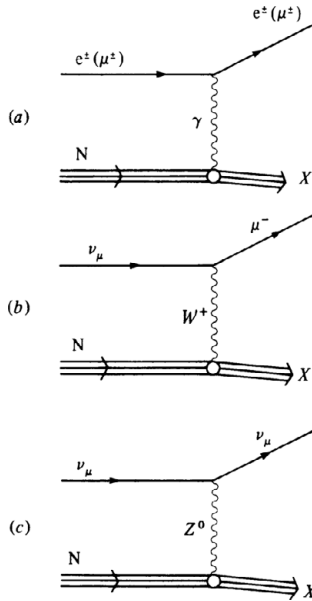


Figure 25.1. (a) Electroproduction via photon exchange. (b) Charged current neutrino production via  $W^\pm$  exchange. (c) Neutral current neutrino production via  $Z^0$  exchange.

### 25.2 Two Key Ideas

Two ideas in particular played an important role in the development of the experiments and in our understanding of them. The two ideas, both put forward in 1969, are those of the parton model and of scaling.

The *parton model* was first put forward by Richard Feynman and is simply a formal statement of the notion that the nucleon is made up of smaller constituents: the partons. No initial assumptions about the partons are necessary, as it is the purpose of the experiments to determine their nature. But obviously we have at the back of our minds the identification of the partons with the quarks of  $SU(3)$ . However, we should not jump to the conclusion that *only* the familiar quarks will be sufficient to describe the composition of the nucleon. For instance, in addition to the proton's three quarks which are required by the internal symmetry scheme (the so-called *valence* quarks), it may be possible for virtual quark–antiquark pairs to

emerge briefly from the vacuum by borrowing energy according to Heisenberg's uncertainty principle. These *sea* quarks may then form an additional material presence within the nucleon and provide a mechanism for the existence, albeit transient, of antimatter inside a 'matter' particle. Because they emerge in quark–antiquark pairs, the sea quarks will have no net effect on the quantum numbers of the nucleon, which are determined by the valence quarks. In addition to the quarks, we may be alert to the fact that quanta of the interquark force field may also be present inside the proton. Just as electrons interact by the exchange of photons, the quanta of the electromagnetic force, so quarks may interact by the exchange of quanta of their force field. These quanta have been called, rather simplistically, gluons, because they glue the quarks together.

*Scaling* is the name given to a phenomenon first predicted by the Stanford physicist James Bjorken. Stated simply, the prediction is that when the momentum carried by the probe becomes very large, then the dependence of the cross-section on parameters such as the energy  $\nu$  and momentum-squared  $q^2$ , transferred by the photon, becomes very simple. In the parton model, the onset of this simple scattering behaviour has a straightforward interpretation. The complicated scattering of the probe off a nucleon of finite spatial extent is, at high momentum, replaced by the scattering of the probe off a point-like parton. The photon ceases to scatter off the nucleon *as a coherent object* and, instead, scatters off the individual point-like partons *incoherently*. We should expect this sort of behaviour to manifest itself when the wavelength of the probe is much less than the nucleon diameter, implying a probe momentum above about 1 GeV.

Observation of this scaling behaviour in 1969 immediately lent support to the parton model of the nucleon, although, as we shall see, the initial discovery was somewhat fortuitous. To understand the concepts of scaling and the parton model further, we must take a more detailed look at the processes involved. The importance of these ideas was acknowledged by the award of the 1990 Nobel Prize to the pioneers of these deep inelastic experiments, Jerome Friedman, Henry Kendall and Richard Taylor.

## *Electron–Nucleon Scattering*

### 26.1 Introduction

Assuming that the electromagnetic interaction between the electron and the nucleon is dominated by the single-photon exchange mechanism, then the mathematics used to describe the reaction becomes relatively simple. To check the experimental observations, we want to derive a formula to explain how the cross-section varies with the energy transfer  $\nu$  and momentum transfer squared  $q^2$  of the intermediate photon. The formula is made up of factors associated with the different parts of the diagram in Figure 25.1(a). It consists of a factor describing the progress of the electron through the reaction (the lepton current), a factor describing the propagation of the virtual photon as a function of  $\nu$  and  $q^2$ , and a factor describing the flow of the nucleon in the reaction including the complicated disintegration process (the hadron current). The factors describing the electron and the photon are well known from QED and present us with no problems. But the factor describing the hadron current is a complicated unknown, describing the evolution of nucleon structure during the reaction. This unknown can be characterised by a number of ‘structure functions’ of which we assume no prior knowledge and which are to be determined by the deep inelastic experiments, see Figure 26.1.

The form of the structure functions is discovered by writing down the most general possible combinations of all the momenta appearing in the reaction and

then simplifying the result using general theoretical principles such as parity and time-reversal invariance.

The two separate functions of  $q^2$  and  $\nu$  that result –  $F_1(q^2, \nu)$  and  $F_2(q^2, \nu)$  – correspond to the scattering of the two possible polarisation states of the virtual photon exchanged: longitudinal and transverse. The longitudinal polarisation state exists only because of the virtual nature of the exchanged photon (because it temporarily has a mass). The virtual photon is ‘off-mass-shell’, meaning that  $E = pc$  is violated, and implying that its mass is non-zero. On the mass shell when the virtual photon becomes real (massless), then the longitudinal polarisation state and its associated structure function disappear. The separate behaviour of the two structure functions can be determined from experiments because they are multiplied by coefficients involving different functions of the electron scattering angle. By observing the reaction at different values of this angle, the two behaviours can be separated out.

### 26.2 The Scaling Hypothesis

The scaling hypothesis mentioned previously is to do with these structure functions. It is important to realise that they are just numbers and have no physical dimension. The cross-section is usually given in units of area which are provided by the simple Rutherford scattering formula for elastic scattering. This has deep implications for the behaviour of the structure functions. If they are to have any dependence on

$$\frac{d^2\sigma}{dq^2 dv} = \frac{4\pi\alpha^2}{q^4} \frac{E_f}{E_i M_p} \left[ \frac{M_p}{v} F_2(q^2, v) \cos^2 \frac{\theta}{2} + 2F_1(q^2, v) \sin^2 \frac{\theta}{2} \right]$$

Figure 26.1. The formula describing the differential cross-section for electron–nucleon scattering with respect to the momentum transfer squared  $q^2$  and the energy lost by the electron  $v$ . The structure functions  $F_1$  and  $F_2$  essentially describe the shape of the nucleon target.

physically dimensional quantities such as the energy  $v$  or momentum-squared  $q^2$  involved in the reaction, then these factors must have their physical dimensionality cancelled out to give structure functions in terms of pure numbers.

In low-energy elastic scattering (corresponding to  $q^2 = 2vM_N$ ), the photon effectively perceives the nucleon as a single extended object and the structure function essentially describes the spatial distribution of electrical charge on the nucleon. This leads to a dependence of the structure function on the momentum of the photon – but the dimensionality of the momentum in the structure function is cancelled out by factors of the nucleon mass:

$$\frac{d\sigma}{dq^2} = \frac{4\pi\alpha^2}{q^4} \cdot F\left(\frac{q^2}{M_N^2}\right),$$

i.e.

$$\text{cross-section} = \text{unit of area} \times \text{pure number.}$$

To signify this cancellation we say that the nucleon mass sets the ‘scale’ of reaction. It provides a scale against which the effect of the photon momentum can be measured.

By contrast, in very high-energy, deep inelastic scattering (i.e.  $q^2, v \rightarrow \infty$ ), the wavelength of the photon is so small that the existence of the complete nucleon is really irrelevant to the reaction: the photon interacts with only a small part of the nucleon and does so independently of the rest of it. This means that there is no justification for using the nucleon mass to determine the scale of the reaction. In fact, there is no justification for using any known mass or any other physically dimensional quantity to determine the scale of the deep inelastic regime. James Björken grasped the consequences of this abstract argument: if the structure functions are to reflect the dependence of the cross-section on the shape of the nucleon as seen by a photon of very high  $q^2$  and  $v$ , and if there exists

no mass scale to cancel out the physical dimensions of these quantities, then the structure functions can only depend on some dimensionless ratio of the two. Choosing such a ratio as  $x$ ,

$$x = \frac{q^2}{2M_N v},$$

then the scaling hypothesis is that the structure functions can depend only on it, and not on either or both of the quantities involved separately. So, as  $q^2, v \rightarrow \infty$ ,

$$F_{1,2}^{eN}(q^2, v) \xrightarrow{q^2, v \rightarrow \infty} F_{1,2}^{eN}(x).$$

The scaling hypothesis becomes rather more accessible when it is combined with the parton model in which the nucleon is regarded as a simple collection of point-like constituents. A point has no dimension, and we are considering the scattering of a photon carrying infinitely high momentum (i.e. one which has a vanishingly small wavelength). In this situation, there are simply no physically significant quantities which are relevant to set the scale of the reaction. So quantities such as the energy and momentum transfer squared in the reaction can only enter into its description in the form of pure numbers, which in turn implies a dimensionless ratio of the two.

Björken’s choice of the ‘scaling’ variable  $x$  has a very significant interpretation. It turns out to be the fraction of the momentum of the nucleon carried by the parton which is struck by the photon. So the structure functions, which depend only on  $x$ , effectively measure the way in which the nucleon momentum is distributed amongst its constituent partons.

Figure 26.2(a) shows how the structure function  $F_2^{ep}(x)$  varies with  $x$ , as measured in early experiments at the Stanford Linear Accelerator Center. As can be seen, the shape implies that the majority of collisions occur with partons carrying a relatively small fraction of the nucleon momentum. Figure 26.2(b) shows a test

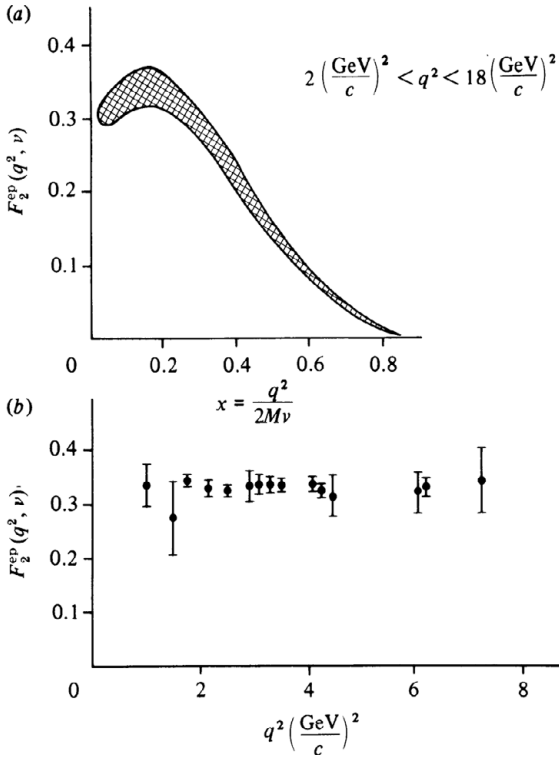


Figure 26.2. Early data on scaling. In (a),  $F_2(q^2, \nu)$  is a universal function of  $x$  for a range of values of  $q^2$  (many experimental points are contained in the shaded area). In (b),  $F_2(q^2, \nu)$  demonstrates approximate constancy over the range of  $q^2$  measured in the early experiments.

of the scaling hypothesis that the structure function depends only on  $x$ , and not on  $q^2$  (or  $\nu$ ) separately.

At a given value of  $x$ , the structure function is found to be almost constant over a  $q^2$  range from 1 to 8  $\text{GeV}^2$ . Data such as this became available in the early 1970s. The apparent validity of the scaling hypothesis and the plausibility of its connection with the existence of point-like constituents within the nucleon led to a more detailed investigation of the structure functions to establish more information about the mysterious partons.

### 26.3 Exploring the Structure Functions

One straightforward exercise is to compare the formula for electron–proton scattering with formulae derived from QED describing the electromagnetic interactions of electrons with other simple, electrically

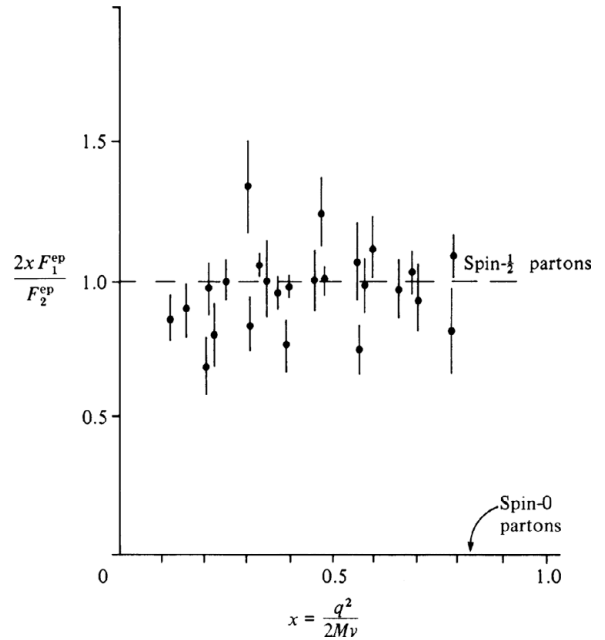


Figure 26.3. The ratio of the structure functions provides a test of the spin- $\frac{1}{2}$  assignment to the partons.

charged particles whose properties, such as spin, may be assumed. In this way it is possible to derive relationships between the structure functions depending on the similarities between the properties of the partons and those of the hypothetical particles assumed in the QED formulae. For instance, by comparing the formula of Figure 26.1 with the QED formula describing the scattering of an electron with an electrically charged particle of spin  $\frac{1}{2}$ , it is possible to derive the relationship between the structure functions:

$$2xF_1(x) = F_2(x).$$

Thus if, experimentally, the ratio  $2xF_1/F_2$  is found to be equal to one, this may be interpreted as evidence for the partons having spin  $\frac{1}{2}$ . Using similar formulae it is possible to show that if the ratio is observed to be zero, then this provides evidence for spin-0 partons. As can be seen from Figure 26.3, the evidence is firmly in favour of spin- $\frac{1}{2}$  partons.

Another lesson to be learnt by comparing the deep inelastic formula with the simpler formula from QED is that the structure functions essentially measure the distribution of electric charge within the nucleon. In low-energy, non-relativistic physics, it is acceptable

to speak of the distribution of charge over the spatial extent of the proton. But in high-energy physics this is better expressed in terms of the conjugate description of how the nucleon momentum is distributed amongst partons of various charges. If we say that the  $i$ th type of parton with charge  $Q_i$  has probability  $f_i(x)$  of carrying a fraction  $x$  of the nucleon momentum, then it is possible to relate the overall structure functions of the nucleon to these individual parton momentum distributions,

$$\begin{aligned} F_1^{\text{eN}}(x) &= \sum_{i \text{ partons}} f_i(x) Q_i^2, \\ F_2^{\text{eN}}(x) &= x \sum_{i \text{ partons}} f_i(x) Q_i^2. \end{aligned} \quad (26.1)$$

Obviously, if we integrate the structure functions over the fractional momentum carried by the partons, we should then expect to provide some measure of the total charge carried by the partons. In fact the most convenient relationship involves the sum of the squares of the parton charges:

$$\int_0^1 \frac{F_2(x)}{x} dx = \sum_i Q_i^2.$$

So, by investigation of relationships involving the structure functions, it is possible to find evidence for the spin of the partons and their charge assignments, which we will investigate further in Chapter 29.

## *The Deep Inelastic Microscope*

### 27.1 Introduction

The physics behind the approach to scaling can be appreciated intuitively by regarding deep inelastic scattering as an extension of the ordinary microscope, whose successor the experiments quite literally are. The distance to which an ordinary microscope can resolve depends ultimately on the wavelength of the light scattered from the object under view. The shorter the wavelength, the smaller the distances which can be resolved. The high-energy photon exchanged between the electron and the nucleon in deep inelastic experiments is simply the logical development of the microscope technique. The origin of the scaling phenomenon becomes clear by considering a succession of snapshots of a virtual photon–nucleon collision.

When the momentum carried by the photon is low, its wavelength is relatively long compared with the dimensions of the nucleon. It will not be able to resolve any structure and will effectively see the nucleon as a point. In this case, structure functions are irrelevant, being represented simply by the nucleon charge in the Rutherford formula for the scattering of two point charges, see Figure 27.1(a).

With higher momentum, the photon will have a wavelength comparable to that of the nucleon. The photon will begin to resolve the finite spatial extent of the nucleon and the structure functions will depend in a non-trivial fashion on the momentum carried by the photon, modifying the Rutherford scattering cross-section (Figure 27.1(b)).

Eventually, with high-momentum transfer the photon will have a very short wavelength and may resolve the internal structure of the nucleon. Scattering off point-like constituents within the nucleon is indicated when the structure function assumes a simple dependence on some dimensionless ‘scaling variable’. In fact, under certain circumstances the structure functions are simply replaced by a constant equal to the sum of the squares of the charges of the constituents, which multiplies the simple Rutherford cross-section (Figure 27.1(c)). This reversion to the simplicity of point-like scattering after a relatively more complicated transitional phase is taken to indicate that we have broken through to a new, more basic level of matter within the nucleon.

### 27.2 Free Quarks and Strong Forces

An essential consequence of the validity of the scaling hypothesis, in which the deep inelastic probe scatters *incoherently* off the individual partons, is that the partons are essentially free from mutual interactions over the space–time distances of the probe–parton interaction. This has important consequences for the nature of interparton forces.

To understand this more fully, we can associate one interaction time with the duration of the probe–parton interaction,  $\tau_1$ , and another time to characterise the duration of interquark forces,  $\tau_2$ . Obviously, for the nucleon to have any sort of collective identity, the nucleon must exist for  $\tau_{\text{life}} > \tau_2$ , so that the partons



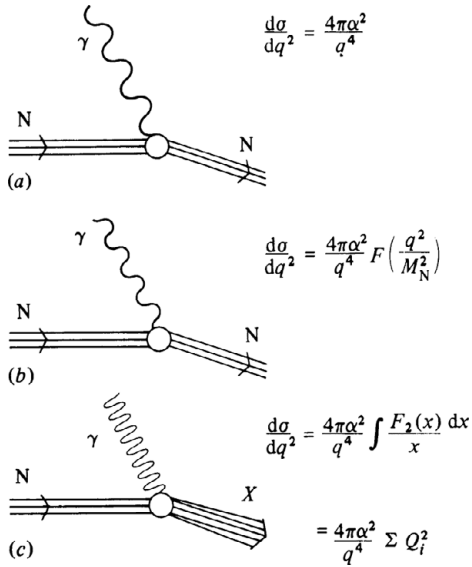


Figure 27.1. An illustration of the approach to scaling as the virtual photon wavelength becomes much smaller than the nucleon diameter. The wavelengths become progressively smaller in (a), (b) and (c).

can become aware of each other's presence. As a rough guide we may estimate the interaction times by the following prescriptions:

$$\tau_1 = \frac{\text{wavelength of probe}}{\text{speed of light}}, \approx \frac{\hbar}{v},$$

$$\tau_2 = \frac{\text{interquark distance}}{\text{speed of light}} < \tau_{\text{life}}.$$

If  $\tau_2 < \tau_1$ , then the interparton forces will have transmitted the effects of the probe collision to all the partons inside the nucleon within the probe's interaction time. In this case, the probe will not be scattering off the individual partons but off the entire nucleon instead. However, in deep inelastic scattering in which the probe wavelength is very small,  $\tau_1 > \tau_2$ , and therefore the probe interaction is completed well before the interparton forces have had time to relay the event to the rest of the nucleon. So the probe-parton interaction occurs well within the lifetime of the nucleon.

For a short time, subsequently, the nucleon exists in an uncomfortable state: one of its partons has been struck hard and flies off with the high momentum imparted by the probe, but the other partons know nothing of this and continue to exist in a quiescent

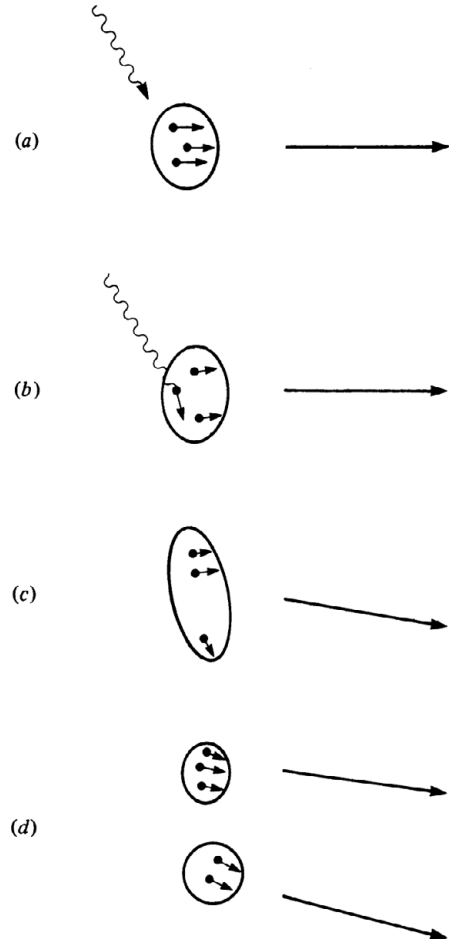


Figure 27.2. When a deep inelastic probe strikes a parton (a) and (b), it flies off with large momentum (c) until some confining mechanism pulls extra parton-antiparton pairs from the vacuum, creating new particles (d).

state, see Figure 27.2. This situation cannot last long, otherwise the struck parton would eventually appear as an isolated particle separated from the rest. This has never been observed. Instead, it is thought that final-state interactions come into effect between the struck parton and the others, turning the energy of the collision into the observed hadrons (Figure 27.2(c), (d)).

An important feature of the parton model is the assumption that the cross-section for deep inelastic scattering can be calculated simply by summing over the individual probe-parton interactions and that the complicated final-state interactions become

important only over longer space–time distances (i.e. greater spatial separations and longer interaction times).

If we think particularly in terms of the strong interactions between quarks, the overall picture emerging from the success of the parton model is that of a force whose strength varies with distance. At the short distances probed by the deep inelastic reactions (say  $10^{-17}$  m) the interquark force is weak and the quarks are essentially free. But as the distance between two quarks grows to the nucleon diameter ( $10^{-15}$  m), the force also grows, confining the quarks, perhaps permanently, within the observed hadrons. Eventually, any energy expended in trying to separate the quarks

will be sufficient to pull a new quark–antiquark pair from the vacuum, so allowing the production of a new hadron, but preventing the emergence of the individual quarks.

The tendency of the interquark forces to become weaker at small distances is known as *asymptotic freedom*. As we shall see in Part VIII, the discovery that this property can be explained naturally in non-Abelian gauge theory was a significant breakthrough in our attempt to understand quark dynamics. The other tendency of the interquark forces, to become increasingly strong as the quarks are separated, is known as confinement or, more colloquially, infrared slavery.

## *Neutrino–Nucleon*

### 28.1 Introduction

Just as in electron– or muon–nucleon scattering the exchanged photon acts as a probe of the electromagnetic structure of the nucleon, so in neutrino (or antineutrino)–nucleon scattering, the exchanged W-boson probes the distribution of ‘weak charges’ within the nucleon. For this purpose, the two most important processes are the charged current inclusive reactions (Figure 25.1). An important feature of these reactions is that they are able to distinguish between the partons and the antipartons of the target nucleon. This is because the space–time structure of the weak interaction ensures that target partons of differing helicities are affected differently. In the relativistic limit, in which the rest mass of a particle is regarded as being negligible, the parton and antiparton helicities are opposite, so they will interact with the W-boson probe differently. Also, because the W-boson probe is electrically charged, the target parton must be able to absorb the charge. As we shall see, this rules out the participation of some types of parton, making the weak interaction a more selective probe of the nucleon’s interior than the indiscriminate photon.

### 28.2 Neutrino Experiments

Although these weak interaction experiments are theoretically more illuminating than their corresponding electromagnetic counterparts, the practical difficulties of dealing with neutrinos tend to spoil their potential.

The electrically neutral, weakly interacting neutrinos cannot be directed by electric and magnetic fields, as can electrons; and building a usable neutrino beam is a complicated process. Firstly, a primary beam of protons is accelerated to a high energy and is made to collide with a stationary target such as a piece of iron. From these collisions, a host of secondary particles, mainly mesons, will emerge in the general direction of the incident proton beam, although with somewhat less energies. These secondary mesons can then decay into neutrinos or antineutrinos and various other particles by decays such as

$$\pi^{\pm} \rightarrow \mu^{\pm} + \nu_{\mu} \text{ (or } \bar{\nu}_{\mu}\text{)}.$$

Because the muon-decay mode of the mesons is generally the most common, it is mainly muon-type neutrinos which make up the beam. Finally, the neutrinos are isolated by guiding the secondary beam through a barrier of, perhaps, 0.5 km of earth. Only the weakly interacting neutrinos can pass through this amount of matter and so the beam emerging from the far side of the barrier is a pure neutrino beam with a typical intensity of about  $10^9$  particles per  $\text{cm}^2$  per second. Unfortunately, the initial proton-target collisions and the subsequent decays of the secondary mesons mean that the resulting energy of the neutrino beam is rather uncertain and often must be inferred by adding up the energies of the products of the neutrino–nucleon interactions under study.

$$\frac{d^2\sigma}{dq^2 dv} = \frac{G_F^2 E_\mu}{2\pi E_{\nu(\bar{\nu})}} \left[ 2F_1^W(q^2, v) \sin^2 \frac{\theta_\mu}{2} + F_2^W(q^2, v) \cos^2 \frac{\theta_\mu}{2} \pm F_3^W(q^2, v) \frac{(E_\mu + E_{\nu(\bar{\nu})})}{M_N} \sin^2 \frac{\theta_\mu}{2} \right]$$

Figure 28.1. The formula describing the differential cross-section for (anti)neutrino–nucleon scattering with respect to momentum transfer squared  $q^2$  and energy lost by the neutrino  $v$ . Three structure functions  $F_1^W$ ,  $F_2^W$  and  $F_3^W$  (functions of  $q^2$  and  $v$  in general) are needed to describe the shape of the nucleon target (i.e. the way in which weak charge is distributed over the nucleon momentum).

To obtain a reasonable rate of interactions, the neutrino beam must be passed through a very massive target. For instance, the Gargamelle bubble chamber at CERN contains about ten tonnes of some heavy liquid such as freon to ensure a satisfactory rate of reactions. Because the beam consists of both neutrinos and antineutrinos, both sorts of reaction will occur during the same experiment. The two can be distinguished by observing the charge of the outgoing muon in any particular reaction.

### 28.3 The Cross-section

For  $\nu_\mu N$  scattering, the cross-section may be written down in a fashion similar to that used for  $e^\pm N$  scattering, by combining various factors describing the different sub-processes which go to make up the collision. Referring back to Figure 25.1, we can see that these factors must include one to describe the transformation of the incoming neutrino into a muon by emitting the W boson (the lepton current); one describing the propagation of the W boson; and one describing the disintegration of the target nucleon under the impact of the W boson (the hadron current). Also analogous to the case of  $e^\pm N$  scattering is that the behaviour of the factors is well known apart from that describing the hadron current. The lepton current and the propagation of the W boson are well known from the gauge theory of the weak interactions – which is well approximated by the simpler Fermi theory at these energies. But, as before, the unknown form of the hadron current must be characterised by a number of structure functions whose nature is the job of the experiments to discover, see Figure 28.1.

The format of the structure functions is found as before by writing down all the possible combinations of momenta involved in the reaction and then appealing to general principles to simplify the result. In contrast to the electromagnetic force, the weak

force does not respect parity invariance and so this simplifying influence cannot be used in neutrino–nucleon scattering. Because of this, a third weak structure function is introduced ( $F_3^W$ ) which enters with a different sign in the formula, depending on whether neutrinos or antineutrinos are being scattered. This is the manifestation of the effect mentioned earlier by which the parity-violating weak interaction will distinguish between matter and antimatter involved in the reactions as a result of helicity effects. In general, the structure functions all depend on  $q^2$  and  $v$  separately. It is interesting to note that it is *only* through the structure functions that these quantities enter the description of the reactions at all.

### 28.4 The Scaling Hypothesis

Although first described in connection with the electromagnetic interactions, the scaling hypothesis is equally valid for the weak force. In this case, all the dimensionality of the cross-section is contained within the Fermi coupling constant  $G_F$  (remember that this was the trouble which provided one of the major motivations for the development of weak interaction field theory). As a result, the structure functions must be pure numbers. In the absence of any ‘scaling factor’ to cancel out the dimensionality of  $q^2$  and  $v$ , the structure functions  $F_{1,2,3}^W$  cannot depend on them as individual quantities, but only on some dimensionless ratio of the two:

$$F_{1,2,3}^W(q^2, v) \xrightarrow{q^2, v \rightarrow \infty} F_{1,2,3}^W(x),$$

where the ratio  $x$  is the same as before. The structure functions can be measured directly as in  $e^\pm N$  scattering and the scaling hypothesis tested. The general shape of the weak structure functions is much the same as that of the electromagnetic example illustrated in Figure 26.2, but because the parameters of the neutrino beam are that much more uncertain than of

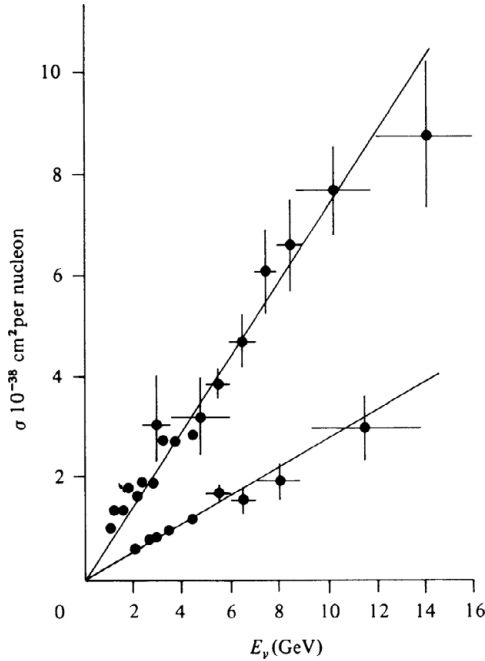


Figure 28.2. Total neutrino–nucleon and antineutrino–nucleon cross-sections plotted against energy support the scaling hypothesis by exhibiting a linear increase.

the electron or muon beam, the experimental errors are much larger, thereby providing a weaker test of the scaling hypothesis.

However, the scaling hypothesis does predict that a far more obvious characteristic will hold in neutrino–nucleon scattering. As mentioned earlier, the cross-section does not depend on  $q^2$  or  $\nu$  apart from through the structure functions. If, then, this dependence is removed by the scaling hypothesis, it means that the cross-section will not depend on these quantities at all. In this case, it is possible to integrate the formula for the cross-section over all possible values of  $q^2$  and  $\nu$  in a very simple fashion to obtain the total cross-section for neutrino– or antineutrino–nucleon scattering:

$$\sigma^{v(\bar{v})N} = \int \frac{d^2\sigma}{dq^2 d\nu} dq^2 \alpha \frac{G_F^2 M E_{\nu(\bar{\nu})}}{\pi}.$$

The resulting scaling prediction for neutrino–nucleon scattering is that the total cross-section should rise linearly with the energy of the incident neutrino. The slope of the rise is given by constants which will be different for neutrino or antineutrino reactions,

because of the different signs in front of  $F_3$  in the formula of Figure 28.1. Experimental measurements of the total cross-sections are consistent with the linear rise with energy predicted by the scaling hypothesis and its interpretation in terms of point-like partons carrying both electric and weak charges, see Figure 28.2.

The difference in the slopes of the energy dependencies of  $\nu$  and  $\bar{\nu}$  scattering is measured to be about a factor of 3. This factor can be easily understood in terms of the underlying neutrino–parton interactions. Recall that neutrinos are exclusively left-handed and antineutrinos are exclusively right-handed (a situation which is possible only because they are massless). Moreover, inasmuch as the mass of the partons may be neglected (i.e. in the relativistic limit), they too are solely left-handed (assuming spin- $\frac{1}{2}$  partons). This follows from the fact that in the Glashow–Weinberg–Salam theory, only left-handed fermions take part in charged current weak interactions (see Part VI). Because partons predominate over antipartons in the nucleon, the neutrino– and antineutrino–parton collisions can be distinguished by the way in which the spins add up. Consider Figure 28.3. If the two colliding spins cancel each other out, as in the case of neutrino–parton scattering, then no restrictions are placed on the angles of emergence of the outgoing particles. If, on the other hand, the two colliding spins add up, as in the case of antineutrino–parton scattering, then the existence of non-zero angular momentum in the system restricts the permitted angles of emergence of the outgoing particles. This means that the cross-section of antineutrino–parton scattering is reduced relative to that of neutrino–parton scattering. This is because the integration over  $q^2$  to obtain the total cross-section is equivalent to summing over all possible angles of emergence (bearing in mind the definition (25.1)) which are more restricted when the angular momentum is non-zero. The mathematics predicts a factor of three between  $\nu N$  and  $\bar{\nu} N$  scattering just as observed.

Neutrino–nucleon scattering provides an independent test of the scaling hypothesis and of the parton model. What is now possible is the comparison of muon–nucleon and neutrino–nucleon scattering to establish that the electromagnetic and weak interactions ‘see’ the same partons.

Bearing in mind that we believe the electromagnetic and weak interactions to be just different

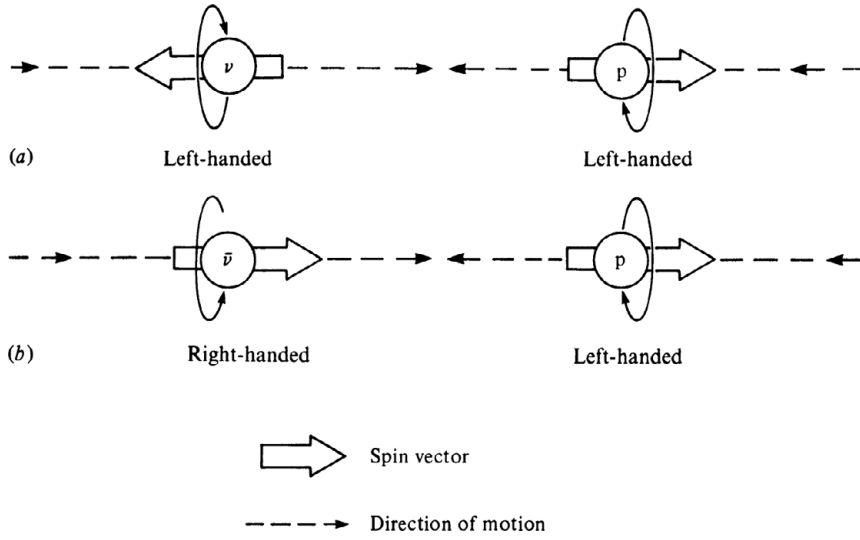


Figure 28.3. (a) Neutrino–parton scattering: spins cancel. There are no restrictions on the directions of the emerging particles. (b) Antineutrino–parton scattering: spins add. Restrictions limit the angles of emergence of the outgoing particles.

manifestations of the same ‘electroweak’ force, we certainly expect that this should be the case. Also, we

should like to compare the properties of the partons with those of the quarks of  $SU(3)$ .

## The Quark Model of the Structure Functions

### 29.1 Introduction

To be more specific about the content of the nucleon (i.e. about what makes up the structure functions), the approach adopted is to assume that the partons have the same properties as the quarks and then to work out the consequences for electron–nucleon and neutrino–nucleon scattering independently. Following this, it is possible to compare the results obtained to see if they are compatible. So we examine, in turn, the electron’s eye view and the neutrino’s eye view of the structure functions.

### 29.2 Electromagnetic Structure Functions

We have mentioned earlier the interpretation of the structure functions as the distribution of the squares of the parton charges within the nucleon, according to the fraction of momentum  $x$  carried by the parton, given by formula (26.1):

$$F_1^{\text{eN}}(x) = \sum_i f_i(x) Q_i^2,$$

$$F_2^{\text{eN}}(x) = x \sum_i f_i(x) Q_i^2.$$

In the simple four-flavour quark model, the quarks and their charges are:

$$u \left( \frac{2}{3}e \right), d \left( -\frac{1}{3}e \right), s \left( -\frac{1}{3}e \right), c \left( \frac{2}{3}e \right),$$

where  $c$  denotes the fourth charmed quark of the GIM scheme with two-thirds of the charge on the electron.

Assuming these values, we can write out the explicit quark content of the structure functions for both the proton and the neutron. In doing so, we must allow for the presence of both quarks and antiquarks from the vacuum sea. This gives, for the proton:

$$F_1^{\text{ep}}(x) = \left[ \left( \frac{2}{3} \right)^2 (f_u(x) + f_{\bar{u}}(x)) + \left( \frac{1}{3} \right)^2 (f_d(x) + f_{\bar{d}}(x)) + \left( \frac{1}{3} \right)^2 (f_s(x) + f_{\bar{s}}(x)) + \left( \frac{2}{3} \right)^2 (f_c(x) + f_{\bar{c}}(x)) \right].$$

The expression for the second structure function  $F_2^{\text{ep}}(x)$  is just the same as above, only multiplied by  $x$ , and the corresponding expressions for the neutron structure functions,  $F_1^{\text{en}}(x)$  and  $F_2^{\text{en}}(x)$ , can be obtained by interchanging  $f_u(x) \leftrightarrow f_d(x)$  in the above expressions, as the distribution of up quarks inside the proton is equivalent to the distribution of down quarks inside the neutron.

The total fractional momentum carried by any particular sort of quark can be obtained simply by integrating over its momentum distribution. For instance, the total share of the momentum carried by the up quarks and antiquarks in the proton is given by:

$$P_u = \int_0^1 x (f_u(x) + f_{\bar{u}}(x)) dx.$$

Similar expressions will obtain for other varieties of quark. The integrals involved are all contained within

the integrals over the total structure functions which are measured in the experiments as the area under the distribution of Figure 26.2(a). This quantity gives a linear combination of the momentum shares of all the possible constituent quarks:

$$\begin{aligned} \int dx F_2^{\text{ep}}(x) &= \left( \frac{4}{9}P_u + \frac{1}{9}P_d + \frac{1}{9}P_s + \frac{4}{9}P_c \right) \\ &= 0.18 \text{ (experiment),} \\ \int dx F_2^{\text{ep}}(x) &= \left( \frac{1}{9}P_u + \frac{4}{9}P_d + \frac{1}{9}P_s + \frac{4}{9}P_c \right) \\ &= 0.12 \text{ (experiment).} \end{aligned}$$

In the above formulae, it is fair to assume that the total fraction of the proton's momentum carried by the strange and charmed quarks is negligible. This assumption leaves us with two equations and two unknowns which may be solved to give:

$$P_u = 0.36, \quad P_d = 0.18.$$

Thus the total fractional momentum carried by the up quarks is measured to be twice that carried by the down quarks, which supports the quark model's picture of the proton as (uud).

However, the measurements also indicate that the total fractional momentum carried by the quarks is only one-half of the total proton momentum. The interpretation of this is that the other half is carried by neutral gluons which are the quanta of the strong nuclear force between the quarks. Because these gluons are electrically neutral, they do not experience the electromagnetic force and so show up only as missing momentum in the overall accounting for the proton.

### 29.3 Weak Interaction Structure Functions

The last section showed that the quark model can lend a great deal more detail to our picture of the structure functions. But the picture cannot be filled in completely because the photon transferred in electron–nucleon scattering differentiates between the quarks only by virtue of their electrical charge. To make further distinctions it is necessary to use the W-boson probe of the weak interaction. It is straightforward to establish which of the W-boson–quark interactions are possible, and which of those possible are dominant.

Because of lepton-number conservation in charged-current reactions, the neutrino must turn into a negatively charged muon and emit a positively charged W boson, and the antineutrino must turn into a positively charged muon emitting a negatively charged

W boson. In principle, the  $W^+$  boson can collide with a down quark, thereby changing it to an up, or it may collide with a strange quark, also turning it into an up. But the  $W^+$  boson cannot be absorbed by an up quark, as this would result in a quark of charge  $\frac{5}{3}e$ , which does not exist. The possible reactions can be simplified further because the weak interactions which change strangeness, such as

$$W^+ + s \left(-\frac{1}{3}e\right) \rightarrow u \left(\frac{2}{3}e\right),$$

are much smaller than the strangeness-conserving weak interactions. This is just the Cabbibo hypothesis mentioned in Chapter 17. Because of it, we may ignore all the strangeness-changing reactions which may occur in principle. In addition to the interactions with quarks mentioned so far, the W bosons can also interact with the antiquarks from the vacuum sea. Combining all these considerations, we can summarise all the significant neutrino–quark interactions which are possible in neutrino–nucleon scattering:

- (1)  $\nu_\mu + d \left(-\frac{1}{3}e\right) \rightarrow \mu^- + u \left(\frac{2}{3}e\right);$
- (2)  $\bar{\nu}_\mu + u \left(\frac{2}{3}e\right) \rightarrow \mu^+ + d \left(-\frac{1}{3}e\right);$
- (3)  $\bar{\nu}_\mu + \bar{d} \left(\frac{1}{3}e\right) \rightarrow \mu^+ + \bar{u} \left(-\frac{2}{3}e\right);$
- (4)  $\nu_\mu + \bar{u} \left(-\frac{2}{3}e\right) \rightarrow \mu^- + \bar{d} \left(\frac{1}{3}e\right).$

We may now proceed to write the cross-section for, say, neutrino–proton scattering as a sum of the neutrino–quark cross-sections involved.

In doing this, it is usual to express the differential cross-section, not as before expressed as varying with the momentum-squared  $q^2$  and energy  $\nu$  carried by the intermediate W boson, but instead expressed as varying with the momentum fraction of the target carried by struck quark  $x$  and the fraction of the incident neutrino energy carried across by the W-boson  $y$ . Mathematically, this requires us to make the following transformation:

$$\frac{d^2\sigma}{dq^2 d\nu} \rightarrow \frac{d^2\sigma}{dx dy},$$

with

$$x = \frac{q^2}{2M_N\nu} \quad \text{and} \quad y = \frac{\nu}{E_i}.$$

This is simple to do, after which the neutrino–proton cross-section can be written in terms of the neutrino–quark cross-sections (1) and (4) of (29.1), the proportions of the two being determined by the distribution of down quarks and anti-up quarks inside the proton.



This gives:

$$\begin{aligned} \frac{d^2\sigma}{dx dy} = & f_d(x) \frac{d^2\sigma}{dx dy} (v_\mu + d \rightarrow \mu^- + u) \\ & + f_{\bar{u}}(x) \frac{d^2\sigma}{dx dy} (v_\mu + \bar{u} \rightarrow \mu^- + \bar{d}). \end{aligned} \quad (29.2)$$

The neutrino–quark scattering is of a very simple point-like kind, and so the cross-sections just provide the usual factors for point-like scattering. Writing the above expression in terms of these factors leaves us with the differential cross-section for deep inelastic scattering, but with the structure functions expressed directly in terms of the distributions of quarks within the target nucleon:

$$\frac{d^2\sigma}{dx dy} = \frac{2MEG_F^2}{\pi} [xf_d(x) + x(1-y^2)f_u(x)].$$

$\uparrow$   
 neutrino–quark  
 point-like  
 scattering

$\uparrow$   
 structure functions  
 expressed as  
 quark distributions

These expressions may be derived for  $\nu$  and  $\bar{\nu}$  scattering off both protons and neutrons. The average of the two gives the scattering of  $\nu$  and  $\bar{\nu}$  off a general nucleon target N consisting of a mixture of protons and nucleons. The ratio of  $\bar{\nu}N$  scattering to  $\nu N$  scattering allows the cancellation of point-like scattering factors leaving only a ratio of the quark distributions. By integrating over the variables  $x$  and  $y$ , the ratio of the total cross-sections is expressed in terms of the total fractional momentum carried by each species of quark:

$$\frac{\sigma^{\bar{\nu}N}}{\sigma^{\nu N}} = \frac{(P_u + P_d) + 3(P_{\bar{u}} + P_{\bar{d}})}{3(P_u + P_d) + (P_{\bar{u}} + P_{\bar{d}})}.$$

If quarks only are present inside the target nucleon, then this ratio will be  $\frac{1}{3}$ , and if antiquarks only are present, then the ratio will be 3. The measured value of  $0.37 \pm 0.02$  suggests a very slight presence of antiquarks from the sea.

The presence of the factors of 3 in the above ratio arises from the allowed helicity states of the  $\nu N$  collisions as described earlier, the integration over the  $y$  variable effectively being the integration over the allowed angles of  $\nu q$  scattering, thus favouring neutrino–quark scattering over neutrino–antiquark scattering (and vice versa for antineutrino scattering).

## 29.4 Electron and Neutrino Structure Functions Compared

The quark content of the neutrino scattering structure functions may be obtained simply by comparing the differential cross-section expressed in terms of quark distribution functions (29.2) with the same quantity expressed in terms of the structure functions. This comparison provides equalities such as:

$$\begin{aligned} F_2^{\nu p} &= 2x(f_d(x) + f_{\bar{u}}(x)), \\ F_3^{\nu p} &= -2(f_d(x) - f_{\bar{u}}(x)), \\ F_2^{\nu n} &= 2x(f_u(x) + f_{\bar{d}}(x)), \\ F_3^{\nu n} &= -2(f_u(x) - f_{\bar{d}}(x)). \end{aligned} \quad (29.3)$$

By comparing the quark content of the neutrino structure functions above with the quark content of the electron structure functions, it is possible to predict a numerical relation between the neutrino– and electron–nucleon scattering structure functions. The resulting relationship is:

$$F_2^{\nu N}(x) = \frac{18}{5} F_2^{eN}(x).$$

Experimentally, the relationship is found to hold very well, as illustrated in Figure 29.1. The significance of this is that the quark content of the target nucleon has been verified independently by two separate interactions and is thus that much more credible.

## 29.5 Sum Rules

We can learn more about the roles played by the various quarks inside the proton by using the expressions above to relate particular quark distribution functions to combinations of the observed structure functions. When integrated over the fractional momentum variable of the quark distributions, these relationships can reveal the total fractional momentum carried by each species of quark:

$$\begin{aligned} \int x(f_s(x) + f_{\bar{s}}(x)) dx &= \int (9F_2^{eN}(x) - \frac{5}{2}F_2^{\nu N}(x)) dx \\ &= 0.05 \pm 0.18, \\ \int x(f_u(x) + f_d(x)) dx &= \frac{1}{2} \int (F_2^{\nu N}(x) - xF_3^{\nu N}(x)) dx \\ &= 0.49 \pm 0.06, \\ \int x(f_{\bar{u}}(x) + f_{\bar{d}}(x)) dx &= \frac{1}{2} \int (F_2^{\nu N}(x) + xF_3^{\nu N}(x)) dx \\ &= 0.02 \pm 0.03. \end{aligned}$$

The numbers show that, as expected, the strange quarks and the various sorts of antiquarks carry

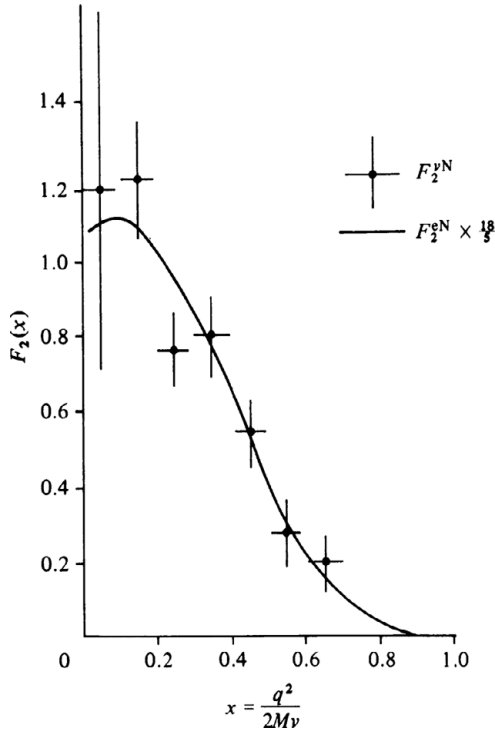


Figure 29.1. The constant relationship between electron and neutrino scattering structure functions is verified by superimposing the two.

only a few per cent of the proton’s total momentum. So the total contribution of ‘sea’ quarks is small. As discovered in electron–nucleon scattering, the up and down quarks together carry about half of the proton’s total momentum. Again the missing half of the proton’s momentum is ascribed to the neutral gluons which are not affected by the weak interactions.

When we integrate the quark distribution functions (or, correspondingly, the structure functions) over the fractional momentum variable  $x$ , it is possible to derive ‘sum rules’ relating these quantities to physically significant numbers. For instance the Gross–Llewellyn Smith sum rule measures the difference between the numbers of quarks and antiquarks in the target. As expected from the quark model, this number is measured to be approximately 3, as illustrated in Figure 29.2,

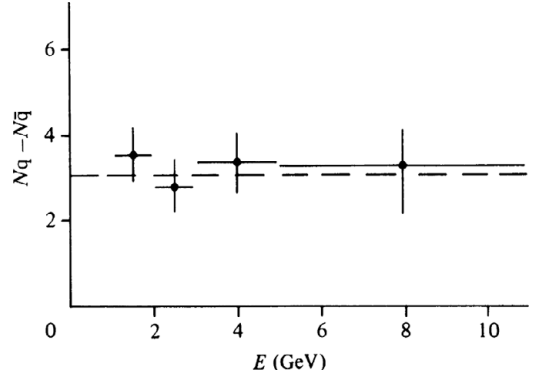


Figure 29.2. The Gross–Llewellyn Smith sum rule measures the difference between the numbers of quarks and antiquarks in the nucleon.

$$\frac{1}{2} \int_0^1 \left( F_3^{\bar{q}}(x) + F_3^q(x) \right) dx = N_q - N_{\bar{q}} = 3.$$

We have spent some time examining the structure functions of deep inelastic scattering, as it is these which have provided us with a direct look inside the proton at its constituent quarks. All the observations are compatible with the standard model of spin- $\frac{1}{2}$  quarks with fractional charges. The big surprise is that these quarks carry only half of the proton’s total momentum – the remainder presumably being carried by the gluons. The next step is to proceed to see if we can learn something of the dynamics of the interactions between quarks and gluons.

### 29.6 Summary

Deep inelastic lepton–nucleon scattering has been an extremely useful tool for probing the structure of hadrons. What we have learnt from these experiments may be stated succinctly as follows. We now know that the nucleon contains point-like constituents (partons), as evidenced by the approximate scale invariance of structure functions:  $F(\nu, q^2) = F(x)$ . That these partons have spin  $\frac{1}{2}$  is clear from the observed relation between the electromagnetic structure functions:  $2xF_1(x) = F_2(x)$ . Furthermore, the behaviour of electroweak cross-sections strongly suggests the identification of these spin- $\frac{1}{2}$  partons with fractionally charged quarks, which account for just one-half of the nucleon momentum (the remainder being due to the gluon constituents).



## **Part VIII**

# **Quantum Chromodynamics – the Theory of Quarks**



## Coloured Quarks

### 30.1 Introduction

The results of deep inelastic scattering experiments are able to tell us a lot about the nature of quarks:

- The scaling behaviour of the cross-sections indicates scattering off point-like quarks with relatively weak interactions between them at short distances.
- The ratio of structure functions  $F_1^{\text{eN}}/F_2^{\text{eN}}$  supports the assignment of half-integer spin for the quarks.
- The comparison of structure functions in electron and neutrino scattering reactions supports the assignment of fractional charges to the quarks.
- The momentum sum rules in both electron– and neutrino–proton scattering suggest that quarks carry only about half of the total proton momentum. The other half is thought to be carried by neutral gluons, the quanta of the interquark force field.

This wealth of information on the structure of the proton was discovered between 1968 and the mid-1970s and represents an experimental triumph similar to the 1911 scattering experiments of Geiger, Marsden and Rutherford, which established the nuclear picture of the atom. In both cases, experimental observation led the way to the development of theories describing the phenomena.

Just as Bohr's early quantum theory of the atom had been advanced to describe Rutherford's discoveries, so quantum chromodynamics (QCD) was put forward as a description of the behaviour of the quarks inside the proton. Pressing the analogy further, just as Bohr's description of the atom was an extension of the quantum theory propounded earlier by Planck, so QCD is an application of the ideas of gauge field theory developed in the 1960s.

QCD was proposed in 1973 by Fritzsch, Leutwyler and Gell-Mann (the last of whom, appropriately enough, was one of the original proponents of quarks in 1963), although a similar idea had been put forward in 1966 by Nambu. The basic idea is to use a new charge called *colour* as the source of the interquark forces, just as electric charge is the source of electromagnetic forces between charged particles.

### 30.2 Colour

Soon after the proposal of the quark model, it was realised that the suggested quark content of some particles clashed with one of the most fundamental principles of quantum mechanics. The Pauli exclusion principle states that no two fermions (particles with half-integer spin) within a particular quantum system can have exactly the same quantum numbers. However, the proposed contents of some particles consist of no less than three identical quarks. For instance, the doubly charged, spin- $\frac{3}{2}$  resonance  $\Delta^{++}$  must consist of three 'up' quarks, all with their spins pointing in the

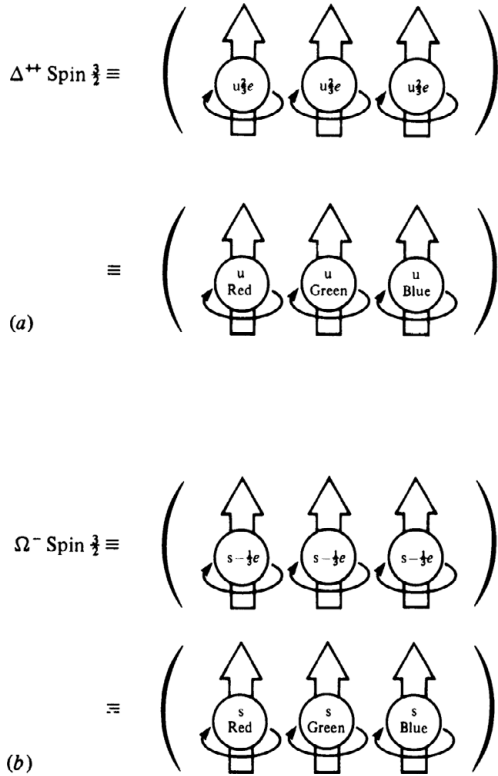


Figure 30.1. The quark content of the  $\Delta^{++}$  in (a) and the  $\Omega^-$  in (b) consists of three identical quarks, which would seem to contradict Pauli's exclusion principle. The introduction of colour distinguishes the quarks and preserves the principle.

same direction (Figure 30.1(a)). Similarly, the famous  $\Omega^-$  particle, the discovery of which first confirmed the validity of the  $SU(3)$  flavour scheme, must consist of three strange quarks (Figure 30.1(b)). These two examples seem to contradict the rules of quantum mechanics.

One immediate way out of this problem is to suggest that the quarks are not fermions at all but, instead, are spinless or integer-spin bosons. However, it was realised early on that only fermionic quarks can account for the spins of the observed hadrons and subsequent observations, say, of structure functions in deep inelastic scattering, have always supported this conjecture. In fact, as we saw in Chapter 3, the mathematical statement of the Pauli exclusion principle deals in terms of the symmetry of the wavefunction which is the total description of the

quantum-mechanical system. The statement that no two fermions can have exactly the same quantum numbers in a particular system is equivalent to the statement that the wavefunction describing a system of fermions must be antisymmetric (i.e. it must change sign) on the interchange of any two of the constituent fermions. The wavefunction which describes a hadron made up of three quarks consists of at least three factors: one describing the positions of the quarks; one describing the spins of the quarks; and one describing the flavours of the quarks. The product of these three factors gives the overall wavefunction:

$$\psi_{\text{TOTAL}} = \psi_{\text{SPACE}} \times \psi_{\text{SPIN}} \times \psi_{\text{FLAVOUR}}.$$

For particles such as the  $\Delta^{++}$ , all quarks have the same flavour, and so the flavour factor of the wavefunction is obviously symmetric under the interchange of any two quarks. The same is true of the spin factor because all quark spins are the same. Because the spins of the quarks add up to give the total overall spin of the particle, it means that there is no orbital angular momentum belonging to the three quarks. This implies that the quarks are positioned symmetrically so that the space factor is symmetric under the interchange of any two quarks. As all the individual factors are symmetric, the total wavefunction must be symmetric and the combination of the three quarks seems to violate the Pauli exclusion principle.

In 1964, Greenberg and, later, Han and Nambu, suggested that the quarks would have to carry another quantum number which would distinguish otherwise identical quarks and so satisfy the demands of the Pauli exclusion principle. This new quantum number they called colour, although it should be stressed that this new property has nothing to do with the normal meaning of the word colour; it is just a label. The total wavefunction will now be multiplied by a new 'colour factor':

$$\psi_{\text{TOTAL}} = \psi_{\text{SPACE}} \times \psi_{\text{SPIN}} \times \psi_{\text{FLAVOUR}} \times \psi_{\text{COLOUR}}.$$

The colour hypothesis is that each of the otherwise identical quarks has a different colour assigned to it and this makes the colour factor, and so the total wavefunction, antisymmetric under interchange of two quarks. The quark model is thus reconciled with the Pauli exclusion principle at the expense of introducing a new quantum number to differentiate between the quarks. Because there are three quarks

Table 30.1. *The flavours and colours of quarks.*

Flavour	Colour		
	Red	Green	Blue
$u (\frac{2}{3}e)$	$u_r$	$u_g$	$u_b$
$d (-\frac{1}{3}e)$	$d_r$	$d_g$	$d_b$
$s (-\frac{1}{3}e)$	$s_r$	$s_g$	$s_b$

inside the proton, three quark ‘colours’ are needed to distinguish them uniquely, say red, green and blue. Each of these labels the three quarks inside the  $\Delta^{++}$  and the  $\Omega^-$  as shown in Figure 30.1(a) and (b). So the net effect of the introduction of colour is to triple the number of quarks; each of the flavours must come in three colours. This is illustrated in the quark table (Table 30.1.)

Obviously, tripling the number of the supposedly fundamental quarks is rather against the spirit of the model, which attempts to make do with as few basic components as possible. So the above arguments for colour, although theoretically compelling, had to be demonstrated by more direct means before the colour quantum number became established as the physical reality on which a theory of quarks could be based. Happily, these more direct means are readily evident.

One piece of evidence supporting the hypothesis that quarks come in three colours is provided by the decay of the neutral  $\pi^0$  meson into two photons (see Figure 7.2). In the quark model, this decay rate is calculated by adding up all the possible varieties of quarks which can act as intermediate states in the decay. The experimental decay rate can be matched by the theory to within a few per cent if the quarks are assumed to come in three different colours. If only one colour of quark is allowed then the answer turns out to be a factor of 9 too small (the number of quarks entering the decay rate formula as the square).

A second piece of evidence for the existence of three different colours of quark is provided by electron–positron annihilations. In these events, the electron and positron annihilate each other to produce a virtual photon loaded with energy. This virtual photon may then decay into either a muon and an antimuon or into a shower of hadrons. The shower of hadrons is the end result of the initial production of a

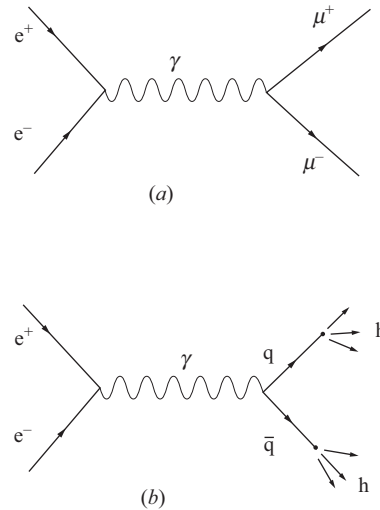


Figure 30.2. An electron–positron pair annihilates into a virtual photon which may then disintegrate into a muon–antimuon pair, as in (a), or into a quark–antiquark pair (which then transforms into hadrons), as in (b).

quark–antiquark pair which subsequently transforms into the conventional hadrons (see Figure 30.2).

The ratio of the cross-sections for these events is a very significant quantity which will be discussed further in Part IX. Suffice it to say at this stage that the ratio is proportional to the sum of the squares of the quark charges, and that only if each flavour of quark comes in three different colours does the number predicted by the quark model agree with the number observed experimentally.

So three separate pieces of evidence attest to the physical reality of the colour quantum number: quark ‘spectroscopy’ (i.e. how quarks build up the known hadrons); the  $\pi^0$  decay; and  $e^+e^-$  annihilations. Two immediate questions arise from this reality. Firstly, if quarks carry these colour charges, is it possible to discover observable hadrons which also carry them? Secondly, given the existence of colour, what is its purpose? Can it form the basis of a theory of quark interactions?

### 30.3 Invisible Colour

The observability of colour and the quark structure of matter are intimately linked. In fact, we will see that the introduction of colour corresponds to a formal method of categorising allowed quark structures. We



saw in Part III how the observed hadrons fit into the multiplets generated by treating the up, down and strange quarks as the elements of the fundamental representation of the symmetry group  $SU(3)$ . However, we saw also that only very specific combinations of the fundamental representation (the quark *flavour* triplet) could generate the correct multiplets for the observed hadrons, see Table 30.2.

The introduction of colour provides a way of categorising which combinations of quarks and anti-quarks are allowed to exist. The first step to realise is that the three colours of any one flavour of quark can be taken as the elements of the fundamental representation of a new symmetry group  $SU(3)_C$ . The methods of group theory then allow us to combine these fundamental representations (the quark colour triplet) into colour multiplets representing the different ways of combining all the colours of the quarks (see Table 30.3). The mathematics of  $SU(3)_C$  is exactly the same as that for  $SU(3)_F$ , although it must be appreciated that these  $SU(3)_C$  colour multiplets represent the various combinations of colour for a given flavour of quark and are completely distinct from the  $SU(3)_F$  flavour multiplets. Just as it was necessary earlier with flavour multiplets to establish which of them could be taken to represent the observed hadrons, it is now necessary

Table 30.2. *Certain combinations of the fundamental quark flavour triplets generate the observed multiplets.*

Fundamental representation	Possible combinations	Multiplets generated	Seen
$q = \begin{pmatrix} u \\ d \\ s \end{pmatrix}$ 'flavour'	$q \times \bar{q}$	<b>1+8</b>	✓
	$q \times q \times q$	<b>1+8+8+10</b>	✓
	$q \times q$	<b>3* × 6</b>	×

Table 30.3. *Combinations of quark colour triplets generate multiplets.*

Fundamental representation	Possible combinations	Multiplets generated	Seen
$q = \begin{pmatrix} r \\ g \\ b \end{pmatrix}$ 'colour'	$q \times \bar{q}$	<b>1+8</b>	?
	$q \times q \times q$	<b>1+8+8+10</b>	?
	$q \times q$	<b>3* × 6</b>	×

to repeat the exercise for the colour multiplets to see which of these are permitted.

The original purpose of the introduction of colour was to ensure that the combinations of quarks representing hadrons are multiplied by factors which are antisymmetric under the interchange of two of the quarks (colours). By using mathematical analysis it is possible to show that some of the multiplets are antisymmetric, just as required, while others are symmetric, and so do not help us. The most obvious and simplest antisymmetric multiplet is the singlet. This observation is then elevated to a hypothesis for explaining the observed quark structure:

All observed hadrons are colour singlets 1.

Under this hypothesis, qq and qqq combinations are allowed because the series of colour multiplets generated includes a singlet. Also allowed are  $q\bar{q}q\bar{q}$  and  $qqq\bar{q}$ . These theoretically allowed combinations are referred to as exotic hadrons and they have long been the subject of experimental searches. Following several false alarms and tentative detections, their existence awaits experimental verification. Also under this hypothesis, combinations such as qq or qqq are not allowed because those combinations of the fundamental representation of the colour group (quark *colour* triplet) do not generate a singlet combination. Experimentally, there has been no suggestion of their existence.

To summarise, let us recap the two parallel descriptions of quark structure, those of flavour and colour. Each combination of quarks generates a set of flavour multiplets and a set of colour multiplets. For certain of these combinations of quarks, the flavour multiplets will correspond to the observed hadrons and these combinations are identified as those which generate a colour singlet. All the observed hadrons are thought to be in colour singlet states.

