

Molecular Modelling

PRINCIPLES AND APPLICATIONS

Second edition

Andrew R. Leach

Glaxo Wellcome Research and Development

Prentice
Hall

An imprint of Pearson Education

Harlow, England London New York Reading, Massachusetts San Francisco Toronto Don Mills, Ontario Sydney
Tokyo Singapore Hong Kong Seoul Taipei Cape Town Madrid Mexico City · Amsterdam · Munich Paris Milan

Pearson Education Limited

Edinburgh Gate
Harlow
Essex CM20 2JE
England

and Associated Companies around the world

Visit us on the World Wide Web at
www.pearsoned.com

First published under the Longman imprint 1996
Second edition 2001

© Pearson Education Limited 1996, 2001

The right of Andrew R. Leach to be identified as the author of
this Work has been asserted by him in accordance with
the Copyright, Designs and Patents Act 1988

All rights reserved. No part of this publication may be reproduced, stored
in a retrieval system, or transmitted in any form or by any means, electronic,
mechanical, photocopying, recording or otherwise without either the prior
written permission of the publisher or a licence permitting restricted copying
in the United Kingdom issued by the Copyright Licensing Agency Ltd,
90 Tottenham Court Road, London W1P 0LP.

ISBN 0-582-38210-6

British Library Cataloguing-in-Publication Data

A catalogue record for this book can be obtained from the British Library

Library of Congress Cataloging-in-Publication Data

Leach, Andrew R.

Molecular modelling principles and applications / Andrew R. Leach. – 2nd ed
p. cm.

Includes bibliographical references and index

ISBN 0-582-38210-6

1 Molecular structure–Computer simulation 2 Molecules–Models–Computer
simulation I. Title.

QD480.L43 2001

541 2'2'0113–dc21

00-046480

10 9 8 7 6 5 4 3 2 1
05 04 03 02 01

Top right-hand cover image © American Institute of Physics

Typeset by 60

Printed in Great Britain by Henry Ling Ltd,
at the Dorset Press, Dorchester, Dorset

Contents

Preface to the Second Edition	xiii
Preface to the First Edition	xv
Symbols and Physical Constants	xvii
Acknowledgements	xxi
1 Useful Concepts in Molecular Modelling	1
1.1 Introduction	1
1.2 Coordinate Systems	2
1.3 Potential Energy Surfaces	4
1.4 Molecular Graphics	5
1.5 Surfaces	6
1.6 Computer Hardware and Software	8
1.7 Units of Length and Energy	9
1.8 The Molecular Modelling Literature	9
1.9 The Internet	9
1.10 Mathematical Concepts	10
Further Reading	24
References	24
2 An Introduction to Computational Quantum Mechanics	26
2.1 Introduction	26
2.2 One-electron Atoms	30
2.3 Polyelectronic Atoms and Molecules	34
2.4 Molecular Orbital Calculations	41
2.5 The Hartree-Fock Equations	51
2.6 Basis Sets	65
2.7 Calculating Molecular Properties Using <i>ab initio</i> Quantum Mechanics	74
2.8 Approximate Molecular Orbital Theories	86
2.9 Semi-empirical Methods	86
2.10 Hückel Theory	99
2.11 Performance of Semi-empirical Methods	102
Appendix 2.1 Some Common Acronyms Used in Computational Quantum Chemistry	104
Further Reading	105
References	105

3 Advanced <i>ab initio</i> Methods, Density Functional Theory and Solid-state Quantum Mechanics	108
3.1 Introduction	108
3.2 Open-shell Systems	108
3.3 Electron Correlation	110
3.4 Practical Considerations When Performing <i>ab initio</i> Calculations	117
3.5 Energy Component Analysis	122
3.6 Valence Bond Theories	124
3.7 Density Functional Theory	126
3.8 Quantum Mechanical Methods for Studying the Solid State	138
3.9 The Future Role of Quantum Mechanics: Theory and Experiment Working Together	160
Appendix 3.1 Alternative Expression for a Wavefunction Satisfying Bloch's Function	161
Further Reading	161
References	162
 4 Empirical Force Field Models: Molecular Mechanics	 165
4.1 Introduction	165
4.2 Some General Features of Molecular Mechanics Force Fields	168
4.3 Bond Stretching	170
4.4 Angle Bending	173
4.5 Torsional Terms	173
4.6 Improper Torsions and Out-of-plane Bending Motions	176
4.7 Cross Terms: Class 1, 2 and 3 Force Fields	178
4.8 Introduction to Non-bonded Interactions	181
4.9 Electrostatic Interactions	181
4.10 Van der Waals Interactions	204
4.11 Many-body Effects in Empirical Potentials	212
4.12 Effective Pair Potentials	214
4.13 Hydrogen Bonding in Molecular Mechanics	215
4.14 Force Field Models for the Simulation of Liquid Water	216
4.15 United Atom Force Fields and Reduced Representations	221
4.16 Derivatives of the Molecular Mechanics Energy Function	225
4.17 Calculating Thermodynamic Properties Using a Force Field	226
4.18 Force Field Parametrisation	228
4.19 Transferability of Force Field Parameters	231
4.20 The Treatment of Delocalised π Systems	233
4.21 Force Fields for Inorganic Molecules	234
4.22 Force Fields for Solid-state Systems	236
4.23 Empirical Potentials for Metals and Semiconductors	240
Appendix 4.1 The Interaction Between Two Drude Molecules	246
Further Reading	247
References	247

5 Energy Minimisation and Related Methods for Exploring the Energy Surface	253
5.1 Introduction	253
5.2 Non-derivative Minimisation Methods	258
5.3 Introduction to Derivative Minimisation Methods	261
5.4 First-order Minimisation Methods	262
5.5 Second Derivative Methods: The Newton–Raphson Method	267
5.6 Quasi-Newton Methods	268
5.7 Which Minimisation Method Should I Use?	270
5.8 Applications of Energy Minimisation	273
5.9 Determination of Transition Structures and Reaction Pathways	279
5.10 Solid-state Systems: Lattice Statics and Lattice Dynamics	295
Further Reading	300
References	301
6 Computer Simulation Methods	303
6.1 Introduction	303
6.2 Calculation of Simple Thermodynamic Properties	307
6.3 Phase Space	312
6.4 Practical Aspects of Computer Simulation	315
6.5 Boundaries	317
6.6 Monitoring the Equilibration	321
6.7 Truncating the Potential and the Minimum Image Convention	324
6.8 Long-range Forces	334
6.9 Analysing the Results of a Simulation and Estimating Errors	343
Appendix 6.1 Basic Statistical Mechanics	347
Appendix 6.2 Heat Capacity and Energy Fluctuations	348
Appendix 6.3 The Real Gas Contribution to the Virial	349
Appendix 6.4 Translating Particle Back into Central Box for Three Box Shapes	350
Further Reading	351
References	351
7 Molecular Dynamics Simulation Methods	353
7.1 Introduction	353
7.2 Molecular Dynamics Using Simple Models	353
7.3 Molecular Dynamics with Continuous Potentials	355
7.4 Setting up and Running a Molecular Dynamics Simulation	364
7.5 Constraint Dynamics	368
7.6 Time-dependent Properties	374
7.7 Molecular Dynamics at Constant Temperature and Pressure	382
7.8 Incorporating Solvent Effects into Molecular Dynamics: Potentials of Mean Force and Stochastic Dynamics	387
7.9 Conformational Changes from Molecular Dynamics Simulations	392
7.10 Molecular Dynamics Simulations of Chain Amphiphiles	394

Appendix 7.1 Energy Conservation in Molecular Dynamics	405
Further Reading	406
References	406
8 Monte Carlo Simulation Methods	410
8.1 Introduction	410
8.2 Calculating Properties by Integration	412
8.3 Some Theoretical Background to the Metropolis Method	414
8.4 Implementation of the Metropolis Monte Carlo Method	417
8.5 Monte Carlo Simulation of Molecules	420
8.6 Models Used in Monte Carlo Simulations of Polymers	423
8.7 'Biased' Monte Carlo Methods	432
8.8 Tackling the Problem of Quasi-ergodicity: J-walking and Multicanonical Monte Carlo	433
8.9 Monte Carlo Sampling from Different Ensembles	438
8.10 Calculating the Chemical Potential	442
8.11 The Configurational Bias Monte Carlo Method	443
8.12 Simulating Phase Equilibria by the Gibbs Ensemble Monte Carlo Method	450
8.13 Monte Carlo or Molecular Dynamics?	452
Appendix 8.1 The Marsaglia Random Number Generator	453
Further Reading	454
References	454
9 Conformational Analysis	457
9.1 Introduction	457
9.2 Systematic Methods for Exploring Conformational Space	458
9.3 Model-building Approaches	464
9.4 Random Search Methods	465
9.5 Distance Geometry	467
9.6 Exploring Conformational Space Using Simulation Methods	475
9.7 Which Conformational Search Method Should I Use? A Comparison of Different Approaches	476
9.8 Variations on the Standard Methods	477
9.9 Finding the Global Energy Minimum: Evolutionary Algorithms and Simulated Annealing	479
9.10 Solving Protein Structures Using Restrained Molecular Dynamics and Simulated Annealing	483
9.11 Structural Databases	489
9.12 Molecular Fitting	490
9.13 Clustering Algorithms and Pattern Recognition Techniques	491
9.14 Reducing the Dimensionality of a Data Set	497
9.15 Covering Conformational Space: Poling	499
9.16 A 'Classic' Optimisation Problem: Predicting Crystal Structures	501

Further Reading	505
References	506
10 Protein Structure Prediction, Sequence Analysis and Protein Folding	509
10.1 Introduction	509
10.2 Some Basic Principles of Protein Structure	513
10.3 First-principles Methods for Predicting Protein Structure	517
10.4 Introduction to Comparative Modelling	522
10.5 Sequence Alignment	522
10.6 Constructing and Evaluating a Comparative Model	539
10.7 Predicting Protein Structures by 'Threading'	545
10.8 A Comparison of Protein Structure Prediction Methods: CASP	547
10.9 Protein Folding and Unfolding	549
Appendix 10.1 Some Common Abbreviations and Acronyms Used in Bioinformatics	553
Appendix 10.2 Some of the Most Common Sequence and Structural Databases Used in Bioinformatics	555
Appendix 10.3 Mutation Probability Matrix for 1 PAM	556
Appendix 10.4 Mutation Probability Matrix for 250 PAM	557
Further Reading	557
References	558
11 Four Challenges in Molecular Modelling: Free Energies, Solvation, Reactions and Solid-state Defects	563
11.1 Free Energy Calculations	563
11.2 The Calculation of Free Energy Differences	564
11.3 Applications of Methods for Calculating Free Energy Differences	569
11.4 The Calculation of Enthalpy and Entropy Differences	574
11.5 Partitioning the Free Energy	574
11.6 Potential Pitfalls with Free Energy Calculations	577
11.7 Potentials of Mean Force	580
11.8 Approximate/'Rapid' Free Energy Methods	585
11.9 Continuum Representations of the Solvent	592
11.10 The Electrostatic Contribution to the Free Energy of Solvation: The Born and Onsager Models	593
11.11 Non-electrostatic Contributions to the Solvation Free Energy	608
11.12 Very Simple Solvation Models	609
11.13 Modelling Chemical Reactions	610
11.14 Modelling Solid-state Defects	622
Appendix 11.1 Calculating Free Energy Differences Using Thermodynamic Integration	630
Appendix 11.2 Using the Slow Growth Method for Calculating Free Energy Differences	631

Appendix 11.3 Expansion of Zwanzig Expression for the Free Energy Difference for the Linear Response Method	631
Further Reading	632
References	633
12 The Use of Molecular Modelling and Chemoinformatics to Discover and Design New Molecules	640
12.1 Molecular Modelling in Drug Discovery	640
12.2 Computer Representations of Molecules, Chemical Databases and 2D Substructure Searching	642
12.3 3D Database Searching	647
12.4 Deriving and Using Three-dimensional Pharmacophores	648
12.5 Sources of Data for 3D Databases	659
12.6 Molecular Docking	661
12.7 Applications of 3D Database Searching and Docking	667
12.8 Molecular Similarity and Similarity Searching	668
12.9 Molecular Descriptors	668
12.10 Selecting 'Diverse' Sets of Compounds	680
12.11 Structure-based <i>De Novo</i> Ligand Design	687
12.12 Quantitative Structure–Activity Relationships	695
12.13 Partial Least Squares	706
12.14 Combinatorial Libraries	711
Further Reading	719
References	720
Index	727

Preface to the Second Edition

The impetus for this second edition is a desire to include some of the new techniques that have emerged in recent years and also extend the scope of the book to cover certain areas that were under-represented (even neglected) in the first edition. In this second volume there are three topics that fall into the first category (density functional theory, bioinformatics/protein structure analysis and chemoinformatics) and one main area in the second category (modelling of the solid state). In addition, of course, a new edition provides an opportunity to take a critical view of the text and to re-organise and update the material. Thus whilst much remains from the first edition, and this second book follows much the same path through the subject, readers familiar with the first edition will find some changes which I hope they will agree are for the better.

As with the first edition we initially consider quantum mechanics, but this is now split into two chapters. Thus Chapter 2 provides an introduction to the *ab initio* and semi-empirical approaches together with some examples of the uses of quantum mechanics. Chapter 3 covers more advanced aspects of the *ab initio* approach, density functional theory and the particular problems of the solid state. Molecular mechanics is the subject of Chapter 4 and then in Chapter 5 we consider energy minimisation and other 'static' techniques. Chapters 6, 7 and 8 deal with the two main simulation methods (molecular dynamics and Monte Carlo). Chapter 9 is devoted to the conformational analysis of 'small' molecules but also includes some topics (e.g. cluster analysis, principal components analysis) that are widely used in informatics. In Chapter 10 the problems of protein structure prediction and protein folding are considered; this chapter also contains an introduction to some of the more widely used methods in bioinformatics. In Chapter 11 we draw upon material from the previous chapters in a discussion of free energy calculations, continuum solvent models, and methods for simulating chemical reactions and defects in solids. Finally, Chapter 12 is concerned with modelling and chemoinformatics techniques for discovering and designing new molecules, including database searching, docking, *de novo* design, quantitative structure-activity relationships and combinatorial library design.

As in the first edition, the inexorable pace of change means that what is currently considered 'cutting edge' will soon become routine. The examples are thus chosen primarily because they illuminate the underlying theory rather than because they are the first application of a particular technique or are the most recent available. In a similar vein, it is impossible in a volume such as this to even attempt to cover everything and so there are undoubtedly areas which are under-represented. This is not intended to be a definitive historical account or a review of the current state-of-the-art. Thus, whilst I have tried to include many literature references it is possible that the invention of some technique may appear to be incorrectly attributed or a 'classic' application may be missing. A general guiding principle has been

to focus on those techniques that are in widespread use rather than those which are the province of one particular research group. Despite these caveats I hope that the coverage is sufficient to provide a solid introduction to the main areas and also that those readers who are 'experts' will find something new to interest them.

A Companion Web Site accompanies *Molecular Modelling: Principles and Applications, Second Edition* by Andrew Leach



Visit the *Molecular Modelling* Companion Web Site at www.booksites.net/leach

The website contains general information about the book, up-to-date hyperlinks to related chemistry sources on the web, reference copies of appendices of relevant acronyms, and twenty-six full screen, full-colour graphical representations of molecular structures.

Preface to the First Edition

Molecular modelling used to be restricted to a small number of scientists who had access to the necessary computer hardware and software. Its practitioners wrote their own programs, managed their own computer systems and mended them when they broke down. Today's computer workstations are much more powerful than the mainframe computers of even a few years ago and can be purchased relatively cheaply. It is no longer necessary for the modeller to write computer programs as software can be obtained from commercial software companies and academic laboratories. Molecular modelling can now be performed in any laboratory or classroom.

This book is intended to provide an introduction to some of the techniques used in molecular modelling and computational chemistry, and to illustrate how these techniques can be used to study physical, chemical and biological phenomena. A major objective is to provide, in one volume, some of the theoretical background to the vast array of methods available to the molecular modeller. I also hope that the book will help the reader to select the most appropriate method for a problem and so make the most of his or her modelling hardware and software. Many modelling programs are extremely simple to use and are often supplied with seductive graphical interfaces, which obviously helps to make modelling techniques more accessible, but it can also be very easy to select a wholly inappropriate technique or method.

Most molecular modelling studies involve three stages. In the first stage a model is selected to describe the intra- and inter-molecular interactions in the system. The two most common models that are used in molecular modelling are quantum mechanics and molecular mechanics. These models enable the energy of any arrangement of the atoms and molecules in the system to be calculated, and allow the modeller to determine how the energy of the system varies as the positions of the atoms and molecules change. The second stage of a molecular modelling study is the calculation itself, such as an energy minimisation, a molecular dynamics or Monte Carlo simulation, or a conformational search. Finally, the calculation must be analysed, not only to calculate properties but also to check that it has been performed properly.

The book is organised so that some of the techniques discussed in later chapters refer to material discussed earlier, though I have tried to make each chapter as independent of the others as possible. Some readers may therefore be pleased to know that it is not essential to completely digest the chapters on quantum mechanics and molecular mechanics in order to read about methods for searching conformational space! Readers with experience in one or more areas may, of course, wish to be more selective.

I have tried to provide as much of the underlying theory as seems appropriate to enable the reader to understand the fundamentals of each method. In doing so I have assumed some background knowledge of quantum mechanics, statistical mechanics, conformational analysis and mathematics. A reader with an undergraduate degree in chemistry should

have covered this material, which should also be familiar to many undergraduates in the final year of their degree course. Full discussion can be found in the suggestions for further reading at the end of each chapter. I have also attempted to provide a reasonable selection of original references, though in a book of this scope it is obviously impossible to provide a comprehensive coverage of the literature. In this context, I apologise in advance if any technique is inappropriately attributed.

The range of systems that can be considered in molecular modelling is extremely broad, from isolated molecules through simple atomic and molecular liquids to polymers, biological macromolecules such as proteins and DNA and solids. Many of the techniques are illustrated with examples chosen to reflect the breadth of applications. It is inevitable that, for reasons of space, some techniques must be dealt with in a rudimentary fashion (or not at all), and that many interesting and important applications cannot be described. Molecular modelling is a rapidly developing discipline and has benefited from the dramatic improvements in computer hardware and software of recent years. Calculations that were major undertakings only a few years ago can now be performed using personal computing facilities. Thus, examples used to indicate the 'state of the art' at the time of writing will invariably be routine within a short time.

Symbols and Physical Constants

This list contains the most frequently used symbols and physical constants ordered according to approximate appearance in the text.

λ	Lagrange multiplier
r, θ, ϕ	spherical polar coordinates
$\mathbf{i}, \mathbf{j}, \mathbf{k}$	orthogonal unit vectors along x, y, z axes
ϕ, θ, ψ	Euler angles
$\langle x \rangle$ or \bar{x}	arithmetic mean value of x
\mathbf{I}	unit matrix
i	square root of -1
$\hat{\mathbf{r}}$	unit vector
α	exponent in Gaussian function (normal distribution)
σ	standard deviation
σ^2	variance
h	Planck's constant ($6.626\ 18 \times 10^{-34}\ \text{J s}$)
\hbar	$h/2\pi$ ($1.054\ 59 \times 10^{-34}\ \text{J s}$)
m	particle mass
Ψ	molecular wavefunction
∇^2	$\partial^2/\partial x^2 + \partial^2/\partial y^2 + \partial^2/\partial z^2$ ('del-squared')
\mathcal{H}	Hamiltonian
ψ	spatial orbital
α, β	spin functions ('spin up' and 'spin down')
χ	spin orbital (product of spatial orbital and a spin function)
ϕ	basis function/atomic orbital (usually labelled $\phi_\mu, \phi_\nu, \phi_\lambda, \phi_\sigma$)
$d\nu$ or $d\mathbf{r}$	indicates an integral over all spatial coordinates
$d\sigma$	indicates an integral over all spin coordinates
$d\tau$	indicates an integral over all spatial and spin coordinates
r_{ij}	distances between two particles i and j (usually electrons in quantum mechanics)
R_{AB}	distance between two nuclei A and B
δ_{ij}	Kronecker delta ($\delta_{ij} = 1$ if $i = j$; $\delta_{ij} = 0$ if $i \neq j$)
\mathcal{K}	exchange operator
\mathcal{J}	Coulomb operator
$\mathcal{H}^{\text{core}}$	core Hamiltonian operator
\mathbf{F}	Fock matrix
\mathbf{S}	overlap matrix
S_{ij}	overlap integral between orbitals i and j
\mathcal{F}	Fock operator
\mathbf{C}	matrix of basis function coefficients

G	metric matrix (in distance geometry)
p_i	i th principal component
Z	variance-covariance matrix
λ	coupling parameter (used in free energy calculations)
$W(r^N)$	weighting function used in umbrella sampling
\mathcal{N}	number density ($= N/V$)
S_{AB}	similarity coefficient between two molecules A and B
D_{AB}	'distance' between two molecules A and B
σ	Hammett substitution constant
P	partition coefficient of solute between two solvents
π	$\log(P_x/P_H)$ for a substituent X relative to a hydrogen substituent
r^2	squared correlation coefficient
R^2	squared correlation coefficient in multiple linear regression
Q^2	cross-validated R^2

Acknowledgements

For this second edition I would like to thank Drs Neil Allan, Paul Bamborough, Gianpaolo Bravi, Richard Bryce, Julian Gale, Richard Green, Mike Hann and Alan Lewis, who commented on various parts of the new text. Julian Gale's suggestions were particularly useful for refining the sections concerning materials science and solid-state applications. I would also like to record my thanks once more to those who gave their time to read and comment on draft copies of various chapters of the first edition, upon which this second edition is based (in alphabetical order): Dr D B Adolf, Dr J M Blaney, Professor A V Chadwick, Dr P S Charifson, Dr C-W Chung, Dr A Cleasby, Dr A Emerson, Dr J W Essex, Dr D V S Green, Dr I R Gould, Dr M M Hann, Dr C A Leach, Dr M Pass, Dr D A Pearlman, Dr C A Reynolds, Dr D W Salt, Dr M Saqi, Professor J I Siepmann, Dr W C Swope, Dr N R Taylor, Dr P J Thomas, Professor D J Tildesley and Mr O Warschkow.

Assistance with the illustrations for the second edition was provided by Drs R Groot, S McGrother and V Milman. Many of the figures from the first edition are also included here and so I would like to thank again Dr S E Greasley, Dr M M Hann, Dr H Jhoti, Dr S N Jordan, Professor G R Luckhurst, Dr P M McMeekin, Dr A Nicholls, Dr P Popelier, Dr A Robinson and Dr T E Klein.

Alexandra Seabrook, Pauline Gillet and Julie Knight at Pearson Education provided the foundation of the publishing team, coping with a steady stream of questions and keeping everyone to schedule. Especial thanks are due to Julie, who did a splendid job as editor.

Any errors that remain are of course my own responsibility. If you do find any, I would like to know! I will also be pleased to receive any constructive suggestions, comments or criticisms. We plan to set up a web site that will provide access to various material from the book (such as electronic versions of the colour images) together with email contacts. This can be accessed via www.booksites.net.

Molecular modelling would not be what it is today without the efforts of those who develop computer hardware and software and I would like to acknowledge the authors of the following computer programs which were used to generate figures and/or data described in the text. All calculations were performed using Silicon Graphics computers.

AMBER D A Pearlman, D A Case, J C Caldwell, G L Seibel, U C Singh, P Weiner and P A Kollman 1991
Amber 3.0, University of California, San Francisco.

Cambridge Structural Database: F H Allen, S A Bellard, M D Brice, B A Cartwright, A Doubleday, H Higgs, T Hummelink, B G Hummelink-Peters, O Kennard, W D S Motherwell, J R Rodgers and D G Watson 1979 The Cambridge Crystallographic Data Centre: Computer-Based Search, Retrieval, Analysis and Display of Information. *Acta Crystallographica B35*:2331-2339. Cambridge Crystallographic Data Centre, Cambridge, United Kingdom.

CASTEP. Molecular Simulations Inc., 9685 Scranton Road, San Diego, California, USA.

Catalyst. Molecular Simulations Inc., 9685 Scranton Road, San Diego, California, USA.

Cerius2. Molecular Simulations Inc., 9685 Scranton Road, San Diego, California, USA.

- COSMIC: J G Vinter, A Davis, M R Saunders 1987. Strategic approaches to drug design. I. An integrated software framework for molecular modeling. *Journal of Computer-Aided Molecular Design*. **1**(1):31–51.
- Dials and Windows. G Ravishankar, S Swaminathan, D L Beveridge, R Lavery and H Sklenar 1989. *Journal of Biomolecular Structure and Dynamics* **6**:669–699. Wesleyan University, USA
- Gaussian 92: M J Frisch, G W Trucks, M Head-Gordon, P M W Gill, M W Wong, J B Foresman, B G Johnson, H B Schlegel, M A Robb, E S Replogle, R Gomperts, J L Andres, K Raghavachari, J S Binkley, C Gonzalez, R L Martin, D J Fox, D J DeFrees, J Baker, J J P Stewart and J A Pople. Gaussian Inc , Pittsburgh, Pennsylvania, USA
- GCG: Genetics Computer Group, Inc., University Research Park, 575 Science Drive, Suite B, Madison, Wisconsin 53711, USA.
- GRASP (Graphical Representation and Analysis of Surface Properties): A Nicholls, Columbia University, New York, USA
- GRID: P J Goodford 1985 A Computational Procedure for Determining Energetically Favorable Binding Sites on Biologically Important Macromolecules. *Journal of Medicinal Chemistry* **28**:849–857. Molecular Discovery Ltd, Oxford, United Kingdom.
- InsightII: Molecular Simulations Inc., 9685 Scranton Road, San Diego, California, USA.
- IsoStar: I J Bruno, J C Cole, J P M Lommerse, R S Rowland, R Taylor and M L Verdonk 1997. IsoStar: a library of information about nonbonded interactions. *Journal of Computer-Aided Molecular Design* **11**:525–537. Cambridge Crystallographic Data Centre, Cambridge, United Kingdom.
- Micromol: S M Colwell, A R Marshall, R D Amos and N C Handy 1985. Quantum chemistry on microcomputers, *Chemistry in Britain* **21**:655–659
- Molscript: P J Kraulis 1991. Molscript – A program to produce both detailed and schematic plots of protein structures. *Journal of Applied Crystallography* **24**:946–950.
- PROCHECK. R Laskowski, M W MacArthur, D S Moss and J M Thornton 1993. Procheck – A program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography* **26**:283–291.
- Quanta: Molecular Simulations Inc., 9685 Scranton Road, San Diego, California, USA.
- Spartan: Wavefunction Inc., 18401 Von Karman, Suite 370, Irvine, California, USA.
- SPASMS (San Francisco Package of Applications for the Simulation of Molecular Systems). D A Spellmeyer, W C Swope, E-R Evensen, T Cheatham, D M Ferguson and P A Kollman. University of California, San Francisco, USA.
- Sybyl: Tripos Inc., 1699 South Hanley Road, St. Louis, Missouri, USA.
- The following programs were used to produce draft copies of the manuscript and diagrams: Microsoft Word (Microsoft Corp.), Gnuplot (T Williams and C Kelley), Kaleidagraph (Abelbeck Software), Chem3D (CambridgeSoft Corp) and Microsoft Excel (Microsoft Corp.).

We are grateful to the following for permission to reproduce copyright material:

- Figure 1.11 from *Mathematical Methods in the Physical Sciences*, 2nd edn, Boas M L, ©1983. Reprinted by permission of John Wiley & Sons, Inc.
- Figure 1.14 from *The FFT Fundamentals and Concepts* by Ramirez ©1985. Reprinted by permission of Prentice-Hall, Inc., Upper Saddle River, NJ.
- Figures 2 7 and 3 3 from *Ab initio Molecular Orbital Theory*, Hehre W J, L Radom, P v R Schleyer, J A Hehre ©1986 Reprinted with permission by John Wiley & Sons, Inc.
- Figure 3 5 from Gerratt J, D L Cooper, P B Karadakov and M Raimondi 1997. Modern valence bond theory. *Chemical Society Reviews* 87–100. Reproduced by permission of The Royal Society of Chemistry
- Figure 3.22 from Needs R J and Mujica 1995. First-principles pseudopotential study of the structural phases of silicon. *Physical Review B* **51**:9652–9660

- Figure 4.18 from Buckingham A D 1959. Molecular Quadrupole Moments *Quarterly Reviews of the Chemical Society* **13**:183–214. Reproduced by permission of The Royal Society of Chemistry.
- Figure 4.29 from *Computer Simulation in Chemical Physics*, edited by Allen M P and D J Tildesley, 1993. Effective Pair Potentials and Beyond, Sprik M, with kind permission from Kluwer Academic Publishers
- Figure 4.49 reprinted with permission from Pranata J and W L Jorgensen. Computational Studies on FK506: Conformational Search and Molecular Dynamics Simulations in Water. *The Journal of the American Chemical Society* **113**:9483–9493 ©1991 American Chemical Society.
- Figure 4.50 from Molecular Parameters for Organosilicon Compounds Calculated from *Ab Initio* Computations, Grigoras S and T H Lane, *Journal of Computational Chemistry* **9**:25–39, ©1988 Reprinted by permission of John Wiley & Sons, Inc.
- Figures 5.4 and 5.8 Press W H, B P Flannery, S A Teukolsky and W T Vetterling, *Numerical Recipes in Fortran*. 1992, Cambridge University Press.
- Figure 5.21 reprinted with permission from Chandrasekhar J, S F Smith and W L Jorgensen. Theoretical Examination of the S_N2 Reaction Involving Chloride Ion and Methyl Chloride in the Gas Phase and Aqueous Solution. *The Journal of the American Chemical Society* **107**:154–163. ©1985 American Chemical Society.
- Figure 5.23 reprinted with permission from Doubleday C, J McIver, M Page and T Zielinski Temperature Dependence of the Transition-State Structure for the Disproportionation of Hydrogen Atom with Ethyl Radical. *The Journal of the American Chemical Society* **107**:5800–5801 ©1985 American Chemical Society.
- Figure 5.29 from Gonzalez C and H B Schlegel 1988 An Improved Algorithm for Reaction Path Following. *The Journal of Chemical Physics* **90**:2154–2161.
- Figure 5.30 reprinted from *Chemical Physical Letters*, 194, Fischer S and M Karplus. Conjugate Peak Refinement: An Algorithm for Finding Reaction Paths and Accurate Transition States in Systems with Many Degrees of Freedom. 252–261, ©1992, with permission from Elsevier Science
- Figure 5.35 reprinted with permission from Houk K N, J González and Y Li. Pericyclic Reaction Transition States. Passions and Punctilios 1935–1995. *Accounts of Chemical Research* **28**:81–90. ©1995 American Chemical Society.
- Figure 6.25 reprinted from *Chemical Physics Letters*, 196, Ding H-Q, N Karasawa and W A Goddard III, The Reduced Cell Multipole Method for Coulomb Interactions in Periodic Systems with Million-Atom Unit Cells, 6–10, ©1992, with permission of Elsevier Science
- Figure 7.2 from Alder B J and TE Wainwright 1959. Studies in Molecular Dynamics. I. General Method. *The Journal of Chemical Physics* **31**:459–466.
- Figure 7.11 from Alder B J and T E Wainwright 1970. Decay of the Velocity Autocorrelation Function. *Physical Review A* **1**:18–21.
- Figure 7.12 from Guillot B 1991 A Molecular Dynamics Study of the Infrared Spectrum of Water. *The Journal of Chemical Physics* **95**:1543–1551.
- Figure 7.13 reprinted with permission from Jorgensen W L, R C Binning Jr and B Bigot Structures and Properties of Organic Liquids: *n*-Butane and 1,2-Dichloroethane and Their Conformational Equilibria. *The Journal of the American Chemical Society* **103**:4393–4399. ©1981 American Chemical Society.
- Figure 7.24 (and on cover) from Groot R D and T J Madden 1998. Dynamic simulation of diblock copolymer microphase separation *The Journal of Chemical Physics* **108**:8713–8724. © American Institute of Physics
- Figure 8.16 from Frantz, D D, D L Freeman and J D Doll 1990. Reducing quasi-ergodic behavior in Monte Carlo simulations by J-walking: applications to atomic clusters. *The Journal of Chemical Physics* **93**:2769–2784

Figure 8.17 from Cracknell R F, D Nicholson and N Quirke 1994. A Grand Canonical Monte Carlo Study of Lennard-Jones Mixtures in Slit Pores; 2: Mixtures of Two-Centre Ethane with Methane. *Molecular Simulation* 13:161–175. ©1994 OPA (Overseas Publishers Association) N.V. Permission to reproduce granted by Gordon and Breach Publishers.

Figure 8.22 reprinted with permission from Smit B and J I Siepmann. Simulating the Adsorption of Alkanes in Zeolites. *Science* 264 1118–1120 ©1994 American Association for the Advancement of Science.

Figure 9.34 from Poling Promoting Conformational Variation, Smellie A S, S L Teig and P Towbin *Journal of Computational Chemistry* 16:171–187, ©1995. Reprinted by permission of John Wiley & Sons, Inc.

Figure 10.18 from Pearson W R and D J Lipman 1988 Improved tools for biological sequence comparison *Proceedings of the National Academy of Sciences USA* 85 2444–2448.

Figure 10.21 reprinted from Current Opinion in Structural Biology, 6, Eddy S R, Hidden Markov Models, 361–365 ©1996, with permission from Elsevier Science

Figure 10.26 Reprinted with permission from Lüthy R, J U Bowie and D Eisenberg Assessment of Protein Models with Three-Dimensional Profiles *Nature* 356:83–85. ©1992 Macmillan Magazines Limited.

Figure 11.6 from Lybrand T P, J A McCammon and G Wipff 1986 Theoretical Calculation of Relative Binding Affinity in Host–Guest Systems. *Proceedings of the National Academy of Sciences USA* 83:833–835

Figure 11.18 reprinted with permission from Guo Z and C L Brooks III Rapid Screening of Binding Affinities. Application of the λ -Dynamics Method to a Trypsin-Inhibitor System. *The Journal of the American Chemical Society* 120:1920–1921. ©1998 American Chemical Society.

Figure 11.24 reprinted with permission from Still W C, A Tempczyrk, R C Hawley and T Hendrickson Semianalytical Treatment of Solvation for Molecular Mechanics and Dynamics. *The Journal of the American Chemical Society* 112 6127–6129 ©1990 American Chemical Society.

Figure 11.35 reprinted with permission from Chandrasekhar J and W L Jorgensen 1985. Energy Profile for a Nonconcerted S_N2 Reaction in Solution. *The Journal of the American Chemical Society* 107:2974–2975. ©1985 American Chemical Society.

Figure 11.37 reprinted with permission from Åqvist J, M Fothergill and A Warshel. Computer Simulation of the CO_2/HCO_3^- Interconversion Step in Human Carbonic Anhydrase I. *The Journal of the American Chemical Society* 115:631–635 ©1993 American Chemical Society

Figure 11.40 reprinted with permission from Saitta A M, P D Sooper, E Wasserman and M L Klein 1999. Influence of a knot on the strength of a polymer strand. *Nature* 399:46–48. ©1999 Macmillan Magazines Limited.

Figure 11.42 from NATO ASI Series C 498 (*New Trends in Materials Chemistry*), 1997, 285–318, Defects and Matter Transport in Solid Materials, Chadwick A V and J Corish, with kind permission of Kluwer Academic Publishers

Whilst every effort has been made to trace the owners of copyright material, we take this opportunity to offer our apologies to any copyright holders whose rights we may have unwittingly infringed.

Useful Concepts in Molecular Modelling

1.1 Introduction

What is molecular modelling? ‘Molecular’ clearly implies some connection with molecules. The *Oxford English Dictionary* defines ‘model’ as ‘a simplified or idealised description of a system or process, often in mathematical terms, devised to facilitate calculations and predictions’. Molecular modelling would therefore appear to be concerned with ways to mimic the behaviour of molecules and molecular systems. Today, molecular modelling is invariably associated with computer modelling, but it is quite feasible to perform some simple molecular modelling studies using mechanical models or a pencil, paper and hand calculator. Nevertheless, computational techniques have revolutionised molecular modelling to the extent that most calculations could not be performed without the use of a computer. This is not to imply that a more sophisticated model is necessarily any better than a simple one, but computers have certainly extended the range of models that can be considered and the systems to which they can be applied.

The ‘models’ that most chemists first encounter are molecular models such as the ‘stick’ models devised by Dreiding or the ‘space filling’ models of Corey, Pauling and Koltun (commonly referred to as CPK models). These models enable three-dimensional representations of the structures of molecules to be constructed. An important advantage of these models is that they are interactive, enabling the user to pose ‘what if ...’ or ‘is it possible to ...’ questions. These structural models continue to play an important role both in teaching and in research, but molecular modelling is also concerned with more abstract models, many of which have a distinguished history. An obvious example is quantum mechanics, the foundations of which were laid many years before the first computers were constructed.

There is a lot of confusion over the meaning of the terms ‘theoretical chemistry’, ‘computational chemistry’ and ‘molecular modelling’. Indeed, many practitioners use all three labels to describe aspects of their research, as the occasion demands! ‘Theoretical chemistry’ is often considered synonymous with quantum mechanics, whereas computational chemistry encompasses not only quantum mechanics but also molecular mechanics, minimisation, simulations, conformational analysis and other computer-based methods for understanding and predicting the behaviour of molecular systems. Molecular modellers use all of these methods and so we shall not concern ourselves with semantics but rather shall consider any theoretical or computational technique that provides insight into the behaviour of molecular systems to be an example of molecular modelling. If a distinction has to be

made, it is in the emphasis that molecular modelling places on the representation and manipulation of the structures of molecules, and properties that are dependent upon those three-dimensional structures. The prominent part that computer graphics has played in molecular modelling has led some scientists to consider molecular modelling as little more than a method for generating ‘pretty pictures’, but the technique is now firmly established, widely used and accepted as a discipline in its own right.

A closely related subject is molecular informatics. This is a rather new term, making a precise definition difficult, but it is usually considered to encompass two disciplines: chemoinformatics and bioinformatics. Of these two areas, chemoinformatics (also written cheminformatics) is the newer name but the older discipline; chemists have been using computers to store, retrieve and manipulate information about molecules almost since computers were invented. Both chemoinformatics and bioinformatics have risen to prominence primarily as a consequence of the introduction of new experimental techniques. For the chemist these experimental techniques are combinatorial library synthesis and high-throughput screening, which enable very large numbers of molecules to be synthesised and tested; for the biologist they are the automated sequencing machines that are being used to determine the human genome. A characteristic feature of molecular informatics is that it is concerned with information about large numbers of molecules, much larger than is typically the case for a traditional molecular modelling study. For this reason, informatics was initially more concerned with less complex representations of molecules that did not fully represent their three-dimensional properties. However, even this distinction is now being eroded and there is increasing use made of more traditional molecular modelling techniques within informatics.

In the rest of this chapter we shall discuss a number of concepts and techniques that are relevant to many areas of molecular modelling and so do not sit comfortably in any individual chapter. We will also define some of the terms that will be used throughout the book.

1.2 Coordinate Systems

It is obviously important to be able to specify the positions of the atoms and/or molecules in the system to a modelling program^{*}. There are two common ways in which this can be done. The most straightforward approach is to specify the Cartesian (x, y, z) coordinates of all the atoms present. The alternative is to use *internal coordinates*, in which the position of each atom is described relative to other atoms in the system. Internal coordinates are usually written as a Z-matrix. The Z-matrix contains one line for each atom in the system. A sample Z-matrix for the staggered conformation of ethane (see Figure 1.1) is

^{*}For a system containing a large number of independent molecules it is common to use the term ‘configuration’ to refer to each arrangement; this use of the word ‘configuration’ is not to be confused with its standard chemical meaning as a different bonding arrangement of the atoms in a molecule

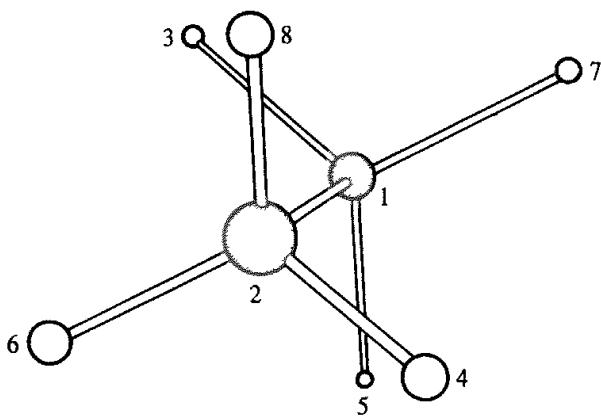


Fig. 1.1 The staggered conformation of ethane

as follows:

1	C							
2	C	1.54	1					
3	H	1.0	1	109.5	2			
4	H	1.0	2	109.5	1	180.0	3	
5	H	1.0	1	109.5	2	60.0	4	
6	H	1.0	2	109.5	1	-60.0	5	
7	H	1.0	1	109.5	2	180.0	6	
8	H	1.0	2	109.5	1	60.0	7	

In the first line of the Z-matrix we define atom 1, which is a carbon atom. Atom number 2 is also a carbon atom that is a distance of 1.54 Å from atom 1 (columns 3 and 4). Atom 3 is a hydrogen atom that is bonded to atom 1 with a bond length of 1.0 Å. The angle formed by atoms 2-1-3 is 109.5°, information that is specified in columns 5 and 6. The fourth atom is a hydrogen, a distance of 1.0 Å from atom 2, the angle 4-2-1 is 109.5°, and the torsion angle (defined in Figure 1.2) for atoms 4-2-1-3 is 180°. Thus for all except the first three atoms, each atom has three internal coordinates: the distance of the atom from one of the atoms previously defined, the angle formed by the atom and two of the previous atoms, and the torsion angle defined by the atom and three of the previous atoms. Fewer internal coordinates are required for the first three atoms because the first atom can be placed anywhere in space (and so it has no internal coordinates); for the second atom it is only necessary to specify its distance from the first atom and then for the third atom only a distance and an angle are required.

It is always possible to convert internal to Cartesian coordinates and vice versa. However, one coordinate system is usually preferred for a given application. Internal coordinates can usefully describe the relationship between the atoms in a single molecule, but Cartesian coordinates may be more appropriate when describing a collection of discrete molecules. Internal coordinates are commonly used as input to quantum mechanics programs, whereas calculations using molecular mechanics are usually done in Cartesian coordinates. The total number of coordinates that must be specified in the internal coordinate system is six fewer

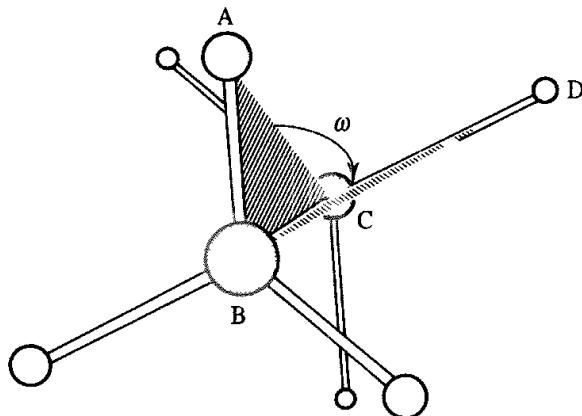


Fig. 1.2 A torsion angle $A-B-C-D$ is defined as the angle between the planes A, B, C and B, C, D . A torsion angle can vary through 360° although the range -180° to $+180^\circ$ is most commonly used. We shall adopt the IUPAC definition of a torsion angle in which an eclipsed conformation corresponds to a torsion angle of 0° and a trans or anti conformation to a torsion angle of 180° . The reader should note that this may not correspond to some of the definitions used in the literature, where the trans arrangement is defined as a torsion angle of 0° . If one looks along the bond $B-C$, then the torsion angle is the angle through which it is necessary to rotate the bond AB in a clockwise sense in order to superimpose the two planes, as shown.

than the number of Cartesian coordinates for a non-linear molecule. This is because we are at liberty to arbitrarily translate and rotate the system within Cartesian space without changing the relative positions of the atoms

1.3 Potential Energy Surfaces

In molecular modelling the Born–Oppenheimer approximation is invariably assumed to operate. This enables the electronic and nuclear motions to be separated; the much smaller mass of the electrons means that they can rapidly adjust to any change in the nuclear positions. Consequently, the energy of a molecule in its ground electronic state can be considered a function of the nuclear coordinates only. If some or all of the nuclei move then the energy will usually change. The new nuclear positions could be the result of a simple process such as a single bond rotation or it could arise from the concerted movement of a large number of atoms. The magnitude of the accompanying rise or fall in the energy will depend upon the type of change involved. For example, about 3 kcal/mol is required to change the covalent carbon–carbon bond length in ethane by 0.1 \AA away from its equilibrium value, but only about 0.1 kcal/mol is required to increase the non-covalent separation between two argon atoms by 1 \AA from their minimum energy separation. For small isolated molecules, rotation about single bonds usually involves the smallest changes in energy. For example, if we rotate the carbon–carbon bond in ethane, keeping all of the bond lengths and angles fixed in value, then the energy varies in an approximately sinusoidal fashion as shown in Figure 1.3, with minima at the three staggered conformations. The energy in this case can be considered a function of a single coordinate only (i.e. the torsion angle of the carbon–carbon bond), and as such can be displayed graphically, with energy along one axis and the value of the coordinate along the other.

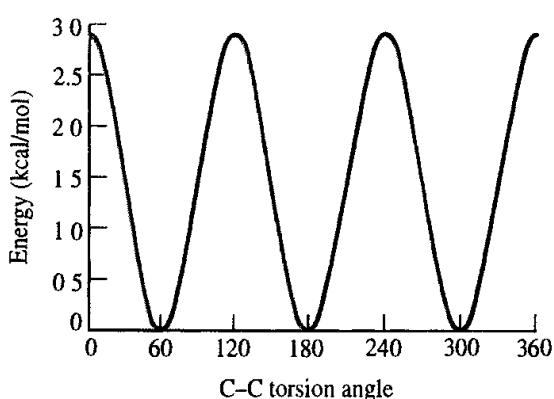


Fig. 1.3 Variation in energy with rotation of the carbon–carbon bond in ethane

Changes in the energy of a system can be considered as movements on a multidimensional ‘surface’ called the *energy surface*. We shall be particularly interested in stationary points on the energy surface, where the first derivative of the energy is zero with respect to the internal or Cartesian coordinates. At a stationary point the forces on all the atoms are zero. Minimum points are one type of stationary point; these correspond to stable structures. Methods for locating stationary points will be discussed in more detail in Chapter 5, together with a more detailed consideration of the concept of the energy surface.

1.4 Molecular Graphics

Computer graphics has had a dramatic impact upon molecular modelling. It should always be remembered, however, that there is much more to molecular modelling than computer graphics. It is the interaction between molecular graphics and the underlying theoretical methods that has enhanced the accessibility of molecular modelling methods and assisted the analysis and interpretation of such calculations.

Molecular graphics systems have evolved from delicate and temperamental pieces of equipment that cost hundreds of thousands of pounds and occupied entire rooms, to today’s inexpensive workstations that fit on or under a desk and yet are hundreds of times more powerful. Over the years, two different types of molecular graphics display have been used in molecular modelling. First to be developed were vector devices, which construct pictures using an electron gun to draw lines (or dots) on the screen, in a manner similar to an oscilloscope. Vector devices were the mainstay of molecular modelling for almost two decades but have now been largely superseded by raster devices. These divide the screen into a large number of small ‘dots’, called pixels. Each pixel can be set to any of a large number of colours, and so by setting each pixel to the appropriate colour it is possible to generate the desired image.

Molecules are most commonly represented on a computer graphics screen using ‘stick’ or ‘space-filling’ representations, which are analogous to the Dreiding and Corey–Pauling–Koltun (CPK) mechanical models. Sophisticated variations on these two basic types have

been developed, such as the ability to colour molecules by atomic number and the inclusion of shading and lighting effects, which give 'solid' models a more realistic appearance. Some of the commonly used molecular representations are shown in Figure 1.4 (colour plate section). Computer-generated models do have some advantages when compared with their mechanical counterparts. Of particular importance is the fact that a computer model can be very easily interrogated to provide quantitative information, from simple geometrical measures such as the distance between two atoms to more complex quantities such as the energy or surface area. Quantitative information such as this can be very difficult if not impossible to obtain from a mechanical model. Nevertheless, mechanical models may still be preferred in certain types of situation due to the ease with which they can be manipulated and viewed in three dimensions. A computer screen is inherently two-dimensional, whereas molecules are three-dimensional objects. Nevertheless, some impression of the three-dimensional nature of an object can be represented on a computer screen using techniques such as depth cueing (in which those parts of the object that are further away from the viewer are made less bright) and through the use of perspective. Specialised hardware enables more realistic three-dimensional stereo images to be viewed. In the future 'virtual reality' systems may enable a scientist to interact with a computer-generated molecular model in much the same way that a mechanical model can be manipulated.

Even the most basic computer graphics program provides some standard facilities for the manipulation of models, including the ability to translate, rotate and 'zoom' the model towards and away from the viewer. More sophisticated packages can provide the scientist with quantitative feedback on the effect of altering the structure. For example, as a bond is rotated then the energy of each structure could be calculated and displayed interactively.

For large molecular systems it may not always be desirable to include every single atom in the computer image; the sheer number of atoms can result in a very confusing and cluttered picture. A clearer picture may be achieved by omitting certain atoms (e.g. hydrogen atoms) or by representing groups of atoms as single 'pseudo-atoms'. The techniques that have been developed for displaying protein structures nicely illustrate the range of computer graphics representation possible (the use of computational techniques to investigate the structures of proteins is considered in Chapter 10). Proteins are polymers constructed from amino acids, and even a small protein may contain several thousand atoms. One way to produce a clearer picture is to dispense with the explicit representation of any atoms and to represent the protein using a 'ribbon'. Proteins are also commonly represented using the cartoon drawings developed by J Richardson, an example of which is shown in Figure 1.5 (colour plate section). The cylinders in this figure represent an arrangement of amino acids called an α -helix, and the flat arrows an alternative type of regular structure called a β -strand. The regions between the cylinders and the strands have no such regular structure and are represented as 'tubes'.

1.5 Surfaces

Many of the problems that are studied using molecular modelling involve the non-covalent interaction between two or more molecules. The study of such interactions is often facilitated

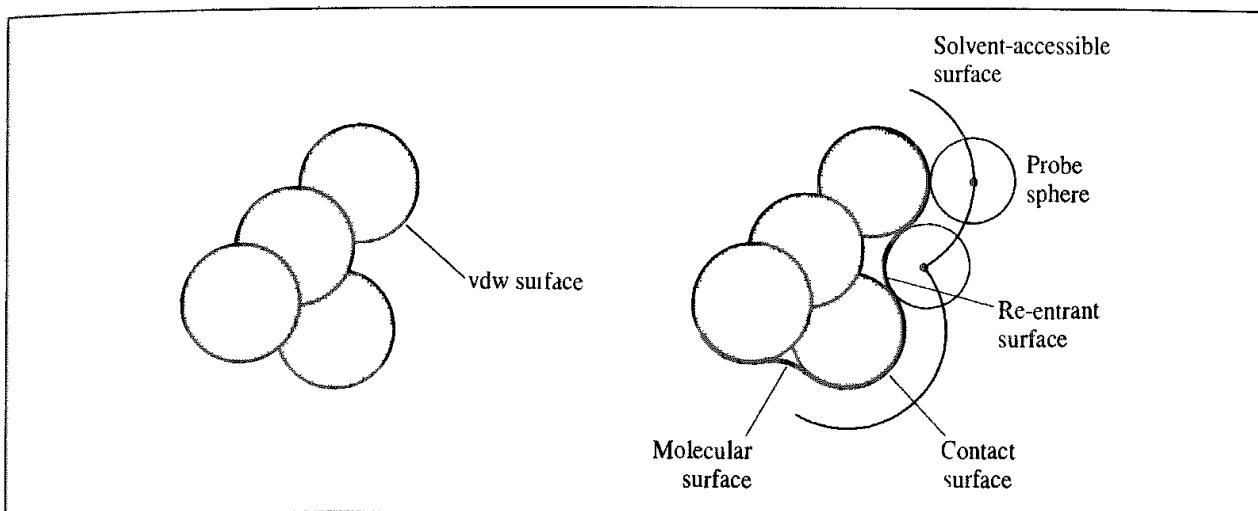


Fig 1.6: The van der Waals (*vdw*) surface of a molecule corresponds to the outward-facing surfaces of the van der Waals spheres of the atoms. The molecular surface is generated by rolling a spherical probe (usually of radius 1.4 Å to represent a water molecule) on the van der Waals surface. The molecular surface is constructed from contact and re-entrant surface elements. The centre of the probe traces out the accessible surface.

by examining the van der Waals, molecular or accessible surfaces of the molecule. The *van der Waals surface* is simply constructed from the overlapping van der Waals spheres of the atoms, Figure 1.6. It corresponds to a CPK or space-filling model. Let us now consider the approach of a small 'probe' molecule, represented as a single van der Waals sphere, up to the van der Waals surface of a larger molecule. The finite size of the probe sphere means that there will be regions of 'dead space', crevices that are not accessible to the probe as it rolls about on the larger molecule. This is illustrated in Figure 1.6. The amount of dead space increases with the size of the probe; conversely, a probe of zero size would be able to access all of the crevices. The *molecular surface* [Richards 1977] is traced out by the inward-facing part of the probe sphere as it rolls on the van der Waals surface of the molecule. The molecular surface contains two different types of surface element. The *contact surface* corresponds to those regions where the probe is actually in contact with the van der Waals surface of the 'target'. The *re-entrant* surface regions occur where there are crevices that are too narrow for the probe molecule to penetrate. The molecular surface is usually defined using a water molecule as the probe, represented as a sphere of radius 1.4 Å.

The *accessible surface* is also widely used. As originally defined by Lee and Richards [Lee and Richards 1971] this is the surface that is traced by the centre of the probe molecule as it rolls on the van der Waals surface of the molecule (Figure 1.6). The centre of the probe molecule can thus be placed at any point on the accessible surface and not penetrate the van der Waals spheres of any of the atoms in the molecule.

Widely used algorithms for calculating the molecular and accessible surfaces were developed by Connolly [Connolly 1983a,b], and others [e.g. Richmond 1984] have described formulae for the calculation of exact or approximate values of the surface area. There are many ways to represent surfaces, some of which are illustrated in Figure 1.7 (colour plate section). As shown, it may also be possible to endow a surface with a translucent quality, which enables the molecule inside the surface to be displayed. Clipping can also be used

to cut through the surface to enable the 'inside' to be viewed. In addition, properties such as the electrostatic potential can be calculated on the surface and represented using an appropriate colour scheme. Useful though these representations are, it is important to remember that the electronic distribution in a molecule formally extends to infinity. The 'hard sphere' representation is often very convenient and has certainly proved very valuable, but it may not be appropriate in all cases [Rouvray 1997, 1999, 2000].

1.6 Computer Hardware and Software

One cannot fail to be amazed at the pace of development in the computer industry, where the ratio of performance-to-price has increased by an order of magnitude every five years or so. The workstations that are commonplace in many laboratories now offer a real alternative to centrally maintained 'supercomputers' for molecular modelling calculations, especially as a workstation or even a personal computer can be dedicated to a single task, whereas the supercomputer has to be shared with many other users. Nevertheless, in the immediate future there will always be some calculations that require the power that only a supercomputer can offer. The speed of any computer system is ultimately constrained by the speed at which electrical signals can be transmitted. This means that there will come a time when no further enhancements can be made using machines with 'traditional' single-processor serial architectures, and parallel computers will play an ever more important role.

A parallel computer couples processors together in such a way that a calculation is divided into small pieces with the results being combined at the end. Some calculations are more amenable to parallel processing than others, and a significant amount of effort is being spent converting existing algorithms to run efficiently on parallel architectures. In some cases completely new methods have been developed to take maximum advantage of the opportunities of parallel processing. The low cost of personal computer chips means that large 'farms' of processors can be constructed to give significant computing power for relatively small outlay.

To perform molecular modelling calculations one also requires appropriate programs (the software). The software used by molecular modellers ranges from simple programs that perform just a single task to highly complex packages that integrate many different methods. There is also an extremely wide variation in the price of software! Some programs have been so widely used and tested that they can be considered to have reached the status of a 'gold standard' against which similar programs are compared. One hesitates to specify such programs in print, but three items of software have been so widely used and cited that they can safely be afforded the accolade. These are the Gaussian series of programs for performing *ab initio* quantum mechanics, the MOPAC/AMPAC programs for semi-empirical quantum mechanics and the MM2 program for molecular mechanics.

Various pieces of software were used to generate the data for the examples and illustrations throughout this book. Some of these were written specifically for the task; some were freely available programs; others were commercial packages. I have decided not to describe specific programs in any detail, as such descriptions rapidly become outdated. Nevertheless,

all items of software are accredited where appropriate. Please note that the use of any particular piece of software does not imply any recommendation!

1.7 Units of Length and Energy

It will be noted that our Z-matrix for ethane has been defined using the angstrom as the unit of length ($1 \text{ \AA} \equiv 10^{-10} \text{ m} \equiv 100 \text{ pm}$). The ångström is a non-SI unit but is a very convenient one to use, as most bond lengths are of the order of 1–2 Å. One other very common non-SI unit found in the molecular modelling literature is the kilocalorie ($1 \text{ kcal} \equiv 4.1840 \text{ kJ}$). Other systems of units are employed in other types of calculation, such as the atomic units used in quantum mechanics (discussed in Chapter 2). It is important to be aware of, and familiar with, these non-standard units as they are widely used in the literature and throughout this book.

1.8 The Molecular Modelling Literature

The number of scientific papers concerned with molecular modelling methods is rising rapidly, as is the number of journals in which such papers are published. This reflects the tremendous diversity of problems to which molecular modelling can be applied and the ever-increasing availability of molecular modelling methods. It does, however, mean that it can be very difficult to remain up to date with the field. A number of specialist journals are devoted to theoretical chemistry, computational chemistry and molecular modelling, each with their own particular emphasis. Relevant papers are also published in the more ‘general’ journals, and there are now a number of books covering aspects of molecular modelling, some aimed at the specialist reader, others at the beginner. Many scientists are now fortunate to have access to electronic catalogues of publications which can be searched to find relevant papers. As many journals are now available over the internet it is possible to perform a literature search and obtain copies of the relevant papers without even having to leave the office. Some of the journals which are devoted to short reviews of recent developments often include molecular modelling sections (such as the ‘Current Opinion’ series); in others, useful review articles appear on an occasional basis. One particularly valuable source of information on molecular modelling methods is the *Reviews in Computational Chemistry*, edited by Lipkowitz and Boyd, beginning in 1990 (see Further Reading). Each of these volumes contains chapters on a variety of subjects, each written by an appropriate expert. A recent addition is the *Encyclopaedia of Computational Chemistry* by Schleyer *et al.* (1998) (see Further Reading), which contains many chapters that cover a wide range of topics.

1.9 The Internet

In the first edition of this book I wrote, ‘A major use of the Internet is for electronic mail, but extremely rapid growth is being observed in other areas, particularly the “World-Wide Web” (WWW) . Such a phrase seems an understatement; despite the ‘hype’, the Internet has certainly made a dramatic impact, not least on the scientific community, where its

origins lie. Anything written about the Internet is almost certain to become obsolete more rapidly than any other topic in this book and so this section will be brief. I will assume that all readers of this book will be familiar with the use of a web browser and the concept of a hyperlink, which enables documents to be linked together. The URL (Uniform Resource Locator) is the currency of the WWW, being the 'electronic address' which enables the particular item to be identified. Most documents are still written using HTML (HyperText Markup Language) but increasingly incorporate more sophisticated features. Given the tremendous growth in the Web it is important to be able to locate relevant information. This is the role of the Internet search engines, which can be used to identify relevant sites of interest via some form of keyword search. Within the molecular modelling context, several trends can be noted. Whilst the Web was initially used to distribute mostly textual information, it is increasingly used for much more sophisticated applications. Interactive molecular graphics are a feature of many sites. Some sites enable calculations or database searches to be performed via the Web, with the results being delivered interactively or via email. This is particularly true for 'intranets' within an organisation. XML (eXtensible Markup Language) is likely to play an increasingly important role in the 'intelligent' exchange of information over the Web, especially in specialist areas such as chemistry [Murray-Rust and Rzepa 1999]. Several 'electronic conferences' have been held with participants from many different countries. Perhaps the only prediction that one can safely make about the Web is that it is here to stay and its use will continue to grow.

1.10 Mathematical Concepts

A full appreciation of all of the techniques of molecular modelling would require a mathematical treatment beyond that appropriate to a book of this size and scope. However, a proper understanding does benefit from some knowledge of mathematical concepts such as vectors, matrices, differential equations, complex numbers, series expansions and Lagrangian multipliers, and some very elementary statistical concepts. There is only space in this book for a cursory introduction to these mathematical concepts and ideas, with very brief descriptions and some key results. The suggestions for further reading provide detailed background information on all of the mathematical topics required.

1.10.1 Series Expansions

There are various series expansions that are useful for approximating functions. Particularly important is the *Taylor series*: if $f(x)$ is a continuous, single-valued function of x with continuous derivatives $f'(x), f''(x), \dots$, then we can expand the function about a point x_0 as follows:

$$f(x_0 + x) = f(x_0) + \frac{x}{1!} f'(x_0) + \frac{x^2}{2!} f''(x_0) + \frac{x^3}{3!} f'''(x_0) + \cdots + \frac{x^n}{n!} f^{(n)}(x_0) \quad (1.1)$$

Taylor series are often truncated after the term involving the second derivative, which makes the function vary in a quadratic fashion. This is a common assumption in many of the minimisation algorithms that we will discuss in Chapter 5.

A *Maclaurin series* is a specific form of the Taylor series for which $x_0 = 0$. Some standard expansions in Taylor series form are:

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \dots \quad (1.2)$$

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \dots \quad (1.3)$$

$$\ln(1 + x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \dots \quad (1.4)$$

The *binomial expansion* is used for functions of the form $(1 + x)^\alpha$:

$$(1 + x)^\alpha = 1 + \alpha x + \alpha(\alpha - 1) \frac{x^2}{2!} + \alpha(\alpha - 1)(\alpha - 2) \frac{x^3}{3!} + \dots \quad (1.5)$$

All these series must have $|x| < 1$ to be convergent.

1.10.2 Vectors

A vector is a quantity with both magnitude and direction. For example, the velocity of a moving body is a vector quantity as it defines both the direction in which the body is travelling and the speed at which it is moving. In Cartesian coordinates a vector such as the velocity will have three components, indicating the contribution to the overall motion from the component motions along the x , y and z directions. The addition and subtraction of vectors can be understood using geometrical constructions, as shown in Figure 1.8. Thus, if we want to calculate the force on an atom due to its interactions with all other atoms in the system (as required in molecular dynamics calculations, see Chapter 7), we would perform a vector sum of all the individual forces.

Some of the common manipulations that are performed with vectors include the scalar product, vector product and scalar triple product, which we will illustrate using vectors \mathbf{r}_1 , \mathbf{r}_2 and \mathbf{r}_3 that are defined in a rectangular Cartesian coordinate system:

$$\begin{aligned} \mathbf{r}_1 &= x_1\mathbf{i} + y_1\mathbf{j} + z_1\mathbf{k} \\ \mathbf{r}_2 &= x_2\mathbf{i} + y_2\mathbf{j} + z_2\mathbf{k} \\ \mathbf{r}_3 &= x_3\mathbf{i} + y_3\mathbf{j} + z_3\mathbf{k} \end{aligned} \quad (1.6)$$

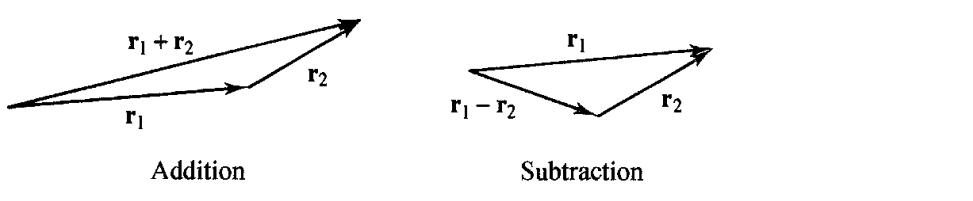


Fig 1.8. The addition and subtraction of vectors

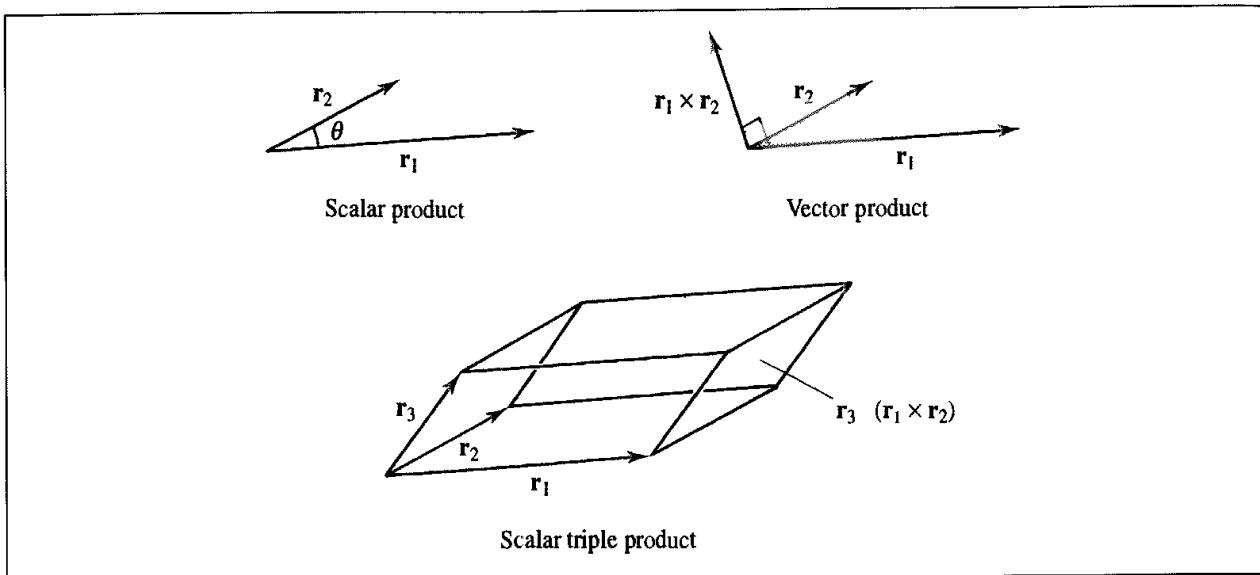


Fig. 1.9. The scalar product, vector product and scalar triple product.

i, **j** and **k** are orthogonal unit vectors along the *x*, *y* and *z* axes. The *scalar product* is defined as:

$$\mathbf{r}_1 \cdot \mathbf{r}_2 = |\mathbf{r}_1| |\mathbf{r}_2| \cos \theta \quad (1.7)$$

$|\mathbf{r}_1|$ and $|\mathbf{r}_2|$ are the magnitudes of the two vectors ($|\mathbf{r}_1| = \sqrt{x_1^2 + y_1^2 + z_1^2}$) and θ is the angle between them (Figure 1.9). The angle can be calculated as follows:

$$\cos \theta = \frac{x_1 x_2 + y_1 y_2 + z_1 z_2}{|\mathbf{r}_1| |\mathbf{r}_2|} \quad (1.8)$$

The scalar product of two vectors is thus a scalar.

The *vector product* of two vectors $\mathbf{r}_1 \times \mathbf{r}_2$ (sometimes written $\mathbf{r}_1 \wedge \mathbf{r}_2$) is a new vector (**v**), in a direction perpendicular to the plane containing the two original vectors (Figure 1.9). The direction of this new vector is such that \mathbf{r}_1 , \mathbf{r}_2 and the new vector form a right-handed system. If \mathbf{r}_1 and \mathbf{r}_2 are three-component vectors then the components of **v** are given by:

$$\mathbf{v} = (y_1 z_2 - z_1 y_2) \mathbf{i} + (z_1 x_2 - x_1 z_2) \mathbf{j} + (x_1 y_2 - y_1 x_2) \mathbf{k} \quad (1.9)$$

Note that the vector product $\mathbf{r}_2 \times \mathbf{r}_1$ is not the same as the vector product $\mathbf{r}_1 \times \mathbf{r}_2$, as it corresponds to a vector in the opposite direction. The vector product is thus not commutative.

The *scalar triple product* $\mathbf{r}_1 \cdot (\mathbf{r}_2 \times \mathbf{r}_3)$ equals the scalar product of \mathbf{r}_1 with the vector product of \mathbf{r}_2 and \mathbf{r}_3 . The result is a scalar. The scalar triple product has a useful geometrical interpretation; it is the volume of the parallelepiped whose sides correspond to the three vectors (Figure 1.9).

1.10.3 Matrices, Eigenvectors and Eigenvalues

A matrix is a set of quantities arranged in a rectangular array. An $m \times n$ matrix has m rows and n columns. A vector can thus be considered to be a one-column matrix. Matrix addition

and subtraction can only be performed with matrices of the same order. For example:

$$\text{If } \mathbf{A} = \begin{pmatrix} 4 & 7 \\ -3 & 5 \\ 8 & -2 \end{pmatrix} \quad \text{and} \quad \mathbf{B} = \begin{pmatrix} -4 & 3 \\ 5 & 2 \\ -5 & 3 \end{pmatrix}$$

$$\text{Then } \mathbf{A} + \mathbf{B} = \begin{pmatrix} 0 & 10 \\ 2 & 7 \\ 3 & 1 \end{pmatrix}; \quad \mathbf{A} - \mathbf{B} = \begin{pmatrix} 8 & 4 \\ -8 & 3 \\ 12 & -5 \end{pmatrix} \quad (1.10)$$

Multiplication of two matrices (\mathbf{AB}) is only possible if the number of columns in \mathbf{A} is equal to the number of rows in \mathbf{B} . If \mathbf{A} is an $m \times n$ matrix and \mathbf{B} is an $n \times o$ matrix then the product \mathbf{AB} is an $m \times o$ matrix. Each element (i, j) in the matrix \mathbf{AB} is obtained by taking each of the n values in the i th row of \mathbf{A} and multiplying by the corresponding value in the j th column of \mathbf{B} . To illustrate with a simple example:

$$\text{If } \mathbf{A} = \begin{pmatrix} 3 & -2 & 5 \\ -3 & 4 & 1 \end{pmatrix} \quad \text{and} \quad \mathbf{B} = \begin{pmatrix} 0 & 3 \\ -2 & 4 \\ 1 & 6 \end{pmatrix}$$

Then

$$\begin{aligned} \mathbf{AB} &= \begin{pmatrix} (3 \times 0) + (-2 \times -2) + (5 \times 1) & (3 \times 3) + (-2 \times 4) + (5 \times 6) \\ (-3 \times 0) + (4 \times -2) + (1 \times 1) & (-3 \times 3) + (4 \times 4) + (1 \times 6) \end{pmatrix} \\ &= \begin{pmatrix} 9 & 31 \\ -7 & 13 \end{pmatrix} \end{aligned} \quad (1.11)$$

We shall often encounter square matrices, which have the same number of rows and columns. A diagonal matrix is a square matrix in which all the elements are zero except for those on the diagonal. The *unit* or *identity* matrix \mathbf{I} is a special type of diagonal matrix in which all the non-zero elements are 1; thus the 3×3 unit matrix is:

$$\mathbf{I} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad (1.12)$$

A matrix is *symmetric* if it is a square matrix with elements such that the elements above and below the diagonal are mirror images; $A_{ij} = A_{ji}$.

Multiplication of a matrix by its inverse gives the unit matrix:

$$\mathbf{A}^{-1} \mathbf{A} = \mathbf{I} \quad (1.13)$$

To compute the inverse of a square matrix it is necessary to first calculate its *determinant*, $|\mathbf{A}|$. The determinants of 2×2 and 3×3 matrices are calculated as follows:

$$\begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bd \quad (1.14)$$

$$\begin{aligned} \begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix} &= a \begin{vmatrix} e & f \\ h & i \end{vmatrix} - b \begin{vmatrix} d & f \\ g & i \end{vmatrix} + c \begin{vmatrix} d & e \\ g & h \end{vmatrix} \\ &= a(ei - hf) - b(di - fg) + c(dh - eg) \end{aligned} \quad (1.15)$$

For example:

$$\begin{vmatrix} 3 & 6 \\ -2 & 3 \end{vmatrix} = 21; \quad \begin{vmatrix} 4 & 2 & -2 \\ 2 & 5 & 0 \\ -2 & 0 & 3 \end{vmatrix} = 28 \quad (1.16)$$

As can be seen, the determinant of a 3×3 matrix can be written as a sum of determinants of 2×2 matrices, obtained by first selecting one of the rows or columns in the matrix (the top row was chosen in our example). For each element A_{ij} in this row, the row and column in which that number appears are deleted (i.e. the i th row and the j th column). This leaves a 2×2 matrix whose determinant is calculated and then multiplied by $(-1)^{i+j}$. The result of this calculation is called the *cofactor* of the element A_{ij} . For example, the cofactor of the element A_{12} in the 3×3 matrix

$$\mathbf{A} = \begin{pmatrix} 4 & 2 & -2 \\ 2 & 5 & 0 \\ -2 & 0 & 3 \end{pmatrix}$$

is -6 . When calculating the determinant the cofactor is multiplied by the element A_{ij} . The determinants of larger matrices can be obtained by extensions of the scheme illustrated above; thus the determinant of a 4×4 matrix is initially written in terms of 3×3 matrices, which in turn can be expressed in terms of 2×2 matrices.

Determinants have many useful and interesting properties. The determinant of a matrix is zero if any two of its rows or columns are identical. The sign of the determinant is reversed by exchanging any pair of rows or any pair of columns. If all elements of a row (or column) are multiplied by the same number, then the value of the determinant is multiplied by that number. The value of a determinant is unaffected if equal multiples of the values in any row (or column) are added to another row (or column).

The vector product and the scalar triple product can be conveniently written as matrix determinants. Thus:

$$\mathbf{r}_1 \times \mathbf{r}_2 = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \end{vmatrix} \quad (1.17)$$

$$\mathbf{r}_1 \cdot (\mathbf{r}_2 \times \mathbf{r}_3) = \begin{vmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ x_3 & y_3 & z_3 \end{vmatrix} \quad (1.18)$$

The *transpose* of a matrix, \mathbf{A}^T , is the matrix obtained by exchanging its rows and columns. Thus the transpose of an $m \times n$ matrix is an $n \times m$ matrix:

$$\text{If } \mathbf{A} = \begin{pmatrix} 4 & 7 \\ -3 & 5 \\ 8 & -2 \end{pmatrix} \quad \mathbf{A}^T = \begin{pmatrix} 4 & -3 & 8 \\ 7 & 5 & -2 \end{pmatrix} \quad (1.19)$$

The transpose of a square matrix is, of course, another square matrix. The transpose of a symmetric matrix is itself. One particularly important transpose matrix is the *adjoint* matrix, $\text{adj}\mathbf{A}$, which is the transpose matrix of cofactors. For example, the matrix of cofactors of the 3×3 matrix

$$\mathbf{A} = \begin{pmatrix} 4 & 2 & -2 \\ 2 & 5 & 0 \\ -2 & 0 & 3 \end{pmatrix} \quad \text{is} \quad \begin{pmatrix} 15 & -6 & 10 \\ -6 & 8 & -4 \\ 10 & -4 & 16 \end{pmatrix} \quad (1.20)$$

In this case the adjoint matrix is the same as the matrix of cofactors (as \mathbf{A} is a symmetric matrix). The *inverse* of a matrix is obtained by dividing the elements of the adjoint matrix by the determinant:

$$\mathbf{A}^{-1} = \frac{\text{adj}\mathbf{A}}{|\mathbf{A}|} \quad (1.21)$$

Thus the inverse of our 3×3 matrix is

$$\mathbf{A}^{-1} = \begin{pmatrix} 15/28 & -3/14 & 5/14 \\ -3/14 & 2/7 & -1/7 \\ 5/14 & -4 & 4/7 \end{pmatrix} \quad (1.22)$$

One of the most common matrix calculations involves finding its *eigenvalues* and *eigenvectors*. An eigenvector is a column matrix \mathbf{x} such that

$$\mathbf{Ax} = \lambda\mathbf{x} \quad (1.23)$$

λ is the associated eigenvalue. The eigenvector problem can be reformulated as follows:

$$\mathbf{Ax} = \lambda\mathbf{x}\mathbf{I} \Rightarrow \mathbf{Ax} - \lambda\mathbf{x}\mathbf{I} = 0 \Rightarrow (\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = 0 \quad (1.24)$$

A trivial solution to this equation is $\mathbf{x} = 0$. For a non-trivial solution, we require that the determinant $|\mathbf{A} - \lambda\mathbf{I}|$ equals zero. One way to determine the eigenvalues and their associated eigenvectors is thus to expand the determinant to give a polynomial equation in λ . For our 3×3 symmetric matrix this gives:

$$\begin{pmatrix} 4 - \lambda & 2 & -2 \\ 2 & 5 - \lambda & 0 \\ -2 & 0 & 3 - \lambda \end{pmatrix} \quad (1.25)$$

or:

$$(4 - \lambda)(5 - \lambda)(3 - \lambda) - 2[2(3 - \lambda)] - 2[2(5 - \lambda)] = 0 \quad (1.26)$$

This can be factorised to give:

$$(1 - \lambda)(7 - \lambda)(4 - \lambda) = 0 \quad (1.27)$$

The eigenvalues are thus $\lambda_1 = 1$, $\lambda_2 = 4$, $\lambda_3 = 7$. The corresponding eigenvectors are:

$$\lambda_1 = 1 : \mathbf{x}_1 = \begin{pmatrix} 2/3 \\ -1/3 \\ 2/3 \end{pmatrix} \quad \lambda_2 = 4 : \mathbf{x}_2 = \begin{pmatrix} -1/3 \\ 2/3 \\ 2/3 \end{pmatrix} \quad \lambda_3 = 7 : \mathbf{x}_3 = \begin{pmatrix} 2/3 \\ 2/3 \\ -1/3 \end{pmatrix} \quad (1.28)$$

Here we have expressed the eigenvectors as vectors of unit length; any multiple of each eigenvector would also be a solution. \mathbf{A} is a real, symmetric matrix. The eigenvalues of such matrices are always real and orthogonal (i.e. the scalar products of all pairs of eigenvectors are zero). This can be easily seen in our example.

As can be readily envisaged, expanding the determinant and solving a polynomial in λ is not the most efficient way to determine the eigenvalues and eigenvectors of larger matrices. Matrix diagonalisation methods are much more common. *Diagonalisation* of a matrix \mathbf{A} involves finding a matrix \mathbf{U} such that:

$$\mathbf{U}^{-1} \mathbf{A} \mathbf{U} = \mathbf{D} \quad (1.29)$$

\mathbf{D} is the diagonal matrix of eigenvalues. When \mathbf{A} is a real symmetric matrix, then \mathbf{U} is the matrix of eigenvectors and \mathbf{U}^{-1} is the inverse matrix of eigenvectors. Thus, for our example:

$$\begin{aligned} & \begin{pmatrix} 2/3 & -1/3 & 2/3 \\ -1/3 & 2/3 & 2/3 \\ 2/3 & 2/3 & -1/3 \end{pmatrix} \begin{pmatrix} 4 & 2 & -2 \\ 2 & 5 & 0 \\ -2 & 0 & 3 \end{pmatrix} \begin{pmatrix} 2/3 & -1/3 & 2/3 \\ -1/3 & 2/3 & 2/3 \\ 2/3 & 2/3 & -1/3 \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 7 \end{pmatrix} \end{aligned} \quad (1.30)$$

Note that for a real symmetric matrix \mathbf{A} , the inverse \mathbf{U}^{-1} is the same as the transpose, \mathbf{U}^T .

Many methods have been devised for diagonalising matrices; some of these are specific to certain classes of matrices such as the class of real symmetric matrices. Many modelling techniques require us to calculate the eigenvalues and eigenvectors of a matrix, including self-consistent field quantum mechanics (Section 2.5), the distance geometry method for exploring conformational space (Section 9.5) and principal components analysis (Section 9.13.1). The class of *positive definite* matrices is important in energy minimisation and when finding transition structures; the eigenvalues of a positive definite matrix are all positive. A *positive semidefinite* matrix of rank m has m positive eigenvalues.

1.10.4 Complex Numbers

A complex number has two components: a real part (a) and an imaginary part (b), as follows:

$$x = a + bi \quad (1.31)$$

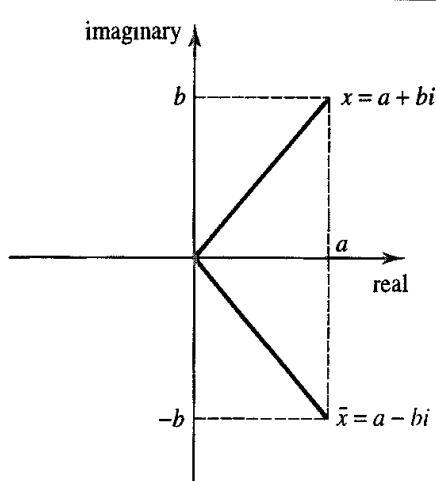


Fig. 1.10 The Argand diagram used to represent complex numbers

i is the square root of -1 ($i = \sqrt{-1}$). Complex numbers enable certain types of equation that have no real solutions to be solved. For example, the roots of the equation $x^2 - 2x + 3 = 0$ are $x = 1 + \sqrt{2}i$ and $x = 1 - \sqrt{2}i$. A complex number can be considered as a vector in a two-dimensional coordinate system. Complex numbers are commonly represented using an *Argand diagram*, in which the x coordinate corresponds to the real part of the complex number and the y coordinate to the imaginary part (Figure 1.10).

Arithmetical operations on complex numbers are performed much as for vectors. Thus, if $x = a + bi$ and $y = c + di$, then:

$$x + y = (a + c) + (b + d)i \quad (1.32)$$

$$x - y = (a - c) + (b - d)i \quad (1.33)$$

$$xy = (ac - bd) + (ad + bc)i \quad (1.34)$$

The *complex conjugate*, \bar{x} , equals $a - bi$ and is obtained by reflecting x in the real axis in the Argand diagram.

A commonly used relationship involving complex numbers is:

$$e^{i\theta} = \cos \theta + i \sin \theta \quad (1.35)$$

where θ is any real number. This relationship is used in Fourier analysis and can be derived from the expansions of the exponential, cosine and sine functions:

$$e^{i\theta} = 1 + i\theta - \frac{\theta^2}{2!} - \frac{i\theta^3}{3!} + \frac{\theta^4}{4!} + \dots \quad (1.36)$$

$$\sin \theta = \theta - \frac{\theta^3}{3!} + \frac{\theta^5}{5!} - \dots \quad (1.37)$$

$$\cos \theta = 1 - \frac{\theta^2}{2!} + \frac{\theta^4}{4!} - \dots \quad (1.38)$$

Various other relationships can be defined. For example:

$$\cos \theta = \frac{e^{i\theta} + e^{-i\theta}}{2} \quad \sin \theta = \frac{e^{i\theta} - e^{-i\theta}}{2i} \quad (1.39)$$

1.10.5 Lagrange Multipliers

Lagrange multipliers can be used to find the stationary points of functions, subject to a set of constraints. Suppose we wish to find the stationary points of a function $f(x, y) = 4x^2 + 3x + 2y^2 + 6y$ subject to the constraint $y = 4x + 2$. In the Lagrange method the constraint is written in the form $g(x, y) = 0$:

$$g(x, y) = y - 4x - 2 = 0 \quad (1.40)$$

To find stationary points $f(x, y)$ subject to $g(x, y) = 0$ we first determine the total derivative df , which is set equal to zero:

$$df = \frac{\partial f}{\partial x} dx + \frac{\partial f}{\partial y} dy = (8x + 3) dx + (4y + 6) dy = 0 \quad (1.41)$$

Without the constraint the stationary points would be determined by setting the two partial derivatives $\partial f / \partial x$ and $\partial f / \partial y$ equal to zero, as x and y are independent. With the constraint, x and y are no longer independent but are related via the derivative of the constraint function g :

$$dg = \frac{\partial g}{\partial x} dx + \frac{\partial g}{\partial y} dy = -4dx + dy = 0 \quad (1.42)$$

The derivative of the constraint function, dg , is multiplied by a parameter λ (the *Lagrange multiplier*) and added to the total derivative df :

$$\left(\frac{\partial f}{\partial x} + \lambda \frac{\partial g}{\partial x} \right) dx + \left(\frac{\partial f}{\partial y} + \lambda \frac{\partial g}{\partial y} \right) dy = 0 \quad (1.43)$$

The value of the Lagrange multiplier is obtained by setting each of the terms in parentheses to zero. Thus for our example we have:

$$8x + 3 - 4\lambda = 0 \quad (1.44)$$

$$4y + 6 + \lambda = 0 \quad (1.45)$$

From these two equations we can obtain a further equation linking x and y :

$$\lambda = 2x + 3/4 = -6 - 4y \quad \text{or} \quad x = -27/8 - 2y \quad (1.46)$$

Combining this with the constraint equation enables us to identify the stationary point, which is at $(-59/72, -23/18)$.

This simple example could, of course, have been solved by simply substituting the constraint equation into the original function, to give a function of just one of the variables. However, in many cases this is not possible. The Lagrange multiplier method provides a powerful approach which is widely applicable to problems involving constraints such as in constraint dynamics (Section 7.5) and in quantum mechanics.