The statement that all the observed hadrons are colour singlets is equivalent to saying that they are colourless. Just as flavour singlet states can carry no net electric charge or strangeness, the colour singlet states can carry no net colour. This can be understood simply by examining the colour combinations of the allowed quark structures:

$$q\bar{q} = \sqrt{\frac{1}{3}} (r\bar{r} + g\bar{g} + b\bar{b}),$$

$$qqq = \sqrt{\frac{1}{6}} (rgb - grb - rbg + gbr + brg - bgr).$$

The laws of quantum mechanics forbid us to say exactly what colour any one quark is at any one time; all we can say is that there is a certain probability of it being red or green or blue. However, what we can say is that in the singlet state of a  $q\bar{q}$  combination the colour of the quark is exactly cancelled by the anticolour of the antiquark, and that in the singlet state of a  $qqq$  combination all the colours mix in equally to provide a 'white' baryon, i.e. one with no net colour quantum number. Under this scheme, the colour of the quarks is permanently hidden from us because all the allowed quark structures are colourless, colour singlets. The confinement of the quarks within the hadrons can accordingly be restated as the confinement of the colour quantum number.

At this point it is perhaps worth making clear that the confinement of quarks and of colour is really only a hypothesis which is occasionally re-examined, and quite properly so. In 1976, when some anomalous events were detected in electron-positron annihilations, Pati and Salam suggested that they may be the signals of unconfined quarks emerging freely and decaying into leptons rather than undergoing their forced transformations into hadrons. In fact, it was later realised that these anomalous events signalled the production of a new heavier brother of the muon. But, at least for a time, the free quark hypothesis was tenable.

In experiments carried out in the late 1970s at Stanford University, William Fairbank and his collaborators claimed to detect fractional electric charges

on supercooled niobium balls. The experiments were modern versions of Millikan's oil-drop experiment (which first established the value of the electric charge,  $e$ ). However, these results have never been confirmed elsewhere. In fact, all the evidence from accelerator experiments and searches for pre-existing quarks suggests that free quarks do not exist. For example, analysis of sea water indicates that if free quarks do exist, they are so rare that there is less than one for every  $10^{24}$  nucleons. Despite this, it is by no means certain that free quarks and/or coloured mesons and baryons will never be seen. For example, it is hoped that experiments such as the Relativistic Heavy Ion Collider (RHIC) at Brookhaven and the ALICE experiment in the LHC at CERN, which collide heavy ions together at high energies, may result in the formation of a new high-temperature state of hadronic matter in which quarks are deconfined. This new phase is called the *quark-gluon plasma*, now the subject of active experimental programmes at the LHC and elsewhere

Worthy of mention in passing is that the effects of new flavours are easily incorporated into the above picture and do not affect the colour of the quarks at all. The only effect of a new flavour is to generate bigger flavour multiplets to accommodate the greater number of quark combinations possible when an extra degree of freedom is present. The number of quark colours is always three, because that is the number required to distinguish the three valence quarks in each baryon.

## *Colour Gauge Theory*

### 31.1 Introduction

The fundamental idea of quantum chromodynamics (or QCD) is that the colour charges of the quarks act as the sources of the strong, so-called ‘chromodynamic’ force between quarks, just as electric charge acts as the source of the electromagnetic force between electrically charged particles. As the quarks carry both colour and electric charge, they experience both the strong and electromagnetic forces, as well as the more feeble weak and gravitational interactions. However, the chromodynamic force is by far the strongest in most regions of interest and so we are justified in examining it in isolation from the others.

In the terms of classical physics, the colour charges may be thought of as giving rise to a chromostatic force, just as electrical charges give rise to the electrostatic force given by Coulomb’s inverse square law. Although there are great similarities between the two cases, the colour force will be a good deal more complicated. In chromodynamics there are three different colour charges between which the force must be attractive to bind together the three different colours inside each baryon, and the force between colour and anticolour must also be attractive to bind together the quark and antiquark in each meson. However, despite these complications the analogy with the theory of electrostatics is worth pursuing as far as possible.

Any theory of quarks, like any other fundamental theory, must be compatible with the laws of quantum mechanics and relativity and the most usual

approach to achieve this is relativistic quantum field theory. Both QED and the Glashow–Weinberg–Salam theory of the electroweak interactions are examples of a particular class of quantum field theory, namely a gauge theory. As this class has enjoyed such striking success in these other two areas, it must have seemed a reasonable candidate for describing the new chromodynamic forces as well.

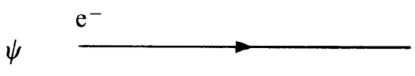
The essence of gauge theory is to explain the origin of the fundamental forces in terms of a symmetry. As we saw in Chapter 20, in QED the form of the interaction between electrons is dictated by the demand that the description of electron behaviour be invariant under arbitrary local redefinitions of the phase of the electron wavefunction. The redefinition of the phase can be expressed mathematically by the action of certain symmetry transformations associated with the group  $U(1)$ . The resulting formulation of QED is summarised in Figure 31.1.

### 31.2 The Formulation of QCD

It will be useful to bear in mind the step-by-step summary of the formulation of QED because the formulation of QCD is remarkably similar. The first step is to identify the symmetry thought to be the fundamental origin of the colour forces. This comes readily to mind.

We have already seen that the quark colour triplet may be used as the fundamental representation of the symmetry group  $SU(3)_C$  and that the colour

(1) A wavefunction describes the propagation of an electron.



(2) Gauge invariance demands that in any theory of electrons the Lagrangian must be invariant under the redefinition of the phase of the electron wave. This is represented by the action of some group of transformations on the Lagrangian.

$$\mathbf{G} \psi \rightarrow \psi^*, \quad \mathbf{G} \mathcal{L}(\psi) \rightarrow \mathcal{L}(\psi^*)$$


(3) Furthermore, it is possible to require that this be true independently at each point in space. This is called local gauge invariance.

$$\mathbf{G}(x) \psi \rightarrow \psi^*, \quad \mathbf{G}(x) \mathcal{L}(\psi) \rightarrow \mathcal{L}(\psi^*)$$

(4) For the Lagrangian to remain invariant under this last operation, a new gauge field must be introduced.

$$\mathbf{G}(x) \mathcal{L}(\psi, A) \rightarrow \mathcal{L}(\psi^*, A^*)$$

(5) This gauge field communicates between the two electron fields their locally defined phase conventions.



(6) Stated in more familiar language, the electromagnetic field mediates the force between two electrons. In quantum theory, this is described as the exchange of photons, the quanta of the electromagnetic field.

Figure 31.1. A summary of the formulation of QED.

multiplet structure generated by this group gives an acceptable way of categorising the known hadrons (i.e. all are in colour singlet states). The fundamental symmetry of the colour force may then be taken as invariance under the redefinition of quark colours.

The redefinition of quark colours is achieved by applying the  $SU(3)_C$  group of transformations to the quark colour triplet. Suppose we define the quark colours initially by the multiplet

$$q \equiv \begin{pmatrix} r \\ g \\ b \end{pmatrix}.$$

We may then choose to change our colour scheme by applying an  $SU(3)_C$  group transformation to this triplet. This will have the effect of mixing up the colours to provide three different combinations each with different proportions of r, g and b:

$$G^{SU(3)} q \rightarrow \begin{pmatrix} c_1(r g b) \\ c_2(r g b) \\ c_3(r g b) \end{pmatrix} \equiv \begin{pmatrix} v \\ y \\ o \end{pmatrix}.$$

However, it is perfectly possible to label these new combinations as new colours, say violet, yellow and orange. The underlying physical requirement is that the theory describing the quark interactions does not depend on whatever ‘colour coding’ is chosen. In fact, the colour coding of the quarks may be different at each point in space and this will require the Lagrangian to be *locally* gauge-invariant under the application of the  $SU(3)_C$  group of transformations.

Just as in the other gauge theories, this requires the introduction of a new gauge field to communicate the local colour conventions from place to place. The quanta of this new colour gauge field are massless spin-1 gauge particles. These are the ‘gluons’ which mediate the chromodynamic forces between quarks. Because the interaction between two quarks corresponds to an interaction between two colour states, gluons must come in colour multiplets which correspond to the colour combinations allowed by group theory. More precisely, since a quark and an antiquark can annihilate into a gluon, the colour quantum numbers of gluons should correspond to a combination of those of quarks and of antiquarks. That is, the colour on a gluon must come from the combination of a quark colour triplet  $\mathbf{3} = (r, g, b)$  with an antiquark anticolour triplet  $\mathbf{3}^* = (\bar{r}, \bar{g}, \bar{b})$ :

$$\mathbf{3} \times \mathbf{3}^* = \mathbf{1} + \mathbf{8},$$

and, indeed, gluons form a colour octet  $\mathbf{8}$  with quantum numbers such as red–antigreen ( $r\bar{g}$ ) and blue–antired ( $b\bar{r}$ ). It is now clear how, in any QCD reaction, redness, greenness and blueness are conserved. For

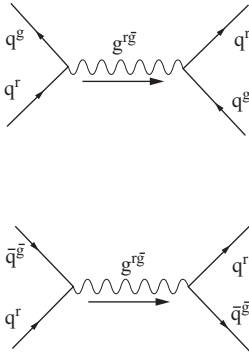
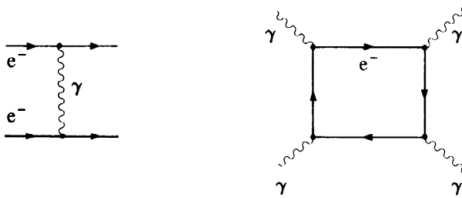


Figure 31.2. Quarks interact by gluon exchange. The colour quantum numbers flowing into the gluon are equivalent to those of a  $q\bar{q}$  pair.

(a) QED



(b) QCD

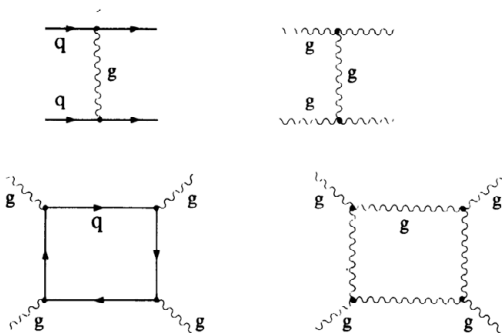
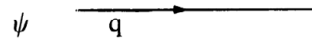


Figure 31.3. (a) In QED, photons cannot interact directly, they must dissociate into an  $e^+ e^-$  pair to do so at all. (b) In QCD, the coloured gluons can interact with each other directly.

instance, in Figure 31.2, we show how a red quark can emit a red–antigreen gluon thereby transforming itself into a green quark.

In QED the photon is electrically neutral and so does not act as a source of electromagnetic fields. This means that photons cannot interact amongst themselves directly. The only way they can interact is for

(1) A wavefunction describes the propagation of a quark.



(2) Gauge invariance demands that the Lagrangian must be invariant under the redefinition of the quarks' colour code.

$$\mathbf{G}^{SU(3)_c} \psi \rightarrow \psi^*, \quad \mathbf{G}^{SU(3)_c} \mathcal{L}(\psi) \rightarrow \mathcal{L}(\psi^*)$$

(3) This gauge invariance may be required to hold locally.

$$\mathbf{G}^{SU(3)_c}(x) \psi \rightarrow \psi^*, \quad \mathbf{G}^{SU(3)_c}(x) \mathcal{L}(\psi) \rightarrow \mathcal{L}(\psi^*)$$

(4) The Lagrangian can remain invariant under this local group only if a new, self-interacting gauge field is introduced.

$$\mathbf{G}^{SU(3)_c}(x) \mathcal{L}(\psi, \tilde{A}) \rightarrow \mathcal{L}(\psi^*, \tilde{A}^*)$$

(5) This gauge field communicates between the quarks their locally defined colour coding. In more familiar terms, the quanta of the colour gauge field, the gluons, mediate the strong force between the quarks and also between themselves.

Figure 31.4. A summary of the formulation of QCD.

each to dissociate into a virtual electron–positron pair or other charged particle pair which may then do the interaction for them. However, this has a much lower probability of occurrence than the direct interaction between electrons. In QCD, the gluons carry colour and so give rise to their own colour fields. This means that they can interact amongst themselves directly. The two cases are illustrated graphically in Figure 31.3.

This difference between the two theories is a very fundamental one and has far-reaching consequences of great importance. The reason for the difference boils down to the number of charges in the theory. As we saw in Part VI, QED is an Abelian gauge theory with a single charge  $Q$ , whereas the more complicated electroweak charges mean that the Glashow–Weinberg–Salam model is a non-Abelian gauge theory, so that the net result of two gauge transformations depends on the order in which they are performed. Similarly, in QCD the colour charges give rise to a non-Abelian gauge theory with the

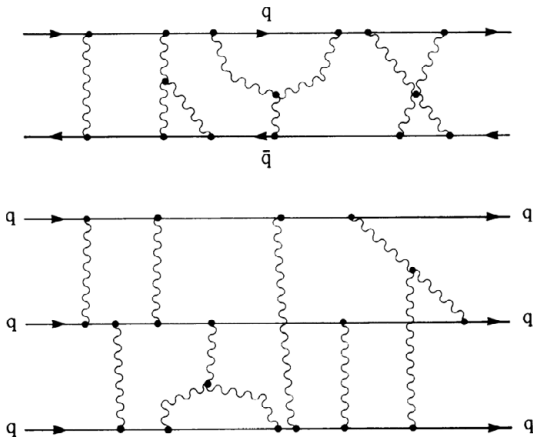
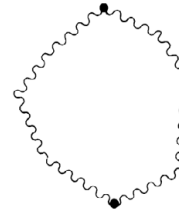


Figure 31.5. The quarks inside the hadrons are bound together by continual gluon exchange.

symmetry group  $SU(3)_C$ . The formulation of QCD is summarised in Figure 31.4.

The picture of the hadrons painted by this theory is one of continual interchange of gluons between the constituent quarks which, as a result, continually keep changing colour but in such a fashion as to always maintain the hadron in its colour singlet state, see Figure 31.5.



$$8 \otimes 8 = 1 + 8 + 8 + 10 + 10 + 27$$

Figure 31.6. The non-Abelian nature of QCD allows the formation of particles made from gluons only.

One interesting consequence of the presence of gluon self-interaction in QCD is the possibility of the existence of particles made only of glue with no quarks. These are referred to as glueballs or gluonium states and are possible because when two colour octets are combined, a colour singlet state always results in addition to non-allowed colour multiplets, see Figure 31.6. Other combinations of more than two gluons are also possible, giving rise to the possibility of a whole spectrum of gluonia. Although theoretically possible, no definitive evidence of their reality has been found to date.

## *Asymptotic Freedom*

### 32.1 Introduction

What we have done so far, formulating a gauge theory of the colour force in analogy to the theory of the electromagnetic interaction, is all very well – but is it correct? Does it explain any of the features of the strong interaction which are observed in the real world? Only by this test, and not by its theoretical elegance or any other criterion, may it be accepted as correct. In particular, we are interested to see if the forces resulting from gluon exchange correspond to the behaviour of the strong interaction as observed in deep inelastic scattering. In these experiments, we saw that when the distances probed were very small (i.e. when the momentum transferred by the probe was very high) then the force between quarks is surprisingly weak and they behave rather like free particles. On the other hand, no free quark has ever been observed, so we may be sure that, over long distances, the force between quarks becomes increasingly strong.

Another, and as it turns out, related theoretical question is whether or not we can actually perform any meaningful calculations in QCD. In QED, calculations of quantities of physical interest are possible because the increasingly complicated higher-order processes become decreasingly important. This is due to the smallness of the electron–photon coupling constant ( $\alpha = \frac{1}{137}$ ). However, in QCD, the strength of the chromodynamic forces may require the quark–gluon and gluon–gluon couplings to be large (greater than one) and this would mean that the increasingly complicated

higher-order processes become increasingly important. In this case, it is impossible to use the same mathematical techniques of perturbation theory to calculate quantities of physical interest.

Resolution of both the experimental and theoretical points mentioned above hinges on the far-from-obvious physical reality that the intrinsic strength of a force (the size of the coupling constant concerned) depends on the distance from which it is viewed. This dependence is in addition to the conventional spatial variation in the strength of forces as given, say, by the inverse square laws of classical physics.

A good example of this phenomenon is provided by the electromagnetic force. In classical physics the electrostatic force between two charged particles is given by Coulomb’s inverse square law:

$$F = K \frac{N_1 e \cdot N_2 e}{r^2}.$$

In this formula, the intrinsic ‘strength’ of the force is fixed by the numerical value of the constant electric charge  $e$ . But when the distance separating the two particles becomes very small, then classical physics is no longer adequate and quantum-mechanical effects must be taken into account.

These quantum-mechanical effects can be described as the polarisation of the vacuum by a sea of virtual electron–positron pairs in the environment of an electric charge. In the region of an electron, say, the virtual positron is attracted towards it and the

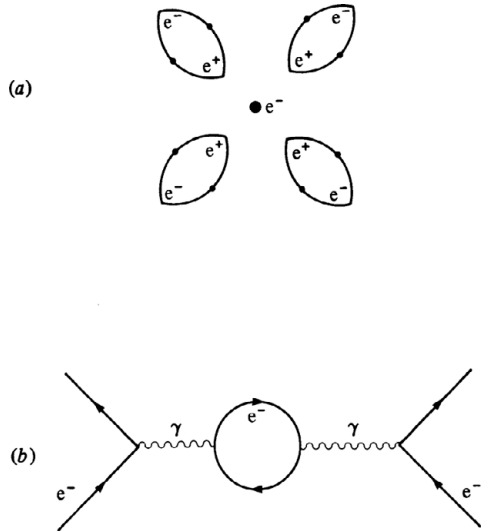


Figure 32.1. (a) Virtual electron–positron pairs shield the ‘bare’ electric charge at very short distances. This effect can be calculated by evaluating Feynman diagrams such as that shown in (b).

virtual electron is repelled away from it. This leads to a cloud of virtual positive charge shielding the ‘bare’ negative charge of the real electron, see Figure 32.1(a). The effect of this is that from a distance the effective negative charge is much reduced compared to its ‘bare’ value. The electric charge appearing in Coulomb’s law is this shielded, effective charge.

The quantum-mechanical shielding effect is known as the ‘renormalisation’ of the bare electrical charge and it can be calculated by evaluating the probabilities of occurrence of the various quantum-mechanical processes such as the one illustrated in Figure 32.1(b). In fact, the probabilities associated with these processes are infinite! This suggests that the bare charge is also infinite, but negative, so that the infinities cancel, leaving a finite value  $e$  for the classical electric charge.

We can carry out a simple thought-experiment to investigate the quantum-mechanical behaviour of electric charge by considering the scattering of two electrons off each other at increasingly high energies. As the distance of approach decreases, the electrons begin to penetrate each other’s virtual charge cloud and to experience each other’s more negative bare charge. The change in the effective value of the

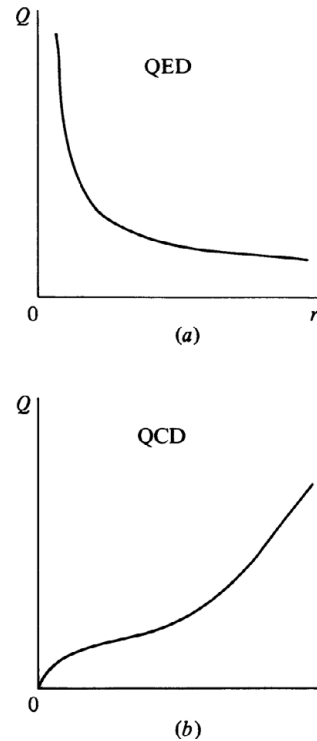


Figure 32.2. (a) The apparent strength of an electrical charge as a function of the distance from which it is viewed. (b) The apparent strength of the colour charge on a quark.

electric charge is shown against distance of approach in Figure 32.2.

A parallel phenomenon exists in QCD. Just as the vacuum can be considered to be a sea of virtual electron–positron pairs, so too can it be considered as a sea of virtual quark–antiquark pairs and gluons, see Figure 32.3(a). The ‘bare’ colour charge on a single quark may then be shielded by the polarisation of this vacuum sea of virtual quarks, antiquarks and gluons. The resulting renormalisation of the bare colour charge may be calculated by evaluating the corresponding probabilities of occurrence of the various quantum-mechanical processes, such as those shown in Figure 32.3(b). The essential new feature in the QCD case is the presence of gluon shielding, possible because of gluon self-interactions. Whereas in QED, the single variety of electron–positron shielding leads to a decrease in the effective electric charge compared with its bare charge, the presence of the gluon shielding effect in QCD provides a greater, opposite effect and



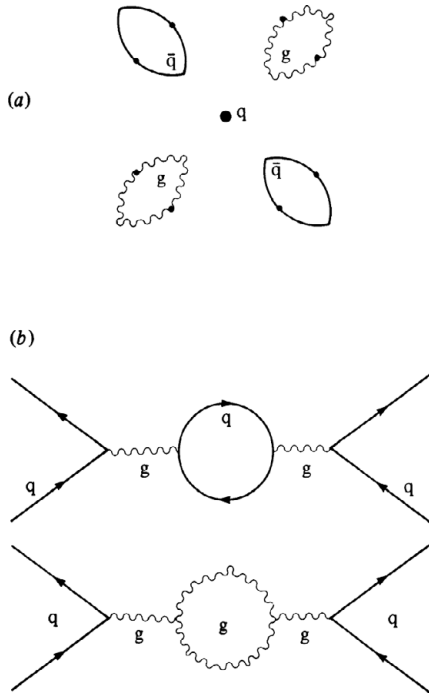


Figure 32.3. (a) Virtual quark–antiquark pairs and gluons combine to ‘antishield’ the colour charges on the quarks. This effect is described by the Feynman diagrams shown in (b).

leads to an increase in the effective colour charge relative to the original bare charge. Conversely stated, the effective strength of the colour charge appears to decrease as the distance from which it is viewed decreases (see Figure 32.2(b)).

This phenomenon is qualitatively similar to the effect noticed in deep inelastic scattering: when the quarks are close together, the chromodynamic forces between them are weak; as the distance between them increases, so too do the forces. The term coined to denote this behaviour is ‘asymptotic freedom’, to denote the fact that when the interquark distances probed become asymptotically small (or, as in the original formulation, when the momentum of the deep inelastic probe becomes asymptotically high), then the chromodynamic forces disappear and the quarks become, effectively, free particles.

This remarkable property of QCD was discovered in 1973 by H. David Politzer of Harvard and independently by David Gross and Frank Wilczek of Princeton. Immediately, the feasibility of developing a field theory for the ‘strong’ interaction received a tremendous boost. For not only does the picture of the forces presented resemble their behaviour as observed in experiment, but also the demonstration that the ‘strong’ interaction coupling constant can, under certain circumstances, be small allows for the application of the traditional methods of perturbation theory to calculate quantities of physical interest.

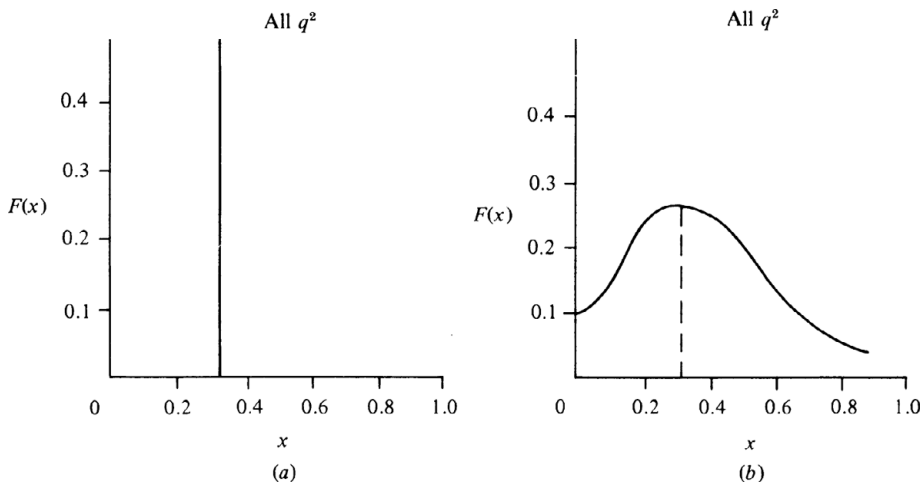


Figure 32.4. The expected behaviour of a nucleon structure function, for (a) free quarks and for (b) confined quarks.

The position of QCD as the candidate theory of strong interactions was strengthened still further in 1974 when Gross and Wilczek went on to prove mathematically that only non-Abelian gauge field theories (of which QCD is an example) can give rise to asymptotically free behaviour. Also, they showed that this behaviour is possible only if there are a limited number of fermions in the theory (no more than 16 quark flavours in QCD) and if there are no Higgs particles involved in any form of spontaneous breaking of the  $SU(3)_C$  symmetry. Thus if we decide that asymptotic freedom is a desirable feature of the theory of chromodynamic forces, certain other possibilities are decided for us automatically.

At this stage, it is desirable that we strengthen the credibility of QCD still further by a demonstration of the validity of the use of perturbation theory in calculating the effects of processes involving quarks and gluons. Fortunately, an example is close at hand.

### 32.2 Violations of Scaling

The description of deep inelastic scattering in Part VII represents a first attempt to gain some understanding of the interior of the proton. In this capacity it served us very well. For not only were we able to interpret the deep inelastic experiments as the first dynamical evidence for the existence of the point-like quarks within the proton, but the implications of the experiments for the interquark forces provided us with the basic material for the formulation of QCD. Armed with this new theory we may now re-examine deep inelastic scattering and provide a more detailed explanation of the structure functions describing the constituents of the nucleon.

If the quarks were truly free particles then each would carry a third of the proton momentum (if we assume there are three valence quarks inside the proton). This would give the very simple form of proton structure function shown in Figure 32.4(a). However, we know that this is not the whole truth as the quarks are confined to within the dimensions of the proton; thus the uncertainty in their position can be no greater than  $2r_p$ . By Heisenberg's uncertainty principle this means that the uncertainty in their momentum must be at least

$$\Delta p \gtrsim \frac{\hbar}{2r_p}, \approx 200 \text{ MeV}/c,$$

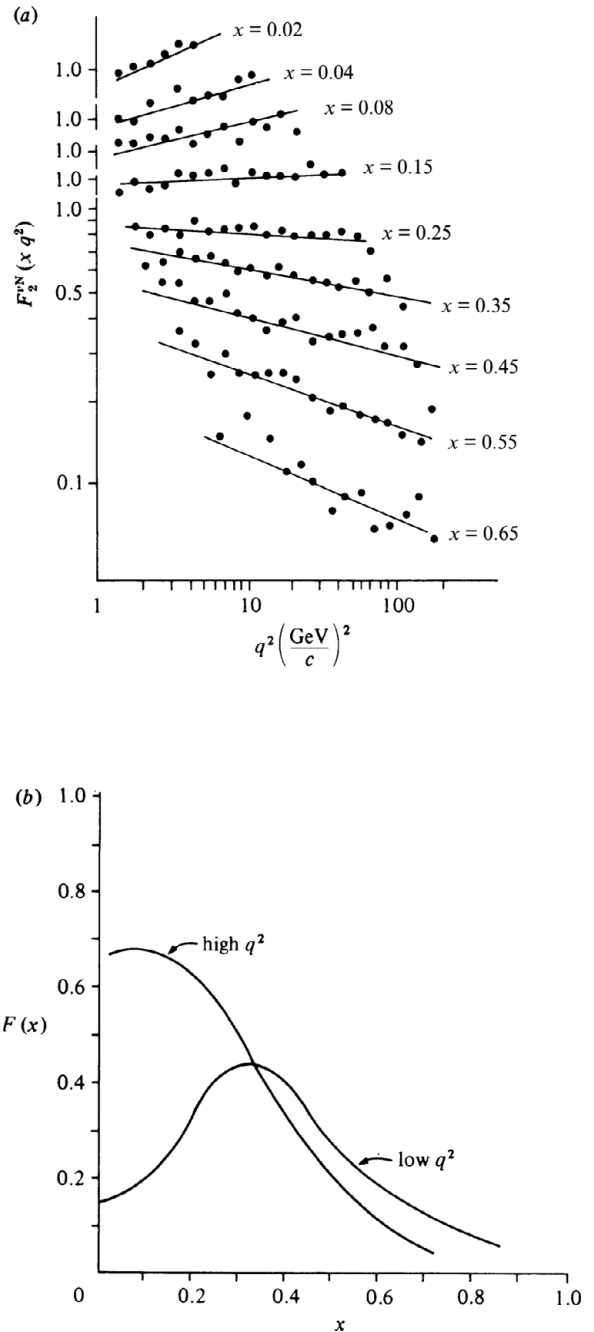


Figure 32.5. (a) The violation of scaling behaviour: the nucleon structure functions vary systematically with momentum transfer squared. Values of  $x$  quoted are in fact the mid-points of ranges centred on those values. (b) The violation of scaling behaviour: the pattern of the variation of the nucleon structure functions.

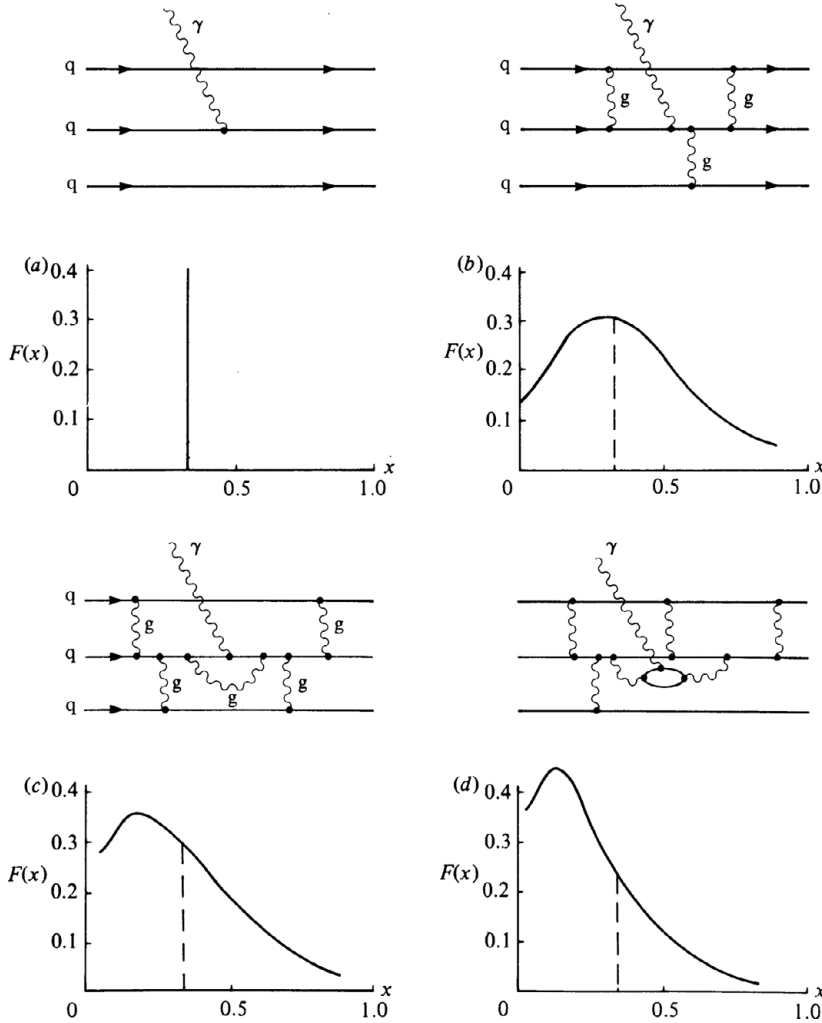


Figure 32.6. The shorter the wavelength of the probe used, the more constituents are seen, each with a smaller fraction of the nucleons' total momentum.

and so the structure function tends to be smeared out as observed in Figure 32.4(b).

A better understanding of deep inelastic scattering is possible when fuller data on the behaviour of the deep inelastic structure functions are examined. These data reveal that, far from being a constant shape for all values of momentum transferred (the original scaling hypothesis), the structure functions vary with it in a very well-defined fashion. This is shown in Figure 32.5(a) and (b). Interestingly, this diagram illustrates why scaling was at first believed to be more exact than it really is. In the early experiments, the structure functions were examined for variations over

only a limited range of  $q^2$ , predominantly in the mid- $x$  regions – where there genuinely is no variation. The important variations in  $q^2$  occur at low and high  $x$  values. Because of this, scaling was credited with more importance than its due. The fuller data show that it is not constancy of the structure functions which is important, but their variation.

The variation of the structure functions is such that at low values of  $x$  they increase with increasing momentum transfer, and that with high values of  $x$  there is a compensating decrease. This means that, as the momentum of the probe increases, it becomes more likely to hit a quark carrying a small fraction

of the total proton momentum and less likely to hit a quark carrying a large fraction. This rather complicated behaviour can be understood by the application of the ‘deep inelastic microscope technique’ to the QCD picture of the proton.

As we have said, if there are no interquark forces then each valence quark will carry a third of the momentum of the proton. The corresponding structure function is shown in Figure 32.6(a). However, to confine the quarks inside the proton, we know that there must be some interquark forces – even if they do weaken in effect as the distance resolved by the probe decreases to less than the proton diameter. In QCD, chromodynamic forces are mediated by the exchange of gluons between the quarks. This continual exchange of gluons transfers momentum between the quarks, so smearing out the deep inelastic structure function (Figure 32.6(b)). As the momentum of the probe increases and the distance it resolves decreases, it begins to see the detailed quantum-mechanical sub-processes of QCD in the environment of the struck quark. For instance, what to a longer-wavelength probe may have appeared to be a

quark may be revealed to a shorter-wavelength probe as a quark accompanied by a gluon (Figure 32.6(c)). What is more, the total momentum of the quark as measured by the long-wavelength probe must now be divided between the quark and the gluon, leaving the quark with a lower fraction of the total proton momentum. So, as the momentum of the probe increases, the average fraction of the total proton momentum carried by the quarks appears to decrease – just as observed in Figure 32.5. As the momentum of the probe increases still further and its resolving distance becomes more minute, it may see the gluon radiated by the valence quark dissociating into a quark–antiquark pair from the vacuum sea. So there will appear to be even more quarks carrying very low fractions of the total proton momentum (Figure 32.6(d)).

Using QCD, it is possible to calculate the probabilities of occurrence of these various quantum-mechanical sub-processes and to derive the way the structure functions vary with the momentum of the probe. As described in part X, all the observed behaviour is completely consistent with the predictions of QCD.

## *Quark Confinement*

### **33.1 Introduction**

The fact that a single quark has never been observed has for years been the single greatest puzzle of elementary particle physics. No matter how energetically protons are collided together in the enormous accelerators at CERN and elsewhere, no quarks are seen to emerge in the debris. Many other varieties of particles are produced, but never any fractionally charged particles which may be identified with the quarks. This means that the forces which bind the quarks together are much stronger than the forces of the collision – which means that they are enormously strong. As an indication, we may note that the energies which bind the electrons into their atomic orbits are of the order of a few electronvolts. The energies binding the protons and neutrons in the nucleus are of the order of a few million electronvolts. Pairs of protons have been collided at energies of hundreds of thousands of millions electronvolts and still no quarks are observed, which means the chromodynamic force between them must be at least that strong.

Not surprisingly, other more bizarre quark hunts have met with no success. Attempts have been made to detect the existence of fractional electric charges in all manner of materials from oysters (because they filter a large amount of sea water) to moon dust, with no convincing record of success. Because of the very delicate nature of the experiments, which are basically modern variants of Millikan's oil-drop experiment, fractional charges are sometimes

reported. But none of these have yet gained general acceptance.

These basic experimental facts have led theorists to conjecture that quarks may be permanently confined within hadrons as a result of the fundamental nature of the chromodynamic force. In contrast to Abelian QED which gives rise to Coulomb's inverse square law of electrostatic attraction, it may well be that the non-Abelian nature of QCD gives rise to a confining force which does not decrease with increasing distance. In fact, the corollary of asymptotic freedom is that the effective strength of the chromodynamic force increases as the quarks are drawn apart, a phenomenon known as 'infrared slavery'. It is not yet known whether QCD gives rise to infrared slavery or whether, after a period of rising, the chromodynamic force tends to a constant strength or even decreases as the quarks are separated. If the force does eventually begin to drop off, then the quarks will eventually be separable and confinement only a temporary phenomenon, apparent because accelerator energies are not yet high enough. The various possibilities are shown in Figure 33.1.

The major hindrance to a straightforward examination of the confinement problem is the difficulty in developing a mathematical description of strong forces. The method of perturbation theory used in QED and in the asymptotically free regime of QCD is valid only because the forces are weak. Attempts have been made to develop other methods such as

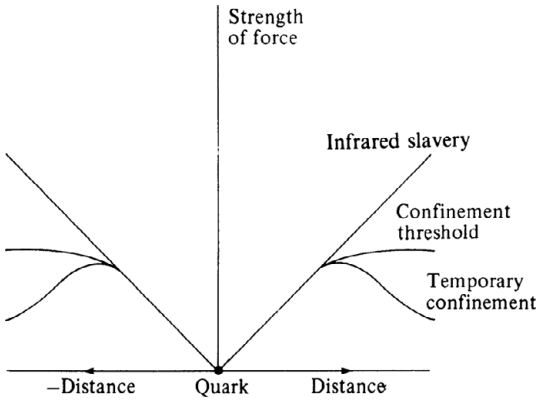


Figure 33.1. Possible behaviours of the chromodynamic force at large distances.

one which divides the space-time continuum into a lattice of discrete points (so-called lattice gauge theory). Quantum field theory calculations can then be done numerically, using a computer. Simulations of this kind clearly indicate confinement, but this is by no means a mathematical proof. (However, such simulations are vital for estimating the amplitudes of hadronic processes needed to disentangle today's complicated experimental data.)

Instead, we will have to content ourselves with an intuitive picture of how the non-Abelian nature of QCD may give rise to the confinement mechanism. As usual, we start off with the familiar case of electrodynamics. The field lines joining two charges spread out to infinity in a spherical fashion. As they are drawn apart the field lines become more spread out. Because the density of field lines at any point is related to the strength of the electrostatic force at that point, this means that the force decreases as the separation increases, see Figure 33.2.

Consider now what may happen in QCD to the chromodynamic force between the quark and anti-quark in a meson. The chromodynamic field lines would like to spread out like the electrodynamic ones, but because the non-Abelian nature of QCD gives rise to self-interactions of the gauge field, the field lines are drawn together instead. This is illustrated in Figure 33.3 by field lines forming a 'flux-tube' between the quarks. As the quarks are separated, the field lines do not spread out but, instead, are drawn out into a tube in which the density of chromodynamic force lines may be constant. This would lead to a constant force existing between the quarks. Eventually, as we

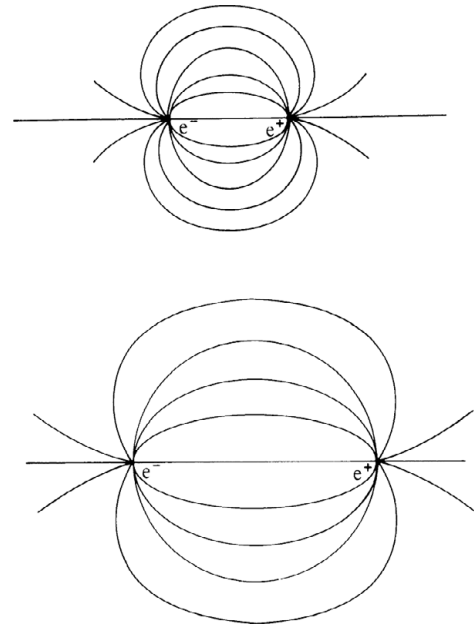


Figure 33.2. Electric field lines spread out as the electric charges are separated.

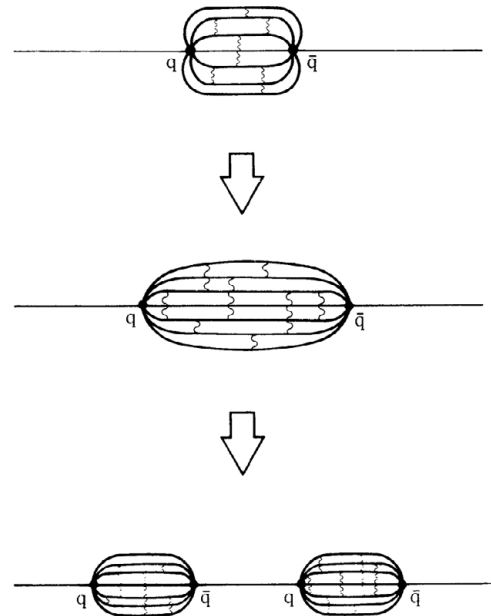


Figure 33.3. Colour force lines between quarks are collimated into a tube-like shape and do not spread out as the quarks are separated. Eventually a single tube will split into two when the force applied has completed enough work.

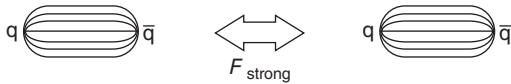
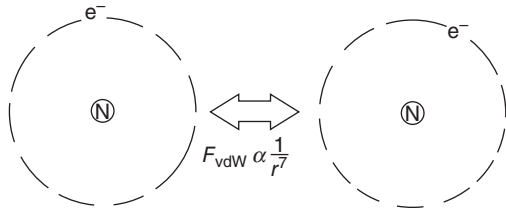


Figure 33.4. The analogy between the van der Waals force between atoms and the long-range colour force between the observed hadrons.

put more and more work into increasing the separation of the quarks, the system will gain enough energy to promote a virtual quark–antiquark pair from the vacuum sea into physical reality. This will give rise to the creation of a new meson. So the energy we expend in attempting to separate the  $q\bar{q}$  pair has, in fact, resulted in the production of another meson, just as occurs in the high-energy collisions!

### 33.2 Quark Forces – Hadron Forces

Having seen how QCD may provide an acceptable picture of the interquark forces, it is worth pausing to relate this picture to that of the forces between the observable hadrons, mentioned briefly in Part III. These are the forces which bind the protons and neutrons together in the nuclei and, when the hadrons are in collision at high energies, produce the numerous secondary particles.

These forces are now seen as the ‘van der Waals’ forces between hadrons. The van der Waals forces between atoms are the very feeble residual electrodynamic effects remaining after the electrons and nucleus

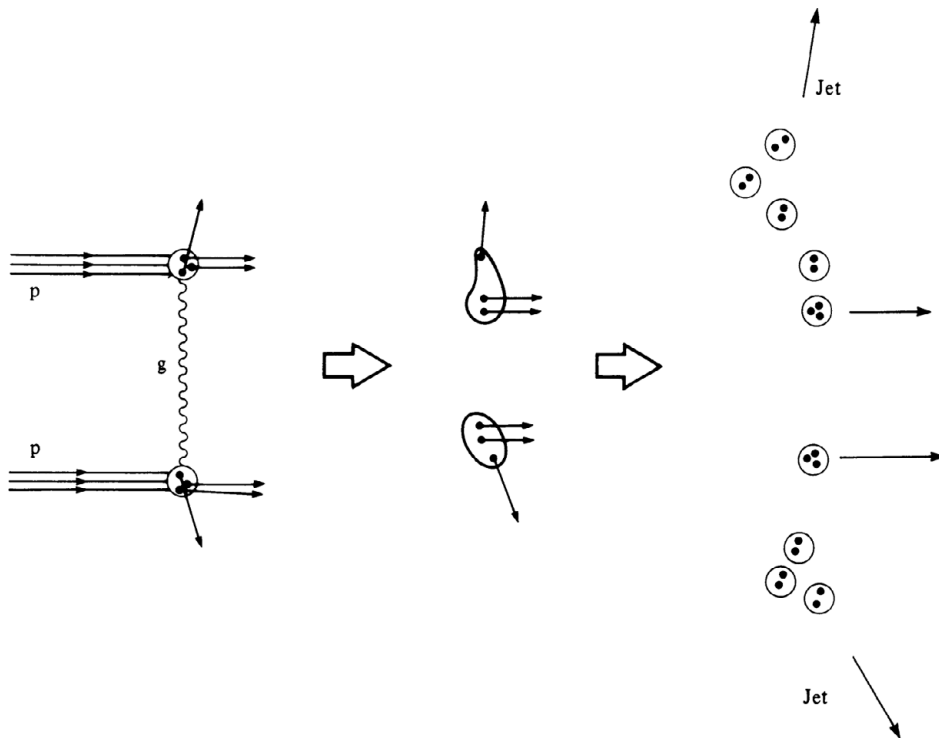


Figure 33.5. Head-on hadron–hadron collisions are described by simple quark and gluon processes, such as one-gluon exchange, which give rise to jets of hadrons emerging from the collisions.

have formed a net electrically neutral atom (Figure 33.4). Analogously, the van der Waals forces between hadrons are the chromodynamic effects remaining between colour singlet states once their colour constituents are bound together. Unlike the electrodynamic case, however, there is no guarantee that these ‘secondary’ forces will be weaker than the ‘primary’ interquark chromodynamic forces. This is because they are predominantly long-range phenomena, which is in the strong coupling regime of QCD.

Hadron collisions can be divided into either of two main classes. In the first are the diffractive collisions which are, in effect, glancing blows between the colliding particles. In the QCD picture, these long-range collisions are complicated affairs involving multiple-gluon exchange with many sub-processes occurring. Because the forces are strong, there is no well-established method of describing the quark and gluon behaviour in these collisions. Indeed, there is no great motivation for examining these collisions at the level of the details of quarks and gluons, as we are unlikely to be able to deduce much about the fundamental nature of the forces from such a complicated event. It is as if we were to attempt to study the electromagnetic force by observing collisions between complex atoms!

In the second main class of hadron collisions are the non-diffractive or ‘head-on’ events. Because

they are much rarer than the diffractive events, they tended to be rather ignored prior to the development of QCD. However, at the levels of quarks and gluons, non-diffractive collisions are rather simple and so became one of the centres of attention in the quest to understand more of the interquark forces. Because the hadrons collide head-on, it means that the quarks in each collision will approach each other very closely. The collisions are then thought to proceed predominantly by the exchange of a single gluon between two passing quarks, all the others acting as passive spectators (Figure 33.5). The result is that the interacting quarks are knocked violently sideways out of their parent hadrons. Of course, they do not emerge as free particles (the confinement mechanism dresses them up as hadrons), but the result is a jet of hadrons emerging along the directions of motion of the original quarks. These ‘high transverse momentum’ jets were observed in the early 1980s at the CERN  $p\bar{p}$  collider.

Jets occur also in other classes of high-energy collisions, such as electron–positron annihilations, where there are no complications due to spectator quarks. These events are altogether cleaner, as we will see in Part IX. But it is encouraging to note common phenomena in two very different circumstances, as this suggests a common, fundamental origin which we take to be the underlying dynamics of quarks and gluons.





**Part IX**  
**Electron–Positron Collisions**



## *Probing the Vacuum*

### 34.1 Introduction

The primary means of studying the fundamental constituents of matter and their interactions is through performing scattering experiments. This has been the case since the very beginning of particle physics. One class of scattering experiments which has, over the past few decades, been extremely fruitful involves collisions between electrons and positrons. These  $e^+ e^-$  experiments have yielded a great deal of information about the nature of the strong, weak and electromagnetic forces, and have played a major role in establishing the ‘Standard Model’ of particle physics: QCD and the Glashow–Weinberg–Salam electroweak theory.

An attractive feature of  $e^+ e^-$  experiments is that because the electron and the positron are antiparticles, they often annihilate into a ‘vacuum’ state of pure energy. All the quantum numbers of the initial particles cancel and so we avoid the inhibiting effects of some conservation laws. The energy resulting from the annihilation (in the form of a virtual photon or  $Z^0$ ) is then free to produce other particle–antiparticle pairs. In this way  $e^+ e^-$  experiments are ideal reactions in which to look for new particles. By bringing beams of electrons and positrons into collision, physicists have also been able to study the couplings of leptons and quarks to the photon and  $Z^0$ , and hence the electroweak properties of these fermions, as well as the nature of the strong interaction between hadrons.

There is a further convenient feature of these reactions. If an electron and positron are collided head-on, with equal and opposite momentum, then the centre of mass of the reaction is absolutely stationary, and the total centre-of-mass collision energy is just the sum of the energy of each particle. (This is in contrast to accelerating electrons into stationary targets where the centre of mass is moving, and the total centre-of-mass collision energy is much less than the energy of the electron.) Furthermore, because the centre of mass is stationary in head-on collisions, the angular distribution of created particles can be measured directly and significant asymmetries detected that much more easily.

Because the  $e^+ e^-$  pair can annihilate to a vacuum state, it means also that the reactions are very clean. There is no debris surviving from the initial state to mask interesting new effects or to confuse the experiments’ detectors. This is in contrast to the deep inelastic experiments in which the photon–quark interaction has to take place in the presence of spectator quarks and leptons.

### 34.2 The Experiments

Naturally, the many benefits of  $e^+ e^-$  collisions are balanced by some penalties. The basic requirement is to collide bunches of electrons with bunches of positrons head-on and this requires comparatively elaborate accelerator technology. The most serious drawback is the numbers of  $e^+$  and  $e^-$  available in

the bunches. The available flux limits the rate at which reactions can be observed, and this in turn limits the accuracy of the measurements obtained.

Another limitation is that the range of phenomena open to study depends upon the total energy of the collision. The total energy is just the sum of the  $e^+$  energy and the  $e^-$  energy and, as the collision is head-on, all of it is available to create the mass of new particles. But while the  $e^+ e^-$  bunches are kept in orbit they emit their energy in the form of synchrotron radiation at a rate which rises rapidly with their energy (as the fourth power) and with the tightness of the bends (as the inverse of the radius of curvature). So, to achieve very high energies, it is necessary to build very large rings and to input a lot of radio frequency (r.f.) power to replenish the lost energy. So the search for new particles led to the construction of increasingly more powerful accelerators. The progress of this construction in the last three decades of the last century is summarised in Table 34.1.

The most recent frontier of  $e^+ e^-$  experiments was reached by the Large Electron–Positron (LEP) collider at CERN. But historically, it was the SPEAR ring at SLAC in California which provided such important advances during the mid-1970s and which may thus serve as the best illustration of  $e^+ e^-$  experiments, see Figure 34.1.

Electrons are created at the end of the two-mile-long linear accelerator tube. After some initial

acceleration by electric fields, some of the electrons can be collided with a target to produce general particle debris from which the positrons can be filtered. Bunches of electrons and positrons are then accelerated down the tube by alternating electric fields and are then injected in opposite directions into the storage ring SPEAR. The bunches can be stored and accelerated in orbit in this ring for several hours by magnetic fields and r.f. power cavities, during which time  $e^+$  and  $e^-$  bunches can be made to pass through each other at specified interaction regions. The individual  $e^+ e^-$  interactions are studied by many different types of detectors, which are usually cylindrical or spherical distributions of sensors wrapped around the interaction region.

### 34.3 The Basic Reactions

It is useful to classify the various possibilities that can occur during an  $e^+ e^-$  collision. Firstly, there are the purely electromagnetic processes. ‘Bhabha scattering’ is the name given to elastic  $e^+ e^- \rightarrow e^+ e^-$  scattering. This can occur by either of the two Feynman diagrams of Figure 34.2(a) corresponding to the possibilities of photon exchange, and of annihilation into a virtual photon with subsequent reproduction of an  $e^+ e^-$  pair.

However, the simplest electromagnetic process is muon-pair production  $e^+ e^- \rightarrow \mu^+ \mu^-$  (Figure 34.2(b)), as this can occur only through the single-photon annihilation mechanism. The single photon must be virtual, as we remarked in Chapter 4, as it is impossible to conserve both energy and momentum. The energy of the initial two-electron state is always greater than  $2m_e$  but its momentum is zero; whereas the energy momentum relationship for real photons is  $E = pc$ .

If the energy of the  $e^+ e^-$  pair is greater than  $2m_\mu$  then the virtual photon can promote a muon pair from the negative-energy sea in the same way as it can reproduce an  $e^+ e^-$  pair. Another possibility is the production of two real photons (Figure 34.2(c)). This is allowed as the photons can emerge with equal and opposite momentum, so both energy and momentum can be conserved simultaneously.

Accurate measurement of these electromagnetic effects allows us to test the validity of QED at very high energies. This is done usually by measuring the angular distribution of particles emerging from the collision and comparing results with predictions. The

Table 34.1. *Important  $e^+ e^-$  storage rings.*

Accelerator	Location	Start	Maximum energy (centre of mass)
SPEAR	Stanford, USA	1972	8 GeV
DORIS	DESY, Hamburg, Germany	1973	12 GeV
PETRA	DESY, Hamburg, Germany	1978	45 GeV
CESR	Cornell, USA,	1979	12 GeV
PEP	Stanford USA	1980	30 GeV
TRISTAN	Tsukuba, Japan	1987	64 GeV
SLC	Stanford, USA	1989	100 GeV
LEP	CERN, Geneva, Switzerland	1989	200 GeV

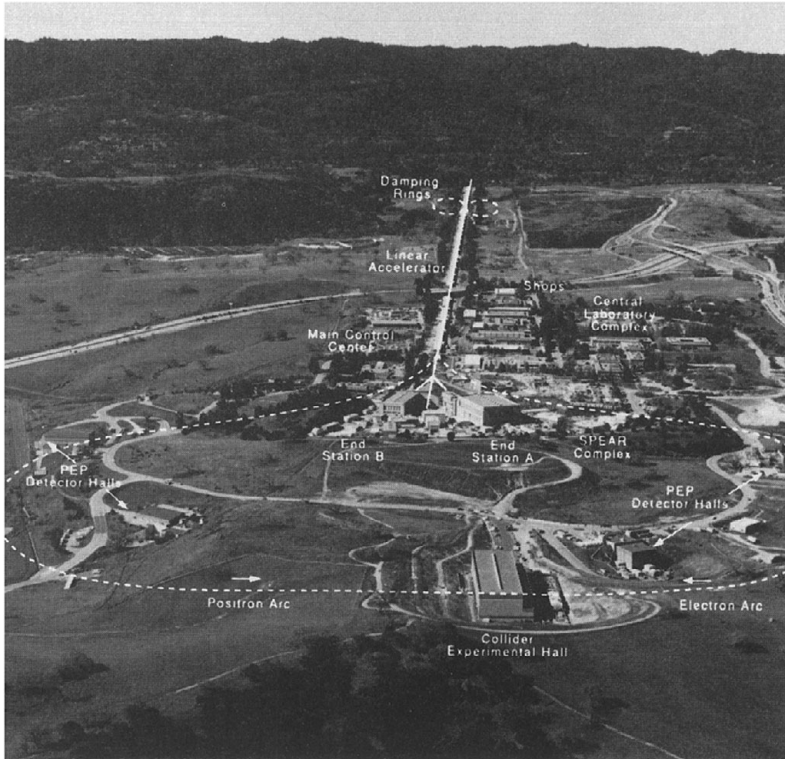


Figure 34.1. An aerial view of the Stanford Linear Accelerator Center. This is the home of the historic SPEAR electron–positron ring and the much larger PEP collider. The 3-kilometre linear accelerator subsequently formed part of the Stanford Linear Collider (SLC). In the SLC, bunches of electrons and positrons were boosted down the linear accelerator and guided around separate arcs, before being brought into collision. (Photo courtesy SLAC.)

correctness of QED has been verified up to the highest accelerator energies. This implies that the leptons are indeed fundamental point-like particles (any possible sub-structure must be smaller than  $10^{-18}$  m in size).

The second class of  $e^+e^-$  collisions comprises those in which hadrons emerge in the final state and which indicate that the strong interaction is involved somewhere, see Figure 34.3(a). One of the most significant quantities in particle physics is the ratio  $R$  of the cross-section for  $e^+e^- \rightarrow \text{hadrons}$  to that for  $e^+e^- \rightarrow \mu^+\mu^-$ , measured as the energy of the collision varies:

$$R = \frac{\sigma(e^+e^- \rightarrow \text{hadrons})}{\sigma(e^+e^- \rightarrow \mu^+\mu^-)}.$$

The significance of the ratio  $R$  is that it compares a reaction we understand very well (muon-pair production) with the class of reactions we wish to understand (hadron production), thus providing a very

useful guide to our thinking about the unknown. Also, the ratio  $R$  is relatively straightforward to observe experimentally. Only two charged ‘prongs’ are seen to emerge in muon-pair production whereas, almost invariably, more emerge from a hadronic final state. So the ratio can be obtained by dividing the number of events detected with more than two prongs by the number with only two prongs, as measured during a given experiment.

Surprisingly, the ratio  $R$  is constant over large energy ranges, indicating that the complicated hadronic state is produced in much the same way as the simple muon pair. The virtual photon is probing the negative-energy sea of hadrons contained in the vacuum instead of electrons or muons. We will see how this can be given a clear interpretation in terms of quarks in the next chapter. Suffice it at this stage to note that the hadrons cannot have been produced by the  $e^+e^-$  pair annihilating into a virtual gluon, as

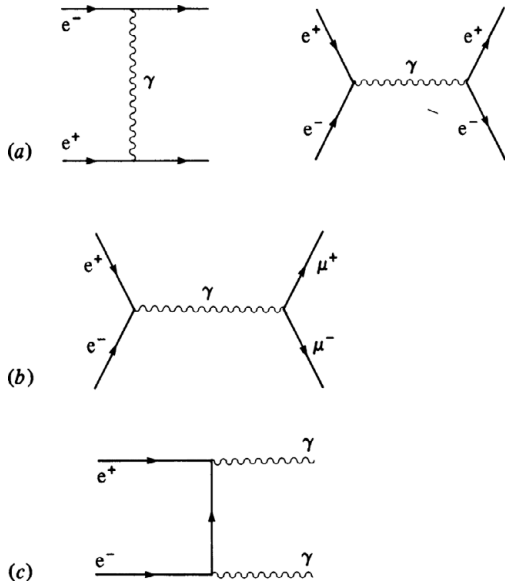


Figure 34.2. Possible electromagnetic effects following an  $e^+ e^-$  collision. (a) Bhabha scattering. (b) Muon-pair production. (c) Two-photon production.

the leptons have no colour and so have no connection with gluons whatsoever.

Before passing on, we must finally identify a third class of  $e^+ e^-$  reaction involving the weak force. This results because the  $e^+ e^-$  do carry weak isospin, which allows them to annihilate into a virtual  $Z^0$  boson. In fact, as we discussed during our look at the Glashow–Weinberg–Salam model, the photon

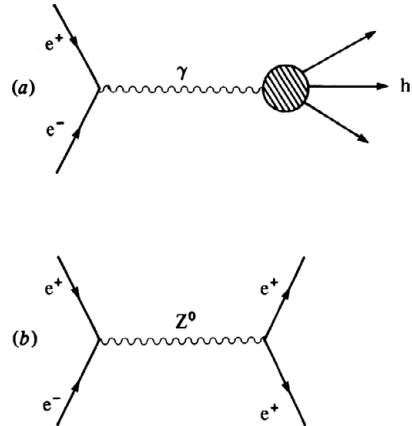


Figure 34.3. Non-electromagnetic effects: (a) hadron production in the final state (the blob); (b) annihilation into a  $Z^0$  boson.

$\gamma$  and the  $Z^0$  boson are simply the rather dissimilar quanta of the unified electroweak force. Thus we might expect weak interaction effects to come into the picture somewhere.

The virtual  $Z^0$  boson is free to explore the negative-energy content of the vacuum just like the photon, see Figure 34.3(b). As a result we should expect to see some uniquely weak interaction effects (such as parity violation) creeping in at higher energies. This we shall discuss further in Chapter 37. Until then, however, we shall ignore these very slight effects. Most of our attention will focus on hadron production and the ratio  $R$ .

## Quarks and Charm

### 35.1 Introduction

The observation of scaling in deep inelastic scattering provides firm evidence for the interaction of the photon with point-like quarks inside the observed hadrons. So when we come to explain the process  $e^+ e^- \rightarrow$  hadrons, the most likely picture is that of the virtual photon interacting with quarks rather than directly with complete hadrons (Figure 35.1(a)). The photon promotes a quark–antiquark ( $q\bar{q}$ ) pair from the vacuum, giving the quark and antiquark a kinetic energy depending on the initial collision energy. The  $q$  and  $\bar{q}$  must separate with equal and opposite momentum to maintain the net momentum of zero and in so doing are ‘dressed up’ into hadrons by the, as yet unknown, quark-confinement mechanism (Figure 35.1(e)). This may be viewed as the potential energy of the long-range attractive force between  $q$  and  $\bar{q}$  being used to promote extra  $q\bar{q}$  pairs from the vacuum.

### 35.2 The Quark Picture

Because the confinement stage of the process *always* occurs (at least assuming the permanently confined quark hypothesis), it enters the calculation of the process  $e^+ e^- \rightarrow$  hadrons only as a final probability of one multiplying the underlying process  $e^+ e^- \rightarrow q\bar{q}$ . As the quarks are observed to be point-like and spin  $\frac{1}{2}$ , the process  $e^+ e^- \rightarrow q\bar{q}$  is very similar to the process  $e^+ e^- \rightarrow \mu^+ \mu^-$ , the only difference being that the charges on the quarks are only some fraction of that on the muons. This explains the constancy of

the ratio  $R$  mentioned earlier and displayed in Figure 35.2. The fundamental dynamics of the two processes are the same, so giving an  $R$  constant with energy, but their magnitudes differ by an amount equal to the ratio of the squares of the charges involved, see Figure 35.3. As several species of quark will be able to act as intermediaries to the creation of hadrons, and as the charge on the muon is 1, then  $R$  is equal to the sum of the squares of the quark charges.

Now the significance of  $R$  is gloriously apparent. It is a directly observable quark-counting opportunity which provides a measure of the number of quarks and their properties. For instance, in the simplest quark scheme with just three quark flavours, namely up ( $\frac{2}{3}e$ ), down ( $-\frac{1}{3}e$ ) and strange ( $-\frac{1}{3}e$ ), the value of  $R$  is predicted to be

$$R_{uds} = \left(\frac{2}{3}\right)^2 + \left(-\frac{1}{3}\right)^2 + \left(-\frac{1}{3}\right)^2 = \frac{2}{3}.$$

As we mentioned in Chapter 30 on QCD, the advent of the colour degree of freedom triples the number of quarks and so predicts the correct value at low energy,  $R = 2$ .

We have now explained the constancy of the ratio  $R$  but have not so far mentioned the very pronounced spikes which punctuate the picture. These shapes are highly reminiscent of the phenomena of resonance particles in Chapter 9 and this in fact is just what they are. At certain energies of the  $e^+ e^-$  collision, the  $q\bar{q}$  pair into which the photon transforms



will have just the correct mass to stay intact as a single-meson resonance. This is signalled by a large increase in the probability of the event occurring compared with non-resonant ‘background’  $q\bar{q}$  production at neighbouring energies and this leads to the observed

spikes in the cross-section. After its brief existence, the resonance particle will then decay by its usual mechanisms into the final-state hadrons observed.

The resonance particles produced are a select sub-set of the hundreds which have been observed in hadron–hadron reactions. The sub-set is defined by the quantum numbers of the virtual photon from which the resonances transform: spin 1, zero charge and strangeness. This defines the allowed quark content of the meson as being that of the quark–antiquark combinations in the vector nonet ( $= \mathbf{8} + \mathbf{1}$ ) of  $SU(3)$  flavour symmetry.

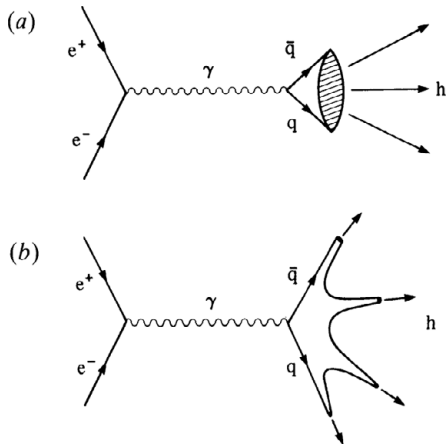


Figure 35.1.  $e^+ e^- \rightarrow$  hadrons proceed by a  $q\bar{q}$  intermediate state, shown in (a). The transformation of this state into the observed hadrons involves the creation of more  $q\bar{q}$  pairs (b).