1.10.6 Multiple Integrals

Many of the theories used in molecular modelling involve multiple integrals. Examples include the two-electron integrals found in Hartree-Fock theory, and the integral over the positions and momenta used to define the partition function, Q . In fact, most of the multiple integrals that have to be evaluated are double integrals.

A 'traditional' or one-dimensional integral corresponds to the area under the curve between the imposed limit, as illustrated in Figure 1.11. Multiple integrals are simply extensions of these ideas to more dimensions. We shall illustrate the principles using a function of two variables, $f(x, y)$. The double integral

$$\iint_A dx dy f(x, y) \equiv \iint_A f(x, y) dx dy \quad (1.47)$$

is the sum of the volume elements $f(x, y)\delta x \delta y$ (see Figure 1.11) over the area A as δx and δy tend to zero. Note that the ' $dx dy$ ' can be put either immediately after the integral sign or at the end; in this book we often use the first method for multiple integrals.

Some multiple integrals can be written as a product of single integrals. This occurs when $f(x, y)$ is itself a product of functions $g(x)h(y)$, in which case the integral can be separated:

$$\iint_A dx dy g(x)h(y) = \int dx g(x) \int dy h(y) \quad (1.48)$$

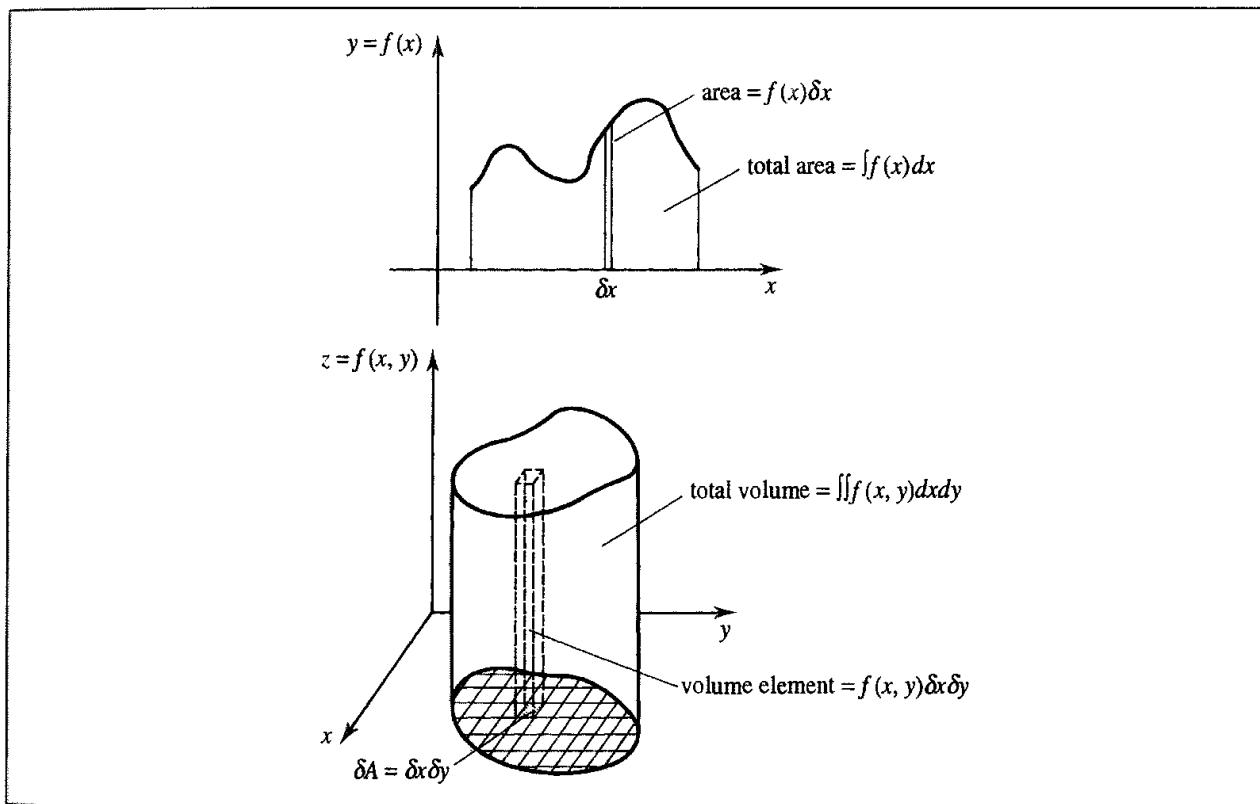


Fig. 1.11. Single and double integrals. (Figure adapted in part from Boas M L, 1983, *Mathematical Methods in the Physical Sciences*. 2nd Edition. New York, Wiley)

For example:

$$\int_{-1}^1 dx \int_{-\pi/2}^{+\pi/2} dy x^2 \cos y = \int_{-1}^1 x^2 dx [\sin y]_{-\pi/2}^{+\pi/2} = 2 \left(\frac{x^3}{3} \right)_{-1}^{+1} = \frac{4}{3} \quad (1.49)$$

We will use the separation of multiple integrals throughout our discussion of quantum mechanics and computer simulation methods (Chapters 2, 3, 6, 7 and 8).

1.10.7 Some Basic Elements of Statistics

Statistics is concerned with the collection and interpretation of numerical data. The subject is a vast and complex one, and all we shall do here is to state some of the definitions commonly used and to explain some of the terminology.

The *arithmetic mean* of a set of observations is the sum of the observations divided by the number of observations:

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i \quad (1.50)$$

N is the number of observations. The mean may also be written $\langle x \rangle$. The *variance*, σ^2 , indicates the extent to which the set of observations cluster around the mean value and equals the average of the squared deviations from the mean:

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2 \quad (1.51)$$

The variance can also be calculated using the following formula, which may be more convenient:

$$\sigma^2 = \frac{1}{N} \left[\sum_{i=1}^N (x_i^2) - \frac{1}{N} \left(\sum_{i=1}^N x_i \right)^2 \right] \quad (1.52)$$

The *standard deviation*, σ , equals the (positive) square root of the variance:

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2} \quad (1.53)$$

It is often desired to compare the distribution of observations in a population with a theoretical distribution. The *normal distribution* (also called the Gaussian distribution) is a particularly important theoretical distribution in molecular modelling. The probability density function for a general normal distribution is:

$$f(x) = \frac{1}{\sigma \sqrt{2\pi}} \exp[-(x - \bar{x})^2 / 2\sigma^2] \quad (1.54)$$

The factor before the exponential ensures that the integral of the function $f(x)$ from $-\infty$ to $+\infty$ equals 1. The distribution is often written in terms of a parameter α :

$$f(x) = \sqrt{\frac{\alpha}{\pi}} e^{-\alpha(x - \bar{x})^2} \quad (1.55)$$

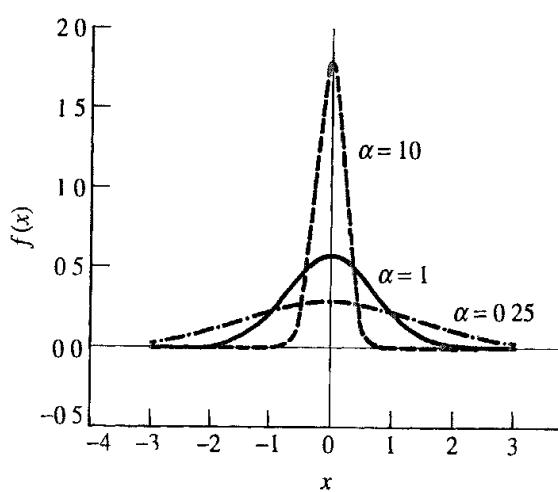


Fig 1.12. Three normal distributions with different values of α (Equation (1.55)) The functions are normalised, so the area under each curve is the same.

In Figure 1.12 we show three normal distributions that all have zero mean but different values of the variance (σ^2). A variance larger than 1 (small α) gives a flatter function and a variance less than 1 (larger α) gives a sharper function.

1.10.8 The Fourier Series, Fourier Transform and Fast Fourier Transform

Consider a periodic function $x(t)$ that repeats between $t = -\tau/2$ and $t = +\tau/2$ (i.e. has period τ). Even though $x(t)$ may not correspond to an analytical expression it can be written as the superposition of simple sine and cosine functions or *Fourier series*, Figure 1.13.

$$x(t) = a_0 + a_1 \cos \omega_0 t + a_2 \cos 2\omega_0 t + \dots + b_1 \sin \omega_0 t + b_2 \sin 2\omega_0 t + \dots \quad (1.56)$$

$$x(t) = a_0 + \sum_{n=1}^{\infty} (a_n \cos n\omega_0 t + b_n \sin n\omega_0 t) \quad (1.57)$$

ω_0 is related to the period of the function by $\omega_0 = 2\pi/\tau$ and to the frequency of the function by $\omega_0 = 2\tau\nu_0$. The frequencies of the contributing harmonics are thus $n\nu_0$ and are separated by $1/\tau$.

The coefficients a_n and b_n can be obtained as follows:

$$a_0 = \frac{1}{\tau} \int_{-\tau/2}^{\tau/2} x(t) dt \quad (1.58)$$

$$a_n = \frac{2}{\tau} \int_{-\tau/2}^{\tau/2} x(t) \cos(2n\pi x/\tau) dx \quad (1.59)$$

$$b_n = \frac{2}{\tau} \int_{-\tau/2}^{\tau/2} x(t) \sin(2n\pi x/\tau) dx \quad (1.60)$$

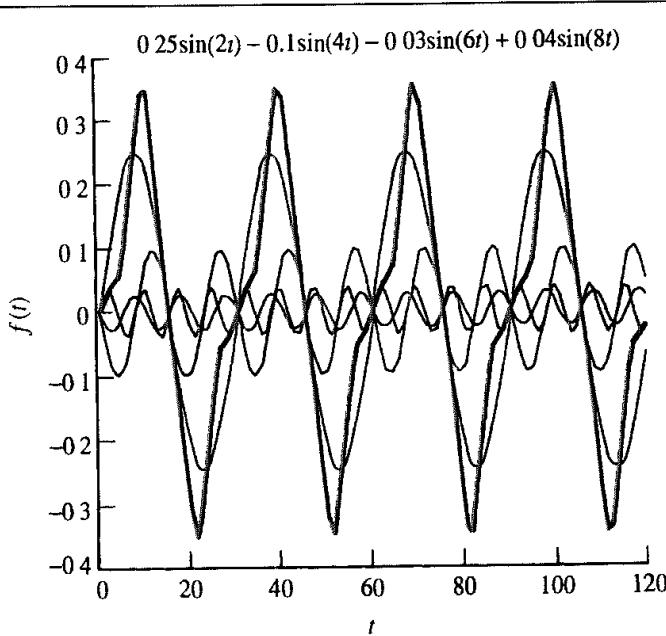


Fig. 1.13 In a Fourier series a periodic function is expressed as a sum of sine and cosine functions

An alternative way to express a Fourier series makes use of the following relationships:

$$\sin \omega_0 t = [\exp(i\omega_0 t) - \exp(-i\omega_0 t)]/2i \quad (1.61)$$

$$\cos \omega_0 t = [\exp(i\omega_0 t) + \exp(-i\omega_0 t)]/2 \quad (1.62)$$

The Fourier series is then written

$$x(t) = \sum_{-\infty}^{+\infty} c_n \exp(in\omega_0 t) \quad (1.63)$$

with

$$c_n = \frac{1}{\tau} \int_{-\tau/2}^{\tau/2} x(t) \exp(in\omega_0 t) dt \quad (1.64)$$

The Fourier series is used to represent a function that is periodic with period τ in terms of frequencies $n\omega_0 = 2\pi n/\tau$. The *Fourier transform* is used when the function has no periodicity. There is a close relationship between the Fourier series and the Fourier transform. One way to demonstrate the gradual change from a Fourier series to a Fourier transform is to consider how the distribution of contributing frequencies changes as the period increases. This is illustrated in Figure 1.14, where the period of a square wavefunction is gradually increased. Also shown are the frequency contributions. It can be seen that an increasing number of frequency components is needed to describe the function as the period increases, and that when the period is infinite, the frequency spectrum is continuous.

The Fourier transform relationship between a function $x(t)$ and the corresponding frequency function $X(\nu)$ is:

$$x(t) = \int_{-\infty}^{+\infty} X(\nu) \exp(2\pi i\nu t) d\nu \quad (1.65)$$

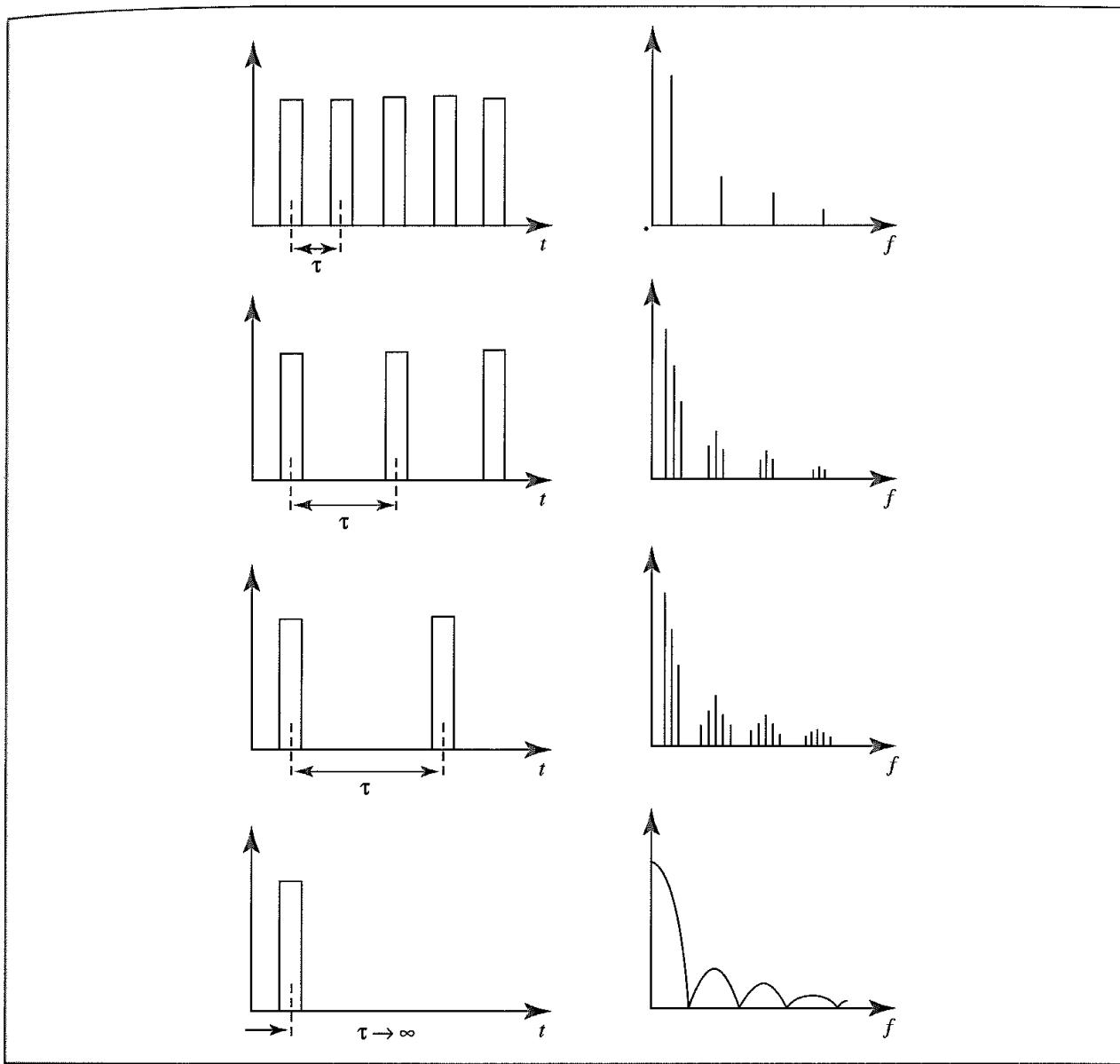


Fig. 1.14: The connection between the Fourier transform and the Fourier series can be established by gradually increasing the period of the function. When the period is infinite a continuous spectrum is obtained (Figure adapted from Ramirez R W, 1985, *The FFT Fundamentals and Concepts*. Englewood Cliffs, NJ, Prentice Hall.)

The frequency function $X(\nu)$ is given by:

$$X(\nu) = \int_{-\infty}^{+\infty} x(t) \exp(-2\pi i \nu t) dt \quad (1.66)$$

In practical applications, $x(t)$ is not a continuous function, and the data to be transformed are usually discrete values obtained by sampling at intervals. Under such circumstances, the discrete Fourier transform (DFT) is used to obtain the frequency function. Let us suppose that the time-dependent data values are obtained by sampling at regular intervals separated by δt and that a total of M samples are obtained (starting at $t = 0$). From M samples, a total of M frequency coefficients can be obtained using the DFT expression

[Press *et al.* 1992]:

$$X(k\delta\nu) = \delta t \sum_{n=0}^{M-1} x(n\delta t) \exp[-2\pi i nk/M] \quad (1.67)$$

Here, $x(n\delta t)$ ($n = 0, 1, \dots, M - 1$) are the experimental values obtained and $X(k\delta\nu)$ is the set of Fourier coefficients ($k = 0, 1, \dots, M - 1$). The separation between the frequencies, $\delta\nu$, depends on the number of samples and the time between samples: $\delta\nu = 1/M\delta t$. An expression for converting frequency data into the time domain is also possible:

$$x(n\delta t) = \frac{1}{M} \sum_{k=0}^{M-1} X(k\delta\nu) \exp[2\pi i nk/M] \quad (1.68)$$

To compute each Fourier coefficient $X(k\delta T)$ (of which there are M) it is therefore necessary to evaluate the summation $\sum_{n=0}^{M-1} x(n\delta t) \exp[-2\pi i nk/M]$ for that value of k . There will be M terms in the summation. A simple algorithm to determine the frequency spectrum would scale with the square of the number of measurements, M . This is a severe limitation, for many problems involve an extremely large number of pieces of data. It is for this reason that the fast Fourier transform (FFT) (ascribed to Cooley and Tukey [Cooley and Tukey 1965] but, in fact, using methods developed much earlier) has made such an impact. The FFT algorithm scales as $M \ln M$. With the FFT algorithm it is possible to derive the Fourier transforms, even with a considerable number of data points.

Further Reading

- Bachrach S M 1996. *The Internet: A Guide for Chemists* Washington, D.C., American Chemical Society
 Boas M L 1983. *Mathematical Methods in the Physical Sciences*. New York, John Wiley & Sons.
 Grant G H and W G Richards 1995. *Computational Chemistry* Oxford, Oxford University Press.
 Goodman J M 1998. *Chemical Applications of Molecular Modelling*. Cambridge, Royal Society of Chemistry.
 Leach A R 1999. Computational Chemistry and the Virtual Laboratory. In *The Age of the Molecule*. Cambridge, Royal Society of Chemistry.
 Lipkowitz K B and D B Boyd (Editors) 1990-. *Reviews in Computational Chemistry* Vols 1-. New York, VCH.
 Ramirez R W 1985. *The FFT Fundamentals and Concepts*. Englewood Cliffs, NJ, Prentice Hall.
 Schleyer, P v R, N L Allinger, T Clark, J Gasteiger, P A Kollman, H F Schaefer III and P R Schreiner 1998. *The Encyclopedia of Computational Chemistry*. Chichester, John Wiley & Sons.
 Stephenson G 1973. *Mathematical Methods for Science Students*. London, Longman.
 Winter M J, H S Rzepa and B J Whitaker 1995. Surfing the Chemical Net. *Chemistry in Britain* **31**: 685-689 and <http://www.ch.ic.ac.uk/rzepa/cib/>

References

- Bolin J T, D J Filman, D A Matthews, R C Hamlin and J Kraut 1982. Crystal Structures of *Escherichia coli* and *Lactobacillus casei* Dihydrofolate Reductase Refined at 1.7 Ångstroms Resolution. I. Features and Binding of Methotrexate. *Journal of Biological Chemistry* **257**: 13650-13662.

- Connolly M L 1983a. Solvent-accessible Surfaces of Proteins and Nucleic Acids. *Science* **221** 709–713
- Connolly M L 1983b Analytical Molecular Surface Calculation. *Journal of Applied Crystallography* **16** 548–558
- Cooley J W and J W Tukey 1965. An Algorithm for the Machine Calculation of Complex Fourier Series. *Mathematics of Computation* **19**:297–301.
- Lee B and F M Richards 1971. The Interpretation of Protein Structures: Estimation of Static Accessibility. *Journal of Molecular Biology* **55**,379–400.
- Murray-Rust P and H Rzepa 1999. Chemical Markup, XML, and the Worldwide Web. 1 Basic Principles. *Journal of Chemical Information and Computer Science* **39**,923–942
- Press W H, B P Flannery, S A Teukolsky and W T Vetterling 1992. *Numerical Recipes in Fortran*. Cambridge, Cambridge University Press.
- Richards F M 1977. Areas, Volumes, Packing and Protein Structure *Annual Review in Biophysics and Bioengineering* **6**:151–176.
- Richmond T J 1984. Solvent Accessible Surface Area and Excluded Volume in Proteins *Journal of Molecular Biology* **178**:63–88.
- Rouvray D 1997. Do Molecular Models Accurately Reflect Reality? *Chemist in Industry* **15**:587–590.
- Rouvray D 1999. Model Answers *Chemistry in Britain* **35**:30–32.
- Rouvray D 2000 Atoms as Hard Spheres. *Chemistry in Britain* **36**:25.

An Introduction to Computational Quantum Mechanics

2.1 Introduction

Our aim in this chapter will be to establish the basic elements of those quantum mechanical methods that are most widely used in molecular modelling. We shall assume some familiarity with the elementary concepts of quantum mechanics as found in most ‘general’ physical chemistry textbooks, but little else other than some basic mathematics (see Section 1.10). There are also many excellent introductory texts to quantum mechanics. In Chapter 3 we then build upon this chapter and consider more advanced concepts. Quantum mechanics does, of course, predate the first computers by many years, and it is a tribute to the pioneers in the field that so many of the methods in common use today are based upon their efforts. The early applications were restricted to atomic, diatomic or highly symmetrical systems which could be solved by hand. The development of quantum mechanical techniques that are more generally applicable and that can be implemented on a computer (thereby eliminating the need for much laborious hand calculation) means that quantum mechanics can now be used to perform calculations on molecular systems of real, practical interest. Quantum mechanics explicitly represents the electrons in a calculation, and so it is possible to derive properties that depend upon the electronic distribution and, in particular, to investigate chemical reactions in which bonds are broken and formed. These qualities, which differentiate quantum mechanics from the empirical force field methods described in Chapter 4, will be emphasised in our discussion of typical applications.

There are a number of quantum theories for treating molecular systems. The first we shall examine, and the one which has been most widely used, is *molecular orbital theory*. However, alternative approaches have been developed, some of which we shall also describe, albeit briefly. We will be primarily concerned with the *ab initio* and semi-empirical approaches to quantum mechanics but will also mention techniques such as Hückel theory and valence bond theory. An alternative approach to quantum mechanics, density functional theory, is considered in Chapter 3. Density functional theory has always enjoyed significant support from the materials science community but is increasingly used for molecular systems.

Quantum mechanics is often considered to be a difficult subject, and a cursory glance at the following pages in this chapter may simply serve to reinforce that view! However, if followed carefully it is possible to see how models that are developed for very simple

systems can be applied to much more complex systems. As a consequence our treatment does require some consideration of the mathematical background to the simplest and most common types of calculation. Our strategy in developing the underlying theory of molecular orbital quantum mechanical calculations is as follows. First, we revise some key features of quantum mechanics, including the hydrogen atom. We then discuss the functional form of an acceptable wavefunction for a molecular system and show how to calculate the energy of such a system from the wavefunction. This leads to the problem of determining the wavefunction itself and how this can be done using routine mathematical methods. We will then be in a position to understand how quantum mechanical calculations can be performed for ‘real’ systems and will have the background necessary to consider more advanced topics.

The starting point for any discussion of quantum mechanics is, of course, the Schrödinger equation. The full, time-dependent form of this equation is

$$\left\{ -\frac{\hbar^2}{2m} \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \right) + \mathcal{V} \right\} \Psi(\mathbf{r}, t) = i\hbar \frac{\partial \Psi(\mathbf{r}, t)}{\partial t} \quad (2.1)$$

Equation (2.1) refers to a single particle (e.g. an electron) of mass m which is moving through space (given by a position vector $\mathbf{r} = xi + yj + zk$) and time (t) under the influence of an external field \mathcal{V} (which might be the electrostatic potential due to the nuclei of a molecule). \hbar is Planck’s constant divided by 2π and i is the square root of -1 . Ψ is the *wavefunction* which characterises the particle’s motion; it is from the wavefunction that we can derive various properties of the particle. When the external potential \mathcal{V} is independent of time then the wavefunction can be written as the product of a spatial part and a time part: $\Psi(\mathbf{r}, t) = \psi(\mathbf{r})T(t)$. We shall only consider situations where the potential is independent of time, which enables the time-dependent Schrödinger equation to be written in the more familiar, time-independent form:

$$\left\{ -\frac{\hbar^2}{2m} \nabla^2 + \mathcal{V} \right\} \Psi(\mathbf{r}) = E\Psi(\mathbf{r}) \quad (2.2)$$

Here, E is the energy of the particle and we have used the abbreviation ∇^2 (pronounced ‘del-squared’).

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \quad (2.3)$$

It is usual to abbreviate the left-hand side of Equation (2.1) to $\mathcal{H}\Psi$, where \mathcal{H} is the *Hamiltonian operator*:

$$\mathcal{H} = -\frac{\hbar^2}{2m} \nabla^2 + \mathcal{V} \quad (2.4)$$

This reduces the Schrödinger equation to $\mathcal{H}\Psi = E\Psi$. To solve the Schrödinger equation it is necessary to find values of E and functions Ψ such that, when the wavefunction is operated upon by the Hamiltonian, it returns the wavefunction multiplied by the energy. The Schrödinger equation falls into the category of equations known as partial differential eigenvalue equations in which an operator acts on a function (the eigenfunction) and returns the

function multiplied by a scalar (the eigenvalue). A simple example of an eigenvalue equation is:

$$\frac{d}{dx}(y) = ry \quad (2.5)$$

The operator here is d/dx . One eigenfunction of this equation is $y = e^{ax}$ with the eigenvalue r being equal to a . Equation (2.5) is a first-order differential equation. The Schrödinger equation is a second-order differential equation as it involves the second derivative of Ψ . A simple example of an equation of this type is

$$\frac{d^2y}{dx^2} = ry \quad (2.6)$$

The solutions of Equation (2.6) have the form $y = A \cos kx + B \sin kx$, where A , B and k are constants. In the Schrödinger equation Ψ is the eigenfunction and E the eigenvalue.

2.1.1 Operators

The concept of an operator is an important one in quantum mechanics. The *expectation value* (which we can consider to be the average value) of a quantity such as the energy, position or linear momentum can be determined using an appropriate operator. The most commonly used operator is that for the energy, which is the Hamiltonian operator itself, \mathcal{H} . The energy can be determined by calculating the following integral:

$$E = \frac{\int \Psi^* \mathcal{H} \Psi d\tau}{\int \Psi^* \Psi d\tau} \quad (2.7)$$

The two integrals in Equation (2.7) are performed over all space (i.e. from $-\infty$ to $+\infty$ in the x , y and z directions). Note the use of the complex conjugate notation (Ψ^*), which reminds us that the wavefunction may be a complex number. This equation can be derived by pre-multiplying both sides of the Schrödinger equation, $\mathcal{H}\Psi = E\Psi$, by the complex conjugate of the wavefunction, Ψ^* , and integrating both sides over all space. Thus:

$$\int \Psi^* \mathcal{H} \Psi d\tau = \int \Psi^* E \Psi d\tau \quad (2.8)$$

E is a scalar and so can be taken outside the integral, thus leading to Equation (2.7). If the wavefunction is normalised then the denominator in Equation (2.7) will equal 1.

The Hamiltonian operator is composed of two parts that reflect the contributions of kinetic and potential energies to the total energy. The kinetic energy operator is

$$-\frac{\hbar^2}{2m} \nabla^2 \quad (2.9)$$

and the operator for the potential energy simply involves multiplication by the appropriate expression for the potential energy. For an electron in an isolated atom or molecule the potential energy operator comprises the electrostatic interactions between the electron and the nucleus and the interactions between the electron and the other electrons. For a

single electron and a single nucleus with Z protons the potential energy operator is thus:

$$\mathcal{V} = -\frac{Ze^2}{4\pi\epsilon_0 r} \quad (2.10)$$

Another operator is that for linear momentum along the x direction, which is

$$\frac{\hbar}{i} \frac{\partial}{\partial x} \quad (2.11)$$

The expectation value of this quantity can thus be obtained by evaluating the following integral:

$$p_x = \frac{\int \Psi^* \frac{\hbar}{i} \frac{\partial}{\partial x} \Psi d\tau}{\int \Psi^* \Psi d\tau} \quad (2.12)$$

2.1.2 Atomic Units

Quantum mechanics is primarily concerned with atomic particles: electrons, protons and neutrons. When the properties of such particles (e.g. mass, charge, etc.) are expressed in ‘macroscopic’ units then the value must usually be multiplied or divided by several powers of 10. It is preferable to use a set of units that enables the results of a calculation to be reported as ‘easily manageable’ values. One way to achieve this would be to multiply each number by an appropriate power of 10. However, further simplification can be achieved by recognising that it is often necessary to carry quantities such as the mass of the electron or electronic charge all the way through a calculation. These quantities are thus also incorporated into the atomic units. The atomic units of length, mass and energy are as follows:

1 unit of charge equals the absolute charge on an electron, $|e| = 1.60219 \times 10^{-19} \text{ C}$

1 mass unit equals the mass of the electron, $m_e = 9.10593 \times 10^{-31} \text{ kg}$

1 unit of length (1 Bohr) is given by $a_0 = \hbar^2/4\pi^2 m_e e^2 = 5.29177 \times 10^{-11} \text{ m}$

1 unit of energy (1 Hartree) is given by $E_a = e^2/4\pi\epsilon_0 a_0 = 4.35981 \times 10^{-18} \text{ J}$

The atomic unit of length is the radius of the first orbit in Bohr’s treatment of the hydrogen atom. It also turns out to be the most probable distance of a 1s electron from the nucleus in the hydrogen atom. The atomic unit of energy corresponds to the interaction between two electronic charges separated by the Bohr radius. The total energy of the 1s electron in the hydrogen atom equals -0.5 Hartree. In atomic units Planck’s constant $\hbar = 2\pi$ and so $\hbar \equiv 1$.

2.1.3 Exact Solutions to the Schrödinger Equation

The Schrödinger equation can be solved exactly for only a few problems, such as the particle in a box, the harmonic oscillator, the particle on a ring, the particle on a sphere and the hydrogen atom, all of which are dealt with in introductory textbooks. A common feature of these problems is that it is necessary to impose certain requirements (often called *boundary*

conditions) on possible solutions to the equation. Thus, for a particle in a box with infinitely high walls, the wavefunction is required to go to zero at the boundaries. For a particle on a ring the wavefunction must have a periodicity of 2π because it must repeat every traversal of the ring. An additional requirement on solutions to the Schrödinger equation is that the wavefunction at a point r , when multiplied by its complex conjugate, is the probability of finding the particle at the point (this is the Born interpretation of the wavefunction). The square of an electronic wavefunction thus gives the electron density at any given point. If we integrate the probability of finding the particle over all space, then the result must be 1 as the particle must be somewhere:

$$\int \Psi^* \Psi d\tau = 1 \quad (2.13)$$

$d\tau$ indicates that the integration is over all space. Wavefunctions which satisfy this condition are said to be *normalised*. It is usual to require the solutions to the Schrödinger equation to be orthogonal:

$$\int \Psi_m^* \Psi_n d\tau = 0 \quad (m \neq n) \quad (2.14)$$

A convenient way to express both the orthogonality of different wavefunctions and the normalisation conditions uses the *Kronecker delta*:

$$\int \Psi_m^* \Psi_n d\tau = \delta_{mn} \quad (2.15)$$

When used in this context, the Kronecker delta can be taken to have a value of 1 if m equals n and zero otherwise. Wavefunctions that are both orthogonal and normalised are said to be *orthonormal*.

2.2 One-electron Atoms

In an atom that contains a single electron, the potential energy depends upon the distance between the electron and the nucleus as given by the Coulomb equation. The Hamiltonian thus takes the following form:

$$\mathcal{H} = -\frac{\hbar^2}{2m} \nabla^2 - \frac{Ze^2}{4\pi\epsilon_0 r} \quad (2.16)$$

In atomic units the Hamiltonian is:

$$\mathcal{H} = -\frac{1}{2} \nabla^2 - \frac{Z}{r} \quad (2.17)$$

For the hydrogen atom, the nuclear charge, Z , equals +1. r is the distance of the electron from the nucleus. The helium cation, He^+ , is also a one-electron atom but has a nuclear charge of +2. As atoms have spherical symmetry it is more convenient to transform the Schrödinger equation to polar coordinates r , θ and ϕ , where r is the distance from the nucleus (located at the origin), θ is the angle to the z axis and ϕ is the angle from the x axis in the xy plane (Figure 2.1). The solutions can be written as the product of a radial function $R(r)$, which depends only on r , and an angular function $Y(\theta, \phi)$ called a *spherical harmonic*, which

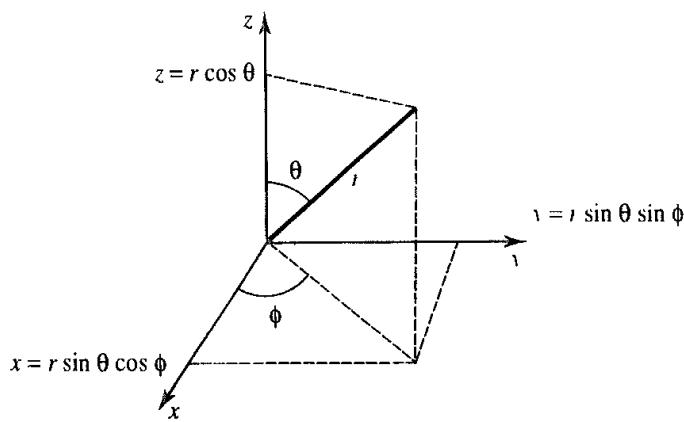


Fig. 2.1: The relationship between spherical polar and Cartesian coordinates

depends on θ and ϕ :

$$\Psi_{nlm} = R_{nl}(r)Y_{lm}(\theta, \phi) \quad (2.18)$$

The wavefunctions are commonly referred to as *orbitals* and are characterised by three quantum numbers n , m and l . The quantum numbers can adopt values as follows:

n : principal quantum number: 0, 1, 2, ...

l : azimuthal quantum number: 0, 1, ..., $(n - 1)$

m : magnetic quantum number: $-l, -(l - 1), \dots, 0, \dots, (l - 1), l$.

The full radial function is:

$$R_{nl}(r) = -\left[\left(\frac{2Z}{na_0}\right)^3 \frac{(n-l-1)!}{2n[(n+l)!]^3}\right]^{1/2} \exp\left(-\frac{\rho}{2}\right) \rho^l L_{n+1}^{2l+1}(\rho) \quad (2.19)$$

$\rho = 2Zr/na_0$, where a_0 is the Bohr radius.* The term in square brackets is a normalising factor. $L_{n+1}^{2l+1}(\rho)$ is a special type of function called a Laguerre polynomial. We shall rarely be interested in any other than the first few members of the series; moreover, they simplify considerably if atomic units are used and we write them in terms of the *orbital exponent* $\zeta = Z/n$. The first few members of the series for low values of n are given in Table 2.1 and are illustrated graphically in Figure 2.2. As can be seen, the radial part of the wavefunction is a polynomial multiplied by a decaying exponential.

The angular part of the wavefunction is the product of a function of θ and a function of ϕ :

$$Y_{lm}(\theta, \phi) = \Theta_{lm}(\theta)\Phi_m(\phi) \quad (2.20)$$

These functions are:

$$\Phi_m(\phi) = \frac{1}{\sqrt{2\pi}} \exp(im\phi) \quad (2.21)$$

$$\Theta_{lm}(\theta) = \left[\frac{(2l+1)}{2} \frac{(l-|m|)!}{(l+|m|)!} \right]^{1/2} P_l^{|m|}(\cos \theta) \quad (2.22)$$

* Strictly, a_0 in this case is given by $a_0 = h^2/\pi^2\mu e$, where μ is the reduced mass, $\mu = m_e M / (m_e + M)$, M is the mass of the nucleus.

<i>n</i>	<i>l</i>	$R_{nl}(r)$
1	0	$2\zeta^{3/2} \exp(-\zeta r)$
2	0	$2\zeta^{3/2}(1 - \zeta r) \exp(-\zeta r)$
2	1	$(4/3)^{1/2}\zeta^{5/2}r \exp(-\zeta r)$
3	0	$(2/3)^{1/2}\zeta^{3/2}(3 - 6\zeta r + 2\zeta^2 r^2) \exp(-\zeta r)$
3	1	$(8/9)^{1/2}\zeta^{5/2}(2 - \zeta r)r \exp(-\zeta r)$
3	2	$(8/45)^{1/2}\zeta^{7/2}r^2 \exp(-\zeta r)$

Table 2.1 Radial function for one-electron atoms.

The functions $\Phi_m(\phi)$ are just the solutions to the Schrödinger equation for a particle on a ring. The term in square brackets for the function $\Theta_{lm}(\theta)$ is a normalising factor. $P_l^{|m|}(\cos \theta)$ is a member of a series of functions called the associated Legendre polynomials (the ‘Legendre polynomials’ are functions for which $|m| = 0$). The total orbital angular momentum of an electron in the orbital is given by $l(l+1)\hbar$ and the component of the angular momentum along the $\theta = 0$ axis is given by $l\hbar$. The energy of each solution is a function of the principal quantum number only; thus orbitals with the same value of n but different l and m are degenerate. The orbitals are often represented as shown in Figure 2.3. These graphical representations are not necessarily the same as the solutions given above. For example, the ‘correct’ solutions for the 2p orbitals comprise one real and two complex functions:

$$2p(+1) = \sqrt{3/4\pi}R(r) \sin \theta e^{i\phi} \quad (2.23)$$

$$2p(0) = \sqrt{3/4\pi}R(r) \cos \theta \quad (2.24)$$

$$2p(-1) = \sqrt{3/4\pi}R(r) \sin \theta e^{-i\phi} \quad (2.25)$$

$R(r)$ is the radial part of the wavefunction and $\sqrt{3/4\pi}$ is a normalisation factor for the angular part. The $2p(0)$ function is real and corresponds to the $2p_z$ orbital that is pictured in Figure 2.3. A linear combination of the two remaining 2p solutions is used to generate two ‘real’ 2p wavefunctions, making use of the relationship $\exp(i\phi) = \cos \phi + i \sin \phi$ (Section 1.10.4). These linear combinations are the $2p_x$ and $2p_y$ orbitals shown in Figure 2.3.

$$2p_x = 1/2[2p(+1) + 2p(-1)] = \sqrt{3/4\pi}R(r) \sin \theta \cos \phi \quad (2.26)$$

$$2p_y = -1/2[2p(+1) - 2p(-1)] = \sqrt{3/4\pi}R(r) \sin \theta \sin \phi \quad (2.27)$$

These linear combinations still have the same energy as the original complex wavefunctions. This is a general property of degenerate solutions of the Hamiltonian operator. The reason why they are labelled $2p_x$ and $2p_y$ is that in polar coordinates the Cartesian coordinates x , y and z have the same angular dependence as the orbitals in Figure 2.3:

$$x = r \sin \theta \cos \phi \quad (2.28)$$

$$y = r \sin \theta \sin \phi \quad (2.29)$$

$$z = r \cos \theta \quad (2.30)$$

The solutions of the Schrödinger equation are either real or occur in degenerate pairs. These pairs are complex conjugates that can then be combined to give energetically equivalent real solutions. It is only when dealing with certain types of operator that it is necessary to retain a

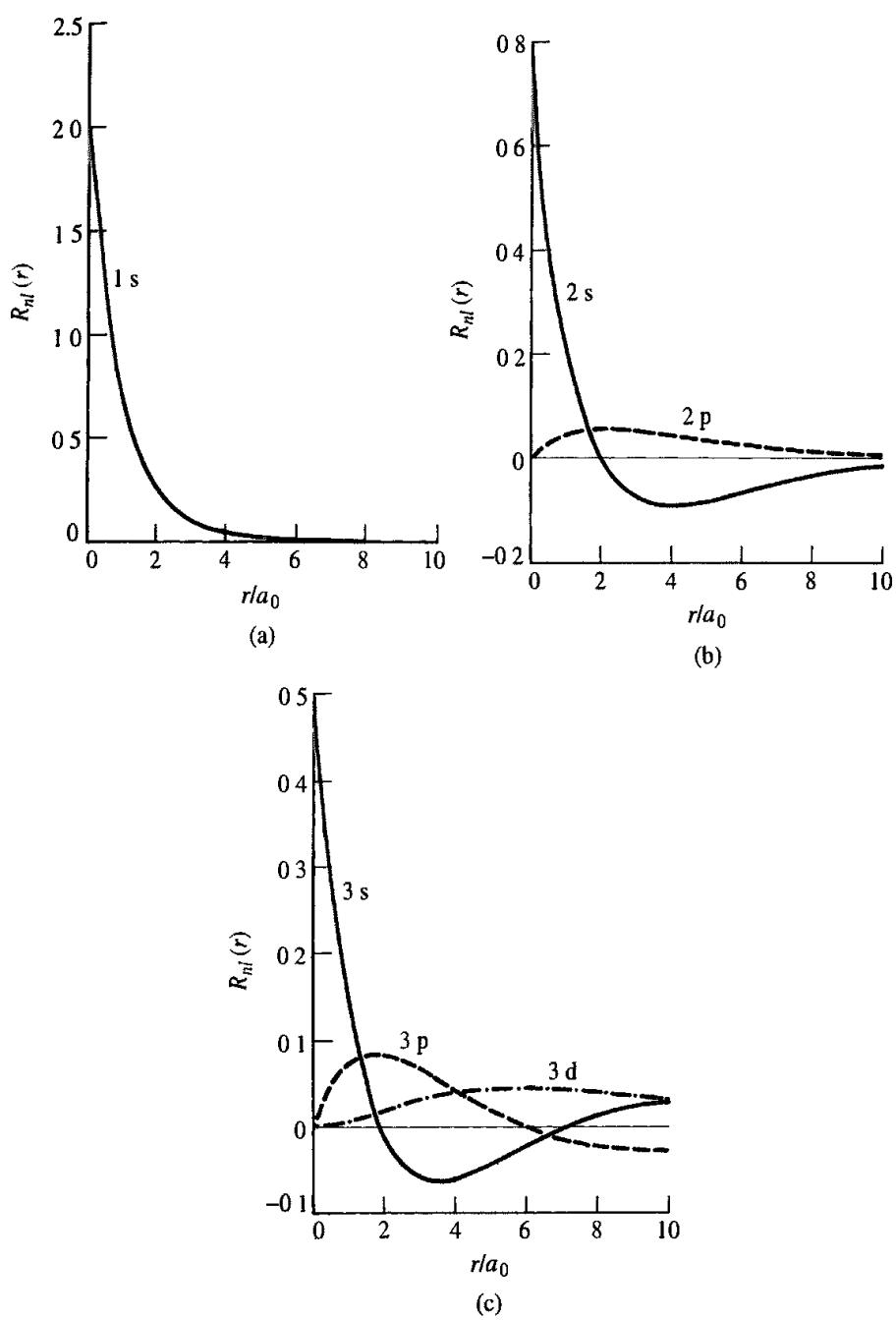


Fig. 2.2. The functions $R_{nl}(r)$ for the first three values of the principal quantum number. (a) 1s; (b) 2s and 2p; (c) 3s, 3p and 3d.

complex wavefunction (for the 2p functions, the operator that corresponds to angular momentum about the z axis falls into this category). In fact, to simplify matters we will almost always ignore the complex notation from now on and will deal with real orbitals.