### 35.3 The Advent of Charm

In what was undoubtedly the most sensational experimental surprise of the 1970s, an extraordinary new resonance spike was discovered in  $e^+e^- \rightarrow$  hadrons at a collision energy of 3.096 GeV, followed quickly by the discovery of a similar spike at 3.687 GeV and a subsequently turbulent rise in the value of the ratio  $R$  to a new plateau, see Figure 35.2. The new resonance was denoted the  $\psi$  (psi) by Burton Richter and his colleagues at SLAC, who observed the particle in  $e^+e^-$  annihilations, but it was also seen simultaneously as a resonance production phenomenon in

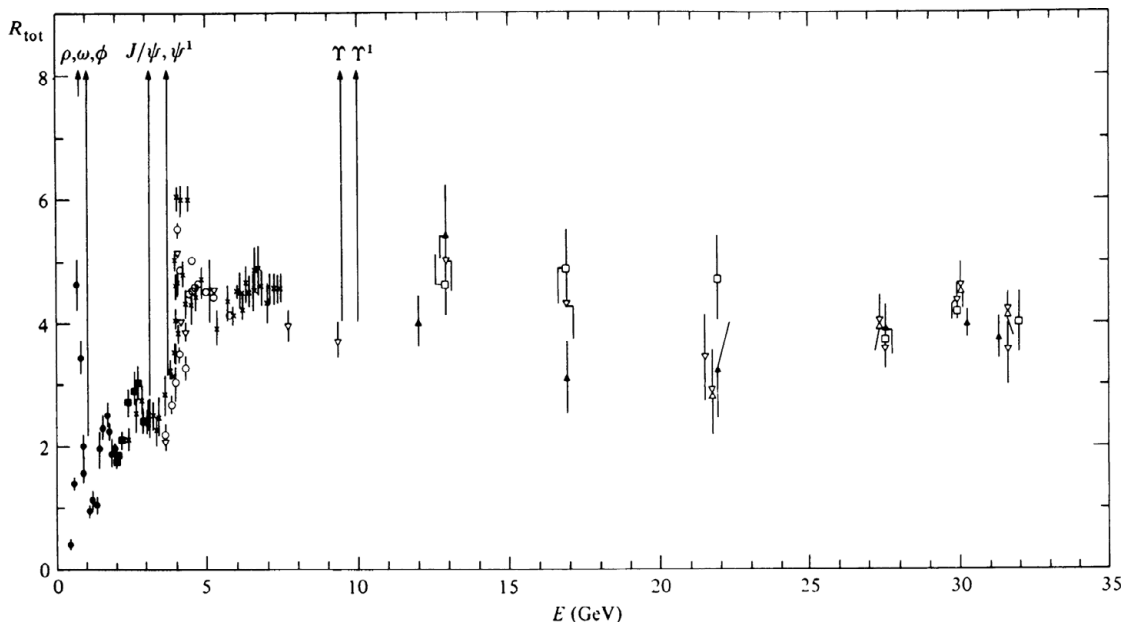


Figure 35.2. The ratio  $R$  of the total hadronic cross-section to  $\sigma(e^+e^- \rightarrow \mu^+\mu^-)$  as a function of the cm energy,  $E$ .

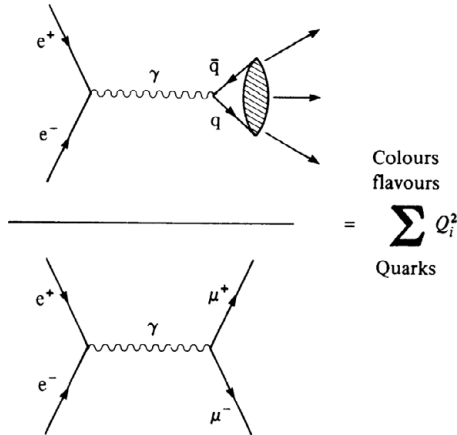


Figure 35.3. The value of the ratio  $R$  is equal to the sum of the squares of the quark charges.

the reaction  $p + p \rightarrow e^+e^- + X$  at Fermilab by Samuel Ting and his team, who denoted it  $J$ . For their discovery both Richter and Ting were awarded the Nobel Prize in 1976. Subsequently,  $J/\psi$  has become the accepted symbol for the particle at 3.097 GeV and  $\psi'$  for that at 3.687 GeV.

After a brief period of speculation, the correct interpretation of the  $J/\psi$  emerged. What had happened was that the increasing energy of the  $e^+e^-$  collision had become sufficiently large to create a new flavour  $q\bar{q}$  pair. It had boosted a new heavier type of quark from its negative-energy sea in the vacuum. The  $J/\psi$  and  $\psi'$  were bound-state mesons consisting of the  $q\bar{q}$  pair and, at energies above the threshold of its production, the pair could contribute to the ratio  $R$ , thus accounting for its observed rise.

This new flavour, called charm, had been anticipated in advance by the GIM theorists attempting to explain the behaviour of hadrons in the Glashow–Weinberg–Salam theory of the weak force. As we saw in Section 23.2, it was put forward as an explanation for the absence of strangeness-changing neutral currents. Also, it was able to complete an aesthetically pleasing matching between the numbers of fundamental leptons and fundamental hadrons. With the advent of charm, it became possible to group the leptons and the quarks into two generations, the second being simply a massive repetition of the quantum numbers of the first.

But the discovery of the charmed quark automatically implied the existence of a horde of new particles

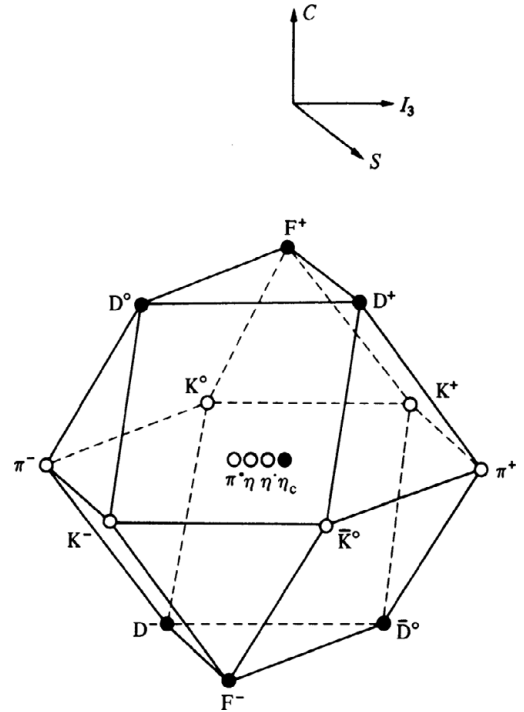


Figure 35.4. The hexadecimetric (16-plet) of spin-0 mesons generated by  $SU(4)$  flavour symmetry. The familiar nonet of  $SU(3)$  flavour is the middle  $C = 0$  plane.

corresponding not only to the excited states formed from the  $cc$  pair, but also to all possible combinations of the charmed quark with the up, down and strange quarks in both mesonic and baryonic configurations and their excited states. In short, the  $SU(3)$  flavour symmetry of the hadrons is enlarged to  $SU(4)$ . Mesons which combine the charmed quark with the up or down antiquarks are denoted the D mesons. These mesons carry explicit charm (i.e. have a non-zero charm quantum number), just as the K mesons carry strangeness. This is in contrast to the  $J/\psi$  itself which, being a  $c\bar{c}$  combination, has the charm of its quark cancelled by the anticharm of its antiquark. There are also mesons consisting of both charmed and strange quarks and antiquarks, which thus possess both charm and strangeness. These mesons were initially designated  $F^\pm$ , but have been recently renamed  $D_s^\pm$ . All possible spin-0 mesons are contained in the hexadecimetric of  $SU(4)$  flavour symmetry, which is illustrated in Figure 35.4. Similarly, there are also new baryons containing both charmed and strange quarks.

With the prospect of such a feast of new particles, both experimenters and theorists busied themselves in the 1970s in confirming the anticipated picture. It often became a race between experimental teams to be the first to detect a particularly tricky candidate, while theorists vied with each other to predict the masses and properties of the particles as accurately as possible. This led to a rapid advance in our understanding of quark behaviour and in the formulation of QCD.

### 35.4 Psychology

What remained a puzzle for some time was the extraordinary size of the resonance (its formation being some 3000 times more probable than the production of a non-resonant  $q\bar{q}$  pair at neighbouring energies), and its extreme narrowness. The  $\psi$  has a width of only 0.002% of its mass compared with the  $\rho$  width of 20% of its mass. By Heisenberg's uncertainty principle this means that the  $\psi$  has a lifetime much longer than that generally associated with hadrons.

This unusual narrowness can be explained in terms of the inhibition of its preferred decay modes because of the masses of the charmed mesons. Its preferred decay mode would normally be expected to be into the charmed mesons  $D^+D^-$  or  $D^0\bar{D}^0$ , proceeding by a quark line diagram rather like that for the decay of the  $\rho^0$  into  $\pi^+\pi^-$  or  $2\pi^0$ , see Figure 35.5(a). However, the  $\rho^0$  decay can proceed only because the  $\rho^0$  mass of 0.77 GeV is larger than the mass of the two-pion state,  $2 \times 0.135$  GeV. The  $J/\psi$  is so narrow because its mass is *less* than that of two charmed mesons. These mesons were detected well after the discovery of the  $J/\psi$  with a mass of 1.86 GeV, thereby

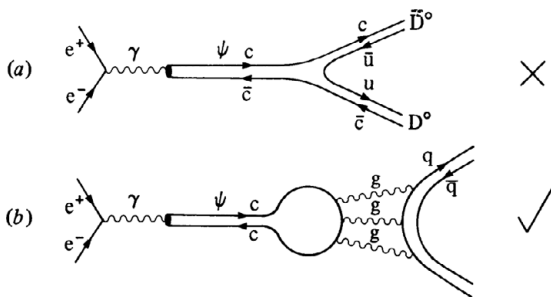


Figure 35.5. The obvious decay mode of the mesons (a) is not possible because the charmed mesons are too massive. The more complicated decay into non-charmed hadrons (b) takes longer.

requiring a particle with a mass of at least 3.72 GeV to produce them.

The  $J/\psi$  can decay only by rather sophisticated means. The  $c\bar{c}$  pair has to annihilate itself into a state of three gluons which must then transform themselves into the observed hadrons by the mechanism of colour confinement similar to that practised by the quarks mentioned earlier, see Figure 35.5(b). Intermediate states of one or two gluons are prohibited by the conservation laws of colour and C parity respectively.

The  $J/\psi$  is but the most obvious of a whole family of mesons consisting of a  $cc$  pair. Some differ only in mass, the differences being due to the increased radial excitation energy of the  $c\bar{c}$  pair. The  $\psi'$  at 3.687 GeV is the lightest of these and, like the  $J/\psi$ , is very

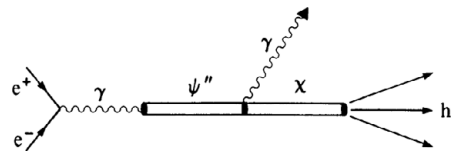


Figure 35.6. If a heavy  $\psi$ -like state emits a photon,  $c\bar{c}$  mesons with new quantum numbers are created.

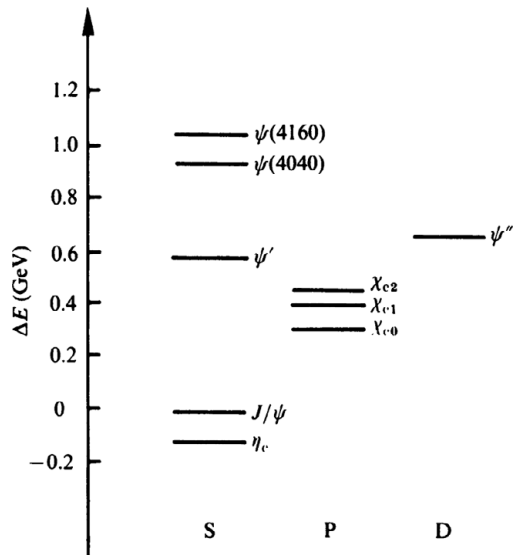


Figure 35.7. The experimentally observed spectrum of  $c\bar{c}$  mesons resulting from the different values possible for the spin and the orbital angular momentum of the constituent quarks. In this notation, S, P and D refer respectively to measured orbital angular momentum equal to 0,  $\hbar$  and  $2\hbar$ .



Figure 35.8. The crystal ball detector at SLAC. Numerous photomultiplier tubes bristle from the surface of the spherical container. They are monitoring the sodium iodide crystals mounted in the interior which detect the photons originating from the interaction point. (Photo courtesy SLAC.)

narrow because it too lies below the  $D\bar{D}$  threshold. The  $\psi''$  is next at 3.77 GeV but this is a hadron of normal width (at 0.7% of mass) as it lies above the  $D\bar{D}$  threshold (but only just!). Above this there are a number of other states.

However, some of the heavier  $c\bar{c}$  mesons have different spin, parity (**P**) and charge conjugation (or **C**-parity, **C**) assignments from those of the  $J/\psi$ . These are denoted  $\chi$  and are produced when one of the heavier members of the  $\psi$  family emits a photon,

thereby allowing the  $c\bar{c}$  pair to change its quantum numbers (Figure 35.6).

A simplified version of the entire  $c\bar{c}$  family (often called ‘charmonium’) is shown in Figure 35.7. Note that  $\eta_c$  with spin 0 is the lightest  $c\bar{c}$  meson. To discover all these ‘secondary’  $c\bar{c}$  states, it is necessary to observe the energies of photons emerging from a process such as that in Figure 35.6. If one energy is preferred above all those possible, this is taken to indicate the mass difference between the heavy  $\psi$ -like particle (the energy of the  $e^+ e^-$  collision) and the secondary  $c\bar{c}$  state with different spin or parity assignments.

To achieve this prodigiously detailed particle-hunting task, experimenters at SLAC built a novel photon detector nicknamed the crystal ball. This consists of a spherical array of sodium iodide crystals pointing towards its centre which is colocated with the interaction region. The sodium iodide crystals are monitored by photomultipliers which can measure the energy deposited in the crystal by an incident photon, see Figure 35.8.

Readers familiar with atomic physics will recognise the pattern of Figure 35.7 as being very similar to the energy level structure of the hydrogen atom. This similarity is understandable because the  $c$  and  $\bar{c}$  have bound themselves together into an exotic sort of elementary particle atom. Recognition of this phenomenon provided an enormous opportunity for particle physicists because such an atomic arrangement of the relatively heavy charmed quarks can be described by well-understood non-relativistic quantum mechanics. The force between the quarks can be formulated as a potential acting in the vicinity of a colour charge, just as in classical electrodynamics an electric potential surrounding an electric charge gives rise to Coulomb’s force law between charges.

The particular form of the potential will determine the splitting of the energy levels or, in the  $c\bar{c}$  case, the mass differences between mesons. As these can be measured experimentally with great accuracy, this can be used to provide a detailed picture of the force between the quarks.

The form of the potential arising from a colour charge which is found to give the most satisfactory match to the spectrum of mass levels is one which combines a simple Coulomb law at short ranges (one corresponding to single gluon exchange in the asymptotically free regime) with an attractive potential rising

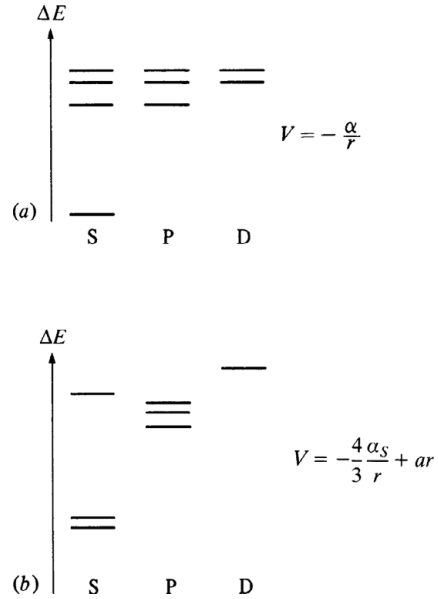


Figure 35.9. The spectrum of energy levels expected from the familiar electric potential is shown in (a). In (b) is shown the spectrum generated by the proposed form of the interquark potential. It is a much closer match to the observed spectrum.

linearly with range at longer ranges, giving rise to the ever-increasing forces of quark confinement. The theoretical pattern of  $c\bar{c}$  mass states generated by this potential is shown in Figure 35.9(b) and, in comparison, the energy levels of positronium, the bound states of  $e^+ e^-$  arising from the Coulomb potential between the two electric charges, is shown in Figure 35.9(a).

So the masses of the  $c\bar{c}$  mesons (sometimes referred to as the spectrum of charmonium) provide direct support for the QCD picture of interquark forces containing both asymptotic freedom at short ranges and confining forces at longer ranges.

### 35.5 Charmed Particles

For several years after the discovery of the  $J/\psi$ , experimenters sought the scores of particles which should be expected to carry explicit charm and their various excited states with ever-increasing spins. These were a lot harder to dig out of the experiments, as they could be found only by searching amongst the final-state hadrons for particular combinations at given masses. When a lot of debris is present in the final state and when the decays of the sought-for particle

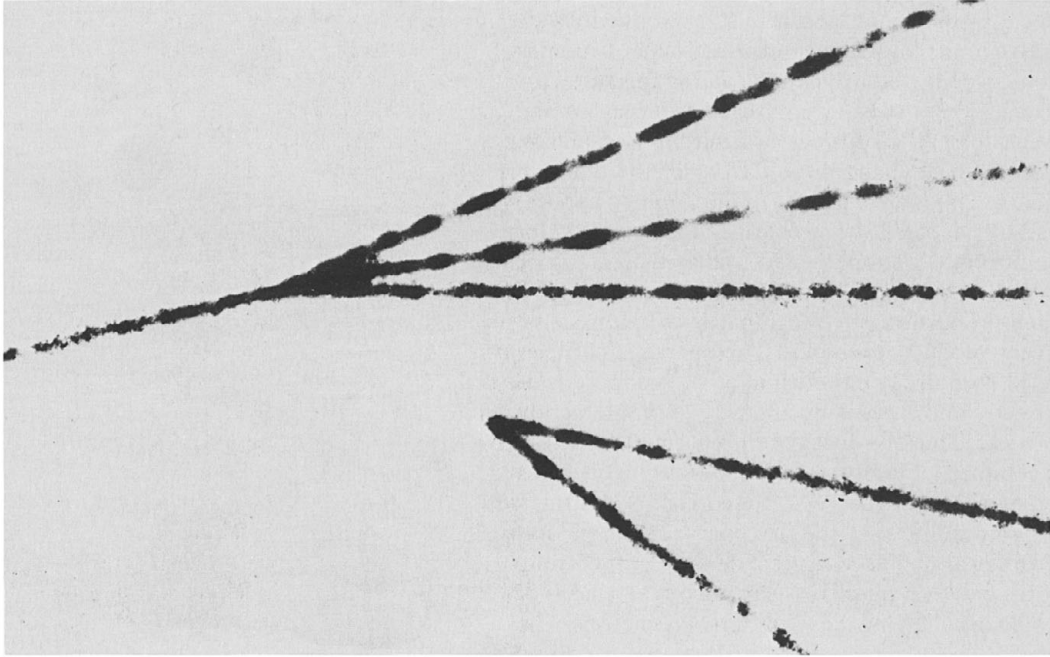


Figure 35.10. A magnified bubble chamber photograph of charmed particles decaying. In the top half, a positively charged charmed meson (track entering from left) decays into three other charged particles. In the lower half, an invisible, neutral charmed meson decays into a pair of charged particles.

are uncertain, this is a tricky business. Eventually, a respectable roll call of the particles was built up which supports their categorisation by  $SU(4)$  flavour symmetry.

Like the strange particles, charmed particles decay by the strong force, emitting pions until they arrive at the lowest-mass charmed state. Charm is conserved by the strong force and so this state, for example a D meson, is obliged to decay by the weak force. This it does by emitting a virtual W boson which changes the flavour of the emitting quark. This is an extension of the Cabbibo hypothesis of Chapter 17. The charmed quark will prefer to turn into a strange quark rather than an up or down, and this is signalled

by the presence of a high proportion of strange particles amongst the decay products of the Ds. These strange particles must then decay either to non-strange mesons or directly into leptons by another weak interaction process. Thus the decay of the charmed particle is a complex laboratory of weak decays involving as many as three in succession, see Figure 35.10.

In summary, the discovery of charm has enabled us to find out a great deal about the strong force between quarks, as carried by the gluons of QCD, by studying the spectrum of  $c\bar{c}$  mesons. The decay of the charmed D and F (i.e.  $D_s$ ) mesons has confirmed our understanding of the weak decays of hadrons as contained in the Glashow–Weinberg–Salam theory.

## Another Generation

### 36.1 Introduction

Soon after physicists had digested the consequences of the  $\psi$  mesons and the charm scheme, the discovery of yet another particle threatened them with elementary particle indigestion. In an experiment similar to Ting's discovery of the  $J/\psi$ , Leon Lederman and his team at Fermilab discovered a new particle in the reaction,

$$p + N \rightarrow \mu^+ \mu^- + X.$$

Lederman and his colleagues observed that this reaction was enhanced slightly for a  $\mu^+ \mu^-$  pair mass of 9.46 GeV compared to its generally declining probability over the neighbouring range, see Figure 36.1. This was taken as the signal of a new, very massive meson resonance consisting of yet another flavour of quark bound to its antiquark. The new meson is denoted by upsilon,  $\Upsilon$ , and its new constituent, the bottom quark, b (after a spirited but doomed effort on the part of a romantic school to call it beauty).

### 36.2 The Upsilon

This interpretation was by no means certain at the beginning and the pN experiment is by no means an ideal reaction in which to study the particle. This is because the hadronic debris  $X$  confuses the final state, and the fact that the very massive  $\mu^+ \mu^-$  pairs are relatively rare makes it difficult to obtain accurate statistics. If its interpretation were correct, then it should be produced also in  $e^+e^-$  annihilations exactly

like the  $J/\psi$  and so this was the obvious way to examine it in more detail. The trouble was that with its mass at 9.46 GeV, the  $\Upsilon$  lay above the energy range of the SPEAR ring at SLAC and *below* the range provided by the new PETRA ring opened at DESY in 1978. Doubtless, the high-energy planners thought that no divine guiding hand would deal such a low card as to stick a particle between 8.4 and 10 GeV

However, it was vital that the  $\Upsilon$  be investigated in the uncluttered environment of  $e^+e^-$  annihilations and so the energy range of the DORIS ring (PETRA's predecessor at DESY) was tweaked to give just enough energy to reach the  $\Upsilon$ .

The  $e^+e^-$  experiments confirmed that the  $\Upsilon$  was indeed a (bb) bound state and confirmed also the existence of its radially excited relative,  $\Upsilon'$  at 10 GeV and  $\Upsilon''$  at 10.40 GeV (Figure 36.2). The width of the states was much harder to establish than that of the  $J/\psi$  as the energy resolution of the storage ring is not as accurate at the very end of its energy range as in the middle. The best value for the  $\Upsilon$  width is about 0.005% of its mass, which indicates that it too, like the  $J/\psi$ , has its preferred decay mode (into explicit bottom mesons) suppressed. It too must annihilate the bottom of its quark with the anti-bottom of its antiquark into a state of three gluons which will then transform into non-bottom hadrons. From measurement of the  $\Upsilon$  width, it is possible to deduce that the most likely charge of the bottom meson is  $-\frac{1}{3}$ , which establishes it as a more massive successor to the

down and strange quarks. The spacing of the masses of the  $\Upsilon$  and  $\Upsilon'$  can be calculated in the same way as the spectrum of  $\psi$  states. The experimental value observed supports the form of the interquark force as described by the potential of Figure 35.9(b).

The existence of yet another flavour of quark of course means that there must exist an entire new family of mesons with explicit bottom for all the various values of isospin, strangeness and charm discussed previously. The  $SU(4)$  flavour symmetry is enlarged to

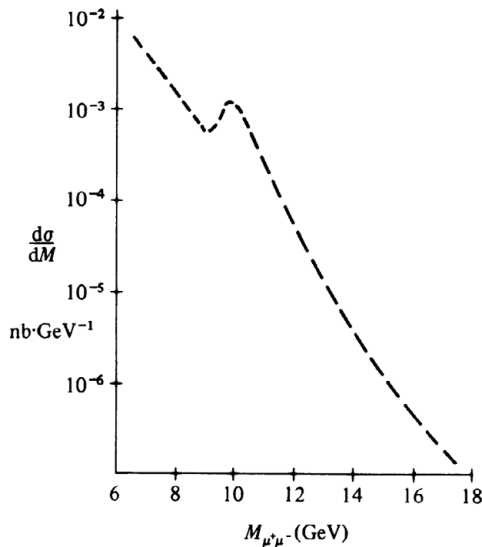


Figure 36.1. The  $\mu^+\mu^-$  mass spectrum in pN collisions, containing the telltale bump of the upsilon.

$SU(5)$  so that the basic multiplet of spin-0 mesons is now expanded from the hexadecimet of Figure 35.4 to a 25-plet. Similarly, baryons with non-zero bottom will augment all the baryonic multiplets. Detection of explicit bottom particles is even harder than that of naked charm as they are much more massive and thus require high-energy collisions. These will contain more debris in the final state from which the suspected decay products of the bottom mesons must be sorted. Despite these difficulties, experiments have detected explicit bottom particles, as shown in Figure 36.3. Bottom particle spectroscopy has provided confirmation of the quark dynamics formulated in the context of the charm spectrum.

In some ways, just the existence of the  $\Upsilon$  meson and bottom quark is of more significance than the details of its properties. For there is no place for the bottom quark in the first two generations. This suggests that it is the herald of a third generation containing yet another quark (denoted, naturally, the top quark) and a new lepton and its neutrino. Indeed, simultaneous with the discovery of the  $\Upsilon$ , evidence for a new lepton was already mounting.

### 36.3 The Tau Heavy Lepton

In 1975, at the time of the  $e^+e^-$  charm experiments, a team of physicists led by Martin Perl, also working on the SPEAR ring at SLAC, reported the existence of 'anomalous  $\mu e$ ' events occurring in  $e^+e^-$  reactions. They suggested that these might signal the existence of a new heavy lepton, denoted  $\tau$ . The 'anomalous  $\mu e$  events' are reactions of the form

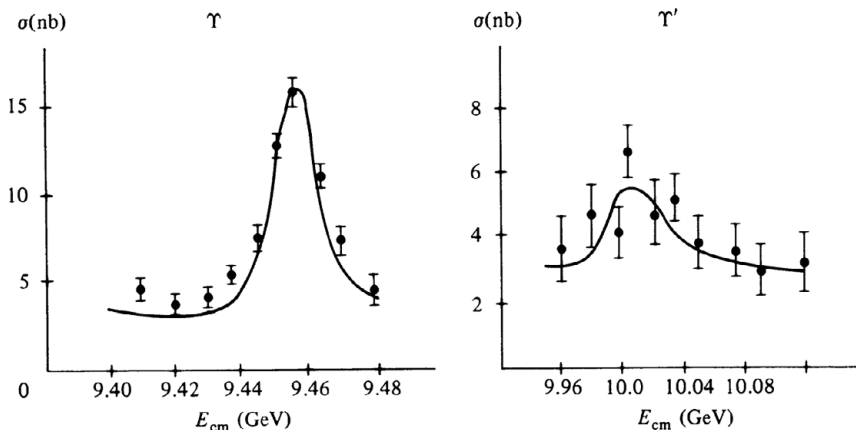


Figure 36.2. Evidence for the  $\Upsilon$  and  $\Upsilon'$  from the total cross-section for  $e^+e^- \rightarrow$  hadrons.



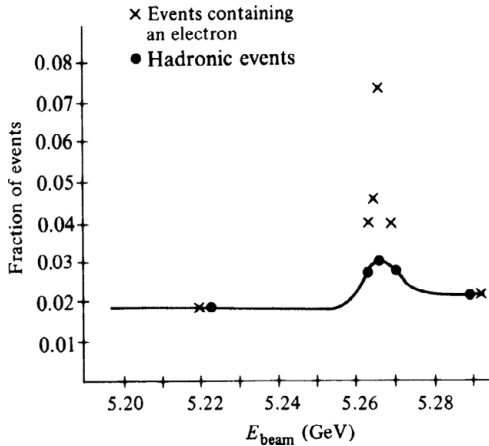


Figure 36.3. Experimental evidence for the production of explicit bottom hadrons. An excess of electron production at the beam energy of an  $\Upsilon$  state suggests that it is decaying into bottom mesons which then produce the electrons in their own weak decays.

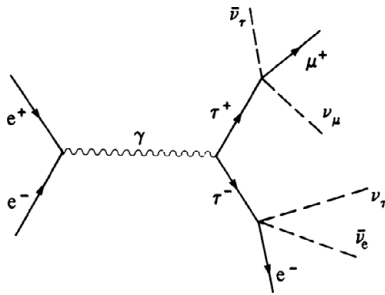


Figure 36.4. Production of a  $\tau^+\tau^-$  heavy-lepton pair in  $e^+e^-$  annihilation gives rise to an ‘anomalous’  $\mu e$  final state.

$e^+e^- \rightarrow e^\pm\mu^\mp + \text{missing energy}$ , and the suggested origin of the final state is that of the separate electronic and muonic decays of the new intermediate pair of heavy leptons (Figure 36.4).

It took some time to establish the truth of Perl’s suggestion, due to several complicating factors. The most serious of these was that the energy threshold for the production of the  $\tau^+\tau^-$  pair is approximately 3.6 GeV (implying a mass for  $\tau$  of about 1.8 GeV). This of course, is very close to the threshold of 3.72 GeV required for the production of a charmed meson pair  $D^0\bar{D}^0$ . As we know, these must decay by the weak interaction and so can quite easily be confused with tau heavy-lepton production and decay. However, in the

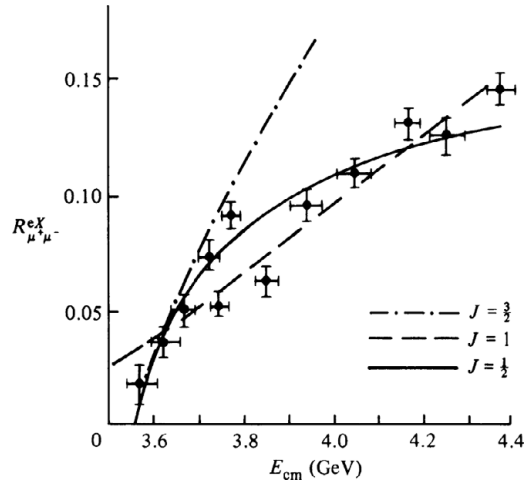


Figure 36.5. The growth of  $\tau^+\tau^-$  production from the threshold as signalled by the ratio of candidate events (those containing an electron and some other charged particle only,  $eX$ ) to known  $\mu^+\mu^-$  production. Note that  $J$  denotes the spin of the  $\tau$ .

case of charmed mesons, one would generally expect other hadronic tracks to be present. It is extremely unlikely that charmed-particle decays will give rise to the final state detected by Perl.

The problems were in ensuring that absolutely no other charged particles had been produced and had slipped past the detector, or that the electrons and muons detected were in fact not hadrons confusing the detectors (misidentification is always possible).

Eventually, Perl was able to place his identification beyond doubt, the final evidence for this being the production of the  $\mu e$  states *below* the threshold for charmed meson-pair production (possible because of the slightly lesser mass of the  $\tau$ ). The evidence for this production is shown in Figure 36.5, which shows the growth of the process away from its theoretical threshold. The shape of the energy dependence also establishes the spin of the  $\tau$  to be  $\frac{1}{2}$ , like that of the electron and the muon, and in contrast to the spin-0 D mesons. Since its confirmation, all the evidence has supported the identification of the  $\tau$  as being a very massive copy of the muon (which is itself simply a massive copy of the electron). Like all leptons, the  $\tau$  does not experience the strong force as do the quarks. However, the  $\tau$  does have one new feature compared with the muon and the electron. Because of its large mass, it can decay into hadrons and this

means that it can add up to one unit to the ratio  $R$ , as defined previously, above its production threshold, over and above the value predicted by the charges of the quarks.

Apart from this new feature, the  $\tau$  behaves exactly like its less massive relatives during interactions. Despite its mass, it shows no deviation from point-like behaviour down to the current experimental limit of  $10^{-18}$  m, and provides us with no hints that the leptons themselves may be composites of even smaller particles.

### 36.4 Completing the Third Generation

It took 25 years after the discovery of the  $\tau$  lepton in 1975 to find the remaining members of the third generation or family. The first to be discovered was the top quark, in March 1995. Tests of the electroweak interaction in  $e^+e^-$  collisions (see Chapter 38) had already indicated that the top quark was very massive and would require a very high centre-of-mass energy if it were to be produced in collisions. It was eventually found at the Tevatron collider at Fermilab, Illinois, a proton–antiproton collider with a centre-of-mass energy of 1.8 TeV. Very occasionally, colliding partons have enough energy to produce (usually via gluons) a  $t\bar{t}$  (top–anti-top) pair which decays almost immediately to a pair of  $W$  bosons and a pair of bottom quarks. These bottom quarks carry away a large amount of momentum in the laboratory frame and consequently move at speeds close to the speed of light. Because of time dilation, their lifetime as measured in the laboratory is much greater than their lifetime of  $10^{-12}$  s as measured in their rest frames and they travel several millimetres before decaying. These events can be observed using a silicon vertex detector surrounding the beam pipe: the tracks from the charged decay products of the  $b$ -quark meet at a point which is displaced from the centre of the beam, where the initial collision occurred. The  $W$  bosons either decay into leptons and neutrinos or jets of hadrons.

The most important experimental signatures of top events are displaced vertices corresponding to  $b$ -quark decays and the presence in the decay products of a large amount of momentum transverse to the beam pipe. This is because the decaying top quarks, which are very massive, impart large momentum to the light decay products which are emitted in all directions. By searching for events with these characteristic signatures and identifying the decay products, two teams using separate detectors at Fermilab (called CDF and D0) were able to obtain conclusive evidence for top quark production. Measurement of the energy and momenta of the decay products allowed the mass of the top quark to be inferred and it did indeed turn out to be very heavy. The current best value is  $172.44 \pm 0.13$  GeV, which is consistent with precision electroweak measurements (see Table 38.1). Quite why the mass of the top quark is so large (the  $b$  quark comes in second with a mass of only 5 GeV!) is something of a mystery. In fact, the mass puts the top quark at about the same mass as an atom of ytterbium with atomic weight of 172 containing, obviously, 172 protons and neutrons and therefore 516 up and down quarks. In quite what sense such a monster could be regarded as a truly point-like elementary particle is open to debate. However, although the possibility of sub-quark structure has occasionally been raised, it has not yet gained any significant traction in the active research community.

The discovery of the last member of the third generation, the  $\tau$  neutrino, did not come until July 2000. It too was discovered at Fermilab by the purpose-built DONUT experiment. This experiment consisted of a neutrino beam (coming from the Tevatron) passing through a three-foot-long target of iron plates sandwiched between layers of emulsion. The neutrino beam contained neutrinos of all types, including  $\tau$  neutrinos. Very occasionally these neutrinos interact, producing a  $\tau$  lepton, which leaves a track about a millimetre long in the emulsion before decaying. Four such events were observed by the DONUT experiment.



## **Part X**

### **The Standard Model**



## *The Model in Summary*

### 37.1 Introduction

The previous chapters have told the long story of the discovery of the various particles and interactions which form what we now call the Standard Model of Particle Physics. The key theoretical elements of the Standard Model were in place by the early 1970s, and the discovery of the W and Z bosons in 1983 convinced physicists that the Standard Model was correct. In the years since then, the Standard Model has been subjected to intense experimental scrutiny. All the constituent particles have been found and a great number of precision tests of the model have been performed. At the time of writing, not one laboratory experiment (barring the discovery of neutrino oscillations – see Chapter 43) has been found to be inconsistent with the predictions of the Standard Model. The model has passed all tests with flying colours, as we shall see in the next chapter. Firstly though, let us summarise the main features.

### 37.2 Summary of the Standard Model

The Standard Model is a gauge quantum field theory, based on the three sacred principles of relativity, quantum mechanics and gauge invariance. There are three distinct sectors of the model, characterised by the spins of the particles in each sector. The principal sector contains the spin-one gauge bosons, which mediate the interactions between all particles. The overall gauge group contains both QCD and the uni-

fied electroweak interaction and is written symbolically as

$$SU(3)_C \times SU(2)_L \times U(1)_Y.$$

The first group,  $SU(3)$ , represents QCD. The subscript  $C$  indicates that the gauge bosons of QCD couple only to colour-charged particles, namely quarks. The eight gauge bosons are called gluons. The  $SU(2) \times U(1)$  part represents the electroweak interaction; the subscripts  $L$  and  $Y$  indicate that the  $SU(2)$  group couples only to left-handed particles and that the  $U(1)$  part couples to weak hypercharged particles. After spontaneous symmetry breaking, the four gauge bosons of  $SU(2) \times U(1)$  become the massive  $W^\pm$  and  $Z^0$  bosons of the weak interaction and the massless photon of QED.

The second sector of the Standard Model is made up of spin-one-half fermions. This sector contains the quarks and leptons which make up ‘ordinary’ matter.<sup>1</sup> These quarks and leptons are grouped into three *generations* or *families*, which are almost identical copies of each other, the only difference being in the masses of corresponding particles in each family. The particles of each family can be further split up into five multiplets, according to the charges they carry with respect to the gauge bosons, or equivalently, how they transform under the three gauge symmetries. The particles within each multiplet are transformed into

<sup>1</sup> Actually, only the up and down quarks and the electron are needed to make all ‘ordinary’ matter, namely atoms.

each other by the gauge symmetries, but particles in different multiplets are not transformed into one another. The multiplets of the lightest fermion family, containing the electron and its neutrino, and the up and down quarks, are shown in Table 37.1.

The left-handed quark multiplet contains, as its name suggests, all the left-handed quarks. It consists of two flavours (a ‘doublet’), with different  $SU(2)$  weak charges (labelled up and down), each of which can have one of three  $SU(3)$  colour charges (labelled red, green and blue), forming a ‘triplet’ representation of  $SU(3)$ . In all, this left-handed quark multiplet contains six states, each labelled by one of three colours and one of two weak charges, up or down.

The next two multiplets are the right-handed up quarks and the right-handed down quarks. Both multiplets contain three colour charges, but in contrast to the left-handed quarks, they do not feel the  $SU(2)$  weak force, which couples only to left-handed fermions.

The fourth multiplet is the left-handed lepton doublet. It contains no colour charge (and therefore the states in it do not feel the strong force) and forms a doublet under the weak  $SU(2)$  force, like the left-handed quarks. The ‘up’ charge is carried by the electron-neutrino, while the ‘down’ charge is carried by the electron.

The fifth and final family multiplet is the right-handed electron. It contains only one state (a ‘singlet’), carrying only weak hypercharge.

In each of the five multiplets, the weak hypercharge is assigned so as to get the measured electric charge for all particles. The rule for determining the hypercharge is that it is given by the average electric charge of particles in the multiplet. So, for example, the left-handed quark multiplet contains three up quarks with electric charge  $+\frac{2}{3}$  and three down quarks with charge  $-\frac{1}{3}$ , giving a hypercharge of

$$\frac{1}{6} \left( 3 \times \frac{2}{3} + 3 \times \frac{-1}{3} \right) = \frac{1}{6}.$$

The hypercharge of each multiplet is displayed in Table 37.1. The assignments look rather ad hoc. We will see in Chapter 44 how the hypercharges are *predicted* in certain theories which go beyond the Standard Model. In total, counting all the different members of each of the five multiplets, we see that the first Standard Model family contains 15 particles. Adding in the other two families gives a total of 45 fermionic particles.

Table 37.1. *The five multiplets of the first fermion family. Colour  $SU(3)$  charges run horizontally, weak  $SU(2)$  charges run vertically and the  $U(1)$  hypercharge is written after each multiplet.*

Multiplet	States	Hypercharge
Left-handed quarks	$\begin{pmatrix} u_L^r & u_L^g & u_L^b \\ d_L^r & d_L^g & d_L^b \end{pmatrix}$	$+\frac{1}{6}$
Right-handed up quarks	$(u_R^r \quad u_R^g \quad u_R^b)$	$+\frac{2}{3}$
Right-handed down quarks	$(d_R^r \quad d_R^g \quad d_R^b)$	$-\frac{1}{3}$
Left-handed leptons	$\begin{pmatrix} \nu_L \\ e_L \end{pmatrix}$	$-\frac{1}{2}$
Right-handed electron	$(e_R)$	$-1$

The third sector of the Standard Model contains a spinless particle – the Higgs boson (or bosons). It is a remnant of the spontaneous symmetry breaking which occurs in the electroweak interaction. The Higgs boson was discovered in 2012 and only now are its properties being measured in detail.

### 37.3 Consistency of the Standard Model

In some senses, the Standard Model has an appealing structure. It is built on very general principles, and incorporates many of the generic features of QFT. In particular, it is satisfying that the Standard Model contains particles of all spins (namely zero, one-half and one) which lead to renormalisable theories. However, in other ways the Standard Model looks to be rather arbitrary. Why, for example, is there such a large number of particles (58 including the Higgs boson)? Why is there so much replication? Why do we ‘need’ three families? Most of these questions have no answer within the context of the Standard Model itself, but there is one that does, namely the question of why the fermions are arranged into families, with five multiplets in each, which are near-identical copies.

The answer comes from a requirement of *consistency* of the theory. Any theory of physics must be self-consistent and the Standard Model is no exception. As we have seen, the biggest obstacle in the search for consistent QFTs is the problem of infinities, which can be solved, for particles of spin less than or equal to

one, by the method of renormalisation. For particles of spin one, an added requirement for renormalisability is that the theory must have a gauge symmetry. This further implies that the spin one bosons must be massless. This appeared to preclude the use of QFT to describe the massive spin one bosons that carry the weak force, until the advent of spontaneously broken gauge theories. It was the proof that these too were renormalisable, in the early 1970s, that heralded their acceptance as serious candidates for theories of particle physics. Towards the end of the 1970s, however, an unexpected cloud appeared on the horizon. It was discovered that gauge theories of the chiral type (such as the Standard Model), in which the gauge coupling to left-handed and right-handed fermions is different, contained an *anomaly*.

Such theories *appeared*, like all gauge theories, to be renormalisable and thus consistent. However, certain loop Feynman diagrams, corresponding to quantum effects, caused the gauge symmetry, built-in at the classical level, to become lost at the quantum level. But the gauge symmetry was crucial for the renormalisability: if gauge invariance was lost, then renormalisability might be lost too. And indeed it was! Particle physics was faced with a disaster.

The dangerous Feynman diagrams consist of corrections to the scattering amplitude for three gauge bosons caused by a loop of virtual fermions (Figure 37.1). In a non-chiral theory, the contributions of left- and right-handed fermions running around the loop automatically cancel. This is no longer true in a theory which is chiral, i.e. in which there is a sense of handedness. However, all is not necessarily lost, because the different chiral fermions in a theory give different contributions depending on their charges, and so there is still a possibility of a cancellation leading to a consistent theory. Such a cancellation looks extremely unlikely though, because of the huge

number of dangerous diagrams, there being one for every combination of three gauge bosons. Amazingly, it was found that, for the gauge group of the Standard Model, all of the diagrams could be cancelled, but *only* if the fermions occurred precisely in complete families! If they did not, the Standard Model would simply not make sense! For the first time, physicists had an explanation for the occurrence of families in the Standard Model: nothing else was possible! Of course, there was (and still is) no explanation for why there should be three families (as opposed to one or a thousand say, which would also be perfectly consistent).

There is another apparent problem associated with gauge theories with chiral fermions like the Standard Model, namely that such fermions must be massless. The reason for this is that a Lagrangian containing such mass terms for chiral fermions cannot be made gauge-invariant.

This too seems to be a catastrophe for the Standard Model, since many of the fermions in the Standard Model are measured to have non-zero masses. The top quark, for example, with mass around 170 GeV, is heavier than most *atoms*. The problem is similar to that encountered with the W- and Z-gauge bosons. Gauge invariance required that they too should be massless, rather than having their observed masses of 80–90 GeV. As we saw in Chapter 21, the solution was provided by the mechanism of spontaneous symmetry breaking – the vacuum expectation value of the Higgs boson gives mass to the gauge bosons. The same thing happens for chiral fermions. Although we have not shown it explicitly, they too can acquire their masses through the Higgs mechanism.

There is, however, a slight difference between gauge boson masses and fermion masses. Whereas the gauge boson masses are related (as in Chapter 22) by

$$M_Z = \frac{M_W}{\cos \theta_W},$$

there are no such predictions in the Standard Model for how the different fermion masses are related: the fermion masses are essentially free parameters in the theory.

Actually, there are a rather large number of these free parameters in the Standard Model – 19 in fact. This is just one reason why we believe that there must be a simpler theory underlying the Standard Model. An analogy can be made with the early

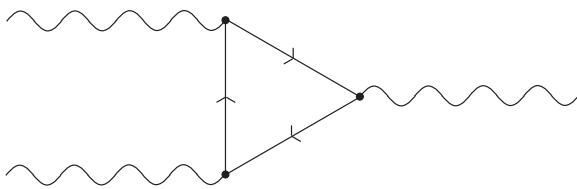


Figure 37.1. An anomalous Feynman diagram with a fermion loop, contributing to three-gauge-boson scattering.



days of chemistry, when the properties of over a hundred elements were catalogued in the periodic table. These properties were later realised to be a direct consequence of the number of electrons in each element. Similarly, the myriad hadrons were found to

be composed of six quarks. Can a similar simplification be achieved for the Standard Model? We will explore this in some detail in later chapters, but first we must discuss the detailed testing of the Standard Model.

## *Precision Tests of the Model*

### 38.1 Introduction

The 1990s saw a decade of particle physics experiments of unprecedented precision, testing the Standard Model in diverse ways and obtaining extremely accurate measurements of the free parameters of the Model. The largest experiments were done at the Stanford Linear Collider (SLC) in California, the Large Electron–Positron Collider (LEP) at CERN in Switzerland, (see Figure 38.1), and at the Tevatron at Fermilab in Illinois. Both SLC and LEP collided electrons and positrons, though in different ways. At SLC, bunches of electrons and positrons were both accelerated to 50 GeV in a *linear* accelerator, before being steered around separate arcs (by a magnetic field) and brought into a head-on collision at a centre-of-mass energy of 100 GeV. In contrast, the LEP collider was a *circular* ring, with bunches of electrons and protons being continually accelerated in opposite directions around the ring. The advantages of a circular collider are that particles can be accelerated repeatedly and that bunches can be made to collide repeatedly as well. There is, however, a serious disadvantage of circular colliders which offsets this. A charged particle moving in a circle radiates photons (this is called synchrotron radiation) and therefore loses energy at a rate proportional to

$$\frac{E^4}{m^4 R^2},$$

where  $E$  is the particle’s energy,  $m$  is its mass and  $R$  is the radius of the ring. Thus, the energy losses become more and more significant as the particle’s energy increases, and the ring must be made larger and larger to compensate. The LEP ring had a circumference of 27 km, running underneath the Franco-Swiss border on the outskirts of Geneva, making it the largest scientific instrument ever built.

SLC and LEP both began by colliding  $e^+e^-$  at, or close to, centre-of-mass energies corresponding to the mass of the Z boson, at 91 GeV. The cross-section for  $e^+e^-$  collisions is enhanced at this threshold, with a characteristic resonance peak. Over a period of a few years, over 20 million Zs were produced around the resonance (Figure 38.2 shows just one event), enabling a very accurate determination of the Z mass and lifetime, and a detailed study of its decay modes. A comparison of the lifetime of the Z (related to the totality of its decay modes) with the observed decay modes enabled the magnitude of hidden decay modes to be inferred. The most important of these hidden decays are decays to neutrinos, and these decay measurements confirmed the number of light neutrinos to be precisely three, as the Standard Model predicts. Figure 38.3 shows the predicted Z resonance peak for 2, 3 or 4 neutrino species.

After exhaustively probing the Z resonance, the centre-of-mass energy of LEP was pushed up to around 160 GeV, above the threshold for production of



Figure 38.1. An aerial view of the CERN accelerator. *Copyright CERN photo.*

*pairs* of  $W^+$  and  $W^-$ . Here again a great deal of precision data were obtained. Because quantum mechanics allows processes in which pairs of virtual particles are produced for a short time before annihilating, even if they are too massive to be produced directly, these measurements on the  $W$  and  $Z$  bosons enabled physicists to make deductions about even heavier particles, as yet unseen. In particular they predicted that the mass of the top quark should be around 150 GeV (it was eventually measured at the Tevatron to be 178 GeV) and that the mass of the lightest Higgs boson should be around 100 GeV (now known to be 125 GeV), with an uncertainty of a factor of two or so.

This caused great excitement, since a Higgs boson of mass 100 GeV *ought* to have been observable at LEP. In the final year or two leading up to the decommissioning of LEP in September 2000, experimentalists worked frantically to tweak up the centre-of-mass energy of LEP in the belief that the Higgs was just around the corner. In the summer of 2000, just as LEP was about to be switched off, a few events were observed with properties characteristic of

those involving the Higgs boson. Figure 38.4 shows one of them.

However, a discovery could not be claimed, since there was simply not enough data to determine unequivocally whether the events really *were* due to the Higgs, or were background effects due to other particles. There followed a great debate as those working on the experiment pleaded for the machine to be kept going, so they could see if they really had found the Higgs. However, the laboratory as a whole was already behind schedule with the construction of the next-generation collider, the Large Hadron Collider (LHC), which was to be built inside the LEP tunnel, once LEP had been switched off. The LHC, due to come online in 2007, would almost certainly find the Higgs and settle the question once and for all. But was it better to take a big risk, and let LEP run for another year or so? In the end, it was decided to press on with the construction of the LHC. As we shall see in Chapter 41, this decision turned out to be the right one: the events seen at LEP were not the Higgs, but rather a statistical fluke.

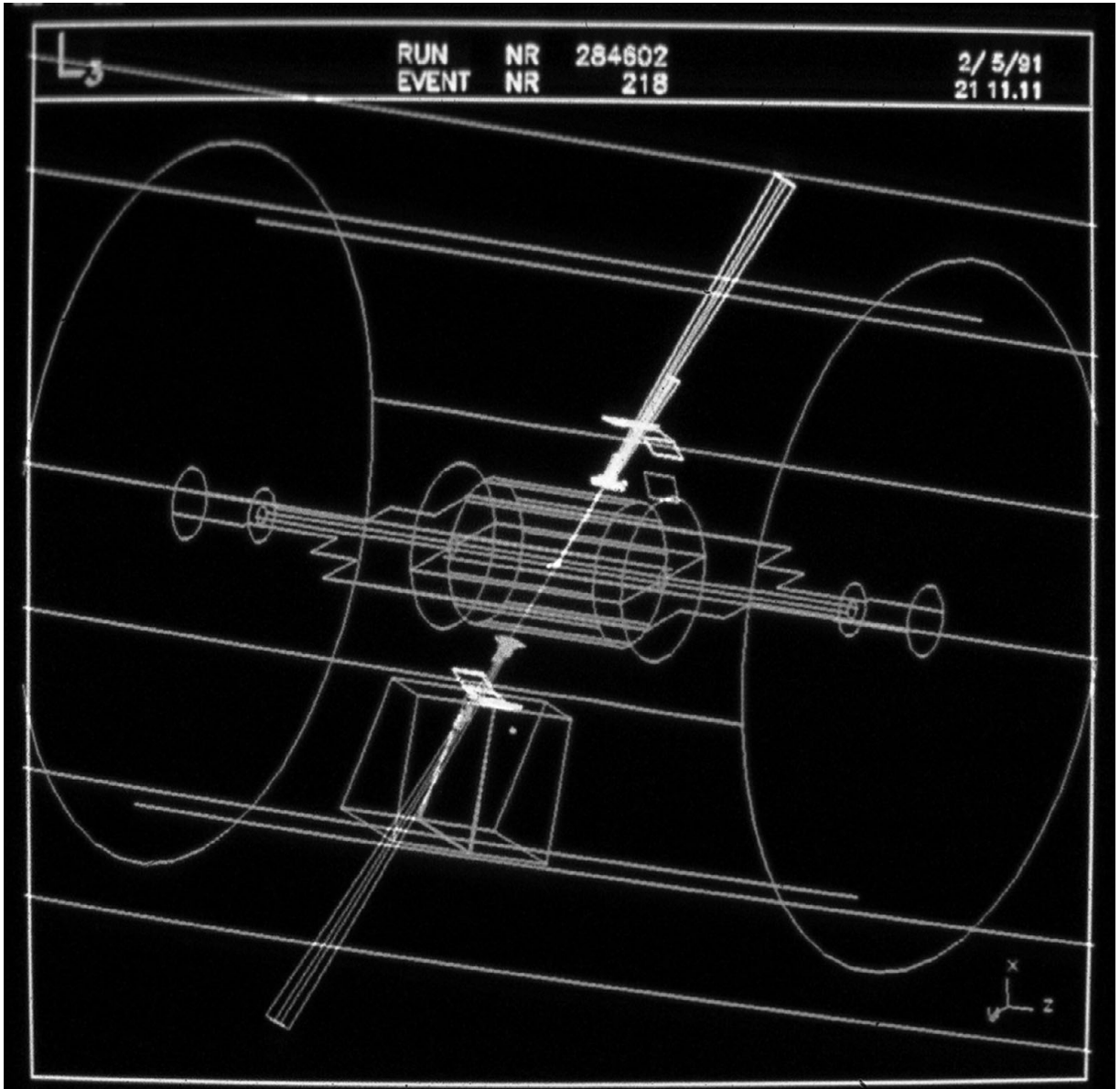


Figure 38.2. Z-boson decay into an electron and a positron observed at DELPHI. *Copyright CERN photo.*

### 38.2 Precision Tests of the Gauge Interactions

We saw in Chapter 37 that the Standard Model is built on the principle of gauge symmetry. It is this principle that ensures a consistent QFT, and so the gauge interactions really represent the core of the Standard Model. This core has just three free parameters, namely the three gauge couplings of the groups  $SU(3)$ ,  $SU(2)$  and  $U(1)$ , which we denote by  $g_3$ ,  $g_2$  and

$g_1$  respectively. If the Standard Model is correct, these three parameters must together account for the totality of experiments testing the gauge sector of the Standard Model. That they do so is truly remarkable, a testament to the triumph of the scientific method.

Over the last 30 years or so, hundreds of such experiments have been performed, some (such as LEP) involving millions of individual measurements. Not

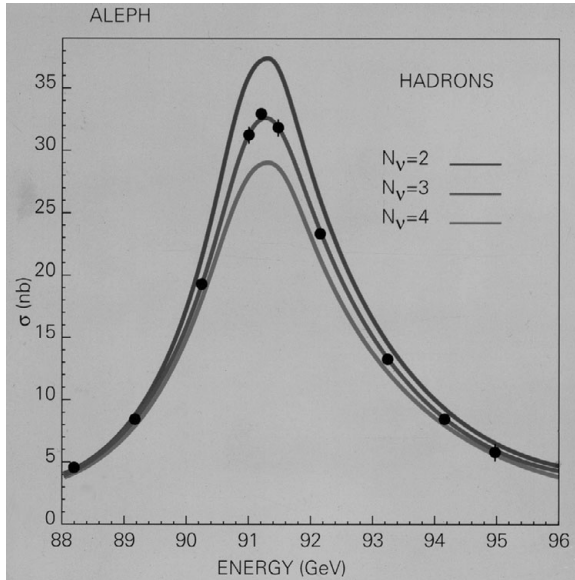


Figure 38.3. The Z-boson resonance measured at ALEPH, with the predictions for varying numbers of light neutrinos superimposed. Copyright CERN photo.

one of these has shown a significant deviation from the Standard Model prediction. The electromagnetic interaction (QED) has long been established as extraordinarily accurate. As Paul Dirac pointed out, QED describes ‘not only all of chemistry, but most of physics as well’ and indeed measurements such as the gyromagnetic ratio of the muon are the most accurate ever performed in the history of science. The measured value is

$$\frac{g_\mu - 2}{2} = 0.001165920(2),$$

where the figure in brackets measures the uncertainty in the last decimal place! The Standard Model prediction is

$$\frac{g_\mu - 2}{2} = 0.001165916(1).$$

There is a minute discrepancy, which may or may not be a hint of new physics.

Tests of the weak interaction (the broken part of the electroweak sector) are not nearly so precise, though there can be no doubt that the Standard Model is correct in the regimes where it has been tested.

Table 38.1. Comparison of precision electroweak measurements with the Standard Model. The first set of measurements all come from independent tests near the Z resonance peak. The remaining ones are the mass and width of the W boson, the mass of the top quark as measured at the Tevatron and the contribution to the electromagnetic coupling coming from the hadronic vacuum polarisation.

Observable	Measurement	SM fit
$m_Z/GeV$	$91.1875 \pm 0.0021$	91.1873
$\Gamma_Z/GeV$	$2.4952 \pm 0.0023$	2.4965
$\sigma_h/nb$	$41.540 \pm 0.037$	41.481
$R_l$	$20.767 \pm 0.025$	20.739
$A_{fb}^l$	$0.0171 \pm 0.0010$	0.0164
$\mathcal{A}_l(SLD)$	$0.1513 \pm 0.0021$	0.1480
$\mathcal{A}_l(P_\tau)$	$0.1465 \pm 0.0033$	0.1480
$R_b$	$0.21644 \pm 0.00065$	0.21566
$R_c$	$0.1718 \pm 0.0031$	0.1723
$A_{fb}^b$	$0.0995 \pm 0.0017$	0.1037
$A_{fb}^c$	$0.0713 \pm 0.0036$	0.0742
$\mathcal{A}_b$	$0.922 \pm 0.020$	0.935
$\mathcal{A}_c$	$0.670 \pm 0.026$	0.668
$\sin^2 \theta_W$	$0.2324 \pm 0.0012$	0.23140
$m_W/GeV$	$80.425 \pm 0.034$	80.398
$\Gamma_W/GeV$	$2.133 \pm 0.069$	2.094
$m_t/GeV$	$178.0 \pm 4.3$	178.1
$\Delta\alpha_{had}$	$0.02761 \pm 0.00036$	0.02768

Table 38.1 shows a recent compilation of nearly 20 very different tests of the weak interaction, all in agreement with the Standard Model to within a per cent or so.

Finally, the strong interaction (QCD) has also been tested exhaustively in a variety of experiments performed over a wide range of energies. A convenient way to present the data is in terms of the predictions for the running coupling constant  $g_3$  as a function of the energy scale (Figure 38.5). Again, the agreement is quite remarkable. All of this data shows beyond doubt that the Standard Model theory of gauge interactions, and the pattern of gauge symmetry breaking, is correct. With just three numbers, all of the data are explained. There remain two areas in which the Standard Model

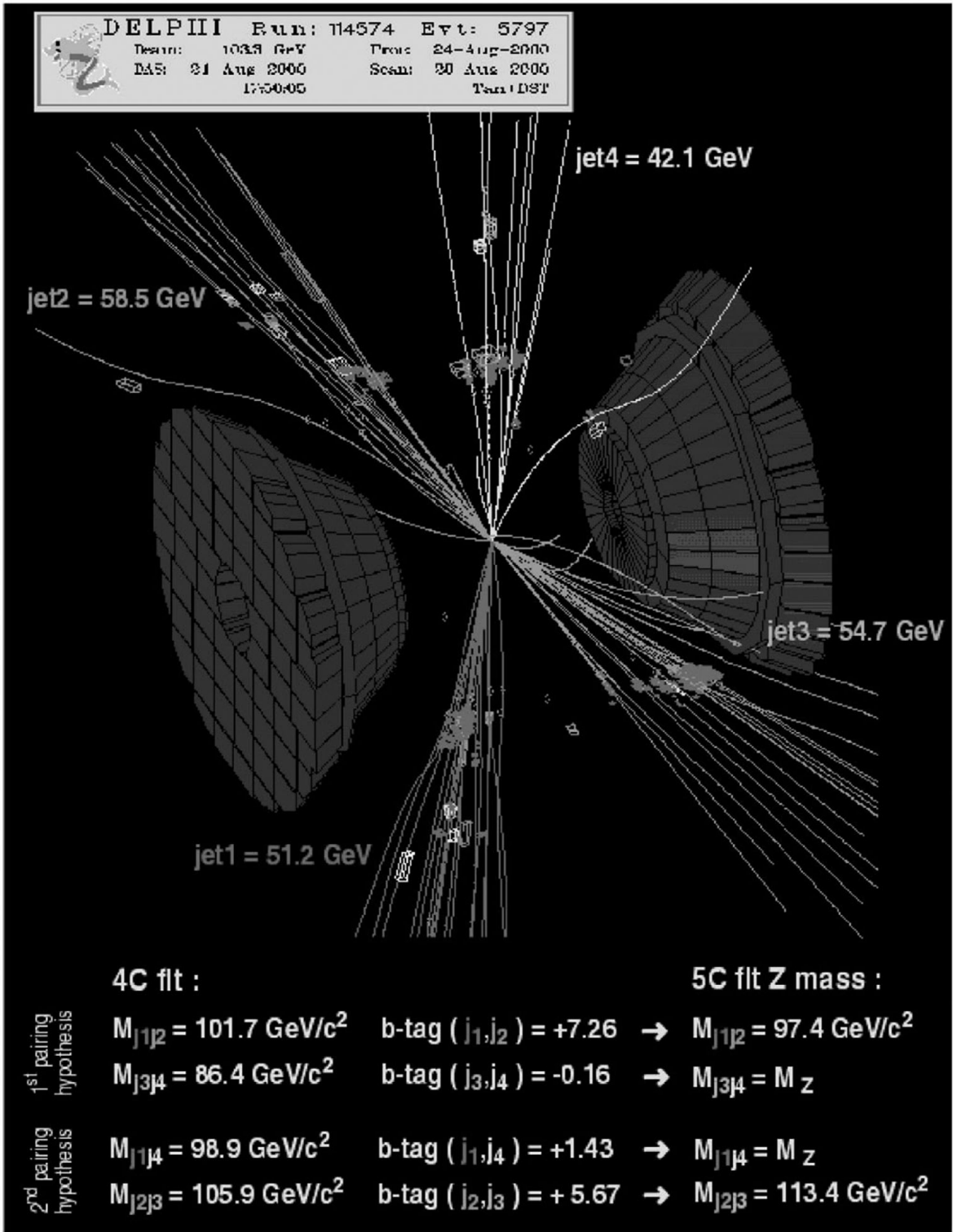


Figure 38.4. Candidate Higgs event collected at DELPHI in August 2000, compatible with the associated production of a Z boson and Higgs boson of mass 113 GeV. A different pairing of the jets could lead to an interpretation compatible with the production of two Z bosons. *Copyright CERN photo.*

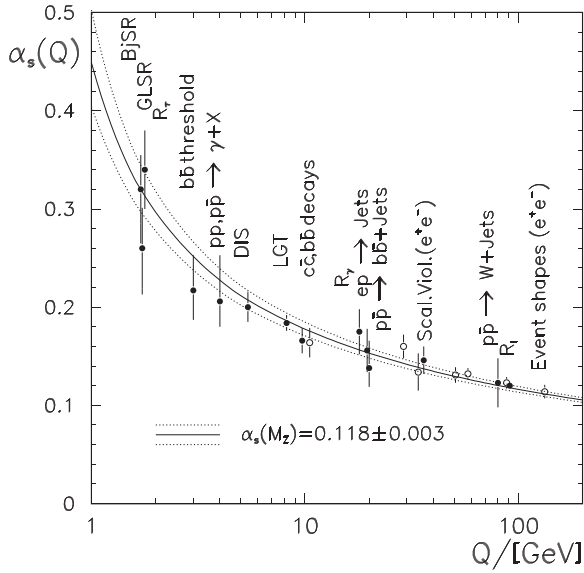


Figure 38.5. Measurement of the running of the strong coupling at various scales by different experiments. Reproduced from *ICHEP '96: Proceedings*, pp. 91–102. Edited by Z. Ajduk and A. K. Wroblewski. River Edge, NJ, World Scientific, 1997. 2 vols.

needs to be tested. The first is the fermion sector – the precise measurement of quark and lepton masses and flavour mixing angles and so on. The second is the Higgs sector. The discussion of experiments in these sectors forms the content of the following chapters.

## Flavour Mixing and CP Violation

### 39.1 Introduction

When we first encountered the weak interactions, one of the things we learnt was that they do not conserve quark flavour and that this results in mixing between the Standard Model fermion families. For example in the decay  $K^- \rightarrow \pi^0 e^- \nu_e$ , a strange quark emits a  $W^-$  boson and changes into an up quark, which combines with an anti-up quark in  $K^-$  to produce a  $\pi$  meson. The  $W$  then decays leptonically. The flavour-changing decay of the strange quark is suppressed by a factor of  $\sin^2 \theta_c \simeq 0.04$ , where  $\theta_c$  is the Cabibbo angle. Flavour-changing can also occur between the first and third families and between the second and third families. Note that all of these flavour-changing processes occur via the exchange of charged  $W$  bosons. Flavour-changing processes involving neutral  $Z$  bosons are not observed. That is, there are no *flavour-changing neutral currents*.