Finally, we should note that the solutions are all orthogonal to each other; if the product of any pair of orbitals is integrated over all space, the result is zero unless the two orbitals are the same. Orthonormality is achieved by multiplying by an appropriate normalisation constant.

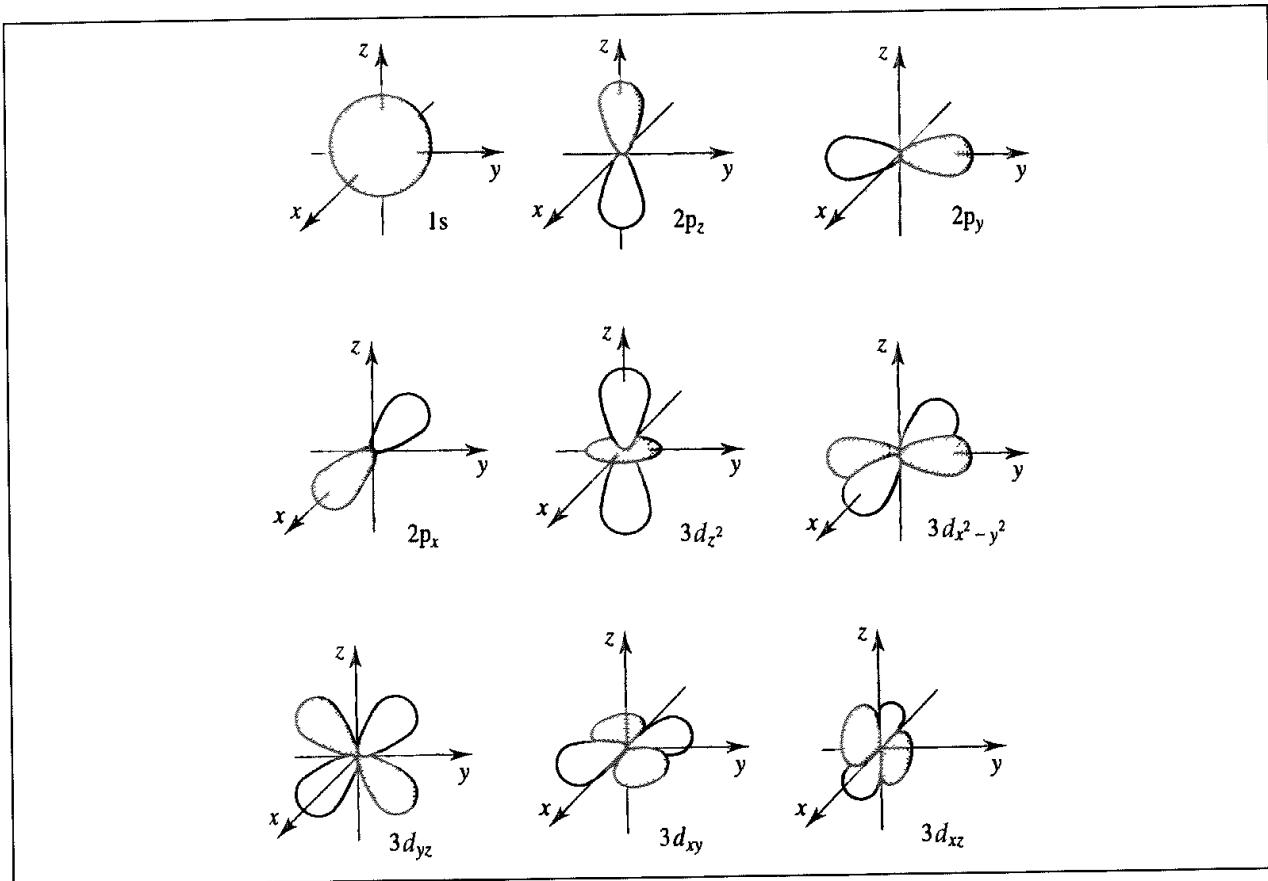


Fig 2.3: The common graphical representations of s, p and d orbitals

The orbital picture has proved invaluable for providing insight and qualitative interpretations into the nature of the bonding in and reactivity of chemical systems. It is one which we would like to retain for polyelectronic systems to provide a unifying theme that links the simplest systems with much more complicated ones.

2.3 Polyelectronic Atoms and Molecules

Solving the Schrödinger equation for atoms with more than one electron is complicated by a number of factors. The first complication is that the Schrödinger equation for such systems cannot be solved exactly, even for the helium atom. The helium atom has three particles (two electrons and one nucleus) and is an example of a *three-body problem*. No exact solutions can be found for systems that involve three (or more) interacting particles. Thus, any solutions we might find for polyelectronic atoms or molecules can only be approximations to the real, true solutions of the Schrödinger equation. One consequence of there being no exact solution is that the wavefunction may adopt more than one functional form; no form is necessarily more 'correct' than another. In fact, the most general form of the wavefunction will be an infinite series of functions.

A second complication with multi-electron species is that we must account for electron spin. Spin is characterised by the quantum number s , which for an electron can only take the

value $\frac{1}{2}$. The spin angular momentum is quantised such that its projection on the z axis is either $+\hbar$ or $-\hbar$. These two states are characterised by the quantum number m_s , which can have values of $+\frac{1}{2}$ or $-\frac{1}{2}$, and are often referred to as ‘up spin’ and ‘down spin’ respectively. Electron spin is incorporated into the solutions to the Schrödinger equation by writing each one-electron wavefunction as the product of a spatial function that depends on the coordinates of the electron and a spin function that depends on its spin. Such solutions are called *spin orbitals*, which we will represent using the symbol χ . The spatial part (which will be referred to as an orbital and represented using ϕ for atomic orbitals and ψ for molecular orbitals) describes the distribution of electron density in space and is analogous to the orbital diagrams in Figure 2.3. The spin part defines the electron spin and is labelled α or β . These spin functions have the value 0 or 1 depending on the quantum number m_s of the electron. Thus $\alpha(\frac{1}{2}) = 1$, $\alpha(-\frac{1}{2}) = 0$, $\beta(+\frac{1}{2}) = 0$, $\beta(-\frac{1}{2}) = 1$. Each spatial orbital can accommodate two electrons, with paired spins. In order to predict the electronic structure of a polyelectronic atom or a molecule, the *Aufbau principle* is employed, in which electrons are assigned to the orbitals, two electrons per orbital. We need to remember that electrons occupy degenerate states with a maximum number of unpaired electrons (Hund’s rules), and that there are certain situations where it is energetically more favourable to place an unpaired electron in a higher-energy spatial orbital rather than pair it with another electron. However, such situations are rare, particularly for molecular systems, and for most of the situations that we shall be interested in the number of electrons, N , will be an even number that occupy the $N/2$ lowest-energy orbitals.

Electrons are indistinguishable. If we exchange any pair of electrons, then the distribution of electron density remains the same. According to the Born interpretation, the electron density is equal to the square of the wavefunction. It therefore follows that the wavefunction must either remain unchanged when two electrons are exchanged, or else it must change sign. In fact, for electrons the wavefunction is required to change sign: this is the *antisymmetry principle*.

2.3.1 The Born–Oppenheimer Approximation

It was stated above that the Schrödinger equation cannot be solved exactly for any molecular systems. However, it is possible to solve the equation exactly for the simplest molecular species, H_2^+ (and isotopically equivalent species such as HD^+), when the motion of the electrons is decoupled from the motion of the nuclei in accordance with the Born–Oppenheimer approximation. The masses of the nuclei are much greater than the masses of the electrons (the resting mass of the lightest nucleus, the proton, is 1836 times heavier than the resting mass of the electron). This means that the electrons can adjust almost instantaneously to any changes in the positions of the nuclei. The electronic wavefunction thus depends only on the positions of the nuclei and not on their momenta. Under the Born–Oppenheimer approximation the total wavefunction for the molecule can be written in the following form:

$$\Psi_{\text{tot}}(\text{nuclei, electrons}) = \Psi(\text{electrons})\Psi(\text{nuclei}) \quad (2.31)$$

The total energy equals the sum of the nuclear energy (the electrostatic repulsion between the positively charged nuclei) and the electronic energy. The electronic energy comprises

the kinetic and potential energy of the electrons moving in the electrostatic field of the nuclei, together with electron-electron repulsion: $E_{\text{tot}} = E(\text{electrons}) + E(\text{nuclei})$.

When the Born-Oppenheimer approximation is used we concentrate on the electronic motions; the nuclei are considered to be fixed. For each arrangement of the nuclei the Schrödinger equation is solved for the electrons alone in the field of the nuclei. If it is desired to change the nuclear positions then it is necessary to add the nuclear repulsion to the electronic energy in order to calculate the total energy of the configuration.

2.3.2 The Helium Atom

We now return to the helium atom, our objective being to find a wavefunction that describes the behaviour of the electrons. The Born-Oppenheimer approximation is not, of course, relevant to systems with just one nucleus, and the wavefunction will be a function of the two electrons (which we shall label 1 and 2 with positions in space \mathbf{r}_1 and \mathbf{r}_2). As noted above, for polyelectronic systems any solution we find can only ever be an approximation to the true solution. There are a number of ways in which approximate solutions to the Schrödinger equation can be found. One approach is to find a simpler but related problem that can be more easily solved and then consider how the differences between the two problems change the Hamiltonian and thereby affect the solutions. This is called *perturbation theory* and is most appropriate when the differences between the real and simple problems are small. For example, a perturbation approach to tackling the helium atom might choose as the related system a 'pseudo atom', containing two electrons that interact with the nucleus but not with each other. Although this is a 'three-body' problem, the lack of any interaction between the electrons means that it can be solved exactly using the method of the separation of variables. The separation of variables technique can be applied whenever the Hamiltonian can be divided into parts that are themselves dependent solely upon subsets of the coordinates. The equation to be solved in this case is:

$$\left\{ -\frac{\hbar^2}{2m} \nabla_1^2 - \frac{Ze^2}{4\pi\epsilon_0 r_1} - \frac{\hbar^2}{2m} \nabla_2^2 - \frac{Ze^2}{4\pi\epsilon_0 r_2} \right\} \Psi(\mathbf{r}_1, \mathbf{r}_2) = E\Psi(\mathbf{r}_1, \mathbf{r}_2) \quad (2.32)$$

Or, in atomic units,

$$\left\{ -\frac{1}{2} \nabla_1^2 - \frac{Z}{r_1} - \frac{1}{2} \nabla_2^2 - \frac{Z}{r_2} \right\} \Psi(\mathbf{r}_1, \mathbf{r}_2) = E\Psi(\mathbf{r}_1, \mathbf{r}_2) \quad (2.33)$$

We can abbreviate this equation to

$$\{\mathcal{H}_1 + \mathcal{H}_2\}\Psi(\mathbf{r}_1, \mathbf{r}_2) = E\Psi(\mathbf{r}_1, \mathbf{r}_2) \quad (2.34)$$

\mathcal{H}_1 and \mathcal{H}_2 are the individual Hamiltonians for electrons 1 and 2. Let us assume that the wavefunction can be written as a product of individual one-electron wavefunctions, $\phi_1(\mathbf{r}_1)$ and $\phi_2(\mathbf{r}_2)$: $\Psi(\mathbf{r}_1, \mathbf{r}_2) = \phi_1(\mathbf{r}_1)\phi_2(\mathbf{r}_2)$. Then we can write:

$$[\mathcal{H}_1 + \mathcal{H}_2]\phi_1(\mathbf{r}_1)\phi_2(\mathbf{r}_2) = E\phi_1(\mathbf{r}_1)\phi_2(\mathbf{r}_2) \quad (2.35)$$

Premultiplying by $\phi_1(\mathbf{r}_1)\phi_2(\mathbf{r}_2)$ and integrating over all space gives:

$$\iint d\tau_1 d\tau_2 \phi_1(\mathbf{r}_1)\phi_2(\mathbf{r}_2)[\mathcal{H}_1 + \mathcal{H}_2]\phi_1(\mathbf{r}_1)\phi_2(\mathbf{r}_2) = \iint d\tau_1 d\tau_2 \phi_1(\mathbf{r}_1)\phi_2(\mathbf{r}_2)\phi_1(\mathbf{r}_1)\phi_2(\mathbf{r}_2) \quad (2.36)$$

or

$$\begin{aligned} & \int d\tau_1 \phi_1(\mathbf{r}_1) \mathcal{H}_1 \phi_1(\mathbf{r}_1) \int d\tau_2 \phi_2(\mathbf{r}_2) \phi_2(\mathbf{r}_2) + \int d\tau_1 \phi_1(\mathbf{r}_1) \phi_1(\mathbf{r}_1) \int d\tau_2 \phi_2(\mathbf{r}_2) \mathcal{H}_2 \phi_2(\mathbf{r}_2) \\ &= E \int d\tau_1 \phi_1(\mathbf{r}_1) \phi_1(\mathbf{r}_1) \int d\tau_2 \phi_2(\mathbf{r}_2) \phi_2(\mathbf{r}_2) \end{aligned} \quad (2.37)$$

If we assume that the wavefunctions are normalised then it can easily be seen that the total energy E is the sum of the individual orbital energies E_1 and E_2 ($E_1 = \int d\tau_1 \phi_1(\mathbf{r}_1) \mathcal{H}_1 \phi_1(\mathbf{r}_1)$ and $E_2 = \int d\tau_2 \phi_2(\mathbf{r}_2) \mathcal{H}_2 \phi_2(\mathbf{r}_2)$). When the separation of variables method is used the solutions for each electron are just those of the hydrogen atom (1s, 2s, etc.) in Equation (2.19) with $Z = 2$.

We now wish to establish the general functional form of possible wavefunctions for the two electrons in this pseudo helium atom. We will do so by considering first the spatial part of the wavefunction. We will show how to derive functional forms for the wavefunction in which the exchange of electrons is independent of the electron labels and does not affect the electron density. The simplest approach is to assume that each wavefunction for the helium atom is the product of the individual one-electron solutions. As we have just seen, this implies that the total energy is equal to the sum of the one-electron orbital energies, which is not correct as it ignores electron-electron repulsion. Nevertheless, it is a useful illustrative model. The wavefunction of the lowest energy state then has each of the two electrons in a 1s orbital:

$$1s(1)1s(2) \quad (2.38)$$

'1s(1)' indicates a 1s function that depends on the coordinates of electron 1 (\mathbf{r}_1) and '1s(2)' indicates a 1s function that depends upon the coordinates of electron 2 (\mathbf{r}_2). This wavefunction satisfies the indistinguishability criterion, for we obtain the same function when we exchange the electrons - 1s(1)1s(2) is the same as 1s(2)1s(1). Its energy is twice that of a single electron in a 1s orbital. What of the first excited state, in which one electron is promoted to the 2s orbital? Two possible wavefunctions for this state are:

$$1s(1)2s(2) \quad (2.39)$$

$$1s(2)2s(1) \quad (2.40)$$

Do these wavefunctions satisfy the indistinguishability criterion? In other words, do we get the same function (or its negative) when we exchange the electrons? We do not, for when the two electrons (1 and 2) are exchanged then a different wavefunction is obtained: '1s(1)2s(2)' and '1s(2)2s(1)' are not the same, nor is one simply minus the other. However, linear combinations of these two wavefunctions do not suffer from the labelling problem and so we might anticipate that functional forms such as the following might constitute acceptable solutions to the Schrödinger equation for the pseudo helium atom:

$$(1/\sqrt{2})[1s(1)2s(2) + 1s(2)2s(1)] \quad (2.41)$$

$$(1/\sqrt{2})[1s(1)2s(2) - 1s(2)2s(1)] \quad (2.42)$$

The factor $(1/\sqrt{2})$ ensures that the wavefunction is normalised. Of the three acceptable spatial forms that we have described so far, two are symmetric (i.e. do not change sign when the electron labels are exchanged) and one is antisymmetric (the sign changes when the electrons are exchanged):

$$1s(1)1s(2) \quad \text{symmetric} \quad (2.43)$$

$$(1/\sqrt{2})[1s(1)2s(2) + 1s(2)2s(1)] \quad \text{symmetric} \quad (2.44)$$

$$(1/\sqrt{2})[1s(1)2s(2) - 1s(2)2s(1)] \quad \text{antisymmetric} \quad (2.45)$$

We now need to consider the effects of electron spin. For two electrons 1 and 2 there are four spin states; $\alpha(1), \beta(1), \alpha(2), \beta(2)$. The indistinguishability criterion holds for the spin components as well, and so the following combinations of spin wavefunctions are possible.

$$\alpha(1)\alpha(2) \quad \text{symmetric} \quad (2.46)$$

$$\beta(1)\beta(2) \quad \text{symmetric} \quad (2.47)$$

$$(1/\sqrt{2})[\alpha(1)\beta(2) + \alpha(2)\beta(1)] \quad \text{symmetric} \quad (2.48)$$

$$(1/\sqrt{2})[\alpha(1)\beta(2) - \alpha(2)\beta(1)] \quad \text{antisymmetric} \quad (2.49)$$

When we combine the spatial and spin wavefunctions, the overall wavefunction must be antisymmetric with respect to exchange of electrons. It is therefore only admissible to combine a symmetric spatial part with an antisymmetric spin part, or an antisymmetric spatial part with a symmetric spin part. The following functional forms are therefore permissible functional forms for the wavefunctions of the ground and first few excited states of the helium atom:

$$(1/\sqrt{2})1s(1)1s(2)[\alpha(1)\beta(2) - \alpha(2)\beta(1)] \quad (2.50)$$

$$(1/2)[1s(1)2s(2) + 1s(2)2s(1)][\alpha(1)\beta(2) - \alpha(2)\beta(1)] \quad (2.51)$$

$$(1/\sqrt{2})[1s(1)2s(2) - 1s(2)2s(1)]\alpha(1)\alpha(2) \quad (2.52)$$

$$(1/\sqrt{2})[1s(1)2s(2) - 1s(2)2s(1)]\beta(1)\beta(2) \quad (2.53)$$

$$(1/2)[1s(1)2s(2) - 1s(2)2s(1)][\alpha(1)\beta(2) + \alpha(2)\beta(1)] \quad (2.54)$$

2.3.3 General Polyelectronic Systems and Slater Determinants

We now turn to the general case. What is an appropriate functional form of the wavefunction for a polyelectronic system (not necessarily an atom) with N electrons that satisfies the anti-symmetry principle? First, we note that the following functional form of the wavefunction is inappropriate:

$$\Psi(1, 2, \dots, N) = \chi_1(1)\chi_2(2) \dots \chi_N(N) \quad (2.55)$$

This product of spin orbitals is unacceptable because it does not satisfy the antisymmetry principle; exchanging pairs of electrons does not give the negative of the wavefunction. This formulation of the wavefunction is known as a *Hartree product*. The energy of a system described by a Hartree product equals the sum of the one-electron spin orbitals. A key conclusion of the Hartree product description is that the probability of finding an electron at a particular point in space is independent of the probability of finding any

other electron at that point in space. In fact, it turns out that the motions of the electrons are correlated. In addition, the Hartree product assumes that specific electrons have been assigned to specific orbitals, whereas the antisymmetry principle requires that the electrons are indistinguishable. Recall that for the helium atom, an acceptable functional form for the lowest-energy state, is:

$$\begin{aligned}\psi &= 1s(1)1s(2)[\alpha(1)\beta(2) - \alpha(2)\beta(1)] \\ &\equiv 1s(1)1s(2)\alpha(1)\beta(2) - 1s(1)1s(2)\alpha(2)\beta(1)\end{aligned}\quad (2.56)$$

This can be written in the form of a 2×2 determinant:

$$\begin{vmatrix} 1s(1)\alpha(1) & 1s(1)\beta(1) \\ 1s(2)\alpha(2) & 1s(2)\beta(2) \end{vmatrix} \quad (2.57)$$

The two spin orbitals are

$$\chi_1 = 1s(1)\alpha(1) \quad \text{and} \quad \chi_2 = 1s(1)\beta(1) \quad (2.58)$$

A determinant is the most convenient way to write down the permitted functional forms of a polyelectronic wavefunction that satisfies the antisymmetry principle. In general, if we have N electrons in spin orbitals $\chi_1, \chi_2, \dots, \chi_N$ (where each spin orbital is the product of a spatial function and a spin function) then an acceptable form of the wavefunction is:

$$\Psi = \frac{1}{\sqrt{N!}} \begin{vmatrix} \chi_1(1) & \chi_2(1) & \cdots & \chi_N(1) \\ \chi_1(2) & \chi_2(2) & \cdots & \chi_N(2) \\ \vdots & \vdots & & \vdots \\ \chi_1(N) & \chi_2(N) & \cdots & \chi_N(N) \end{vmatrix} \quad (2.59)$$

As before, $\chi_1(1)$ is used to indicate a function that depends on the space and spin coordinates of the electron labelled '1'. The factor $1/\sqrt{N!}$ ensures that the wavefunction is normalised; we shall see later why the normalisation factor has this particular value. This functional form of the wavefunction is called a *Slater determinant* and is the simplest form of an orbital wavefunction that satisfies the antisymmetry principle. The Slater determinant is a particularly convenient and concise way to represent the wavefunction due to the special properties of determinants. Exchanging any two rows of a determinant, a process which corresponds to exchanging two electrons, changes the sign of the determinant and therefore directly leads to the antisymmetry property. If any two rows of a determinant are identical, which would correspond to two electrons being assigned to the same spin orbital, then the determinant vanishes. This can be considered a manifestation of the Pauli principle, which states that no two electrons can have the same set of quantum numbers. The Pauli principle also leads to the notion that each spatial orbital can accommodate two electrons of opposite spins.

When the Slater determinant is expanded, a total of $N!$ terms results. This is because there are $N!$ different permutations of N electrons. For example, for a three-electron system with spin orbitals χ_1 , χ_2 and χ_3 the determinant is

$$\Psi = \frac{1}{\sqrt{12}} \begin{vmatrix} \chi_1(1) & \chi_2(1) & \chi_3(1) \\ \chi_1(2) & \chi_2(2) & \chi_3(2) \\ \chi_1(3) & \chi_2(3) & \chi_3(3) \end{vmatrix} \quad (2.60)$$

Expansion of the determinant gives the following expression (ignoring the normalisation constant):

$$\begin{aligned} & \chi_1(1)\chi_2(2)\chi_3(3) - \chi_1(1)\chi_3(2)\chi_2(3) + \chi_2(1)\chi_3(2)\chi_1(3) \\ & - \chi_2(1)\chi_1(2)\chi_3(3) + \chi_3(1)\chi_1(2)\chi_2(3) - \chi_3(1)\chi_2(2)\chi_1(3) \end{aligned} \quad (2.61)$$

This expansion contains six terms ($\equiv 3!$). The six possible permutations of three electrons are: 123, 132, 213, 231, 312, 321. Some of these permutations involve single exchanges of electrons; others involve the exchange of two electrons. For example, the permutation 132 can be generated from the initial permutation by exchanging electrons 2 and 3. If we do so then the following wavefunction is obtained:

$$\begin{aligned} & \chi_1(1)\chi_2(3)\chi_3(2) - \chi_1(1)\chi_3(3)\chi_2(2) + \chi_2(1)\chi_3(3)\chi_1(2) \\ & - \chi_2(1)\chi_1(3)\chi_3(2) + \chi_3(1)\chi_1(3)\chi_2(2) - \chi_3(1)\chi_2(3)\chi_1(2) \\ & = -\chi_1(1)\chi_2(2)\chi_3(3) + \chi_1(1)\chi_3(2)\chi_2(3) - \chi_2(1)\chi_3(2)\chi_1(3) \\ & + \chi_2(1)\chi_1(2)\chi_3(3) - \chi_3(1)\chi_1(2)\chi_2(3) + \chi_3(1)\chi_2(2)\chi_1(3) \\ & = -\Psi \end{aligned} \quad (2.62)$$

By contrast, the permutation 312 requires that electrons 1 and 3 are exchanged and then electrons 1 and 2 are exchanged. This gives rise to an unchanged wavefunction. In general, an odd permutation involves an odd number of electron exchanges and leads to a wavefunction with a changed sign; an even permutation involves an even number of electron exchanges and returns the wavefunction unchanged.

For any sizeable system the Slater determinant can be tedious to write out, let alone the equivalent full orbital expansion, and so it is common to use a shorthand notation. Various notation systems have been devised. In one system the terms along the diagonal of the matrix are written as a single-row determinant. For the 3×3 determinant we therefore have:

$$\begin{vmatrix} \chi_1(1) & \chi_2(1) & \chi_3(1) \\ \chi_1(2) & \chi_2(2) & \chi_3(2) \\ \chi_1(3) & \chi_2(3) & \chi_3(3) \end{vmatrix} \equiv |\chi_1 \ \chi_2 \ \chi_3| \quad (2.63)$$

The normalisation factor is assumed. It is often convenient to indicate the spin of each electron in the determinant; this is done by writing a bar when the spin part is β (spin down); a function without a bar indicates an α spin (spin up). Thus, the following are all commonly used ways to write the Slater determinantal wavefunction for the beryllium atom (which has the electronic configuration $1s^2 2s^2$):

$$\begin{aligned} \Psi &= \frac{1}{\sqrt{24}} \begin{vmatrix} \phi_{1s}(1) & \bar{\phi}_{1s}(1) & \phi_{2s}(1) & \bar{\phi}_{2s}(1) \\ \phi_{1s}(2) & \bar{\phi}_{1s}(2) & \phi_{2s}(2) & \bar{\phi}_{2s}(2) \\ \phi_{1s}(3) & \bar{\phi}_{1s}(3) & \phi_{2s}(3) & \bar{\phi}_{2s}(3) \\ \phi_{1s}(4) & \bar{\phi}_{1s}(4) & \phi_{2s}(4) & \bar{\phi}_{2s}(4) \end{vmatrix} \\ &\equiv |\phi_{1s} \ \bar{\phi}_{1s} \ \phi_{2s} \ \bar{\phi}_{2s}| \\ &\equiv |1s \ \bar{1}s \ 2s \ \bar{2}s| \end{aligned} \quad (2.64)$$

An important property of determinants is that a multiple of any column can be added to another column without altering the value of the determinant. This means that the spin orbitals are not unique; other linear combinations give the same energy. To illustrate this, consider the first excited state configuration of the helium atom ($1s^2 2s^2$), which can be written as the following 2×2 determinant:

$$\begin{vmatrix} 1s(1)\alpha(1) & 2s(1)\alpha(1) \\ 1s(2)\alpha(2) & 2s(2)\alpha(2) \end{vmatrix} = 1s(1)\alpha(1)2s(2)\alpha(2) - 1s(2)\alpha(2)2s(1)\alpha(1) \quad (2.65)$$

We now introduce two new 'spin orbitals':

$$\chi'_1 = \frac{1s + 2s}{\sqrt{2}}\alpha; \quad \chi'_2 = \frac{1s - 2s}{\sqrt{2}}\alpha \quad (2.66)$$

With these new orbitals the value of the determinant is as follows:

$$\begin{aligned} \begin{vmatrix} \chi'_1(1) & \chi'_2(1) \\ \chi'_1(2) & \chi'_2(2) \end{vmatrix} &= \frac{[1s(1) + 2s(1)][1s(2) - 2s(2)]\alpha(1)\alpha(2)}{2} \\ &\quad - \frac{[1s(1) - 2s(1)][1s(2) + 2s(2)]\alpha(1)\alpha(2)}{2} \\ &\equiv -\Psi \end{aligned} \quad (2.67)$$

This can be helpful because it may enable more meaningful sets of orbitals to be generated from the original solutions. Molecular orbital calculations may give solutions that are 'smeared out' throughout the entire molecule, whereas we may find orbitals that are localised in specific regions (e.g. in the bonds between atoms) to be more useful.

2.4 Molecular Orbital Calculations

2.4.1 Calculating the Energy from the Wavefunction: the Hydrogen Molecule

In our treatment of molecular systems we first show how to determine the energy for a given wavefunction, and then demonstrate how to calculate the wavefunction for a specific nuclear geometry. In the most popular kind of quantum mechanical calculations performed on molecules each molecular spin orbital is expressed as a linear combination of atomic orbitals (the LCAO approach*). Thus each molecular orbital can be written as a summation of the following form:

$$\psi_i = \sum_{\mu=1}^K c_{\mu i} \phi_{\mu} \quad (2.68)$$

ψ_i is a (spatial) molecular orbital, ϕ_{μ} is one of K atomic orbitals and $c_{\mu i}$ is a coefficient. In a simple LCAO picture of the lowest energy state of molecular hydrogen, H_2 , there are two electrons with opposite spins in the lowest energy spatial orbital (labelled $1\sigma_g$), which is

* Computational quantum chemistry is well endowed with acronyms and abbreviations. A list of some of the more common ones can be found in Appendix 2.1

formed from a linear combination of two hydrogen-atom 1s orbitals:

$$1\sigma_g = A(1s_A + 1s_B) \quad (2.69)$$

A is the normalisation factor, whose value is not important in our present discussion. To calculate the energy of the ground state of the hydrogen molecule for a fixed internuclear distance we first write the wavefunction as a 2×2 determinant:

$$\Psi = \begin{vmatrix} \chi_1(1) & \chi_2(1) \\ \chi_1(2) & \chi_2(2) \end{vmatrix} = \chi_1(1)\chi_2(2) - \chi_1(2)\chi_2(1) \quad (2.70)$$

where

$$\begin{aligned} \chi_1(1) &= 1\sigma_g(1)\alpha(1) \\ \chi_2(1) &= 1\sigma_g(1)\beta(1) \\ \chi_1(2) &= 1\sigma_g(2)\alpha(2) \\ \chi_2(2) &= 1\sigma_g(2)\beta(2) \end{aligned} \quad (2.71)$$

For the hydrogen molecule, the Hamiltonian comprises the kinetic energy operator for each electron plus the potential energy operator due to the Coulomb attraction between the two electrons and the two nuclei, and the repulsion between the two electrons. In atomic units the Hamiltonian is thus

$$\mathcal{H} = -\frac{1}{2}\nabla_1^2 - \frac{1}{2}\nabla_2^2 - \frac{Z_A}{r_{1A}} - \frac{Z_B}{r_{1B}} - \frac{Z_A}{r_{2A}} - \frac{Z_B}{r_{2B}} + \frac{1}{r_{12}} \quad (2.72)$$

The electrons have been labelled 1 and 2 and the nuclei have been labelled A and B. For H_2 the nuclear charges Z_A and Z_B are both equal to 1. First we need to consider how to calculate the energy of this hydrogen molecule. This is obtained using Equation (2.7):

$$E = \frac{\int \Psi \mathcal{H} \Psi d\tau}{\int \Psi \Psi d\tau} \quad (2.73)$$

In general, a quantum mechanical calculation provides molecular orbitals that are normalised but the total wavefunction is not. The normalisation constant for the wavefunction of the two-electron hydrogen molecule is $1/\sqrt{2}$ and so the denominator in Equation (2.73) is equal to 2.

We now substitute the hydrogen molecule wavefunction into Equation (2.73) to provide the following:

$$\begin{aligned} E = \frac{1}{2} \iint d\tau_1 d\tau_2 &\{ [\chi_1(1)\chi_2(2) - \chi_2(1)\chi_1(2)] [-\frac{1}{2}\nabla_1^2 - \frac{1}{2}\nabla_2^2 - (1/r_{1A}) - (1/r_{1B}) \\ &- (1/r_{2A}) - (1/r_{2B}) + (1/r_{12})] [\chi_1(1)\chi_2(2) - \chi_2(1)\chi_1(2)] \} \end{aligned} \quad (2.74)$$

$d\tau_i$ indicates that the integration is over the spatial and spin coordinates of electron i . It is useful to separate the Hamiltonian operator into two H_2^+ Hamiltonians plus the inter-electronic repulsion term:

$$\begin{aligned} E = \frac{1}{2} \iint d\tau_1 d\tau_2 &\{ [\chi_1(1)\chi_2(2) - \chi_2(1)\chi_1(2)] [\mathcal{H}_1 + \mathcal{H}_2 + (1/r_{12})] \\ &\times [\chi_1(1)\chi_2(2) - \chi_2(1)\chi_1(2)] \} \end{aligned} \quad (2.75)$$

where

$$\mathcal{H}_1 = -\frac{1}{2}\nabla_1^2 - \frac{1}{r_{1A}} - \frac{1}{r_{1B}} \quad \text{and} \quad \mathcal{H}_2 = -\frac{1}{2}\nabla_2^2 - \frac{1}{r_{2A}} - \frac{1}{r_{2B}} \quad (2.76)$$

We can now start to separate the integral in Equation (2.74) into individual terms and identify the various contributions to the electronic energy:

$$\begin{aligned} E = & \int \int d\tau_1 d\tau_2 \chi_1(1)\chi_2(2)(\mathcal{H}_1)\chi_1(1)\chi_2(2) \\ & - \int \int d\tau_1 d\tau_2 \chi_1(1)\chi_2(2)(\mathcal{H}_1)\chi_2(1)\chi_1(2) + \dots \\ & + \int \int d\tau_1 d\tau_2 \chi_1(1)\chi_2(2)(\mathcal{H}_2)\chi_1(1)\chi_2(2) \\ & - \int \int d\tau_1 d\tau_2 \chi_1(1)\chi_2(2)(\mathcal{H}_2)\chi_2(1)\chi_1(2) + \dots \\ & + \int \int d\tau_1 d\tau_2 \chi_1(1)\chi_2(2) \left(\frac{1}{r_{12}} \right) \chi_1(1)\chi_2(2) \\ & - \int \int d\tau_1 d\tau_2 \chi_1(1)\chi_2(2) \left(\frac{1}{r_{12}} \right) \chi_2(1)\chi_1(2) + \dots \end{aligned} \quad (2.77)$$

Each of these individual terms can be simplified if we recognise that terms dependent upon electrons other than those in the operator can be separated out. For example, the first term in the expansion, Equation (2.77), is:

$$\int \int d\tau_1 d\tau_2 \chi_1(1)\chi_2(2)(\mathcal{H}_1)\chi_1(1)\chi_2(2) \quad (2.78)$$

The operator \mathcal{H}_1 is a function of the coordinates of electron 1 only, so terms involving electron 2 can be separated out as follows:

$$\begin{aligned} & \int \int d\tau_1 d\tau_2 \chi_1(1)\chi_2(2)(\mathcal{H}_1)\chi_1(1)\chi_2(2) \\ &= \int d\tau_2 \chi_2(2)\chi_2(2) \int d\tau_1 \chi_1(1) \left(-\frac{1}{2}\nabla_1^2 - \frac{1}{r_{1A}} - \frac{1}{r_{1B}} \right) \chi_1(1) \end{aligned} \quad (2.79)$$

If the molecular orbitals are normalised, the integral $\int d\tau_2 \chi_2(2)\chi_2(2)$ equals 1. Further simplification can be achieved by splitting the integral involving electron 1 into separate integrals over the spatial and spin parts; the integral over spin orbitals is equal to the product of an integral over the spatial coordinates and an integral over the spin coordinates:

$$\begin{aligned} & \int d\tau_1 \chi_1(1) \left(-\nabla_1^2 - \frac{1}{r_{1A}} - \frac{1}{r_{1B}} \right) \chi_1(1) \\ &= \int d\nu_1 1\sigma_g(1) \left(-\frac{1}{2}\nabla_1^2 - \frac{1}{r_{1A}} - \frac{1}{r_{1B}} \right) 1\sigma_g(1) \int d\sigma_1 \alpha(1)\alpha(1) \end{aligned} \quad (2.80)$$

$d\nu$ indicates integration over spatial coordinates and $d\sigma$ indicates integration over the spin coordinates. The integral over the spin coordinates equals 1. This expression corresponds

to the sum of the kinetic and potential energy of an electron in the orbital $1\sigma_g$ in the electrostatic field of the two bare nuclei. This integral can in turn be expanded by substituting the atomic orbital combination for $1\sigma_g$:

$$\begin{aligned} & \int d\nu_1 1\sigma_g(1) \left(-\frac{1}{2} \nabla_1^2 - \frac{1}{r_{1A}} - \frac{1}{r_{1B}} \right) 1\sigma_g(1) \\ &= A^2 \int d\nu_1 \{1s_A(1) + 1s_B(1)\} \left(-\frac{1}{2} \nabla_1^2 - \frac{1}{r_{1A}} - \frac{1}{r_{1B}} \right) \{1s_A(1) + 1s_B(1)\} \end{aligned} \quad (2.81)$$

A is the normalisation constant. The integral in Equation (2.81) can in turn be factorised to give a sum of integrals, each of which involves a pair of atomic orbitals:

$$\begin{aligned} & \int d\nu_1 \{1s_A(1) + 1s_B(1)\} \left(-\frac{1}{2} \nabla_1^2 - \frac{1}{r_{1A}} - \frac{1}{r_{1B}} \right) \{1s_A(1) + 1s_B(1)\} \\ &= \int d\nu_1 1s_A(1) \left(-\frac{1}{2} \nabla_1^2 - \frac{1}{r_{1A}} - \frac{1}{r_{1B}} \right) 1s_A(1) \\ &+ \int d\nu_1 1s_A(1) \left(-\frac{1}{2} \nabla_1^2 - \frac{1}{r_{1A}} - \frac{1}{r_{1B}} \right) 1s_B(1) + \dots \end{aligned} \quad (2.82)$$

Let us now apply the same procedure to the second term in Equation (2.77):

$$\iint d\tau_1 d\tau_2 \chi_1(1) \chi_2(2) (\mathcal{H}_1) \chi_2(1) \chi_1(2) = \int d\tau_1 \chi_1(1) (\mathcal{H}_1) \chi_2(1) \int d\tau_2 \chi_2(2) \chi_1(2) \quad (2.83)$$

This particular integral is zero because the molecular orbitals are orthogonal and so the integral over the coordinates of electron 2 equals zero:

$$\int d\tau_2 \chi_2(2) \chi_1(2) = 0 \quad (2.84)$$

A similar procedure can be applied to the other integrals involving electron–nuclear interactions; it turns out that there are four non-zero integrals, each of which is equal to the energy of a single electron in the field of the two hydrogen nuclei.

There remain four integrals arising from electron–electron interactions. These are:

$$\begin{aligned} & \iint d\tau_1 d\tau_2 \chi_1(1) \chi_2(2) \left(\frac{1}{r_{12}} \right) \chi_1(1) \chi_2(2) + \iint d\tau_1 d\tau_2 \chi_2(1) \chi_1(2) \left(\frac{1}{r_{12}} \right) \chi_2(1) \chi_1(2) \\ & - \iint d\tau_1 d\tau_2 \chi_1(1) \chi_2(2) \left(\frac{1}{r_{12}} \right) \chi_2(1) \chi_1(2) - \iint d\tau_1 d\tau_2 \chi_2(1) \chi_1(2) \left(\frac{1}{r_{12}} \right) \chi_1(1) \chi_2(2) \end{aligned} \quad (2.85)$$

The first two of these can be simplified as follows:

$$\begin{aligned} & \iint d\tau_1 d\tau_2 \chi_1(1) \chi_2(2) \left(\frac{1}{r_{12}} \right) \chi_1(1) \chi_2(2) = \iint d\nu_1 d\nu_2 1\sigma_g(1) 1\sigma_g(2) \left(\frac{1}{r_{12}} \right) 1\sigma_g(1) 1\sigma_g(2) \\ & \quad \times \int d\sigma_1 \alpha(1) \alpha(1) \int d\sigma_2 \beta(2) \beta(2) \\ & = \iint d\nu_1 d\nu_2 1\sigma_g(1) 1\sigma_g(1) \left(\frac{1}{r_{12}} \right) 1\sigma_g(2) 1\sigma_g(2) \end{aligned} \quad (2.86)$$

According to the Born interpretation of the wavefunction, $1\sigma_g(r_1)1\sigma_g(r_1)$ equals the electron density of electron 1 in orbital $1\sigma_g$ at a position \mathbf{r}_1 . Similarly, $1\sigma_g(r_2)1\sigma_g(r_2)$ is the electron density of electron 2. The electrostatic repulsion between these regions of electron density thus equals $1\sigma_g(r_1)1\sigma_g(r_1) \times (1/r_{12}) \times 1\sigma_g(r_2)1\sigma_g(r_2)$, where r_{12} is the distance between the two electrons. The integral of this function over all space thus corresponds to the electrostatic (Coulomb) repulsion between the two orbitals.

If we substitute the atomic orbital expansion, we obtain a series of two-electron integrals, each of which involves four atomic orbitals:

$$\begin{aligned} & \int \int d\nu_1 d\nu_2 1\sigma_g(1)1\sigma_g(2) \left(\frac{1}{r_{12}} \right) 1\sigma_g(1)1\sigma_g(2) \\ &= \int \int d\nu_1 d\nu_2 1s_A(1)1s_A(2) \left(\frac{1}{r_{12}} \right) 1s_A(1)1s_A(2) \\ &+ \int \int d\nu_1 d\nu_2 1s_A(1)1s_A(2) \left(\frac{1}{r_{12}} \right) 1s_A(1)1s_B(2) + \dots \end{aligned} \quad (2.87)$$

The remaining two integrals from Equation (2.85) are:

$$\begin{aligned} \int \int d\tau_1 d\tau_2 \chi_1(1)\chi_2(2) \left(\frac{1}{r_{12}} \right) \chi_2(1)\chi_1(2) &= \int \int d\nu_1 d\nu_2 1\sigma_g(1)1\sigma_g(2) \left(\frac{1}{r_{12}} \right) 1\sigma_g(1)1\sigma_g(2) \\ &\times \int d\sigma_1 \alpha(1)\beta(1) \int d\sigma_2 \beta(2)\alpha(2) \end{aligned} \quad (2.88)$$

$$\begin{aligned} \int \int d\tau_1 d\tau_2 \chi_2(1)\chi_1(2) \left(\frac{1}{r_{12}} \right) \chi_1(1)\chi_2(2) &= \int \int d\nu_1 d\nu_2 1\sigma_g(1)1\sigma_g(2) \left(\frac{1}{r_{12}} \right) 1\sigma_g(1)1\sigma_g(2) \\ &\times \int d\sigma_1 \beta(1)\alpha(1) \int d\sigma_2 \alpha(2)\beta(2) \end{aligned} \quad (2.89)$$

Both of these integrals are zero due to the orthogonality of the electron spin states α and β .

The triplet excited state of H_2 is obtained by promoting an electron to a higher-energy molecular orbital. This higher-energy (antibonding) orbital is written $1\sigma_u$ and can be considered to arise from two 1s orbitals as follows:

$$1\sigma_u = A(1s_A - 1s_B) \quad (2.90)$$

The triplet state has two unpaired electrons with the same spin (α) and so the wavefunction state is:

$$\begin{vmatrix} 1\sigma_g\alpha(1) & 1\sigma_u\alpha(1) \\ 1\sigma_g\alpha(2) & 1\sigma_u\alpha(2) \end{vmatrix} \quad (2.91)$$

If we now expand the expression for the energy as for the ground state, terms analogous to the electron–nucleus and electron–electron interactions can again be obtained. However, the cross-terms are no longer equal to zero as was the case for the ground state, because the

electron spins are now the same (both α). For example, compare with Equation (2.88):

$$\begin{aligned} \iint d\tau_1 d\tau_2 \chi_1(1) \chi_2(2) \left(\frac{1}{r_{12}} \right) \chi_2(1) \chi_1(2) &= \iint d\nu_1 d\nu_2 1\sigma_g(1) 1\sigma_u(2) \left(\frac{1}{r_{12}} \right) 1\sigma_g(2) 1\sigma_u(1) \\ &\quad \times \int d\sigma_1 \alpha(1) \alpha(1) \int d\sigma_2 \alpha(2) \alpha(2) \end{aligned} \quad (2.92)$$

This contribution is called the *exchange interaction*. This appears with a minus sign in the expression for the total energy and so acts to stabilise the triplet $1s^1 2s^1$ state over the analogous singlet state. The exchange term is only non-zero for electrons of the same spin. It has the effect of making electrons of the same spin 'avoid' each other. As a result of this each electron can be considered to have a 'hole' associated with it. This hole is known as the *exchange hole* or the *Fermi hole*.

2.4.2 The Energy of a General Polyelectronic System

The hydrogen molecule is such a small problem that all of the integrals can be written out in full. This is rarely the case in molecular orbital calculations. Nevertheless, the same principles are used to determine the energy of a polyelectronic molecular system. For an N -electron system, the Hamiltonian takes the following general form:

$$\mathcal{H} = \left(-\frac{1}{2} \sum_{i=1}^N \nabla_i^2 - \frac{1}{r_{1A}} - \frac{1}{r_{1B}} \dots + \frac{1}{r_{12}} + \frac{1}{r_{13}} + \dots \right) \quad (2.93)$$

As with the hydrogen molecule, we have adopted the convention that the nuclei are labelled using capital letters A, B, C, etc., and the electrons are labelled 1, 2, 3,

Recall that the Slater determinant for a system of N electrons in N spin orbitals can be written:

$$\begin{vmatrix} \chi_1(1) & \chi_2(1) & \chi_3(1) & \dots & \chi_N(1) \\ \chi_1(2) & \chi_2(2) & \chi_3(2) & \dots & \chi_N(2) \\ \chi_1(3) & \chi_2(3) & \chi_3(3) & \dots & \chi_N(3) \\ \vdots & \vdots & \vdots & & \vdots \\ \chi_1(N) & \chi_2(N) & \chi_3(N) & \dots & \chi_N(N) \end{vmatrix} \quad (2.94)$$

Each term in the determinant can thus be written $\chi_i(1)\chi_j(2)\chi_k(3)\dots\chi_u(N-1)\chi_v(N)$ where i, j, k, \dots, u, v is a series of N integers.

As usual, the energy can be calculated from $E = \int \Psi \mathcal{H} \Psi / \int \Psi \Psi$:

$$\begin{aligned} \int \Psi \mathcal{H} \Psi &= \int \dots \int d\tau_1 d\tau_2 \dots d\tau_N \left\{ [\chi_i(1)\chi_j(2)\chi_k(3)\dots] \right. \\ &\quad \times \left(-\frac{1}{2} \sum_i \nabla_i^2 - (1/r_{1A}) - (1/r_{1B}) \dots + (1/r_{12}) + (1/r_{13}) + \dots \right) \\ &\quad \left. \times [\chi_i(1)\chi_j(2)\chi_k(3)\dots] \right\} \end{aligned} \quad (2.95)$$

$$\int \Psi \Psi = \int \cdots \int d\tau_1 d\tau_2 \cdots d\tau_N \{ [\chi_i(1)\chi_j(2)\chi_k(3) \cdots] [\chi_i(1)\chi_j(2)\chi_k(3) \cdots] \} \quad (2.96)$$

We can now see why the normalisation factor of the Slater determinantal wavefunction is $1/\sqrt{N!}$. If each determinant contains $N!$ terms then the product of two Slater determinants, [determinant][determinant], contains $(N!)^2$ terms. However, if the spin orbitals form an orthonormal set then only products of identical terms from the determinant will be non-zero when integrated over all space. We can illustrate this with the three-electron example. Considering just the first two terms in the expansion we obtain the following:

$$\begin{aligned} & \int \int \int d\tau_1 d\tau_2 d\tau_3 [\chi_1(1)\chi_2(2)\chi_3(3) - \chi_1(1)\chi_3(2)\chi_2(3) + \cdots] \\ & \times [\chi_1(1)\chi_2(2)\chi_3(3) - \chi_1(1)\chi_3(2)\chi_2(3) + \cdots] \end{aligned} \quad (2.97)$$

When multiplied out this gives:

$$\begin{aligned} & \int \int \int d\tau_1 d\tau_2 d\tau_3 [\chi_1(1)\chi_2(2)\chi_3(3)][\chi_1(1)\chi_2(2)\chi_3(3)] \\ & - \int \int \int d\tau_1 d\tau_2 d\tau_3 [\chi_1(1)\chi_2(2)\chi_3(3)][\chi_1(1)\chi_3(2)\chi_2(3)] + \cdots \\ & + \int \int \int d\tau_1 d\tau_2 d\tau_3 [\chi_1(1)\chi_3(2)\chi_2(3)][\chi_1(1)\chi_3(2)\chi_2(3)] + \cdots \end{aligned} \quad (2.98)$$

The first of the integrals in Equation (2.98) equals 1 (if the spin orbitals are normalised). The second term is zero because the terms involving both electrons 2 and 3 are different (for example, the integral $\int d\tau_2 \chi_2(2)\chi_3(2)$ will be zero due to the orthogonality of the spin orbitals χ_2 and χ_3). The third term in Equation (2.98) will be equal to 1, and so on. It turns out that there are $N!$ such non-zero terms. Thus if each individual term in the determinant is normalised, then:

$$\int \Psi \Psi = N! \quad (2.99)$$

Hence the normalisation factor for the determinantal wavefunction is $1/\sqrt{N!}$.

Turning now to the numerator in the energy expression (Equation (2.95)), this can be broken down into a series of one-electron and two-electron integrals, as for the hydrogen molecule. Each of these individual integrals has the general form:

$$\int \cdots \int d\tau_1 d\tau_2 \cdots [term1]operator[term2] \quad (2.100)$$

[term1] and [term2] each represent one of the $N!$ terms in the Slater determinant. To simplify this integral, we first recognise that all spin orbitals involving an electron that does not appear in the operator can be taken outside the integral. For example, if the operator is $1/r_{1A}$, then all spin orbitals other than those that depend on the coordinates of electron 1 can be separated from the integral. The orthogonality of the spin orbitals means that the integral will be zero unless all indices involving these other electrons are the same in

[term1] and [term2]. Again, to use our three-electron system as an example:

$$\begin{aligned} & \iiint d\tau_1 d\tau_2 d\tau_3 [\chi_1(1)\chi_2(2)\chi_3(3)] \left(-\frac{1}{r_{1A}}\right) [\chi_1(1)\chi_2(2)\chi_3(3)] \\ &= \iint d\tau_2 d\tau_3 [\chi_2(2)\chi_3(3)][\chi_2(2)\chi_3(3)] \int d\tau_1 \chi_1(1) \left(-\frac{1}{r_{1A}}\right) \chi_1(1) \\ &= \int d\tau_1 \chi_1(1) \left(-\frac{1}{r_{1A}}\right) \chi_1(1) \end{aligned} \quad (2.101)$$

But:

$$\begin{aligned} & \iiint d\tau_1 d\tau_2 d\tau_3 [\chi_1(1)\chi_2(2)\chi_3(3)] \left(-\frac{1}{r_{1A}}\right) [\chi_1(1)\chi_3(2)\chi_2(3)] \\ &= \iint d\tau_2 d\tau_3 [\chi_2(2)\chi_3(3)][\chi_3(2)\chi_2(3)] \int d\tau_1 \chi_1(1) \left(-\frac{1}{r_{1A}}\right) \chi_1(1) \\ &= 0 \end{aligned} \quad (2.102)$$

For integrals that involve two-electron operators (i.e. $1/r_{ij}$), only those terms that do not involve the coordinates of the two electrons can be taken outside the integral. For example:

$$\begin{aligned} & \iiint d\tau_1 d\tau_2 d\tau_3 [\chi_1(1)\chi_2(2)\chi_3(3)] \left(\frac{1}{r_{12}}\right) [\chi_1(1)\chi_2(2)\chi_3(3)] \\ &= \iint d\tau_1 d\tau_2 [\chi_1(1)\chi_2(2)] \left(\frac{1}{r_{12}}\right) [\chi_1(2)\chi_2(2)] \int d\tau_3 \chi_3(3) \chi_3(3) \\ &= \iint d\tau_1 d\tau_2 [\chi_1(1)\chi_2(2)] \left(\frac{1}{r_{12}}\right) [\chi_1(2)\chi_2(2)] \end{aligned} \quad (2.103)$$

But:

$$\begin{aligned} & \iiint d\tau_1 d\tau_2 d\tau_3 [\chi_1(1)\chi_2(2)\chi_3(3)] \left(\frac{1}{r_{12}}\right) [\chi_1(1)\chi_3(2)\chi_2(3)] \\ &= \iint d\tau_1 d\tau_2 [\chi_1(1)\chi_2(2)] \left(\frac{1}{r_{12}}\right) [\chi_1(2)\chi_3(2)] \int d\tau_3 \chi_3(3) \chi_2(3) \\ &= 0 \end{aligned} \quad (2.104)$$

As a consequence of these results, most of the individual integrals in the expansion will be zero. Nevertheless, it can be readily envisaged that there will still be an extremely large number of integrals to consider for all except the smallest problems. It is thus more convenient to write the energy expression in a concise form that recognises the three types of interaction that contribute to the total electronic energy of the system.

First, there is the kinetic and potential energy of each electron moving in the field of the nuclei. The energy associated with this contribution for the molecular orbital χ_i is often written H_{ii}^{core} and for M nuclei is given by:

$$H_{ii}^{\text{core}} = \int d\tau_1 \chi_i(1) \left(-\frac{1}{2} \nabla_i^2 - \sum_{A=1}^M \frac{Z_A}{r_{iA}}\right) \chi_i(1) \quad (2.105)$$

For N electrons in N molecular orbitals this contribution to the total energy is:

$$E_{\text{total}}^{\text{core}} = \sum_{i=1}^N \int d\tau_1 \chi_i(1) \left(-\frac{1}{2} \nabla_i^2 - \sum_{A=1}^M \frac{Z_A}{r_{iA}} \right) \chi_i(1) = \sum_{i=1}^N H_{ii}^{\text{core}} \quad (2.106)$$

Here we have followed convention and have used the label '1' wherever there is an integral involving the coordinates of a single electron, even though the actual electron may not be 'electron 1'. Similarly, when it is necessary to consider two electrons then the labels 1 and 2 are conventionally employed. H_{ii}^{core} makes a favourable (i.e. negative) contribution to the electronic energy.

The second contribution to the energy arises from the electrostatic repulsion between pairs of electrons. This interaction depends on the electron-electron distance and, as we have seen, is calculated from integrals such as:

$$J_{ij} = \iint d\tau_1 d\tau_2 \chi_i(1) \chi_j(2) \left(\frac{1}{r_{12}} \right) \chi_i(1) \chi_j(2) \quad (2.107)$$

The symbol J_{ij} is often used to represent this Coulomb interaction between electrons in spin orbitals i and j , and is unfavourable (i.e. positive). The total electrostatic interaction between the electron in orbital χ_i and the other $N - 1$ electrons is a sum of all such integrals, where the summation index j runs from 1 to N , excluding i :

$$\begin{aligned} E_i^{\text{Coulomb}} &= \sum_{j \neq i}^N \int d\tau_1 d\tau_2 \chi_i(1) \chi_j(2) \frac{1}{r_{12}} \chi_j(2) \chi_i(1) \\ &\equiv \sum_{j \neq i}^N \int d\tau_1 d\tau_2 \chi_i(1) \chi_i(1) \frac{1}{r_{12}} \chi_j(2) \chi_j(2) \end{aligned} \quad (2.108)$$

The total Coulomb contribution to the electronic energy of the system is obtained as a double summation over all electrons, taking care to count each interaction just once:

$$E_{\text{total}}^{\text{Coulomb}} = \sum_{i=1}^N \sum_{j=i+1}^N \int d\tau_1 d\tau_2 \chi_i(1) \chi_i(1) \frac{1}{r_{12}} \chi_j(2) \chi_j(2) = \sum_{i=1}^N \sum_{j=i+1}^N J_{ij} \quad (2.109)$$

The third contribution to the energy is the exchange 'interaction'. This has no classical counterpart and arises because the motions of electrons with parallel spins are correlated: whereas there is a finite probability of finding two electrons with opposite (i.e. paired) spins at the same point in space, where the spins are the same then the probability is zero. This can be considered a manifestation of the Pauli principle, for if two electrons occupied the same region of space and had parallel spins then they could be considered to have the same set of quantum numbers. Electrons with the same spin thus tend to 'avoid' each other, and they experience a lower Coulombic repulsion, giving a lower (i.e. more favourable) energy. The exchange interaction involves integrals of the form:

$$K_{ij} = \iint d\tau_1 d\tau_2 \chi_i(1) \chi_j(2) \left(\frac{1}{r_{12}} \right) \chi_i(2) \chi_j(1) \quad (2.110)$$

This integral is only non-zero if the spins of the electrons in the spin orbitals χ_i and χ_j are the same. The energy due to exchange is often represented as K_{ij} . The exchange energy between the electron in spin orbital χ_i and the other $N - 1$ electrons is:

$$E_i^{\text{exchange}} = \sum_{j \neq i}^N \int \int d\tau_1 d\tau_2 \chi_i(1) \chi_j(2) \left(\frac{1}{r_{12}} \right) \chi_i(2) \chi_j(1) \quad (2.111)$$

The total exchange energy is calculated thus:

$$E_{\text{total}}^{\text{exchange}} = \sum_{i=1}^N \sum_{j'=i+1}^N \int \int d\tau_1 d\tau_2 \chi_i(1) \chi_j(2) \left(\frac{1}{r_{12}} \right) \chi_i(2) \chi_j(1) = \sum_{j=1}^N \sum_{j'=i+1}^N K_{ij} \quad (2.112)$$

The prime on the counter j' indicates that the summation is only over electrons with the same spin as electron i .

2.4.3 Shorthand Representations of the One- and Two-electron Integrals

Various shorthand ways have been devised to represent the integrals involved in an electronic structure calculation. The two-electron integrals J_{ij} and K_{ij} are particularly long-winded to write out. In one scheme the Coulomb interaction J_{ij} is written as:

$$\left\langle \chi_i^* \chi_j^* \left| \frac{1}{r_{12}} \right| \chi_i \chi_j \right\rangle \quad (2.113)$$

In this notation the complex parts are written on the left-hand side and the real parts on the right. Sometimes the χ symbol is eliminated:

$$\left\langle ij \left| \frac{1}{r_{12}} \right| ij \right\rangle \quad (2.114)$$

The exchange integrals would be written:

$$\left\langle ij \left| \frac{1}{r_{12}} \right| ji \right\rangle \quad (2.115)$$

in this notation.

A notation that is widely used in the chemical literature writes the orbitals that are functions of electron 1 on the left-hand side (with the complex conjugate orbital first, if appropriate) and the orbitals that are functions of electron 2 on the right-hand side (again with the complex conjugate orbital first). In this notation, which is the one that we will adopt, the Coulomb integral is written $(ii|jj)$ and the exchange integral $(ij|ji)$. The one-electron integrals such as Equation (2.105) are written as follows:

$$\left(i \left| -\frac{1}{2} \nabla_i^2 - \sum_{A=1}^M \frac{Z_A}{r_{iA}} \right| j \right) \equiv \int d\tau_1 \chi_i(1) \left(-\frac{1}{2} \nabla_i^2 - \sum_{A=1}^M \frac{Z_A}{r_{iA}} \right) \chi_j(1) \quad (2.116)$$

When calculating the total energy of the system, we should not forget the Coulomb interaction between the nuclei; this is constant within the Born–Oppenheimer approximation for a given spatial arrangement of nuclei. When it is desired to change the nuclear positions,

it is of course necessary to take the internuclear repulsion energy into account, which is calculated using the Coulomb equation:

$$\sum_{A=1}^M \sum_{B=A+1}^M \frac{Z_A Z_B}{R_{AB}} \quad (2.117)$$

2.4.4 The Energy of a Closed-shell System

In molecular modelling we are usually concerned with the ground states of molecules, most of which have closed-shell configurations. In a closed-shell system containing N electrons in $N/2$ orbitals, there are two spin orbitals associated with each spatial orbital ψ_i : $\psi_i\alpha$ and $\psi_i\beta$. The electronic energy of such a system can be calculated in a manner analogous to that for the hydrogen molecule. First, there is the energy of each electron moving in the field of the bare nuclei. For an electron in a molecular orbital χ_i , this contributes an energy H_{ii}^{core} . If there are two electrons in the orbital then the energy is $2H_{ii}^{\text{core}}$ and for $N/2$ orbitals the total contribution to the energy will be:

$$\sum_{i=1}^{N/2} 2H_{ii}^{\text{core}} \quad (2.118)$$

If we consider the electron-electron terms, the interaction between each pair of orbitals ψ_i and ψ_j involves a total of four electrons. There are four ways in which two electrons in one orbital can interact in a Coulomb sense with two electrons in a second orbital, thus giving $4J_{ij}$. However, there are just two ways to obtain paired electrons from this arrangement, giving a total exchange contribution of $-2K_{ij}$. Finally, the Coulomb interaction between each pair of electrons in the same orbital must be included; there is no exchange interaction because the electrons have paired spins. The total energy is thus given as:

$$E = 2 \sum_{i=1}^{N/2} H_{ii}^{\text{core}} + \sum_{i=1}^{N/2} \sum_{j=i+1}^{N/2} (4J_{ij} - 2K_{ij}) + \sum_{i=1}^{N/2} J_{ii} \quad (2.119)$$

A more concise form of this equation can be obtained if we recognise that $J_{ii} = K_{ii}$:

$$E = 2 \sum_{i=1}^{N/2} H_{ii}^{\text{core}} + \sum_{i=1}^{N/2} \sum_{j=1}^{N/2} (2J_{ij} - K_{ij}) \quad (2.120)$$

2.5 The Hartree–Fock Equations

In our hydrogen molecule calculation in Section 2.4.1 the molecular orbitals were provided as input, but in most electronic structure calculations we are usually trying to calculate the molecular orbitals. How do we go about this? We must remember that for many-body problems there is no ‘correct’ solution; we therefore require some means to decide whether one proposed wavefunction is ‘better’ than another. Fortunately, the *variation theorem* provides us with a mechanism for answering this question. The theorem states that the

energy calculated from an approximation to the true wavefunction will always be greater than the true energy. Consequently, the better the wavefunction, the lower the energy. The ‘best’ wavefunction is obtained when the energy is a minimum. At a minimum, the first derivative of the energy, δE , will be zero. The Hartree–Fock equations are obtained by imposing this condition on the expression for the energy, subject to the constraint that the molecular orbitals remain orthonormal. The orthonormality condition is written in terms of the *overlap integral*, S_{ij} , between two orbitals i and j . Thus

$$S_{ij} = \int \chi_i \chi_j d\tau = \delta_{ij} \quad (\delta_{ij} \text{ is the Kronecker delta}) \quad (2.121)$$

This type of constrained minimisation problem can be tackled using the method of Lagrange multipliers. In this approach (see Section 1.10.5 for a brief introduction to Lagrange multipliers) the derivative of the function to be minimised is added to the derivatives of the constraint(s) multiplied by a constant called a Lagrange multiplier. The sum is then set equal to zero. If the Lagrange multiplier for each of the orthonormality conditions is written λ_{ij} , then:

$$\delta E + \delta \sum_i \sum_j \lambda_{ij} S_{ij} = 0 \quad (2.122)$$

In the Hartree–Fock equations the Lagrange multipliers are actually written $-2\varepsilon_{ij}$ to reflect the fact that they are related to the molecular orbital energies. The equation to be solved is thus:

$$\delta E - 2\delta \sum_i \sum_j \varepsilon_{ij} S_{ij} = 0 \quad (2.123)$$

We will not describe in detail how this equation is solved, as it is rather complicated. However, a qualitative picture is possible. The major difference between polyelectronic systems and systems with single electrons is the presence of interactions between the electrons, which, as we have seen, are expressed as Coulomb and exchange integrals. Suppose we are given the task of finding the ‘best’ (i.e. lowest energy) wavefunction for a polyelectronic system. We wish to retain the orbital picture of the system, in which single electrons are assigned to individual spin orbitals. The problem is to find a solution which simultaneously enables all the electronic motions to be taken into account, as a change in the spin orbital for one electron will influence the behaviour of an electron in another spin orbital due to the coupling of the electronic motions. We concentrate on a single electron in a spin orbital χ_i in the field of the nuclei and the other electrons in their (fixed) spin orbitals χ_j . The Hamiltonian operator for the electron in χ_i contains three terms appropriate to the three different contributions to the energy that were identified above (core, Coulomb, exchange). The result can be written as an integro-differential equation for χ_i that has the following form:

$$\begin{aligned} & \left[-\frac{1}{2} \nabla_i^2 - \sum_{A=1}^M \frac{Z_A}{r_{iA}} \right] \chi_i(1) + \sum_{j \neq i} \left[\int d\tau_2 \chi_j(2) \chi_j(2) \frac{1}{r_{12}} \right] \chi_i(1) \\ & - \sum_{j \neq i} \left[\int d\tau_2 \chi_j(2) \chi_i(2) \frac{1}{r_{12}} \right] \chi_i(1) = \sum_j \varepsilon_{ij} \chi_j(1) \end{aligned} \quad (2.124)$$

This expression can be tidied up by introducing three operators that represent the contributions to the energy of the spin orbital χ_i in the ‘frozen’ system:

The core Hamiltonian operator, $\mathcal{H}^{\text{core}}(1)$:

$$\mathcal{H}^{\text{core}}(1) = -\frac{1}{2}\nabla_1^2 - \sum_{A=1}^M \frac{Z_A}{r_{1A}} \quad (2.125)$$

In the absence of any interelectronic interactions this would be the only operator present, corresponding to the motion of a single electron moving in the field of the bare nuclei.

The Coulomb operator, $\mathcal{J}_j(1)$:

$$\mathcal{J}_j(1) = \int d\tau_2 \chi_j(2) \frac{1}{r_{12}} \chi_j(2) \quad (2.126)$$

This operator corresponds to the average potential due to an electron in χ_j .

The exchange operator $\mathcal{K}_j(1)$:

$$\mathcal{K}_j(1)\chi_i(1) = \left[\int d\tau_2 \chi_j(2) \frac{1}{r_{12}} \chi_i(2) \right] \chi_j(1) \quad (2.127)$$

The form of this operator is rather unusual, insofar as it must be defined in terms of its effect when acting on the spin orbital χ_i .

Equation (2.124) can thus be written:

$$\mathcal{H}^{\text{core}}(1)\chi_i(1) + \sum_{j \neq i}^N \mathcal{J}_j(1)\chi_i(1) - \sum_{j \neq i}^N \mathcal{K}_j(1)\chi_i(1) = \sum_j \varepsilon_{ij}\chi_j(1) \quad (2.128)$$

Making use of the fact that $\{\mathcal{J}_i(1) - \mathcal{K}_i(1)\}\chi_i(1) = 0$ leads to the following form:

$$\left[\mathcal{H}^{\text{core}}(1) + \sum_{j=1}^N \{\mathcal{J}_j(1) - \mathcal{K}_j(1)\} \right] \chi_i(1) = \sum_{j=1}^N \varepsilon_{ij}\chi_j(1) \quad (2.129)$$

Or, more simply:

$$\mathcal{F}_i\chi_i = \sum_j \varepsilon_{ij}\chi_j \quad (2.130)$$

\mathcal{F}_i is called the *Fock operator*:

$$\mathcal{F}_i(1) = \mathcal{H}^{\text{core}}(1) + \sum_{j=1}^N \{\mathcal{J}_j(1) - \mathcal{K}_j(1)\} \quad (2.131)$$

For a closed-shell system, the Fock operator has the following form:

$$\mathcal{F}_i(1) = \mathcal{H}^{\text{core}}(1) + \sum_{j=1}^{N/2} \{2\mathcal{J}_j(1) - \mathcal{K}_j(1)\} \quad (2.132)$$

The Fock operator is an effective one-electron Hamiltonian for the electron in the poly-electronic system. However, written in this form of Equation (2.130), the Hartree-Fock

equations do not seem to be particularly useful: on the left-hand side we have the Fock operator acting on the molecular orbital χ_i , but this returns, not the molecular orbital multiplied by a constant as in a normal eigenvalue equation, but rather a series of orbitals χ_j multiplied by some unknown constants ε_{ij} . This is because the solutions to the Hartree–Fock equations are not unique. We have already seen that the value of a determinant is unaffected when the multiple of any column is added to another column. If such a transformation is performed on the Slater determinant, then a different set of constants ε'_{ij} would be obtained with the spin orbitals χ'_i being linear combinations of the first set. Certain transformations give rise to localised orbitals, which are particularly useful for understanding the chemical nature of the system. These localised orbitals are no more ‘correct’ than a delocalised set. Fortunately, it is possible to manipulate Equations (2.130) mathematically so that the Lagrangian multipliers are zero unless the indices i and j are the same. The Hartree–Fock equations then take on the standard eigenvalue form:

$$\mathcal{F}_i \chi_i = \varepsilon_i \chi_i \quad (2.133)$$

Recall that in setting up these equations, each electron has been assumed to move in a ‘fixed’ field comprising the nuclei and the other electrons. This has important implications for the way in which we attempt to find a solution, for any solution that we might find by solving the equation for one electron will naturally affect the solutions for the other electrons in the system. The general strategy is called a *self-consistent field* (SCF) approach. One way to solve these equations is as follows. First, a set of trial solutions χ_i to the Hartree–Fock eigenvalue equations are obtained. These are used to calculate the Coulomb and exchange operators. The Hartree–Fock equations are solved, giving a second set of solutions χ_i' , which are used in the next iteration. The SCF method thus gradually refines the individual electronic solutions that correspond to lower and lower total energies until the point is reached at which the results for all the electrons are unchanged, when they are said to be *self-consistent*.

2.5.1 Hartree–Fock Calculations for Atoms and Slater’s Rules

The Hartree–Fock equations are usually solved in different ways for atoms and for molecules. For atoms, the equations can be solved numerically if it is assumed that the electron distribution is spherically symmetrical. However, these numerical solutions are not particularly useful. Fortunately, analytical approximations to these solutions, which are very similar to those obtained for the hydrogen atom, can be used with considerable success. These approximate analytical functions thus have the form:

$$\psi = R_{nl}(r) Y_{lm}(\theta, \phi) \quad (2.134)$$

Y is a spherical harmonic (as for the hydrogen atom) and R is a radial function. The radial functions obtained for the hydrogen atom cannot be used directly for polyatomic atoms due to the screening of the nuclear charge by the inner shell electrons, but the hydrogen atom functions are acceptable if the orbital exponent is adjusted to account for the screening effect. Even so, the hydrogen atom functions are not particularly convenient to use in molecular orbital calculations due to their complicated functional form. Slater [Slater 1930] suggested

a simpler analytical form for the radial functions:

$$R_{nl}(r) = (2\zeta)^{n+1/2} [(2n)!]^{-1/2} r^{n-1} e^{-\zeta r} \quad (2.135)$$

These functions are universally known as *Slater type orbitals* (STOs) and are just the leading term in the appropriate Laguerre polynomials. The first three Slater functions are as follows.

$$R_{1s}(r) = 2\zeta^{3/2} e^{-\zeta r} \quad (2.136)$$

$$R_{2s}(r) = R_{2p}(r) = \left(\frac{4\zeta^5}{3}\right)^{1/2} r e^{-\zeta r} \quad (2.137)$$

$$R_{3s}(r) = R_{3p}(r) = R_{3d}(r) = \left(\frac{8\zeta^7}{45}\right)^{1/2} r^2 e^{-\zeta r} \quad (2.138)$$

To obtain the whole orbital we must multiply $R(r)$ by the appropriate angular part. For example, we would use the following expressions for the 1s, 2s and 2p_z orbitals:

$$\phi_{1s}(\mathbf{r}) = \sqrt{\zeta^3/\pi} \exp(-\zeta r) \quad (2.139)$$

$$\phi_{2s}(\mathbf{r}) = \sqrt{\zeta^5/3\pi} \mathbf{r} \exp(-\zeta r) \quad (2.140)$$

$$\phi_{2p_z}(\mathbf{r}) = \sqrt{\zeta^5/\pi} \exp(-\zeta r) \cos \theta \quad (2.141)$$

Slater provided a series of empirical rules for choosing the orbital exponents ζ , which are given by:

$$\zeta = \frac{Z - \sigma}{n^*} \quad (2.142)$$

Z is the atomic number and σ is a *shielding constant*, determined as below. n^* is an effective principal quantum number, which takes the same value as the true principal quantum number for $n = 1, 2$ or 3 , but for $n = 4, 5, 6$ has the values $3.7, 4.0, 4.2$, respectively. The shielding constant is obtained as follows:

First, divide the orbitals into the following groups:

$$(1s); (2s, 2p); (3s, 3p); (3d); (4s, 4p); (4d); (4f); (5s, 5p); (5d) \quad (2.143)$$

For a given orbital, σ is obtained by adding together the following contributions:

- (a) zero from an orbital further from the nucleus than those in the group;
- (b) 0.35 from each other electron in the same group, but if the other orbital is the 1s then the contribution is 0.3;
- (c) 1.0 for each electron in a group with a principal quantum number 2 or more fewer than the current orbital;
- (d) for each electron with a principal quantum number 1 fewer than the current orbital: 1.0 if the current orbital is d or f; 0.85 if the current orbital is s or p.

The shielding constant for the valence electrons of silicon is obtained using Slater's rules as follows. The electronic configuration of Si is $(1s^2)(2s^2 2p^6)(3s^2 3p^2)$. We therefore count

3×0.35 under rule (b), 2.0 under rule (c) and 8×0.85 under rule (d), giving a total of 9.85. When subtracted from the atomic number (14) this gives 4.15 for the value of $Z - \sigma$.

2.5.2 Linear Combination of Atomic Orbitals (LCAO) in Hartree–Fock Theory

Direct solution of the Hartree–Fock equations is not a practical proposition for molecules and so it is necessary to adopt an alternative approach. The most popular strategy is to write each spin orbital as a linear combination of single electron orbitals:

$$\psi_i = \sum_{\nu=1}^K c_{\nu i} \phi_{\nu} \quad (2.144)$$

The one-electron orbitals ϕ_{ν} are commonly called *basis functions* and often correspond to the atomic orbitals. We will label the basis functions with the Greek letters μ , ν , λ and σ . In the case of Equation (2.144) there are K basis functions and we should therefore expect to derive a total of K molecular orbitals (although not all of these will necessarily be occupied by electrons). The smallest number of basis functions for a molecular system will be that which can just accommodate all the electrons in the molecule. More sophisticated calculations use more basis functions than a minimal set. At the *Hartree–Fock limit* the energy of the system can be reduced no further by the addition of any more basis functions; however, it may be possible to lower the energy below the Hartree–Fock limit by using a functional form of the wavefunction that is more extensive than the single Slater determinant.

In accordance with the variation theorem we require the set of coefficients $c_{\nu i}$ that gives the lowest-energy wavefunction, and some scheme for changing the coefficients to derive that wavefunction. For a given basis set and a given functional form of the wavefunction (i.e. a Slater determinant) the best set of coefficients is that for which the energy is a minimum, at which point

$$\frac{\partial E}{\partial c_{\nu i}} = 0 \quad (2.145)$$

for all coefficients $c_{\nu i}$. The objective is thus to determine the set of coefficients that gives the lowest energy for the system.

2.5.3 Closed-shell Systems and the Roothaan–Hall Equations

We shall initially consider a closed-shell system with N electrons in $N/2$ orbitals. The derivation of the Hartree–Fock equations for such a system was first proposed by Roothaan [Roothaan 1951] and (independently) by Hall [Hall 1951]. The resulting equations are known as the Roothaan equations or the Roothaan–Hall equations. Unlike the integro-differential form of the Hartree–Fock equations, Equation (2.124), Roothaan and Hall recast the equations in matrix form, which can be solved using standard techniques and can be applied to systems of any geometry. We shall identify the major steps in the Roothaan approach,

starting with the expression for the Hartree–Fock energy for our closed-shell system, Equation (2.120):

$$E = 2 \sum_{i=1}^{N/2} H_{ii}^{\text{core}} + \sum_{i=1}^{N/2} \sum_{j=1}^{N/2} (2J_{ij} - K_{ij}) \quad (2.146)$$

The corresponding Fock operator is (Equation (2.132)):

$$\mathcal{F}_i(1) = \mathcal{H}^{\text{core}}(1) + \sum_{j=1}^{N/2} \{2J_j(1) - K_j(1)\} \quad (2.147)$$

We now introduce the atomic orbital expansion for the orbitals ψ_i and substitute for the corresponding spin orbital χ_i into the Hartree–Fock equation, $\mathcal{F}_i(1)\chi_i(1) = \varepsilon_i\chi_i(1)$:

$$\mathcal{F}_i(1) \sum_{\nu=1}^K c_{\nu i} \phi_{\nu}(1) = \varepsilon_i \sum_{\nu=1}^K c_{\nu i} \phi_{\nu}(1) \quad (2.148)$$

Pre-multiplying each side by $\phi_{\mu}(1)$ (where ϕ_{μ} is also a basis function) and integrating gives the following matrix equation:

$$\sum_{\nu=1}^K c_{\nu i} \int d\nu_1 \phi_{\mu}(1) \mathcal{F}_i(1) \phi_{\nu}(1) = \varepsilon_i \sum_{\nu=1}^K c_{\nu i} \int d\nu_1 \phi_{\mu}(1) \phi_{\nu}(1) \quad (2.149)$$

$\int d\nu_1 \phi_{\mu}(1) \phi_{\nu}(1)$ is the overlap integral between the basis functions μ and ν , written $S_{\mu\nu}$. Unlike the molecular orbitals, which will be required to be orthonormal, the overlap between two basis functions is not necessarily zero (for example, they may be located on different atoms).

The elements of the *Fock matrix* are given by

$$F_{\mu\nu} = \int d\nu_1 \phi_{\mu}(1) \mathcal{F}_i(1) \phi_{\nu}(1) \quad (2.150)$$

The Fock matrix elements for a closed-shell system can be expanded as follows by substituting the expression for the Fock operator:

$$F_{\mu\nu} = \int d\nu_1 \phi_{\mu}(1) \mathcal{H}^{\text{core}}(1) \phi_{\nu}(1) + \sum_{j=1}^{N/2} \int d\nu_1 \phi_{\mu}(1) [2J_j(1) - K_j(1)] \phi_{\nu}(1) \quad (2.151)$$

The elements of the Fock matrix can thus be written as the sum of core, Coulomb and exchange contributions. The core contribution is:

$$\int d\nu_1 \phi_{\mu}(1) \mathcal{H}^{\text{core}}(1) \phi_{\nu}(1) = \int d\nu_1 \phi_{\mu}(1) \left[-\frac{1}{2} \nabla^2 - \sum_{A=1}^M \frac{Z_A}{|r_1 - R_A|} \right] \phi_{\nu}(1) \equiv H_{\mu\nu}^{\text{core}} \quad (2.152)$$

The core contributions thus require the calculation of integrals that involve basis functions on up to two centres (depending upon whether ϕ_{μ} and ϕ_{ν} are centred on the same nucleus or not). Each element $H_{\mu\nu}^{\text{core}}$ can in turn be obtained as the sum of a kinetic energy integral and a potential energy integral corresponding to the two terms in the one-electron Hamiltonian.

The Coulomb and exchange contributions to the Fock matrix element $F_{\mu\nu}$ are together given by:

$$\sum_{j=1}^{N/2} \int d\nu_1 \phi_\mu(1) [2\mathcal{J}_j(1) - \mathcal{K}_j(1)] \phi_\nu(1) \quad (2.153)$$

Recall that the Coulomb operator $\mathcal{J}_j(1)$ due to interaction with a spin orbital χ_j is given by:

$$\mathcal{J}_j(1) = \int d\tau_2 \chi_j(2) \frac{1}{r_{12}} \chi_j(2) \quad (2.154)$$

We need to write each of the two occurrences of the spin orbital χ_j in this integral in terms of the appropriate linear combination of basis functions:

$$\mathcal{J}_j(1) = \int d\tau_2 \sum_{\sigma=1}^K c_{\sigma j} \phi_\sigma(2) \frac{1}{r_{12}} \sum_{\lambda=1}^K c_{\lambda j} \phi_\lambda(2) \quad (2.155)$$

We have used the indices σ and λ for the basis functions here. Similarly, the exchange contribution can be written:

$$\mathcal{K}_j(1) \chi_i(1) = \left[\int d\tau_2 \sum_{\sigma=1}^K c_{\sigma j} \phi_\sigma(2) \frac{1}{r_{12}} \chi_i(2) \right] \sum_{\lambda=1}^K c_{\lambda j} \phi_\lambda(2) \quad (2.156)$$

When the Coulomb and exchange operators are expressed in terms of the basis functions and the orbital expansion is substituted for χ_i , then their contributions to the Fock matrix element $F_{\mu\nu}$ take the following form:

$$\begin{aligned} & \sum_{j=1}^{N/2} \int d\nu_1 \phi_\mu(1) [2\mathcal{J}_j(1) - \mathcal{K}_j(1)] \phi_\nu(1) \\ &= \sum_{j=1}^{N/2} \sum_{\lambda=1}^K \sum_{\sigma=1}^K c_{\lambda j} c_{\sigma j} \left[\begin{aligned} & 2 \int d\nu_1 d\nu_2 \phi_\mu(1) \phi_\nu(1) \frac{1}{r_{12}} \phi_\lambda(2) \phi_\sigma(2) \\ & - \int d\nu_1 d\nu_2 \phi_\mu(1) \phi_\lambda(1) \frac{1}{r_{12}} \phi_\nu(2) \phi_\sigma(2) \end{aligned} \right] \\ &\equiv \sum_{j=1}^{N/2} \sum_{\lambda=1}^K \sum_{\sigma=1}^K c_{\lambda j} c_{\sigma j} [2(\mu\nu|\lambda\sigma) - (\mu\lambda|\nu\sigma)] \end{aligned} \quad (2.157)$$

We have used the shorthand notation for the integrals in the final expression. Note that the two-electron integrals may involve up to four different basis functions ($\mu, \nu, \lambda, \sigma$), which may in turn be located at four different centres. This has important consequences for the way in which we try to solve the equations.

It is helpful to simplify Equation (2.157) by introducing the *charge density matrix*, \mathbf{P} , whose elements are defined as:

$$P_{\mu\nu} = 2 \sum_{i=1}^{N/2} c_{\mu i} c_{\nu i} \quad \text{and} \quad P_{\lambda\sigma} = 2 \sum_{i=1}^{N/2} c_{\lambda i} c_{\sigma i} \quad (2.158)$$

Note that the summations are over the $N/2$ occupied orbitals. Other properties can be calculated from the density matrix; for example, the electronic energy is:

$$E = \frac{1}{2} \sum_{\mu=1}^K \sum_{\nu=1}^K P_{\mu\nu} (H_{\mu\nu}^{\text{core}} + F_{\mu\nu}) \quad (2.159)$$

The electron density at a point \mathbf{r} can also be expressed in terms of the density matrix:

$$\rho(\mathbf{r}) = \sum_{\mu=1}^K \sum_{\nu=1}^K P_{\mu\nu} \phi_{\mu}(\mathbf{r}) \phi_{\nu}(\mathbf{r}) \quad (2.160)$$

The expression for each element $F_{\mu\nu}$ of the Fock matrix elements for a closed-shell system of N electrons then becomes:

$$F_{\mu\nu} = H_{\mu\nu}^{\text{core}} + \sum_{\lambda=1}^K \sum_{\sigma=1}^K P_{\lambda\sigma} [(\mu\nu|\lambda\sigma) - \frac{1}{2} (\mu\lambda|\nu\sigma)] \quad (2.161)$$

This is the standard form for the expression for the Fock matrix in the Roothaan–Hall equations

2.5.4 Solving the Roothaan–Hall Equations

The Fock matrix is a $K \times K$ square matrix that is symmetric if real basis functions are used. The Roothaan–Hall equations (2.149) can be conveniently written as a matrix equation:

$$\mathbf{FC} = \mathbf{SCE} \quad (2.162)$$

The elements of the $K \times K$ matrix \mathbf{C} are the coefficients $c_{\nu i}$:

$$\mathbf{C} = \begin{pmatrix} c_{1,1} & c_{1,2} & \dots & c_{1,K} \\ c_{2,1} & c_{2,2} & \dots & c_{2,K} \\ \vdots & \vdots & & \vdots \\ c_{K,1} & c_{K,2} & \dots & c_{K,K} \end{pmatrix} \quad (2.163)$$

\mathbf{E} is a diagonal matrix whose elements are the orbital energies:

$$\mathbf{E} = \begin{pmatrix} \varepsilon_1 & 0 & \dots & 0 \\ 0 & \varepsilon_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \varepsilon_K \end{pmatrix} \quad (2.164)$$

Let us consider how we might solve the Roothaan–Hall equations and thereby obtain the molecular orbitals. The first point we must note is that the elements of the Fock matrix, which appear on the left-hand side of Equation (2.162), depend on the molecular orbital coefficients $c_{\nu i}$, which also appear on the right-hand side of the equation. Thus an iterative procedure is required to find a solution.