In the context of the Standard Model, we say that as far as the weak interactions are concerned, a  $u$  quark can be converted into a  $d$ ,  $s$  or  $b$  quark by emitting a  $W^+$ . The same is true for the  $c$  and  $t$  quarks, so one can write the schematic equations

$$\begin{aligned} u &= V_{ud}d + V_{us}s + V_{ub}b, \\ c &= V_{cd}d + V_{cs}s + V_{cb}b, \\ t &= V_{td}d + V_{ts}s + V_{tb}b, \end{aligned}$$

where the  $V_{ij}$  parameterise the relative amplitudes for the transition from quark  $i$  into quark  $j$ . The  $V_{ij}$

form a matrix, called the *CKM matrix* after Cabibbo, Kobayashi and Maskawa. The parameters of the CKM matrix are free parameters of the Standard Model. However, if there are to be no flavour-changing neutral currents, then the CKM matrix must satisfy a *unitarity* condition. This is just the extension of the GIM mechanism (see Chapter 23) to three families. Unitarity implies relations between the nine parameters  $V_{ij}$  of the CKM matrix. The most important of these conditions as far as experiments are concerned is the relation

$$V_{ub}V_{ud}^* + V_{cb}V_{cd}^* + V_{tb}V_{td}^* = 0.$$

Essentially, this equation says that the sum of three complex numbers must equal zero. Equivalently, if the three complex numbers are represented as lines in a plane, then the three lines must close up to form a triangle. This is the *unitarity triangle*. By a re-parameterisation, one can put one vertex of the triangle at the origin in the plane and one at the point with coordinates  $(1, 0)$ . The only remaining degree of freedom is the position of the apex of the triangle (Figure 39.1). Finding the position of the apex experimentally is very important for testing the consistency of the Standard Model, as we shall see below.

### 39.2 CP-Violation in the Standard Model

There is some remarkable physics associated with the unitarity triangle. Most striking is that, if the triangle really is a triangle (rather than three points



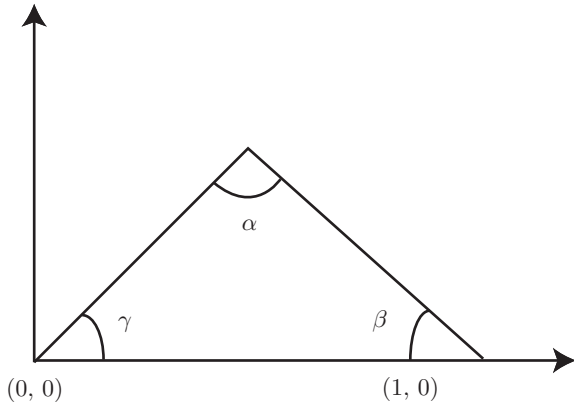


Figure 39.1. The unitarity triangle. The position of the apex is fixed by various experiments and provides a consistency test of the Standard Model.

in a straight line), then the Standard Model exhibits CP violation. Recall from Chapter 14 that CP is the combined discrete symmetry of charge conjugation (in which each particle is replaced by its antiparticle) and parity (reflection of space). Up until the 1950s, both of these were assumed to be sacrosanct, as was the symmetry under time-reversal, denoted  $T$ . The discovery that both  $C$  and  $P$  were violated maximally in weak interactions came as a huge shock. In some ways, the discovery in 1964 by Christensen, Cronin, Fitch and Turlay that the combined symmetry CP was also violated came as even more of a shock, because at the time no one knew how to even write down a theory without CP invariance.

It turns out that in order to have a theory without CP invariance, the theory must contain couplings which are complex numbers. However, that is not really the end of the story. In most theories with complex couplings, the imaginary parts can be scaled away by some redefinition of the parameters. This means that the complex couplings are not really physical. In particular, in the weak interaction theory with only two fermion families (such as the one that existed in 1964), CP invariance is automatic. However, by adding a third fermion family, CP violation becomes a possibility. In the Standard Model, which does indeed have three fermion families (and a single Higgs doublet), there is just one CP-violating parameter, contained in the CKM matrix.

So the Standard Model does allow for the CP violation which is observed in nature, but it contains only one free parameter which causes it. As far as

Table 39.1. Neutral flavoured mesons, which can mix with their antiparticles.

Meson $M^0$	Quark Content
$K^0$	$\bar{s}d$
$D^0$	$\bar{c}u$
$B_d^0$	$\bar{b}d$
$B_s^0$	$\bar{b}s$

testing the Standard Model goes, this is very desirable. There are many processes in which one might be able to observe CP violation experimentally, and if the Standard Model is correct, all of them must be explained by a single number. If the data from one of these experiments were inconsistent with the data from any other, then we would have evidence that the Standard Model is incomplete. Along with neutrino experiments (to be discussed in Chapter 43) and the Large Hadron Collider (to be discussed in Chapter 40), experiments searching for CP violation (as well as other rare effects in flavour physics) offer one of our best hopes for discovering physics beyond the Standard Model.

### 39.3 CP-Violation Experiments

The most important experiments searching for CP-violation effects involve neutral, flavoured mesons. These are made up of a quark and an antiquark of different flavours, but equal and opposite charge. The different possibilities are listed in Table 39.1. Each pair of neutral flavoured mesons,  $M^0$  and  $\bar{M}^0$ , are CP conjugates. This means that the operation of CP turns one into the other. For example,

$$CP(M^0) = \bar{M}^0.$$

Now because the neutral flavoured mesons have the same electric charge (zero), they can mix quantum-mechanically. This is analogous to the mixing of electrons in the famous double-slit experiment. Classically, the electron goes through one slit or the other, but quantum-mechanically it goes through both, and the ‘two’ electrons can interfere. Thus, the states  $M^0$  and  $\bar{M}^0$ , in which the mesons are produced by the strong interactions, are not necessarily the same states in which they propagate through space. Indeed, for a CP-invariant theory, the propagating states must be invariant under CP. It is straightforward to construct

these states from the CP conjugate states. They are given by

$$M^0 + \overline{M}^0 \text{ and } M^0 - \overline{M}^0.$$

To check that these are CP-invariant, we act on them with CP:

$$\begin{aligned} CP(M^0 + \overline{M}^0) &= M^0 + \overline{M}^0, \\ CP(M^0 - \overline{M}^0) &= -(M^0 - \overline{M}^0). \end{aligned}$$

Thus, the action of CP on each of these states produces the same state (up to a factor of  $\pm 1$ ). For a theory which is not CP-invariant, the propagating states can be different, since they need not be invariant under CP.

There are three ways in which CP violation can show up in the physics of neutral flavoured mesons. The first is CP violation in the propagating states, as discussed above. The second is CP violation in CP-conjugate decay processes. This is signalled by a difference in the rates for a decay, say  $M^0 \rightarrow f$ , and its CP conjugate  $\overline{M}^0 \rightarrow \overline{f}$  and is also known as *direct* CP violation. The third possibility occurs when both  $M^0$  and  $\overline{M}^0$  can decay to the same state, which must therefore be its own CP conjugate ( $f = \overline{f}$ ). Then one can get interference between the decay processes  $M^0 \rightarrow f$  and  $\overline{M}^0 \rightarrow \overline{f}$ .

CP violation can in principle show up in all three ways and experiments have been designed to search for all of them. The CP violation originally observed in K mesons has now been shown to be predominantly (but not entirely) of the first kind. CP violation of the third kind is particularly suitable for comparing the Standard Model with experiment. Theoretical predictions for most processes involving neutral flavoured mesons are hard to calculate because of inherent hadronic effects. These involve QCD processes at strong coupling, where one is unable to use the tools of perturbation theory. Tremendous advances have been made recently in doing the necessary calculations numerically using a computer. In this scheme, known as *lattice gauge theory*, continuous space-time is replaced by a discrete lattice of points. However, these calculations require unprecedented computing power, and the errors introduced by the discretisation process are rather large. However, some processes of the third kind are such that the strong interaction effects cancel out. The so-called *golden decay mode* for B mesons,  $B_d \rightarrow J/\Psi + K_s$  is particularly useful, because both the theoretical and experimental errors

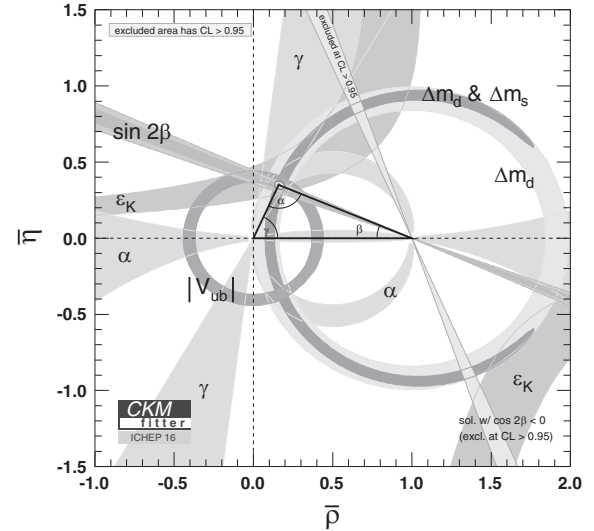


Figure 39.2. Combined data on the unitarity triangle. The position of the apex is fixed by a number of different experiments and must lie in the intersection of the shaded regions, while the golden decay mode gives a precise measurement of the angle  $\beta$  of the lower right-hand vertex. *Courtesy of CKM fitter group.*

are small. Intensive study of this decay mode has led to a precise determination of the angle  $\beta$  at the lower right-hand corner of the unitarity triangle in Figure 39.1. Figure 39.2 shows the latest global fit to the experimental data. The diverse variety of complementary experiments, together with the heroic efforts in the last decades to reduce the experimental and calculational uncertainties, have led to an impressive agreement with the predictions of the Standard Model.

### 39.4 B-Physics Experiments

The first decade of the millennium saw two experiments, BaBar at Stanford and Belle in Japan, studying neutral B mesons. Both of these *B-factory* experiments collided electrons and positrons together at a centre-of-mass energy equal to the mass of the  $\Upsilon$  resonance. These resonances have quark content  $b\overline{b}$  and decay to either  $B^+B^-$  or  $B^0\overline{B}^0$ . In order to look for CP violation, it is necessary to measure the decay rate of  $B^0$ s or  $\overline{B}^0$ s as a function of time. The *B-factory* experiments do this using an ingenious idea put forward by the Peruvian physicist Pier Oddone in 1987. The colliding electrons and positrons are contained in two separate storage rings, with different

energies. When they collide, the centre-of-mass frame does not coincide with the laboratory frame, and so the B mesons, which are approximately stationary in the centre-of-mass frame, are moving in the laboratory with large momentum. Because of the time dilation effect of special relativity, the lifetimes of the B mesons measured in the laboratory are significantly enhanced and the physical separation of the two decaying B mesons (typically a centimetre or so) can be measured, using a *silicon vertex detector*. The time of flight can then be inferred.

One looks for decays in which one of the B mesons decays into the mode being studied and the other decays into something which enables it to be identified unambiguously as either  $B^0$  or  $\bar{B}^0$ . For example, if the B meson decays leptonically, then the sign of the charge of the lepton determines the type of B meson. With this information, one knows the type of the other meson, and how long it took to decay. By analysing a number of events with different decay times, one can measure the time asymmetry of the decays and the relevant CKM parameters.

BaBar and Belle both observed millions of B-meson decay events. Figure 39.3 shows a golden decay event observed at BaBar. Although BaBar has now ceased data taking, the Belle experiment has been upgraded to a new version, which will allow approximately 40 times more data to be recorded. Belle II began taking data in late 2018.

The Large Hadron Collider, discussed in the next chapter, includes a dedicated *B*-physics experiment, LHCb, which also probes the physics of B mesons, but differs in that they are produced via hadronic collisions. While the hadronic environment makes observing rare decays more difficult, this is offset by the far greater rate at which B mesons are produced. Thus, LHCb provides a foil to BaBar and Belle, with greater sensitivity in certain processes and lesser in others.

Among the standout results obtained by BaBar, Belle, and LHCb are the observation of CP violation in a variety of B-meson systems. This included observation of *direct* CP violation in decays involving charmed mesons, such as  $B^- \rightarrow D^0 K^-$ , which allowed, for the first time, a clean measurement of the angle  $\gamma$  in the unitarity triangle, represented by the lower left-hand vertex in Figure 39.2.

Compared to the tiny amount of CP violation observed in K-meson experiments, the amount of CP violation observed in B-meson systems is rather large. This is in precise accord with the predictions of the

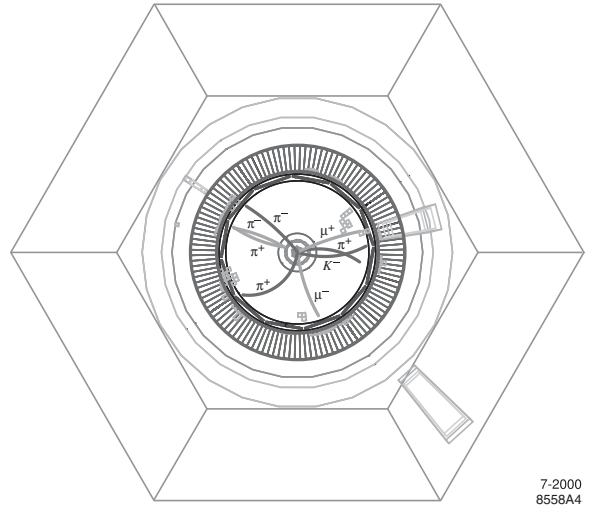


Figure 39.3. Computer reconstruction of a ‘golden event’ in the BABAR Detector. An electron and positron have annihilated at the centre of the vertex detector in this cross-sectional view, producing a B and an anti-B meson. One of them decays into a pair of muons and a pair of pions, while the other (the ‘tagging’ B) decays into a kaon and three pions. *Courtesy of Stanford Linear Accelerator Center.*

Standard Model. As we shall see in Chapter 50, this makes it difficult to explain the huge matter–antimatter asymmetry in the Universe today.

The real power of the current generation of *B*-physics experiments lies arguably not in their ability to make more precise measurements of the Standard Model parameters, but rather in their ability to look for deviations from the Standard Model in completely different places. This, physicists believe, will be the key to discovering the new physics, beyond the Standard Model, which is needed to resolve various puzzles in our understanding of the Universe. We discuss these in depth in Parts XI and XII.

### 39.5 K-Meson Experiments

In 2015, the NA62 experiment began taking data at CERN. This experiment studies the decay properties of charged K mesons. The main goal is a measurement of the decay rate for the process  $K^+ \rightarrow \pi^+ \nu \bar{\nu}$ , which will enable a measurement of the CKM matrix element  $V_{td}$  to a precision of better than 10%. But, as for the *B*-physics experiments, the real benefit of this experiment is that it provides us with a completely different way to test the Standard Model and, in doing so, to search for the physics that lies beyond it.

## *The Large Hadron Collider*

### 40.1 Introduction

In the last few chapters, we saw that, by the turn of the millennium, a wealth of evidence had been amassed pointing to the Standard Model as a correct and consistent theory of particle physics. Indeed, no experiment we have discussed so far has been found to be inconsistent with the Standard Model. That will have to wait until the next chapter. However, a key element of the Standard Model was missing. Despite great hopes (and desperate efforts!) at LEP, no experiment had yet detected the Higgs particle which was believed to be responsible for the spontaneous breakdown of gauge symmetry, and for giving masses to the gauge bosons and fermions. The Higgs boson is, in a sense, the keystone of the Standard Model; while its discovery would be a triumphant confirmation of the Standard Model in its entirety, its absence would leave a significant question mark hanging over the theory.

The existence of the Higgs boson had been predicted at the end of the 1960s, so why had no experiment yet found it? The answer must simply be that the Higgs boson (or bosons) is too massive to have been produced in previous experiments. So, in order to have a chance of seeing the Higgs, higher energies were needed, and a new particle physics collider: the Large Hadron Collider.

### 40.2 Historical Constraints on the Higgs Boson Mass

Early generations of experiments had shown that the Higgs boson must be massive. But how massive?

The best constraint came from the LEP experiment at CERN, which in its final results of 2001 excluded a Higgs boson of mass below 114.4 GeV. So in order to detect the Higgs, a machine was needed which was capable of producing Higgs bosons of this mass or above. But there was a problem: What if the Higgs boson is much heavier than 114.4 GeV, say a TeV or more? Then we would still have no chance of detecting the Higgs, even with the next generation of experiments. Fortunately, there was a way in which physicists could estimate an upper bound for the mass of the Higgs boson.

As we have seen throughout this book, physical processes which we observe in experiments (such as scattering amplitudes) receive quantum corrections, corresponding to Feynman diagrams with loops of virtual particles. The contribution of such diagrams decreases as the masses of the virtual particles increase. In particular, all Standard Model processes have quantum corrections with loops containing virtual Higgs bosons. The sizes of these quantum corrections due to Higgs bosons are determined by the mass of the Higgs. Thus, by precision measurements of Standard Model processes, one can estimate the Higgs mass. A combined estimate suggests that the most likely value for the Higgs is around 117 GeV, which is just above the lower bound of 114.4 GeV obtained from the LEP data! The combined Standard Model data suggest that LEP really was close to finding the Higgs (as we saw in Chapter 38, many believed that LEP may have actually seen it!) and that

the Higgs was probably just around the corner. This gave physicists great confidence that the Higgs boson could be found by the next generation particle physics experiment, the Large Hadron Collider.

### 40.3 The Large Hadron Collider Concept

The Large Hadron Collider (or LHC) was built at CERN, in the tunnel which formerly housed LEP. In contrast to LEP, LHC collides not electrons and positrons, but protons (and occasionally heavy ions, such as lead). The energy losses due to synchrotron radiation are much lower for these heavier particles (see Chapter 38) and this enables much higher collision energies to be reached. The beam energy of LHC was designed to be 7 TeV, corresponding to a centre-of-mass energy of 14 TeV.

The construction of the LHC involved massive engineering and computing challenges, as well as financial ones, with a total budget of around 7.5 billion euros. The magnetic field used to guide the protons around the tunnel has a strength of 8.3 Tesla. Huge electric currents (2000 Amps) are required to generate these magnetic fields. In order to carry such currents without power loss, superconducting cables are used, which must be cooled to temperatures just a few degrees above absolute zero with liquid helium. The computing challenges are particularly acute. High-energy collisions between protons typically generate hundreds of secondary particles. Consequently, vast amounts of data must be stored and analysed. It is estimated that tens of petabytes of data are generated per year at the LHC. If stored on compact discs, one year's worth of data would result in a stack twice the size of Mount Everest and it would require the equivalent of  $10^5$  of today's highest-performance personal computers to process data at this rate!

In order to deal with such a huge amount of data, a new concept in scientific computing was employed, known as the *Grid*. The Grid is a global computing infrastructure, based on 170 computing sites spread around Europe, North America and Asia and connected by a super-high-bandwidth telecommunications network, which processes and stores the LHC data. The Grid allowed, for the first time, computing power around the world to be pooled, making it possible to perform computations that no single existing computer could ever hope to perform. Such distributed computing power has applications in many other areas of science and technology, and it is

fair to say that the Grid has resulted in a revolution in computing and telecommunications, just like the World Wide Web (see the end of this chapter).

### 40.4 Construction Timeline

Conceived in 1984 and commissioned in the mid-1990s, work on the collider did not actually begin until 2001, and delays in the construction of the superconducting dipole magnets that bend the beams meant that beams of particles did not begin to circulate until 10 September 2008. The first particle collisions were planned for 30 September, but on the 19th, disaster struck! A problem with faulty electrical connections between the superconducting magnets led to a quench (in which the magnets revert from the superconducting to the ordinary state) which led to extensive mechanical damage. The need to repair the damage and to prevent subsequent incidents led not only to a substantial delay before the first particle collisions occurred (on 23 November 2009), but also to significant limitations on the physics programme. It was decided to carry out a first physics run from 2009–13 at a much lower energy of 3.5–4 TeV per beam (7–8 TeV total), before a two-year shutdown in which the electrical connections were improved and the magnets re-trained to allow a higher energy run with a beam energy of 6.5 TeV (13 TeV total). The machine has been running at this energy without problems since April 2015, although it seems unlikely that the initial design energy of 14 TeV total will be reached.

### 40.5 The LHC Experiments

Five separate experiments make use of the LHC: ATLAS, CMS, ALICE, LHCb and TOTEM. ATLAS and CMS consist of multi-purpose detectors which probe physics at the high-energy frontier. Table 40.1 shows the event rates per year for various processes for either of the two detectors, compared with the total number of such events observed in all previous experiments. The first four rows contain previously observed Standard Model processes. The number of such events observed at the LHC is overwhelmingly large compared to the numbers observed previously, allowing even more precise tests of the Standard Model. The fifth row concerns the Higgs boson, and shows (for a mass around the observed value) around  $10^5$  are produced per year! In fact the LHC was designed to be able to detect a Higgs boson with any

Table 40.1. *Expected event rates at the LHC for various processes (including new physics) and comparison with existing experiments.*

Process	Events per year	Total events at earlier facilities
$W \rightarrow e\nu$	$10^8$	$10^4$ LEP, $10^7$ Tevatron
$Z \rightarrow e^+e^-$	$10^7$	$10^7$ LEP
$t\bar{t}$	$10^7$	$10^4$ Tevatron
$b\bar{b}$	$10^{12}$	$10^9$ Belle/BaBar
Higgs boson	$10^5$	0
Gluino pairs, mass 1 TeV (Ch. 45)	$10^4$	0
Black holes from Large Extra Dimensions, mass > 3 TeV (Ch. 60)	$10^3$	0

mass between the LEP lower bound of 114.4 GeV and around a TeV, meaning that the Higgs of any plausible mass would be discoverable at the LHC. The LHC was also designed to be able to measure the mass of the Higgs to within a per cent or so, and to test a crucial prediction of the Standard Model: that the coupling of Higgs bosons to other particles should be proportional to the masses of those particles. This prediction follows from the fact that it is the Higgs boson which gives particles their masses, and is a key signature of the Higgs mechanism.

The LHC is far more than just a ‘Higgs-hunting’ machine however. The ATLAS and CMS detectors are also able to search for new physics, beyond the Standard Model, which we shall discuss in great detail in later chapters. In particular, LHC is able to look for *supersymmetry* (see Chapter 45), *large extra dimensions* (Chapter 60), and evidence of *Higgs compositeness* (Chapter 46). The ALICE detector studies collisions between heavy ions (such as lead) in the hope of observing the deconfinement of quarks and the formation of a new state of matter, the quark–gluon plasma, discussed in Chapter 30.

The LHCb experiment focuses on the physics of  $b$  quarks, discussed in the previous chapter, providing complementary measurements to the electron–positron  $B$ -factory experiments already underway (BaBar and Belle). This too allows for further precision tests of the Standard Model, in particular regarding the CKM matrix and CP violation.

The TOTEM experiment measures deep-inelastic scattering and diffraction processes, as well as providing a calibration for the other experiments.

All in all, the LHC and its constituent experiments, which probe the frontiers of many aspects of particle physics, are the current cornerstone of particle physics experiments and will remain so for at least the next decade or so.

#### 40.6 CERN and the World Wide Web

It is now well known that the World Wide Web was originated at CERN by the Oxford physicist Tim Berners-Lee in 1989. Although not in the mainstream of fundamental physics itself, it demands a place in ‘The Ideas of Particle Physics’ as perhaps the most striking example of a scientific spin-off in recent history. It is perhaps not too great a hyperbole to observe that simply managing the data and dataflows of high-energy physics experiments led directly to the internet revolution of the last decade or so.

Although not precisely quantifiable, the financial market value of the Web in the form of Web-enabled enterprises such as Amazon, e-Bay and Google, as well as the value added by the Web to the rest of enterprise and the wider world, currently amounts to many hundreds of billions of pounds, dollars or euros; thereby providing ample financial justification, if such were ever actually required, of the merits of truly fundamental research.

The invention of the Web resulted from the meeting of the variety of the world’s data networks with the computing and data storage requirements of high-energy physics experiments. Involving, as they do, contributions from dozens of universities and government laboratories, all focused on the few major experimental centres such as CERN in Switzerland

and SLAC and Fermilab in the US, the flows of both experimental data and computational code around the globe are immense.

The physical internet itself had its beginnings in the interconnection of a variety of scientific data networks which had been created following the invention of packet-switching in the mid-1960s. Foremost amongst these was the ARPANET of the US Department of Defense. In these networks, data are shipped over conventional phone lines or fibres in small packets which are then reassembled at the intended destination to form the original message. Transmission across these networks had become a routine feature of international physics from the 1970s onwards. But using the networks was a complex task hampered by a cumbersome system for addressing the messages.

However, the origin of the Web was in Berners-Lee's desire simply to manage the information within CERN itself. The relatively rapid turnover of staff,

experiments and data, and the physical distribution of all of these elements in CERN itself and in distant universities and laboratories, meant that the standard hierarchical forms of database design would have been exceptionally difficult to navigate, given the complexities of the system. Very early on in his CERN career in 1980, Berners-Lee created his own relational programme ENQUIRE, precisely to keep track of the many elements in CERN on which he needed to keep data. Key to the development of the Web was the concept of hypertext, first introduced in 1965 by the futurologist, Ted Nelson. This was conceived of as non-sequential text elements which could themselves be used as navigational beacons in conventional sequential text. The hypertext concept achieved a gradual acceptance in the database industry to the point that by the late 1980s several hypertext editing programs were commercially available.

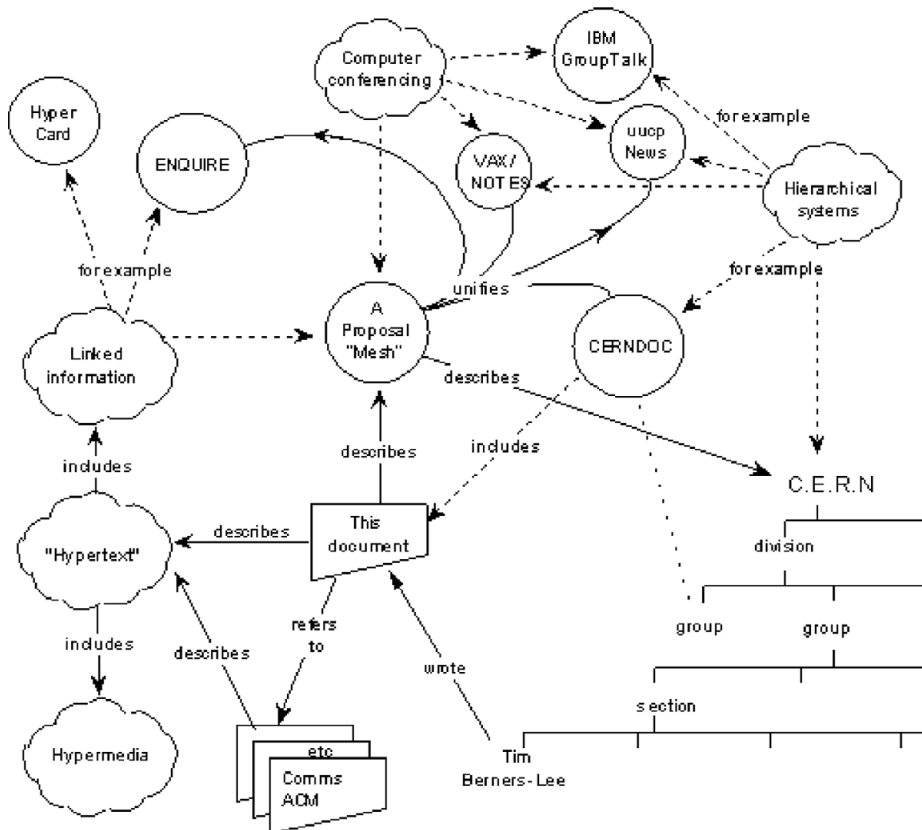


Figure 40.1. Berners-Lee's proposal for the Web.



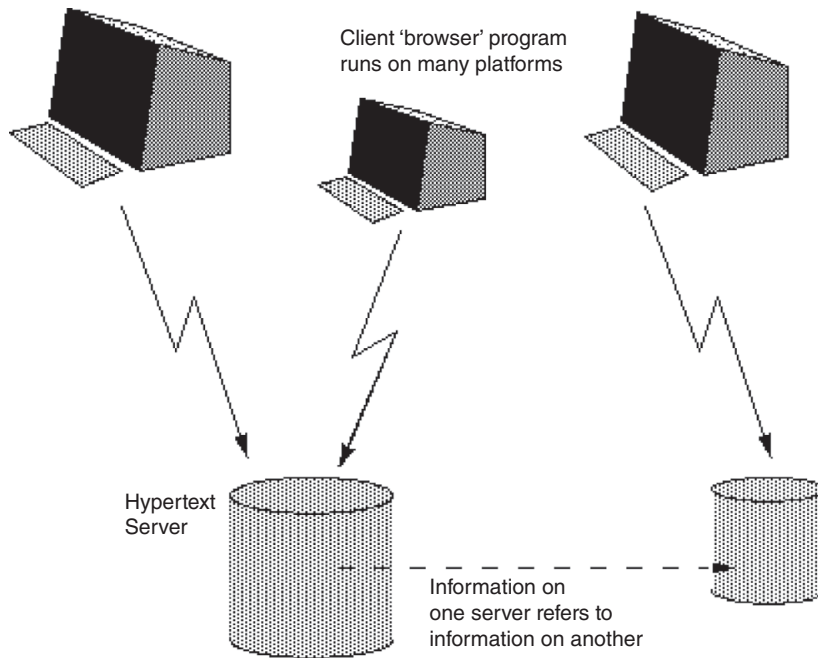


Figure 40.2. Client/server model for a distributed hypertext system.

The effort to create the Web began in earnest in 1989 with Berners-Lee's first formal proposal to CERN managers, subsequently reissued with Robert Cailliau in 1990. The figure used in the proposal to illustrate the document's own relationships with other elements both in the world of CERN IT and the real world is shown in Figure 40.1. The creation of the core code for the Web was achieved quickly by a small team led by Berners-Lee and Cailliau in the last quarter of 1990. This required writing the code of a hypertext editor in the role of Web browser, and its integration with the TCP/IP protocols of the internet

allowing communication between Web client and Web servers. By creating Web client front-ends for the range of machines used at CERN, it was thus possible to deploy the first Web application by Christmas Day 1990; this was the CERN telephone directory at [www.info@cern.ch](mailto:www.info@cern.ch).

Since then the Web has grown into the global phenomenon we know today well documented, as we might expect, not least of all in Berners-Lee's own book (see bibliography in Appendix D) but also at sites such as [www.w3.org](http://www.w3.org). Figure 40.2 shows a client/server model for a distributed hypertext system.



# 41

## *Discovery and Properties of the Higgs Boson*

### 41.1 Introduction

As we saw in the last chapter, the LHC had been designed so that, if the Higgs boson was there, the LHC would be able to find it! In this chapter, we tell the story of the search for the Higgs boson and the momentous announcement of its discovery on 4 July 2012. We also describe the rich programme of Higgs physics (which continues to this day) that followed as a result.

### 41.2 Decays of the Higgs Boson

As we have detailed in previous chapters, the Higgs boson plays a special rôle in the Standard Model, in that it gives particles their mass. This is true not only for the W and Z electroweak gauge bosons, but also for the three families of quarks and leptons that make up matter. Indeed, the mechanism of spontaneous symmetry breaking discussed in Chapter 21 allows the Higgs boson to give mass to any particle to which it couples. What is more, the theory predicts that the mass that results is directly proportional to the strength of the coupling.

This proportionality allows us to gain a rough understanding of the collider physics properties of the Higgs boson. Let us look first at how it decays. Because the Higgs couples to many different particles, it can, potentially, decay in many different ways. Naively, since the Higgs couplings are stronger when the other particles are heavier, we might expect decays to the heaviest particle, i.e. the top quark, to dominate.

But this decay process is in fact forbidden, because the Higgs couples only to a *pair* consisting of a top quark *and* an anti-top quark. These have a combined mass of c. 350 GeV, so the decay of a Higgs boson of mass 125 GeV (which turns out to be the true mass) to such a pair is energetically forbidden. The only way for the decay to take place is if both the top quark and antiquark are produced virtually, but this leads to a large suppression of the decay rate.

So, in fact, the dominant decay of a Higgs boson of mass 125 GeV is to a pair of bottom quarks (which are the heaviest particles of which a pair is nevertheless lighter than the Higgs boson itself). Roughly 60% of Higgs bosons decay in this way. Unfortunately, such a decay is very difficult to observe in a hadron collider, where many secondary hadronic particles are typically produced in collision events. The only hope of seeing it is when the Higgs boson is produced in association with an (easily observed) W or Z boson. This happens infrequently, and so evidence for the decay of the Higgs to *b* quarks was not found until 2017.

The next most prolific decay of the Higgs is to a pair of W bosons. Because  $2m_W > 125 \text{ GeV} > m_W$ , this is a process in which one of the two W bosons is necessarily produced virtually, with a consequent suppression of the decay rate. In fact, roughly 20% of Higgs bosons decay in this way. The W bosons that result subsequently decay into either a pair of quarks or a pair of leptons. The former are hard to see in a hadron collider and the latter are hard to see

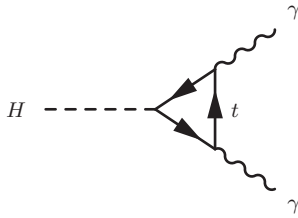


Figure 41.1. Feynman diagram with a loop of top quarks, contributing to the process  $H \rightarrow \gamma\gamma$ .

in any collider, because one of the leptons is always an invisible neutrino. The leptonic contributions nevertheless made a small, but significant contribution to the discovery in 2012.

Paradoxically, the main decay channels that contributed to the Higgs boson discovery in 2012 are processes which occur only very rarely. One of these is the decay to a pair of Z bosons (which occurs around 1% of the time) and the other was the decay to a pair of photons (which occurs less than 0.1% of the time). The contributions, though tiny, are significant because the decay products are easy to see: not only are the decay products non-hadronic (if the Zs both decay to leptons), but also one can use them to reconstruct the Higgs as a resonance in the centre-of-mass spectrum. Thus, even just a few events may suffice to allow a discovery to be claimed.

The fact that a Higgs boson can decay to a pair of photons, even weakly, seems paradoxical, because photons have no mass, and so it seems that the Higgs should not couple to them at all! But this neglects the fact that the Higgs particle can couple to photons via a loop of virtual top quarks, as in Figure 41.1.

Of the other possible decay channels, the only one seen so far is the decay to a pair of  $\tau$  leptons, observed in 2017.

All of the decay modes allow for measurements of the decay rates, all of which are consistent with Standard Model predictions.

### 41.3 Production of the Higgs Boson

Since protons are mostly made up of light quarks (up and down) and massless gluons, Higgs boson production rates at the LHC tend to be low. By far the largest comes from the *gluon fusion* process, in which a pair of gluons annihilates to produce a Higgs boson via a diagram similar to that in Figure 41.1, but with the direction of time reversed and with the photons replaced by gluons.

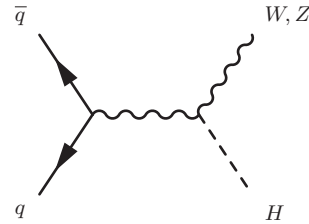


Figure 41.2. The Higgs-strahlung production process.

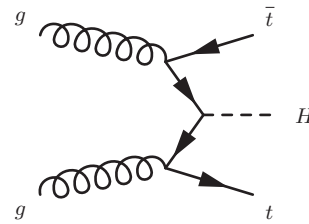


Figure 41.3. The  $t\bar{t}H$  production process.

Higgs boson production can also arise via a variety of other processes, but at much lower rates. These processes can, nevertheless, be important, either because they lead to other, more easily detected, particles produced in tandem with the Higgs, or because they allow us to measure different properties of the Higgs boson. As an example of the former, the production of Higgs bosons in association with W and Z bosons via the *Higgs-strahlung* process in Figure 41.2 allowed the decay into bottom quarks to be observed, while the  $t\bar{t}H$  production process in Figure 41.3, observed in 2018, allowed measurements of the Higgs coupling to top quarks to be performed for the first time.

### 41.4 Discovery of the Higgs Boson

After initial hints at the LHC in 2011 for a slightly heavier Higgs boson (with a mass of 140 GeV) turned out to be spurious, the discovery of a Higgs boson with a mass of 125 GeV was eventually claimed, simultaneously, by the ATLAS and CMS experiments, on 4 July 2012.

As described above, the main evidence for the discovery came from the observation of resonances at an invariant mass of around 125 GeV in the spectrum of pairs of Z bosons (followed by decays of each Z boson into a pair of either electrons or muons) and pairs of photons. Figure 41.4 shows the ZZ spectrum

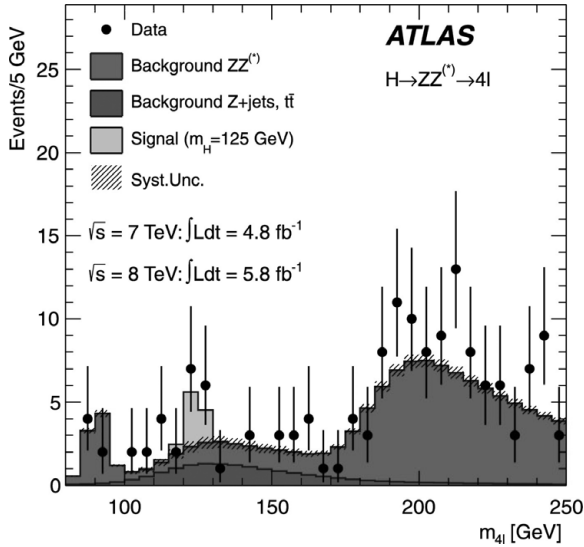


Figure 41.4. Invariant mass spectrum of  $ZZ$  obtained by ATLAS, showing the Higgs boson as a resonance near 125 GeV. The Standard Model prediction for a Higgs boson mass of 125 GeV is also shown. Reproduced from ATLAS Collaboration, Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC, *Physical Letters B*, **716** (2012) 1, <https://doi.org/10.1016/j.physletb.2012.08.020>

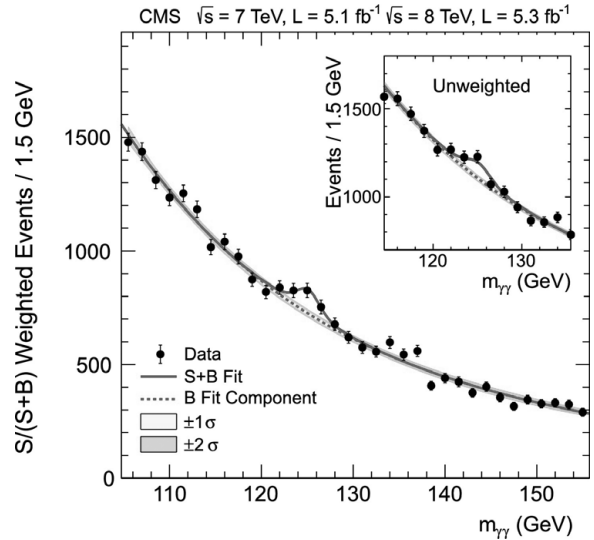


Figure 41.5. Invariant di-photon mass spectrum obtained by CMS, showing the Higgs boson as a resonance near 125 GeV. The Standard Model prediction is shown. The data in the main figure have been re-weighted to make the resonance more visible, while the inset box shows the unweighted data. Reproduced from: CMS Collaboration, Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC, *Physical Letters B*, **716** (2012) 30, <https://doi.org/10.1016/j.physletb.2012.08.021>

obtained by ATLAS, while Figure 41.5 shows the di-photon spectrum obtained by CMS.

The discovery of the Higgs boson was followed by the award of the Nobel Prize to François Englert and Peter Higgs, two of the surviving discoverers of the mechanism of spontaneous symmetry breaking, in October 2013.

### 41.5 Properties of the Higgs Boson

The discovery of the Higgs boson opened the door to a new experimental programme in particle physics, namely that of making detailed measurements of the Higgs boson’s properties and comparing them with the predictions of the Standard Model.

The mass of the Higgs boson is not predicted in the Standard Model, but is rather a free parameter. But once it has been measured, there are no free parameters left in the theory, meaning that the results of all other possible measurements can be predicted (at least if the required calculations are tractable!). To give an example, once the Higgs mass is known to be 125 GeV, it is possible to predict the production

cross-section and decay rates of the Higgs boson to all other particles, and these may then be compared with the results of experimental searches in the various decay channels. Thus far, a quantitative agreement is observed in all possible channels, though the measured precision on the rates is no more than 20% at best, and is often much worse.

What other properties of the Higgs boson can we test in experiment? Firstly, the observation of decays only to pairs of fermions (rather than an odd number), confirms that the Higgs boson is indeed a boson, rather than a fermion. As such, its spin could, a priori, take any integer value. But spin one (i.e. a massive vector boson) is ruled out by a mathematical theorem of Landau and Yang, which says that a massive vector boson cannot decay to two photons. Similarly, spin 2 (or larger) is ruled out by studying the angular distributions of the four decay products that arise in Higgs decays to pairs of W or Z bosons, which in turn decay to pairs of leptons or quarks. The same observables also confirm that the Higgs boson has

positive, rather than negative, parity. The fact that the Higgs boson decays in two photons, each of which have  $C = -1$ , indicates that it must have  $C = +1$ . All of this is consistent with the Standard Model.

#### 41.6 The Future of Higgs Physics

While one of the future goals of the LHC is to improve, as much as possible, the precision of existing measurements in the Higgs sector, there is also a hope that genuinely new tests of the Standard Model can be carried out.

One hope, for example, is that yet further decay modes can be observed. For example, the sensitivity on the  $Z\gamma$  and  $\mu^+\mu^-$  modes is now at the level of just a few times the Standard Model prediction.

A more speculative hope is that it may be possible to measure the strength of the Higgs coupling to itself. In particular, the Standard Model contains interaction processes featuring either three or four Higgs bosons. While the processes involving four Higgs bosons seem completely hopeless, there is some hope that, given enough data, the LHC may ultimately be sensitive to the triple Higgs coupling, via observation of Higgs pair production processes such as those in the diagram of Figure 41.6. Observation of these processes would allow us to test directly that the scalar

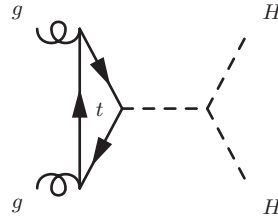


Figure 41.6. Feynman diagram with a loop of top quarks, contributing to Higgs boson pair production.

potential in the Standard Model has the form required for spontaneous symmetry breaking.

One can also search for processes where the sensitivity is too low to test the Standard Model prediction. Observation of a signal would provide convincing evidence for theories of physics that go beyond the Standard Model, and which we discuss in Chapter 42 onwards. Most spectacular amongst these would be flavour-changing processes such as decay of the top quark via the Higgs,  $t \rightarrow Hc$  or a lepton-flavour-violating decay of the Higgs itself, e.g.  $H \rightarrow \tau\mu$ . Such processes are forbidden in the Standard Model, and so observation of just one of them, even with a tiny rate, would provide a clear signal that a new theory is needed.



**Part XI**  
**Beyond the Standard Model**



## *Reasons to Go Beyond*

### 42.1 Introduction

Part X of this book described the triumph of the Standard Model, culminating in the 2012 discovery of the Higgs boson. An intense effort over a period of decades has led to all of the constituent particles of the Standard Model being found and their detailed properties and interactions with each other have been shown to be in precise agreement (often to an accuracy of one part in a thousand or more) with the predictions of the theory.

So, the question arises: is particle physics a ‘done deal’? As we shall see in the remainder of this book, the answer is a resounding no! For one thing, as we shall see in the next chapter, there is direct, definitive experimental evidence for physics beyond the Standard Model, in the form of non-vanishing neutrino masses and mixings.

In fact, there are a number of reasons why physicists believe that there must be physics beyond the Standard Model, and lots of it! Some of these, like neutrino masses and mixings, are based on concrete experimental evidence, while some are based more on theoretical puzzles. We list them here, in no particular order. In later chapters, we shall discuss in detail theories which go beyond the Standard Model in an attempt to resolve these mysteries, and the ongoing experiments which may or may not corroborate them.

- *neutrino masses and mixings*  
We shall see in the [next chapter](#) that, whereas the masses and mixings of quarks can be (and

are) explained by the Standard Model, neutrino masses and mixings cannot. The simplest explanation involves new physics at an enormous energy scale, around  $10^{14}$  GeV.

- *the presence of dark matter*  
As we shall learn in Chapter 51, roughly a fifth of the energy density of the Universe is known to be made up of *dark matter*, which is electrically neutral, colourless and non-baryonic. The only such particles in the Standard Model are neutrinos. The [next chapter](#) shows that they do have a mass, but they are too light to make up more than a few per cent of the observed dark matter. There have been many suggestions for new particles, beyond the Standard Model, making up dark matter. These are the subject of intense current experimental searches.
- *inflation*  
We will show in Chapter 48 that many recent cosmological observations suggest that the early Universe underwent a period of rapid inflation. Nothing in the Standard Model can explain this.
- *the observed abundance of matter over antimatter*  
The Universe appears to be full of matter, with very little antimatter. As we will review in Chapter 50, this requires, among other things, more CP violation than is present in the Standard Model.
- *the inability to describe physics at Planckian scales*



Einstein's theory of general relativity describes gravity as a classical, rather than a quantum theory. In fact, general relativity also makes sense as a theory of quantum gravity, but only at length scales much greater than the Planck scale. Beyond that we need a theory of quantum gravity, such as string theory, described in Chapter 57 onwards.

- *dark energy and the cosmological constant problem*

While dark matter makes up 20% of the energy density of the Universe, normal matter (such as protons and electrons) makes up only 5%. The nature of the remaining 75% is completely unknown. It could simply be *vacuum energy*, represented by a *cosmological constant* in the Lagrangian, but then its value is  $10^{120}$  times smaller than a naive prediction. We discuss this further in Chapter 52.

- *the Higgs hierarchy problem*

Just like the cosmological constant, the measured value of the Higgs mass, 125 GeV, is far smaller than a naive prediction based on the existence of much higher energy scales, such as the Planck scale, at which gravity becomes quantum mechanical.

- *the comparable values of matter, radiation and vacuum energy densities in the Universe*

The Universe is observed to contain roughly comparable amounts of matter, radiation, and vacuum energy densities. These three quantities scale with different power laws during the Universe's evolution, so why are they roughly the same today?

- *the structure in fermion masses and mixings*

The masses and mixing parameters of quarks and leptons vary over huge ranges (e.g. the top quark mass is  $10^5$  times larger than the electron mass), but also show a degree of regularity (e.g. the other fermion masses are distributed roughly evenly in between). In the Standard Model they

are just free parameters, so why do they exhibit this structure?

- *the smallness of measured electric dipole moments*

These violate CP, but are at least  $10^{-10}$  smaller than a naive prediction. As we shall see in Chapter 47, one solution for this is the *axion*, which also provides an explanation for dark matter.

- *the comparable size of the three gauge couplings*

These are all numbers of order one and different from each other, but not *so* different. As we shall see in Chapter 44, this can be quantitatively explained by the theory of *grand unification*.

- *the quantisation of electric charges*

The Standard Model contains the hypercharge gauge group, and the values of the hypercharges of the individual quarks and leptons could have *any* real number values. Why then do we only find simple, rational ratios between them? This can also be explained by *grand unification*.

- *the number of fermion families*

Why are there three fermion families? Why not two or 1000?

- *the number of space–time dimensions*

Why do we live in four space–time dimensions and not greater or fewer? For that matter, why are there three space dimensions and one time dimension? Why not two space and two time dimensions?

An explanation (together with an experimental confirmation, of course) of any one of these issues would surely merit a Nobel Prize, not least because they are such deep issues, but also because it seems so hard to extend the Standard Model in such a way as to furnish a solution, without contradicting one of the many successful experimental tests of the Standard Model that we have described in earlier chapters. We now discuss some of the many possibilities suggested by theorists, together with their experimental status.

## *Neutrino Masses and Mixing*

### 43.1 Introduction

We first met the neutrino back in Chapter 5, where we saw that it was postulated by Pauli in order to avoid violation of energy conservation in beta decays of radioactive nuclei. The neutrino was integral to Fermi's theory of weak interactions – he gave them their name, meaning 'little neutral one' – and now takes its place in the Standard Model on an equal footing with the other leptons and quarks. Because the neutrino carries no electric charge, and because the weak interaction is so weak, the probability of an interaction involving neutrinos is extremely small. This means that neutrinos are extremely difficult to detect. They are, to a large extent, 'invisible'. Few people realise, for example, that *billions* of neutrinos pass through their bodies every second!

Because neutrinos are so difficult to detect, experiments designed to elucidate their properties are exceedingly difficult to implement, typically involving huge detectors and very low event rates. For this reason, we know very little about neutrinos, far less than we know about the other fermions. We know (from the Z-boson decay width – see Chapter 38) that there must be three light neutrinos in the Standard Model, one for each of the fermion families. We also know that the only neutrinos we observe are left-handed, with extremely small upper limits on their masses. All of this is consistent with the Standard Model. But in the last few decades, a variety of neutrino experiments have been performed which

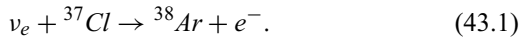
reveal a lot more about the nature of neutrinos. Most importantly, these experiments reveal that the neutrinos are in fact massive (though their masses are miniscule) and that, like the quarks, the different neutrino flavours mix. These discoveries perhaps do not seem all that great, but we will see that they imply that the Standard Model is incomplete and hint at new physics at incredibly high energy scales. They represent the first concrete evidence that there is physics 'beyond the Standard Model'. First though, let us discuss the physics of neutrino masses and mixing.

### 43.2 The Solar Neutrino Problem

The original motivation for current neutrino experiments comes from the Homestake experiment begun in the late 1960s in South Dakota by Ray Davies (who shared the 2002 Nobel Prize for Physics for his work). The experiment was designed to detect the neutrinos produced in the Sun's core as a result of the nuclear reactions which power the Sun. Once produced, the weakly interacting neutrinos simply fly straight out of the Sun and into the cosmos, interacting very occasionally on the way. The neutrinos thus allow us to 'see' right into the solar core!

The Homestake experiment consisted of 600 tons of  $C_2Cl_4$  (better known as dry-cleaning fluid!) buried deep underground in a mine to shield it from all other radiation: only neutrinos can penetrate to such a depth, and any reactions observed would therefore be due to neutrinos. Every 35 days, the fluid was removed

and processed to search for electrons produced in the reaction



Only a few events were observed in each cycle. Neutrinos with sufficient energy to cause this reaction are produced by the elements beryllium and boron in the solar core. The nuclear reactions occurring in the Sun were believed to be well understood, since the standard solar model, developed by John Bahcall and others, gave a very good description of other solar properties, such as seismic data. However, the flux of electron-neutrinos observed in the Homestake experiment was smaller than the flux predicted by the standard solar model by a factor of one-third. This discrepancy became known as the ‘solar neutrino problem’ and for many years its solution was unknown. If the standard solar model was correct, then the electron-neutrinos must have somehow ‘disappeared’ in the eight minutes or so between being produced in the Sun and being detected on Earth!

### 43.3 Neutrino Oscillations

The electron-neutrinos cannot simply disappear, so where do they go? It is possible that they interact somehow between the Sun and Earth, but such an interaction would have to be completely new, because the only known interactions involving neutrinos are very weak.

A much more convincing explanation is that the electron-neutrinos disappear not by *interaction*, but by *oscillation* into other neutrino flavours. In order for this to happen, neutrinos must have masses and the neutrino flavours must mix. This occurs in much the same way as the mixing of quark flavours. The neutrinos, like the quarks, are produced and detected in weak interactions in definite flavours:  $\nu_e$ ,  $\nu_\mu$  and  $\nu_\tau$ . However, it is not necessarily the case that these flavour states are the same states in which the neutrinos propagate through free space. The two sets of states could be ‘rotated’ or *mixed* relative to one another.

Let us consider the simplest case of not three, but two neutrinos, with flavour states  $\nu_e$  and  $\nu_\mu$ , and propagating states  $\nu_1$  and  $\nu_2$ . These two sets of states are related by a ‘rotation’ through an angle  $\theta$ , such that

$$\begin{aligned} \nu_e &= \nu_1 \cos \theta + \nu_2 \sin \theta, \\ \nu_\mu &= -\nu_1 \sin \theta + \nu_2 \cos \theta. \end{aligned}$$

The neutrinos are produced in the Sun as  $\nu_e$ , but propagate through space–time as a mixture of  $\nu_1$  and  $\nu_2$ . The  $\nu_1$  and  $\nu_2$  can have different masses and therefore propagate in different ways. When a neutrino arrives on Earth, it may (with a certain probability) have changed into a muon-neutrino, which cannot be detected via the reaction (43.1). Thus, the number of electron-neutrinos detected will in general be less than the number predicted ignoring oscillations. If the masses of the propagating states  $\nu_1$  and  $\nu_2$  are  $m_1$  and  $m_2$  respectively, then the probability that an electron-neutrino of energy  $E$  will have oscillated into a muon-neutrino is given by

$$P_{e \rightarrow \mu} = \sin^2 2\theta \sin^2 \left( \frac{\Delta m^2 L}{4E} \right),$$

where  $L$  is the Earth–Sun distance and  $\Delta m^2 = m_1^2 - m_2^2$  is the mass-squared difference of the neutrinos. From this formula, we can see why the neutrinos need to be massive if they are to oscillate: if  $m_1 = m_2 = 0$ , then the mass-squared difference is automatically zero and  $P_{e \rightarrow \mu}$  vanishes. Intuitively, the explanation for this is that if the neutrino masses are the same, there is no difference in the way the states propagate.

In the real case of three neutrinos, the mixing and oscillations are more complicated, but the general principle is the same. The most important difference is that there is no longer just one mixing angle  $\theta$ , but three mixing angles and three CP-violating phases. This is similar to (but not identical to) the mixing between two quarks (parameterised by the single Cabibbo angle  $\theta_c$ ) and between three quarks (parameterised by three angles and just one CP-violating phase).

### 43.4 Neutrino Oscillation Experiments

The two-neutrino oscillation formula above shows that the ability of neutrino experiments to measure the mixing angle  $\theta$  and the mass difference depends on the energy  $E$  of the detected neutrinos and the distance  $L$  over which they propagate, called the *baseline*. Experiments can be divided into roughly three categories, depending on whether the baseline is short, long, or very long. Table 43.1 shows their typical characteristics and sensitivity.

### 43.5 Solar Experiments

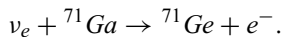
As we saw at the beginning of this chapter, the first solar neutrino experiment was the Homestake

Table 43.1. Typical characteristics of neutrino experiments with different baselines.

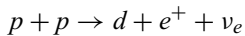
Type	Baseline	Sensitivity $\Delta m^2$	Examples
Short baseline	10 m–100 km	$\gtrsim 0.1 eV^2$	LSND, Bugey
Long baseline and atmospheric	1 km–1000 km	$\gtrsim 10^{-4} eV^2$	CHOOZ, SuperK
Very long baseline and solar	200 km– $10^8$ km	$\gtrsim 10^{-12} eV^2$	KamLAND, SNO

experiment begun in the late 1960s, which observed a flux equal to  $0.34 \pm 0.03$  of the expected flux of electron-neutrinos from beryllium and boron. This deficit was confirmed in the late 1980s by the Japanese Kamiokande experiment which detected the neutrinos via their elastic scattering off electrons in 3000 tons of water. The recoiling electrons move through the water at a speed faster than the speed of light in water, and this causes a cone of radiation, called Cherenkov radiation, to be produced. The cone of radiation gives an indication of the direction of the incoming neutrino. The Kamiokande experiment was only sensitive to neutrinos from boron and (together with the 50 000 ton Super Kamiokande experiment which followed it) observed an electron-neutrino flux of  $0.465 \pm 0.015$  of the predicted flux.

In the early 1990s, two experiments called GALLEX and SAGE were devised, based on the detection of  $\nu_e$  via scattering off gallium to produce germanium,

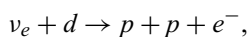


This reaction has a much lower threshold energy, and enabled neutrinos produced in the solar nuclear reaction

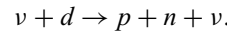


to be detected, as well as those coming from boron and beryllium. Again, the flux was lower than expected, by a factor of  $0.56 \pm 0.03$ .

Thus, by the 1990s, there was clear evidence of a  $\nu_e$  deficit. The clinching evidence came from the Sudbury Neutrino Oscillation experiment (SNO), consisting of 1000 tons of heavy water (in which hydrogen is replaced by deuterium, containing a neutron as well as a proton in its nucleus) on loan from Atomic Energy Canada Ltd, see Figure 43.1. The use of heavy water means that SNO was able to detect not only electron-neutrinos through the weak charged-current reaction



but also all other flavours of neutrino through the neutral-current reaction



So SNO was able to measure not just the electron-neutrino deficit, but also the total neutrino flux. If this agreed with the standard solar model prediction, then one would have clear evidence for the oscillation of  $\nu_e$  into  $\nu_\mu$  or  $\nu_\tau$ .

By April 2002, the observation of the charged- and neutral-current reactions clearly indicated an oscillation of electron-neutrinos. After further data had been gathered, SNO announced its best results in September 2003. The total flux of neutrinos from  ${}^8\text{B}$  decay was measured to be

$$5.21 \pm 0.47 \times 10^6 \text{cm}^{-2} \text{s}^{-1},$$

consistent with the standard solar model prediction of

$$5 \pm 1 \times 10^6 \text{cm}^{-2} \text{s}^{-1}.$$

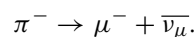
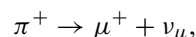
The flux of  $\nu_e$  was measured to be just

$$1.6 \pm 0.1 \times 10^6 \text{cm}^{-2} \text{s}^{-1},$$

giving a ratio of  $0.31 \pm 0.04$ . The measured mixing angle and mass-squared difference are  $\tan^2 \theta = 0.4$  and  $\Delta m_{\text{sol}}^2 = 7 \times 10^{-5} eV^2$ , respectively. The mass-squared difference is minuscule.

### 43.6 Atmospheric Experiments

Solar neutrino experiments detect neutrinos produced in the Sun, but neutrinos are also produced in the Earth's atmosphere by incoming cosmic rays (mostly protons), which produce pions, which decay into muons- and muon-neutrinos:



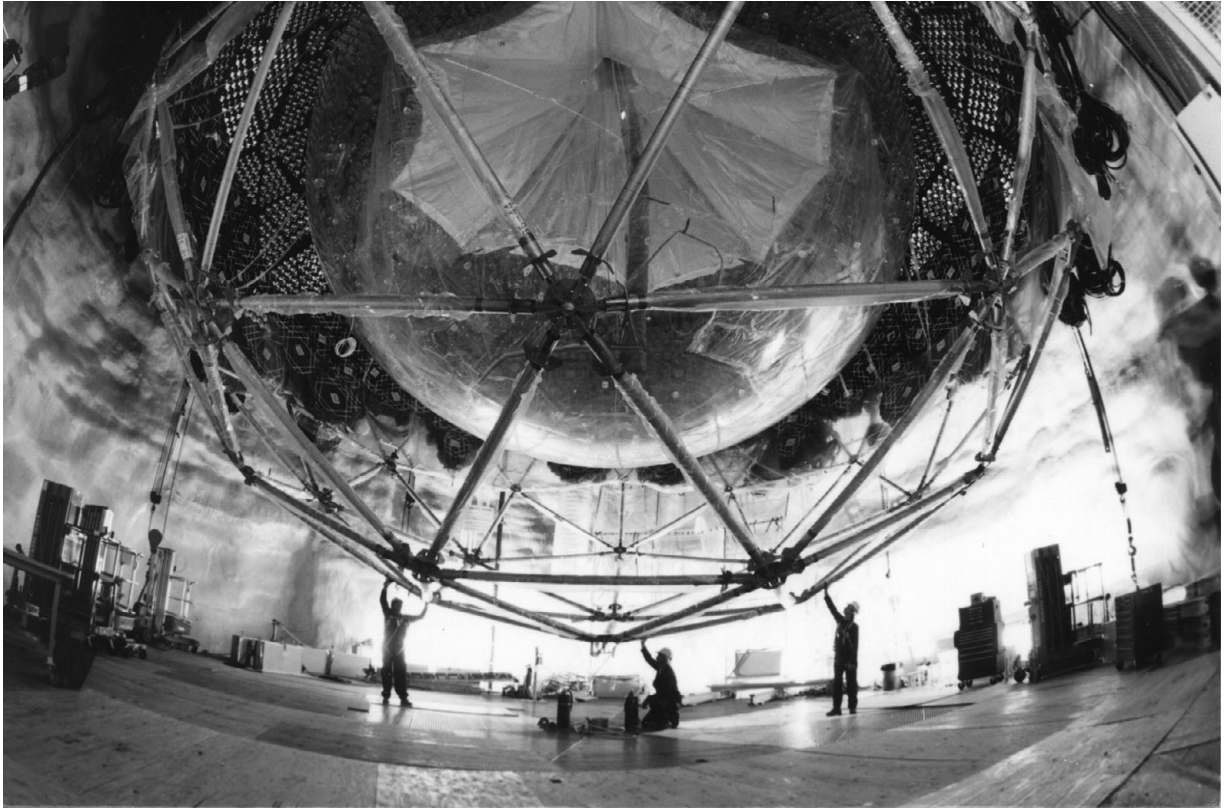


Figure 43.1. The SNO experiment. *Photo courtesy SNO.*

The muons themselves then decay to electrons and neutrinos:

$$\mu^+ \rightarrow e^+ + \nu_e + \bar{\nu}_\mu,$$

$$\mu^- \rightarrow e^- + \bar{\nu}_e + \nu_\mu.$$

This chain of decays suggests that the ratio of atmospheric muon-neutrinos to electron-neutrinos should be about two.

In the late 1980s and 1990s, several experiments, including Kamiokande (and its progeny Super Kamiokande), IMB and Soudan 2, measured the ratio to be less than two, suggesting the disappearance of muon-neutrinos (rather than electron-neutrinos, as in the solar case). The real breakthrough came in 1998, when Super Kamiokande measured an asymmetry in the number of detected  $\nu_\mu$  depending on whether the neutrinos were moving upwards or downwards through the water detector. The asymmetry as a function of the neutrinos' energies is shown in Figure 43.2. The overall asymmetry is

$$\frac{N^{\text{up}} - N^{\text{down}}}{N^{\text{up}} + N^{\text{down}}} = -0.31 \pm 0.04.$$

The explanation for this asymmetry is as follows. Downwards-moving neutrinos produced in the atmosphere travel only 20 km or so before being detected. Upwards-moving neutrinos must travel through the Earth, around 12 000 km, before they are detected and so have a much greater distance in which to oscillate. The oscillations result in the up–down asymmetry. Oscillations also explain why the up–down asymmetry in Figure 43.2 disappears at low energies. As we saw, the oscillation probability depends on energy, and at low energies, both up- and down-moving neutrinos have sufficient time to oscillate fully, so no asymmetry is detected. In contrast, at high energies, the downwards-moving neutrinos hardly oscillate at all, and the deficit becomes small.

The data suggest the oscillation of  $\nu_\mu$  into  $\nu_\tau$  with  $\Delta m_{\text{atm}}^2 \simeq 2.6 \times 10^{-3} eV^2$  and  $\sin^2 \theta_{\text{atm}} \simeq 1$ . This is called *maximal* mixing, because  $\sin^2 \theta_{\text{atm}}$  takes its

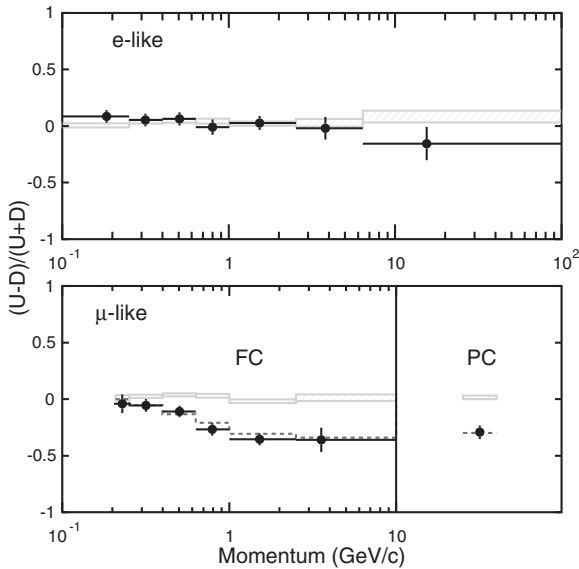


Figure 43.2. The up-down asymmetry measured at SuperK for (top)  $e$ -like neutrinos and (bottom)  $\mu$ -like neutrinos. The shaded rectangles show the prediction without oscillations and the dashed line shows the best fit to the data allowing oscillations. Reproduced from *Proceedings, 8th International Workshop on Neutrino Telescopes*, vol. 1, pp. 183–201. Edited by M. Baldo Ceolin. Padova, Papergraf, 1999. 2 vols.

largest allowed value, namely one. The atmospheric mass-squared difference is much larger than in the solar case, but is still minuscule.