The one-electron contributions $H_{\mu\nu}^{\text{core}}$ due to the electrons moving in the field of the bare nuclei do not depend on the basis set coefficients and remain unchanged throughout the calculation. However, the Coulomb and exchange contributions do depend on the coefficients and we would expect these to vary throughout the calculation. The individual two-electron integrals $(\mu\nu|\lambda\sigma)$ are, however, constant throughout the calculation. An obvious strategy is thus to calculate and store these integrals for later use.

Having written the Roothaan–Hall equations in matrix form we would obviously like to solve them using standard matrix eigenvalue methods (discussed in Section 1.10.3). However, standard eigenvalue methods would require an equation of the form $\mathbf{FC} = \mathbf{CE}$. The Roothaan–Hall equations only adopt such a form if the overlap matrix, \mathbf{S} , is equal to the unit matrix, \mathbf{I} (in which all diagonal elements are equal to 1 and all off-diagonal elements are zero). The functions ϕ are usually normalised but they are not necessarily orthogonal (for example, because they are located on different atoms) and so there will invariably be non-zero off-diagonal elements of the overlap matrix. To solve the Roothaan–Hall equations using standard methods they must be transformed. This corresponds to transforming the basis functions so that they form an orthonormal set. We seek a matrix \mathbf{X} , such that $\mathbf{X}^T \mathbf{S} \mathbf{X} = \mathbf{I}$. \mathbf{X}^T is the transpose of \mathbf{X} , obtained by interchanging rows and columns. There are various ways in which \mathbf{X} can be calculated; in *symmetric orthogonalisation*, the overlap matrix is diagonalised. Diagonalisation involves finding the matrix \mathbf{U} such that

$$\mathbf{U}^T \mathbf{S} \mathbf{U} = \mathbf{D} = \text{diag}(\lambda_1 \dots \lambda_K) \quad (2.165)$$

\mathbf{D} is the diagonal matrix containing the eigenvalues λ_i of \mathbf{S} , and \mathbf{U} contains the eigenvectors of \mathbf{S} . \mathbf{U}^T is the transpose of the matrix \mathbf{U} . (This expression is often written $\mathbf{U}^{-1} \mathbf{S} \mathbf{U} = \mathbf{D}$ since for real basis functions $\mathbf{U}^{-1} = \mathbf{U}^T$.) Then the matrix \mathbf{X} is given by $\mathbf{X} = \mathbf{U} \mathbf{D}^{-1/2} \mathbf{U}^T$, where $\mathbf{D}^{-1/2}$ is formed from the inverse square roots of \mathbf{D} . We shall write \mathbf{X} as $\mathbf{S}^{-1/2}$, as it can be considered to be the inverse square root of the overlap matrix: $\mathbf{S}^{-1/2} \mathbf{S} \mathbf{S}^{-1/2} = \mathbf{I}$.

The Roothaan–Hall equations can now be manipulated as follows. Both sides of Equation (2.162) are pre-multiplied by the matrix $\mathbf{S}^{-1/2}$:

$$\mathbf{S}^{-1/2} \mathbf{F} \mathbf{C} = \mathbf{S}^{-1/2} \mathbf{S} \mathbf{C} \mathbf{E} = \mathbf{S}^{1/2} \mathbf{C} \mathbf{E} \quad (2.166)$$

Inserting the unit matrix, in the form $\mathbf{S}^{-1/2} \mathbf{S}^{1/2}$, into the left-hand side gives:

$$\mathbf{S}^{-1/2} \mathbf{F} (\mathbf{S}^{-1/2} \mathbf{S}^{1/2}) \mathbf{C} = \mathbf{S}^{1/2} \mathbf{C} \mathbf{E} \quad (2.167)$$

or

$$\mathbf{S}^{-1/2} \mathbf{F} \mathbf{S}^{-1/2} (\mathbf{S}^{1/2} \mathbf{C}) = (\mathbf{S}^{1/2} \mathbf{C}) \mathbf{E} \quad (2.168)$$

Equation (2.168) can be written $\mathbf{F}' \mathbf{C}' = \mathbf{C}' \mathbf{E}$, where $\mathbf{F}' = \mathbf{S}^{-1/2} \mathbf{F} \mathbf{S}^{-1/2}$ and $\mathbf{C}' = \mathbf{S}^{1/2} \mathbf{C}$.

The matrix equation $\mathbf{F}' \mathbf{C}' = \mathbf{C}' \mathbf{E}$ can be solved using standard methods; a solution only exists if the determinant $|\mathbf{F}' - \mathbf{E}\mathbf{I}|$ equals zero. In simple cases this can be done by multiplying out the determinant to give a polynomial (the secular equation) whose roots are the eigenvalues ε_i , but for large matrices a much more practical approach involves the diagonalisation of \mathbf{F}' . The matrix of coefficients, \mathbf{C}' , are the eigenvectors of \mathbf{F}' . The basis function coefficients \mathbf{C} can then be obtained from \mathbf{C}' using $\mathbf{C} = \mathbf{S}^{-1/2} \mathbf{C}'$. A common scheme for

solving the Roothaan–Hall equations is thus as follows:

1. Calculate the integrals to form the Fock matrix, \mathbf{F} .
2. Calculate the overlap matrix, \mathbf{S} .
3. Diagonalise \mathbf{S} .
4. Form $\mathbf{S}^{-1/2}$.
5. Guess, or otherwise calculate, an initial density matrix, \mathbf{P} .
6. Form the Fock matrix using the integrals and the density matrix \mathbf{P} .
7. Form $\mathbf{F}' = \mathbf{S}^{-1/2} \mathbf{F} \mathbf{S}^{-1/2}$.
8. Solve the secular equation $|\mathbf{F}' - \mathbf{E}\mathbf{I}| = 0$ to give the eigenvalues \mathbf{E} and the eigenvectors \mathbf{C}' by diagonalising \mathbf{F}' .
9. Calculate the molecular orbital coefficients, \mathbf{C} from $\mathbf{C} = \mathbf{S}^{-1/2} \mathbf{C}'$.
10. Calculate a new density matrix, \mathbf{P} , from the matrix \mathbf{C} .
11. Check for convergence. If the calculation has converged, stop. Otherwise repeat from step 6 using the new density matrix, \mathbf{P} .

This procedure requires an initial guess of the density matrix, \mathbf{P} . The simplest approach is to use the null matrix, which corresponds to ignoring all the electron–electron terms so that the electrons just experience the bare nuclei. This can sometimes lead to convergence problems, which may be prevented if a lower level of theory (such as semi-empirical or extended Hückel) is used to provide the initial guess. Moreover, a better guess may enable the calculation to be performed more quickly. A variety of criteria can be used to establish whether the calculation has converged or not. For example, the density matrix can be compared with that from the previous iteration, and/or the change in energy can be monitored together with the basis set coefficients.

The result of a Hartree–Fock calculation is a set of K molecular orbitals, where K is the number of basis functions in the calculation. The N electrons are then fed into these orbitals in accordance with the Aufbau principle, two electrons per orbital, starting with the lowest-energy orbitals. The remaining orbitals do not contain any electrons; these are known as the *virtual orbitals*. Alternative electronic configurations can be generated by exciting electrons from the occupied orbitals to the virtual orbitals; these excited configurations are used in more advanced calculations that will be discussed in Chapter 3.

A Hartree–Fock calculation provides a set of orbital energies, ε_i . What is the significance of these? The energy of an electron in a spin orbital is calculated by adding the core interaction H_{ii}^{core} to the Coulomb and exchange interactions with the other electrons in the system:

$$\varepsilon_i = H_{ii}^{\text{core}} + \sum_{j=1}^{N/2} (2J_{ij} - K_{ij}) \quad (2.169)$$

The total electronic energy of the ground state is given by Equation (2.120):

$$E = 2 \sum_{i=1}^{N/2} H_{ii}^{\text{core}} + \sum_{i=1}^{N/2} \sum_{j=1}^{N/2} (2J_{ij} - K_{ij}) \quad (2.170)$$

The total energy is therefore not equal to the sum of the individual orbital energies but is related as follows:

$$E = \sum_{i=1}^N \varepsilon_i - \sum_{i=1}^{N/2} \sum_{j=1}^{N/2} (2J_{ij} - K_{ij}) \quad (2.171)$$

The reason for the discrepancy is that the individual orbital energies include contributions from the interaction between that electron and all the nuclei and all other electrons in the system. The Coulomb and exchange interactions between pairs of electrons are therefore counted twice when summing the individual orbital energies.

2.5.5 A Simple Illustration of the Roothaan–Hall Approach

We will illustrate the stages involved in the Roothaan–Hall approach using the helium hydrogen molecular ion, HeH^+ , as an example. This is a two-electron system. Our objective here is to show how the Roothaan–Hall method can be used to derive the wavefunction, for a fixed internuclear distance of 1 Å. We use HeH^+ rather than H_2 as our system as the lack of symmetry in HeH^+ makes the procedure more informative. There are two basis functions, $1s_A$ (centred on the helium atom) and $1s_B$ (on the hydrogen). The numerical values of the integrals that we shall use in our calculation were obtained using a Gaussian series approximation to the Slater orbitals (the STO-3G basis set, which is described in Section 2.6). This detail need not concern us here. Each wavefunction is expressed as a linear combination of the two 1s atomic orbitals centred on the nuclei A and B:

$$\psi_1 = c_{1A} 1s_A + c_{1B} 1s_B \quad (2.172)$$

$$\psi_2 = c_{2A} 1s_A + c_{2B} 1s_B \quad (2.173)$$

First, it is necessary to calculate the various one- and two-electron integrals and to formulate the Fock and overlap matrices, each of which will be a 2×2 symmetric matrix (as there are two orbitals in the basis set). The diagonal elements of the overlap matrix, S , are equal to 1.0 as each basis function is normalised; the off-diagonal elements have smaller, but non-zero, values that are equal to the overlap between $1s_A$ and $1s_B$ for the internuclear distance chosen. The matrix S is:

$$S = \begin{pmatrix} 1.0 & 0.392 \\ 0.392 & 1.0 \end{pmatrix} \quad (2.174)$$

The core contributions $H_{\mu\nu}^{\text{core}}$ can be calculated as the sum of three 2×2 matrices comprising the kinetic energy (T) and nuclear attraction terms for the two nuclei A and B (V_A and V_B). The elements of these three matrices are obtained by evaluating the following integrals:

$$\begin{aligned} T_{\mu\nu} &= \int d\nu_1 \phi_\mu(1) \left(-\frac{1}{2} \nabla^2\right) \phi_\nu(1) \\ V_{A,\mu\nu} &= \int d\nu_1 \phi_\mu(1) \left(-\frac{Z_A}{r_{1A}}\right) \phi_\nu(1) \\ V_{B,\mu\nu} &= \int d\nu_1 \phi_\mu(1) \left(-\frac{Z_B}{r_{1B}}\right) \phi_\nu(1) \end{aligned} \quad (2.175)$$

The matrices are:

$$\mathbf{T} = \begin{pmatrix} 1.412 & 0.081 \\ 0.081 & 0.760 \end{pmatrix} \quad \mathbf{V}_A = \begin{pmatrix} -3.344 & -0.758 \\ -0.758 & -1.026 \end{pmatrix} \quad \mathbf{V}_B = \begin{pmatrix} -0.525 & -0.308 \\ -0.308 & -1.227 \end{pmatrix} \quad (2.176)$$

\mathbf{H}^{core} is the sum of these three:

$$\mathbf{H}^{\text{core}} = \begin{pmatrix} -2.457 & -0.985 \\ -0.985 & -1.493 \end{pmatrix} \quad (2.177)$$

As far as the two-electron integrals are concerned, with two basis functions there are a total of 16 possible two-electron integrals. There are however only six unique two-electron integrals, as the indices can be permuted as follows:

- (i) $(1s_A 1s_A | 1s_A 1s_A) = 1.056$
- (ii) $(1s_A 1s_A | 1s_A 1s_B) = (1s_A 1s_A | 1s_B 1s_A) = (1s_A 1s_B | 1s_A 1s_A)$
 $= (1s_B 1s_A | 1s_A 1s_A) = 0.303$
- (iii) $(1s_A 1s_B | 1s_A 1s_B) = (1s_A 1s_B | 1s_B 1s_A) = (1s_B 1s_A | 1s_A 1s_B)$
 $= (1s_B 1s_A | 1s_B 1s_A) = 0.112$
- (iv) $(1s_A 1s_A | 1s_B 1s_B) = (1s_B 1s_B | 1s_A 1s_A) = 0.496$
- (v) $(1s_A 1s_B | 1s_B 1s_B) = (1s_B 1s_A | 1s_B 1s_B) = (1s_B 1s_B | 1s_A 1s_B)$
 $= (1s_B 1s_B | 1s_B 1s_A) = 0.244$
- (vi) $(1s_B 1s_B | 1s_B 1s_B) = 0.775$

To reiterate, these integrals are calculated as follows:

$$(\mu\nu|\lambda\sigma) = \int \int d\nu_1 d\nu_2 \phi_\mu(1) \phi_\nu(1) \frac{1}{r_{12}} \phi_\lambda(2) \phi_\sigma(2) \quad (2.178)$$

Having calculated the integrals, we are now ready to start the SCF calculation. To formulate the Fock matrix it is necessary to have an initial guess of the density matrix, \mathbf{P} . The simplest approach is to use the null matrix in which all elements are zero. In this initial step the Fock matrix \mathbf{F} is therefore equal to \mathbf{H}^{core} .

The Fock matrix must next be transformed to \mathbf{F}' by pre- and post-multiplying by $\mathbf{S}^{-1/2}$:

$$\mathbf{S}^{-1/2} = \begin{pmatrix} -1.065 & -0.217 \\ -0.217 & 1.065 \end{pmatrix} \quad (2.179)$$

\mathbf{F}' for this first iteration is thus:

$$\mathbf{F}' = \begin{pmatrix} -2.401 & -0.249 \\ -0.249 & -1.353 \end{pmatrix} \quad (2.180)$$

Diagonalisation of \mathbf{F}' gives its eigenvalues and eigenvectors, which are:

$$\mathbf{E} = \begin{pmatrix} -2.458 & 0.0 \\ 0.0 & -1.292 \end{pmatrix} \quad \mathbf{C}' = \begin{pmatrix} 0.975 & -0.220 \\ 0.220 & 0.975 \end{pmatrix} \quad (2.181)$$

The coefficients \mathbf{C} are obtained from $\mathbf{C} = \mathbf{S}^{-1/2} \mathbf{C}'$ and are thus:

$$\mathbf{C} = \begin{pmatrix} 0.991 & -0.446 \\ 0.022 & 1.087 \end{pmatrix} \quad (2.182)$$

To formulate the density matrix P we need to identify the occupied orbital(s). With a two-electron system both electrons occupy the orbital with the lowest energy (i.e. the orbital with the lowest eigenvalue). At this stage the lowest-energy orbital is:

$$\psi = 0.991 \text{ } 1s_A + 0.022 \text{ } 1s_B \quad (2.183)$$

The orbital is composed largely of the s orbital on the helium nucleus; in the absence of any electron-electron repulsion the electrons tend to congregate near the nucleus with the larger charge. The density matrix corresponding to this initial wavefunction is:

$$P = \begin{pmatrix} 1.964 & 0.044 \\ 0.044 & 0.001 \end{pmatrix} \quad (2.184)$$

The new Fock matrix is formed using P and the two-electron integrals together with H^{core} . For example, the element F_{11} is given by:

$$\begin{aligned} F_{11} = H_{11}^{\text{core}} &+ P_{11}[(1s_A 1s_A | 1s_A 1s_A) - \frac{1}{2}(1s_A 1s_A | 1s_A 1s_A)] \\ &+ P_{12}[(1s_A 1s_A | 1s_A 1s_B) - \frac{1}{2}(1s_A 1s_A | 1s_A 1s_B)] \\ &+ P_{21}[(1s_A 1s_A | 1s_B 1s_B) - \frac{1}{2}(1s_A 1s_B | 1s_A 1s_B)] \\ &+ P_{12}[(1s_A 1s_A | 1s_B 1s_B) - \frac{1}{2}(1s_A 1s_B | 1s_A 1s_B)] \end{aligned} \quad (2.185)$$

The complete Fock matrix is:

$$F = \begin{pmatrix} -1.406 & -0.690 \\ -0.690 & -0.618 \end{pmatrix} \quad (2.186)$$

The energy that corresponds to this Fock matrix (calculated using Equation (2.159)) is -3.870 Hartree. In the next iteration, the various matrices are as follows:

$$\begin{aligned} F' &= \begin{pmatrix} -1.305 & -0.347 \\ -0.347 & -0.448 \end{pmatrix} & E &= \begin{pmatrix} -1.427 & 0.0 \\ 0.0 & -0.325 \end{pmatrix} \\ C' &= \begin{pmatrix} 0.943 & -0.334 \\ 0.334 & 0.943 \end{pmatrix} & C &= \begin{pmatrix} 0.931 & -0.560 \\ 0.150 & 1.076 \end{pmatrix} \\ P &= \begin{pmatrix} 1.735 & 0.280 \\ 0.280 & 0.045 \end{pmatrix} & F &= \begin{pmatrix} -1.436 & -0.738 \\ -0.738 & -0.644 \end{pmatrix} \\ \text{Energy} &= -3.909 \text{ Hartree} \end{aligned} \quad (2.187)$$

The calculation proceeds as illustrated in Table 2.2, which shows the variation in the coefficients of the atomic orbitals in the lowest-energy wavefunction and the energy for the first four SCF iterations. The energy is converged to six decimal places after six iterations and the charge density matrix after nine iterations.

The final wavefunction still contains a large proportion of the 1s orbital on the helium atom, but less than was obtained without the two-electron integrals.

Iteration	$c(1s_A)$	$c(1s_B)$	Energy
1	0.991	0.022	-3.870
2	0.931	0.150	-3.909
3	0.915	0.181	-3.911
4	0.912	0.187	-3.911

Table 2.2 Variation in basis set coefficients and electronic energy for the HeH^+ molecule.

2.5.6 Application of the Hartree–Fock Equations to Molecular Systems

We are now in a position to consider how the Hartree–Fock theory we have developed can be used to perform practical quantum mechanical calculations on molecular systems. This is an appropriate place in our discussion to distinguish the two major categories of quantum mechanical molecular orbital calculations: the *ab initio* and the semi-empirical methods. *Ab initio* strictly means ‘from the beginning’, or ‘from first principles’, which would imply that a calculation using such an approach would require as input only physical constants such as the speed of light, Planck’s constant, the masses of elementary particles, and so on. *Ab initio* in fact usually refers to a calculation which uses the full Hartree–Fock/Roothaan–Hall equations, without ignoring or approximating any of the integrals or any of the terms in the Hamiltonian. The *ab initio* methods do rely upon calibration calculations, and this has led some quantum chemists, notably Dewar (who has played a large part in the development of semi-empirical methods), to claim that any real difference between the *ab initio* and semi-empirical methods is entirely pedagogical. By contrast, semi-empirical methods simplify the calculations, using parameters for some of the integrals and/or ignoring some of the terms in the Hamiltonian. First we shall consider *ab initio* methods.

2.6 Basis Sets

The basis sets most commonly used in quantum mechanical calculations are composed of atomic functions. An obvious choice would be the Slater type orbitals. Unfortunately, Slater functions are not particularly amenable to implementation in molecular orbital calculations. This is because some of the integrals are difficult, if not impossible, to evaluate, particularly when the atomic orbitals are centred on different nuclei. It is relatively straightforward to calculate integrals involving one or two centres, such as $(\mu\mu|\nu\nu)$, $(\mu\nu|\nu\nu)$ and $(\mu\nu|\mu\nu)$. Three- and four-centre integrals are also feasible with Slater functions if the atomic orbitals are located on the same atom. However, three- and four-centre integrals are very difficult if the atomic orbitals are based on different atoms. It is common in *ab initio* calculations to replace the Slater orbitals by functions based upon Gaussians. A Gaussian function has the form $\exp(-\alpha r^2)$, and *ab initio* calculations use basis functions comprising integral powers of x , y and z multiplied by $\exp(-\alpha r^2)$:

$$x^a y^b z^c \exp(-\alpha r^2) \quad (2.188)$$

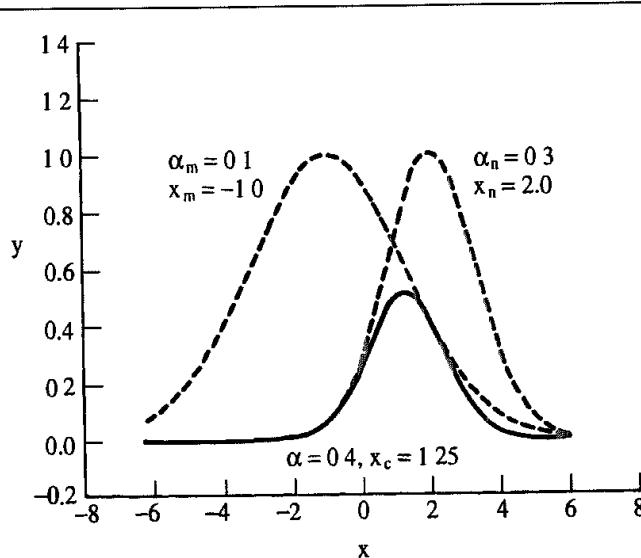


Fig 2.4: The product of two Gaussian functions is another Gaussian centred along the line joining their centres. In this case the equations of the two functions are $y = \exp[-0.1(x + 1.0)^2]$ and $y = \exp[-0.3(x - 2.0)^2]$ and the equation of the product is $y = \exp(-27/40)[-0.4(x - 1.25)^2]$ (Equation (2.189))

α determines the radial extent (or 'spread') of a Gaussian function; a function with a large value of α does not spread very far, whereas a small value of α gives a large spread. The *order* of these Gaussian-type functions is determined by the powers of the Cartesian variables; a zeroth-order function has $a + b + c = 0$; a first-order function has $a + b + c = 1$, and so on. There is thus one zeroth-order function, three first-order functions and six second-order functions. The idea of using Gaussian functions in quantum mechanical calculations is often ascribed to Boys [Boys 1950]. A major advantage of Gaussian functions is that the product of two Gaussians can be expressed as a single Gaussian, located along the line joining the centres of the two Gaussians m and n (Figure 2.4):

$$\exp(-\alpha_m r_m^2) \exp(-\alpha_n r_n^2) = \exp\left(-\frac{\alpha_m \alpha_n}{\alpha_m + \alpha_n} r_{mn}^2\right) \exp(-\alpha r_c^2) \quad (2.189)$$

r_{mn} is the distance between the centres m and n , and the orbital exponent α of the combined function is related to the exponents α_m and α_n by:

$$\alpha = \alpha_m + \alpha_n \quad (2.190)$$

r_C is the distance from point C, which has coordinates:

$$x_c = \frac{\alpha_m x_m + \alpha_n x_n}{\alpha_m + \alpha_n}; \quad y_c = \frac{\alpha_m y_m + \alpha_n y_n}{\alpha_m + \alpha_n}; \quad z_c = \frac{\alpha_m z_m + \alpha_n z_n}{\alpha_m + \alpha_n} \quad (2.191)$$

x_m, y_m, z_m and x_n, y_n, z_n are the centres of the two original Gaussians m and n respectively.

Thus, in a two-electron integral of the form $(\mu\nu|\lambda\sigma)$, the product $\phi_\mu(1)\phi_\nu(1)$ (where ϕ_μ and ϕ_ν may be on different centres) can be replaced by a single Gaussian function that is centred at the appropriate point C. For Cartesian Gaussian functions the calculation is more complicated than for the example we have stated above, due to the presence of the Cartesian functions, but even so, efficient methods for performing the integrals have been devised.

The zeroth-order Gaussian function g_s has s-orbital angular symmetry; the three first-order Gaussian functions have p-orbital symmetry. In normalised form these are:

$$g_s(\alpha, r) = \left(\frac{2\alpha}{\pi} \right)^{3/4} e^{-\alpha r^2} \quad (2.192)$$

$$g_x(\alpha, r) = \left(\frac{128\alpha^5}{\pi^3} \right)^{1/4} x e^{-\alpha r^2} \quad (2.193)$$

$$g_y(\alpha, r) = \left(\frac{128\alpha^5}{\pi^3} \right)^{1/4} y e^{-\alpha r^2} \quad (2.194)$$

$$g_z(\alpha, r) = \left(\frac{128\alpha^5}{\pi^3} \right)^{1/4} z e^{-\alpha r^2} \quad (2.195)$$

The six second-order functions have the following form, exemplified by two of the functions:

$$g_{xx}(\alpha, r) = \left(\frac{2048\alpha^7}{9\pi^3} \right)^{1/4} x^2 e^{-\alpha r^2} \quad (2.196)$$

$$g_{xy}(\alpha, r) = \left(\frac{2048\alpha^7}{9\pi^3} \right)^{1/4} xy e^{-\alpha r^2} \quad (2.197)$$

These second-order functions do not all have the same angular symmetry as the 3d atomic orbitals, but a set comprising g_{xy} , g_{xz} and g_{yz} , together with two linear combinations of the g_{xx} , g_{yy} and g_{zz} , does give the desired result:

$$g_{3zz-rr} = \frac{1}{2}(2g_{zz} - g_{xx} - g_{yy}) \quad (2.198)$$

$$g_{xx-yy} = \sqrt{\frac{3}{4}}(g_{xx} - g_{yy}) \quad (2.199)$$

The remaining sixth linear combination has the symmetry properties of an s function:

$$g_{rr} = \sqrt{5}(g_{xx} + g_{yy} + g_{zz}) \quad (2.200)$$

The advantages of Gaussian functions are countered by some serious shortcomings. This can be readily seen from a graphical comparison of the 1s Slater function and its 'best' Gaussian approximation, Figure 2.5. Unlike the Slater functions the Gaussian functions do not have a cusp at the origin and they also decay towards zero more quickly. It is found that replacing a Slater type orbital by a single Gaussian function leads to unacceptable errors. However, this problem can be overcome if each atomic orbital is represented as a linear combination of Gaussian functions. Each linear combination has the following form:

$$\phi_\mu = \sum_{i=1}^L d_{i\mu} \phi_i(\alpha_{i\mu}) \quad (2.201)$$

$d_{i\mu}$ is the coefficient of the primitive Gaussian function ϕ_i , which has exponent $\alpha_{i\mu}$. L is the number of functions in the expansion. For example, the linear combinations of Gaussian 1s functions that can be used to represent a 1s Slater type orbital with exponent $\xi = 1$ are given in Table 2.3.

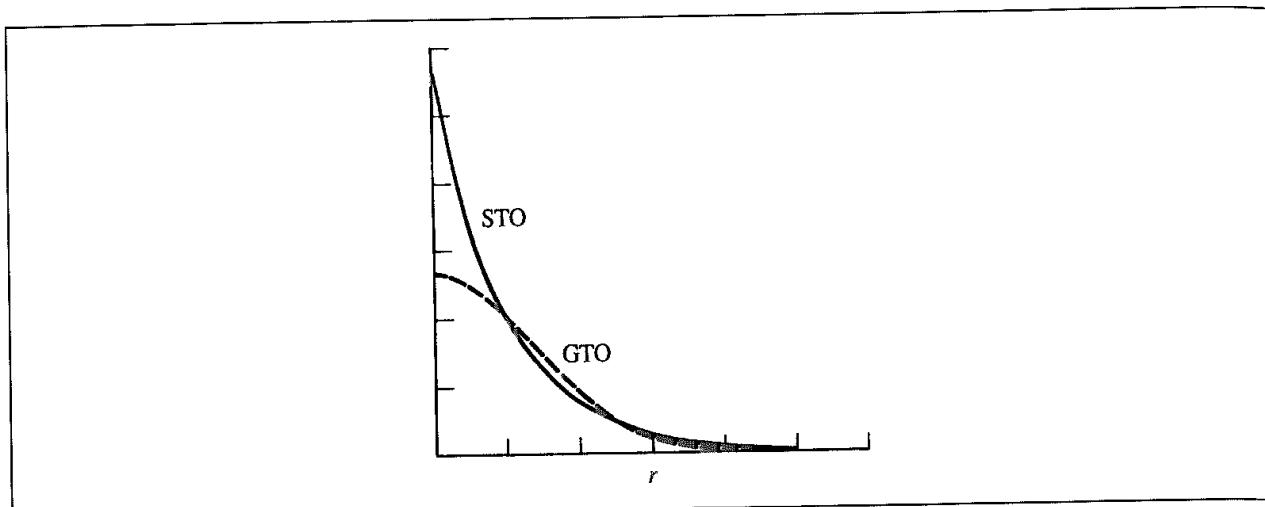


Fig. 2.5. The 1s Slater type orbital and the best Gaussian equivalent.

The coefficients and the exponents are found by least-squares fitting, in which the overlap between the Slater type function and the Gaussian expansion is maximised. Thus, for the 1s Slater type orbital we seek to maximise the following integral:

$$S = \frac{1}{\sqrt{\pi}} \left(\frac{2\alpha}{\pi} \right)^{3/4} \int dr e^{-r} e^{-\alpha r^2} \quad (2.202)$$

A graphical comparison of the 1s Slater type orbital and the four Gaussian expansions in Table 2.3 is shown in Figure 2.6. It is clear that the fit improves as the number of Gaussian functions increases, but even so, the addition of many more Gaussian functions cannot properly describe the exponential tail in the 'true' function and the cusp at the nucleus. This means that Gaussian functions underestimate the long-range overlap between atoms and the charge and spin density at the nucleus.

A Gaussian expansion contains two parameters: the coefficient and the exponent. The most flexible way to use Gaussian functions in *ab initio* molecular orbital calculations permits both of these parameters to vary during the calculation. Such a calculation is said to use

Number of Gaussians	Exponent, α	Expansion coefficient, d
1	0.270 950	1.00
2	0.151 623 0.851 819	0.678 914 0.430 129
3	0.109 818 0.405 771 2.227 66	0.444 635 0.535 28 0.154 329
4	0.088 0187 0.265 204 0.954 620 5.216 86	0.291 626 0.532 846 0.260 141 0.056 7523

Table 2.3 Coefficients and exponents for best-fit Gaussian expansions for the 1s Slater type orbital [Hehre et al 1969]

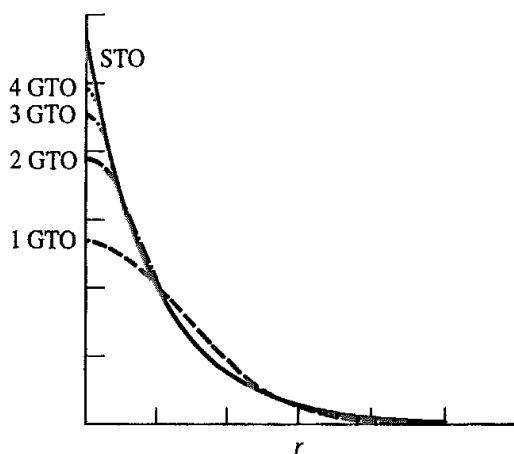


Fig 2.6: Comparison of 1s Slater type orbital and Gaussian expansions with up to four terms

uncontracted or *primitive* Gaussians. However, calculations with primitive Gaussians require a significant computational effort and so basis sets that consist of *contracted* Gaussian functions are most commonly employed. In a contracted function the contraction coefficients and exponents are pre-determined and remain constant during the calculation. The series of Gaussian functions in such cases is commonly referred to as a *contraction*, with the *contraction length* being the number of terms in the expansion. A further approximation that is often employed for the sake of computational efficiency is to use the same Gaussian exponents for the s and p orbitals in a given shell. This clearly restricts the flexibility of the basis set, but it does have the advantage of significantly reducing the number of numerically different integrals that need to be calculated.

Quantum chemists have devised efficient short-hand notation schemes to denote the basis set used in an *ab initio* calculation, although this does mean that a proliferation of abbreviations and acronyms are introduced. However, the codes are usually quite simple to understand. We shall concentrate on the notation used by Pople and co-workers in their Gaussian series of programs (see also the appendix to this chapter).

A *minimal basis set* is a representation that, strictly speaking, contains just the number of functions that are required to accommodate all the filled orbitals in each atom. In practice, a minimal basis set normally includes all of the atomic orbitals in the shell. Thus, for hydrogen and helium a single s-type function would be required; for the elements from lithium to neon the 1s, 2s and 2p functions are used, and so on. The basis sets STO-3G, STO-4G, etc. (in general, STO- n G) are all minimal basis sets in which n Gaussian functions are used to represent each orbital. It is found that at least three Gaussian functions are required to properly represent each Slater type orbital and so the STO-3G basis set is the 'absolute minimum' that should be used in an *ab initio* molecular orbital calculation. In fact, there is often little difference between the results obtained with the STO-3G basis set and the larger minimal basis sets with more Gaussian functions, although for hydrogen-bonded complexes STO-4G can perform significantly better. The STO-3G basis set does perform remarkably well in predicting molecular geometries, though this is due in part to

a fortuitous cancellation of errors. Of course, the computational effort increases with the number of functions in the Gaussian expansion.

The minimal basis sets are well known to have several deficiencies. There are particular problems with compounds containing atoms at the end of a period, such as oxygen or fluorine. Such atoms are described using the same number of basis functions as the atoms at the beginning of the period, despite the fact that they have more electrons. A minimal basis set only contains one contraction per atomic orbital and as the radial exponents are not allowed to vary during the calculation the functions cannot expand or contract in size in accordance with the molecular environment. The third drawback is that a minimal basis set cannot describe non-spherical aspects of the electronic distribution. For example, for a second-row element such as carbon the only functions that incorporate any anisotropy are the $2p_x$, $2p_y$ and $2p_z$ functions. As the radial components of these functions are required to be the same, no one component (x , y or z) can differ from another.

These problems with minimal basis sets can be addressed if more than one function is used for each orbital. A basis set which doubles the number of functions in the minimal basis set is described as a *double zeta* basis. Thus, a linear combination of a 'contracted' function and a 'diffuse' function gives an overall result that is intermediate between the two. The basis set coefficients of the contracted and the diffuse functions are automatically calculated by the SCF procedure, which thus automatically determines whether a more contracted or a more diffuse representation of that particular orbital is required. Such an approach can provide a solution to the anisotropy problem because it is then possible to have different linear combinations for the p_x , p_y and p_z orbitals.

An alternative to the double zeta basis approach is to double the number of functions used to describe the valence electrons but to keep a single function for the inner shells. The rationale for this approach is that the core orbitals, unlike the valence orbitals, do not affect chemical properties very much and vary only slightly from one molecule to another. The notation used for such *split valence* double zeta basis sets is exemplified by 3-21G. In this basis set three Gaussian functions are used to describe the core orbitals. The valence electrons are also represented by three Gaussians: the contracted part by two Gaussians and the diffuse part by one Gaussian. The most commonly used split valence basis sets are 3-21G, 4-31G and 6-31G.

Simply increasing the number of basis functions (triple zeta, quadruple zeta, etc.) does not necessarily improve the model. In fact, it can give rise to wholly erroneous results, particularly for molecules with a strongly anisotropic charge distribution. All of the basis sets we have encountered so far use functions that are centred on atomic nuclei. The use of split valence basis sets can help to surmount the problems with non-isotropic charge distribution but not completely. The charge distribution about an atom in a molecule is usually perturbed in comparison with the isolated atom. For example, the electron cloud in an isolated hydrogen atom is symmetrical, but when the hydrogen atom is present in a molecule the electrons are attracted towards the other nuclei. The distortion can be considered to correspond to mixing p-type character into the 1s orbital of the isolated atom to give a form of sp hybrid. In a similar manner, the unoccupied d orbitals introduce asymmetry into p orbitals (Figure 2.7). The most common solution to this problem is to introduce

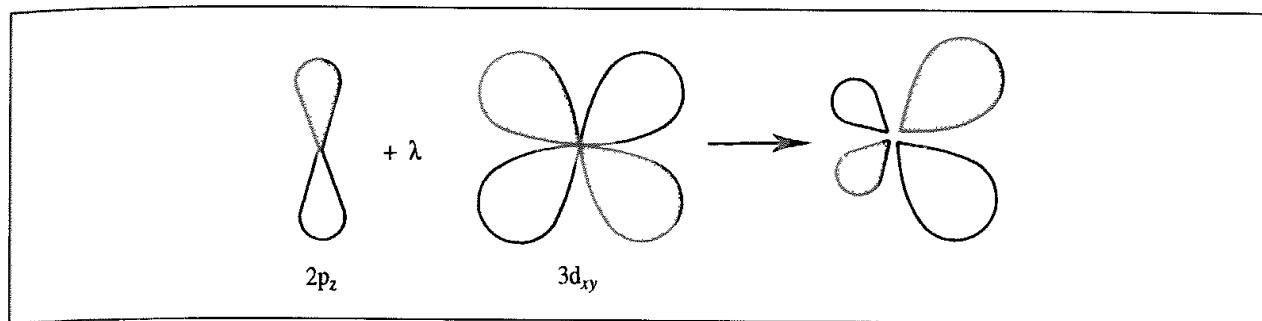


Fig. 2.7. The addition of a $3d_{xy}$ orbital to $2p_z$ gives a distorted orbital. (Figure adapted from Hehre W J, L Radom, P v R Schleyer and J A Hehre 1986 Ab initio Molecular Orbital Theory. New York, Wiley)

polarisation functions into the basis set. The polarisation functions have a higher angular quantum number and so correspond to p orbitals for hydrogen and d orbitals for the first- and second-row elements.

The use of polarisation basis functions is indicated by an asterisk (*). Thus, 6-31G* refers to a 6-31G basis set with polarisation functions on the heavy (i.e. non-hydrogen) atoms. Two asterisks (e.g. 6-31G**) indicate the use of polarisation (i.e. p) functions on hydrogen and helium. The 6-31G** basis set is particularly useful where hydrogen acts as a bridging atom. Partial polarisation basis sets have also been developed. For example, the 3-21G⁽⁺⁾ basis set has the same set of Gaussians as the 3-21G basis set (i.e. three functions for the inner shell, two contracted functions and one diffuse function for the valence shell) supplemented by six d-type Gaussians for the second-row elements. This basis set therefore attempts to account for d-orbital effects in molecules containing second-row elements. There are no special polarisation functions on first-row elements, which are described by the 3-21G basis set.

A deficiency of the basis sets described so far is their inability to deal with species such as anions and molecules containing lone pairs which have a significant amount of electron density away from the nuclear centres. This failure arises because the amplitudes of the Gaussian basis functions are rather low far from the nuclei. To remedy this deficiency highly diffuse functions can be added to the basis set. These basis sets are denoted using a '+'; thus the 3-21+G basis set contains an additional single set of diffuse s- and p-type Gaussian functions. '++' indicates that the diffuse functions are included for hydrogen as well as for heavy atoms. At these levels the terminology starts to become a little unwieldy. For example, the 6-311++G(3df, 3pd) basis set uses a single zeta core and triple zeta valence representation with additional diffuse functions on all atoms. The '(3df, 3pd)' indicates three sets of d functions and one set of f functions for first-row atoms and three sets of p functions and one set of d functions for hydrogen. This latter convention is probably the most generic; one commonly encountered example is the 6-31G(d) basis set, which is synonymous with 6-31G*.

The basis sets that we have considered thus far are sufficient for most calculations. However, for some high-level calculations a basis set that effectively enables the basis set limit to be achieved is required. The *even-tempered* basis set is designed to achieve this; each function in this basis set is the product of a spherical harmonic and a Gaussian function multiplied

by a power of the distance from the origin:

$$\chi_{klm}(\rho, \theta, \phi) = \exp(-\zeta_k^2 r^l) Y_{lm}(\theta, \phi) \quad (2.203)$$

The orbital exponent ζ_k is expressed as a function of two parameters α and β as follows.

$$\zeta_k = \alpha \beta^k \quad k = 1, 2, 3, \dots, N \quad (2.204)$$

The even-tempered basis set consists of the following sequence of functions: 1s, 2p, 3d, 4f, ..., which correspond to increasing values of k . The advantage of this basis set is that it is relatively easy to optimise the exponents for a large sequence of basis functions.

2.6.1 Creating a Basis Set

There is no definitive method for generating basis sets, and the construction of a new basis set is very much an art. Nevertheless, there are a number of well-established approaches that have resulted in widely used basis sets. We have already seen how linear combinations of Gaussian functions can be fitted to Slater type orbitals by minimising the overlap (see Figure 2.6 and Table 2.3). The Gaussian exponents and coefficients are derived by least-squares fitting to the desired functions, such as Slater type orbitals. When using basis sets that have been fitted to Slater orbitals it is often advantageous to use Slater exponents that are different to those obtained from Slater's rules. In general, better results for molecular calculations are obtained if larger Slater exponents are used for the valence electrons; this has the effect of giving a 'smaller', less diffuse orbital. For example, a value of 1.24 is widely used for the Slater exponent of hydrogen rather than the 1.0 that would be suggested by Slater's rules. It is straightforward to derive a basis set for a different Slater exponent if the Gaussian expansion has been fitted to a Slater type orbital with $\zeta = 1.0$. If the Slater exponent ζ is replaced by a new value, ζ' , then the respective Gaussian exponents α and α' are related by:

$$\frac{\alpha'}{\alpha} = \frac{\zeta'^2}{\zeta^2} \quad (2.205)$$

A doubling of the Slater exponent thus corresponds to a quadrupling of the Gaussian exponent. The expansion coefficients remain the same. For example, to obtain the exponents of the Gaussian functions for hydrogen in the STO-3G basis set we need to multiply the appropriate values in Table 2.3 by 1.24^2 , giving exponents of 0.168 856, 0.623 913 and 3.425 25. This strategy can be quite powerful; the STO- n G basis sets were originally defined with exponents that reproduce 'best atom' values for the core orbitals, but the exponents for the valence electrons were values that give optimal performance for a selected set of small molecules. For example, the suggested exponent for the valence orbitals in carbon was 1.72 rather than the 1.625 predicted by Slater's rules. The core orbitals have a Slater exponent of 5.67.

Basis sets can be constructed using an optimisation procedure in which the coefficients and the exponents are varied to give the lowest atomic energies. Some complications can arise when this approach is applied to larger basis sets. For example, in an atomic calculation the diffuse functions can move towards the nucleus, especially if the core region is described

by only a few basis functions. This is contrary to the role of diffuse functions, which is to enhance the description in the internuclear region. It may therefore be necessary to construct the basis set in stages, first determining the diffuse functions, using many basis functions for the core, and then optimising the basis functions for the core region, keeping the diffuse functions fixed. In many of the popular Gaussian basis sets the coefficients and exponents of the core orbitals are designed to reproduce calculations on atoms, whereas the valence basis functions are parametrised to reproduce the properties of a carefully selected set of molecular data.