The atmospheric neutrino oscillation data were further supported in 2002 by the K2K experiment, which measured the disappearance of  $\nu_\mu$  produced in an Earth-based accelerator over a distance of 250 km. Only 56 muon-neutrinos out of an expected number of 80 were detected.

The combined solar and atmospheric neutrino data are explained well by the mixing of the three Standard Model neutrino flavours  $\nu_e$ ,  $\nu_\mu$  and  $\nu_\tau$  into propagating states  $\nu_1$ ,  $\nu_2$  and  $\nu_3$  with masses  $m_1$ ,  $m_2$  and  $m_3$ . There are two ways in which the combined experimental data can be explained, called the *normal* and *inverted mass hierarchies*, shown in Figure 43.3. It is not yet known which hierarchy is the correct one, or what the absolute masses of the neutrinos are. Attempts to infer the neutrino masses (via energy-momentum conservation) in  $\beta$ -decay of tritium suggest that they should have mass of less than 2.2 eV.

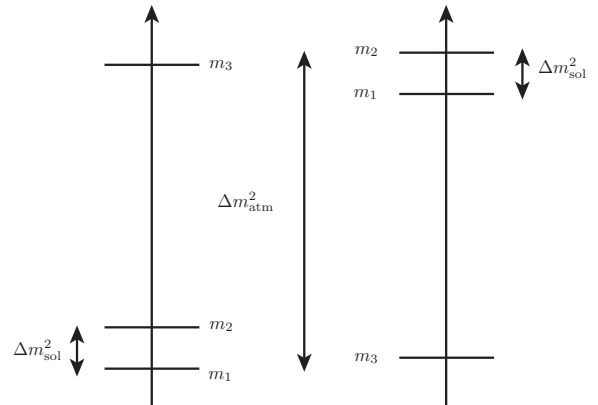


Figure 43.3. Normal (left) and inverted (right) neutrino mass hierarchies, with solar and atmospheric mass differences indicated.

Furthermore, there are cosmological constraints on the absolute neutrino masses, obtained by considering the contribution of neutrinos to the current density of the Universe and their effect on structure formation in the early Universe. The most recent constraint from the Planck experiment (see Chapter 48) suggests that the sum of the absolute neutrino masses should be less than 0.23 eV. So each neutrino mass could be as large as a small fraction of an electron volt.

### 43.7 Short Baseline Experiments

While solar and atmospheric neutrino experiments have a consistent interpretation in terms of a model in which the three neutrinos of the Standard Model mix with each other, short baseline experiments suggest that a more complicated picture may be required. The experiments in question, LSND at Los Alamos and miniBOONE at Fermilab, use neutrino beams produced by ‘dumping’ a beam of protons onto a fixed target, such as graphite or beryllium. Doing so produces large numbers of charged pions, which are focused into a broad beam by magnetic fields, before decaying into neutrinos, antineutrinos, and other particles. Large blocks of steel or aluminium are then used to filter out everything but the weakly interacting neutrinos.

In 1995, LSND claimed to have observed  $\bar{\nu}_\mu \rightarrow \bar{\nu}_e$  oscillations. Although the level of statistical significance was low, this caused excitement among physicists because such oscillations cannot be explained by

three-neutrino mixing. The miniBOONE experiment was commissioned to check the LSND result. While its initial results in 2007 were inconclusive, later results (May 2018) using a much larger dataset claim to reproduce the LSND anomaly, with a combined statistical significance of over  $6\sigma$ . If true, the most likely explanation seems to involve additional *sterile* neutrinos with a mass of around an eV. Sterile neutrinos do not feel the weak force, so this is not inconsistent with the cosmological results from Planck.

A different short baseline experiment uses anti-neutrinos produced by the Daya Bay nuclear reactor near Hong Kong. In 2014, the Daya Bay experiment announced for the first time a measurement of the mixing angle between the 1st and 3rd neutrino families, with  $\theta_{13} \sim 9$  degrees. This relatively large value has raised hopes that it may soon be possible to observe CP violation in the leptonic sector. Indeed, much like in the quark sector, CP violation can lead to a difference in the oscillation rates between an oscillation process involving neutrinos and the corresponding process involving antineutrinos. In a three-neutrino mixing model there is one CP-violating phase that can be accessed in this way, but both mass differences and all three mixing angles must be sizable in order to get an effect. The measured value of  $\theta_{13}$  suggests that a measurement may be possible and two experiments, DUNE at Fermilab and HyperKamiokande in Japan (a larger version of SuperK) are currently being developed in the hope of achieving it.

### 43.8 Theory of Neutrino Masses and Mixings

The claim that neutrino masses and mixings cannot be explained in the Standard Model may seem surprising. After all, the quarks have masses and mix according to the CKM matrix (see Chapter 39) so why shouldn't the same thing occur in the lepton sector? To understand why neutrino masses and mixing really are radical discoveries, we need to go back and analyse the difference between quarks and leptons. Referring back to Table 37.1, showing the quark and lepton multiplets of the Standard Model, we see that the differences are twofold. Firstly, the leptons do not carry colour charges. Secondly, whereas all the quarks (and the *charged* leptons) come in both left- and right-handed versions, there are no right-handed neutrinos in the Standard Model.

In the Standard Model, quarks (and electrons, muons and taus) get their masses from gauge-invariant terms in the Lagrangian coupling a left-handed quark, a right-handed quark and a Higgs boson. Normally, such terms correspond not to mass terms, but to three-particle interactions. However, in the electroweak symmetry breaking process, the Higgs boson sits at the bottom of the wine-bottle-shaped potential (see Chapter 41) where it has a non-zero average value in the vacuum. Inserting this average value for the Higgs boson into the would-be three-particle interaction terms mentioned above leads to mass terms for the quarks as well as three-particle interactions between two quarks and a Higgs boson. For neutrinos, there is no right-handed particle and so such terms simply do not exist. Neutrinos cannot acquire a mass in this way and so in the Standard Model they are strictly massless.

This leads us onto mixing. In Chapter 39, we saw that flavour mixing in the quarks corresponds to the fact that the quarks which participate in the weak interaction are not the same as the quarks which propagate through space-time. The weak interaction states are determined by the flavour, whereas the propagating states are determined by the masses. The two sets of quark states are 'rotated' or mixed relative to each other. In the neutrino sector, there are no masses, so the notion of a rotation or mixing is meaningless.

It is clear then that neutrino masses and mixing are in conflict with the Standard Model, yet they have been observed to be a feature of nature! So the Standard Model must be modified somehow. That is, there must be a theory which goes *beyond* the Standard Model, but which reproduces the results of the Standard Model in the regimes where the Standard Model has been shown experimentally to be correct.

### 43.9 A Minimal Extension of the Standard Model

The neutrino data clearly necessitate an extension of the Standard Model, and so we begin by showing how this can be done. The neutrino data can in fact be explained by a very simple modification to the Standard Model. We saw in the last section that there are no right-handed neutrinos in the Standard Model, the reason being that no right-handed neutrino has ever been seen. But this does not necessarily mean that there are no right-handed neutrinos, only that we

have never seen one! Indeed, let us ask what would happen if we added a ‘right-handed neutrino multiplet’ to the five Standard Model multiplets in each family, making six multiplets in all. In order to ensure that the right-handed neutrino is invisible, we declare that it carries no charge with respect to *any* of the Standard Model gauge groups.

Adding a right-handed neutrino in this way seems pointless, since we cannot see it directly, but such a particle does have indirect, observable effects. Firstly, we can now write down a term in the Lagrangian which generates a mass for the neutrinos, provided the Higgs boson acquires a vev. This mass term is the same as the mass terms of all the other quarks and leptons. It is called a *Dirac* mass term.

Thus, the addition of a right-handed neutrino allows the neutrinos to gain a mass in the same way as all other fermions in the Standard Model. However, there is something rather unsatisfactory about this. Excepting the neutrinos, the lightest fermion in the Standard Model is the electron, with a mass of 0.511 MeV. Although we do not yet know the absolute masses of the neutrinos, we do know they are very small. We saw above that the largest mass-squared difference is around  $10^{-3}(\text{eV})^2$ , and the upper bound for the sum of absolute masses is about 0.7 eV. So the masses of the neutrinos really are *tiny*. If they do acquire their mass in the same way as all the other fermions, why should they be so much lighter?

There is a rather beautiful resolution of this problem, which goes by the name of the *see-saw mechanism*. Because the right-handed neutrinos carry

no charges, they can acquire a mass from another allowed (i.e. gauge-invariant) term in the Lagrangian, known as a *Majorana* mass term. Suppose this mass is  $M$ , whereas the Dirac mass term is  $m$ . The immediate question is: which of these is the actual physical mass of the neutrino? Is it  $M$ , or  $m$ , or any combination of them? Now neither  $m$  nor  $M$  is expected to be small in the Standard Model, and so it appears that the problem of the small neutrino masses has only been further compounded. However, it turns out that the physical neutrino mass is not  $m$  or  $M$ , but is instead given by

$$\frac{m^2}{M}.$$

This combination of masses need not be large, even if  $m$  and  $M$  are themselves large. Let us ask how large the Majorana mass needs to be if a natural size for the Dirac mass, say  $m \sim 1$  GeV, leads to the observed neutrino masses of a fraction of an eV. The answer is that the Majorana mass must be very large indeed, say around  $10^{15}$  GeV! This, of course, is way above any of the scales in the Standard Model, and looks very unnatural. However, we will see in the next chapter that this is the relevant mass scale in a very compelling class of ‘beyond-the-Standard Model’ theories, called Grand Unified Theories, for which there is other indirect evidence besides. So if the see-saw mechanism is correct, we already have a hint of the existence of theories existing not just ‘beyond the Standard Model’, but *way beyond* the Standard Model, at extremely large energy scales.



## *Grand Unification*

### 44.1 Introduction

In the previous chapter, we saw how neutrino experiments necessitated a modification of the Standard Model, and how this could be achieved by adding right-handed neutrinos to the model.

A critic might argue that this is an ad hoc extension to a model which is in itself ad hoc. Indeed, while the core structure of the Standard Model (with or without massive neutrinos), based on the paradigms of quantum mechanics, Lorentz invariance and gauge invariance is almost forced upon us by consistency requirements, the details of the theory are something of a hotch-potch of unexplained structure, albeit a consistent hotch-potch. This raises a number of questions.

For example, we might ask why there are *three* gauge groups, and why they are different? Why are the three gauge couplings different in strength at the energy scales we observe, but not so very different? Why is there such a hierarchy of fermion masses, with neutrinos at a fraction of an electron volt, the electron at 0.5 MeV and the top quark at 178 GeV? Why do the quarks mix, but not very much? Why do the neutrinos mix a lot?

At the heart of all this are the free parameters in the Standard Model, 19 in all. They are allowed to take any values, and this is arbitrariness that allows the hierarchies in the Standard Model to occur. It seems very unsatisfactory, though, to have so many free parameters in a fundamental theory of physics, and one might hope that they are somehow related to

each other in a more fundamental theory. This hope is not a blind one either. The hierarchies indicate that the Standard Model *does* have a great deal of structure, and the presence of structure in itself invites a deeper explanation. It is hoped that this may come from some over-arching theory.

In order to find such a theory, one would ideally like to be confronted with experimental data which contradict the Standard Model and point towards a new theory. Apart from the results of the neutrino oscillation experiments, such data are simply not available to us at the present time. Given this state of affairs, a pragmatic way to go looking for such a ‘higher theory’ is to ask how some of the unexplained structure of the Standard Model *could* be explained. Let us start then, with our first question, namely why are there three gauge groups, and why are the couplings so different?

We already know, from the electroweak theory, that the *apparent* gauge symmetry of a theory is not necessarily the *actual* gauge symmetry of the theory. This is because gauge symmetry can be spontaneously broken in the vacuum. In the electroweak theory, the pattern of symmetry breaking is

$$SU(2)_L \times U(1)_Y \rightarrow U(1)_{\text{em}}.$$

Could it be that the full  $SU(3)_c \times SU(2)_L \times U(1)_Y$  gauge symmetry of the Standard Model is itself just the remnant of some larger, broken, gauge symmetry? One appealing proposal is that the larger gauge symmetry

is actually just a single group, with a single coupling constant, not a combination of three groups, like the Standard Model. Such theories, going by the name of Grand Unified Theories (or GUTs), were first proposed by Howard Georgi and Sheldon Glashow in 1974.

At first sight, such a proposal seems to be impossible: if the three Standard Model groups are all embedded in a larger group, then they should all have the same coupling constant. In GUTs, this problem is evaded in the following way. The larger gauge symmetry is necessarily broken at an energy scale which is larger than that which we currently observe in experiments, since we do not see the full symmetry. Let us call this energy, or mass, scale  $M_{\text{GUT}}$ . At this scale, Grand Unified Theories postulate that the  $SU(3)$ ,  $SU(2)$  and  $U(1)$  Standard Model couplings  $g_1$ ,  $g_2$  and  $g_3$  are indeed equal. But as we know, the coupling ‘constants’ run with energy, because of the screening and anti-screening effects of virtual particles. The different couplings, which as we know couple to matter in different ways, have different running behaviour. This leads to very different couplings at the low energies we observe, despite the fact that they coincide at the scale  $M_{\text{GUT}}$ . Indeed, one finds the general pattern  $g_3 \gg g_2 > g_1$ , which is what we observe! Since there are three Standard Model gauge couplings, and only two free parameters in a GUT, namely the scale  $M_{\text{GUT}}$  and the value of the unified coupling at that scale, we are able to actually *test* the GUT prediction. If we plot the measured running couplings as a function of energy and extrapolate to high energies, they should all meet at a point. Figure 44.1 shows that the running couplings do very nearly meet, at an extraordinarily high energy scale of  $M_{\text{GUT}} \sim 10^{15}$  GeV. There is a slight mismatch, which tells us that the simple theory that we have outlined here cannot be quite correct. Of course, there are a number of things which could change the running of the couplings at high energies, most notably the appearance of new massive particles, as yet undetected in accelerators. It turns out that most modifications involving adding new particles make the prediction even worse. We will see in the next chapter that a very special modification of GUTs causes the mismatch to disappear.

There is more to GUTs than just unification of the gauge couplings. The Standard Model matter multiplets must also be amalgamated into multiplets transforming under the unified gauge group. This too results in a simplification of the Standard Model.

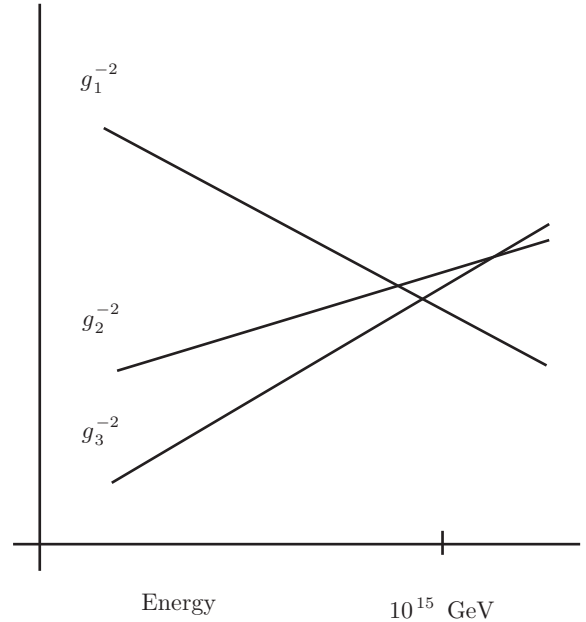


Figure 44.1. Running of the (inverse-square) Standard Model gauge couplings with energy. The couplings almost unify, but not quite.

For example, in the simplest GUT, the Standard Model gauge group is embedded in the larger group  $SU(5)$ . This is just large enough to fit all of the Standard Model gauge groups into it. The five multiplets of each fermion family in the Standard Model (15 states in all) can be amalgamated into just two multiplets of the group  $SU(5)$ . One of the multiplets has five states and the other has ten. There is much more to this than a simple counting of states – the multiplets have to be grouped in such a way that  $SU(3)$ ,  $SU(2)$  and  $U(1)$  Standard Model charges are assigned correctly. It is remarkable, in particular, that the seemingly arbitrary hypercharge assignments of Table 37.1 (1/6 for the left-handed quark multiplet,  $-1/2$  for the left-handed lepton multiplet and so on) are just what is required for  $SU(5)$  unification.

The next simplest GUT is based on a slightly larger gauge group called  $SO(10)$ . In the  $SO(10)$  GUT, all of the Standard Model fermions can be put into a *single* multiplet of  $SO(10)$ , containing 16 states. The 16th state, which is missing in the Standard Model, has just the right quantum numbers to be a right-handed neutrino, which is required for the observed neutrino masses and mixing (Chapter 43). Moreover, the

right-handed neutrino Majorana mass is then expected to be around  $M_{\text{GUT}} \sim 10^{15}$  GeV, which is roughly what is required for the see-saw mechanism!

Finally, we should mention one other interesting property of GUTs. Because quarks (which make up baryons) and leptons get grouped into the same multiplets in the unified theory, there are allowed processes which convert them into one another. So in GUTs, baryon and lepton number can be violated. In particular, it is a generic prediction of GUTs that the lightest baryon, the proton, can decay. The predicted rate of proton decay is very small, going as the inverse-

fourth power of  $M_{\text{GUT}}$ . This is just as well, because proton decay has never been observed, despite intensive searches. The current lower bound on the proton lifetime is around  $10^{33}$  years, which is enough to exclude the simplest  $SU(5)$  GUT, but not others. The possibility of lepton- and baryon-number violations in GUTs could even be a positive virtue of the theories, even though we do not observe such processes. We do of course observe a huge predominance of matter over antimatter in the Universe, and it may be that this has arisen because of lepton- and baryon-number violation in the early Universe (see Chapter 50).

## *Supersymmetry*

### 45.1 Introduction

In the previous chapter, we saw how enlarging the gauge symmetry of the Standard Model led to appealing theories of physics beyond the Standard Model which unified the interactions of the Model. The Standard Model has other symmetries of course, most importantly the symmetry under the Lorentz transformations of special relativity (discussed in Chapter 2), and it is pertinent to ask if we can somehow enlarge this symmetry of the Standard Model.

In 1971, Yuri Gol'fand and Evgeny Likhtman, Pierre Ramond, and Andre Neveu and John Schwarz independently discovered models with an extended symmetry of this type, now called *supersymmetry* (or SUSY). As we saw in Chapter 4, the marriage of special relativity and quantum mechanics in QFT gave rise to particles with spin, that is, intrinsic angular momentum. So it should not come as a surprise to learn that supersymmetry results in relations between particles of different spins. So in a supersymmetric theory, for example, a spin-1/2 particle could be related to a spin-0 or a spin-1 particle. Following their discovery, supersymmetric QFTs were studied intensively, and many more remarkable properties emerged, which we now discuss.

### 45.2 Miracles of SUSY

We saw above that SUSY relates particles of different spin. More precisely, when SUSY currents

act on particles of integer spin, they transform them into particles of half-integer spin. Likewise, they transform half-integer-spin particles into integer-spin particles. But integer-spin particles are bosons, which are symmetric under interchange, and half-integer-spin particles are fermions, which are antisymmetric under interchange. Thus it follows that in any theory which is SUSY, every particle is accompanied by a particle of opposite spin and symmetry under interchange (also called *statistics*), called a *superpartner* or *superparticle*. A particle and its superparticle share the same mass, charge and all other quantum numbers. They differ only in their spin and statistics.

The most important consequence of the pairing-up of particles and superparticles appears when we consider the virtual particles which appear in the loops of Feynman diagrams for quantum-mechanical processes. It is these loops, remember, which lead to the infinities of QFT. Now a particle and its superparticle have similar properties, and indeed they contribute to many loop diagrams in exactly the same way (because they have the same charge and mass and so on), except that there is a relative minus sign because of the different statistics. This means that all such loop diagrams *cancel* in a pair-wise fashion. The result is that a great many of the infinities which are present in a non-SUSY QFT simply disappear, without the need for renormalisation. There are still infinities, but far fewer.

### 45.3 SUSY and the Real World

These and other miracles of SUSY are very interesting, but they give no hint of how SUSY may be relevant to nature. Indeed, a casual inspection of the masses of observed particles shows that the world cannot be supersymmetric, for SUSY requires that each particle be accompanied by a superpartner of opposite spin and the same mass. No such pairs of equal mass but opposite spin particles are observed.

This is not necessarily the death knell for SUSY, however. We have seen in the electroweak theory of Glashow, Salam and Weinberg that the full gauge symmetry of the Standard Model is broken in the world we see. The same notion of broken symmetry is used in GUTs. Could it be that SUSY too is somehow broken at the low energies at which we are currently able to do experiments?

### 45.4 The Hierarchy Problem

We have not yet said why SUSY, broken or otherwise, is a desirable feature of physics beyond the Standard Model, beyond the fact that it has an undeniable aesthetic elegance. In fact there is a very good reason why SUSY is desirable, related to the mass of the Higgs boson. As we saw in Chapter 37, all the other particles in the Standard Model, the gauge bosons, quarks and leptons, acquire their masses via spontaneous symmetry breaking of the electroweak gauge symmetry. The Higgs boson, by contrast, does not. Its mass is a free parameter in the theory, and is simply put in by hand. Because of quantum effects, the Higgs boson will receive corrections to its mass from loop diagrams containing virtual particles in the loops. These corrections to the mass are very large. Indeed, if we regard QFT as having some ultraviolet cutoff, say the GUT scale of  $10^{15}$  GeV, then the quantum corrections to the Higgs mass will be around  $10^{15}$  GeV, and so we expect the Higgs mass itself to be

around  $10^{15}$  GeV. The only way this could be avoided would be if there were very delicate (and unnatural) cancellations between the quantum corrections.

The problem is that the mass of the Higgs boson is not  $10^{15}$  GeV, but 125 GeV, i.e. a factor  $10^{13}$  times smaller than the Higgs mass we naively expect. If the Standard Model is valid all the way up to the higher mass scale, this enormous hierarchy of scales can only be explained by a fantastically unlikely fine-tuning of the free parameters of the Standard Model. This is the so-called *hierarchy problem*.

### 45.5 SUSY as a Resolution of the Hierarchy Problem

SUSY can provide precisely the delicate cancellation needed for a small Higgs mass, but in a very natural way. In such a theory, for every loop of virtual particles providing a correction to the Higgs mass, there is a loop containing virtual superparticles which exactly cancels it. There are then no quantum corrections to the Higgs mass. Figure 45.1 shows Feynman diagrams with loops containing top quarks and their superpartners, called stops, which give equal and opposite contributions to the Higgs boson mass.

Actually, this argument is only correct if SUSY is unbroken. If SUSY is broken at low energies as it must be, then a mass difference can arise between a particle and its superpartner. The loop corrections no longer exactly cancel, and lead to corrections to the mass of the Higgs boson. These corrections lead to a Higgs boson mass of the right size, 125 GeV say, only if the masses of all the superpartners are less than about 1 TeV.

Thus SUSY, if it solves the hierarchy problem, provides a staggering prediction: superparticles should be within the energy reach of the LHC! This simple prediction has motivated intense, ongoing searches for superparticles at the LHC. Before discussing these

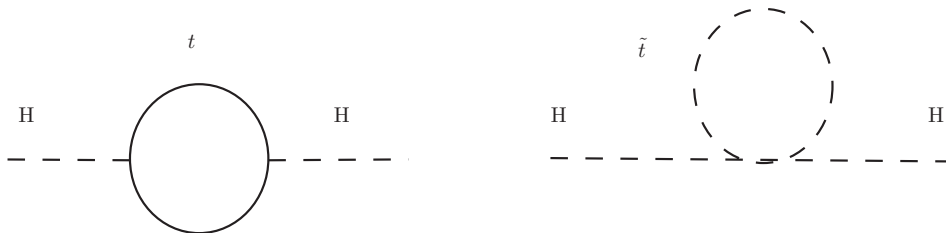


Figure 45.1. Feynman diagrams showing contributions to the Higgs boson mass coming from top quarks on the left and their spin-zero superpartners, the stops, on the right. In a supersymmetric theory, the contributions cancel.

searches, let us describe how to supersymmetrise the Standard Model and some other predictions that result.

#### 45.6 Supersymmetrising the Standard Model

The first step in constructing a supersymmetric version of the Standard Model is to add a superpartner for each Standard Model particle. Conventionally, the scalar superpartners of the Standard Model fermions are prefixed by an s- (so an electron is partnered by a *selectron*, a quark by a *squark* and so on) and the fermionic superpartners of Standard Model bosons are suffixed by -ino (so a gluon is accompanied by a *gluino*, and a Higgs by a *higgsino*).

#### 45.7 Supersymmetry Breaking and the MSSM

Assuming supersymmetry really is a part of nature, a key question is: what causes the breaking of SUSY at low energies? Although a number of viable mechanisms for this have been put forward, it is not at all clear which, if any, is the correct one. Because of this ignorance, it is impossible to make concrete predictions for the masses of superparticles, their couplings and so on. The best one can do is to start by writing down the SUSY theory, and then add by hand all terms (with arbitrary parameters) consistent with the low-energy breaking of SUSY. The resulting theory is called the *minimal supersymmetric Standard Model* or MSSM for short, and has around 120 free parameters. Even before the advent of the LHC, the parameters were somewhat constrained by LEP and other colliders. Indeed, even though superpartners in the expected mass range could not be produced directly by these low-energy machines, they could still make contributions through their virtual effects in loop diagrams. But the LHC was needed to explore the heart of the parameter space.

#### 45.8 Another Prediction of SUSY

When combined with the idea of Grand Unification, SUSY leads to another startling prediction. We saw in the previous chapter on GUTs how the running of the three Standard Model gauge couplings with energy indicated that they meet at an enormously high energy scale of  $10^{15}$  GeV and this was interpreted as indirect evidence for unification. In fact, precision measurements of the electroweak coupling constants at LEP (see Chapter 38) show that the three couplings do not quite meet (Figure 44.1). In the SUSY version

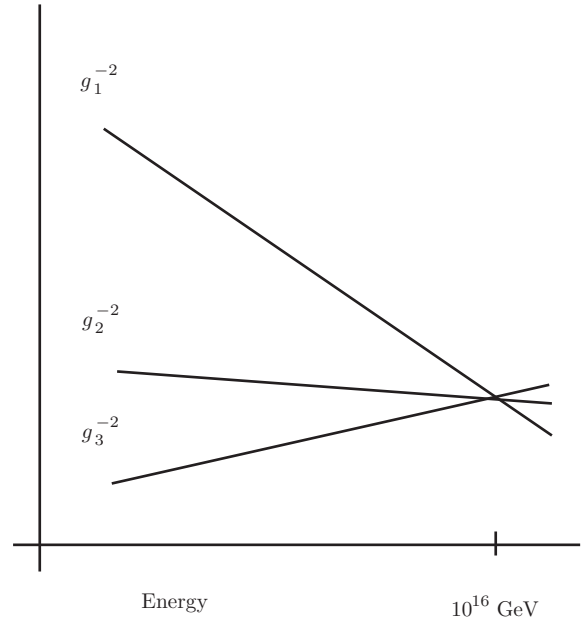


Figure 45.2. Running of the (inverse-square) gauge couplings with energy in the MSSM. The couplings unify at a higher scale.

of a GUT, there are more particles in the theory (the superpartners) which contribute, via loop corrections, to the running of the couplings. This adjustment of the running leads to two desirable changes, see Figure 45.2. The first is that the three couplings now meet at the same point, as required by grand unification. The second change is that the unification scale is shifted to an energy scale which is an order of magnitude larger, at around  $10^{16}$  GeV. This consequently suppresses the probability of proton decay, and moves the predicted proton lifetime beyond current experimental bounds. Here then is experimental evidence (albeit indirect) for supersymmetric grand unification!

#### 45.9 Supersymmetry and Dark Matter

In a supersymmetric theory, it is necessary (in order to prevent much larger contributions to proton decay) to impose an additional discrete symmetry (like parity or baryon number), called *R-parity*, defined such that each Standard Model particle has *R-parity* equal to 1 and each superparticle has *R-parity* equal to  $-1$ . The fact that *R-parity* must be conserved in interaction processes has two immediate consequences. The first is that superpartners must be produced in pairs in

collisions between Standard Model particles. This makes them even harder to produce at the LHC, since they are already believed to be very massive, with masses of hundreds or even thousands of GeV. Secondly,  $R$ -parity predicts that the lightest superparticle will be stable, since it could only decay into an even lighter superparticle, of which there would be none. This has great significance for cosmology. In the high temperatures of the early Universe, such particles would presumably have been produced in great abundance. Since they cannot decay, they must still be present today. If the lightest superparticle happens to be a Higgsino, a Bino, or a Wino (the latter being the partners of the electroweak gauge bosons) or similar, it will interact only weakly with other matter and will be essentially invisible. Could it be the dark matter?

This idea, though highly speculative, has quantitative support: given the standard Big Bang cosmology described in Chapter 48, the superpartners would have been in thermal equilibrium with the Standard Model particles in the high temperatures of the early Universe. As the Universe expanded and cooled, the weakly interacting superpartners would have eventually decoupled (or ‘frozen out’, to use the jargon), leading to a *thermal relic density* which can be calculated using the Boltzmann equations of statistical mechanics. The result of the calculation is that the thermal relic density obtained coincides with the observed dark matter density if the superpartner mass lies roughly at the weak scale, which is precisely what one expects on the basis of the hierarchy problem. The precise value of the superpartner mass depends on the precise details of the model (for example, if dark matter is a pure bino, the mass should be roughly 100 GeV, while for a pure Wino it should be roughly 3 TeV), but many physicists nevertheless considered it a miracle that a calculation based purely on astrophysics and cosmological input data could result in an output prediction so close to the weak scale of particle physics. This *WIMP miracle* only further fuelled the excitement that superparticles would be discovered at the LHC.

#### 45.10 Supersymmetry and the LHC

Though superpartners that solve the hierarchy problem are within reach of the LHC, the fact that even the minimal supersymmetric extension of the Standard Model has 120 parameters means that searching for

them is not straightforward. Since the LHC collides hadrons, the superpartners carrying colour charges, such as the gluino and the squarks, are likely to be the easiest to produce. But their subsequent decay modes, and hence their experimental signatures, depend to a great extent on the exact spectrum of superpartners. To counter this, a huge variety of different searches are being carried out at the LHC, each targeted at different corners of the possible parameter space.

Despite this variety, there are some common aspects in the search strategies. As we have already seen,  $R$ -parity implies that superpartners must be pair produced, so search strategies are often based on looking for pairs of objects in the final state. Moreover,  $R$ -parity implies that all decays of superpartners terminate in the lightest superpartner, which must be colourless and electrically neutral, hence invisible in the LHC detectors. Thus a key signature of superpartners is the presence of apparent *missing energy* in the detector.

Unfortunately, there are plenty of processes involving Standard Model particles which have similar or even identical signatures. For example, neutrinos also manifest themselves as missing energy, while even charged particles can be missed if they strike a ‘dead’ area of the detector, or escape in the forward region of the beam. Thus, an important part of looking for signals of supersymmetry at the LHC is a careful estimate of the various background processes that are present.

Sadly, the hundreds of searches that have been carried out for supersymmetry at the LHC so far share one other common aspect: they have shown no evidence whatsoever for supersymmetry! This was perhaps not so surprising during the first, lower-energy LHC run, where the sensitivity to superpartners only reached up to a TeV or so, even in the most favourable scenarios. But the negative results from the second, higher-energy run that are now flooding in have become a serious thorn in the side of a theory that was once regarded by some physicists as ‘too good not to be true’.

What, then, are the future prospects for supersymmetry? One possibility is that the superpartners are simply heavier than the LHC can reach. But, as we discussed above, the heavier they are, the less they provide us with a solution of the Higgs hierarchy problem. So, one problem with this approach is that, if the superpartners have masses around 10 TeV (putting

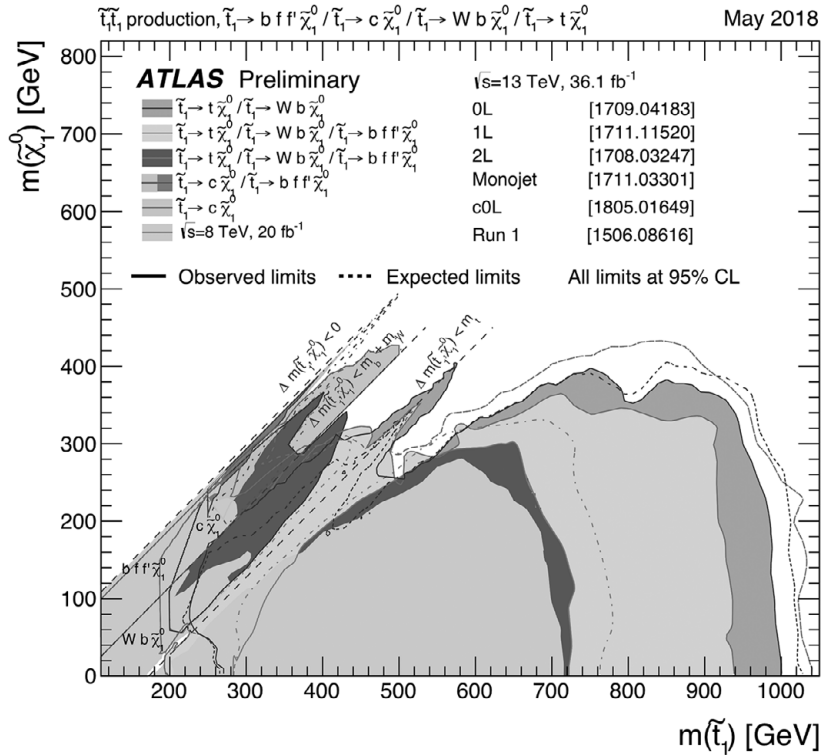


Figure 45.3. Summary of searches in the ATLAS experiment for stop squarks decaying to neutralinos in a variety of SUSY models at the LHC, showing that stop masses up to around 1 TeV are excluded, unless the stop is roughly degenerate in mass with the neutralino. Copyright: CERN.

them out of the reach of the LHC) then we still need to fine-tune the parameters to around one part in a thousand or so in order to get the observed Higgs mass, leaving us with a residual *little hierarchy problem*. The other problem with this approach is that if we are willing to accept a tuning of one part in a thousand, why not one part in a million, or a billion? If the latter are acceptable, then we can forget about trying to build a collider to search for the much heavier superpartners.

Another possibility is that the superpartners (or at least those, such as the partners of the top quark, which most contribute to the fine-tuning of parameters) are light enough to be within reach of the LHC, but that the spectrum of the theory is such that they remain hidden from our suite of searches. Given that

the theory has so many free parameters, this possibility is hard to definitively rule out. But great efforts have been made in the last decade to identify such difficult regions of parameter space and devise new searches to probe them. For example, dedicated searches for the top squark, see Figure 45.3, now show that it cannot have a mass below a TeV or so, unless its mass is roughly degenerate with that of the lightest superpartner.

A third possibility is that supersymmetry is a symmetry of nature, but only at much higher energy scales, such that it does not play a rôle in the solution of the hierarchy problem. If so, we must look elsewhere for an explanation of the hierarchy. In the [next](#) chapter, we describe a completely different idea: the *composite Higgs*.



## *Composite Higgs Models*

### 46.1 Introduction

In Chapter 45, we saw how supersymmetry provides an elegant explanation of the Higgs hierarchy problem, but that a multitude of searches at the LHC have failed to find any evidence for it. In this chapter, we turn to a completely different explanation, which starts from the idea that the Higgs boson is not an elementary particle, but rather a composite state built out of more elementary particles bound together by some new force.

One can motivate this idea in a very basic way by observing that every one of the many scalar bosons that were discovered before the Higgs boson turned out to be composite. Thus, for example, hydrogen turned out to be a bound state of two fermions, an electron and a proton, bound together by the electromagnetic force. Similarly, an alpha particle turned out to be two protons and two neutrons bound together by the strong nuclear force and the pions turned out to be quark–antiquark pairs bound together by QCD. This is in sharp distinction to, say, the spin-half electron, which despite being probed over a vast range of energy scales since its discovery over a century ago, shows no evidence of being a bound state: it is elementary, rather than composite.

This line of argument alone suggests that our ‘null hypothesis’ for the recently-discovered Higgs boson should be that, like all other scalars, it is not elementary, but composite. But there is a much more compelling argument for Higgs compositeness, which

is that it could provide us with an explanation of the hierarchy between the weak scale and higher energy scales that are believed to exist in nature, such as the Planck scale of quantum gravity.

The easiest way to convince oneself of this is to consider, somewhat paradoxically, what the world would look like if there were no Higgs boson at all.

### 46.2 A World without the Higgs

What would physics look like if we took the Standard Model (with just one family of quarks and leptons, for simplicity), but removed the Higgs? Clearly, the weak interactions would look very different at the 100 GeV scale: with no Higgs, there would be no electroweak symmetry breaking and so the W and Z gauge bosons would not acquire masses of around 100 GeV. But the strong nuclear force, which is immune to the presence or absence of the Higgs, would look much the same. As we saw in Chapter 44, the strong coupling constant would run slowly (logarithmically in fact), all the way down from very high energy scales (such as the GUT scale of  $10^{15-16}$  GeV) down to the GeV scale, where it would become strong enough to cause confinement of quarks into mesons and baryons. The one difference, of course, is that the quarks themselves would be massless, since their masses in the Standard Model come (via spontaneous symmetry breaking) from the coupling to the Higgs. So the three pions made up of up and down quarks would, a priori, be massless.

But this conclusion ignores the electroweak gauge interactions. The up and down quarks transform as doublets and singlets under the  $SU(2)$  gauge symmetry and carry hypercharges as in Table 37.1 and their confinement into hadrons would cause the  $SU(2) \times U(1)$  electroweak symmetry to be broken! The pattern of breaking is, in fact, almost an exact copy of the breaking achieved by the Higgs boson in the Standard Model. That is, the  $SU(2) \times U(1)$  symmetry gets broken down to the electromagnetic symmetry, and the  $W^\pm$  and  $Z$  bosons acquire masses in the same way. The three Goldstone bosons that get eaten to give masses to the  $W^\pm$  and  $Z$  are no longer three of the four components of the Higgs field, but rather are the three massless pions.

Thus, even without the Higgs, we obtain a pattern of electroweak symmetry breaking that looks very similar to that of the Standard Model. But there are three notable differences. The first is that the gauge boson masses are not around the electroweak scale of 100 GeV, but rather around the confinement scale of 1 GeV. So this theory cannot possibly describe reality. The second is that there is no analogue of the Higgs boson: the Higgs boson in the Standard Model is the leftover component arising from the fact that only three of the four components of the Higgs field get eaten, while here all three pions get eaten. The third difference, which is the most interesting one for us, is that this theory has no hierarchy problem. There *is* a hierarchy in the theory, namely the one between the confinement scale of 1 GeV and any other scale in the theory, such as the GUT scale at  $10^{15-16}$ . But there is no hierarchy problem, because this large ratio of scales is explained in a completely natural way by the slow logarithmic running of the strong coupling constant: a small, but perfectly reasonable value (say one-tenth) for the coupling at the GUT scale results in electroweak symmetry breaking at a scale a factor of  $10^{15-16}$  lower, when the running coupling constant finally reaches values of order one.

### 46.3 Technicolour

The model just described cannot describe nature, but one can easily make a similar model which looks much more realistic, yet still explains the hierarchy. The idea is to add to the model a copy of the strong nuclear force, called the *technicolour* force, with identical properties, except that its coupling constant is adjusted so that it becomes confining not at 1 GeV,

but at 100 GeV. At that scale, the *techniquarks* (which are just like the usual quarks, except that they feel the technicolour force rather than the usual colour force) become bound into three *technipions*, which act as the Goldstone bosons for the electroweak symmetry breaking, giving the  $W^\pm$  and  $Z$  bosons their usual masses.

Technicolour was a wonderful idea, but it cannot be correct, not least because it does not feature a Higgs boson, in contradiction with experiment. In fact, technicolour was known to be wrong even before the discovery of the Higgs boson, because it leads to contradictions with the electroweak precision tests described in Chapter 38. But there is a relatively easy way to extend technicolour to get a more acceptable theory, which goes by the name of the *composite Higgs*.

### 46.4 Composite Higgs

Suppose we do not restrict the technicolour force to be an exact copy of the strong nuclear force, but rather allow any gauge theory coupled to technifermions, subject only to the restriction that it becomes confining at an energy scale of 100 GeV. The pattern of symmetry breaking that results from confinement of the technifermions can then be different from the one obtained in the usual technicolour. There, we have an  $SU(3)$  gauge theory (like in QCD) coupled to two flavours of techniquarks (like the up and down quarks in QCD), resulting in a pattern of symmetry breaking  $SU(2) \times SU(2) \rightarrow SU(2)$ , where the two  $SU(2)$ s on the left just correspond to separate rotations of the left- and right-handed up and down quarks amongst themselves. This pattern of breaking results in three Goldstone bosons and the right pattern of electroweak gauge symmetry breaking  $SU(2) \times U(1) \rightarrow U(1)$ . But with a different technicolour gauge theory, we could arrange for a pattern of symmetry breaking  $SO(6) \rightarrow SO(5)$ , where  $SO(n)$  is the group of  $n \times n$  orthogonal matrices with determinant one. This pattern of breaking results in five Goldstone bosons, rather than three. Similarly, a pattern of breaking  $SO(5) \rightarrow SO(4)$  would result in four Goldstone bosons.

Now we must introduce another subtlety. In QCD, the pions are not true Goldstone bosons, because they are not massless. They receive contributions to their masses from the underlying quark masses and from the electromagnetic interaction (which is

responsible for the few MeV splitting between the masses of the neutral and charged pions). A similar phenomenon happens in composite Higgs models: although the strongly interacting sector on its own has  $SO(5)$  or  $SO(6)$  symmetry, these symmetries are only approximate once we add in the couplings to the rest of the Standard Model. As a result, only three of the four or five Goldstone bosons remain massless (and get eaten to form the  $W^\pm$  and  $Z$  bosons), while the other one or two degrees of freedom become massive scalar bosons.

In the case of the  $SO(5)/SO(4)$  model, we end up with a model which closely resembles the Standard Model at low energies: we have a pattern of electroweak symmetry breaking which is exactly the same as that in the Standard Model, together with an extra scalar boson that behaves much like the Standard Model Higgs. The  $SO(6)/SO(5)$  model is similar, except that there is an extra scalar, gauge singlet boson. But, in stark contrast to the Standard Model, neither model has a hierarchy problem.

The advantage of composite Higgs models over the original theory of technicolour is not only that they feature a Higgs boson, as data requires, but also that the contributions of the new strongly interacting sector to precision electroweak tests can be brought under control. Indeed, they can be made arbitrarily small by raising the confinement scale of the strongly interacting sector relative to the weak scale of 100 GeV. The price for doing so is that raising the confinement scale in this way reintroduces a fine-tuning into the parameters of the theory. The need to suppress contributions to electroweak tests already requires reintroducing a little hierarchy corresponding to tuning at the level of one part in 20 or so. So composite Higgs models are no panacea!

#### 46.5 Composite Higgs at Colliders

If composite Higgs models closely resemble the Standard Model at low energies, how are we to tell

them apart in experiments? A glib answer is to look at higher energies, corresponding to the confinement scale of the new interaction, where one should presumably see a rich spectrum of broad resonances, just as in QCD. But there is a problem, in that the fine-tuning required to suppress contributions to electroweak precision tests is already likely to put the resonances just beyond the reach of the LHC.

Nevertheless, one may hope that some of the resonances of the strong sector are anomalously light and within reach of the LHC. In particular, the observed value of the Higgs mass suggests that there ought to be partners of the top quark that are within reach. These partners differ from superpartners in supersymmetry in that they have the same spin and same statistics as the top quark, rather than spin differing by one-half and opposite statistics. But like searches for top squarks in supersymmetry, the lower bounds on top partner masses are getting close to a TeV.

Another way to look for evidence of Higgs compositeness at the LHC is to look for deviations in the couplings of the Higgs boson to other particles, which arise due to its composite nature. Unfortunately, the deviations are believed to be at most 10% or so, meaning that there is not much sensitivity at the LHC.

A third way to look for compositeness is to note that, in these scenarios, the Higgs is a composite object bound by a new strong force. That being so, the Higgs will couple strongly to itself as the confinement scale is approached, rather than weakly to itself as in the Standard Model. Thus there are hopes that one could see large deviations in double Higgs production, as discussed at the end of Chapter 41.

In all these cases, we are fighting against the fact that the deviations from the Standard Model only become significant as we get close to the confinement scale. Since precision electroweak tests already tell us that this scale is more likely to be 10 TeV than a TeV, we really need a future, higher-energy collider to fully test the compositeness paradigm.

## *Axions and the Strong CP Problem*

### 47.1 The Strong CP Problem

The identification of QCD, the gauge theory of quarks and gluons, as the correct description of the strong nuclear forces brought great successes. But it also brought a very subtle problem, which goes by the name of the *strong CP problem*.

The general philosophy of quantum field theory says that we should include all terms in the Lagrangian that are compatible with gauge symmetry and renormalisability. In QCD there is one such term involving the gluon fields, called the  $\theta$ -term because its coefficient is an angle, traditionally labelled by  $\theta$ . For a long time, the  $\theta$ -term was ignored, for the simple reason that it appears completely benign. In particular, it has no effects at any order in perturbation theory. But that does not mean that it has no effect at all. At low energies, when the QCD interaction becomes non-perturbative, the  $\theta$ -term can have physical effects. For example, there is a contribution to the vacuum energy density given by

$$E(\theta) = -m_\pi^2 f_\pi^2 \frac{m_u m_d}{(m_u + m_d)^2} \cos^2 \theta, \quad (47.1)$$

where  $m_u$  and  $m_d$  are the up and down quark masses,  $m_\pi$  is the pion mass and  $f_\pi$  is another parameter in pion physics called the pion decay constant.

In fact, the  $\theta$ -term violates CP meaning that its effects, if present, can be rather spectacular. In particular, the  $\theta$ -term gives rise to electric dipole moments of the neutron and other nucleons. No such

dipole moments have been observed, leading to the conclusion that  $\theta$  can be no larger than  $10^{-9}$  (in units of radians). In the context of the Standard Model, it is hard to understand why such a small value occurs, when any angle (in particular an angle of order 1, in units of radians) would do.

### 47.2 The Axion

In 1977, Roberto Peccei and Helen Quinn put forward an elegant solution to the strong CP problem. The idea, in essence, was to promote the parameter  $\theta$  to a scalar field. If this can be achieved, the expression in Equation (47.1) becomes a potential for the scalar field, whose minimum is seen to be at  $\theta = 0$ . Thus, by making  $\theta$  into a field, its vacuum expectation value is zero, meaning that there is no CP violation and no contribution to electric dipole moments.

Steven Weinberg and Frank Wilczek independently realised that the mechanism of Peccei and Quinn implied the presence of a new particle, termed the *axion*, which could be sought in experiments. The simplest explicit model merely involved adding another Higgs doublet to the Standard Model. The properties of the axion in this model were fixed, and soon shown to be ruled out. But it did not take long for other models to be put forward, which were not so easy to rule out. These models contain essentially one free parameter, called the *axion decay constant* and denoted  $f_a$ , which controls both the mass and the couplings of the axion to matter. A substantial research

programme to either discover the axion, or rule out all possible values of  $f_a$  was launched.

### **47.3 The Axion Window**

An experimental lower bound on  $f_a$  of around  $10^9$  GeV comes from the fact that, for lower values, the axion is sufficiently strongly coupled to matter to be able to efficiently transport energy out of cooling stars, leading to observable effects in supernovae

(see Chapter 52) and red giants. An upper bound of  $f_a \simeq 10^{12}$  GeV comes from the fact that axions are an excellent candidate for the dark matter that makes up most of the matter in the Universe (see Chapter 51). But if  $f_a \simeq 10^{12}$  is too large, the Universe would contain too much matter. The remaining allowed window of  $f_a$  is tantalisingly small and many new ideas are being put forward for experiments that can help close it.

**Part XII**  
**Particle Physics and Cosmology**



## *The Big Bang and Inflation*

### 48.1 Introduction

Recently, physics has witnessed the convergence of two of its most fascinating and most fundamental branches: elementary particle physics and cosmology. These two subjects, dealing with the Universe on the smallest and largest possible scales, are now thought to be inextricably intertwined within the framework of the Big Bang theory of the origin of the Universe. This intimate interrelationship between particle physics and cosmology is revealed in the profound implications each discipline holds for the other. According to the Big Bang theory, the Universe began some  $10^{10}$  years ago from a space–time singularity, a single point of infinite energy–density and infinite space–time curvature. The act of creation – the Big Bang – was an enormous explosion from which an extremely hot and dense, rapidly expanding Universe came into being. The early Universe was a thick, hot primordial ‘soup’, filled with a great abundance of elementary particles of every kind, its evolution governed by the fundamental forces between them. Consequently the early Universe was also the ultimate particle accelerator. Its extremely high temperature and high density offer an unrivalled opportunity to probe physics beyond the reach of terrestrial accelerators and test ideas such as grand unified theories, supersymmetry and string theory.

### 48.2 Big Bang Cosmology

Three observations form the basis of Big Bang cosmology. The first of these is that of the expansion

of the Universe, which was first discovered in 1929 by Edwin Hubble. He observed that distant galaxies are moving away from us, and moreover, the farther away a galaxy is, the faster it is receding. This discovery is embodied in the equation known as Hubble’s law:

$$v = Hr,$$

where  $v$  is the galaxy’s recessional velocity,  $r$  is its distance from us, and  $H$  is a constant of proportionality called Hubble’s constant. We now know that  $H$  is not strictly constant but changing very slowly with time. Its present value is now known fairly precisely and is

$$H = 100h \text{ km s}^{-1}/\text{Mpc}$$

where  $h \simeq 0.7$ . (A megaparsec is given by  $1\text{Mpc} = 3 \times 10^6$  light years  $= 3 \times 10^{24}$  cm.) So, a typical galaxy 1 megaparsec away will be moving away from us at a speed of  $71 \text{ km s}^{-1}$ . A galaxy 10 megaparsecs away will be receding at ten times this speed. Actually, because the distances to galaxies are very difficult to determine accurately, the best way to determine the Hubble constant is from observations of the cosmic microwave background, see the next chapter.

The second observation is that of the relative abundance in the cosmos of the light elements, namely hydrogen, helium, deuterium and lithium. In the late 1940s, George Gamow and his collaborators explained these observed abundances in terms of an early Universe which was very hot and dense. The light elements, they proposed, were synthesised when the



Universe was at an absolute temperature of 109 K (on the Kelvin scale  $0\text{ K} = -273^\circ\text{C}$ ). This temperature is equivalent to a thermal energy per particle of about 0.1 MeV. (It is often convenient to express temperatures in electronvolts. Note that  $1\text{ eV} = 1.2 \times 10^4\text{ K}$ .) This process is called *nucleosynthesis* and accounts only for the light elements – heavier elements were formed much later inside stars and distributed throughout the cosmos by supernova explosions.

The third principal observation is that of the cosmic microwave background radiation (the CMB), discovered by chance in 1965 by Arno Penzias and Robert Wilson. This radiation, in which we are constantly bathed from all directions, is a residue of the hot Universe. However, it now has a temperature of only 2.7 K, owing to the cooling effect of the Universe's expansion. The photons that make up this radiation have been propagating freely through space–time ever since electrons and nucleons combined into neutral atoms around  $10^5$  years after the Big Bang, and provide a snapshot of the Universe at that time. We discuss the CMB and its observation in the next chapter.

#### 48.2.1 Friedmann Models

It was Einstein who should have predicted an expanding Universe. However, he was unsettled by the fact that he was unable to find a static cosmological solution to general relativity, and so modified the theory by introducing a new term, a 'cosmological constant', into his equations. A cosmological constant acts as a repulsive antigravity force which is not connected with the presence of matter: it corresponds to an energy in empty space. It is a property of space–time itself, and, Einstein argued, exactly balances the gravitational attraction of all the matter in the Universe. The net result is a static cosmological model.

In 1922, working with Einstein's unmodified equations, the Russian Alexandre Friedmann (and much later Howard Robertson and Arthur Walker independently) considered expanding cosmologies based on two assumptions: the Universe is (1) isotropic (i.e. looks the same in all directions); and (2) homogeneous (i.e. looks the same from every point in the cosmos). These assumptions give rise to the so-called Friedmann models which seem to describe our Universe to a very good approximation. Although on smallish scales the Universe appears very different in different directions, on very large

length scales (much greater than the distances between galaxies) it is indeed remarkably uniform and isotropic. Distant galaxies are distributed more or less uniformly. Moreover, the cosmic microwave background radiation is extremely uniform, indicating that the Universe was even more isotropic in the past. However, anisotropies have been detected at the level of one part in  $10^5$  in the temperature of the radiation. These fluctuations are believed to come from fluctuations in the matter density in the early Universe, which were themselves enhanced by gravity to form the dramatic structures such as galaxies and clusters of galaxies that surround us today. Observing the CMB is very important for studying the Universe, see the next chapter.

An expanding Universe raises the question of whether the expansion will continue forever or eventually end. In the framework of Friedmann models, the answer depends on (1) how fast the Universe is expanding, and (2) how much matter there is. If the mass/energy density of the Universe is greater than a certain critical value, then gravitational attraction will eventually overcome the expansion and the Universe will collapse. If, on the other hand, the density is less than this critical value, expansion will continue ad infinitum. The critical density is

$$\rho_{\text{crit}} = \frac{3H^2}{8\pi G} = 2 \times 10^{-29} h^2 \text{gcm}^{-3}, \quad (48.1)$$

$$= 10^4 h^2 \text{eVcm}^3, \quad (48.2)$$

and is equivalent to about ten hydrogen atoms per cubic metre throughout the Universe. Current observations suggest that the density of the Universe is equal to  $\rho_{\text{crit}}$  to within a per cent or so. From the point of view of the Big Bang theory alone, the fact that the observed value lies so close to the critical value is a mystery, but is easily explained by the theory of inflation, which we discuss below.

General relativity is above all about geometry, see Figure 48.1. If the density is greater than the critical density, then space (not space–time) is positively curved like the surface of a sphere, and the Universe is said to be 'closed', expanding for a certain time before contracting again. But if the density is less than critical, then space is negatively curved like a saddle, and the Universe is said to be 'open', expanding forever. Finally, if the density just so happens to be exactly equal to the critical density, then space is not curved at all but 'flat' (but space–time is still curved).

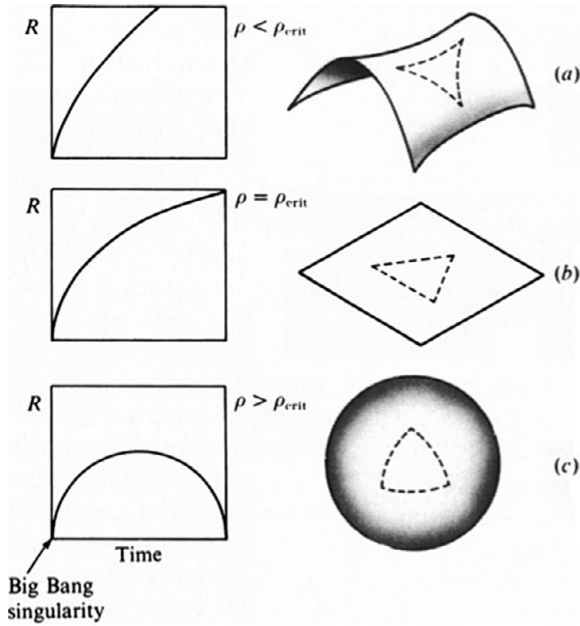


Figure 48.1. Open (a), flat (b) and closed (c) Friedmann cosmologies, showing how the size of the Universe changes with time. The spatial curvature is such that the three angles of a triangle add up to less than  $180^\circ$  in an open Universe and to greater than  $180^\circ$  in a closed Universe.

These Friedmann models form the basis of standard Big Bang cosmology. Which one describes our Universe depends on the actual rate of expansion (i.e.  $H$ ) and the density (i.e.  $\rho$ ). However, despite their vastly different predictions for the eventual fate of the cosmos, these models nevertheless paint very similar pictures of the Universe at early times.

#### 48.2.2 Chronology of Big Bang Cosmology

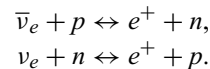
According to Hubble's law, the Universe is expanding in such a way that any two points are separating at a velocity proportional to the distance between them. This expansion is summarised in the behaviour of the cosmological scale factor,  $R$ : all cosmological distances increase with  $R$ . Furthermore, the temperature decreases in inverse proportion:

$$T \propto 1/R.$$

(Note that this means that there are points so far apart that they are separating faster than the speed of light! There is no contradiction here: this is perfectly consistent with general relativity. Locally, special relativity remains valid and  $c$  remains the limiting velocity.)

Now, let us chart the Universe's evolution from a time shortly after the Big Bang.

- $t \simeq 10^{-43}$  s,  $T \simeq 10^{32}$  K  $\simeq 10^{19}$  GeV: The Universe emerges from the Planck era in which quantum gravity was dominant; it is perhaps described by a grand unified theory. The energy density is dominated by very relativistic particles and is falling as  $\rho \propto 1/R^4 \propto T^4$ .
- $t \simeq 10^{-35}$  s,  $T \simeq 10^{28}$  K  $\simeq 10^{15}$  GeV: The grand unified symmetry is broken. What we know as our observable Universe – a region which is today some  $10^{10}$  light years (or  $10^{28}$  cm) in size – is, at this time, contained in a region of space only a millimetre across. A little later, baryogenesis leads to the cosmological excess of matter over antimatter (see below).
- $t \simeq 10^{-10}$  s,  $T \simeq 10^{15}$  K  $\simeq 10^2$  GeV: Electroweak symmetry breaking takes place. The presently observable Universe is contained in a region  $10^{14}$  cm in size.
- $t \simeq 10^{-5}$  s,  $T \simeq 3 \times 10^{12}$  K  $\simeq 300$  MeV: QCD becomes confining: free quarks combine to form hadrons.
- $t \simeq 10^{-2}$  s,  $T \simeq 10^{11}$  K  $\simeq 10$  MeV: The Universe consists mainly of photons, electrons, positrons, neutrinos and antineutrinos. There are small numbers of protons and neutrons which undergo rapid inter-conversions at this high temperature:



Our observable Universe is one light year (or  $10^{18}$  cm) in size. The density is over a billion times that of water.

- $t \simeq 0.1$  s,  $T \simeq 3 \times 10^{10}$  K  $\simeq 3$  MeV: At this temperature it is much easier for the heavier neutrons to turn into lighter protons than vice versa. There are 1.5 times more protons than neutrons.
- $t \simeq 1$  s,  $T \simeq 10^{10}$  K  $\simeq 1$  MeV: Neutrinos and antineutrinos begin to behave as free particles. They decouple from the rest of matter and evolve independently. Electrons and positrons begin to annihilate into photons, increasing the temperature of photons relative to neutrinos:  $T_\gamma = 1.4T_\nu$ .
- $t \simeq 10^2$  s,  $T \simeq 10^9$  K  $\simeq 0.1$  MeV: The Universe is almost entirely made up of photons, neutrinos and antineutrinos, with a small number of

electrons and nucleons. There are now six times as many protons as neutrons. Our observable Universe is about 100 light years (or  $10^{20}$  cm) in size. The density of the Universe is 40 times that of water.

- $t \simeq 3\text{--}4$  minutes,  $T \simeq 8 \times 10^8$  K: Nucleosynthesis begins. During nucleosynthesis, all free neutrons and some free protons are synthesised into the nuclei of light elements: chiefly deuterium ( $^2\text{D}$ ); helium ( $^3\text{He}$  and  $^4\text{He}$ ); and lithium ( $^7\text{Li}$ ). Within a few hours the synthesis is completed, leaving 24% helium by weight and 76% hydrogen (i.e. unused protons), plus smaller amounts of other light elements. However, the Universe is still mostly photons and neutrinos.
- $t \simeq 10^4$  years,  $T \simeq 10^5$  K: The energy density becomes dominated by non-relativistic matter and now only falls as  $\rho \propto 1/R^3 \propto T^3$ .
- $t \simeq 10^5$  years,  $T \simeq 4000$  K: Electrons combine with nuclei to form electrically neutral atoms. With the disappearance of charged particles the Universe becomes transparent (there are now no charged particles to scatter photons). In particular, the cosmic microwave radiation was last scattered at this time. Optical and radio astronomy cannot see back beyond this time.
- $t \simeq 10^9\text{--}10^{10}$  years,  $T \simeq 10$  K: Galaxies form.
- $t \simeq 10^{10}$  years,  $T \simeq 2.7$  K: Today. Size of observable Universe is  $10^{10}$  light years (or  $10^{28}$  cm).

### 48.3 Beyond the Big Bang

The original Big Bang theory was tremendously successful in explaining the origin and gross structure of the Universe. But several questions went unanswered, such as: What is the reason for the large-scale isotropy and homogeneity which the Big Bang theory takes for granted? How did matter come to dominate over antimatter? How did the structures we see, such as galaxies and clusters of galaxies, grow out of the tiny initial fluctuations in the matter density? And what caused the initial fluctuations themselves? The quest to answer these questions has stimulated a new era in observational cosmology in the last few decades, in which precision measurements have led to a consistent, but quite unexpected, picture of the Universe, dominated by radically new sources of matter and energy. The telling of this story forms the

basis of the next few chapters. We begin by describing the theory of *inflation*.

### 48.4 Inflation

The modern period of cosmology could be said to begin with the theory of inflation, a radical theory developed by Alan Guth, Andrei Linde, Paul Steinhardt, Andy Albrecht and others around 1980. The idea of the theory is that the early Universe underwent a period of exponential expansion, increasing in size by a factor of  $10^{30}$  or so. It is not yet known exactly when this period of inflation took place, the only constraint being that it must have taken place *before* the period of nucleosynthesis (after which, as we have seen, the original Big Bang theory gives a coherent and successful description of the Universe's evolution). Inflation was driven by some as-yet-unknown source of energy called the *inflaton*.

Inflation solves a number of problems arising in the original Big Bang theory, as we now describe.

#### 48.4.1 The Flatness Problem

As the Universe evolves in time, any deviation of the energy density from the critical density is exacerbated. The fact that the Universe is flat today, to within a per cent or two, implies that it must have been flat to within one part in  $10^{25}$  at a time  $10^{-10}$  seconds or so after the Big Bang! Such an incredible degree of flatness is easily explained by inflation, since the Universe's exponential expansion would have smoothed out any curvature of space-time, resulting in the flatness we perceive today.

#### 48.4.2 The Horizon Problem

Another mystery is why the observable Universe appears so isotropic and homogeneous on large scales. This problem is exacerbated by the fact that, if we simply extrapolate the standard Big Bang cosmological evolution back in time, we find that regions of the Universe that are observed to be isotropic and homogeneous could never have been in causal contact with each other. That is, the regions are so far apart that there would not have been enough time to send light signals (or any other form of communication between them). So no possible physical process could have occurred to make them isotropic and homogeneous with respect to one another! Inflation again solves this problem with ease: provided inflation lasted for long enough, the entire observable Universe could have

been generated from a single causally connected patch. So, for example, we can understand why the temperature of the CMB is roughly the same everywhere in the sky: we are observing regions that were once contained in a tiny volume of space, in which the radiation was in thermal equilibrium with the same temperature.