The basis sets of Dunning [Dunning 1970] are obtained in a rather different way to those of Pople and co-workers. The first step is to perform an atomic SCF calculation using a set of primitive Gaussian functions in which the exponents are optimised to give the lowest energy for the atom. This set of primitive Gaussian functions (usually far too many for general use in molecular calculations) is then contracted to a smaller number of Gaussian functions, so drastically reducing the number of integrals that need be calculated. For example, Huzinga optimised the exponents of an uncontracted basis set that contained nine functions of s symmetry and five functions of p symmetry for the first-row elements [Huzinga 1965]. This (9s5p) basis set represents the 1s, 2s and three 2p orbitals and in fact corresponds to 24 basis functions per atom ($9 + 3 \times 5$). The primitive Gaussians in this uncontracted basis set are then apportioned to the basis functions in the new, contracted basis set, which contains three s functions and two p functions and is written [3s2p]. No primitive is assigned to more than one of the contracted basis functions. The 1s orbital is constructed from six primitives, the 2s orbital from one set of two primitives and one set containing just one primitive, and the 2p orbitals are represented by one contracted function containing five primitives and one contracted function that contains the remaining primitive. The final basis set, which is illustrated in Table 2.4 for nitrogen, contains a total of nine basis functions rather than the original 24. Each of the primitive functions appears

Exponent 1s	Coefficient	Exponent 2s	Coefficient	Exponent 2s	Coefficient
5900	0.001 190	7.193	-0.160 405	0.2133	1 000 000
887.5	0.009 099	1.707	1 058 215		
204.7	0 044 145				
59.84	0 150 464				
20.00	0.356 741				
7.193	0.446 533				
2.686	0 145 603				
2p		2p			
26.79	0 018 254	0.1654	1.000 000		
5.956	0.116 461				
1.707	0 390 178				
0.5314	0 637 102				

Table 2.4 Exponents and contraction coefficients for the three s-type and the two p-type Gaussian functions in the basis set of Dunning for nitrogen [Dunning 1970].

in just one basis function with its original exponent. The ratios of the coefficients of the primitives in the contracted basis set are equal to the ratios of the coefficients determined in the atomic SCF calculation. The major advantage of this approach is that calculations with the smaller basis set give results that are almost as good as calculations using the full basis set but with much less computational effort.

2.7 Calculating Molecular Properties Using *ab initio* Quantum Mechanics

We have now considered the key features of the *ab initio* approach to quantum mechanical calculations and so, as an antidote to the rather theoretical nature of the chapter so far, it is appropriate to consider how the method might be used in practice. Quantum mechanics can be used to calculate a wide range of properties. In addition to thermodynamic and structural values, quantum mechanics can be used to derive properties dependent upon the electronic distribution. Such properties often cannot be determined by any other method. In this section we shall provide a flavour of the ways in which quantum mechanics is used in molecular modelling. Other applications, such as the location of transition structures and the use of quantum mechanics in deriving force field parameters, will be discussed in later chapters. Many different computer programs are now available for performing *ab initio* calculations; probably the best known of these is the Gaussian series of programs which originated in the laboratory of John Pople, who has made numerous contributions to the field, recognised by the award of the Nobel Prize in 1998.

2.7.1 Setting Up the Calculation and the Choice of Coordinates

The traditional way to provide the nuclear coordinates to a quantum mechanical program is via a Z-matrix, in which the positions of the nuclei are defined in terms of a set of internal coordinates (see Section 1.2). Some programs also accept coordinates in Cartesian format, which can be more convenient for large systems. It can sometimes be important to choose an appropriate set of internal coordinates, especially when locating minima or transition points or when following reaction pathways. This is discussed in more detail in Section 5.7.

2.7.2 Energies, Koopman's Theorem and Ionisation Potentials

The energy of an electron in an orbital (Equation (2.169)) is often equated with the energy required to remove the electron to give the corresponding ion. This is *Koopman's theorem*. Two important caveats must be remembered when applying Koopman's theorem and comparing the results with experimentally determined ionisation potentials. The first of these is that the orbitals in the ionised state are assumed to be the same as in the unionised state; they are 'frozen'. This neglects the fact that the orbitals in the ionised state will be different from those in the unionised state. The energy of the ionised state will thus tend to be higher than it 'should' be, giving too large an ionisation potential. The second caveat is that the Hartree-Fock method does not include the effects of electron correlation.

The correction due to electron correlation would be expected to be greater for the unionised state than for the ionised state, as the former has more electrons. Fortunately, therefore, the effect of electron correlation often opposes the effect of the frozen orbitals, resulting in many cases in good agreement between experimentally determined ionisation potentials and calculated values.

A Hartree–Fock SCF calculation with K basis functions provides K molecular orbitals, but many of these will not be occupied by any electrons; they are the ‘virtual’ spin orbitals. If we were to add an electron to one of these virtual orbitals then this should provide a means of calculating the electron affinity of the system. Electron affinities predicted by Koopman’s theorem are always positive when Hartree–Fock calculations are used, because the virtual orbitals always have a positive energy. However, it is observed experimentally that many neutral molecules will accept an electron to form a stable anion and so have negative electron affinities. This can be understood if one realises that electron correlation would be expected to add to the error due to the ‘frozen’ orbital approximation, rather than to counteract it as for ionisation potentials.

2.7.3 Calculation of Electric Multipoles

Some of the most important properties that a quantum mechanical calculation provides are the electric multipole moments of the molecule. The electric multipoles reflect the distribution of charge in a molecule. The simplest electric moment (apart from the total net charge on the molecule) is the dipole. The dipole moment of a distribution of charges q_i located at positions \mathbf{r}_i is given by $\sum q_i \mathbf{r}_i$. If there are just two charges $+q$ and $-q$ separated by a distance r then the dipole moment is qr . A dipole moment of 4.8 Debye corresponds to two charges equal in magnitude to the electronic charge e separated by 1 Å. The dipole moment is a vector quantity, with components along the three Cartesian axes. The dipole moment of a molecule has contributions from both the nuclei and the electrons. The nuclear contributions can be calculated using the formula for a system of discrete charges:

$$\mu_{\text{nuclear}} = \sum_{A=1}^M Z_A \mathbf{R}_A \quad (2.206)$$

The electronic contribution arises from a continuous function of electron density and must be calculated using the appropriate operator:

$$\mu_{\text{electronic}} = \int d\tau \Psi_0 \left(\sum_{i=1}^N -\mathbf{r}_i \right) \Psi_0 \quad (2.207)$$

The dipole moment operator is a sum of one-electron operators \mathbf{r}_i , and as such the electronic contribution to the dipole moment can be written as a sum of one-electron contributions. The electronic contribution can also be written in terms of the density matrix, P , as follows:

$$\mu_{\text{electronic}} = \sum_{\mu=1}^K \sum_{\nu=1}^K P_{\mu\nu} \int d\tau \phi_\mu(-\mathbf{r}) \phi_\nu \quad (2.208)$$

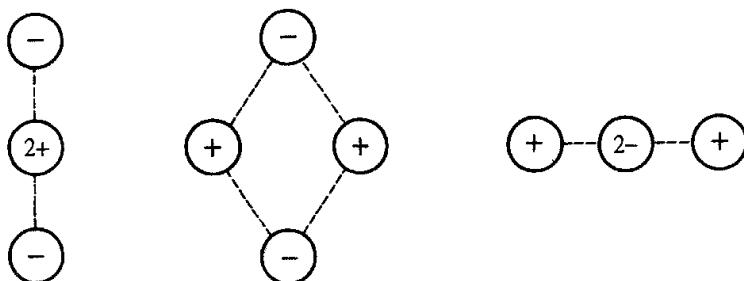


Fig. 2.8. A quadrupole moment can be obtained from various arrangements of two positive and two negative charges

The electronic contribution to the dipole moment is thus determined from the density matrix and a series of one-electron integrals $\int d\tau \phi_\mu(-\mathbf{r})\phi_\nu$. The dipole moment operator, \mathbf{r} , has components in the x , y and z directions, and so these one-electron integrals are divided into their appropriate components; for example, the x component of the electronic contribution to the dipole moment would be determined using:

$$\mu_x = \sum_{\mu=1}^K \sum_{\nu=1}^K P_{\mu\nu} \int d\tau \phi_\mu(-x)\phi_\nu \quad (2.209)$$

The quadrupole is the next electric moment. A molecule has a non-zero electric quadrupole moment when there is a non-spherically symmetrical distribution of charge. A quadrupole can be considered to arise from four charges that sum to zero which are arranged so that they do not lead to a net dipole. Three such arrangements are shown in Figure 2.8. Whereas the dipole moment has components in the x , y and z directions, the quadrupole has nine components from all pairwise combinations of x and y and is represented by a 3×3 matrix as follows:

$$\Theta = \begin{pmatrix} \sum q_i x_i^2 & \sum q_i x_i y_i & \sum q_i x_i z_i \\ \sum q_i y_i x_i & \sum q_i y_i^2 & \sum q_i y_i z_i \\ \sum q_i z_i x_i & \sum q_i z_i y_i & \sum q_i z_i^2 \end{pmatrix} \quad (2.210)$$

The three moments higher than the quadrupole are the octopole, hexapole and decapole. Methane is an example of a molecule whose lowest non-zero multipole moment is the octopole. The entire set of electric moments is required to completely and exactly describe the distribution of charge in a molecule. However, the series expansion is often truncated after the dipole or quadrupole as these are often the most significant.

Extensive comparisons have been made of experimental and calculated dipole moments (and in some cases the higher moments, though these are difficult to determine accurately by experiment). Factors such as the basis set and electron correlation can have a significant impact on the accuracy of the results, but it is found in many cases that the errors are systematic and that a simple scaling factor can be used to convert the results of a calculation with a small basis set to those obtained from experiment or with a much larger basis set. To illustrate how calculated dipole moments can vary, Table 2.5 provides the dipole moments for formaldehyde calculated at the experimental geometry using a variety of basis sets. It is

STO-3G	1.5258	3-21G	2.2903	4-31G	3.0041
6-31G*	2.7600	6-31G**	2.7576	6-311G**	2.7807
Expt.	2.34				

Table 2.5 Dipole moments calculated for formaldehyde using various basis sets at the experimental geometry.

also important to note that the dipole moment can be very sensitive to the geometry from which it is calculated.

2.7.4 The Total Electron Density Distribution and Molecular Orbitals

The electron density $\rho(\mathbf{r})$ at a point \mathbf{r} can be calculated from the Born interpretation of the wavefunction as a sum of squares of the spin orbitals at the point \mathbf{r} for all occupied molecular orbitals. For a system of N electrons occupying $N/2$ real orbitals, we can write:

$$\rho(\mathbf{r}) = 2 \sum_{i=1}^{N/2} |\psi_i(\mathbf{r})|^2 \quad (2.211)$$

If we express the molecular orbital ψ_i as a linear combination of basis functions, then the electron density at a point \mathbf{r} is given as:

$$\begin{aligned} \rho(\mathbf{r}) &= 2 \sum_{i=1}^{N/2} \left(\sum_{\mu=1}^K c_{\mu i} \phi_\mu(\mathbf{r}) \right) \left(\sum_{\nu=1}^K c_{\nu i} \phi_\nu(\mathbf{r}) \right) \\ &= 2 \sum_{i=1}^{N/2} \sum_{\mu=1}^K c_{\mu i} c_{\nu i} \phi_\mu(\mathbf{r}) \phi_\nu(\mathbf{r}) + 2 \sum_{i=1}^{N/2} \sum_{\mu=1}^K \sum_{\nu=\mu+1}^K 2c_{\mu i} c_{\nu i} \phi_\mu(\mathbf{r}) \phi_\nu(\mathbf{r}) \end{aligned} \quad (2.212)$$

Equation (2.212) can be tidied up considerably if it is written in terms of the elements of the density matrix:

$$\begin{aligned} \left(P_{\mu\nu} = 2 \sum_{i=1}^{N/2} c_{\mu i} c_{\nu i} \right) \\ \rho(\mathbf{r}) &= \sum_{\mu=1}^K \sum_{\nu=1}^K P_{\mu\nu} \phi_\mu(\mathbf{r}) \phi_\nu(\mathbf{r}) \\ &= \sum_{\mu=1}^K P_{\mu\mu} \phi_\mu(\mathbf{r}) \phi_\mu(\mathbf{r}) + 2 \sum_{\mu=1}^K \sum_{\nu=\mu+1}^K P_{\mu\nu} \phi_\mu(\mathbf{r}) \phi_\nu(\mathbf{r}) \end{aligned} \quad (2.213)$$

The integral of $\rho(\mathbf{r})$ over all space equals the number of electrons in the system, N :

$$N = \int d\mathbf{r} \rho(\mathbf{r}) = 2 \sum_{i=1}^{N/2} \int d\mathbf{r} |\psi_i(\mathbf{r})|^2 \quad (2.214)$$

If the overlap between two orbitals ϕ_μ and ϕ_ν is written as $S_{\mu\nu}$, and if the basis functions are

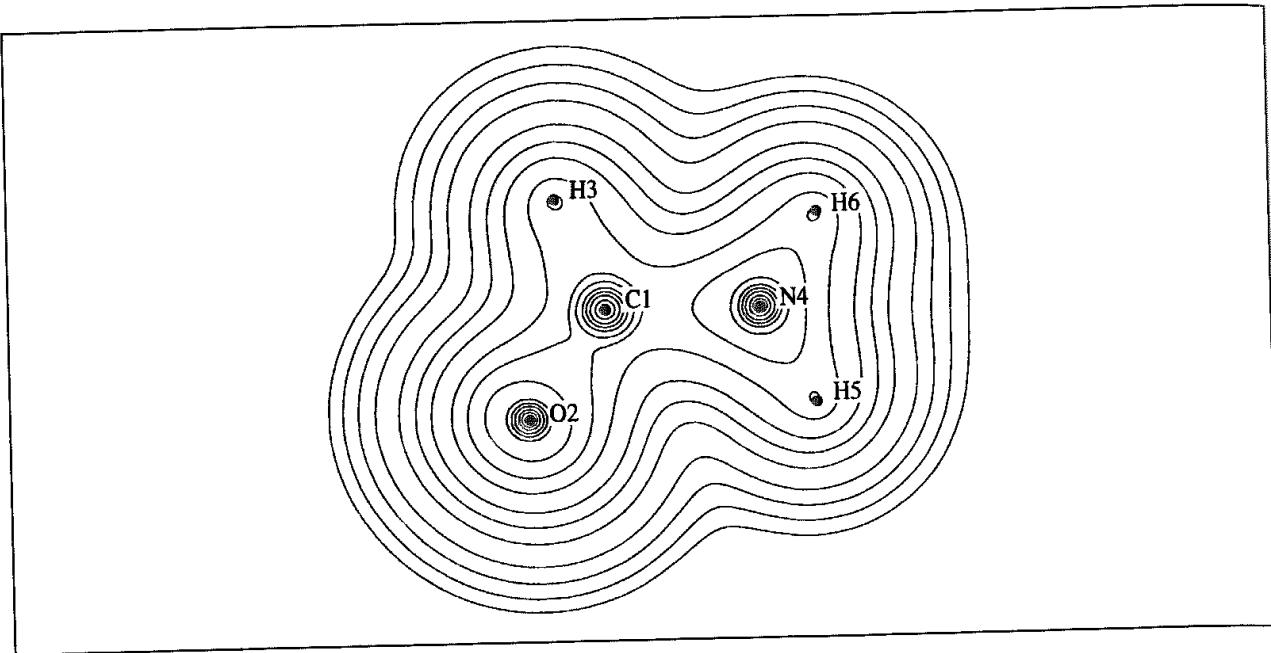


Fig. 2.9: Contour map showing the variation in electron density around formamide

assumed to be normalised ($S_{\mu\mu} = 1$), then:

$$N = \sum_{\mu=1}^K P_{\mu\mu} + 2 \sum_{\mu=1}^K \sum_{\nu=\mu+1}^K P_{\mu\nu} S_{\mu\nu} \quad (2.215)$$

The electron density can be visualised in several ways. One approach is to construct contours on slices through the molecule, such that each contour connects points of equal density, as shown in Figure 2.9 for formamide. The electron density can also be represented as an isometric projection (or a 'relief map', Figure 2.10), in which the height above the plane

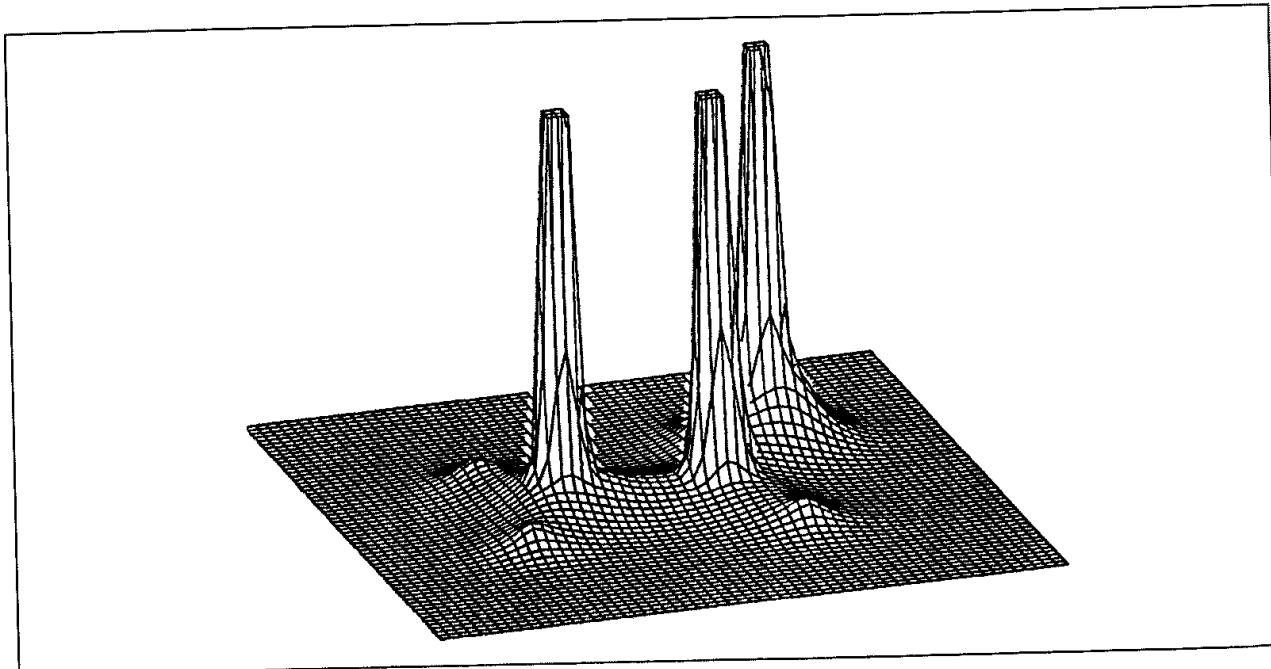


Fig. 2.10: Isometric projection of the electron density around formamide

represents the magnitude of the electron density. These diagrams show that the electron density tends to be greatest near the nuclei, as would be expected. The electron density can also be represented as a solid object, whose surface connects points of equal density. The surface shown in Figure 2.11 (colour plate section) corresponds to an electron density of 0.0001 a.u. around formamide. Other properties such as the electrostatic potential can be mapped onto this surface, as we shall see in Section 2.7.9.

The electron density distribution of individual molecular orbitals may also be determined and plotted. The highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO) are often of particular interest as these are the orbitals most commonly involved in chemical reactions. As an illustration, the HOMO and LUMO for formamide are displayed in Figures 2.12 and 2.13 (colour plate section) as surface pictures.

2.7.5 Population Analysis

Population analysis methods partition the electron density between the nuclei so that each nucleus has a ‘number’ (not necessarily an integral number) of electrons associated with it. Such a partitioning provides a way to calculate the atomic charge on each nucleus. It should be noted that there is no quantum mechanical operator for the atomic charge and so any partitioning scheme must be arbitrary. Hence many methods have been devised. Here we will consider Mulliken and Löwdin analysis and Bader’s theory of atoms in molecules. The alternatives include natural population analysis [Reed *et al.* 1985; Bachrach 1994]. Wiberg and Rablen have compared a number of methods for calculating atomic charges, and we refer to some of their results in the following discussion [Wiberg and Rablen 1993]. To illustrate the variation that can be obtained in the results, for methane they found that the charge on the carbon atom varied from -0.473 to $+0.244$, depending upon the method chosen! We will also consider the problem of calculating atomic charges in more detail in Chapter 4 on molecular mechanics.

2.7.6 Mulliken and Löwdin Population Analysis

RS Mulliken suggested a widely used method for performing population analysis [Mulliken 1955]. The starting point is Equation (2.215), which relates the total number of electrons to the density matrix and to the overlap integrals. In the Mulliken method, all of the electron density ($P_{\mu\mu}$) in an orbital is allocated to the atom on which ϕ_μ is located. The remaining electron density is associated with the overlap population, $\phi_\mu\phi_\nu$. For each element $\phi_\mu\phi_\nu$ of the density matrix, half of the density is assigned to the atom on which ϕ_μ is located and half to the atom on which ϕ_ν is located. The net charge on an atom A is then calculated by subtracting the number of electrons from the nuclear charge, Z_A :

$$q_A = Z_A - \sum_{\mu=1; \mu \text{ on } A}^K P_{\mu\mu} - \sum_{\mu=1; \mu \text{ on } A}^K \sum_{\nu=1; \nu \neq \mu}^K P_{\mu\nu} S_{\mu\nu} \quad (2.216)$$

Mulliken population analysis is a trivial calculation to perform once a self-consistent field has been established and the elements of the density matrix have been determined

However, there are some serious shortcomings to the method, as Mulliken himself pointed out.

A Mulliken analysis depends upon the use of a balanced basis set, in which an equivalent number of basis functions is present on each atom in the molecule. For example, it is possible to calculate a wavefunction for a molecule such as water in which all of the basis functions reside on the oxygen atom; if a large enough basis set is used then a quite reasonable wavefunction for the whole molecule can be obtained. However, the Mulliken analysis would put all of the charge on the oxygen. This is an extreme example of a general problem; p, d and f orbitals are spread quite far from the nucleus with which they are associated and so may be very close to other atoms, yet the charge associated with electron occupation of such orbitals is assigned to the atom on which the orbital is centred. The equal apportioning of electrons between pairs of atoms, even if their electronegativities are very different, can lead in some cases to quite unrealistic values for the net atomic charge. *In extremis*, some orbitals may ‘contain’ a negative number of electrons and others more than two electrons, in clear contradiction of the Pauli principle. A Mulliken analysis assumes that each basis function can be associated with an atomic centre and so is not applicable if basis functions not centred on the nuclei are used. The atomic charges can be very dependent upon the basis set; for example, Wiberg and Rablen found that the charge on the central carbon in isobutene changed from +0.1 with a 6-31G* basis set to +1.0 for a 6-311++G** basis set.

In the Löwdin approach to population analysis [Löwdin 1970; Cusachs and Politzer 1968] the atomic orbitals are transformed to an orthogonal set, along with the molecular orbital coefficients. The transformed orbitals ϕ'_μ in the orthogonal set are given by:

$$\phi'_\mu = \sum_{\nu=1}^K (\mathbf{S}^{-1/2})_{\nu\mu} \phi_\nu \quad (2.217)$$

The electron population associated with an atom becomes:

$$q_A = Z_A - \sum_{\mu=1, \mu \text{ on } A}^K (\mathbf{S}^{1/2} \mathbf{P} \mathbf{S}^{1/2})_{\mu\mu} \quad (2.218)$$

Löwdin population analysis avoids the problem of negative populations or populations greater than 2. Some quantum chemists prefer the Löwdin approach to that of Mulliken as the charges are often closer to chemically intuitive values and are less sensitive to basis set.

2.7.7 Partitioning Electron Density: The Theory of Atoms in Molecules

R F W Bader’s theory of ‘atoms in molecules’ [Bader 1985] provides an alternative way to partition the electrons between the atoms in a molecule. Bader’s theory has been applied to many different problems, but for the purposes of our present discussion we will concentrate on its use in partitioning electron density. The Bader approach is based upon the concept of a *gradient vector path*, which is a curve around the molecule such that it is always perpendicular to the electron density contours. A set of gradient paths is drawn in Figure 2.14 for formamide. As can be seen, some of the gradient paths terminate at the atomic nuclei. Other gradient paths are attracted to points (called critical points) that are

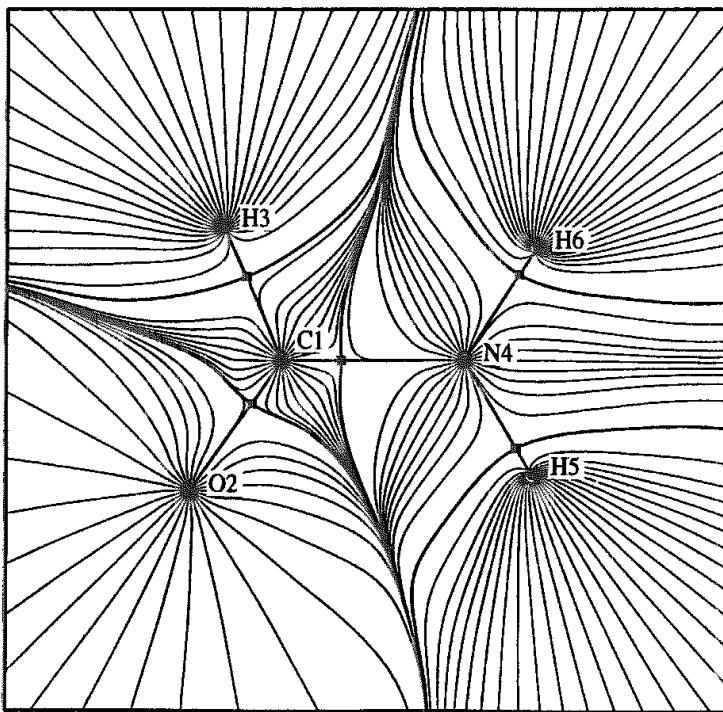


Fig 2.14: Gradient vector paths around formamide. The paths terminate at atoms or at bond critical points (indicated by squares)

not located at the nuclei; particularly common are the bond critical points, which are located between bonded atoms. Other types of critical point can occur; for example, a *ring critical point* is found in the centre of a benzene ring.

The bond critical points are points of minimum electron charge density between two bonded atoms. If we follow the contour in three-dimensional space from such a point down the gradient path along which the density decreases most rapidly then this gives a means of partitioning the density. This is shown in Figure 2.15 for hydrogen fluoride and in Figure 2.16 for formamide. This procedure can be performed for each bond, resulting in a three-dimensional partitioning of the electron density. The electron population that is assigned to each atom is then calculated by numerically integrating the charge density within the region surrounding that atom.

Wiberg and Rablen found that the charges obtained with the atoms in molecules method were relatively invariant to the basis set. The charges from this method were also consistent with the experimentally determined C-H bond dipoles in methane (in which the carbon is positive) and ethyne (in which the carbon is negative), unlike most of the other methods they examined.

2.7.8 Bond Orders

As with atomic charges, the bond order is not a quantum mechanical observable and so various methods have been proposed for calculating the bond orders in a molecule.

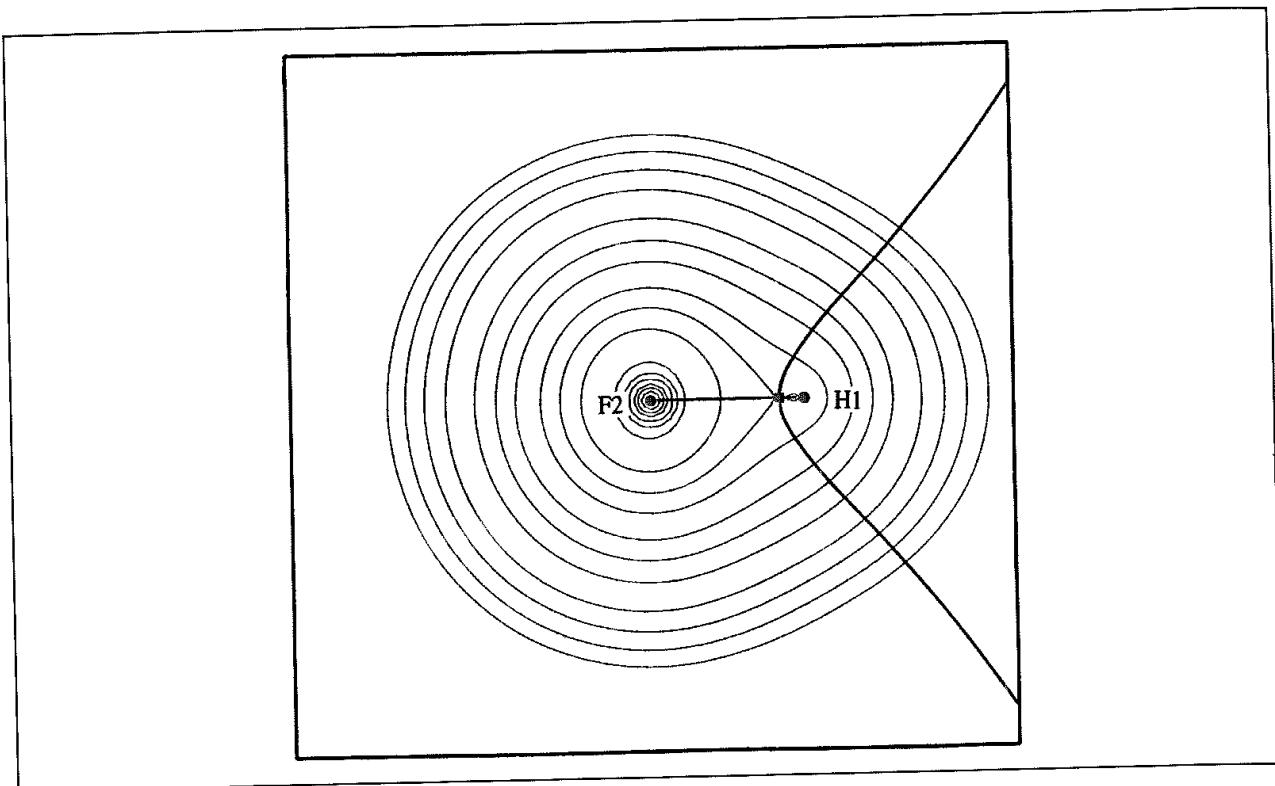


Fig. 2 15. Partitioning the electron density in hydrogen fluoride

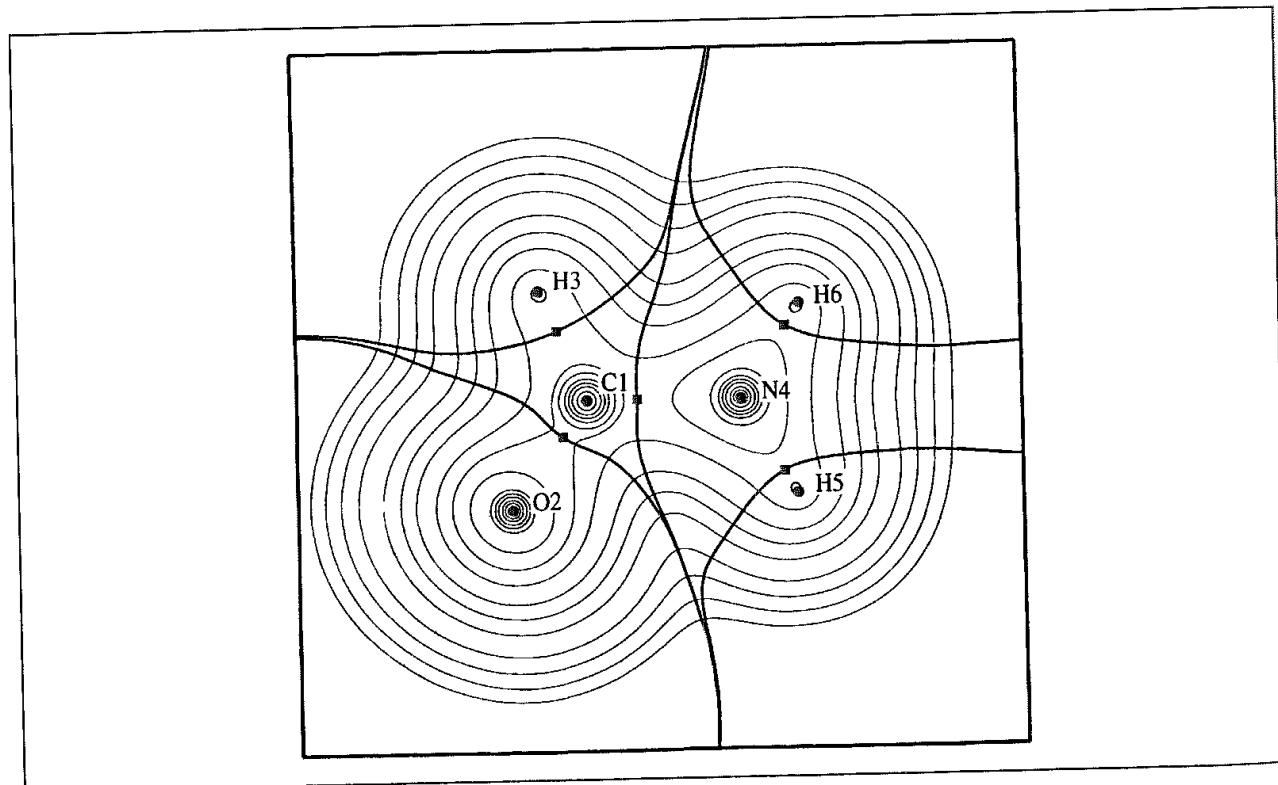


Fig. 2 16. Partitioning the electron density in formamide

Molecule	Bond	STO-3G	4-31G
H ₂	H–H	1.0	1.0
Methane	C–H	0.99	0.96
Ethene	C=C	2.01	1.96
	C–H	0.98	0.96
Ethyne	C≡C	3.00	3.27
	C–H	0.98	0.86
Water	O–H	0.95	0.80
N ₂	N≡N	3.0	2.67

Table 2.6 Bond order obtained from the Mayer bond order scheme [Mayer 1983]

Mayer defined the bond order between two atoms as follows [Mayer 1983]:

$$B_{AB} = \sum_{\mu \text{ on } A} \sum_{\nu \text{ on } B} [(\mathbf{PS})_{\mu\nu} (\mathbf{PS})_{\nu\mu} + (\mathbf{P}^s \mathbf{S})_{\mu\nu} (\mathbf{P}^s \mathbf{S})_{\nu\mu}] \quad (2.219)$$

\mathbf{P} is the total spinless density matrix ($\mathbf{P} = \mathbf{P}^\alpha + \mathbf{P}^\beta$) and \mathbf{P}^s is the spin density matrix ($\mathbf{P}^s = \mathbf{P}^\alpha - \mathbf{P}^\beta$). For a closed-shell system Mayer's definition of the bond order reduces to:

$$B_{AB} = \sum_{\mu \text{ on } A} \sum_{\nu \text{ on } B} (\mathbf{PS})_{\mu\nu} (\mathbf{PS})_{\nu\mu} \quad (2.220)$$

The bond orders obtained from Mayer's formula often seem intuitively reasonable, as illustrated in Table 2.6 for some simple molecules. The method has also been used to compute the bond orders for intermediate structures in reactions of the form $\text{H} + \text{XH} \rightarrow \text{HX} + \text{H}$ and $\text{X} + \text{H}_2 \rightarrow \text{XH} + \text{H}$ ($\text{X} = \text{F}, \text{Cl}, \text{Br}$). The results suggested that bond orders were a useful way to describe the similarity of the transition structure to the reactants or to the products. Moreover, the bond orders were approximately conserved along the reaction pathway.

As with methods for allocating electron density to atoms, the Mayer method is not necessarily 'correct', though it appears to be a useful measure of the bond order that conforms to accepted pictures of bonding in molecules.

2.7.9 Electrostatic Potentials

The electrostatic potential at a point \mathbf{r} , $\phi(\mathbf{r})$, is defined as the work done to bring unit positive charge from infinity to the point. The electrostatic interaction energy between a point charge q located at \mathbf{r} and the molecule equals $q\phi(\mathbf{r})$. The electrostatic potential has contributions from both the nuclei and from the electrons, unlike the electron density, which only reflects the electronic distribution. The electrostatic potential due to the M nuclei is:

$$\phi_{\text{nucl}}(\mathbf{r}) = \sum_{A=1}^M \frac{Z_A}{|\mathbf{r} - \mathbf{R}_A|} \quad (2.221)$$

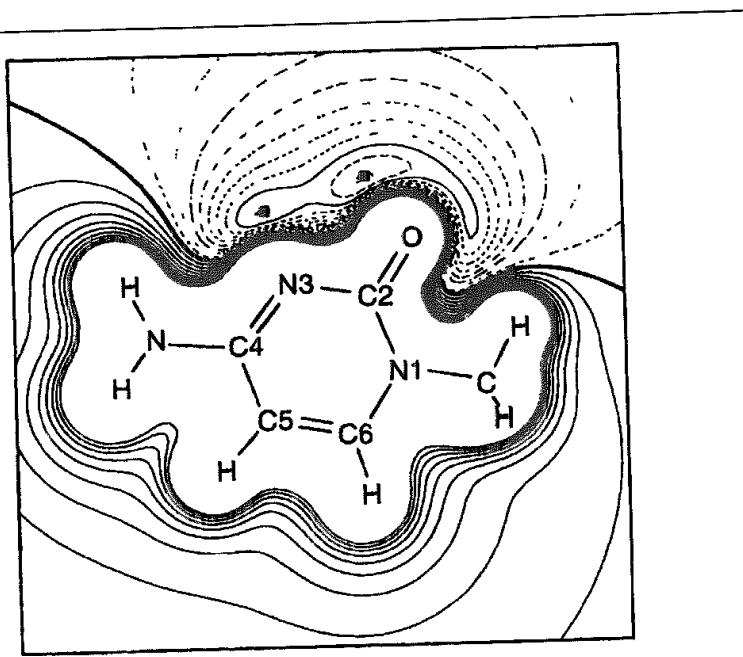


Fig 2.17: Electrostatic potential contours around cytosine. Negative contours are dashed, the zero contour is bold. The minima near N3 and O are marked

The potential due to the electrons is obtained from the appropriate integral of the electron density:

$$\phi_{\text{elec}}(\mathbf{r}) = - \int \frac{d\mathbf{r}' \rho(\mathbf{r}')}{|\mathbf{r}' - \mathbf{r}|} \quad (2.222)$$

The total electrostatic potential equals the sum of the nuclear and the electronic contributions:

$$\phi(\mathbf{r}) = \phi_{\text{nuc}}(\mathbf{r}) + \phi_{\text{elec}}(\mathbf{r}) \quad (2.223)$$

The electrostatic potential has proved to be particularly useful for rationalising the interactions between molecules and molecular recognition processes. This is because electrostatic forces are primarily responsible for long-range interactions between molecules. The electrostatic potential varies through space, and so it can be calculated and visualised in the same way as the electron density. Electrostatic potential contours can be used to propose where electrophilic attack might occur; electrophiles are often attracted to regions where the electrostatic potential is most negative. For example, the experimentally determined position of electrophilic attack at the nucleic acid cytosine is at N3 (Figure 2.17). This atom is next to a minimum in the electrostatic potential (also shown in Figure 2.17), as pointed out by Politzer and Murray [Politzer and Murray 1991].

Non-covalent interactions between molecules often occur at separations where the van der Waals radii of the atoms are just touching and so it is often most useful to examine the electrostatic potential in this region. For this reason, the electrostatic potential is often calculated at the molecular surface (defined in Section 1.5) or the equivalent isodensity surface as shown in Figure 2.18 (colour plate section). Such pictorial representations

can be used to qualitatively assess the degree of electrostatic similarity between two molecules.

2.7.10 Thermodynamic and Structural Properties

The total energy of a system is equal to the sum of the electronic energy and the Coulombic nuclear repulsion energy:

$$E_{\text{tot}} = E_{\text{elec}} + \sum_{A=1}^M \sum_{B=A+1}^M \frac{Z_A Z_B}{R_{AB}} \quad (2.224)$$

A more useful quantity for comparison with experiment is the heat of formation, which is defined as the enthalpy change when one mole of a compound is formed from its constituent elements in their standard states. The heat of formation can thus be calculated by subtracting the heats of atomisation of the elements and the atomic ionisation energies from the total energy. Unfortunately, *ab initio* calculations that do not include electron correlation (which we will discuss in Chapter 3) provide uniformly poor estimates of heats of formation with errors in bond dissociation energies of 25–40 kcal/mol, even at the Hartree–Fock limit for diatomic molecules.

When combined with an energy minimisation algorithm, quantum mechanics can be used to calculate equilibrium geometries of molecules. The results of such calculations can be compared with the structures obtained from gas-phase experiments using microwave spectroscopy, electronic spectroscopy and electron diffraction. Extensive tables listing comparisons between calculations and experiment for many molecules have been published in several reviews. Not surprisingly, the agreement between theory and experiment for *ab initio* calculations generally improves as one increases the size of the basis set. Hehre *et al.* suggest that the 3-21G basis set offers a good compromise between performance and applicability [Hehre *et al.* 1986]. It is often found that errors in structural predictions are systematic rather than random. For example, STO-3G bond lengths are generally too long, whilst 6-31G* bond lengths tend to be too short. By analysing the trends in such calculations it can be possible to derive scaling factors which enable more accurate predictions to be made for each level of theory.

Quantum mechanics can be used to calculate the relative energies of conformations and the energy barriers between them. Experimental data is available for both relative stabilities and barrier heights in some cases, though this tends to be limited to relatively simple molecules. Butane is one molecule that has been investigated in great detail, with its *gauche* and *anti* conformations and the barriers that separate them. The energy difference between the *syn* and *anti* conformations of butane (Figure 2.19) was found to fall significantly with increasing basis set size, particularly when correlated levels of theory were employed [Wiberg and Murcko 1988; Allinger *et al.* 1990; Smith and Jaffe 1996]. However, the smaller energy difference between the minimum energy *anti* and *gauche* conformations can be calculated quite accurately even with a relatively small basis set. Quantum mechanics calculations of the change in energy as a bond is rotated are often used to parametrise the torsional terms in molecular mechanics force fields, as will be discussed in Section 4.18.