#### 48.4.3 *The Origin of Large-scale Structure*

Inflation can explain not only why the Universe is broadly homogeneous and isotropic, but also why we see structure in matter (such as galaxies and clusters of galaxies). The idea is that the inflaton would (like everything else) have been subject to quantum fluctuations and that these fluctuations in the energy density would have been stretched to macroscopic size during inflation and so became the seeds for the matter structures we see today. Since the initial fluctuations were small and have remained so (at least on large distance scales), one can study them using perturbation theory and the resulting theory of *cosmological perturbations* gives a very successful description of the observations of the matter structure on large scales and the anisotropies in the CMB (discussed in the next chapter). Thus, we have compelling quantitative evidence that the Universe we see grew out of a quantum fluctuation in the early Universe!

#### 48.4.4 *The Magnetic Monopole Problem*

One of the remarkable predictions of grand unified theories is that they contain stable *magnetic monopoles*. (The magnetic monopoles arise as subtle configurations in which the grand unified gauge field is wrapped in a topologically non-trivial way around space–time and represent just one of the many ways in which ‘modern’ abstract mathematical ideas such as topology have become central to particle physics.) Unlike electric ‘monopoles’ (i.e. electric charges), a magnetic monopole has never been observed: magnets only seem to come in dipole form. But if grand unification is correct, the Universe ought to be full of magnetic monopoles, which would have been formed as the Universe’s temperature dropped below the GUT scale. Roughly, the monopoles would have been formed as follows: at high temperatures, the gauge field would have been fluctuating in a random way

in space–time. As the phase transition took place, the random variations in the field would have been ‘frozen in’. In some regions of space–time, the field variations would be topologically non-trivial and constitute a monopole. Once formed, such monopoles are stable, so persist to the present day.

Provided inflation happened *after* the GUT symmetry, the absence of monopoles is easily explained: inflation would have diluted them away. Even starting from a high concentration of monopoles, we would be left with fewer than one per patch of observable Universe.

### 48.5 Theories of Inflation

A theory which results in an accelerating expansion is easy to build: as we shall see in Chapter 52, one simply needs to add a cosmological constant term to the Lagrangian. But this is not enough for a theory of inflation, because such a term would dominate the cosmic expansion at all later times, while the Big Bang theory shows that inflation was followed by periods of radiation and then matter domination. What is needed is a theory in which an accelerated expansion occurs and then ends.

This can easily be achieved by coupling general relativity to a theory with a scalar field (the inflaton) which moves in a potential, similar to what we discussed when describing spontaneous symmetry breaking in Chapter 21. If the field begins its motion not at the minimum of the potential, but displaced from it, then it will roll towards the minimum. Provided that the field rolls slowly enough, the value of the potential will remain roughly constant, leading to the required period of exponential inflation. Eventually, inflation will end as the field reaches the minimum of the potential.

Originally, it was hoped that the Higgs could play the role of the inflaton, but it soon became clear that the Higgs potential (whose form is fixed by the observed properties of electroweak symmetry breaking) is too steep to satisfy the slow roll conditions. So an extra scalar field is needed. Even then, it is difficult to understand why the inflaton potential is so flat: again, an unexplained fine-tuning of the parameters of the theory seems to be required. Many more sophisticated models have been put forward to try to address this and other issues.

## *The Cosmic Microwave Background*

### 49.1 Introduction

When the plasma constituting the early Universe cooled sufficiently for electrons and nuclear ions to form atoms, the Universe went from being opaque to transparent. From that point on, radiation, in the form of photons, decoupled from matter and remains today, with a spectrum described approximately by that of black-body radiation (acquired from when they were in thermal equilibrium with matter), but with an apparent temperature that is cooled by the amount that their wavelengths have been increased by the subsequent expansion of the Universe. The spectrum peaks in the microwave region.

### 49.2 Observations of the CMB Anisotropy

Early observations of the CMB suggested it to be isotropic, with a temperature of  $T \simeq 2.7$  K. In fact, there is a small dipole anisotropy (meaning that the temperature is larger in one hemisphere of the sky than another), corresponding to the fact that the Earth (from which we observe the CMB) is moving with respect to the cosmic rest frame. The size of the dipole (around 0.003 K), tells us that this *peculiar velocity* is  $370 \text{ km s}^{-1}$ . But there are also yet smaller anisotropies in the temperature, at the level of one part in  $10^5$ , which reveal much more fundamental information about the Universe.

The anisotropies, which had been predicted long before, were not observed until 1984 by the NASA

COBE satellite, a discovery which was awarded the Nobel Prize in 2006. Since then, their form has been probed in ever finer detail.

The CMB anisotropy data are commonly presented in terms of the *angular power spectrum*, in which fluctuations in the temperature over the sky are resolved into their angular sizes measured by a parameter  $\ell$ . Thus, the dipole caused by the Earth's peculiar velocity corresponds to  $\ell = 2$ , while increasing values of  $\ell$  correspond to fluctuations on decreasing angular scales.

The two most significant observations of the CMB over the whole sky were those carried out by the NASA WMAP satellite from 2001 to 2010 and by the ESA Planck satellite from 2009 to 2013. These enabled measurements of the angular power spectrum with unprecedented accuracy, down to increasing small angular scales. Figure 49.1 shows the temperature fluctuations observed by Planck and Figure 49.2 shows the resulting angular power spectrum.

In addition, dedicated experiments such as the South Pole Telescope and the Atacama Cosmology Telescope probe the CMB in smaller regions of the sky, but with angular resolution down to the arcminute scale (corresponding to the extreme right-hand side of Figure 49.2). These telescopes are located in regions of the Earth where the atmosphere is thin and dry, so as to minimise the effect that microwave ovens exploit, namely that water vapour absorbs microwaves!

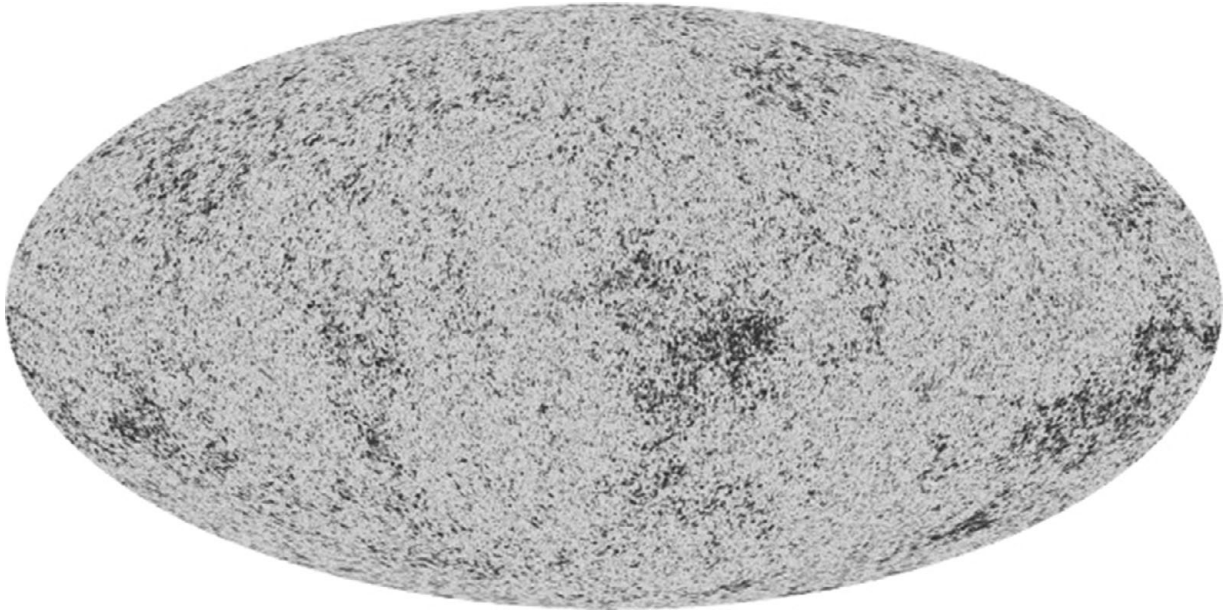


Figure 49.1. A map of the sky showing temperature fluctuations in the microwave background. The average temperature is 2.73 K and the fluctuations are around a millionth of a degree. *Courtesy: ESA and the Planck Collaboration.*

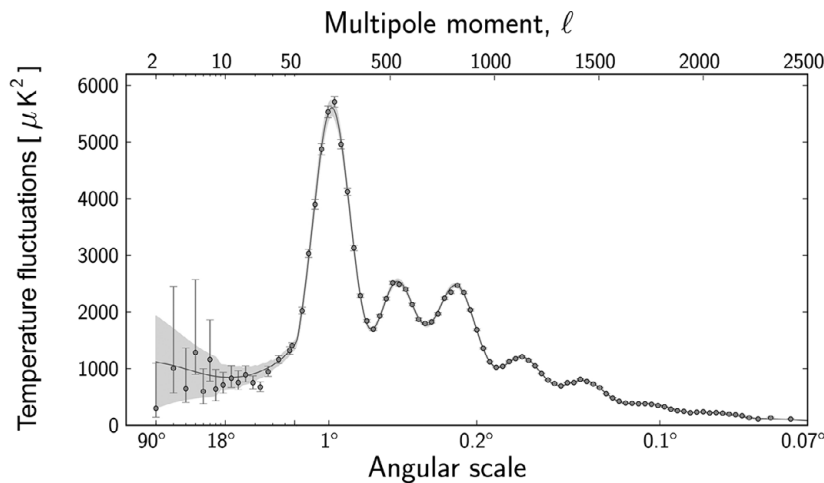


Figure 49.2. Angular power spectrum of the CMB observed by Planck. The peaks reflect acoustic oscillations in the primordial plasma. *Courtesy: ESA and the Planck Collaboration.*

### 49.3 Physics of the CMB Anisotropy

The peak structure of the angular power spectrum in Figure 49.2 appears complicated, but is qualitatively easy to understand. The CMB carries a snapshot of the Universe at the period when the photons decoupled from matter. During this period, the

Universe consisted of a hot plasma whose dynamics were dominated by photons and baryons.

The competing effects of photon pressure (which tends to erase the fluctuations imprinted by inflation) and the gravitational attraction of baryons (which tends to reinforce fluctuations) lead to

oscillations of the plasma medium, i.e. sound waves. The peaks in the angular power spectrum correspond to scales at which the photon effects are minimised.

The detailed structure of the peaks (for example their positions and relative heights) depends on the detailed composition of the plasma. We have already seen this for the photons and baryons. But now suppose that the plasma contains a matter component which does not interact with the photons. This matter, like baryonic matter, clusters because of gravity, but, unlike baryonic matter, this tendency to clump is not countered by the pressure of photons via interactions with them. Thus, the different matter components behave differently in the plasma and it is possible to determine how much of each is present from the resulting power spectrum. A similar story applies for dark energy (see Chapter 52), which also does not clump, but rather causes the plasma to expand.

In this way, precise measurements of the angular power spectrum enable us to build up a detailed picture of the plasma's evolution and of its constituents. The results are remarkable: the 2013 Planck data, for example, tell us not only the precise age of the Universe ( $13.8 \pm 0.04$  billion years old) and value

of the Hubble constant ( $67.8 \pm 0.8 \text{ km s}^{-1} \text{ Mpc}^{-1}$ ), but also its composition:  $4.82 \pm 0.05\%$  ordinary matter,  $25.8 \pm 0.4\%$  dark matter and  $69 \pm 1\%$  dark energy. It seems that the Universe is dominated by dark matter and dark energy, whose form we do not yet understand!

#### 49.4 CMB Polarisation

As well as exhibiting anisotropies in temperature, the CMB shows anisotropies in its *polarisation*. For the CMB, it is convenient to decompose the polarisation into so-called 'E-modes' and 'B-modes' (in analogy with electric or magnetic fields, an E-mode can be thought of as having no rotation, while a B-mode has no source). B-modes are especially interesting because of an as-yet-untested prediction of the theory of inflation, which is that it creates gravitational waves, which can then act as a source of B-modes in the CMB. Both E- and B-modes in the CMB have recently been detected. Unfortunately, there are other sources of B-modes beyond gravitational waves, and up until now it has not been possible to disentangle these effects. There are great hopes for the future however.

## *The Matter–Antimatter Asymmetry*

### 50.1 Introduction

Along with the discovery of antimatter in the 1930s came a new mystery: if matter and antimatter are so similar (for example every particle has its antiparticle, with equal mass and spin, and opposite charge), why does the Universe appear to be made up exclusively of matter with little or no antimatter?

Indeed, it now seems likely that there are no large concentrations of antimatter in the Universe, and almost certainly none within our local cluster of galaxies. The question naturally arises of how this has come about. Until the advent of inflation, it was logically possible (if philosophically unattractive) that at the beginning of the Big Bang some divine guiding hand arbitrarily decided on one or another. But inflation means that any initial asymmetry would be diluted away, so some dynamical mechanism is certainly needed.

By matter we mean, of course, ordinary baryonic matter; so, we must explain how a net excess of baryons over antibaryons was generated. In 1967, the Russian physicist (and political dissident) Andrei Sakharov showed that the generation of a net baryon number requires: (1) baryon-number violation; (2) CP and  $C$  violation (otherwise, the rates of reactions producing quarks and antiquarks will be equal); and (3) non-equilibrium (otherwise in equilibrium, CPT conservation requires the number of baryons and antibaryons to be equal).

Now, it turns out that all of these ingredients are present in the Standard Model: the weak interactions violate both  $C$  and CP, while (at least if the Higgs mass is light enough) the electroweak phase transition can be strongly first-order, proceeding out of equilibrium through the nucleation and growth of bubbles of the broken electroweak phase. Even baryon number, which is conserved at the perturbative level in the Standard Model, can be violated by non-perturbative quantum mechanical processes. The reason for this is that both baryon- and lepton-number symmetries suffer from a quantum anomaly (see Chapter 37), with only their difference  $B - L$  being exactly conserved. Thus violation can occur through non-perturbative quantum tunnelling processes, discovered by Klinkhamer and Manton in 1984 and termed *sphalerons*. These processes are exponentially suppressed at low energies (which is why we see no evidence of baryon-number violation around us) but can occur freely above the electroweak phase transition.

Unfortunately, while all three effects are present in the Standard Model, none of them is large enough to generate the observed baryon asymmetry. Thus, physics beyond the Standard Model is needed. We outline three very different theories below.

### 50.2 GUT Baryogenesis

Grand unified theories treat leptons and baryons as the same (they live in the same GUT multiplets),



so certainly possess the requisite baryon-number violation. Furthermore, CP and C violation are already known to be present, even in the Standard Model, and if the Universe goes through a stage in which reactions occur out of equilibrium, then a non-zero baryon number can be generated. This process is called ‘GUT baryogenesis’ and was once regarded as a major selling-point for GUTs.

The chain of reasoning went roughly as follows. At a time less than  $10^{-35}$  s after the Big Bang, the temperature of the fledgling Universe would have been higher than  $10^{28}$  K, corresponding to an average energy of the material particles more than  $10^{15}$  GeV ( $\simeq M_X$ ). In this regime the super-heavy gauge bosons,  $X$ , and their antiparticles,  $\bar{X}$ , would have been produced with ease in particle collisions and the equal populations of  $X$  and  $\bar{X}$  would have remained in thermal equilibrium (i.e. as many would be produced in collisions as would be annihilated):

$$X + \bar{X} \leftrightarrow \text{matter and radiation.}$$

But, as the Universe expanded it also cooled. Very soon the temperature would have fallen below  $M_X$  and the  $X$  and  $\bar{X}$  bosons could no longer be produced, as the average collision energy would have been too low. By the same token, they would have been unable to annihilate, as the expansion of the Universe destroyed equilibrium. That is, the Universe expanded faster than the bosons could interact. Then, the large numbers of  $X$  and  $\bar{X}$  bosons would have begun to decay. Because of the presence of the CP-violating effects described in Chapters 39 and 47, there is no guarantee that the average value of the baryon number of the states into which the  $X$ s decayed would be exactly opposite to that of the states into which the  $\bar{X}$ s decayed.

So the CP-violating decays of the  $X$  and  $\bar{X}$  bosons could generate a net baryon number for the Universe from an initial state consisting of equal numbers of  $X$ s and  $\bar{X}$ s. As the average energy of the particles in the Universe would then have continued to fall, baryon-number-violating processes would have become increasingly insignificant and the net baryon number would thus have become frozen. This is generally stated as a ratio of the net number density of baryons  $n_B$  to the number density of cosmological Big Bang photons  $n_\gamma$  ( $\simeq 400$  per cubic centimetre). The observed value of this ratio, namely

$$n_B/n_\gamma = (4 \pm 1) \times 10^{-10},$$

could be reproduced in many GUTs.

Unfortunately, it was eventually realised that the baryon asymmetry created in this way in standard GUTs would be completely removed (or ‘washed out’) by sphaleron processes during the later electroweak phase transition. The reason for this is that the standard GUTs conserve  $B - L$ , so in fact any baryon number generated is accompanied by an equal lepton-number asymmetry. Hence  $B - L = 0$  and so we can write  $B = (B + L)/2$ . But  $B + L$  is precisely the combination which is anomalous in the Standard Model and will be destroyed by sphaleron processes in the electroweak phase transition! So, any baryon asymmetry created in the early Universe will have disappeared by the time the electroweak phase transition is complete, and cannot persist to this day.

### 50.3 Baryogenesis via Leptogenesis

The washout argument killed baryogenesis in the standard GUT scenarios, but it also opened the door to completely new mechanisms of baryogenesis. For example, it implies that if an asymmetry in *lepton* number can be generated in the early Universe, then sphaleron processes during the electroweak phase transition will convert half of this into a baryon asymmetry.

As Fukugita and Yanagida recognised in 1986, this mechanism can be realised in the simple extension of the Standard Model in Chapter 43, where we introduced heavy right-handed neutrinos to generate light neutrino masses via the see-saw mechanism. The combination of Dirac and Majorana mass terms for neutrinos means that lepton number is necessarily violated in the model. Moreover, the mass terms can also feature CP-violating phases and decays of the heavy neutrinos can provide the necessary out-of-equilibrium dynamics.

In more detail, the idea is that the heavy neutrinos  $N$  can decay to either a light neutrino or a light antineutrino (along with a Higgs boson), violating lepton number. The interference between tree-level and loop diagrams for these decay processes allows C and CP violation to enter, such that all of Sakharov’s conditions are satisfied.

Leptogenesis provides a simple and successful theory of the origin of the matter–antimatter asymmetry. Unfortunately, because it involves processes at the see-saw scale of c.  $10^{15}$  GeV, it is hard to imagine a definitive experimental test that will confirm or refute the proposal.

#### **50.4 Electroweak Baryogenesis**

Although it is known that the Standard Model cannot account for baryogenesis, it is possible that some electroweak-scale extension of it could. This is particularly interesting given the motivation from the Higgs hierarchy problem for new particles at or below the TeV scale.

Even simple modifications of the Standard Model can be enough to allow electroweak

baryogenesis. For example, adding an additional Higgs field, or even just an additional singlet scalar boson can be enough. These minimal extensions can also be accommodated in more ambitious theories which attempt to solve the hierarchy problem. The composite Higgs model based on  $SO(6)/SO(5)$ , for example, which contains an additional singlet, has been shown to be capable of doing the job.

# 51

## *Dark Matter*

### 51.1 Introduction

One of the conclusions of the analysis of the recent CMB data in Chapter 50 is that the energy density stored in baryonic matter is dwarfed by a factor of five by that of non-baryonic matter. In fact, evidence for a significant component of non-baryonic, or dark matter, in the Universe has been around for decades, from a variety of sources, as we now review.

### 51.2 Gravitational Evidence for Dark Matter

#### 51.2.1 Galaxies

One way in which one can deduce the mass of the Sun is by measuring the orbital velocity  $v$  of the Earth and the radius  $r$  of its orbit. Then by equating the force required for a body to move in a circle with the gravitational force given by Newton's law, one finds that the velocity is given by

$$v = \sqrt{\frac{GM}{r}}, \quad (51.1)$$

where  $G$  is Newton's gravitational constant and  $M$  is the solar mass. By using the same trick of comparing the orbital velocities and radii of objects in a galaxy, astronomers are able to determine how matter is distributed within the galaxy. In doing so, they invariably find that the amount of luminous matter (stars, dust, gas and so on) is too small to account for the measured orbital velocities. For our own galaxy, the Milky Way, the visible mass is too small by a factor of ten or so. Nowadays, some of this mass is believed to be

accounted for by black holes<sup>1</sup> near the centres of galaxies, but most of the unexplained matter must be distributed throughout the galaxy in order to explain the data.

One explanation for this is to propose that galaxies are permeated by a *halo* of additional, invisible matter, called *dark matter*. The halo is typically much larger than the region occupied by the bulk of the visible matter and spherically shaped rather than planar. This suggests that the dark matter is largely collisionless, interacting with gravity in the normal way, but only very weakly with itself.

#### 51.2.2 Galaxy Clusters

The first evidence for dark matter actually came from studies of the dynamics of clusters of galaxies in the 1920s and 30s. Based on observations of the brightness of galaxy clusters and the motions of galaxies near their edges, it was possible to estimate the visible and gravitational masses. Though the original measurements have proven to be unreliable, observations today suggest that as little as 2 % of the total mass of a galaxy cluster is visible.

#### 51.2.3 Gravitational Lensing

More evidence for dark matter in galaxy clusters was provided by the observation of *gravitational*

<sup>1</sup> Black holes are objects which are so massive that not even light can escape their gravitational pull (see Part XIII).

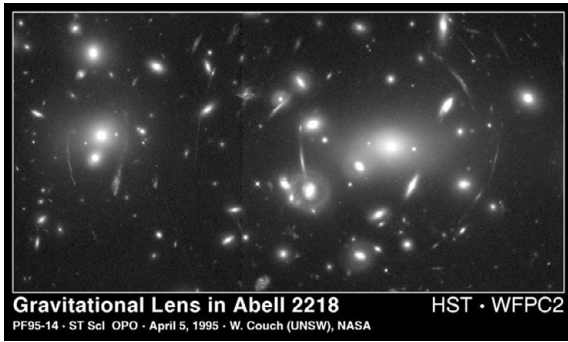


Figure 51.1. Image from the Hubble Space Telescope showing gravitational lensing of distant light sources. The light comes from proto-galaxies which are in a very early stage of formation. *Source: NASA and STSci.*

*lensing* effects by the Hubble Space Telescope. As we saw in Chapter 5, in General Relativity all sources of matter and energy feel the influence of space–time curvature, including light. Thus light itself can be bent, or lensed, as it propagates through the curved space–time surrounding massive bodies. Figure 51.1 shows a picture from Hubble in which the starlight from distant sources has been smeared (or lensed) by the gravitational effects of intervening, invisible matter. Perhaps the most spectacular evidence for dark matter from gravitational lensing comes from the *bullet cluster*, which is actually a pair of clusters which underwent a collision with each other. It is possible to track separately the distributions of stars (through their visible light), hot gas (through X-ray emissions) and dark matter (via gravitational lensing). The observations shows that while the stars and dark matter passed through each other largely without interacting, the accompanying clouds of hot gas (which contain the bulk of the baryons and interact electromagnetically) were slowed by the collision. Figure 51.2 shows a composite image with the X-ray distribution superimposed with the inferred matter distribution from lensing. One can see how the X-ray gas not only has slowed relative to the gravitational matter due to the collision, but also that a shock wave (visible as the cone-shaped structure in the gas on the right) has formed as a result of the collision.

#### 51.2.4 CMB

As we saw in Chapter 50, the CMB also provides compelling evidence for dark matter. Together

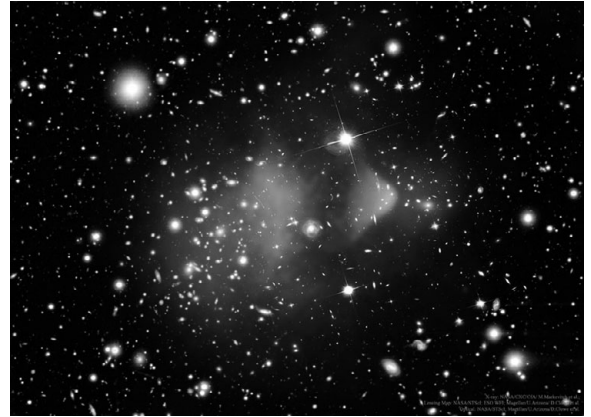


Figure 51.2. Composite image of the bullet cluster, showing the X-ray distribution (central bright region) superimposed with the inferred matter distribution from lensing (outer bright region). *Source: NASA.*

with observations on the large scale structure of matter, astronomers have been able to build up a more detailed picture of the properties of dark matter. In particular, it is now known that dark matter is *cold*, meaning that dark matter became non-relativistic (slowly-moving) early in the Universe. This has the effect that the structure we see today is formed by a *bottom-up* process, in which small scale structures grow into larger and larger structures via gravitational attraction, in accordance with observations.

There are, by now, many other different observations inferring the existence of dark matter through its gravitational effects. All are compatible with the CMB result that dark matter makes up a quarter of the energy density of the universe. Thus, it has become a most pressing question to figure out what dark matter actually is! We now discuss some possible candidates.

### 51.3 Dark Matter Candidates

#### 51.3.1 Neutrinos

An early idea for dark matter was that it consists of massive neutrinos (see Chapter 43), but this has been ruled out. Not only is the contribution to the matter density too small, but also neutrinos behave like warm dark matter rather than cold dark matter and result in a pattern of structure formation that disagrees with observations.

Thus it seems that dark matter requires an explanation in terms of physics beyond the Standard Model.

### 51.3.2 Primordial Black Holes

Primordial black holes are black holes that are supposed to have formed in the early universe due to sufficiently overdense fluctuations in the energy density. Such black holes differ from the usual black holes formed by stellar collapse in that they can be much lighter. But such black holes emit light quantum-mechanically by the process of Hawking radiation and so if they are too light would have evaporated before the present day. Thus, primordial black holes making up the dark matter must be rather heavy, compact objects, meaning that they focus (via gravitational lensing) the light from distant objects, making them appear brighter. Searches for such *microlensing events* have put strong constraints on primordial black holes as dark matter candidates.

### 51.3.3 Supersymmetric Particles

As we saw in Chapter 45, until recently one of the most popular explanations for dark matter was the lightest supersymmetric particle. If stable, weakly interacting, and with a mass in the TeV range required to solve the Higgs hierarchy problem, such particles can have a thermal relic density in just the right ballpark to explain the observed dark matter. But the popularity of supersymmetric dark matter has waned in recent years, not only because of the absence of evidence for them at the LHC, but also because direct and indirect searches for dark matter (see below) have also imposed strong constraints on WIMPS.

### 51.3.4 Axions

In Chapter 47, we saw how the axion could solve the strong CP problem of the Standard Model. The axion is uncharged and extremely weakly interacting so makes for a natural dark matter candidate. There is also an appealing mechanism for how the axion generates the observed dark matter density. The idea is that, if the axion field is misaligned from its minimum in the early universe (e.g. by the dynamics of inflation), then it will subsequently undergo damped oscillations about the minimum. The oscillations make a contribution to the energy density which precisely mimics that of non-relativistic matter.

The rather peculiar properties of the axion mean that it cannot be searched for in conventional ways. Rather, dedicated searches are required, many of which are still under development. One common theme is that the axion  $a$  couples to electromagnetic fields via an interaction of the form  $a\mathbf{E} \cdot \mathbf{B}$ . So for example, one can look for an axion by shining light through a wall in the presence of a background electromagnetic field: without the axion, the light will be blocked, but if the axion exists, a photon can convert into an axion, pass through the wall, and reconvert into a photon to be detected on the other side!

## 51.4 Searches for Dark Matter

Conventional weakly-interacting dark matter particles are searched for in one of two main ways. In *direct detection* experiments, one looks for recoils of galactic dark matter particles in the Milky Way with nuclei in an Earth-based detector. Such interactions are expected to be extremely rare and the recoil energies are low, meaning that the experiments demand very careful control of other backgrounds. Nevertheless, a slew of such experiments in recent decades have managed to put strong constraints on weakly-interacting particles with masses in the range 10–500 GeV. Unfortunately, the experiments are starting to reach the level at which the background from scattering of neutrinos (from e.g. the Sun or the atmosphere) on nuclei becomes significant, making it difficult for the limits to be pushed much further.

*Indirect detection* experiments search for the particles produced when pairs of dark matter particles annihilate in space. Such annihilations are expected to occur in regions such as the centre of our galaxy where the concentration of dark matter is high. For example, the supersymmetric partners of the  $W^\pm$  gauge bosons could annihilate to form a pair of photons, which would form a spectacular signal. The difficulty with such experiments lies in discriminating dark matter events from the many other possible astrophysical sources. Nevertheless, indirect detection experiments have managed to set impressive limits on various dark matter scenarios.

## *Dark Energy*

### 52.1 Introduction

As well as the inference that there is roughly five times more matter than antimatter in the Universe, the CMB measurements deliver the remarkable conclusion that most of the energy density in the Universe, nearly 70%, is not matter at all! While matter has the effect of causing the expansion of the Universe to *decelerate* (due to gravitational attraction), this mysterious *dark energy* causes the expansion to *accelerate*. Let us now tell its story.

### 52.2 Einstein's Cosmological Constant

The story of dark energy really begins with Einstein himself who, as we saw in Chapter 48, introduced a cosmological constant into the equations of general relativity, in order to try to stabilise the Universe: while the gravitational attraction of normal matter (which certainly is present in the Universe) causes the cosmos to contract, a cosmological constant causes it to expand. So by carefully adjusting the value of the constant in the theory, the Universe can be made static.

The cosmological constant was quickly discarded once it was realised that the Universe was not static but rather expanding. But, as theoretical physicists came to better understand quantum field theory, they realised that every such theory has a vacuum energy and that such a vacuum energy has precisely the effect of the cosmological constant when the theory is coupled to gravity. Thus, a cosmological

constant seemed inevitable, from the theorists' point of view. Theorists could even estimate how large the vacuum energy should naturally be. For example, in a theory with a particle of mass  $m$ , the energy density would be at least  $m^4$ . Thus, the very existence of the proton alone would suggest a vacuum energy density  $\simeq (\text{GeV})^4$ .

The trouble with this is that, even before dark energy was found, measurements of the expansion rate indicated that the *total* energy density was more like  $(10^{-3}\text{eV})^4$ . As a result, many physicists assumed that the cosmological constant must be exactly zero. For example, it was hoped that supersymmetry might be able to do the job (though it turns out that it cannot, even in the absence of the needed low-energy supersymmetry breaking). Over the years, thousands of scientific papers were written trying to prove the vanishing of the cosmological constant. But this industry came to an abrupt halt in the 1990s with the experimental discovery that the expansion of the Universe is, in fact, currently accelerating, meaning that a non-vanishing vacuum energy (or something with very similar properties to it, given the generic name of dark energy) was needed.

### 52.3 Supernovae and Dark Energy

The discovery of the accelerated expansion of the Universe originally came by extrapolating Hubble's original measurements of the velocity–distance

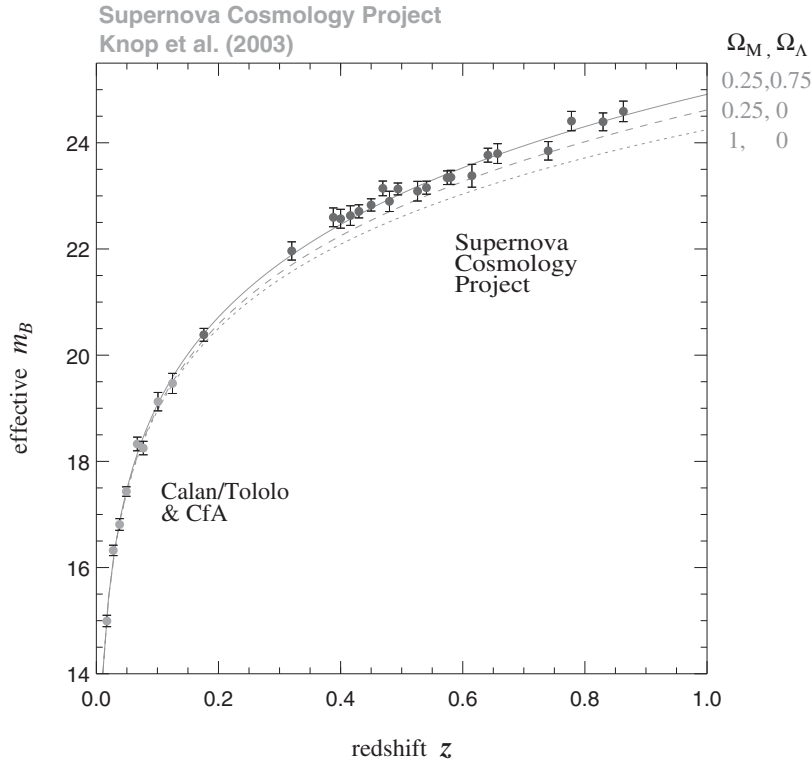


Figure 52.1. Data showing the apparent brightness (a measure of the distance) versus redshift of type IA supernovae. The solid line shows a fit in which the Universe has critical energy density, with one-quarter matter ( $\Omega_M = 0.25$ ) and three-quarters dark energy ( $\Omega_\Lambda = 0.75$ ). Source: *Supernova Cosmology Project*.

relation of galaxies (which provided the first evidence for the universal expansion and the Big Bang) to objects which are much, much further away from us. Such distant objects tell us about the pattern of the Universe's expansion over much longer timescales, up to eight billion years ago, and thus enable changes in the velocity of the expansion, i.e. accelerations or decelerations, to be observed.

While the idea is simple in principle, the necessary measurements are extremely difficult in practice. The main obstacle is making reliable measurements of the distances of objects. To enable such measurements to be made, astronomers observed not galaxies, but *supernovae*: catastrophic explosions caused by the gravitational collapse of very massive stars, once the nuclear processes that sustain them (the same as those that power our Sun) have been exhausted. Over the years, it has been observed that, at least for a sub-class of supernovae (called *type IA supernovae*) this process

appears to happen in an identical way. Thus, the distances of such supernovae can be estimated by assuming that every supernova is equally bright intrinsically, such that their apparent brightness viewed from Earth is inversely proportional to the square of their distance. The recession velocity of a supernova is determined by the redshifting of the light which is emitted.

Two separate experiments, led by Saul Perlmutter and Adam Riess, observed hundreds of such supernovae. They both found that the supernovae are dimmer than expected on the basis of the standard Friedmann–Robertson–Walker cosmologies described earlier and hence further away. Figure 52.1 shows data from the Supernova Cosmology Project, along with fits to the data with and without dark energy. Not only is there a clear preference for dark energy, but also we see that the best fit is obtained for a Universe with the critical energy density divided into three-quarters dark energy and one-quarter matter (dark and baryonic), in

precise accord with observations made on the CMB and elsewhere.

### 52.4 The Cosmological Hierarchy Problem

With the observation of dark energy or a cosmological constant with energy density  $(10^{-3}\text{eV})^4$ , it was pointless for theorists to continue seeking a beautiful theory in which it exactly vanished. Rather, a theory was needed which predicted how the cosmological constant could be very small compared to the scales of particle physics (much smaller than, say, the proton mass), but very large compared to the scales of cosmology, such that it dominates the expansion of the Universe today.

This dichotomy is very similar to the *Higgs hierarchy problem* for the mass of the Higgs boson that we encountered in Chapter 41. Without supersymmetry, we would expect a Higgs boson mass of around  $10^{15}$  GeV, rather than the mass of 125 GeV which experiments find. A similar, but far worse, hierarchy problem occurs for the cosmological constant: the quantum corrections coming from, say, quantum gravity at the Planck scale suggest a cosmological constant over  $10^{120}$  times larger than the measured value! Even the contributions of size  $(\text{GeV})^4$  expected from the strong nuclear force are far too large.

### 52.5 The ‘Why Now?’ Problem

This problem of why the dark energy is so small compared to the scales of particle physics has no satisfactory explanation and indeed is regarded by many as the most fundamental problem in physics today. However, there is *another* problem with dark energy, namely: why is the density of dark energy today (three-quarters of the total) about the same as the density of matter (one-quarter of the total)? This similarity is peculiar for the following reason. As the Universe evolves, the density of dark energy stays the same (as expected for something which is like a cosmological constant) but the density of the other matter decreases rapidly as the Universe in which it is contained expands. So why is it that, just at the epoch when humankind happens to gaze at the sky, the contributions of dark energy and all other matter and energy happen to be about the same? This *cosmic coincidence* (which surely can be no coincidence!) also goes by the name of the ‘*why now?*’ *problem*; it too lacks a satisfactory explanation.

### 52.6 The Anthropic Principle

Despite thousands of scientific papers on the subject, it is fair to say that no one has yet come up with a satisfactory dynamical explanation of the size of the cosmological constant. This has led some physicists to drastic ideas. One idea is that perhaps the value of the cosmological constant is set not by some fundamental dynamical mechanism, but rather by the fact that, were it not to have such a value, we would not be around to observe it. Indeed, as Steven Weinberg pointed out in the late 1980s, if the constant were much larger, it would cause the Universe to fly apart before galaxies and other structures necessary for life could form, while if it were much smaller and negative, the Universe would recontract too soon. Thus, the idea has come about, going under the name of the *anthropic principle*, that perhaps the small value of the cosmological constant is not fundamental, but rather a manifestation of *environmental selection*: it takes the value it does, because this is a value which allows life to exist.

To give an analogy, centuries ago scientists struggled to come up with theories of the Earth–Sun distance. But with the discovery of the other planets, and other stars (which we now know have their own planets), it became clear that the Earth–Sun distance is not some fundamental parameter of physics. It is set by the fact that it provides a temperature on Earth in which water can exist in the liquid state, which is very convenient for the evolution of life (as least, for life as we know it!).

But how could it be that the cosmological constant is environmentally selected? By analogy with our Earth–Sun example (where there are lots of ‘planet–star distances’ and the one which is selected is one which supports life), we need a theory in which there exist lots of possible values of ‘the’ cosmological constant, including the observed one, and a mechanism by which the observed one can be selected. Remarkably, there are hints that string theory, to which we now turn, is just such a theory.

### 52.7 Summary

The current era of precision cosmology has thrown up several new sources of matter and energy which, if they are real, must have some explanation in terms of elementary particle physics. Moreover, as we have seen in earlier chapters, the unprecedented



precision of current cosmological experiments means that cosmology is, for the first time, able to give quantitative information about particle physics. Examples we have discussed include bounds on the neutrino masses (Chapter 41) and constraints on the nature of the lightest supersymmetric particles (Chapter 44), but

there are many more. As the technological limit on what is achievable with Earth-based particle physics experiments looms closer, and as cosmological experiments become more and more precise, it is likely that we shall see these two sciences, the study of the very small and the very large, converge.

## **Part XIII**

# **Gravity and Gravitational Waves**



## *From General Relativity to Gravitational Waves*

### 53.1 Introduction

The previous few parts have summarised the progress of particle physics in the opening decades of the twenty-first century. The chapters in Part X described the success of the Standard Model and its eventual triumphal confirmation with the discovery of the Higgs boson in 2012. Part XI went on to describe the tantalising glimpses of physics beyond the Standard Model such as experiments demonstrating the reality of neutrino oscillations and went on to explain the range of theories thereby vying to succeed it. Next in Part XII we surveyed the advances in astroparticle physics and cosmology which have provided invaluable observational data to reinforce that from terrestrial accelerators but which have also added to the list of outstanding mysteries requiring explanation in future.

In this Part XIII, however, we survey the parallel advances in gravitational physics leading to the most spectacular confirmation of one of physics most long-standing predictions: gravitational waves. Occurring as recently as 2015, it is clear that the discovery of the waves and its consequences will have a major impact on the fields of astroparticle physics and cosmology which will contribute directly to connecting the phenomena in the cosmos with those in our accelerators here on Earth.

Arising as a direct consequence of Einstein's general theory of relativity first formulated in 1915, the waves remained a hypothetical curiosity for

decades until the formulation of a believable theory of how they might be created and until the advent of technologies capable of detecting them. Whereas the Higgs boson took 45 years from its proposal to discovery, Einstein's waves took a century. So, confirmation of some of the more recent exotic proposals should be regarded as potentially very long-term projects!

We saw in Section 5.2 that Newton's theory of gravity, although perfectly adequate for most practical purposes of calculating gravitational effects in the everyday world, is not compatible with the special theory of relativity. It requires instantaneous action at a distance whereas special relativity imposes a universal speed limit of the speed of light.

Also by the turn of the last century, Newton's theory was found unable to account for certain astronomical observations, the most famous being the advance of the perihelion of the elliptical orbit of Mercury (i.e. the planet's closest point of approach to the Sun). Although small at a minute 0.12 degrees of arc per century, such a discrepancy had been measured and remained unexplained.

In Einstein's astonishing paper of 1915 he starts from the principle of equivalence between gravitational and inertial mass which, when cast in mathematical form, leads to the field equations shown in Section 5.2.1. Using this and with only the meagre calculating aids then available, Einstein was able to account for the discrepancy in the prediction of the advance in

the perihelion of the orbit of Mercury's orbit. Also, the paper predicted an entirely new effect, the bending of light around massive objects such as the Sun. This latter effect was duly observed in 1919 by the world-famous expeditions led by Sir Arthur Eddington to observe the eclipse of the Sun which duly found the deflection of light exactly in accord with Einstein's prediction. As we saw in Chapter 51, this gravitational lensing can now be used as an observational tool to measure mass densities in the Universe.

The derivation of gravitational waves (GWs) from the field equations of general relativity (GR) follows a very similar mathematical path to the derivation of the propagation of electromagnetic waves from Maxwell's equations. Just as the oscillations of an electrical charge under the influence of an alternating voltage applied lead to the emission of alternating electric and magnetic fields (electromagnetic radiation), so oscillating masses acting under the influence of gravity will lead to gravitational radiation. The key difference, however, is that whereas EM waves propagate through familiar fixed three-dimensional space, GWs are the propagation of ripples in Minkowski's four-dimensional space-time itself.

Having derived the propagation of such waves, their signature would be the tiny oscillation in everyday spatial dimensions detectable by the variations in length between two defined points. For many years, the prospect of discovering gravity waves was regarded as a faint hope. Indeed, in the first edition of this book published in 1984, gravitational waves were rated as tenth out of ten in the list of prospective Nobel achievements and awarded 'extremely long odds'.

This may have been due to the controversy of Joseph Weber's resonant mass experiments of the 1960s and 1970s in which large blocks of aluminium were fitted with piezo-electric detectors in the hope of monitoring the disturbances resulting from passing gravity waves. Although Weber claimed to have positive observations, which he stood by until the end of his life, no other experiments could repeat his findings. In light of the actual modern detection, his apparatus would have been many orders of magnitude below the sensitivity required for actual detection.

### 53.2 Hulse–Taylor Variation in Binary Pulsar Periodicity

A clue to the best hope of detecting the waves arrived in the 1960s and 1970s with the dawn of

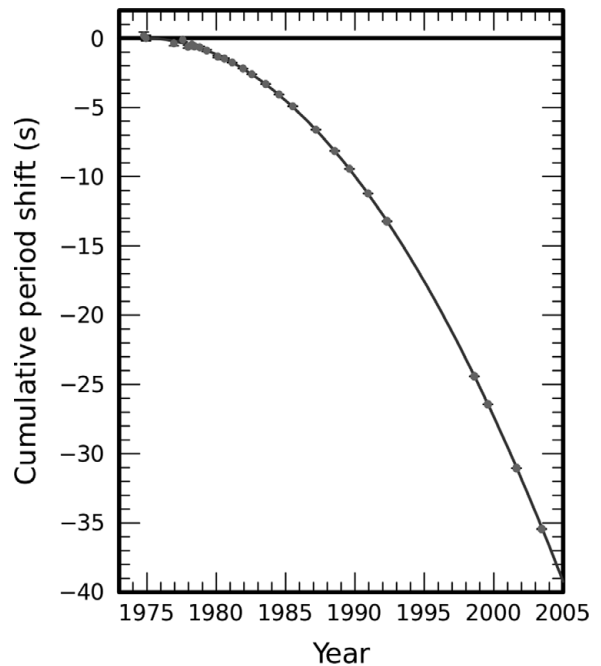


Figure 53.1. Almost 30 years of data illustrate the changing size of the orbit of the Hulse–Taylor binary pulsar. The shrinking orbit leads to a change in the orbital period that is measured each year. The measured points are plotted along with the theoretical prediction. Credit: J. M. Weisberg and J. H. Taylor, *ASP Proceedings*, 2005.

the modern era of radio astronomy. It was to be the cosmos, not the laboratory, that would be the most promising arena for their detection.

The discovery in 1970 of a binary system of pulsars (see following Section 53.3) rotating around each other was for decades the best indirect sign of the existence of gravitational waves. As the pulsars rotate the system emits energy in the form of gravitational waves which results in a decrease in the rate of rotation. Over the decades since, the observation of the one system discovered shows a decrease in the rate of rotation exactly as predicted by the emission of GWs according to GR, see Figure 53.1. For this discovery the American astronomers R. A. Hulse and J. H. Taylor were awarded the Nobel Prize in 1993.

### 53.3 Modern Astrophysics

In Part XII, we reviewed our current understanding of the cosmos at the largest scales such as the observed expansion, and acceleration of this

expansion, of the Universe resulting in the increasing velocities of recession of distant galaxies.

We also summarised in Chapter 48, the chronology of the Big Bang from the supposed instant of creation through an initial shock of inflation to the era of cosmic particle physics culminating at the time of last scattering of the photons now imprinted in the CMB. This was then followed by the billions of years of galaxy and stellar formation.

The final element in this story of the Universe necessary for the understanding of GWs is that of stellar evolution. At the risk of brutal over-simplification, this can be summarised as follows:

Stars are formed when clouds of interstellar gas of atoms, molecules and dust begin to concentrate under gravitational attraction, thereby attracting more and more matter, becoming denser and denser and hotter and hotter. Eventually seeds of stars coalesce, perhaps involving thousands of such seeds in a vast cloud of the gas. Some seeds fall into orbit around each other and lead to the formation of a binary star system, other seeds remain loners.

When the temperature has risen to around  $10^6$  K, hydrogen nuclei begin to fuse into deuterium and helium. Hence a star is born and begins to shine. What happens subsequently depends crucially on the mass of the star.

For stars of around our Sun's mass,  $M_{\odot}$  or less, hydrogen and helium fusion is as far as the fusion process goes. The very stable equilibrium between the attractive force of gravity and the repelling force of radiation resulting from the fusion reactions allows these lighter suns to exist for billions of years. Eventually these stars pass through a red giant phase marking the transition to the dominance of helium fusion in the star's core during which time the star becomes enormously bigger. This can then lead to the inflation of an outer-shell of gas we observe as planetary nebulae, leaving a small bright core as the inner remnant in which the fusion process begins to tail off and the star is left as a small white dwarf.

For heavier stars of a few times  $M_{\odot}$  or greater, the greater attractive gravitational forces cause a higher rate of fusion reactions leading to much shorter-lived stars. Heavy stars burn brighter and burn out in much shorter lifetimes of, typically, hundreds of millions of years. The fusion reactions do not stop at helium but continue up the periodic table producing in turn nitrogen, oxygen, carbon and up through the table

until creating the most stable of nuclei, iron (recall Section 5.4). At this point, fusion beyond iron begins to consume energy and the iron core begins to contract. Eventually the core collapses in a violent explosion and blows off a huge shell of the star's outer layers, an event we observe as a supernova, the radiation from which can temporarily outshine the radiation output of an entire galaxy.

Supernovae are crucial cosmic events for a number of reasons. Firstly, the production of trans-ferric nuclei in the star's core such as those of tungsten, gold, lead and uranium and their expulsion in the explosion is the main source of heavy elements in the Universe. Secondly, the explosion involves not only an intense burst of EM radiation, but of neutrinos also, a fact we will investigate further in Chapter 55. Thirdly, supernovae act for us as beacons in the cosmos. It was the plotting of thousands of supernovae remnants in distant galaxies that led to the relatively recent discovery of the acceleration of the expansion of the Universe by S. Perlmutter, B. Schmidt and A. Reiss for which they were awarded the Nobel Prize in 2011.

To continue to the endgame for massive stars, what happens after the explosion depends once more on mass. For less massive cores below some critical value, the still intense pressure of gravitation leads to the capture of electrons by the protons of the iron nuclei resulting in the entire core becoming a single ultra-dense mass of neutrons. This neutron star (NS) is effectively one huge, bizarre atomic nucleus with a typical radius of a few kilometres.

Neutron stars typically rotate at a rate between a few hundred times a second up to a few seconds emitting radiation along the axis of their magnetic poles. When this direction is pointing towards Earth, we can detect the very regular pulses which we categorise as pulsars, the discovery of which in 1959 led to the Nobel Prize of 1974 for British astronomer Anthony Hewish (shared with Martin Ryle for his development of the technique of aperture synthesis), but not, infamously, for co-worker Jocelyn Bell Burnell.

Above the critical mass referred to above, the gravitational force becomes so enormously strong that even neutrons cannot survive and the structure collapses to a point in space-time of infinite energy density, a black hole (BH). The critical mass at which this occurs is still uncertain but is currently estimated at between  $1.4\text{--}3.0 \times M_{\odot}$ , the larger part of the

initially much higher-mass star having been blown off in the explosion.

Despite its initially highly speculative nature, black hole physics has become a huge field of study led by some of the most famous names in physics (e.g. Oppenheimer, Wheeler, Thorne, Hawking, Penrose *et al.*) and generating thousands of articles and books, scholarly or otherwise. Apart from the other-worldly, sci-fi fascination, a sounder justification is that BH physics is, along with the Big Bang itself, the only field for the study of gravity in the quantum regime, the crucial next step in our route to the Theory of Everything.

To end our brutal over-simplification with yet another, there are several key facts of BHs for our exposition.

Initially it was believed nothing, none of matter, light nor information could escape from the gravitational pull of a BH.

Secondly, BHs (specifically, non-rotating BHs) are surrounded by a boundary, an event horizon, called the Schwarzschild radius, across which everything will fall into the BH without possibility of escape. The size of the radius is given by the formula  $R_s = 2GM/c^2$ .

Thirdly, a BH can be categorised by only a very few parameters: its mass, spin and charge. This is known as the No Hair Theorem.

Matter falling into a BH will increase mass and so the surface area. This surface area is taken as the entropy of the BH in that, classically, it can only increase.

There are thought to be supermassive BHs, greater than, say,  $10^3 \times M_\odot$  at the centres of most galaxies around which the observable stars will orbit and into which they may eventually be absorbed.

It is possible there may be primordial BHs originating from the time of the Big Bang which may still be wandering freely across the Universe, or may have merged with other BHs and conceivably led and still lead to galactic formation around themselves.

It is conceivable that micro-BHs could be produced in terrestrial accelerators such as the LHC but their signatures have not yet been satisfactorily explained.

Having mentioned these key facts, there is one putative phenomenon which would seem to contradict many of them. This is the phenomenon of Hawking Radiation (HR) proposed by the highly celebrated, British physicist, Stephen Hawking. HR posits that the random appearance of a pair of virtual particles on the event horizon of a BH may see one such particle sucked into the BH with negative energy, leaving its previously virtual antiparticle on the other side of the horizon as a real particle to escape with positive energy. The effect of this would be for the in-falling negative energy particle to reduce the mass of the BH while the escaping, positive energy particle would carry the same amount of energy into the wider Universe. Hence BHs will act as cosmic torches, radiating particles and shrinking in size. As it happens, as the BH shrinks, the rate of HR increases leading to an effective explosive evaporation of the BH!

This proposed phenomenon of HR is one of the very few conceivable instances of gravity acting in a quantum regime. So far it has not been observed as, yet again, a clear signature is not obvious. But its unambiguous detection would be a very major step forward (and, little doubt, another Nobel Prize)!

But amongst more exotic observations we can now observe with advanced telescopes are stars in distant galaxies dancing around the very massive black holes in their galactic centres. However, looking for the most dramatic, hence energetic events by which we might observe GWs here on Earth, we arrive at the idea of colliding BHs or NSs. In fact, a straightforward collision as between the head-on collisions of billiard balls on a table would be highly unlikely. More plausible is a pair of BHs or NSs or a BH plus NS binary spiralling around each other in a decaying orbit until they finally coalesce.

## *The Discovery of Gravitational Waves*

### 54.1 Introduction

One of the key developments in the modern discovery of the waves was when the American physicist Rainer Weiss became motivated to pursue the search for the waves following the contentious, but ultimately false, claims of Joseph Weber using his laboratory-based resonant mass experiment.

In 1972 Weiss suggested that a more promising method to detect the minute strains expected from the waves would be to search for interference patterns between beams of light rather than the monitoring of solid matter. A meeting with the distinguished gravitational theorist Kip Thorne in 1975 led, over a decade later, to the formulation of the 1987 proposal by CALTECH and MIT for the construction of a Laser Interferometer GW Observatory (LIGO). Construction began in the US at sites in Hanford, Washington State and at Livingston, Louisiana in 1994 led by Barry Barrish as principal investigator.

The search for the waves had grown from a single laboratory experiment to a multi-billion dollar undertaking in the same league as the CERN laboratory eventually involving over 1000 scientists and scores of laboratories across the world.

### 54.2 LIGO

The basic idea is to detect the variations in length between a light source and a fixed test mass due to the passage of the waves. This is done by arranging for two laser beams from the same source

traversing otherwise identical perpendicular distances, bouncing the beams off massive and extremely sophisticated mirrors acting as the test masses, and observing the interference patterns of the recombined beams, see Figure 54.1. If GWs traverse the system, the interference pattern will change to reveal the minute variations of the length of one of the beams relative to the other. This variation is measured as the strain in space defined as the variation in the length divided by the length itself:

$$\text{Strain} = dL/L.$$

As an indication of the extreme sensitivity required, the order of magnitude of the strain expected from a passing GW is of order of one part in  $10^{-21}$ , equivalent to the width of a human hair variation in the distance between the Earth and our nearest star alpha centauri, some 4.4 light years distant. To achieve the necessary sensitivity required state of art technology for the suspension and coatings of the reflecting mirrors, the vacuum in the arms (maintained since construction) and the laser and display technology.

### 54.3 The Detection of GW150914

LIGO began operating in 2002 and operated until 2010 with no positive results in wave detection although providing much information on the astrophysical limits of likely GW sources. Also during this time LIGO established a collaboration with VIRGO, a French-Italian-led project operating a similar



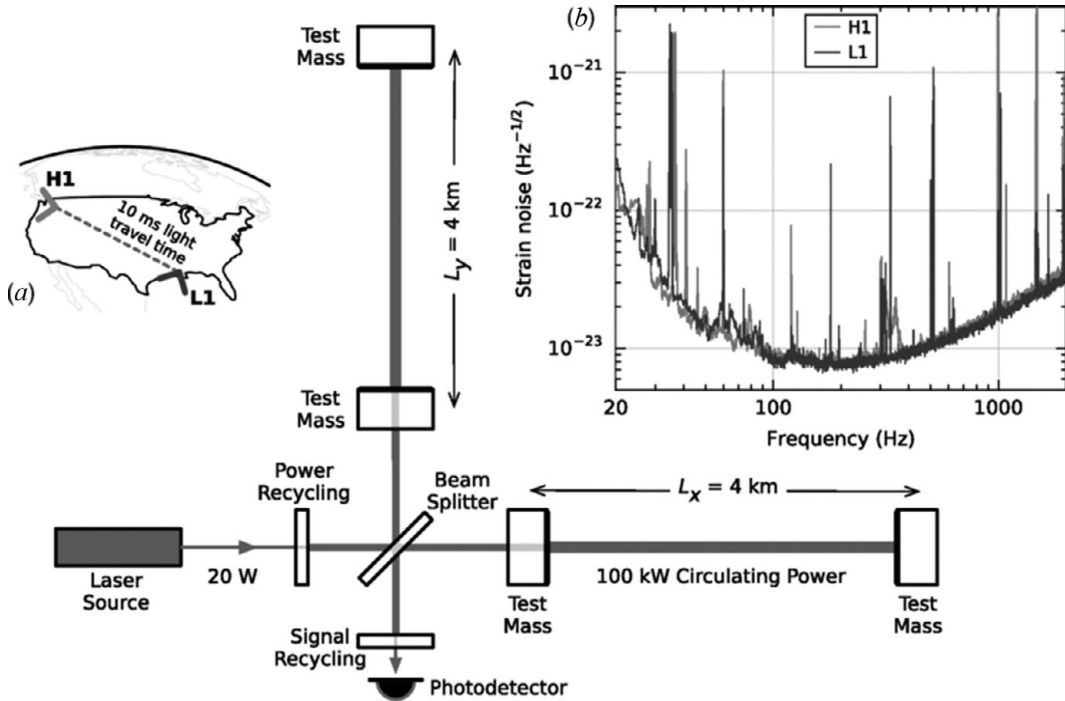


Figure 54.1. Simplified Advanced LIGO detector (not to scale). A gravitational wave propagating orthogonally to the detector plane and linearly polarised parallel to the 4 km optical cavities will have the effect of lengthening one 4 km arm and shortening the other during one half-cycle of the wave; these length changes are reversed during the other half-cycle. The output photodetector records these differential cavity length variations. While a detector’s directional response is maximal for this case, it is still significant for most other angles of incidence or polarisations (gravitational waves propagate freely through the Earth). Inset (a): Location and orientation of the LIGO detectors at Hanford, WA (Figure 54.2) (H1) and Livingston, LA (L1). Inset (b): The instrument noise for each detector near the time of the signal detection; this is an amplitude spectral density, expressed in terms of equivalent gravitational-wave strain amplitude. *Source: LIGO.*

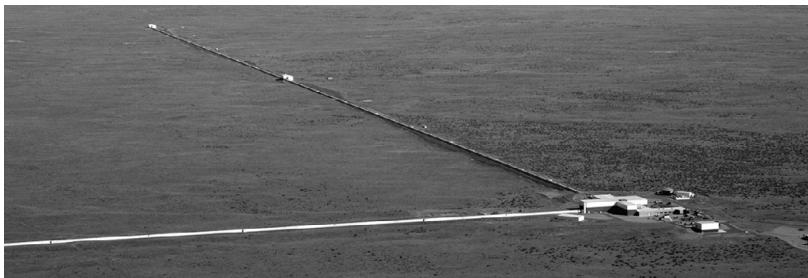


Figure 54.2. LIGO detector, Hanford, WA site.

interferometer device located at Cascina in Italy. As a result of the first null run, an improved Advanced aLIGO was approved in 2008, involving the upgrading of many of the crucial elements to improve sensitivity. aLIGO commenced operations in early 2015 and almost immediately hit the jackpot! It was during

the engineering run ahead of planned full operation that a clear, statistically very significant signal on 14 September 2015 was seen at both the Hanford and Livingston detectors. The signal was duly named GW150914 (i.e. the date, 14 September 2015), see Figure 54.3.

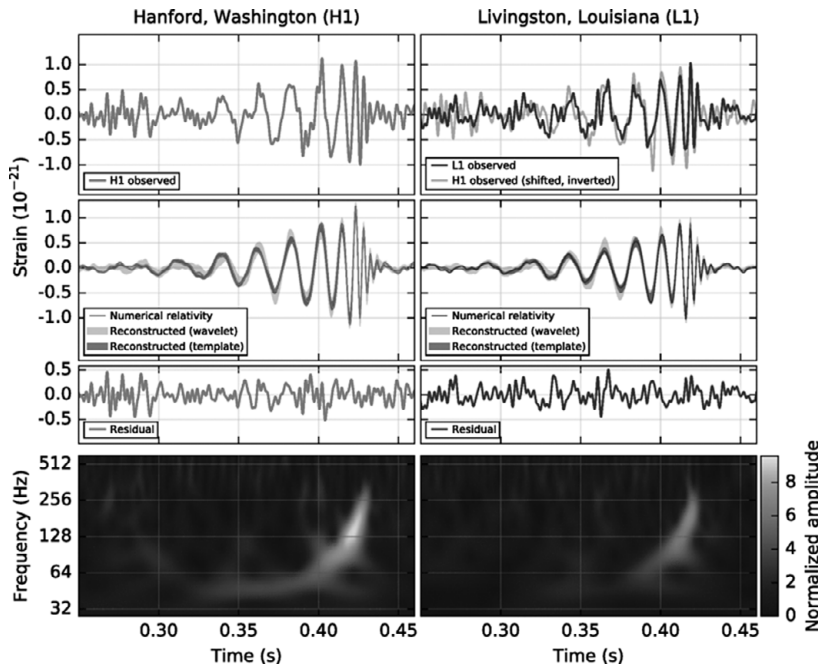


Figure 54.3. The GW event GW150914 observed by the LIGO Hanford (left column panels) and Livingston (right-hand panels) detectors. Second row: GW strain projected onto each detector confirmed to 99.9% by independent calculation. Third row: Residual noise after subtracting the waveform. Bottom row: A time-frequency representation of the strain data. *Credit: LIGO.*

The signal is identified by comparing the interference pattern observed with a database of hundreds of thousands of such patterns generated by the machine intensive calculations of numerical general relativity – each pattern representing the merger of black holes with different masses and spins and a range of other parameters.

Two human observations are perhaps worth noting to celebrate this 100th anniversary prize for Einstein’s marvellous theory. First is that the frequency of the waves measured in the hundreds of hertz range can be represented audibly by sound as well as graphically. (The result can be found on LIGO Lab video: *The Sound of Two Black Holes Colliding.*) This sounds for all the world like the chirp of a songbird (perhaps rather a tired one having travelled s.b. 1.3 billion light years in a vacuum!).

A second observation is that, given the rigidity of space–time and the duration of the signal, a better analogy for the waves, rather than ripples on the surface of a pond, might be the sounding of a huge cosmic gong!

The results of the parameterisation of the signal are summarised in Table 54.1.

Table 54.1. *Important parameters for GW150914.*

Time detected	14 September 2015 09:50:45 UTC	
Mass (in units of solar mass)	black hole 1	$36^{+5}_{-4}$
	black hole 2	$29 \pm 4$
	final mass	$62 \pm 4$
GW energy	$3.0 \pm 0.5 M_{\odot}c^2$	
Distance	$410^{+160}_{-180}$ Mpc $\sim 1.34 \times 10^9$ light years	
Redshift	$410^{+0.03}_{-0.04}$	
Observing band	35–350 Hz	
Peak strain $h$	$1.0 \times 10^{-21}$	

The numerical profile of the table does scant justice to the truly cosmic nature of the happening. The two black holes, one 29 times, the other 36 times the mass of our Sun, approach each other in a terminal spiral into coalescence, see Figure 54.4. In the final seconds, the two objects are orbiting each other at a rate of hundreds of hertz (i.e. orbits per second)

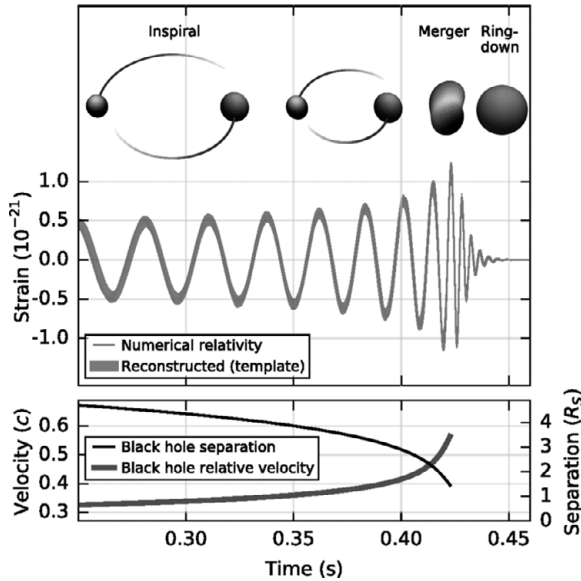


Figure 54.4. GW150914 models. Credit: LIGO.

thereby travelling at an appreciable fraction of the speed of light. Each of the black holes of these masses will have a Schwarzschild radius of c. 100 km which will be the effective radius of their mutual orbit around their centre of gravity.

On coalescence, the resulting mass of the merged super black hole of 62 times  $M_{\odot}$  implies an energy radiated of three times  $M_{\star}$ . This vast amount of energy then propagated out spherically for 1.34 billion light years until being detected here on Earth. Using Einstein’s famous mass–energy equivalence formula (2.5) gives an energy density at the Earth’s surface variously estimated as equivalent to that of a full moon illumination of the Earth, or of a mobile phone emission one metre from the human ear!

The significance of GW150914 can hardly be overstated. Simply confirming the existence of GWs would be significant enough. But the event was also the first incontrovertible evidence for the existence of BHs themselves and also the first direct evidence for binary systems of BHs. In short, it took the existence of BHs from inferred existence to direct, measurable reality.

#### 54.4 Subsequent Events

Having searched and waited for such a long period for the first observation, gravity waves went on to exhibit a behaviour similar to London buses: three appear more or less all at once.

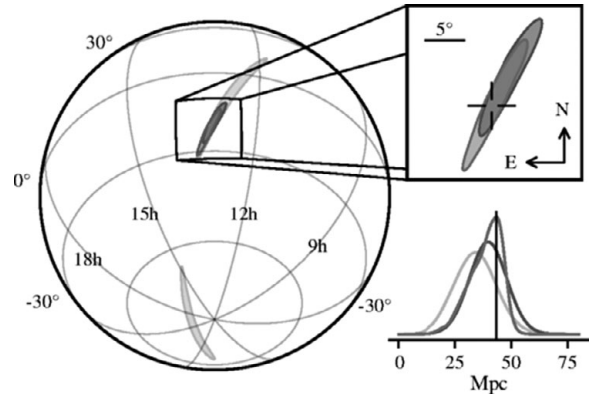


Figure 54.5. Sky location reconstructed for GW170817 by a rapid localisation algorithm from a Hanford–Livingston (190 deg<sup>2</sup>, lighter contours) and Hanford–Livingston–Virgo (31 deg<sup>2</sup>, darker contours) analysis. In the top-right inset panel, the reticle marks the position of the apparent host galaxy NGC 4993. The bottom-right panel shows the a posteriori luminosity distance distribution from the three gravitational-wave localisation analyses. The distance of NGC 4993, assuming the redshift from the NASA/IPAC Extragalactic Database and standard cosmological parameters, is shown with a vertical line. *Source: LIGO.*

GW 151226 (colloquially known as the Boxing Day event) occurred towards the end of observation run 1. This involved a similar BBH (Binary Black Hole) coalescence of  $14M_{\odot} + 7M_{\odot}$  at a distance of 1.45 billion light years or 440 Mpc.

GW 170104 occurred just after the start of observation run 2 and was another BBH coalescence of  $M_{\odot}31+M_{\odot}19$  at the extremely remote distance of 2.9 billion light years or 880 Mpc.

GW 170608 was the lightest and closest BBH coalescence of  $12M_{\odot} + 7M_{\odot}$  much closer at 1.1 billion light years.

Finally, GW 170814 was noteworthy in that the BBH coalescence of  $30M_{\odot} + 25M_{\odot}$  at a distance of 1.8 billion light years was observed not only by the two LIGO observatories in the US but also by the advanced VIRGO observatory in Italy. VIRGO had largely tracked the LIGO project but trailing by some two years in making this, its first GW observation. Crucially, to have a third observation point at a considerable distance (at least in terrestrial terms) allows for the triangulation and hence the much more accurate localisation of the event in the cosmic sphere (see Figure 54.5, which shows the slightly later event GW170817), a fact which will take on increasing significance in the next section and in the years ahead.

## *Gravitational-wave and Multi-messenger Astronomy*

### **55.1 Introduction**

Until recent decades, our entire knowledge of the Universe was based on telescopes receiving ordinary visible light. This allowed determination of stellar and planetary orbits and galactic dynamics using Newton's theory of gravity, the observations confirming Einstein's general theory of relativity and Hubble's observation of the expansion of the Universe.

In the 1950s, the advent of radio astronomy saw the discovery of pulsars confirming the reality of neutron stars. Over the subsequent decades, the whole spectrum of EM radiation has been monitored by an increasing array of terrestrial telescopes and satellite detectors allowing us to build a comprehensive understanding of stellar, galactic and cosmological phenomena as far back as the time of last scattering (i.e. time of formation of the CMB some 300 000 years after the Big Bang) prior to which the Universe was opaque. (Earlier than this the photons could not escape the hot, dense quark–gluon plasma.)

This monopoly of EM radiation was broken rather accidentally in 1987 on the occasion of the supernova SN1987a. A typically huge supernova explosion, the event saw not only the huge increase in the visibility of the exploding star but also the emission of a burst of neutrinos, 25 of which were detected in a variety of detectors on Earth. The opportunity

to cross check the implications of the EM radiation with the parameters of the arriving neutrinos allowed for a considerable advance in our understanding of supernova dynamics. (It is sometimes jokingly observed that each neutrino resulted in the publication of over a thousand research articles!)

The significance of the detection of GWs is that they will now join EM radiation and neutrino fluxes as observation tools to provide complementary signals from the Universe. As all three traverse the Universe at the speed of light they can provide the evidence from the three separate forces of electromagnetism, the weak nuclear force and gravity of the same events. Such a coincidence of signals from the different forces is now referred to as multi-messenger astronomy. Only the strong force cannot join the fun. Although its massive material particles (fermions) arrive in the upper atmosphere as cosmic rays, they cannot be identified with any distant cosmic events due to their time of travel.

### **55.2 Gravitational-wave Astronomy**

Although relatively early days, the planned increasing sensitivity of the aLIGO and other detectors will increase the range, hence the observable volume, of the Universe amenable to observation, hence increase the rate and variety of event detection. Thus far, all the events have occurred well within

the distance of a gigaparsec ( $3.3 \times 10^9$  light years). Current expectations are for an event rate of c. 12–200 per year per cubic gigaparsec. So, the events observed thus far would be almost compatible with the lower end of this range. Increases in sensitivity of the aLIGO detector expected in the near term may double the accessible distance, thus increase the event rate by a factor of eight times. In the medium term an increase in sensitivity of a further factor of five may take the observable event rate into the range of 12 000–200 000 per year.

Such a large event rate will allow an intensive study of BHs in the cosmos, from which a great deal may be learnt. For example, it is possible that primordial BHs could account for a significant proportion of the dark matter in the Universe, so we may be able to learn more about dark matter in this way.

Another area where GW astronomy will be of use is the area of BH dynamics. The shapes of the waveforms of the GWs detected are sensitive to key BH parameters, the leading of which is the angular momentum. So observation of the shape of the waveforms will allow study of the rotational effects on BH dynamics. Next, as we shall soon see, in addition to the BH coalescences observed to date, we have already observed the GWs from two colliding neutron stars (see following section). In future we can reasonably expect to see BH/NS coalescences and other GW-emitting phenomena.

One example of the latter is the possibility of using GWs to look beyond the time of last scattering, which gave rise to the CMB. One such idea is that the putative theory of inflation described in Chapter 48 may have given rise to the emission of GWs which would then make their presence felt by the polarisation of the EM CMB radiation. Indeed, some claims have already been made of detecting this effect but remain currently unconfirmed. Whether primordial GWs resulting from the Big Bang itself may be detectable remains an open question. But in principle there would be every reason to think the Universe is bathed in primordial GW emission in the same way it is by the later CMB radiation.

Another even more exotic possibility is to look back in time, possibly before the Big Bang itself. A theory by the name of conformal cyclic cosmology proposed by Roger Penrose posits that our Universe

was preceded by earlier ones which then collapsed into what became our Big Bang. Penrose and others propose that such an evolution should be detectable by patterns in the CMB reflecting the demise of BHs in the previous eon. Whether GWs resulting from that previous time may in principle be detectable after tunnelling though the bang remains an open – and highly speculative – question!

### 55.3 GW170817 – a Binary Neutron Star Merger

The most recent of the GW events observed thus far involved the merger of two neutron stars of  $2.3 M_{\odot} + 0.9 M_{\odot}$  at a distance of just 132 million light years (40 Mpc). By far the closest of the GW events to Earth involving the lightest objects, well below the threshold of BH formation, it was thus identified as the coalescence of a neutron star binary (NSB). In fact, the event was first noticed as a burst of gamma rays by NASA's Fermi gamma-ray observatory. As the gamma rays arrived after the associated GWs, communications between the Fermi satellite and the two aLIGO detectors warned of the likely occurrence of a GW event some 2 seconds before the arrival of the gammas. This same communications system alerted a network of dozens of other observatories covering the whole range of the EM spectrum which detected, and are still detecting, the event and its aftermath. The GWs are the first to be emitted followed closely by the gammas. A neutrino burst might have been expected between the two but was not detected, possibly owing simply to the fact that they were emitted in a direction away from Earth. Subsequently, EM of increasing wavelengths arrives at increasingly delayed times due to the origin and then the passage of the radiation though the detritus of the merger.

The merger of two NSs is very different from that between two BHs. When the two NSs merge, the resulting NS is likely to be above the mass at which it must collapse into a BH, currently estimated at about  $2.2 M_{\odot}$ . Although most of the masses of the NSs go into the BH, some will be scattered into what forms as an accretion disc around it. As this disc is then sucked into the BH, a jet of matter is shot out along an axis perpendicular to the disc, including an array of heavy atoms that will have formed in the neutron plasma of the NSs. (This is now thought to be the

dominant mechanism of heavy element production in the Universe ahead of the mechanism of supernovae as previously discussed.) As this jet travels through the interstellar gas surrounding the BH it emits first gamma rays, then X-rays and so down through the spectrum of UV, visible, infrared and eventually radio EM radiation, all of which are detectable by

the various observatories focusing on their respective parts of the spectrum.

Having thus surveyed the excitements of the first GW detections, and the first dual messenger events in the instances of SN1987a and GW 170817, the world keenly awaits the arrival of the first tri-messenger event, necessarily either NS + NS or NS + BH.

## *The Future: Super LIGO and LISA*

The planned improvements in existing detector sensitivities together with the arrival of new terrestrial observatories such as the KAGRA detector in Japan at Karioka and a possible aLIGO detector in India, both due by 2020 and collectively dubbed Super LIGO, will greatly improve both the event rates as described earlier and the triangulation capabilities to enable all three of GW, EM and, hopefully, neutrino observations of the same events.

But in the medium to longer term the real game-changer will be the European Space Agency's (ESA) Laser Interferometer Space Antenna (LISA), to be executed with NASA and quite probably a variety of other agencies worldwide. The basic idea is to launch three satellites into a helio-centric orbit some 50 million km from Earth. This position is the so-called Laplace point at which the gravitational forces of the Sun and the Earth are in balance, allowing the observatory to conduct its own orbit around the Sun but in a fixed position trailing behind the Earth. The observatory is to consist of three arms each of 2.5 million km with six laser links between the three identical spacecraft in a triangular formation, see Figure 56.1. Each of the spacecraft is to carry two test-mass mirrors in free-fall each of which will act as the endpoint to two of the interferometer arms. The spacecraft themselves are simply containers for the orbiting test-mass mirrors,

their positions relative to the mirrors being controlled electrostatically.

As ambitious as this apparatus looks, its main technology elements have already been tested in the LISA Pathfinder mission between 2015 and 2017 during which the essential elements of laser operation, mirror test-mass positioning and interferometer optics were all tested with results reported to be superior to those specified for the full LISA system.

LISA will allow study of effectively the entire Universe with GWs. First it will study the numerous population of compact binary stars in the Milky Way including white dwarf, NS and BH binaries of which up to 25 000 are expected. Next will be the ability to investigate the origin, growth and mergers of massive BHs in galactic centres which may weigh in at exceptional masses of up to  $10^5 M_{\odot}$ . The GW signals from such BHs may last for months during the inspiral leading up to the merger giving LISA ample time to alert EM observatories to pinpoint the same events for multi-messenger observation. The plethora of observations will allow LISA also to determine more accurately cosmic expansion parameters such as the Hubble constant by providing another metric alongside current measurements derived from supernova and CMB observations. Also, by looking back inside the CMB, LISA will probe the multi TeV universe existing in the instants following

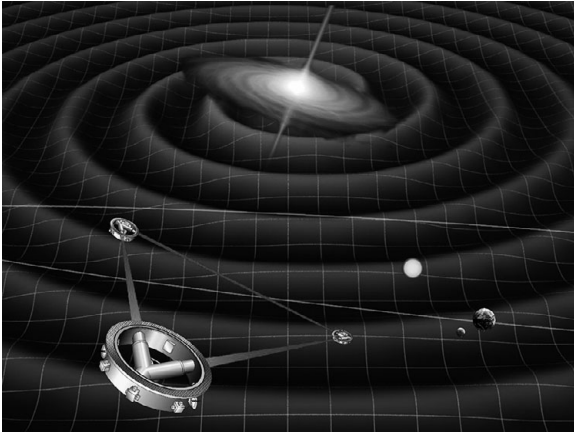


Figure 56.1. Laser Interferometer Space Antenna.

Big Bang. Finally, of great interest, would be any real surprises such as previously unknown sources of either GWs or gamma-ray bursts, or signs of stringiness in the cosmos.

As the mission is expected to launch in 2034, there should be plenty of time for future editions of this book and others to highlight these and other prospects!





**Part XIV**  
**String Theory**



## *Origins – the Hadronic String*

### 57.1 The Success of QFT

In the foregoing chapters, we have seen how the framework of quantum field theory in general, and the gauge theories of the Standard Model in particular, enable us to describe all of the diverse phenomena observed in particle physics thus far. What is more, we saw in Chapters 42–47 how we can use the language of quantum field theory to go beyond the Standard Model and build theories, such as those with SUSY, Higgs compositeness, and unification of the strong and weak nuclear forces and electromagnetism.

It would seem, then, that QFT can do everything that we ask of it. There is, however, one force that we have not yet discussed at all in the context of QFT – gravity.

### 57.2 The Problem of Gravity

Although, *we* have not discussed it in any detail yet, gravity was actually the first force to be ‘understood’, according to Newton’s law of universal gravitation (superseded by Einstein’s theory of General Relativity as we saw in Chapter 5). As we remarked then, gravity is a rather weak force – the gravitational force between two protons, say, is  $10^{36}$  times weaker than the electrostatic force between them. Indeed, for all laboratory particle physics experiments, gravitational forces are utterly negligible. The reason that gravity is so important in everyday life is that gravity is always attractive and so the forces between particles are cumulative. Thus,

while the gravitational force between two protons is tiny, the gravitational force between an apple and the Earth (both of which contain very many protons) is rather large. In contrast, electric and other forces can be attractive or repulsive, and so the forces between the individual particles in a large body tend to cancel.

General Relativity is a classical theory of gravity, in that it takes no account of the principles of quantum mechanics described in Chapter 3. Because gravity is only significant on macroscopic scales, where quantum mechanics is unimportant, a classical theory is perfectly adequate for describing all the gravitational phenomena we observe – the falling of apples off trees, the orbits of planets, and so on. However, our observations on microscopic scales (in atomic and sub-atomic physics) tell us that the world is really quantum-mechanical and so a classical theory of gravity won’t do. We need a quantum theory of gravity!

One might wonder whether this *really* is the case. After all, the theory of gravity (General Relativity) works perfectly well on large scales when QM is negligible, and the quantum-mechanical theories of the other forces work perfectly well on small scales where gravity is negligible, so why can’t they coexist happily?

The answer is that there are situations where neither gravity nor quantum-mechanical effects are negligible. These situations occur whenever one is dealing with high energies. The gravitational coupling constant is, like the other couplings in nature, a

running coupling ‘constant’. However, unlike the other running couplings, which change very slowly with the energy (logarithmically in fact, e.g.  $g \propto \log E$ ), the gravitational coupling grows as the square of the energy. So at short distances, corresponding to high energies, the coupling of gravity can become very large.

The reason we do not see this strong coupling of gravity is that we cannot yet probe such energies in the laboratory – the scale at which gravity becomes strong is about  $10^{19}$  GeV, the so-called Planck scale.<sup>1</sup> However, we know of at least two situations where such enormous energies occur – in the interior of black holes and in the early Universe. In such situations, our classical theory of gravity *must* break down.

So a quantum theory of gravity really is mandatory. Since we were able to construct quantum field theories of all the other fundamental forces, the reader may wonder why it is so difficult to make a quantum theory of gravity. The answer is that it is precisely because the gravitational coupling becomes strong at high energies. In quantising the other interactions of nature, we saw that infinities were encountered due to virtual processes occurring at high energy. We saw though, that by a redefinition of the physical parameters of the theory (the renormalisation idea), we were able to absorb (or rather hide) these infinities, producing a finite and sensible theory. In gravity, because the coupling becomes so strong at the high energies where the infinities occur, there are simply too many infinities to absorb. There is no carpet big enough to sweep them under! The upshot is that, while we can make sense of gravity as a quantum theory (a so-called *effective field theory*) up to Planck-scale energies, such a theory inevitably cannot describe physics at or above the Planck scale.

It appears then that the problem of constructing a complete quantum theory of gravity is insurmountable within the framework of QFT, and that QFT cannot be the be-all-and-end-all. We need a radically different kind of theory. The theory we shall discuss in this chapter – String Theory – provides an elegant resolution of the problem of quantum gravity, and is appealing on many other grounds besides. In particular, it

offers the hope of unifying *all* of the ideas we have discussed in this book, and is, as such, the best (and only) candidate we have for a Theory of Everything at the present time.

### 57.3 Strings versus Particles

The fundamental entities in the Standard Model (and all quantum field theories) are point particles: quarks, leptons, gauge bosons and so on. The starting point of string theory is to replace point particles by one-dimensional objects, called strings, as the fundamental entities of nature. This modification results in a theory with completely different properties, as we shall see below, but which reproduces QFT at low energies. The fundamental strings can either form closed loops, called closed strings, or can have their ends free, called open strings. In QFT, a point particle traces out a line called its *worldline* as it moves through space–time. The Feynman diagrams of Chapter 4 represent the interacting worldlines of point particles. The amplitude for a physical scattering process is found by summing over all the Feynman diagrams contributing to that process. Similarly, in string theory, a moving string traces out a surface or *worldsheet* in space–time. Worldsheets can split apart or join up, and such processes constitute the interactions between strings. The quantum-mechanical amplitude for a string scattering process is found by summing over all worldsheets contributing to that process.

Already, we can see two very important things about string theory. The first important thing is that, if we look on distance scales much greater than the characteristic size of a string (i.e. at low energies), we will not be able to see that strings are strings. They just look like point particles, and the worldsheet diagrams look like Feynman diagrams. So we see intuitively that string theory looks like QFT at low energies. Moreover, as we see in Figure 57.1, a single worldsheet diagram actually incorporates several Feynman diagrams. The different Feynman diagrams are just different limits of the worldsheet diagram, in which the worldsheet diagram is stretched out in different ways. In this sense, string theory is already a simplification of QFT. In fact, it will turn out to be much more of a simplification, producing all the observed particles and forces from a single type of string. The second important thing concerns the infinities of QFT. What happens to the problem of short-distance divergences

<sup>1</sup> In a sense, the Planck energy is not very large at all – it corresponds to the kinetic energy of a slug moving on a lettuce leaf! However, in the case of a slug, this energy is not concentrated on a single fundamental particle.

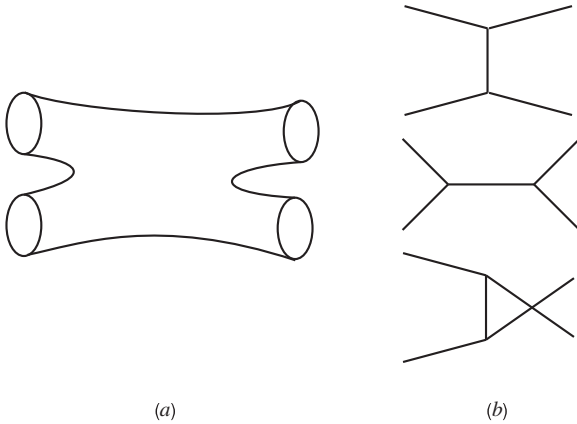


Figure 57.1. (a) Shows a worldsheet diagram contributing to the scattering of four closed strings. By stretching the diagram in various directions, one can see that it incorporates all of the Feynman diagrams in (b).

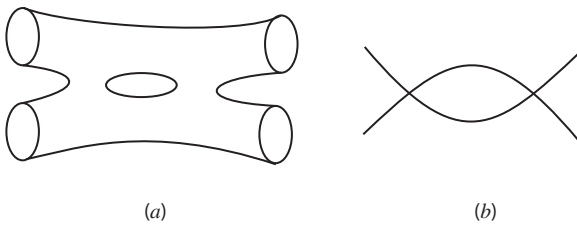


Figure 57.2. Loop diagrams in string theory (a) and quantum field theory (b). In the string diagram, there is no notion of interaction points coinciding.

in QFT in string theory? In QFT, these divergences come from Feynman diagrams with loops of virtual particles, and occur when two or more interaction points coincide (such that the exchanged energy and momentum, which is inversely proportional to the separation, become very large). If we look at the corresponding worldsheet diagram (Figure 57.2), we see that it is no longer possible to say when the interaction ‘points’ coincide, because the interactions, like the strings themselves, are smeared out. There is no longer any notion of a point in space–time at which an interaction occurs. Since the notion of interaction points is also lost, it is just possible that some or all of the infinities of QFT will disappear in string theory. Remarkably, this is indeed the case! There are supersymmetric versions of string theory – superstring theory – with no infinities whatsoever. Strings

have another very important property which point particles do not. Whereas point particles can, by definition, have no internal structure, strings, which are extended objects, can. In particular, they can vibrate. Just as a plucked guitar or violin string vibrates at an infinite number of frequencies (the fundamental frequency and harmonics), so a fundamental string has infinitely many vibrational modes, corresponding to the different allowed frequencies. The different modes manifest themselves in nature as the different particles we see. Miraculously, we will see that string theory can generate all of the different particles of the Standard Model: bosons and fermions, leptons and quarks, the whole lot.

What is more, because there is only one sort of string in string theory, there is only (at most) one free parameter – the coupling constant. From this, one should be able, if string theory is correct, to predict all of the coupling constants, masses and mixing parameters in the Standard Model!

### 57.4 The Hadronic String

While string theory now tilts at being a Theory of Everything, its origins in the 1960s were rather more humble, coming from attempts to describe hadrons and the strong nuclear force. While we now know that the hadrons are described by a quantum field theory, QCD, it seemed in those early days that this could not be the case. For example, while the use of perturbation theory in quantum electrodynamics had yielded a successful calculation of the anomalous magnetic moment of the electron (as we saw in Chapter 4), the magnetic moments of nucleons were known to deviate significantly from 2, making a perturbative calculation in any quantum field theory out of the question. Moreover, unlike in quantum electrodynamics where we find a small number of particles (e.g. the electron and the muon) which appear to behave like fundamental point-like particles, in the strongly-interacting sector hundred of hadrons had been found, which were clearly composite, extended objects.

In retrospect, it seems natural to describe a plethora of extended objects as modes of vibrating strings, but in fact such a description was only stumbled upon via a circuitous route. To begin with, physicists tried to describe hadrons using only basic quantum-mechanical constraints, such as the insistence that probabilities always sum to one and that

cause precedes effect, together with inputs from data, such as the 1961 observation of Geoffrey Chew and Steven Frautschi that a plot of the spins of different mesons versus their masses squared follows a straight line. Eventually, in 1968, Gabriele Veneziano came up with a model that described this behaviour and, a year or two later, Yoishiro Nambu, Holger Nielsen and Leonard Susskind showed that it described vibrating strings. But the theory fell foul of many other observations and, with the arrival of QCD and

asymptotic freedom in the early 1970s, was soon forgotten as a theory of hadrons, only to later resurface as a theory of quantum gravity.

Ironically, there is a precise sense in which a certain string theory really *is* a theory of hadrons, albeit a supersymmetric one with an infinite number of colours. We will return to this fascinating subject when we discuss the AdS-CFT correspondence in Chapter 59. First, we turn to the emergence of string theory as a possible Theory of Everything.

## *String Theory to M-theory*

### 58.1 The Search for a Consistent String Theory

The notion that string theory could be theory of quantum gravity began in 1974, when Joel Scherk and John Schwarz realised that the theory contained a boson of spin two, whose properties closely resembled those of the graviton, the quantum excitation of gravity. Even then, the theory was beset by problems, which theoretical physicists worked heroically to overcome. Firstly, it was found that a consistent quantum theory of strings could only exist in either twenty-six or ten space–time dimensions, not the four which we observe. Secondly, it was found that many string theories contained tachyons, particles which move faster than light.<sup>1</sup> This obstacle was surmounted by the discovery that certain supersymmetric versions of string theories, now called superstring theories, do not contain tachyons. To understand why this is this case, recall from Chapter 45 that supersymmetry is an enlargement of the space–time symmetry which corresponds to the transformations of Special Relativity. One consequence of Special Relativity is Einstein’s celebrated equation  $E = mc^2$  relating the energy and mass of a particle at rest. Now supersymmetry is an extension of Special Relativity and it has even stronger consequences for the energy. Supersymmetry dictates

that the energy of a particle should always be positive. This automatically excludes tachyons, which can carry negative amounts of energy. Thirdly, it was found that, even with these caveats, most string theories were anomalous (see Chapter 37 for a discussion of anomalies in quantum field theory), leading to the reappearance of infinities and the breakdown of the theory. However, in what became known as the ‘first string revolution’ in August 1984, Michael Green and John Schwarz showed that string theories could be free of anomalies, but only if the gauge symmetry group of the theory was one of two special groups, called  $SO(32)$  and  $E_8 \times E_8$ .<sup>2</sup> In a sense, these restrictions were rather satisfying. Theoretical physicists hoped that string theory might be the Theory of Everything, and if it was going to be *the* Theory of Everything, then there had better be only one of them! After all, how or why should one choose between two Theories of Everything? However, there was also a constant fear that there might turn out to be not one consistent string theory, but none at all.

By the early 1990s, five superstring theories had been found which passed all the consistency checks. It was not at all clear which, if any, might describe nature. All of them exist in ten dimensions, are supersymmetric and have a large gauge symmetry. They were given the names Type I, Type IIA, Type IIB,

<sup>1</sup> Such faster-than-light particles travel backwards in time, thus allowing the possibility of changing history. In such a theory it might even be possible to kill one’s grandmother before one’s own birth!

<sup>2</sup> At the time, a string theory with gauge symmetry group  $E_8 \times E_8$  was unknown, but one was later found.



Table 58.1. The five superstring theories and their properties.

Theory	Supersymmetries	Open/closed strings	Chiral?
Type I	1	Open and closed	Yes
Type IIA	2	Closed	No
Type IIB	2	Closed	Yes
Heterotic $SO(32)$	1	Closed	Yes
Heterotic $E_8 \times E_8$	1	Closed	Yes

Heterotic  $SO(32)$  and Heterotic  $E_8 \times E_8$ . In these exotic names, the roman numerals refer to the number of ten-dimensional space–time supersymmetries. There are two theories which have two supersymmetries (which really corresponds to an extended supersymmetry of the type discussed in Chapter 45) and these are distinguished (for no good reason) by the letters A and B. The heterotic theories are so-called because the waves that move to the left along the string are very different in character to those which move to the right, unlike on a violin string! Again, there are two possible theories, which are distinguished by their gauge group:  $SO(32)$  or  $E_8 \times E_8$ . Table 58.1 catalogues the properties of the five superstring theories, including the number of ten-dimensional supersymmetries and whether the theories are chiral, distinguishing left from right.

**58.2 String Theories Contain More Than String**

All that changed, however, in 1995, when the Fields Medallist<sup>3</sup> Edward Witten started what is now known as the *second* string revolution. In the years leading up to the second revolution, it was realised that strings could behave very differently when the coupling between strings (measuring the propensity of worldsheets to split and recombine) became large. Like quantum field theories, string theories are easiest to analyse when the coupling constant is small and interactions are infrequent. In this regime, the tools of perturbation theory can be employed (see Chapter 4). The five string theories described in the last section had *only* been studied in this way. In QFT, it was already well known that when coupling constants are large, quite different phenomena occur. The most striking example of this is the confinement of quarks into hadrons in QCD, which occurs at low energy when the coupling constant becomes large due to asymptotic freedom. Was it possible that similar unexpected things happened in string theories at strong coupling?

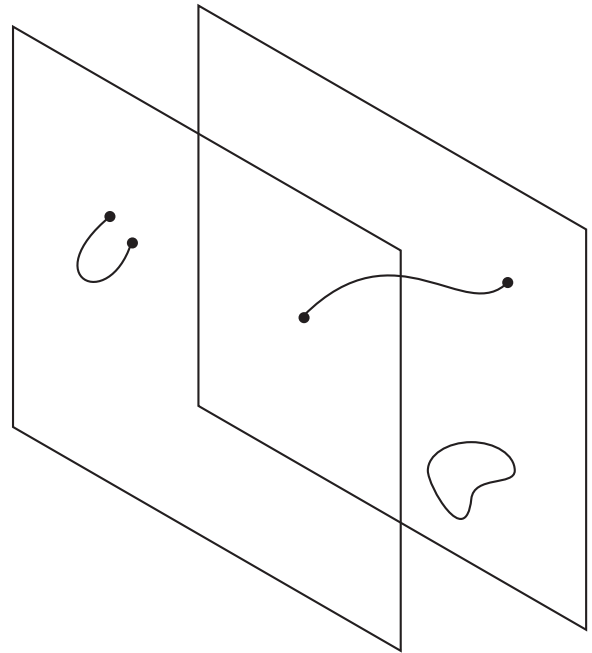


Figure 58.1. Open strings can have both ends terminating on the same brane or be stretched between two distinct branes. Closed strings can move freely throughout the bulk space–time.

The first unexpected realisation was made in the early 1990s by Joseph Polchinski, who found that there was more to string theory than just strings. Polchinski saw that string theory contained other, higher-dimensional extended objects, called membranes, or branes for short. In the perturbative regime (weak coupling) considered up until then in string theory, these membranes were very massive and therefore static. Their only role in the theory was to act as places where open strings could end, see Figure 58.1. Thus the open strings were always stretched between branes. However, when the string coupling increases, the masses of these branes decrease, and they become dynamical objects in their

<sup>3</sup> Mathematics’ equivalent of the Nobel Prize.

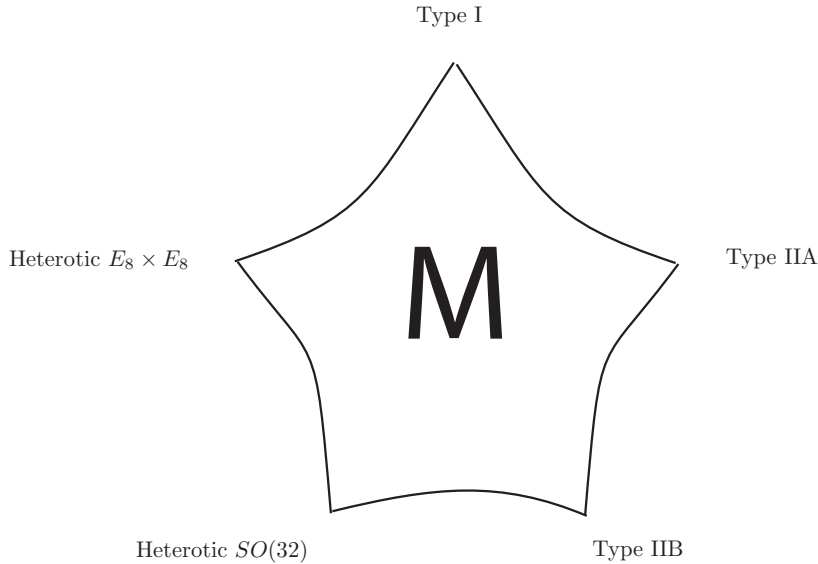


Figure 58.2. The five known superstring theories are the ‘corners’ of M-theory.

own right, just like the strings. So string theory is not just a theory of strings, but a theory of many extended objects. There then followed an even more unexpected realisation. The five superstring theories did indeed become very different as the coupling became strong. In fact, they turned into one another! That is, one of the theories with large coupling  $g$ , turned out to be equivalent, or *dual* to, another one (which could in fact be the same theory) with weak coupling  $g'$ . The couplings were related in the following simple way:

$$g \propto \frac{1}{g'}$$

For example, the Type I theory turned into the Heterotic  $SO(32)$  theory under this duality, while the Type IIB theory turned into itself. This type of weak–strong coupling duality became known as S-duality.

Two other sorts of duality were also found. One, called T-duality, related one theory compactified (see below) on a large space–time volume with another compactified on a small volume. The third duality, U-duality, was essentially a combination of S- and T-duality, mixing up both the couplings and the compactification volumes.

Many such dualities were found. All of the string theories were found to be dual to each other in some way, and, hence, somehow connected. These mysterious dualities prompted Witten to propose that the five string theories were not really fundamental at all, but that they too descended from some as

yet unknown theory in *eleven* dimensions, which he called *M-theory*.<sup>4</sup> Each of the five string theories with ten dimensions could, he claimed, be obtained from M-theory by curling up, or *compactifying* the extra eleventh dimension in a certain way. The string theories would then be recovered at low energies, below which the size of the extra dimension could not be resolved. For example, compactifying the eleventh dimension into a circle resulted in the Type IIA theory, whereas compactifying it into a line interval resulted in the Heterotic  $E_8 \times E_8$  theory.

The M-theory idea is certainly appealing, and there are several indications that it is correct, the best evidence being the observed string theory dualities. Unfortunately, we have no *real* idea at the present time of what M-theory actually is. We know that it exists in eleven dimensions, we know how the extra dimensions must be curled up to obtain the five string theories, and we know the nature of the five string theories themselves. In this way we have information about the five ‘corners’ of M-theory shown in Figure 58.2. However, we know very little besides. We do not even know what the fundamental degrees of freedom of M-theory are, let alone their dynamics. Are they particles, strings, branes or something completely different?

<sup>4</sup> Nobody really knows what Witten’s M stands for; Witten himself says ‘Magic, Mystery, or Matrix’, but others say it is an upside-down ‘W’!

## *The AdS–CFT Correspondence*

### 59.1 The Miracle of Duality

As we saw in Chapter 58, one of the miracles of string theory is that theories which appear to be superficially different in fact make identical predictions for all physical processes. The theories are thus regarded as different descriptions of the same thing and are said to be *dual* to each other. Over the years, many examples of such dualities have been found, both in quantum field theory and string theory. There are even examples of theories that are *self-dual*, in the sense that two different choices of the parameters of the theory lead to the same physics.

In this chapter we describe what is arguably the most astonishing duality of all, discovered by Juan Maldacena in 1997. This duality, called the *AdS–CFT correspondence* relates a string theory living in ten space–time dimensions to a quantum field theory living in four dimensions. Superficially, such a duality is unthinkable. In everyday life, for example, it is obvious to us that life in, say, one space dimension is very different to life in more than one space dimension: in one dimension, for example, we cannot help but collide with our fellow beings as we move around, while in more than one dimension we are free to go around them. But this naive picture is built on the concept of objects (such as ourselves) which interact only weakly with one another. When couplings become strong, what actually happens is much less clear.

### 59.2 The String Theory Side

On the string theory side, the theory is described by the Type IIB strings mentioned in the last chapter, but propagating in a non-trivial, ten-dimensional space–time. The space–time consists of a five-dimensional space part which is just a five-dimensional sphere (hence compact), together with a five-dimensional non-compact space–time called anti-de Sitter space–time. This space–time can be obtained as a solution of Einstein’s equation (in five rather than four space–time dimensions) with a *negative* cosmological constant. Unlike for a positive cosmological constant, where the slices of space at constant time are spheres, now the spatial slices are hyperbolic spaces, with geometry like a saddle, as in Figure 48.1 of Chapter 48.

One of the many strange properties of the five-dimensional anti-de Sitter space–time is that its boundary looks like the usual four-dimensional Minkowski space–time. Thus, one way to picture the duality is that the quantum field theory on the other side of the correspondence ‘lives’ on the boundary of the five-dimensional anti-de Sitter space–time.

Because string theory is gravitational, this theory exhibits many of the phenomena that we observe in general relativity. For example, the theory features black hole solutions. One of the miracles of the AdS–CFT correspondence is that all of these phenomena

must have an interpretation in terms of the quantum field theory living on the boundary.

### 59.3 The Quantum Field Theory Side

Now we describe the quantum field theory that lives on the boundary. In some ways, it is a theory which is familiar, being an  $SU(N)$  gauge theory similar to those in the Standard Model. But unlike the Standard Model, it is supersymmetric. Highly supersymmetric in fact, with the maximal possible amount of supersymmetry in four space–time dimensions. As such this is a very special theory, with very special properties. The most important among these is that the coupling constant of the theory neither becomes bigger at low energies (in QCD) nor smaller (as in QED): rather, it stays exactly the same. The properties of the theory are, in fact, completely independent of the scale at which we probe it. Such a theory is called a *conformal field theory* or CFT.

### 59.4 The AdS–CFT Dictionary

The AdS–CFT correspondence is often described as a *holographic duality*, because it relates a theory with five non-compact dimensions (the AdS space–time on the string theory side) to a theory with four non-compact space–time dimensions (the field theory on the boundary). Thus, all of the information about the physics of the extra dimension must be subtly encoded in the four-dimensional field theory, just as a two-dimensional hologram encodes information about a three-dimensional image. Qualitatively, this can happen because of the scale invariance of the field theory: the information about where in the extra dimension a process takes place on

the string theory side is encoded in the field theory in the scale at which the process takes place.

This procedure is an example of the *AdS–CFT dictionary*: given that the theories are equivalent, there must be a way to translate physical quantities in one theory to the other, and the dictionary provides the concrete description for how to do so. The dictionary provides physicists with a powerful tool: in cases where one can calculate quantities on both sides of the correspondence, one can use the calculations to check that the correspondence is correct (and this is how Maldacena originally found the correspondence). But in other cases, one can use calculations carried out on one side to infer results on the other side in regimes where it is not known how to compute.

### 59.5 Applications of AdS–CFT

Since its discovery in 1997, the AdS–CFT correspondence has found many applications. Within string theory, the power to calculate in new regimes has given us a much better understanding of string and M-theory in general. But AdS–CFT has also been applied in completely different areas of physics. Indeed, wherever one encounters the problem of strong interactions and an ability to calculate perturbation theory, one can try to get an understanding, at least at the qualitative level, using an appropriate modification of AdS–CFT. As a result, AdS–CFT has found applications in particle physics (for example to the strong interactions that bind the Higgs in composite Higgs models, discussed in Chapter 46), to studies of nuclear physics and heavy-ion collisions at the LHC, and to strongly-interacting systems in condensed matter physics. No doubt there are yet more applications still waiting to be found.

## *Consequences of the Theory*

### **60.1 The Richness of String and M-theory**

Even if string theory or M-theory is correct, then there remains the important question of how it results in the Universe we see around us. This question is, of course, made difficult to answer by the fact that we do not yet know precisely what M-theory is! But even so, we can study it at its various string theory limits, to see if we can find something that resembles the real world.

String theory certainly contains all the elements we need, such as gravity, gauge fields, bosons, fermions and so on, but the theory still appears very different to the world we live in. The most obvious discrepancy is that string or M-theory has either ten or eleven space–time dimensions and yet we see only four. There are two intuitive ways in which this can be explained. The first is that the six or seven extra dimensions must be rolled up, or compactified, and are so small that they cannot be resolved by experiments at currently accessible energies. When such compactifications occur in string theory, new particles appear in the theory corresponding to strings wrapped around the rolled-up dimensions (Figure 60.1). The second possibility is that the extra dimensions are rather large (and possibly infinite), but that we simply cannot ‘see’ them. As we saw above, string theories typically contain branes on which open strings must end. Open strings, and the particles they correspond to at low energies, are thus trapped on the brane and cannot probe the extra dimensions. Thus it could be that we live on a four-dimensional brane and that the Standard

Model particles we observe are trapped on the brane. We are simply unaware of the large extra dimensions around us. This so-called *braneworld hypothesis* has another intriguing consequence. Suppose we live on such a brane. Then it is reasonable to suppose that there are other parallel braneworlds separated from us in the extra dimensions. Could these too support life?

These simple explanations raise many questions. What causes some dimensions to be rolled up while others are not? Why are four so large and the others so small? These are still very much open questions. One direction in which a great deal of progress has been made in the last few years is in studying which compactifications of the extra dimensions are allowed if we are to end up with a realistic four-dimensional theory which looks like the Standard Model at low energies. Are such compactifications possible at all? For a long time, this was not clear. Again it was hoped that there might be just one or two special compactifications which led to the Standard Model. However, the discovery of branes has made it clear that string theory is much richer than we first thought, with many more ways of compactifying the extra dimensions, and many more states (particles) at low energy. Consequently, the number of possible compactifications has increased dramatically. An estimate suggests that there are around  $10^{500}$  compactifications of string theory, of which  $10^{150}$  may result in something akin to the Standard Model at low energies! This problem of the myriad possible compactifications of string theory has

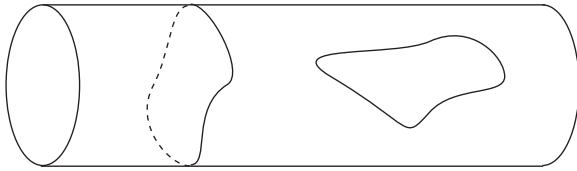


Figure 60.1. Compactification of an extra space–time dimension into a circle. Closed strings winding around the circle (like the string on the left) generate new states in the theory.

become known as the ‘vacuum selection problem’ and there are very few ideas at the moment as to how it may be resolved. Even if one doesn’t start from a unique theory at high energies, the many theories one can end up with at low energy appear to be very different and so one’s ability to make predictions about the low-energy physics is completely lost.

## 60.2 Back to the Anthropic Principle

The realisation that string theory contains so many vacua was regarded as a disaster by many physicists, but others saw it as an opportunity. With  $10^{500}$  different vacua, it seems likely that there exists a vacuum in which the cosmological constant, or vacuum energy density is  $10^{120}$  times smaller than the Planck scale. Indeed, if the distribution of the values of the cosmological constant in this *landscape* of vacua is uniform, we would expect not just one such vacuum, but  $10^{380}$ ! This raises an extraordinary hope for solving the cosmological constant problem. Suppose that there exists a mechanism by which all vacua in the landscape are realised during the evolution of the Universe, somehow, with different corners of the Universe corresponding to different vacua. If so, then in corners where the cosmological constant is too large (either positive or negative), no intelligent life will evolve and no one will be around to measure a natural value of the cosmological constant. But in those special vacua where the cosmological constant is very small, galaxies may form, life may evolve, the cosmological constant in that corner of the Universe may be measured, and a cosmological constant problem formulated.

What is more, the landscape idea can also be applied to the Higgs hierarchy problem. Now the argument says that without a light Higgs mass, the electroweak force would behave completely differently. In particular, atoms would not form, and the Sun would not produce light via nuclear interactions.

Again, it is difficult to imagine how life (at least as we know it) could form without complex atoms (in particular carbon) and without an energy source in the form of the Sun. But with a landscape of vacua, we merely need a mechanism which allows many vacua to be realised; we will then find ourselves occupying one conducive to our existence, with a light Higgs mass.

Remarkably, there is a mechanism by which this could occur, going by the name of *eternal inflation*. The idea is that the inflaton starts out at some point in the landscape and begins rolling towards the nearest local minimum, leading to a period of inflation. But because of quantum fluctuations, there is a non-vanishing probability for the inflaton to tunnel, in some region of space, towards a different local minimum of the landscape, with lower energy. If this happens, then that region of space will also start to inflate, but at a different rate, fixed by the new value of the vacuum energy at or near the local minimum. And so the process continues, leading to the Universe exploring many different local minima (each with different values of the cosmological constant), by an eternal process of inflation and quantum tunnelling.

A key feature of this scenario is that because the different corners of the Universe are expanding exponentially fast, they are causally disconnected from one another. That is, there is no way to send light signals between them and so no way for observers, like ourselves, to carry out observations in other corners.

Such a scenario represents, in one sense, just one more step along the road first trod half a millennium ago by Copernicus, in which we humans have gone from regarding ourselves as being at the very centre of everything, to being the inhabitants of just one among planets, orbiting one among many stars, in one among many galaxies and so on. If the landscape idea is correct, our entire observable Universe is just one corner among  $10^{500}$  others. But there is another sense in which the landscape idea represents a rubicon: for the first time, the new extent of our Universe cannot be confirmed by observation, because the other corners are not causally connected to us. But if we cannot possibly observe the other corners of the Universe, can we still claim to be doing Science?

## 60.3 A Theory in Search of Experiment

String theory has now been around as a serious candidate for a Theory of Everything for two decades.

Theoretical physicists have devoted a great deal of effort in that time to overcome setbacks which seemed repeatedly to thwart the theory. Finally, we seem to be getting close to a consistent theory which *can* describe nature, including solutions of great mysteries such as the cosmological constant problem.

The big question is *does* it actually describe nature? That is, is string theory the Theory of Everything which it purports to be? That question, like all questions in physics, can only be settled ultimately by experiment. The obstacle in the case of string theory is that the theory only reveals its true stringy nature near the Planck scale of  $10^{19}$  GeV. By comparison, the state of the art accelerator experiments operate at around  $10^3$  GeV which is a long, long way away from the Planck scale. So a direct test of string theory seems to be beyond us, at least for the foreseeable future. Although string theory effects could be observed indirectly (e.g. via virtual string processes), the effects are likely to be suppressed by factors of  $10^3 \text{ GeV} / 10^{19} \text{ GeV} \sim 10^{-16}$ . To measure such effects would require experiments far more accurate than any yet performed in the history of science, even in QED.

Another hope is that, since string theory is supposed to be unique, it could be possible to *predict* one or more of the 19 parameters in the Standard Model. However, as we saw above, the process of compactification into one of many possible vacua introduces a huge number of parameters into the theory, which may or may not be fixed by environmental/anthropic issues. So this hope too seems to be stymied.

The absence of any experimental confirmation (or denial) of string theory has led to a polarisation of opinion amongst particle physicists. Some champion it as the Theory of Everything, some dismiss it as a theory of nothing. Nevertheless, it is still the subject of intense scrutiny. Great progress has been made in the last decades using the guiding principle of internal consistency alone. It would be surprising, too, if the second string revolution proved to be the last, and

this brings hope that a dramatic turn of events (e.g. singling out the compactification) could lead to an experimentally testable prediction.

#### 60.4 Conclusion

In the last 50 years, particle physics has come a long way. Quantum field theories such as QCD and the electroweak theory, once considered esoteric and treated with suspicion, have now become the Standard Model, routinely tested and confirmed by precision experiments to many decimal places. But precision experiments, both on the ground and in the sky, have also raised questions.

As regards ‘pure’ particle physics, the discovery that neutrinos are not quite massless, and many other puzzles besides, have taught us that there *must* be physics beyond the Standard Model. Despite many elegant theories being put forward in an attempt to describe it, none has found experimental confirmation so far. On the contrary, heroic efforts at the LHC and elsewhere have ‘merely’ confirmed the predictions of the Standard Model, time after time. Of course, there are always small anomalies present in the data and one of these may turn out to be the key to unlocking the door to new physics. If so, particle physicists must ensure that they have the experimental facilities in place to exploit it.

As regards cosmology, things are perhaps even more exciting. Gravitational waves have opened a new window on the Universe, while other experiments have entered a new era of precision and the puzzles they raise are truly profound. What is the real nature of the dark matter and energy which make up 96% of the Universe? What caused the fantastic inflation that we believe occurred very close to the beginning of time? Will grand unification, supersymmetry or string theory (whatever it really is) answer any or all of these questions? Nobody knows what the future may bring, but the only real surprise would be if there were no surprises.

## **Part XV**

### **The Future: To Boldly Go!**





## *Accelerators, Observatories and Other Experiments*

### 61.1 Accelerators

Clearly, the last decade of particle physics has been dominated by the operation of the LHC. Undoubtedly its greatest success was the discovery of the Higgs boson in 2012 and the greatest mystery, the absence of any signs of the widely anticipated supersymmetric particles first suggested in the early 1970s, or indeed any other unambiguous evidence for physics beyond the Standard Model. But as described in Parts X and XI, the LHC has proved the spectacular accuracy of the Standard Model in a host of measurements, and has also provided some intriguing hints of physics beyond it. So where next?

Given the very long timescales and very high cost of any brand new accelerator facilities the near term is dominated by the upgrading or re-purposing of existing facilities:

- The high-luminosity large hadron collider (HL-LHC). Following its first decade of operation, the LHC has now ceased operation to upgrade the machine to circulate more intense beams of proton–antiprotons to achieve a luminosity of ten times the current level generating 25 times as much data. The greatly improved statistics will allow far more exhaustive searches for physics beyond the Standard Model, as well as more detailed studies of the properties of the Higgs boson itself, such as its coupling to the fermionic

sector, and the study of various other phenomena, such as new hadrons.

- The Fermilab Tevatron. The DUNE (Deep Underground Neutrino Experiment) (Figure 61.1) will continue to use the high-energy proton beams from the Tevatron to generate neutrino beams to be detected by the DUNE detectors some 1300 km distant. This will allow deeper study of neutrino oscillations, hence of their masses and possible role in the dark matter make-up and CP violation in the Universe.

But in the longer term, several projects are under consideration for totally new accelerators around the world. Clearly, all such projects will be subject to years of planning approval and construction. As a consequence, none is likely to see operation until the mid-2030s.

- The International Linear Collider (ILC), currently planned to be sited in Japan, is intended as the much larger replacement of the Stanford Linear Accelerator Center (SLAC) collider which achieved so much in the 1970s and 1980s. In this electrons and positrons are to be collided at c. 500 GeV at the end of a 30 km tunnel to be located some 400 km north of Tokyo.
- The Chinese Electron Positron Collider (CEPC) (Figure 61.2) is proposed also as an electron

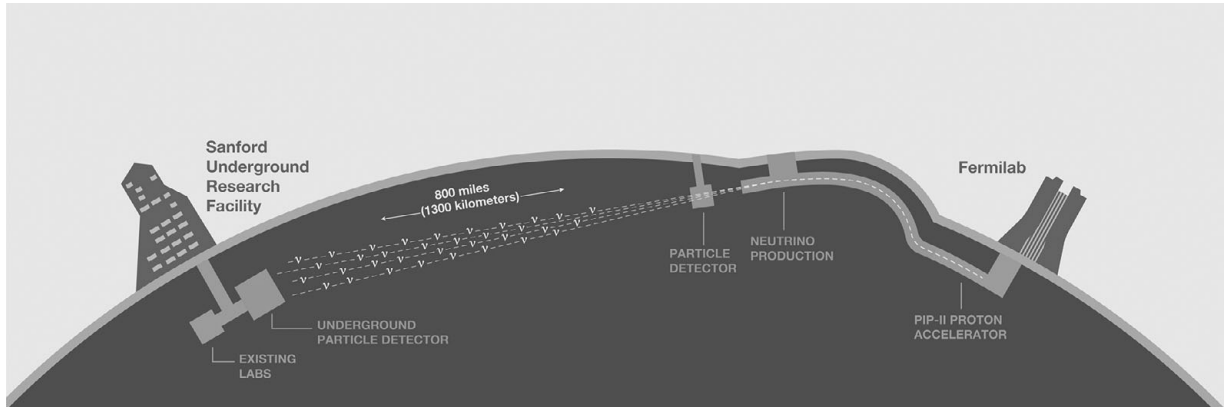


Figure 61.1. The DUNE experiment. (Image: Courtesy Fermilab.)

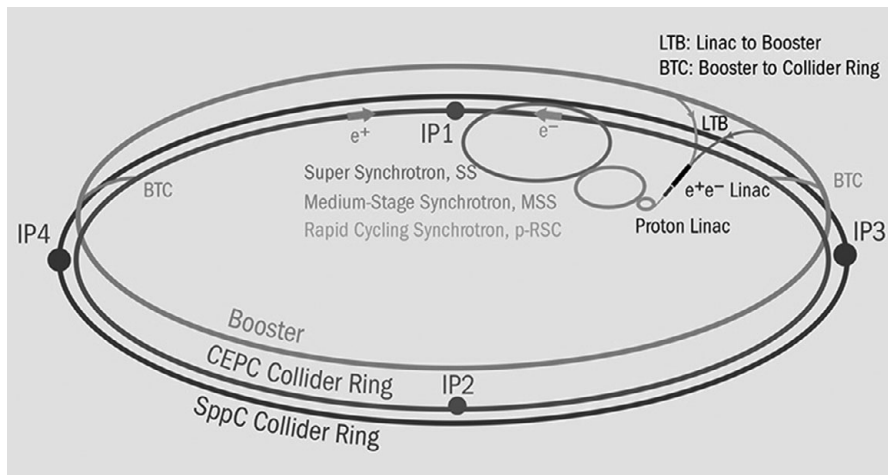


Figure 61.2. The proposed Chinese Electron Positron Collider. (Image: Courtesy CEPC.)

positron collider in a circular tunnel of some 100 km circumference at a site to be decided in mainland China.

- The Future Circular Collider (FCC) (Figure 61.3) at CERN is suggested as the successor to the current LHC. Formally launched in 2014, an initial study calls for a 100 km tunnel at the CERN site in Geneva which may achieve proton–proton collisions at up to 100 TeV.
- The Electron–Ion Collider (EIC) is a proposal in the US to continue the studies of the quark–gluon plasma inside the nucleons/nuclei, thereby continuing the investigation of the LHCb experiment at CERN and the relativistic heavy ion accelerator at Brookhaven in the US.

Also worthy of note is the AWAKE experiment at CERN which is demonstrating a technique for accelerating electrons by guiding them to surf on top of a longitudinal plasma of protons generating regions of alternating positive and negative charge. In contrast to the usual r.f. acceleration technology currently in use, the AWAKE method can accelerate electrons achieving ten times the energy/distance gradient of existing accelerators. Although early days, this could provide the mechanism for attaining ultra-high energies in coming decades.

### 61.2 Observatories and Other Experiments

Although less well known to the public, various varieties of observatories have achieved some

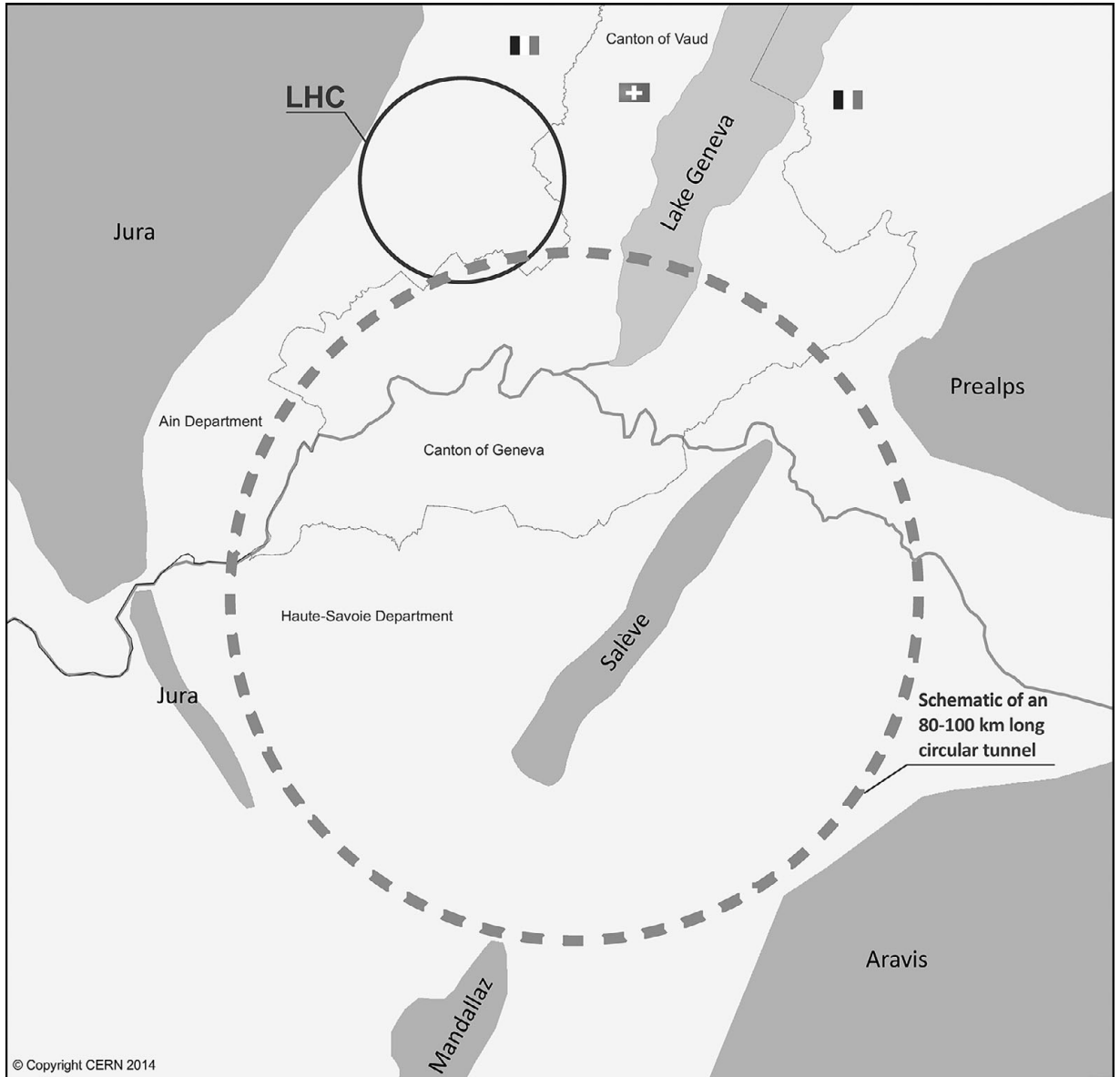


Figure 61.3. The proposed Future Circular Collider at CERN. (Image: Courtesy CERN.)

of the most intriguing physics discoveries of the last few decades, short of the actual discoveries of new particles.

- Kamiokande/Super-K/Hyper-K. This experiment in Kamioko in Japan monitors a volume of water to detect the Cherenkov radiation emitted by a particle struck by a neutrino from a beam generated at the J-PARC facility in Tokai some

295 km distant. It was at Super-K that neutrino oscillations were first discovered in 1998.

- Ice Cube. This array of a 1 km cubic lattice of photomultiplier tubes buried in the Antarctic ice is focused on the detection of high-energy neutrinos from deep space (Figure 61.4). Its most recent achievement was the detection of a 300 TeV neutrino from a blazar galaxy (TXS 0506), the particle emission jet of which points directly



Figure 61.4. The Ice Cube experiment in Antarctica. (Image: Courtesy IceCube University of Wisconsin-Madison.)

towards Earth. Its localisation in the sky allowed the Ice Cube alert system to communicate with other observatories. As a result, both the Fermi gamma-ray satellite and the MAGIC gamma-ray telescope in La Palma Spain detected associated gamma rays, making it one of the few examples thus far of multi-messenger astronomy.

- GADMC. The Global Argon Dark Matter Collaboration is the combined effort of the ArDM laboratory in Spain, Gran Sasso in Italy and SNOLAB in Canada. The experiment is attempting to detect massive dark matter candidate particles, so-called WIMPS (weakly interacting massive particles for which the neutralino, the more massive supersymmetric partner of the obviously ultra-low mass Standard Model neutrinos, is a leading candidate) by detecting

their collisions with argon nuclei marked by the emission of a burst of scintillation light. Predecessor experiments based on xenon detectors reported null results; the argon-based experiments are yet to report.

These facilities listed above are just a few of the myriad laboratories around the world conducting dozens of experiments on many aspects of the Standard Model and the physics which must lie beyond. Most countries in North America, Europe, Southern Asia and the Far East have one or more national laboratories engaged in experiments to test the 14 issues listed in Chapter 42, and more besides. Those countries not able to sustain an independent effort very often join CERN or other multinational facilities as associate members allowing their scientists to contribute to the global effort.

## *Known Unknowns*

### 62.1 The Current In-tray

In the introduction to Part XI: Reasons to Go Beyond, we listed 14 main mysteries not explained by the Standard Model. Below we list the current experimental attempts to address some of these issues, any of which would mark a significant step into the unknown territory lying beyond it:

- Supersymmetric particles. Clearly the subject of the most intensive searches at the LHC. So vast is the volume of data at the LHC experiments, the search will continue even as the machine is shut down by trawling the data from LHC run 2 which ended in December 2018 until the start of LHC run 3, scheduled for 2021 ahead of the eventual upgrade to the HL/LHC scheduled for operation in 2025.
- Anomalies in rare B-decays. Given the universality of lepton flavour (i.e. the notion that electrons, muons, and tau mesons are indistinguishable apart from in their mass), decays of neutral B mesons (consisting of a down quark and a bottom antiquark) should result (to a very good approximation) in equal numbers of electron–positron and muon–antimuon pairs being produced – but instead an unequal number is observed. Similarly, decays of charged B mesons show inequalities in decays to tau mesons (plus a neutrino) versus decays to muons. Such anomalies have been observed both at LHCb and at the dedicated

B-physics experiments BaBar and Belle. As the statistics (and significance) of these anomalies grow, they are increasingly being interpreted as the effect of intermediate, virtual, beyond the Standard Model states.

- Axions. Three classes of experiments are looking for axions, thought to be the ultra-light products of the additional symmetry required to suppress CP violation in the strong interaction, as discussed in Chapter 47. The axion mass is currently anticipated in the range  $10^{-6}$  to  $10^{-3}$  eV. All experiments rely on the coupling of axions to photons in the presence of a strong magnetic field. The first class looks for axions emitted from the Sun which may be converted to photons in a custom-designed helioscope (e.g. CAST – the CERN Axion Solar Telescope). A second class results from the conversion of axions into photons in a resonant microwave cavity (e.g. the ADMX – Axion Dark Matter Experiment at the University of Washington in the US). A third class involves shining laser light at an opaque wall and seeking a signal on the other side (due to photon conversion to axions, which can penetrate the wall, followed by reconversion to photons on the other side of the wall (e.g. the ALPS 2 experiment at the DESY laboratory in Germany). Thus far, no signs of axions have been detected, although the experimental limits are compatible with the mass range quoted above.

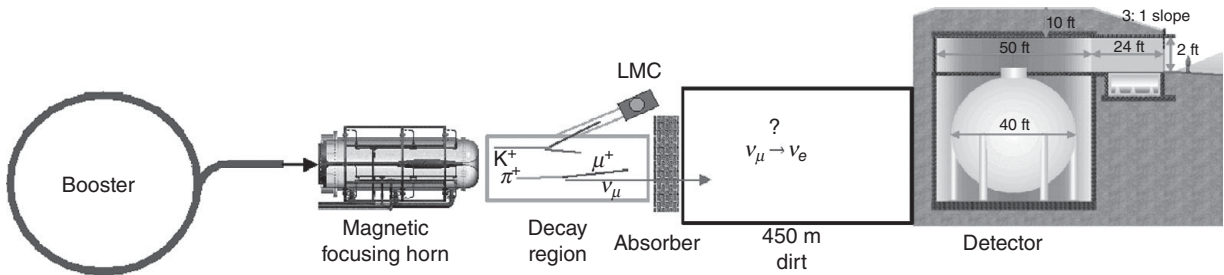


Figure 62.1. The miniBOONE experiment. (Image: Courtesy Fermilab.)

- Sterile neutrinos. The three generations of conventional neutrinos have left-handed helicity and interact via the weak interaction with the other fermions and bosons. However, as discussed in Chapter 43, Section 43.7, an extension of the Standard Model can accommodate right-handed helicity neutrinos which do not interact via the weak interaction but only by gravity, but nevertheless make their presence felt by affecting the oscillations of the more familiar neutrinos. The effect is to speed up the oscillations thereby shortening the distances over which they occur. The current miniBOONE (Booster Neutrino Experiment) (Figure 62.1) experiment at Fermilab has detected just such an effect, confirmation of which would provide yet another hint of physics beyond the vanilla Standard Model.
- The anomalous magnetic moment of the muon. This is a measure of the strength of the magnetic field generated by the electrically charged spinning muon. The Dirac equation predicts the exact value of 2. However, contributions are due also from loop corrections in QED, also from the electroweak force and finally from hadronic effects. Currently the value deviates marginally from the Standard Model prediction. If this deviation were confirmed, then it could imply contributions to the calculation from beyond the Standard Model particles. The

current experiment at Fermilab is taking data and is regarded as a key test of the Standard Model.

- The electric dipole moment of the neutron. This is the extent to which the neutron may appear as a dipole of regions of positive and negative electric charge, despite its electrically neutral behaviour overall. The existence of a neutron dipole moment would imply the breaking of CP symmetry which is required to account for the matter–antimatter asymmetry observed in the Universe. There are currently at least six experiments in progress in national laboratories in Europe and the US which are increasingly restricting the dipole moment to ever more minimal levels. Failure to detect this dipole moment would imply deeper mysteries in our understanding of CP violation and its role in the matter–antimatter make-up of the Universe.

This list is just a sample of the main experiments currently of interest. There are scores of others studying a host of questions ranging from the very small (e.g. putting limits on the lifetime of the proton) to the very large (e.g. using the gravitational lensing of light from distant galaxies to elucidate the nature of dark energy). Also of interest is a range of more conventional laboratory experiments which are exploring further some of the more subtle effects of quantum physics, e.g. entanglement.

## *Glittering Prizes*

Although the examples of Newton, Maxwell, Hawking and others demonstrate that the prime motive for the progress of physics is that of intellectual curiosity, there is no doubt that the lure of recognition is a strong motivation to excel.

### **63.1 The Class of 1984**

The light-hearted conclusion of the first edition of this book proposed a list of the then forthcoming likely milestones deserving, in the authors' opinion, of Nobel recognition. In a similar light-hearted fashion, it is interesting to see how the list turned out. In order then written the list was:-

W, Z discoveries. In process during publication. Nobel Prize awarded to Carlo Rubbia and Simon van der Meer in 1984.

Glueballs. Difficult to establish conclusively, due to the broad nature of hadronic resonances appearing in QCD. However, this field has developed into the study of the quark-gluon plasma, now the subject of experiments in lead ion collisions at CERN and elsewhere.

Proton decay. Not detected, but possible in some theories. Remains on the list.

Higgs bosons. Detected in 2012 in the LHC at CERN some 44 years after their proposal. Nobel Prize in 2013 for Francois Englert and Peter Higgs.

Top quark. No prize granted for the particle detection per se, but half the Nobel in 2008 was awarded to Makoto Kobayashi and Toshihide Maskawa for their work predicting the existence of a third

generation of the Standard Model from flavour physics considerations.

Magnetic monopole. Not detected, possible in some theories. Remains on the list.

Neutrino oscillations and mass. Detected at the Kamiokande laboratory in Japan, Nobel Prize awarded to Takaaki Kajita and Arthur B. McDonald in 2015.

Free quarks. Not detected in the sense originally intended, i.e. as a fractionally charged particle in free space. But the discovery of asymptotic freedom in QCD led to the discovery that quarks behave as if free but within the confines of the nucleon. Poetically, quarks are asymptotically free but everywhere in (gluonic) chains. The Nobel Prize was awarded in 2004 to David J. Gross, Frank Wilczek and H. David Politzer for their formulation of asymptotic freedom.

Supersymmetric or technicolour particles. Not discovered so far despite the greatest particle hunt in history. Remains on the list.

Gravitational waves. Discovery suggested as very long odds. Detected in 2016 at LIGO. Nobel Prize to Rainer Weiss, Kip Thorne and Barry Barish in 2017.

Having drawn the above Nobel quest to a close after some some three decades with 60% success, 30% pending, and one void, we will avoid speculating again on what the future may hold, other than to note that the list in Chapter 42 gives at least another 14 outstanding matters, any one of which could achieve a similar recognition to those listed above.



## *Unknown Unknowns: It Must Be Beautiful*

There is not a lot that can be written about the unknown unknowns other than, perhaps, to draw some lessons from history. In the closing years of the nineteenth century, all would have seemed reasonably well with the world of what we now call classical physics. Apart, that is, from some curious discrepancies. Firstly, there was the puzzling constancy of the speed of light and the failure to detect any medium through which it might have propagated. Secondly, there was the difficulty in explaining the spectrum of light radiated from a heated body. Thirdly, there was a measurable discrepancy in the advance in the perihelion of the planet Mercury with the calculations of the Newtonian theory of gravity.

As we now know, these three discrepancies led to the formulation of three completely new theories which in turn led to the then unimaginable consequences we now take for granted. The special theory of relativity led to a comprehensive revision of our notions of physical dynamics and, most famously, to the equivalence of mass and energy. The advent of quantum mechanics to explain the problem of black-body radiation led to a comprehensive revision of our ideas of matter and radiation, allowing the invention of devices now ubiquitous on the planet and in use on a daily basis as a matter of course, e.g. nuclear power, lasers, computers, mobile phones, etc.

The combination of special relativity and quantum mechanics led to the prediction of antimatter

which opened up, literally, a parallel universe, necessary to explain the one we actually inhabit.

Finally, Einstein's formulation of general relativity revised our understanding of the nature of space and time and the behaviour of the large-scale nature of the Universe, the manifestations of which we are only now beginning to observe and, increasingly, to comprehend.

### **64.1 The Challenges of Quantum Gravity**

Obviously, the major problem outstanding is the successful unification of general relativity with quantum theory. This challenge has occupied the most capable of the subject's minds for the last four decades with no verifiable results. This suggests historic parallels and what we might infer from them.

String theory as described in Part XIV has been the main field of activity to address the challenge of quantum gravity and the only one which has had any real success, even at the theoretical level. But even though much has been learnt about what string theory is (and what it is not), we still lack clear confirmation of any of its predictions, necessary for promotion to the status of a physical theory. In this regard, string theory occupies a position similar perhaps to the development of non-Euclidean geometry in the mid-nineteenth century: an interesting theoretical extension of the real world, but of no practical application until a discovery made it so. In the case of non-Euclidean

geometry it was Einstein's formulation of general relativity as a theory of gravity which brought the previous mathematical work to the fore as a theory of physics.

The copious amount of mathematical work of the last four decades may now be in a similar position, awaiting the kiss of a practical application to turn them into physical theory. We might also be alert to the idea that this practical application may not be in the area of particle or astroparticle physics but in the field of laboratory physics, quantum technologies or further afield.

Also, we should remember that Einstein and others spent some decades of the mid-twentieth century attempting to unite general relativity with electromagnetism, then a perfectly understandable idea. But even now this is unachieved due to the difficulties in combining quantum theory with gravity. Also, it was not until the discovery of the neutron in 1932 that the full extent of nuclear forces became apparent, which made the restricted combination of gravity and electromagnetism only a sub-set of the full theory that would be required.

## **64.2 The Beautiful Equations**

A striking feature of the successful ideas in physics is their simplicity of expression, once known. The great theories can nearly all be summarised in equations which look very simple. The Newtonian formulae for both dynamics and gravitational attraction (Chapter 5, Section 5.2) are cases in point. Maxwell's formulae for electromagnetism can be summarised in a seemingly simple form very similar to Einstein's formula describing gravity in general relativity (shown in Chapter 5, Section 5.2.1).

The formulation of special relativity leads to the famous mass–energy formula (Chapter 2), the very model of simplicity. Likewise, the founding hypothesis of quantum mechanics is of the proportionality of the energy and frequency of radiation, the constant of proportionality being Planck's constant (Chapter 3).

The combination of special relativity and quantum mechanics into the apparently simple Dirac equation (described in Chapter 4, Section 4.2) gives little hint of the profundity of the phenomena that it implies, including antimatter and much more.

In contrast, as we saw in Chapter 37, Section 37.3, the working formula of the Standard Model is complicated, requiring separate expressions for the gauge, fermionic and Higgs sectors (and possibly another for the axions) and involves approximately 19 separate parameters. How this model can be expanded to include gravity is obviously the challenge currently at the forefront of the subject; perhaps, in doing so, the fundamental equations of physics will take on a new-found simplicity.

Looking back over the achievements summarised in the last few paragraphs rather suggests that there is some as yet undiscovered facet of the physical world to be discovered and articulated to bring simplicity to the currently complex and incomplete picture we have today.

The idea that there are fundamental ideas yet to be discovered which will explain the current puzzles, both large (dark matter and dark energy) and small (the list of Chapter 42), suggests that the next 40 years of particle physics, astroparticle physics and cosmology will be every bit as exciting as the last.



## **Appendices**



## APPENDIX A

# *Units and Constants*

*Energy:* The most common unit of energy in the microworld is the electronvolt, eV. This is defined as the energy possessed by an electron after it has been accelerated through a potential difference of one volt.

$$1 \text{ eV} = 1.602 \times 10^{-19} \text{ J},$$

$$1 \text{ keV} = 10^3 \text{ eV}, \quad 1 \text{ MeV} = 10^6 \text{ eV},$$

$$1 \text{ GeV} = 10^9 \text{ eV}, \quad 1 \text{ TeV} = 10^{12} \text{ eV}.$$

*Mass:*

$$\begin{aligned} \text{electron } m_e &= 9.109 \times 10^{-31} \text{ kg} \\ &= 0.511 \text{ MeV}/c^2, \end{aligned}$$

$$\begin{aligned} \text{proton } M_p &= 1.673 \times 10^{-27} \text{ kg} \\ &= 938.27 \text{ MeV}/c^2. \end{aligned}$$

*Charge:*

$$\text{electron } e = 1.602 \times 10^{-19} \text{ C}.$$

*Speed of light:*

$$c = 2.998 \times 10^8 \text{ ms}^{-1}.$$

*Planck's constant* ( $\hbar = \frac{h}{2\pi}$ ):

$$\begin{aligned} \hbar &= 1.055 \times 10^{-34} \text{ J s} \\ &= 6.582 \times 10^{-22} \text{ MeV s}. \end{aligned}$$

*Fine structure constant* ( $\alpha = \frac{e^2}{4\pi\epsilon_0\hbar c}$ ):

where  $\epsilon_0$  is the permittivity constant of the vacuum,

$$\alpha = \frac{1}{137.036}.$$

## APPENDIX B

# *Glossary*

**Abelian group** A mathematical group of transformations with the property that the end result of a series of transformations does not depend on the order in which they are performed.

**absolute temperature** Temperature measured on the Kelvin scale: 0 Kelvin =  $-273.15^\circ$  Celsius. Absolute temperature is directly related to (kinetic) energy via the equation  $E = k_B T$ , where  $k_B$  is Boltzmann's constant. So, a temperature of 0 K corresponds to zero energy, and room temperature, 300 K =  $27^\circ\text{C}$ , corresponds to an energy of 0.025 eV.

**alpha ( $\alpha$ ) particles** Particles first discovered in radioactive  $\alpha$  decay, and later identified as helium nuclei (two protons and two neutrons bound together).

**amplitude** See quantum-mechanical amplitude.

**angular momentum** The rotational equivalent of ordinary momentum, being mass  $\times$  velocity  $\times$  orbital radius. It is a vector quantity directed along the axis of rotation. In quantum mechanics, (orbital) angular momentum is quantised in integer multiples of  $\hbar$ . This corresponds classically to only certain frequencies of rotation being allowed.

**antiparticles** Particles predicted by combining the theories of special relativity and quantum mechanics.

For each particle, there must exist an antiparticle with the opposite charge, magnetic moment and other internal quantum numbers (e.g. lepton number, baryon number, strangeness, charm, etc.), but with the same mass, spin and lifetime. Note that certain neutral particles (such as the photon and  $\pi^0$ ) are their own antiparticles.

**asymptotic freedom** A term used to describe the observed *decrease* in the intrinsic strength of the colour force between quarks as they are brought closer together. At asymptotically small separations, the quarks are virtually free. This is in contrast to the electromagnetic force whose intrinsic strength increases as two charged particles approach each other.

**B-factory** Asymmetric electron-positron collider experiments tuned to the  $\Upsilon$  resonance for the study of bottom quarks.

**baryogenesis** The process by which the Universe's net baryon number was generated. This explains why the Universe is made predominantly of baryons and not antibaryons.

**baryon** The generic term for any strongly-interacting particle with half-integer spin in units of  $\hbar$  (e.g. the proton, neutron and all their more massive excited resonance states).

**beta ( $\beta$ ) particles** Particles first discovered in radioactive  $\beta$  decay – later identified as electrons.

**Big Bang theory** The most widely accepted theory of the origin of the Universe. It asserts that the Universe began some  $10^{10}$  years ago from a space–time point of infinite energy density (a singularity). The expansion of the Universe since that time is akin to the expansion of the *surface* of an inflating balloon: every point on the balloon’s surface is moving away from every other point. So, microbes living on the surface see their two-dimensional world expanding, yet there is no centre to the expansion which is everywhere uniform.

**boson** Any particle with integer spin: 0,  $\hbar$ ,  $2\hbar$ , etc.

**Cabbibo angle ( $\theta_C$ )** The measure of the probability that one flavour of quark (u) will change into other flavours (d or s) under the action of the weak force.

**CERN** The European Laboratory for Particle Physics (formerly the Conseil Européen pour la Recherche Nucléaire), located near Geneva in Switzerland. Here, the resources of the European member nations are pooled to construct the large particle accelerators needed for high-energy experiments. The major project at CERN currently is the LHC; see Chapter 40.

**charm** The fourth flavour (i.e. type) of quark, the discovery of which in 1974 contributed both to the acceptance of the reality of quarks and to our understanding of their dynamics. The charmed quark exhibits a property called ‘charm’ which is conserved in strong interactions.

**chirality** The handedness of a relativistic fermion, defined by how it transforms under Lorentz transformations.

**colour** An attribute which distinguishes otherwise identical quarks of the same flavour. Three colours – red, green and blue – are required to distinguish the three valence quarks of which baryons are composed. It must be stressed that these colours are just labels and have nothing to do with ordinary colour. Colour is the source of the strong force which binds quarks

together inside baryons and mesons, and so the three colours (r, g, b) can be thought of as three different colour charges analogous to electric charge.

**cosmological constant** A term added by Einstein to the gravitational field equations of his theory of general relativity. Such a term would produce a repulsive antigravity force. There is, at present, some evidence for the existence of a cosmological constant.

**cosmological principle** The hypothesis that the Universe is isotropic and homogeneous on very large distance scales.

**coupling constant** A measure of the intrinsic strength of a force. The coupling constant of a particular force determines how strongly a particle couples to the associated field. For example,  $\alpha = e^2/\hbar c = \frac{1}{137}$  (or, equivalently, electric charge  $e$ ) specifies the strength of the coupling of charged particles to the electromagnetic field.

**cross-section ( $\sigma$ )** The basic measure of the probability that particles will interact. It corresponds to the effective target area (in, for example,  $\text{cm}^2$ ) seen by the ingoing particles. It can be derived from the quantum-mechanical interaction probability. A convenient unit for measuring cross-section is the barn (symbol: b), defined as  $1 \text{ b} = 10^{-24} \text{ cm}^2$ . Typical hadronic cross-sections are measured in millibarns;  $1 \text{ mb} = 10^{-27} \text{ cm}^2$ . However, neutrino collision cross-sections are typically much smaller,  $10^{-39} \text{ cm}^2$ .

**dark matter and energy** New sources of energy density and matter required to explain cosmological observations.

**DESY** The German national laboratory for high-energy physics, located near Hamburg. It is the home of the  $e^+e^-$  storage rings DORIS and PETRA, and the electron–proton machine, HERA.

**deuteron** The nucleus of deuterium, an isotope of hydrogen. It consists of one proton and one neutron bound together.

**diffraction** A property which distinguishes wave-like motions. When a wave is incident upon a barrier



which is broken by a narrow slit (of comparable size to the wavelength), then the slit will act as a new source of secondary waves.

**dimensions** Physically significant quantities usually have dimensions associated with them. The fundamental dimensions are those of mass  $M$ , length  $L$  and time  $T$ . The dimensions of other quantities can be expressed in terms of these fundamental dimensions. So, for example, momentum has dimensions of mass  $\times$  velocity ( $MLT^{-1}$ ) and energy dimensions of force  $\times$  distance ( $ML^2T^{-2}$ ). One may define certain quantities in which the dimensions cancel out. These dimensionless quantities are significant in that they are independent of the conventions used to define units of mass, length and time.

**duality** The equivalence between different superstring theories.

**eigenstate, eigenvalue** The eigenvalue of a matrix  $M$  is a number  $\lambda$  which satisfies the equation

$$M\psi = \lambda\psi, \text{ with } \psi \neq 0.$$

In quantum mechanics, the matrix  $\mathbf{M}$  will correspond to a particular dynamical variable (such as position, energy or momentum) and  $\lambda$  will correspond to the value obtained by measuring that dynamical variable if the system is in the state described by  $\psi$ .  $\psi$  is called an eigenstate of the system.

**elastic scattering** Particle reactions in which the same particles emerge from the reaction as entered it (e.g.  $\pi^- p \rightarrow \pi^- p$ ). In *inelastic* scattering, where different and/or new particles emerge, energy is used to create new particles.

**electron** A negatively charged spin- $\frac{1}{2}$  particle, which interacts via the electromagnetic, weak and gravitational forces. It has a mass of  $0.511 \text{ MeV}/c^2$ , some 1800 times lighter than the proton.

**entropy** A quantitative measure of order (or equivalently information) in a physical system. The units are those of energy/temperature.

**family** See generation.

**Fermilab** The Fermi National Accelerator Laboratory, in Batavia, Illinois, USA. Fermilab is the home

of the Tevatron, once the world's most powerful accelerator, a  $p\bar{p}$  collider with a maximum collision energy of  $1.8 \text{ TeV}$  ( $= 1800 \text{ GeV} = 1.8 \times 10^{12} \text{ eV}$ ).

**fermion** Any particle with half-integer spin:  $\frac{1}{2}\hbar$ ,  $\frac{3}{2}\hbar$ ,  $\frac{5}{2}\hbar$ , etc. All fermions obey Pauli's exclusion principle.

**flavour** The term used to describe different quark types. There are six quark flavours: up, down, strange, charm, bottom and top.

**gamma ( $\gamma$ ) rays** Rays first discovered in radioactive material, and later identified as very high energy photons.

**gauge theory** A theory whose dynamics originate from a symmetry. That is, the formulae describing the theory (in particular, the Lagrangian) are unchanged under certain symmetry transformations, called 'gauge' transformations. For example, the equations of classical electrodynamics are invariant under local redefinitions of the electrostatic potential. This symmetry is ultimately responsible for the conservation of electric charge. However, in quantum electrodynamics this gauge symmetry is reinterpreted as invariance under local redefinitions of the phase of the electron wavefunction. The term 'gauge theory' is an archaic one, coming from earlier theories which were based on invariance under transformation of scale (i.e. gauge).

**generation** Leptons and quarks come in three related sets, called generations or families, consisting of two leptons and two quarks. The first generation consists of (e,  $\nu_e$ ; u, d). The second and third generations consist of ( $\mu$ ,  $\nu_\mu$ ; c, s) and ( $\tau$ ,  $\nu_\tau$ ; t, b) respectively.

**gluon, glueball** Gluons are the massless gauge bosons of QCD which mediate the strong colour force between quarks. Because of the non-Abelian structure of the theory, gluons can interact with themselves, and may form particles consisting of gluons bound together. The existence of these 'glueballs' has yet to be confirmed.

**Goldstone boson** A massless spin-0 particle which arises whenever a (continuous) global symmetry is spontaneously broken.

**graviton** A massless spin-2 particle which is the hypothetical quantum of the gravitational field. It mediates the force of gravity in a similar way to that in which the spin-1 gauge bosons (i.e. the photon,  $W^\pm$ ,  $Z^0$ , and gluons) mediate the other forces.

**group theory** The branch of mathematics which describes symmetry. A mathematical group  $\mathbf{G}$  is defined as a collection of elements  $\{a, b, c, \dots\}$  with the properties:

- (1) if  $a$  and  $b$  are in group  $\mathbf{G}$ , then the product of the two elements,  $ab$ , is also in  $\mathbf{G}$ ;
- (2) there is a unit element  $e$  such that  $ae = a$  for all elements  $a$  in  $\mathbf{G}$ ;
- (3) each element  $a$  has an inverse  $a^{-1}$  such that  $aa^{-1} = e$ .

So, for instance, the rotations of an  $(x, y)$  coordinate system about the  $z$ -axis form a group, since the effect of two rotations  $\theta_1$  and  $\theta_2$  is equivalent to the effect of one big rotation  $\theta_3$ . Such a group is called a continuous group as the angles of rotation can vary continuously.

In general, the elements of a group may be represented by matrices, which form various ‘representations’ of the group. These representations may be used to determine how a physical system changes under the action of symmetry transformations. Moreover, when a system possesses a symmetry which is described by a group  $\mathbf{G}$  (i.e. when its equations of motion are left invariant under group transformations), then the various representations specify the symmetry properties of the relevant degrees of freedom.

For example, hadrons appear to possess an  $SU(3)$  ‘flavour’ symmetry. The fundamental three-dimensional representation of  $SU(3)$  contains the three flavour degrees of freedom associated with the up, down and strange quarks:  $\mathbf{3} = (u, d, s)$ . The eight-dimensional representation  $\mathbf{8}$  of  $SU(3)$  contains the eight flavour degrees of freedom associated with the meson and baryon octets (see Chapter 10). Furthermore, a given representation completely specifies the flavour quantum numbers of associated particles.

The same is true for local gauge (i.e. dynamical) symmetries. For example, the theory of QCD possesses a local  $SU(3)_C$  colour symmetry. The three-dimensional representation of  $SU(3)_C$  contains the three colour degrees of freedom associated with the colour charges red, green and blue:  $\mathbf{3} = (r, g, b)$ .

The eight-dimensional representation contains the eight colour degrees of freedom associated with the eight gluons – the gauge bosons of QCD.

Discrete groups having a finite number of elements are associated with discrete symmetry transformations, such as parity. For example, the discrete group corresponding to parity has only two elements:  $\mathbf{P}$  and  $\mathbf{P}^2 = e$ .

**hadron** The generic name for any particle which experiences the strong nuclear force.

**helicity** The projection of a particle’s spin along its direction of motion. See Section 12.3.

**Higgs boson** A hypothetical, spinless particle that plays an important role in the Glashow–Weinberg–Salam electroweak theory (and in other theories involving spontaneous symmetry breaking, e.g. GUTs).

**Higgs mechanism** A mechanism by which gauge bosons acquire mass through spontaneous symmetry breaking. In the Glashow–Weinberg–Salam electroweak model, for example, Higgs fields are introduced into the theory in a gauge-invariant way. However, the state of minimum energy breaks the local gauge symmetry, generating masses for the  $W^\pm$  and  $Z^0$  bosons, and giving rise to a real, observable Higgs boson,  $\phi'$ .

**hyperon** A baryon with non-zero strangeness.

**isotopic spin or isospin** A concept introduced by Heisenberg in 1932 to describe the charge independence of the strong nuclear force. Since the strong force cannot distinguish between a proton and a neutron, Heisenberg proposed that these particles were actually different states of a single particle – the nucleon. He argued that just as the electron comes in two different spin states, so the nucleon comes in two different ‘isospin’ states. So, isospin is a concept analogous to spin which is conserved by the strong interaction. The nucleon is an isospin- $\frac{1}{2}$  particle, and its third component of isospin determines whether we are talking about a proton ( $I_3 = +\frac{1}{2}$ ) or a neutron ( $I_3 = -\frac{1}{2}$ ).

**K meson or kaon** The name of particular spin-0 mesons with non-zero strangeness quantum numbers.

**Kelvin** Unit of absolute temperature.

**Lagrangian** A mathematical expression summarising the properties and interactions of a physical system. It is essentially the difference between the kinetic energy and potential energy of the system. Moreover, one can derive the system's dynamical equations of motion directly from the Lagrangian.

**lepton** The generic name for any spin- $\frac{1}{2}$  particle which does not feel the strong nuclear force. The six known leptons are the electron, the muon, the tau lepton, and their respective neutrinos. The name was originally coined to refer to light particles.

**lifetime** The time it takes for a sample of identical particles to decay to  $1/e$  of its initial population ( $e \approx 2.718$ ). A related concept is 'half-life', being the time it takes for the number of particles to halve. Half-life,  $\tau_{1/2}$ , is related to lifetime,  $\tau$ , by  $\tau_{1/2} = (\ln 2) \tau$ .

**M-theory** An as-yet-unknown theory in eleven space-time dimensions, believed to encompass the five known superstring theories in ten space-time dimensions.

**magnetic moment** A measure of the extent to which a physical system (e.g. an atom, or nucleus, or particle) behaves like a tiny magnet. It is generally measured in units of magnetons, i.e.  $e\hbar/2mc$ .

**magnetic monopole** A hypothetical particle that carries an isolated north or south magnetic pole. This is in contrast to magnets which are north-south-pole pairs. If magnetic monopoles exist, they must be very massive.

**mass-shell** In quantum mechanics, a particle's energy and momentum are essentially independent of each other. A particle is said to be 'on mass-shell' when its energy and momentum satisfy the formula from special relativity:

$$E^2 = p^2 c^2 + m_0^2 c^4,$$

which is necessary for it to exist as a real observable particle. Otherwise, the particle is 'virtual'.

**meson** The generic name for any strongly interacting particle with integer spin in units of  $\hbar$  (e.g. the pion and kaon).

**MSSM** The most simple supersymmetric extension of the Standard Model; see Chapter 44.

**muon ( $\mu$ )** A second-generation lepton. It is essentially a more massive electron.

**natural units** Units of length, time, mass, etc. in which the fundamental constants  $c$  (the speed of light),  $\hbar$  (Planck's constant) and  $k_B$  (Boltzmann's constant) are equal to unity. That is,  $c$ ,  $\hbar$  and  $k_B$  have the numerical value 1. (For example, if we measure length in light years and time in years, then  $c = 1$  light year per year.) The use of natural units allows these constants to be omitted from mathematical equations, leading to less-cluttered calculations. In natural units,  $E = mc^2$  becomes  $E = m$  and  $E = k_B T$  becomes  $E = T$ , so that both mass and temperature can be expressed in units of energy. (Of course, the correct factors of  $c$ ,  $\hbar$  and  $k_B$  must be inserted at the end of a calculation to obtain measurable quantities.)

**neutral-current reactions** Weak-interaction reactions in which no electric charge is exchanged between the colliding particles. Observation of such reactions in 1973 provided important support for the then-developing gauge theory of the weak interactions. We now know that these reactions are mediated by the exchange of a massive, neutral gauge boson – the  $Z^0$ .

**neutrino** An electrically neutral, nearly massless particle of spin  $-\frac{1}{2}$ , which interacts only by the weak force and gravity. It was first postulated by Pauli in 1930 to ensure conservation of energy and angular momentum in nuclear  $\beta$  decay. Three different types of neutrinos are known to exist corresponding to the three massive leptons:  $\nu_e$ ,  $\nu_\mu$  and  $\nu_\tau$ .

**neutron** One of the constituents of the atomic nucleus discovered in 1932. It is bound into atomic nuclei by the strong nuclear force. Free neutrons decay slowly via the weak nuclear force. Despite being electrically neutral, the neutron possesses both an electric dipole moment (as if it were made of positive and negative charges separated by a minute distance) and a magnetic moment, indicating some internal sub-structure.

**Noether's theorem** A mathematical theorem that states that for every symmetry of the Lagrangian of a physical system (i.e. for every set of transformations

under which the Lagrangian is invariant), there will be some quantity which is conserved by the dynamics of the system.

**nucleon** The generic name for the proton and the neutron.

**nucleosynthesis** The process by which the light elements (deuterium, helium, lithium) were synthesised in the first few minutes after the Big Bang. See Chapter 45.

**parity** The operation of spatial inversion, i.e.  $(x, y, z) \rightarrow (-x, -y, -z)$ .

**parton** A generic term used to describe any particle which may be present inside nucleons. It includes quarks, antiquarks and gluons.

**Pauli's exclusion principle** Two identical fermions cannot occupy the same quantum state (i.e. cannot have the same charge, spin, momentum, quantum numbers etc. within the same region of space).

**phase** A number (usually expressed as an angle between  $0^\circ$  and  $360^\circ$ ) which characterises a wave. The phase of a wave corresponds to the position in its cycle relative to an arbitrary reference point. It is a measure of how far away a wave crest or trough is.

**photon ( $\gamma$ )** The quantum of the electromagnetic field. It is the massless spin-1 gauge boson of QED. Virtual photons mediate the electromagnetic force between charged particles.

**Planck units** Fundamental units of length, time, mass, energy, etc. involving Planck's quantum constant,  $\hbar$ , Newton's gravitational constant,  $G$ , and the speed of light,  $c$ . As they incorporate both the quantum and gravitational constants, the Planck units play a key role in theories of quantum gravity. The Planck energy is  $10^{19}$  GeV.

**positron** The antiparticle of the electron, discovered by Anderson in 1934. It has the same mass and spin as the electron, but opposite charge and magnetic moment.

**propagator** The mathematical expression used to describe the propagation in space-time of virtual particles.

**proton** One of the constituents of the atomic nucleus. It is a spin  $-\frac{1}{2}$  particle carrying positive electric charge. The proton is the lightest baryon and, as a result, is the particle into which all other baryons eventually decay. It is believed to be absolutely stable, but certain theories (GUTs) predict it will decay very, very slowly.

**quantum chromodynamics (QCD)** The quantum field theory describing the interactions of quarks through the strong 'colour' field (whose quanta are gluons). QCD is a gauge theory with the non-Abelian gauge symmetry group  $SU(3)_C$ .

**quantum electrodynamics (QED)** The quantum field theory describing the interactions between electrically charged particles through the electromagnetic field (whose quantum is the photon). QED is a gauge theory with the Abelian gauge symmetry group  $U(1)$ .

**quantum field theory** The theory used to describe the physics of elementary particles. According to this theory, particles are localised quanta of these fields.

**quantum-mechanical amplitude** A mathematical quantity in quantum mechanics whose absolute square determines the probability of a particular process occurring.

**quantum theory** The theory used to describe physical systems which are very small, of atomic dimensions or less. A feature of the theory is that certain quantities (e.g. energy, angular momentum, light) can only exist in certain discrete amounts, called quanta.

**quark** A spin  $-\frac{1}{2}$  particle with fractional electric charge ( $+\frac{2}{3}$  or  $-\frac{1}{3}$ ). Baryons are composed of three (valence) quarks which are bound together by the strong colour forces, and mesons consist of a bound quark and antiquark. Quarks come in six flavours (u, d, s, c, b, t) and three colours (r, g, b).

**renormalisation** The process which ensures that the basic quantities in quantum field theory (e.g. in QED: the photon, electron and electric charge) are well defined and not infinite.

**resonance particles or resonances** Hadronic particles which exist for only a very brief time ( $10^{-23}$  seconds) before decaying into hadrons.

**r.f. (radio-frequency) power** Electromagnetic fields alternating at the frequencies of radio waves (up to  $10^{10}$  Hz), which can be used to accelerate charged particles in accelerators.

**scaling** The phenomenon observed in deep inelastic scattering, and predicted by James Björken, whereby the structure functions which describe the shape of the nucleon depend not on the energy or momentum involved in the reaction, but on some dimensionless ratio of the two. The structure functions are hence independent of any dimensional scale.

**see-saw mechanism** The mechanism by which neutrinos may acquire a very small mass due to new physics at very high energy scales; see Chapter 42.

**singularity** A point in space–time at which the space–time curvature and other physical quantities become infinite and the laws of physics break down.

**SLAC** The acronym for the Stanford Linear Accelerator Center at Stanford University in California, USA. It is distinguished by having a 2-mile-long linear accelerator in which electrons and positrons can be accelerated for subsequent injection into storage rings such as PEP, an  $e^+e^-$  collider which was commissioned in 1980. It was in the SPEAR rings at SLAC that the  $J/\psi$  (psi) meson and the  $\tau$  (tau) lepton were first observed in the mid-1970s. SLAC was also the home of the SLC (Stanford Linear Collider), consisting of the old linear accelerator together with two collider arcs and now hosts the B-factory experiment BaBar.

**spin** The intrinsic angular momentum possessed by many particles. It can be thought of as resulting from the particles spinning about an axis through their centres. In contrast to orbital angular momentum, spin is quantised in integer and half-integer units of  $\hbar$ . Fundamentally, spin describes how quantum fields transform under the transformations of special relativity.

**spontaneous symmetry breaking** Any situation in physics in which the ground state (i.e. the state of minimum energy) of a system has less symmetry than the system itself. For example, the state of minimum energy for an iron magnet is that in which the atomic

spins are all aligned in the same direction, giving rise to a net macroscopic magnetic field. By selecting a particular direction in space, the magnetic field has broken the rotational symmetry of the system. However, if the energy of the system is raised, the symmetry may be restored (e.g. the application of heat to an iron magnet destroys the magnetic field and restores rotational symmetry).

**Standard Model** This refers collectively to the successful theories of QCD and the Glashow–Weinberg–Salam electroweak model.

**strangeness** A quantum number associated with the strange quark. Strangeness is conserved by the strong nuclear force.

**string theory** A theory in which the fundamental constituents of matter are not particles but tiny one-dimensional objects, which we can think of as strings. These strings are so minute (only  $10^{-33}$  cm long) that, even at current experimental energies, they seem to behave just like particles. So, according to string theory, what we call ‘elementary particles’ are actually tiny strings, each of which is vibrating in a way characteristic of the particular ‘elementary particle’.

**supersymmetry** An extension of Lorentz symmetry, relating fermions and bosons. If supersymmetry is a true symmetry of nature, then every ‘ordinary’ particle has a corresponding ‘superpartner’ which differs in spin by half a unit.

**top** The sixth, and most massive, flavour of quark.

**vacuum** The state of minimum energy (or ground state) of a quantum theory. It is the quantum state in which no real particles are present. However, because of Heisenberg’s uncertainty principle, the vacuum is actually seething with *virtual* particles which constantly materialise, propagate a short distance and then disappear. See Section 4.9.

**virtual particles** Particles which take part in virtual processes. They are said to be ‘off mass-shell’, meaning that the relation  $E^2 = p^2c^2 + m_0^2c^4$  does *not* hold.

**virtual processes** Quantum-mechanical processes which do not conserve energy and momentum over microscopic timescales, in accordance with Heisenberg's uncertainty principle. These processes cannot be observed.

**wavefunction** A mathematical function  $\psi$  describing the behaviour of a particle according to quantum

mechanics. The wavefunction satisfies Schrödinger's wave equation. Furthermore, the probability of finding the particle at a particular point in space is given by the absolute square of the wavefunction.

**WIMP** Weakly interacting massive particle, hypothesised to make up dark matter.

APPENDIX C

***List of Symbols***

$a$	acceleration	$I$	isospin
$A$	electromagnetic gauge field	$I_3$	third component of isospin
$b$	bottom quark flavour; blue quark colour	$J$	spin
$B$	gauge field of $U(1)^W$	$J^W$	total weak current
$B$	baryon number	$K$	kaon
$c$	charm quark flavour	$K$	constant of proportionality
$c$	speed of light	$l_e, l_\mu$	electronic or muonic lepton
$C$	charge conjugation operator	$L$	Lagrangian; length; lepton number
$d$	down quark flavour	$L^W$	weak leptonic current
$D$	meson	$\mathcal{L}$	Lagrangian density
$e$	electron	$m$	mass; quantum-mechanical amplitude of sub-processes
$e$	electronic charge	$M$	total quantum-mechanical amplitude of a process
$E$	energy; electric field strength	$\mathbf{M}$	matrix
$f$	parton probability distribution	$n$	neutron
$F$	force	$n$	an integer
$F_1 F_2 F_3$	deep inelastic structure functions	$N$	nucleon
$g$	green quark colour; gluon	$N^*$	baryon resonances
$g$	$g$ -factor of the electron; general coupling constant	$N$	numbers of . . .
$G$	Newton's constant	$N_L, N_R$	number of left-spinning (right-spinning)
$G_F$	Fermi's coupling constant	$p$	proton
$\mathbf{G}$	Group	$p$	magnitude of momentum
$h$	hadron	$\mathbf{p}$	momentum vector
$h$	Planck's constant; strangeness-conserving weak hadronic current	$P$	probability (of occurrence of quantum-mechanical event); polarisation; fraction of nucleon momentum carried by the quark
$H$	magnetic field strength		
$H^W$	weak hadronic current		

<b>P</b>	parity operator	$\Delta$	infinitesimal amount of ...; delta baryon
q	quark	$\varepsilon$	(epsilon) small number
$q^2$	momentum transfer squared in deep inelastic scattering	$\zeta$	(zeta)
$Q$	electric charge (in units of $e$ )	$\eta$	(eta)
r	red quark colour	$\theta$	(theta) theta meson (archaic); an angle
$r$	magnitude of distance	$\theta_C$	the Cabbibo angle
<b>R</b>	Reggeon	$\theta_W$	the weak angle
$R$	vacuum expectation value of Higgs fields; ratio of cross-sections for $e^+e^-$ into hadrons to $e^+e^-$ into muon-antimuon pair	$\iota$	(iota)
s	strange quark flavour	$\kappa$	(kappa)
$s$	space-time interval; spin	$\lambda$	(lambda) wavelength
$s_z$	third component of spin	$\Lambda$	lambda hyperon
$s^\pm$	strangeness-changing weak hadronic current	$\mu$	(mu) muon
$S$	strangeness quantum number	$\nu$	(nu) neutrino; deep inelastic scattering energy transfer; frequency
$SU(2)$	special unitary groups of	$\xi$	(xi)
$SU(3)$	transformations of order 2, 3 and 5	$\Xi$	xi hyperon
$SU(5)$	respectively	$o$	(omicron)
t	top quark flavour	$\pi$	(pi) 3.141 5927; pion
$t$	time	$\rho$	(rho) rho meson; hypothetical hadronic isospin gauge particle
$T$	temperature	$\sigma$	(sigma) cross-section
<b>T</b>	time-reversal operator	$\sum$	summation over ...
u	up quark flavour	$\Sigma$	sigma hyperon
$u$	magnitude of velocity	$\tau$	(tau) tau heavy lepton; tau meson (archaic); lifetime, duration
$v$	magnitude of velocity	$\upsilon$	(upsilon)
<b>v</b>	velocity vector	$\Upsilon$	upsilon meson
$W^\pm$	W boson	$\phi$	(phi) Higgs particles
$x$	deep inelastic scattering variable; spatial separation	$\chi$	(chi) $\chi$ meson
<b>x</b>	position vector	$\psi$	(psi) quantum-mechanical wavefunction; psi meson
$X$	massive gauge bosons predicted by GUTs; unspecified final state of reaction	$\omega$	(omega) omega meson
$y$	fraction of energy transferred in deep inelastic scattering	$\Omega^-$	omega minus baryon
$Y$	hypercharge quantum number	=	is equal to
$Z^0$	Z boson	$\equiv$	is identical with
$\alpha$	(alpha) $\alpha$ radiation or particles (helium nuclei); electromagnetic fine structure constant	$\approx$	is approximately equal to
$\beta$	(beta) $\beta$ radiation or particles (electrons)	$a > b$	$a$ is greater than $b$
$\gamma$	(gamma) $\gamma$ radiation or particles (photons)	$a < b$	$a$ is less than $b$
$\Gamma$	Fermi's weak interaction couplings	$a \leq b$	$a$ is equal to, or less than, $b$
$\delta$	(delta) infinitesimal increment in variable in calculus	$a \geq b$	$a$ is equal to, or greater than, $b$
		$a \supset b$	$a$ contains $b$
		$\langle f M i \rangle$	initial $i$ and final $f$ states connected by a quantum-mechanical amplitude



## APPENDIX D

# *Bibliography*

This bibliography provides two types of references on most of the subjects dealt with in this book. The first, non-specialist, category includes articles and books accessible to the audience of books such as this one. The specialist category includes material for the professional student of physics and is generally aimed at the level of a third-year undergraduate or first-year post-graduate.

(ns) denotes the non-specialist category.

(s) denotes the specialist category.

### **Part I**

- (ns) *From X-rays to Quarks*, E. Segrè. Freeman, San Francisco, 1980.
- (ns) *Relativity*, Albert Einstein. Methuen, London, 1920.
- (ns) ‘The classical vacuum’, T. H. Boyer. *Scientific American*, **253** (2), 56–62, August 1985.
- (s) *Special Relativity*, A. P. French. Nelson, London, 1968.
- (s) *Simple Quantum Physics*, P. Landshoff & A. Metherell. Cambridge University Press, Cambridge, 1979.
- (s) *Relativistic Quantum Mechanics*, I. J. R. Aitchison. Macmillan, London, 1972.
- (s) *Quantum Field Theory*, F. Mandl & G. Shaw. Wiley, Chichester, 1984.

### **Part II and general**

- (ns) *The Forces of Nature*, P. C. W. Davies. Cambridge University Press, Cambridge, 1979.
- (ns) *The Particle Play*, J. C. Polkinghorne. Freeman, Oxford, 1979.
- (ns) *The Nature of Matter*, J. H. Mulvey (ed.). Clarendon, Oxford, 1981.
- (ns) *The Cosmic Onion*, F. Close. Heinemann, London, 1983.
- (ns) *The Particle Explosion*, F. Close, M. Marten & C. Sutton. Oxford University Press, Oxford, 1987.
- (s) *An Introduction to High-energy Physics*, D. H. Perkins. Addison-Wesley, Reading, MA (3rd edn), 1987. (An excellent textbook on most of the material mentioned in this book.)
- (s) *Symmetry Principles in Elementary Particle Physics*, W. M. Gibson & B. R. Pollard. Cambridge University Press, Cambridge, 1976.

### **Part III**

- (ns) ‘Resonance particles’, R. D. Hill. *Scientific American*, **208** (1), January 1963.
- (ns) ‘Strongly interacting particles’, G. F. Chew, M. Gell-Mann & A. H. Rosenfeld. *Scientific American*, **210** (2), February 1964.

- (ns) ‘Dual-resonance models of elementary particles’, J. H. Schwarz. *Scientific American*, **232** (5), 61–67, February 1975.
- (s) ‘High energy hadron collisions: a point of view’, J. P. Aurenche & J. E. Paton. *Reports on Progress in Physics*, **39** (2), February 1976.

#### Parts IV and V

- (ns) ‘The weak interactions’, S. B. Treiman. *Scientific American*, March 1959.
- (ns) ‘The two-neutrino experiment’, L. M. Lederman. *Scientific American*, **208** (3), March 1963.
- (ns) ‘Violations of symmetry in physics’, E. P. Wigner. *Scientific American*, **213** (6), 28–36, December 1965.
- (ns) ‘The detection of weak neutral currents’, D. B. Cline, A. K. Mann & C. Rubbia. *Scientific American*, **231** (6), 108–119, December 1974.
- (ns) ‘Weak interactions’, M. K. Gaillard. *Nature*, **279**, 585–589, June 1979.

#### Part VI

- (ns) ‘Unified theories of elementary-particle interaction’, S. Weinberg. *Scientific American*, **231** (1), 50–59, July 1974.
- (ns) ‘Gauge theories of the forces between elementary particles’, G. ’t Hooft. *Scientific American*, **242** (6), 104–138, June 1980.
- (ns) ‘The search for intermediate vector bosons’, D. B. Cline, C. Rubbia & S. van der Meer. *Scientific American*, **246** (3), 38–49, March 1982.
- (ns) *Story of the W and Z*, P. Watkins. Cambridge University Press, Cambridge, 1986.
- (s) *Gauge Theories in Particle Physics*, I. J. R. Aitchison & A. J. G. Hey. Hilger, Bristol, 1982.
- (s) *Gauge Theories of Weak Interactions*, J. C. Taylor. Cambridge University Press, Cambridge, 1978.

#### Part VII

- (s) ‘Inelastic lepton–nucleon scattering’, D. H. Perkins. *Reports on Progress in Physics*, **40**, 409–481, 1977.

- (s) *An Introduction to Quarks and Partons*, F. E. Close. Academic Press, London, 1979.

#### Part VIII

- (ns) ‘Quantum chromodynamics’, W. Marciano & H. Pagels. *Nature*, **279**, 479–483, June 1979.
- (ns) ‘Quark confinement’, R. L. Jaffe. *Nature*, **268**, 201–209, July 1977.
- (s) ‘Quantum chromodynamics’, W. Marciano & H. Pagels. *Physics Reports*, **36C**, 137–276, 1978.

#### Part IX

- (ns) ‘Electron–positron collisions’, A. M. Litke & R. Wilson. *Scientific American*, **229** (4), 104–113, October 1973.
- (ns) ‘Fundamental particles with charm’, R. F. Schwitters. *Scientific American*, **237** (4), 56–70, October 1977.
- (ns) ‘The upsilon particle’, L. M. Lederman. *Scientific American*, **239** (4), 60–68, October 1978.
- (ns) ‘The tau heavy lepton’, M. L. Perl. *Nature*, **275**, 273–277, September 1978.
- (ns) ‘Particles with naked beauty’, N. B. Mistry, R. A. Poling & E. H. Thorndike. *Scientific American*, **249** (1), 98–107, July 1983.
- (ns) ‘The Stanford linear collider’, J. R. Rees. *Scientific American*, **261** (4), 36–43, October 1989.

#### Part X

- (ns) *A Zeptospace Odyssey: A Journey into the Physics of the LHC*, G. F. Giudice. Oxford University Press, Oxford, 2009.
- (ns) *Voyage to the Heart of Matter: The Atlas Experiment at CERN*, A. Radevsky & E. Sanders. CERN, 2010.
- (ns) *LHC: Large Hadron Collider*, P. Ginter *et al.* Edition Lammerhuber, Baden, Austria, 2013.
- (ns) *Cern: How We Found the Higgs Boson*, M. Krause. World Scientific, Singapore, 2014.
- (s) *The Standard Model in a Nutshell*, D. Goldberg. Princeton University Press, Princeton, NJ, 2017.
- (s) *The Standard Model: A Primer*, C. Burgess & G. Moore. Cambridge University Press, Cambridge, 2012.

**Part XI**

- (ns) ‘Unified theory of elementary-particle forces’, H. Georgi & S. L. Glashow. *Physics Today*, **33** (9), September 1980.
- (ns) ‘A unified theory of elementary particles and forces’, H. Georgi. *Scientific American*, **244** (4), 40–55, April 1981.
- (s) *Grand Unified Theories*, G. G. Ross. Benjamin/Cummings, San Francisco, 1985.
- (ns) ‘Is nature supersymmetric?’, H. E. Haber & G. L. Kane. *Scientific American*, **254** (6), 42–50, June 1986.
- (ns) ‘Evidence of supersymmetry’, G. G. Ross. *Nature*, **352**, 21–22, 1991.
- (ns) *Beyond the Standard Model of Elementary Particle Physics*, Y. Nagashima. Wiley, New York, 2014.
- (s) *The Composite Nambu–Goldstone Higgs*, G. Panico & A. Wulzer. Springer Lecture Notes in Physics, Springer, Berlin, 2016.
- (s) *Supersymmetry in Particle Physics: An Elementary Introduction*, I. Aitchison. Cambridge University Press, Cambridge, 2012.
- (s) *Supersymmetry and String Theory: Beyond the Standard Model*, M. Dine. Cambridge University Press, Cambridge, 2016.

**Part XII**

- (ns) ‘The cosmic asymmetry between matter and antimatter’, F. Wilczek. *Scientific American*, **243** (6), 60–8, December 1980.
- (ns) *The First Three Minutes*, S. Weinberg. Basic Books, New York, 1977.
- (ns) ‘Cosmology and elementary particle physics’, M. S. Turner & D. N. Schramm. *Physics Today*, **32** (9), September 1979.
- (ns) ‘Particle accelerators test cosmological theory’, D. N. Schramm & G. Steigman. *Scientific American*, **258** (6), 44–50, June 1988.
- (ns) ‘Dark matter in the universe’, L. M. Krauss. *Scientific American*, **255** (6), 50–60, December 1986.
- (ns) ‘The search for dark matter in the laboratory’, B. Moskowit. *New Scientist*, 39–42, 15 April 1989.
- (ns) ‘How a supernova explodes’, H. A. Bethe & G. Brown. *Scientific American*, **252** (5), 40–8, May 1985.

- (ns) ‘The great supernova of 1987’, S. Woosley & T. Weaver. *Scientific American*, **261** (2), 24–32, August 1989.
- (ns) ‘The inflationary universe’, A. H. Guth & P. J. Steinhardt. *Scientific American*, **250** (5), May 1984.
- (ns) ‘Solitons’, C. Rebbi. *Scientific American*, **240** (2), 76–92, February 1979.
- (ns) ‘Superheavy magnetic monopoles’, R. A. Carrigan & W. P. Trower. *Scientific American*, **246** (4), 91–99, April 1982.

**Part XIII**

- (ns) *A Brief History of Time*, S. W. Hawking. Bantam Press, London, 1988.
- (ns) *Poetry of the Universe*, Robert Osserman. Anchor Books, New York, 1996.
- (ns) *Gravitational Waves*, Brian Clegg. Icon Books, London, 2018.
- (ns) *Gravity’s Kiss*, Harry Collins. S.MIT Press, Cambridge, MA, 2017.
- (s) ‘Observation of gravitational waves from a binary black hole merger’. *Physical Review Letters*, 061102, 2016.
- (ns /s) All news, references at [www.ligo.caltech.edu](http://www.ligo.caltech.edu)

**Part XIV**

- (ns) ‘Quantum gravity’, B. C. De Witt. *Scientific American*, **249** (6), 104–115, December 1983.
- (ns) ‘The quantum mechanics of black holes’, S. W. Hawking. *Scientific American*, **236** (1), 34–40, January 1977.
- (ns) ‘Black-hole thermodynamics’, J. D. Bekenstein. *Physics Today*, **33** (1), 24–31, January 1980.
- (ns) ‘The hidden dimensions of space–time’, D. Z. Freedman & P. van Nieuwenhuizen. *Scientific American*, **252** (3), 62–69, March 1985.
- (ns) ‘Superstrings’, M. B. Green. *Scientific American*, **255** (3), 44–56, September 1986.
- (ns) *Superstrings: A Theory of Everything?*, P. C. W. Davies & J. Brown. Cambridge University Press, Cambridge, 1988.
- (s) ‘Why superstrings?’, D. Bailin. *Contemporary Physics*, **30**, 237–250, 1989.
- (ns) *The Elegant Universe: Superstrings, Hidden Dimensions, and the Quest of the Ultimate Theory*, B. R. Greene. Norton, New York, 1999.

- (ns) *The Road to Reality: A Complete Guide to the Laws of the Universe*, R. Penrose. Jonathan Cape, London, 2004.
- (ns) ‘The myth of the beginning of time’, G. Veneziano. *Scientific American*, **16** (15), 72–81, May 2004.
- (ns) *The Edge of Physics*, Scientific American Special Edition, **13** (15), 2003.
- (ns) ‘A unified physics by 2050?’, S. Weinberg. *Scientific American*, **281** (6), 68–75, December 1999.
- (ns) ‘The theory formerly known as strings’, M. Duff. *Scientific American*, **278** (2), 64–69, February 1998.
- (ns) ‘The string theory landscape’, J. Polchinski & R. Bousso. *Scientific American*, **291** (3), 78–87, September 2004.
- (ns) ‘Out of the darkness’, G. Dvali. *Scientific American*, **290** (2), 68–75, February 2004.
- (ns) ‘The universe’s unseen dimensions’, N. Arkani-Hamed, S. Dimopoulos & G. Dvali, sidebar by G. P. Collins. *Scientific American*, **283** (2), 62–69, August 2000.
- (ns) ‘Strings in four dimensions’, J. Ellis. *Nature*, **329**, 488–489, 1987.
- (ns) ‘Physics: brane new worlds’, J. Gauntlett. *Nature*, **404**, 28–29, 2000.
- (ns) ‘High-energy physics: into the fifth dimension’, J. Maldacena. *Nature*, **423**, 695–696, 2003.
- (ns) ‘The holes are defined by the string’, E. Witten. *Nature*, **383**, 215–216, 1996.
- (ns) ‘From “not wrong” to (maybe right)’, F. Wilczek. *Nature*, **428**, 261, 2004.
- (s) *String Theory*, J. Polchinski. Cambridge University Press, Cambridge, 2004.
- (s) *A First Course in String Theory*, B. Zwiebach. Cambridge University Press, Cambridge, 2009.
- (ns) *Why String Theory*, Joseph Conlon. CRC Press, London, 2016.

### Part XV

- (ns) *Fashion, Faith and Fantasy in the New Physics of the Universe*, Roger Penrose. Princeton University Press, Princeton, NJ, 2016.
- (ns) *Reality is Not What It Seems*, Carlo Rovelli. Penguin Books, London, 2017.
- (s) ‘The future of particle physics’, Ian Shipsey. *The 38th International conference on High Energy Particle Physics, Chicago USA 2016* arxiv.org/1707.03711
- (s) ‘European strategy for particle physics’ at [www.europeanstrategy.cern](http://www.europeanstrategy.cern)

APPENDIX E

***Elementary Particle Data***

An up-to-date catalogue of elementary particle data can be found in *The Review of Particle Physics*, at the website <http://pdg.lbl.gov/pdg.html>

# *Name Index*

- Amber, E., 71  
Anderson, C.D., 28, 52, 303
- Bahcall, J., 214  
Balmer, J., 18  
Barrish, B., 259  
Becquerel, A.H., 5  
Berners-Lee, T., 201  
Bohr, N., 17–19, 21, 23, 24  
Born, M., 23  
Bose, S., 25  
Bjorken, J., 126  
Burnell, J., 257  
Butler, C. C., 56
- Cailliau, R., 203  
Casimir H., 33  
Chadwick, J., 43  
Christenson, J.H., 83, 84  
Cline, D., 116  
Cronin, J. W., 83, 196  
Curie, P. & M., 5
- Davies, R., 91, 213  
Davisson, C., 19  
de Broglie, L., 30  
Dirac, P. A. M., 25–30, 47, 192,  
219, 244, 290  
Dyson, F., 32
- Eddington, A., 256  
Einstein, A., 8, 11, 16–18, 25,  
26, 30, 40, 41, 47  
Englert, F., 291
- Fairbank, W.M., 149  
Fermi, E., 46, 58, 75, 77–79, 89,  
92, 94
- Feynman, R. P., 31, 32, 75, 126  
Fitch, V.L., 83, 196  
Fitzgerald, G. F., 10  
Frauenfelder, F., 77  
Friedman, J., 126  
Fritzsche, H., 145
- Geiger, H., 6, 145  
Gell-Mann M., 56, 58, 65, 66,  
75, 145  
Germer, L., 19  
Glashow, S., 108, 114  
Goldhaber, M., 77  
Goldstone, J., 106  
Goudsmit, S., 23  
Green, M., 275  
Greenberg, O. W., 146  
Gross, D. J., 141, 156, 157
- Han, M-Y., 146  
Heisenberg, W., 26, 55  
Hewish, A., 257  
Higgs, P., 291  
Hubble, E., 235  
Hulse, A. R., 256
- Iliopoulos, J., 114
- Kajita, T., 291  
Kendall, H., 126  
Kobayashi, M., 291  
Konopinski, E. J., 79
- Lattes, C., 53  
Lederman, L., 81, 178  
Lee, T. D., 71  
Lenard, P.E., 17
- Leutwyler, H., 145  
Llewellyn Smith, C. H., 141  
Lorentz, H. A., 10, 11  
Lyman, T., 18
- Mahmoud, H. M., 79  
Maiani, L., 114  
Marsden, E., 6, 145  
Maskawa, T., 291  
McDonald, B. A., 291  
Michelson, A. A., 10  
Mills, R. L., 103  
Morley, E. W., 10
- Nambu, Y., 145, 146  
Ne'eman, 66  
Ne'eman, Y., 65  
Nelson, E., 202  
Newton, I., 47  
Nishijima, K., 56  
Noether, E., 47, 48
- Occhialini, G., 53  
Oddone, P., 197
- Paschen, F., 18  
Pati, J., 149  
Pauli, W., 46  
Penrose, Sir Roger, 264  
Perl, M., 179, 180  
Perlmutter, S., 257  
Planck, M., 7, 16–18, 30  
Poincaré, H., 48  
Poltzer, H. D., 156  
Powell, C., 53
- Röntgen, W., 5  
Reines, F., 46  
Reiss, A., 257
- Richter, B., 172  
Rochester, G. D., 56  
Rubbia, C., 116, 121  
Rutherford, E., 145  
Ryle, M., 257
- Salam, A., 108, 115, 149  
Schmidt, B., 257  
Schrodinger, E., 26  
Schwartz, M., 81  
Schwinger, J., 32  
Shaw, R., 103  
Steinberger, J., 81
- 't hooft, G., 108, 111  
Taylor, H., J., 256  
Taylor, R., 126  
Thomson, G. P., 4, 6, 19  
Thorne, K., 291  
Ting, S., 173, 178  
Tomonaga, S., 32  
Turlay, R., 83, 196
- van der Meer, S., 121  
Von Laue, M., 5
- Weinberg, S., 108, 115  
Weiss, R., 259  
Wilczek, F., 156, 157  
Wilson, C. T. R., 5  
Wu, C.S., 71, 72
- Yang, C. N., 71, 103, 206  
Yukawa, H., 45, 52, 53
- Zweig, G., 66

# Subject Index

- Advanced Proton Driven Plasma  
Wakefield Acceleration  
Experiment (AWAKE), 286
- Angular momentum, 18, 46
- Astronomy, multi-messenger,  
263
- Asymptotic freedom  
introduction, 154
- Atom  
atomic spectra, 18  
Bohr's model, 18  
Rutherford model, 6  
Thomson model, 4
- B-physics, 197
- Bhabba scattering, 168
- Brookhaven accelerator, 56
- Bubble chambers, 63
- Casimir effect, 33
- Charge conjugation symmetry,  
82
- Charm  
incorporation of hadrons-,  
113–115  
introduction, 51
- Charmonium, 176
- Chirality, 76
- Chromo-static force, 150
- Coordinates, 8
- Collaboration, Global Argon  
Dark Matter, 288
- Collider, Chinese Electron  
Positron, 285
- Collider, Electron-Ion, 286
- Collider, Future Circular,  
286
- Collider, International Linear,  
285
- Colour  
confinement of, 149  
evidence for, 146  
introduction of, 145  
invisible, 147–149
- Coulomb's Law, 42
- CP  
symmetry, 50  
violation, 82–84
- CPT theorem, 50  
symmetry, 50
- Current–current theory  
hadronic current, 92  
leptonic current, 87  
of the weak force, 87
- Deep inelastic scattering  
electron–nucleon, 127  
introduction to, 125, 131  
structure functions in,  
127–128
- Deuteron, 54
- Diffraction  
electron waves, 19  
introduction, 5  
light as a wave, 10
- Dirac equation, 26–27
- DUNE, 285
- Einstein's field equations of  
gravity, 40, 256
- Eigenstate, 82
- Electromagnetism, 7, 41
- Electrons  
collisions with positrons, 167  
discovery, 4  
electron definition, 4  
electron number, 103  
gyromagnetic ratio, 27  
magnetic moment, 24  
polarisation in decay, 76–77  
scattering off nucleons,  
127–128  
scattering with neutrinos, 91  
spin, 23–24
- Electroweak theory, 116
- Exclusion principle  
bosons, 25  
Pauli's formulation of, 24–25  
role in QCD, 145–146  
role in the nucleus, 45
- Experiment, Michelson–Morley,  
10
- Fermilab, 173, 178, 181, 285
- Field theory, quantum, 29–30
- Fine structure constant, 42
- Fourier sum, 30
- Gamma rays, 42, 264, 288
- Glashow–Weinberg–Salam  
model  
consequences of, 112–115  
formulation of, 108–111
- Glueballs, 153
- Graviton, 41
- Gravity, 39–41, 228
- Group theory, 47
- Gauge theory  
formulation of QED, 101  
generalised invariance,  
102–103  
introduction, 101  
of the weak interactions, 99  
and weak force, 103
- Hadron  
collisions in QCD, 162  
dynamics, 65–66  
introduction, 66
- Half-life, 45
- Helicity  
introduction-, 76  
neutrino-, 77
- Higgs mechanism, particles, 107
- Hyper-charge Y, 59
- Hyperon, 59
- Ice Cube Neutrino Observatory  
(ICE CUBE), 288
- Infrared slavery, 133, 160
- Isotopic spin, 55
- KamioKande, 215, 287
- Kaons, 57
- Laboratory, Sudbury Neutrino  
Observatory (SNOLAB),  
288
- Lagrangian, 30
- Lamb shift, 34
- Large Hadron Collider, 107,  
190, 200
- Laser Interferometer GW  
Observatory (LIGO), 259
- LEP collider, 116, 168
- Lepton  
definition, 52  
number, 55
- Mass- shell, 32
- Matter waves, 20–21
- Maxwell's equations, 42
- mechanics, relativistic field, 30
- Mercury, perihelion of orbit, 255
- Model, Rutherford Atomic, 6

- Muon  
   decay, 75  
   discovery of, 52  
   muon number, 80
- Neutral currents  
   and charm, 114  
   discovery, 112  
   introduction, 88
- Neutrino  
   helicity, 77  
   oscillation, 214  
   Pauli's hypothesis, 45  
   types, 195
- Neutron  
   discovery, 43  
   Noether's theorem, 48  
   quark structure of, 148
- Neutron star, 257
- Particle-wave duality, 19
- Periodic table, Mendeleef's, 18
- Perturbation theory of quantum fields, 30–32
- Photoelectric effect, 17
- Photon, 17
- Pi-Meson, 52
- Planck's constant, 16
- Poincare group transformations, 48
- Positron, 28
- Propagator, 34
- Proton  
   decay, 222  
   quark structure of, 66, 139
- Psi-meson, 172
- Pulsar binary period,  
   Hulse–Taylor variation, 256
- Quantum  
   chromodynamics (QCD), 145  
   electrodynamics (QED), 34  
   field theory amplitude, 31  
   gravity, 41  
   introduction of, 16
- mechanics, 16–25  
   mechanics, relativistic field, 30  
   vacuum, 33
- Quarks  
   and partons, 126  
   bottom quark, 178  
   charmed quark, 171–177  
   colour, 147  
   confinement, 160–163  
   flavours, 148  
   forces, 162  
   free, 149  
   interquark forces, 131  
   introduction of, 65  
   line diagrams, 174  
   model, 66, 138  
   top quark, 179  
   weak quantum numbers, 109
- Radiation, Hawking, 258
- Radioactivity  
    $\alpha$ -decay, 5  
    $\beta$ -decay, 6  
    $\gamma$ -decay, 6  
   discovery, 5
- Radius, Schwarzschild, 258
- Relativistic invariants, 15
- Relativity  
   Galilean, 8  
   general theory, 40, 223, 236  
   theory, 8–15
- Renormalisation  
   introduction, 32  
   of Glashow–Weinberg–salam model, 111  
   physical picture, 155
- Rules, 31
- Second quantisation, 30
- Solar neutrino problem, 213
- Space-time diagrams, 14
- Speed of light  
   constancy, 7  
   introduction, 8
- limiting-velocity, 11
- Spinor, 27
- Spontaneous symmetry breaking, 105–107, 109
- SPS ring, 119
- Standard model  
   consistency, 186  
   introduction, 185  
   summary, 185
- Stanford Linear Accelerator Center (SLAC), 68, 128, 169, 189, 198, 285
- Storage rings, 168
- Strangeness, 51, 56
- Strong nuclear forces  
   description, 39  
   internal symmetry, 65  
   nuclear stability, 39
- Structure functions  
   description in QCD, 157  
   in deep elastic scattering, 127  
   in quark model, 138  
   sum rules, 141
- Sudbury Neutrino Oscillation experiment (SNO), 215
- Supernova, 257
- Symmetry  
   and conservation laws, 47–50  
   broken, 51  
   charge conjugation, 49  
   discrete, 48  
   dynamical, 50  
   group theory of, 47  
   internal, 51  
   space-time, 47
- Synchrotron radiation, 168
- Theorem, No Hair, 258
- Time dilation, 13, 181
- Time reversal, 50
- Uncertainty Principle  
   Heisenberg's formulation, 21  
   role in resonances, 63
- and Yukawa's pion, 52
- Upsilon meson, 178
- V-Particles, 56
- Vacuum state  
   fluctuations, 33  
   quantum vacuum, 33  
   van der Waals forces, 162
- Vectors  
   axial, 92  
   four-vectors, 14  
   three-vectors, 14
- VIRGO, 259
- Virtual particles, 32, 223
- Virtual processes, 32, 272
- W boson  
   detection of, 116–121  
   difficulties in theory of, 99  
   introduction, 94  
   mass prediction, 110  
   propagation, 99
- Wavefunction  
   interpretation, 22  
   introduction, 20  
   relation to quantum field, 29  
   Schrodinger's, 20–21
- Weak nuclear forces  
   current–current theory, 87, 89  
   description, 45  
   Fermi's theory, 75  
   weak hypercharge, 109  
   weak isospin, 103, 109
- Weakly interacting massive particles, (WIMPS), 288
- White dwarf, 257
- X-rays  
   chamber, cloud, 5
- $z^0$  boson  
   detection of, 116  
   introduction, 110  
   mass predictions, 110  
   role in  $e^+e^-$ , 170
- Zero-point vibrations, energy, 33