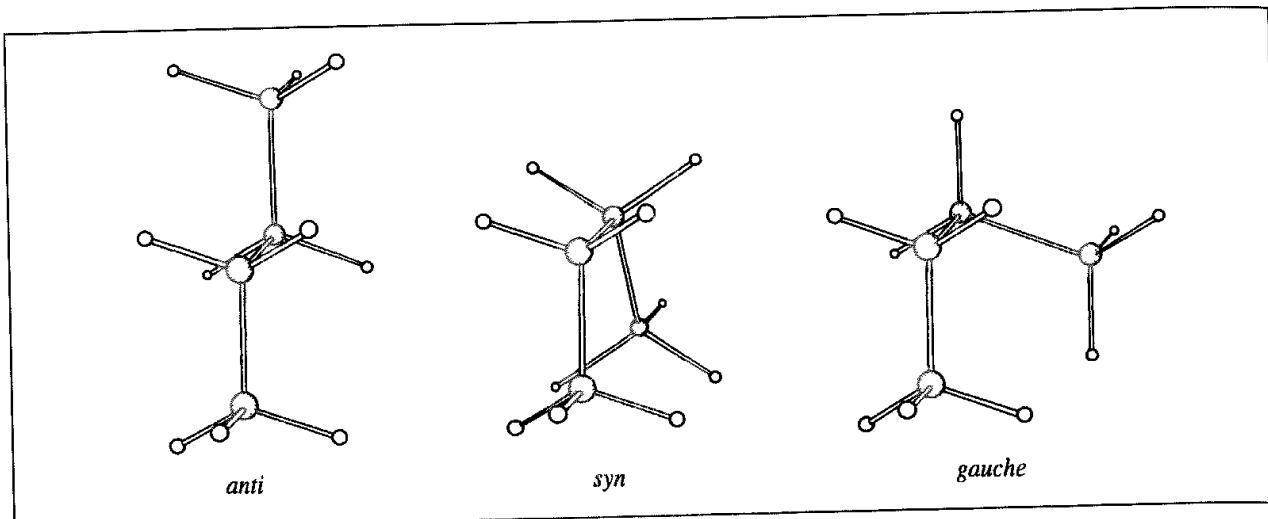


Fig. 2.19. syn, anti and gauche conformations of butane (C-C-C-C torsion angles 0° , 180° and $\pm 60^\circ$ respectively).

2.8 Approximate Molecular Orbital Theories

Ab initio calculations can be extremely expensive in terms of the computer resources required. Nevertheless, improvements in computer hardware and the availability of easy-to-use programs have helped to make *ab initio* methods a widely used computational tool. The approximate quantum mechanical methods require significantly less computational resources. Indeed, the earliest approximate methods such as Hückel theory predate computers by many years. Moreover, by their incorporation of parameters derived from experimental data some approximate methods can calculate certain properties more accurately than even the highest level of *ab initio* methods.

Many approximate molecular orbital theories have been devised. Most of these methods are not in widespread use today in their original form. Nevertheless, the more widely used methods of today are derived from earlier formalisms, which we will therefore consider where appropriate. We will concentrate on the semi-empirical methods developed in the research groups of Pople and Dewar. The former pioneered the CNDO, INDO and NDDO methods, which are now relatively little used in their original form but provided the basis for subsequent work by the Dewar group, whose research resulted in the popular MINDO/3, MNDO and AM1 methods. Our aim will be to show how the theory can be applied in a practical way, not only to highlight their successes but also to show where problems were encountered and how these problems were overcome. We will also consider the Hückel molecular orbital approach and the extended Hückel method. Our discussion of the underlying theoretical background of the approximate molecular orbital methods will be based on the Roothaan-Hall framework we have already developed. This will help us to establish the similarities and the differences with the *ab initio* approach.

2.9 Semi-empirical Methods

A discussion of semi-empirical methods starts most appropriately with the key components

of the Roothaan–Hall equations, which for a closed-shell system are:

$$\mathbf{FC} = \mathbf{SCE} \quad (2.225)$$

$$F_{\mu\nu} = H_{\mu\nu}^{\text{core}} + \sum_{\lambda=1}^K \sum_{\sigma=1}^K P_{\lambda\sigma} [(\mu\nu|\lambda\sigma) - \frac{1}{2}(\mu\lambda|\nu\sigma)] \quad (2.226)$$

$$P_{\lambda\sigma} = 2 \sum_{i=1}^{N/2} c_{\lambda i} c_{\sigma i} \quad (2.227)$$

$$H_{\mu\nu}^{\text{core}} = \int d\nu_1 \phi_\mu(1) \left[-\frac{1}{2} \nabla^2 - \sum_{A=1}^M \frac{Z_A}{|r_1 - R_A|} \right] \phi_\nu(1) \quad (2.228)$$

In *ab initio* calculations all elements of the Fock matrix are calculated using Equation (2.226), irrespective of whether the basis functions ϕ_μ , ϕ_ν , ϕ_λ and ϕ_σ are on the same atom, on atoms that are bonded or on atoms that are not formally bonded. To discuss the semi-empirical methods it is useful to consider the Fock matrix elements in three groups: $F_{\mu\mu}$ (the diagonal elements), $F_{\mu\nu}$ (where ϕ_μ and ϕ_ν are on the same atom) and $F_{\mu\nu}$ (where ϕ_μ and ϕ_ν are on different atoms).

We have mentioned several times that the greatest proportion of the time required to perform an *ab initio* Hartree–Fock SCF calculation is invariably spent calculating and manipulating integrals. The most obvious way to reduce the computational effort is therefore to neglect or approximate some of these integrals. Semi-empirical methods achieve this in part by explicitly considering only the valence electrons of the system; the core electrons are subsumed into the nuclear core. The rationale behind this approximation is that the electrons involved in chemical bonding and other phenomena that we might wish to investigate are those in the valence shell. By considering all the valence electrons the semi-empirical methods differ from those theories (e.g. Hückel theory) that explicitly consider only the π electrons of a conjugated system and which are therefore limited to specific classes of molecule. The semi-empirical calculations invariably use basis sets comprising Slater type s, p and sometimes d orbitals. The orthogonality of such orbitals enables further simplifications to be made to the equations.

A feature common to the semi-empirical methods is that the overlap matrix, \mathbf{S} (in Equation (2.225)), is set equal to the identity matrix \mathbf{I} . Thus all diagonal elements of the overlap matrix are equal to 1 and all off-diagonal elements are zero. Some of the off-diagonal elements would naturally be zero due to the use of orthogonal basis sets on each atom, but in addition the elements that correspond to the overlap between two atomic orbitals on different atoms are also set to zero. The main implication of this is that the Roothaan–Hall equations are simplified: $\mathbf{FC} = \mathbf{SCE}$ becomes $\mathbf{FC} = \mathbf{CE}$ and so is immediately in standard matrix form. It is important to note that setting \mathbf{S} equal to the identity matrix does not mean that all overlap integrals are set to zero in the calculation of Fock matrix elements. Indeed, it is important specifically to include some of the overlaps in even the simplest of the semi-empirical models.

2.9.1 Zero-differential Overlap

Many semi-empirical theories are based upon the zero-differential overlap approximation (ZDO). In this approximation, the overlap between pairs of different orbitals is set to zero for all volume elements $d\nu$:

$$\phi_\mu \phi_\nu d\nu = 0 \quad (2.229)$$

This directly leads to the following result for the overlap integrals:

$$S_{\mu\nu} = \delta_{\mu\nu} \quad (2.230)$$

If the two atomic orbitals ϕ_μ and ϕ_ν are located on different atoms then the differential overlap is referred to as diatomic differential overlap; if ϕ_μ and ϕ_ν are on the same atom then we have monatomic differential overlap. If the ZDO approximation is applied to the two-electron repulsion integral $(\mu\nu|\lambda\sigma)$ then the integral will equal zero if $\mu \neq \nu$ and/or if $\lambda \neq \sigma$. This can be written concisely using the Kronecker delta:

$$(\mu\nu|\lambda\sigma) = (\mu\mu|\lambda\lambda)\delta_{\mu\nu}\delta_{\lambda\sigma} \quad (2.231)$$

It can immediately be seen that all three- and four-centre integrals are set to zero under the ZDO approximation. If the ZDO approximation is applied to all orbital pairs then the Roothaan-Hall equations for a closed-shell molecule (Equation (2.226)) simplify considerably to give the following for $\mu \equiv \nu$:

$$F_{\mu\mu} = H_{\mu\mu}^{\text{core}} + \sum_{\lambda=1}^K P_{\lambda\lambda}(\mu\mu|\lambda\lambda) - \frac{1}{2}P_{\mu\mu}(\mu\mu|\mu\mu) \quad (2.232)$$

The summation over λ includes $\lambda = \mu$, and the terms in $(\mu\mu|\mu\mu)$ can be separated to give:

$$F_{\mu\mu} = H_{\mu\mu}^{\text{core}} + \frac{1}{2}P_{\mu\mu}(\mu\mu|\mu\mu) + \sum_{\lambda=1; \lambda \neq \mu}^K P_{\lambda\lambda}(\mu\mu|\lambda\lambda) \quad (2.233)$$

For $\nu \neq \mu$ we have,

$$F_{\mu\nu} = H_{\mu\nu}^{\text{core}} - \frac{1}{2}P_{\mu\nu}(\mu\mu|\nu\nu) \quad (2.234)$$

Sensible results cannot be obtained by simply applying the ZDO approximation to all pairs of orbitals *carte blanche*. There are two major reasons for this.

The first consideration is that the total wavefunction and the molecular properties calculated from it should be the same when a transformed basis set is used. We have already encountered this requirement in our discussion of the transformation of the Roothaan-Hall equations to an orthogonal set. To reiterate: suppose a molecular orbital is written as a linear combination of atomic orbitals:

$$\psi_i = \sum_{\mu} c_{\mu i} \phi_{\mu} \quad (2.235)$$

If an alternative basis set is used in which the basis functions are just linear combinations of the original basis functions, then the same wavefunction can be written as a linear

combination of these new transformed functions:

$$\psi_i = \sum_{\alpha} c_{\alpha i} \phi'_{\alpha} \quad (2.236)$$

$$\phi'_{\alpha} = \sum_{\mu_{\alpha}} t_{\mu_{\alpha}} \phi_{\mu} \quad (2.237)$$

$t_{\mu_{\alpha}}$ are the coefficients of the original basis functions in the linear expansion of the transformed basis set. Different types of transformation are possible; for example, some transformations mix orbitals with the same principal and azimuthal quantum numbers (e.g. mixing $2p_x$, $2p_y$ and $2p_z$); others mix orbitals with the same principal quantum number but different azimuthal quantum numbers (e.g. mixing $2s$, $2p_x$, $2p_y$ and $2p_z$ orbitals to give sp^3 hybrid orbitals); yet other transformations mix orbitals located on different atoms. Suppose we mix $2p_x$ and $2p_y$ atomic orbitals on the same atom. The differential overlap between these two orbitals is $2p_x 2p_y$. We now introduce the following two new coordinates, which correspond to a rotation in the xy plane:

$$x' = \frac{1}{\sqrt{2}}(x + y) \quad (2.238)$$

$$y' = \frac{1}{\sqrt{2}}(-x + y) \quad (2.239)$$

The overlap between the $2p'_x$ and $2p'_y$ orbitals in this new coordinate system is $\frac{1}{2}(2p_y^2 - 2p_x^2)$. If the zero differential overlap approximation were applied, then different results would be obtained for the two coordinate systems unless the overlap in the new, transformed system was also ignored.

The second reason why the ZDO approximation is not applied to all pairs of orbitals is that the major contributors to bond formation are the electron–core interactions between pairs of orbitals and the nuclear cores (i.e. $H_{\mu\nu}^{\text{core}}$). These interactions are therefore not subjected to the ZDO approximation (and so do not suffer from any transformation problems).

2.9.2 CNDO

The complete neglect of differential overlap (CNDO) approach of Pople, Santry and Segal was the first method to implement the zero-differential overlap approximation in a practical fashion [Pople *et al.* 1965]. To overcome the problems of rotational invariance, the two-electron integrals ($\mu\mu|\lambda\lambda$), where μ and λ are on different atoms A and B, were set equal to a parameter γ_{AB} which depends only on the nature of the atoms A and B and the internuclear distance, and not on the type of orbital. The parameter γ_{AB} can be considered to be the average electrostatic repulsion between an electron on atom A and an electron on atom B. When both atomic orbitals are on the same atom the parameter is written γ_{AA} and represents the average electron–electron repulsion between two electrons on an atom A.

With this approximation we can divide the elements of the Fock matrix into three groups: $F_{\mu\mu}$ (the diagonal elements), $F_{\mu\nu}$ (where μ and ν are on different atoms) and $F_{\mu\nu}$ (where μ

and ν are on the same atom). To obtain $F_{\mu\mu}$ we substitute γ_{AB} for the two-electron integrals $(\mu\mu|\lambda\lambda)$ where μ and λ are on different atoms and γ_{AA} where μ and λ are on the same atom into the Fock matrix equations, Equations (2.240)–(2.242):

$$F_{\mu\mu} = H_{\mu\mu}^{\text{core}} + \sum_{\lambda=1; \lambda \text{ on A}}^K P_{\lambda\lambda} \gamma_{AA} - \frac{1}{2} P_{\mu\mu} \gamma_{AA} + \sum_{\lambda=1; \lambda \text{ not on A}}^K P_{\lambda\lambda} \gamma_{AB} \quad (2.240)$$

$$F_{\mu\nu} = H_{\mu\nu}^{\text{core}} - \frac{1}{2} P_{\mu\nu} \gamma_{AA}; \quad \mu \text{ and } \nu \text{ both on atom A} \quad (2.241)$$

$$F_{\mu\nu} = H_{\mu\nu}^{\text{core}} - \frac{1}{2} P_{\mu\nu} \gamma_{AB}, \quad \mu \text{ and } \nu \text{ on different atoms, A and B} \quad (2.242)$$

Equation (2.240) is rather untidy, involving summations over basis functions on atom A and basis functions not on atom A. It is often simplified by writing P_{AA} as the total electron density on atom A, where:

$$P_{AA} = \sum_{\lambda \text{ on A}}^A P_{\lambda\lambda} \quad (2.243)$$

A similar expression can also be introduced for P_{BB} . With this notation $F_{\mu\mu}$ simplifies to:

$$F_{\mu\mu} = H_{\mu\mu}^{\text{core}} + (P_{AA} - \frac{1}{2} P_{\mu\mu}) \gamma_{AA} + \sum_{B \neq A} P_{BB} \gamma_{AB} \quad (2.244)$$

The core Hamiltonian expressions, $H_{\mu\mu}^{\text{core}}$ and $H_{\mu\nu}^{\text{core}}$, correspond to electrons moving in the field of the parent nucleus and the other nuclei. In semi-empirical methods the core electrons are subsumed into the nucleus and so the nuclear charges are altered accordingly (for example, carbon has a nuclear ‘charge’ of +4).

In CNDO $H_{\mu\mu}^{\text{core}}$ is separated into an integral involving the atom on which ϕ_μ is situated (labelled A), and all the others (labelled B). Thus:

$$H_{\mu\mu}^{\text{core}} = U_{\mu\mu} - \sum_{B \neq A} V_{AB} \quad (2.245)$$

where:

$$U_{\mu\mu} = \left(\mu \left| -\frac{1}{2} \nabla^2 - \frac{Z_A}{|\mathbf{r}_1 - \mathbf{R}_A|} \right| \mu \right) \quad \text{and} \quad V_{AB} = \left(\mu \left| \frac{Z_B}{|\mathbf{r}_1 - \mathbf{R}_B|} \right| \mu \right) \quad (2.246)$$

$U_{\mu\mu}$ is thus the energy of the orbital ϕ_μ in the field of its own nucleus (A) and core electrons; $-V_{AB}$ is the energy of the electron in the field of another nucleus (B). To maintain consistency with the way in which the two-electron integrals are treated, the terms

$$\left(\mu \left| \frac{Z_B}{|\mathbf{r}_1 - \mathbf{R}_B|} \right| \mu \right) \quad (2.247)$$

must be the same for all orbitals ϕ_μ on atom A (i.e. the interaction energy between any electron in an orbital on atom A with the core of atom B is equal to V_{AB}).

We next consider $H_{\mu\nu}^{\text{core}}$, where ϕ_μ and ϕ_ν are both on the same atom, A. In this case the core Hamiltonian has the following form:

$$\begin{aligned} H_{\mu\nu}^{\text{core}} &= \left(\mu \left| -\frac{1}{2} \nabla^2 - \frac{Z_A}{|\mathbf{r}_1 - \mathbf{R}_A|} \right| \nu \right) - \sum_{B \neq A} \left(\mu \left| \frac{Z_B}{|\mathbf{r}_1 - \mathbf{R}_B|} \right| \nu \right) \\ &= U_{\mu\nu} - \sum_{B \neq A} \left(\mu \left| \frac{Z_B}{|\mathbf{r}_1 - \mathbf{R}_B|} \right| \nu \right) \end{aligned} \quad (2.248)$$

As ϕ_μ and ϕ_ν are on the same atom, $U_{\mu\nu}$ is zero due to the orthogonality of atomic orbitals. The term

$$\left(\mu \left| \frac{Z_B}{|\mathbf{r}_1 - \mathbf{R}_B|} \right| \nu \right) \quad (2.249)$$

is zero in accordance with the zero-differential overlap approximation. Thus $H_{\mu\nu}^{\text{core}}$ is zero in CNDO.

Finally, if ϕ_μ and ϕ_ν are on two different atoms A and B, then we can write:

$$H_{\mu\nu}^{\text{core}} = \left(\mu \left| -\frac{1}{2} \nabla^2 - \frac{Z_A}{|\mathbf{r}_1 - \mathbf{R}_A|} - \frac{Z_B}{|\mathbf{r}_1 - \mathbf{R}_B|} \right| \nu \right) - \sum_{C \neq A, B} \left(\mu \left| -\frac{Z_C}{|\mathbf{r}_1 - \mathbf{R}_C|} \right| \nu \right) \quad (2.250)$$

The second term corresponds to the interaction of the distribution $\phi_\mu \phi_\nu$ with the atoms C ($\neq A, B$). These interactions are ignored. The first part (known as the *resonance integral* and commonly written $\beta_{\mu\nu}$) is not subject to the ZDO approximation, because it is the main cause of bonding. In CNDO the resonance integral is made proportional to the overlap integral, $S_{\mu\nu}$:

$$H_{\mu\nu}^{\text{core}} = \beta_{AB}^0 S_{\mu\nu} \quad (2.251)$$

where β_{AB}^0 is a parameter which depends on the nature of atoms A and B.

With these approximations the Fock matrix elements for CNDO become:

$$F_{\mu\mu} = U_{\mu\mu} + \sum_{B \neq A} V_{AB} + (P_{AA} - \frac{1}{2} P_{\mu\mu}) \gamma_{AA} + \sum_{B \neq A} P_{BB} \gamma_{AB} \quad (2.252)$$

$$F_{\mu\nu} = -\frac{1}{2} P_{\mu\nu} \gamma_{AA}; \quad \mu \text{ and } \nu \text{ on the same atom, A} \quad (2.253)$$

$$F_{\mu\nu} = \beta_{AB}^0 S_{\mu\nu} - \frac{1}{2} P_{\mu\nu} \gamma_{AB}, \quad \mu \text{ on A and } \nu \text{ on B} \quad (2.254)$$

To perform a CNDO calculation requires the following to be calculated or specified: the overlap integrals, $S_{\mu\nu}$, the core Hamiltonians $U_{\mu\mu}$, the electron–core interactions V_{AB} , the electron repulsion integrals γ_{AB} and γ_{AA} and the bonding parameters β_{AB}^0 . The CNDO basis set comprises Slater type orbitals for the valence shell with the exponents being chosen using Slater's rules (except for hydrogen, where an exponent of 1.2 is used as this value is more appropriate to hydrogen atoms in molecules). Thus the basis set comprises 1s for hydrogen and 2s, 2p_x, 2p_y and 2p_z for the first-row elements. The overlap integrals are calculated explicitly (the overlap between two basis functions on the same atom is, of course, zero with an s, p basis set). The electron repulsion integral parameter γ_{AB} is

calculated using valence s functions on the two atoms A and B:

$$\gamma_{AB} = \iint d\nu_1 d\nu_2 \phi_{s,A}(1) \phi_{s,A}(1) \left(\frac{1}{r_{12}} \right) \phi_{s,B}(2) \phi_{s,B}(2) \quad (2.255)$$

The use of spherically symmetric s orbitals avoids the problems associated with transformations of the axes. The core Hamiltonians ($U_{\mu\mu}$) are not calculated but are obtained from experimental ionisation energies. This is because it is important to distinguish between s and p orbitals in the valence shell (i.e. the 2s and 2p orbitals for the first-row elements), and without explicit core electrons this is difficult to achieve. The resonance integrals, β_{AB}^0 , are written in terms of empirical single-atom values as follows:

$$\beta_{AB}^0 = \frac{1}{2} (\beta_A^0 + \beta_B^0) \quad (2.256)$$

The β^0 values are chosen to fit the results of minimal basis set *ab initio* calculations on diatomic molecules.

The electron–core interaction, V_{AB} , is calculated as the interaction between an electron in a valence s orbital on atom A with the nuclear core of atom B:

$$V_{AB} = \int d\nu_1 \phi_{s,A}(1) \frac{Z_B}{|\mathbf{r}_1 - \mathbf{R}_B|} \phi_{s,A}(1) \quad (2.257)$$

CNDO is rightly recognised as the first in a long line of important semi-empirical models. However, there were some important limitations with the model. One especially serious deficiency of the first version of CNDO (introduced in 1965 [Pople and Segal 1965, Pople *et al.* 1965] and now known as CNDO/1) is that two neutral atoms show a significant (and incorrect) attraction, even when separated by several ångströms. The predicted equilibrium distances for diatomic molecules are also too short and the dissociation energies too large. These effects are due to electrons on one atom penetrating the valence shell of another atom and so experiencing a nuclear attraction. This penetration effect can be quantified more explicitly as follows. The net charge on an atom B equals the difference between its nuclear charge and the total electron density: $Q_B = Z_B - P_{BB}$. If we now substitute for P_{BB} ($= Z_B - Q_B$) in the diagonal elements of the Fock matrix, Equation (2.252), we obtain:

$$F_{\mu\mu} = U_{\mu\mu} + (P_{AA} - \frac{1}{2}P_{\mu\mu})\gamma_{AA} + \sum_{B \neq A} [-Q_B\gamma_{AB} + (Z_B\gamma_{AB} - V_{AB})] \quad (2.258)$$

$-Q_B\gamma_{AB}$ is the contribution from the total charge on atom B; this is zero if the atomic charge is exactly balanced by the electron density. $Z_B\gamma_{AB} - V_{AB}$ is called the *penetration integral*. It was this contribution that caused the anomalous results for two neutral atoms at large separation. In the second version of CNDO (CNDO/2 [Pople and Segal 1966]) the penetration integral effect was eliminated by putting $V_{AB} = Z_B\gamma_{AB}$. The core Hamiltonian $U_{\mu\mu}$ was also defined differently in CNDO/2, using both ionisation energies and electron affinities.

2.9.3 INDO

CNDO makes no allowance for the fact that the interaction between two electrons depends upon their relative spins. This effect can be particularly severe for electrons on the same

atom. Thus, in CNDO all two-electron integrals ($\mu\nu|\lambda\nu$) are set to zero, and integrals ($\mu\mu|\nu\nu$) and ($\mu\mu|\mu\mu$) are forced to be equal (to γ_{AA}). The next development was the intermediate neglect of differential overlap model (INDO [Pople *et al.* 1967]), which includes monatomic differential overlap for one-centre integrals (i.e. for integrals involving basis functions centred on the same atom). This enables the interaction between two electrons on the same atom with parallel spins to have a lower energy than the comparable interaction between electrons with paired spins. For this reason the Fock matrix elements are usually written with the spin (α or β) explicitly specified. The elements $F_{\mu\mu}$ and $F_{\mu\nu}$ (where μ and ν are located on atom A) then change from their CNDO/2 values as follows:

$$F_{\mu\mu}^{\circ} = U_{\mu\mu} + \sum_{\lambda \text{ on A}} \sum_{\sigma \text{ on A}} [P_{\lambda\sigma}(\mu\mu|\lambda\sigma) - P_{\lambda\sigma}^{\alpha}(\mu\lambda|\mu\sigma)] + \sum_{B \neq A} (P_{BB} - Z_B) \gamma_{AB} \quad (2.259)$$

$$F_{\mu\nu}^{\alpha} = U_{\mu\nu} + \sum_{\lambda \text{ on A}} \sum_{\sigma \text{ on A}} [P_{\lambda\sigma}(\mu\nu|\lambda\sigma) - P_{\lambda\sigma}^{\alpha}(\mu\lambda|\nu\sigma)]; \quad \mu \text{ and } \nu \text{ both on atom A} \quad (2.260)$$

In Equation (2.259) we have included the CNDO/2 approximation $V_{AB} = Z_B \gamma_{AB}$. The matrix element $F_{\mu\nu}$, where μ and ν are on different atoms, is the same as in CNDO/2:

$$F_{\mu\nu}^{\alpha} = \frac{1}{2} (\beta_A^0 + \beta_B^0) S_{\mu\nu} - P_{\mu\nu}^{\alpha} \gamma_{AB} \quad (2.261)$$

In a closed-shell system, $P_{\mu\nu}^{\alpha} = P_{\mu\nu}^{\beta} = \frac{1}{2} P_{\mu\nu}$ and the Fock matrix elements can be obtained by making this substitution. If a basis set containing s, p orbitals is used, then many of the one-centre integrals nominally included in INDO are equal to zero, as are the core elements $U_{\mu\nu}$. Specifically, only the following one-centre, two-electron integrals are non-zero: ($\mu\mu|\mu\mu$), ($\mu\mu|\nu\nu$) and ($\mu\nu|\mu\nu$). The elements of the Fock matrix that are affected can then be written as follows:

$$F_{\mu\mu} = U_{\mu\mu} + \sum_{\nu \text{ on A}} [P_{\nu\nu}(\mu\mu|\nu\nu) - \frac{1}{2} P_{\nu\nu}(\mu\nu|\mu\nu)] + \sum_{B \neq A} (P_{BB} - Z_B) \gamma_{AB} \quad (2.262)$$

$$F_{\mu\nu} = \frac{3}{2} P_{\mu\nu}(\mu\nu|\mu\nu) - \frac{1}{2} P_{\mu\nu}(\mu\mu|\nu\nu); \quad \mu, \nu \text{ on the same atom} \quad (2.263)$$

Some of the one-centre two-electron integrals in INDO are semi-empirical parameters, obtained by fitting to atomic spectroscopic data. The core integrals $U_{\mu\mu}$ are obtained in a slightly different fashion to that of CNDO/2, to take into account the new electronic configurations under the INDO model for atoms and their cations and anions. An INDO calculation requires little additional computational effort compared with the corresponding CNDO calculation and has the key advantage that states of different multiplicities can be distinguished. For example, in CNDO the singlet and triplet configurations $1s^2 2s^2 2p^2$ of carbon have the same energy, whereas these can be distinguished using INDO. Two of the systems considered in the original INDO publication were the methyl and ethyl radicals, the unpaired electron density being compared with experimentally determined hyperfine coupling constants. INDO gave a much more favourable result for these systems than CNDO.

2.9.4 NDDO

The next level of approximation is the neglect of diatomic differential overlap model (NDDO [Pople *et al.* 1965]); this theory only neglects differential overlap between atomic orbitals on

different atoms. Thus all of the two-electron, two-centre integrals of the form $(\mu\nu|\lambda\sigma)$, where μ and ν are on the same atom and λ and σ are also on the same atom, are retained. The Fock matrix elements become:

$$F_{\mu\mu} = H_{\mu\mu}^{\text{core}} + \sum_{\lambda \text{ on A}} \sum_{\sigma \text{ on A}} [P_{\lambda\sigma}(\mu\mu|\lambda\sigma) - \frac{1}{2}P_{\lambda\sigma}(\mu\lambda|\mu\sigma)] + \sum_{B \neq A} \sum_{\lambda \text{ on B}} \sum_{\sigma \text{ on B}} P_{\lambda\sigma}(\mu\mu|\lambda\sigma) \quad (2.264)$$

$$\begin{aligned} F_{\mu\nu} = H_{\mu\nu}^{\text{core}} &+ \sum_{\lambda \text{ on A}} \sum_{\sigma \text{ on A}} [P_{\lambda\sigma}(\mu\nu|\lambda\sigma) - \frac{1}{2}P_{\lambda\sigma}(\mu\lambda|\nu\sigma)] \\ &+ \sum_{B \neq A} \sum_{\lambda \text{ on B}} \sum_{\sigma \text{ on B}} P_{\lambda\sigma}(\mu\nu|\lambda\sigma); \quad \mu \text{ and } \nu \text{ both on A} \end{aligned} \quad (2.265)$$

$$F_{\mu\nu} = H_{\mu\nu}^{\text{core}} - \frac{1}{2} \sum_{\lambda \text{ on B}} \sum_{\sigma \text{ on A}} P_{\lambda\sigma}(\mu\sigma|\nu\lambda); \quad \mu \text{ on A and } \nu \text{ on B} \quad (2.266)$$

It is again possible to tidy up equations (2.264) and (2.265) when an s, p basis set is used:

$$F_{\mu\mu} = H_{\mu\mu}^{\text{core}} + \sum_{\nu \text{ on A}} [P_{\nu\nu}(\mu\mu|\nu\nu) - \frac{1}{2}P_{\nu\nu}(\mu\nu|\mu\nu)] + \sum_{B \neq A} \sum_{\lambda \text{ on B}} \sum_{\sigma \text{ on B}} P_{\lambda\sigma}(\mu\mu|\lambda\sigma) \quad (2.267)$$

$$F_{\mu\nu} = H_{\mu\nu}^{\text{core}} + \frac{3}{2}P_{\mu\nu}(\mu\nu|\mu\nu) - \frac{1}{2}P_{\mu\nu}(\mu\mu|\nu\nu) + \sum_{B \neq A} \sum_{\lambda \text{ on B}} \sum_{\sigma \text{ on B}} P_{\lambda\sigma}(\mu\nu|\lambda\sigma) \quad (2.268)$$

Whereas the computation required for an INDO calculation is little more than for the analogous CNDO calculation, in NDDO the number of two-electron, two-centre integrals is increased by a factor of approximately 100 for each pair of heavy atoms in the system.

2.9.5 MINDO/3

The CNDO, INDO and NDDO methods, as originally devised and implemented, are now little used, in comparison with the methods subsequently developed by Dewar and colleagues, but they were of considerable importance in showing how a systematic series of approximations could be used to develop methods of real practical value. Moreover, the calculations could be performed in a fraction of the time required to solve the full Roothaan–Hall equations. However, they did not produce very accurate results, largely because they were parametrised upon the results from relatively low-level *ab initio* calculations, which themselves agreed poorly with experiment. They were also limited to small classes of molecule, and they often required a good experimental geometry to be supplied as input because their geometry optimisation algorithms were not very sophisticated.

It was through the introduction of the MINDO/3 method by Bingham, Dewar and Lo [Bingham *et al.* 1975a–d] that a wider audience was able to apply semi-empirical methods in their own research. MINDO/3 was not so much a significant change in the theory, being based upon INDO (MINDO stands for modified INDO), but it did differ significantly in the way in which the method was parametrised, making much more use of experimental data. It also incorporated a geometry optimisation routine (the Davidon–Fletcher–Powell method; see Chapter 5), which enabled the program to accept crude initial geometries as input and derive the associated minimum energy structures.

MINDO/3 uses an s, p basis set and its Fock matrix elements are:

$$F_{\mu\mu} = U_{\mu\mu} + \sum_{\nu \text{ on A}} (P_{\nu\nu}(\mu\mu|\nu\nu) - \frac{1}{2}P_{\nu\nu}(\mu\nu|\mu\nu)) + \sum_{B \neq A} (P_{BB} - Z_B)\gamma_{AB} \quad (2.269)$$

$$F_{\mu\nu} = -\frac{1}{2}P_{\mu\nu}(\mu\nu|\mu\nu); \quad \mu \text{ and } \nu \text{ both on the same atom A} \quad (2.270)$$

$$F_{\mu\nu} = H_{\mu\nu}^{\text{core}} - \frac{1}{2}P_{\mu\nu}(\mu\nu|\mu\nu) = H_{\mu\nu}^{\text{core}} - \frac{1}{2}P_{\mu\nu}\gamma_{AB}; \quad \mu \text{ on A and } \nu \text{ on B} \quad (2.271)$$

The two-centre repulsion integrals γ_{AB} in MINDO/3 are calculated using the following function.

$$\gamma_{AB} = \frac{e^2}{\left[R_{AB}^2 + \frac{1}{4} \left(\frac{e^2}{\bar{g}_A} + \frac{e^2}{\bar{g}_B} \right)^2 \right]^{1/2}} \quad (2.272)$$

\bar{g}_A is the average of the one-centre, two-electron integrals $g_{\mu\nu}$ on atom A (i.e. $g_{\mu\nu} \equiv (\mu\mu|\nu\nu)$) and \bar{g}_B is the equivalent average for atom B. This seemingly complex function for γ_{AB} is, in fact, quite simple; at large R_{AB} it tends towards the Coulomb's law expression e^2/R_{AB} and as R_{AB} tends to zero it approaches the average of the one-centre integrals on the two atoms. The two-centre, one-electron integrals $H_{\mu\nu}^{\text{core}}$ are given in MINDO/3 by:

$$H_{\mu\nu}^{\text{core}} = S_{\mu\nu}\beta_{AB}(I_\mu + I_\nu) \quad (2.273)$$

$S_{\mu\nu}$ is the overlap integral, I_μ and I_ν are ionisation potentials for the appropriate orbitals and β_{AB} is a parameter dependent upon both of the two atoms A and B.

The core–core interaction between pairs of nuclei was also changed in MINDO/3 from the form used in CNDO/2. One way to correct the fundamental problems with CNDO/2 such as the repulsion between two hydrogen atoms (or indeed any neutral molecules) at all distances is to change the core–core repulsion term from a simple Coulombic expression ($E_{AB} = Z_A Z_B / R_{AB}$) to:

$$E_{AB} = Z_A Z_B \gamma_{AB} \quad (2.274)$$

In fact, while this correction gives the desired behaviour at relatively long separations, it does not account for the fact that as two nuclei approach each other the screening by the core electrons decreases. As the separation approaches zero the core–core repulsion should be described by Coulomb's law. In MINDO/3 this is achieved by making the core–core interaction a function of the electron–electron repulsion integrals as follows:

$$E_{AB} = Z_A Z_B \{ \gamma_{AB} + [(e^2/R_{AB}) - \gamma_{AB}] \exp(-\alpha_{AB}R_{AB}) \} \quad (2.275)$$

α_{AB} is a parameter dependent upon the nature of the atoms A and B. For OH and NH bonds a slightly different core–core interaction was found to be more appropriate:

$$E_{XH} = Z_X Z_H \{ \gamma_{XH} + [(e^2/R_{XH}) - \gamma_{XH}] \alpha_{XH} \exp(-R_{XH}) \} \quad (2.276)$$

The parameters for MINDO/3 were obtained in an entirely different way from previous semi-empirical methods. Some of the values that were fixed in CNDO, INDO and NNDO were permitted to vary during the MINDO/3 parametrisation procedure. For example, the exponents of the Slater atomic orbitals were allowed to vary from the values given by Slater's rules, and indeed the exponents for s and p orbitals were not required to be the

same. $U_{\mu\mu}$ and β_{AB} were also regarded as variable parameters. Another key difference was that the MINDO/3 parametrisation used experimental data such as molecular geometries and heats of formation, rather than theoretical values from *ab initio* calculations or data from atomic spectra. The parametrisation effort was a considerable undertaking, and it was only at the fourth attempt that an acceptable model was obtained (as is implicit in the appearance of the '3' in the name). For example, just to parametrise two atoms such as carbon and hydrogen using a set of 20 molecules required between 30 000 and 50 000 SCF calculations for each parametrisation scheme that was investigated.

2.9.6 MNDO

MINDO/3 proved to be very successful when it was introduced; it is important to realise that even simple *ab initio* calculations were beyond the computational resources of all but a few research groups in the 1970s. However, there were some significant limitations. For example, heats of formation of unsaturated molecules were consistently too positive, the errors in calculated bond angles were often quite large, and the heats of formation for molecules containing adjacent atoms with lone pairs were too negative. Some of these limitations were due to the use of the INDO approximation, and in particular the inability of INDO to deal with systems containing lone pairs. Dewar and Thiel therefore introduced the modified neglect of diatomic overlap (MNDO) method, which was based on NDDO [Dewar and Thiel 1977a, b]. The Fock matrix elements in MNDO were as follows:

$$F_{\mu\mu} = H_{\mu\mu}^{\text{core}} + \sum_{\nu \text{ on A}} [P_{\nu\nu}(\mu\mu|\nu\nu) - \frac{1}{2}P_{\nu\nu}(\mu\nu|\mu\nu)] + \sum_{B \neq A} \sum_{\lambda \text{ on B}} \sum_{\sigma \text{ on B}} P_{\lambda\sigma}(\mu\mu|\lambda\sigma) \quad (2.277)$$

$$\text{where } H_{\mu\mu}^{\text{core}} = U_{\mu\mu} - \sum_{B \neq A} V_{\mu\mu B} \quad (2.278)$$

$$\begin{aligned} F_{\mu\nu} = & H_{\mu\nu}^{\text{core}} + \frac{3}{2}P_{\mu\nu}(\mu\nu|\mu\nu) - \frac{1}{2}P_{\mu\nu}(\mu\mu|\nu\nu) \\ & + \sum_{B \neq A} \sum_{\lambda \text{ on B}} \sum_{\sigma \text{ on B}} P_{\lambda\sigma}(\mu\nu|\lambda\sigma); \quad \mu \text{ and } \nu \text{ both on A} \end{aligned} \quad (2.279)$$

$$\text{where } H_{\mu\nu}^{\text{core}} = - \sum_{B \neq A} V_{\mu\nu B} \quad (2.280)$$

$$F_{\mu\nu} = H_{\mu\nu}^{\text{core}} - \frac{1}{2} \sum_{\lambda \text{ on B}} \sum_{\sigma \text{ on A}} P_{\lambda\sigma}(\mu\sigma|\nu\lambda); \quad \mu \text{ on A and } \nu \text{ on B} \quad (2.281)$$

$$\text{where } H_{\mu\nu}^{\text{core}} = \frac{1}{2}S_{\mu\nu}(\beta_\mu + \beta_\nu) \quad (2.282)$$

The similarity with the NDDO expressions, Equations (2.264)–(2.266), can clearly be seen; the major new features are the appearance of terms $V_{\mu\mu B}$ and $V_{\mu\nu B}$ and a new form for the two-centre, one-electron core resonance integrals, which depend upon the overlap $S_{\mu\nu}$ and parameters β_μ and β_ν as shown in Equation (2.282). $V_{\mu\mu B}$ and $V_{\mu\nu B}$ are two-centre, one-electron attractions between an electron distribution $\phi_\mu \phi_\mu$ or $\phi_\mu \phi_\nu$, respectively, on atom A and the core of atom B. These are expressed as follows:

$$V_{\mu\mu B} = -Z_B(\mu_A \mu_A | s_B s_B) \quad (2.283)$$

$$V_{\mu\nu B} = -Z_B(\mu_A \nu_A | s_B s_B) \quad (2.284)$$

The core–core repulsion terms are also different in MNDO from those in MINDO/3, with OH and NH bonds again being treated separately:

$$E_{AB} = Z_A Z_B (s_A s_A | s_B s_B) \{1 + \exp(-\alpha_A R_{AB}) + \exp(-\alpha_B R_{AB})\} \quad (2.285)$$

$$E_{XH} = Z_X Z_H (s_X s_X | s_H s_H) \{1 + R_{XH} \exp(-\alpha_X R_{XH}) / R_{AB} + \exp(-\alpha_H R_{XH})\} \quad (2.286)$$

Perhaps the most significant advantage of MNDO over MINDO/3 is the use throughout of monatomic parameters; MINDO/3 requires diatomic parameters in the resonance integral (β_{AB}) and the core–core repulsion (α_{AB}). It has been possible to expand MNDO to cover a much wider variety of elements such as aluminium, silicon, germanium, tin, bromine and lead. However, the use of an (s, p) basis set in the original MNDO method did mean that the method could not be applied to most transition metals, which require a basis set containing d orbitals. In addition, hypervalent compounds of sulphur and phosphorus are not modelled well. In more recent versions of the MNDO method d orbitals have been explicitly included for the heavier elements [Thiel and Voityuk 1994]. Another serious limitation of MNDO is its inability to accurately model intermolecular systems involving hydrogen bonds (for example, the heat of formation of the water dimer is far too low in MNDO). This is because of a tendency to overestimate the repulsion between atoms when they are separated by a distance approximately equal to the sum of their van der Waals radii. Conjugated systems can also present difficulties for MNDO. An extreme example of this occurs with compounds such as nitrobenzene in which the nitro group is predicted to be orthogonal to the aromatic ring rather than conjugated with it. In addition, MNDO energies are too positive for sterically crowded molecules and too negative for molecules containing four-membered rings.

2.9.7 AM1

The Austin Model 1 (AM1) model was the next semi-empirical theory produced by Dewar's group [Dewar *et al.* 1985]. AM1 was designed to eliminate the problems with MNDO, which were considered to arise from a tendency to overestimate repulsions between atoms separated by distances approximately equal to the sum of their van der Waals radii. The strategy adopted was to modify the core–core term using Gaussian functions. Both attractive and repulsive Gaussian functions were used; the attractive Gaussians were designed to overcome the repulsion directly and were centred in the region where the repulsions were too large. Repulsive Gaussian functions were centred at smaller internuclear separations. With this modification the expression for the core–core term was related to the MNDO expression by:

$$E_{AB} = E_{MNDO} + \frac{Z_A Z_B}{R_{AB}} \times \left\{ \sum_i K_{A_i} \exp[-L_{A_i} (R_{AB} - M_{A_i})^2] + \sum_j K_{B_j} \exp[-L_{B_j} (R_{AB} - M_{B_j})^2] \right\} \quad (2.287)$$

The additional terms are spherical Gaussian functions with a width determined by the parameter L . It was found that the values of these parameters were not critical and many

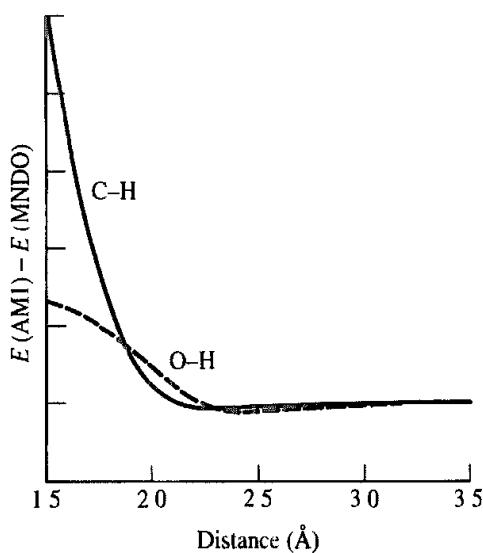


Fig 2.20. The difference in the core–core energy for AM1 and MNDO for carbon–hydrogen and oxygen–hydrogen interactions

were set to the same value. The M and K parameters were optimised for each atom, together with the α parameters in the exponential terms in Equations (2.285) and (2.286). In the original parametrisation of AM1 there are four terms in the Gaussian expansion for carbon, three for hydrogen and nitrogen and two for oxygen (both attractive and repulsive Gaussians were used for carbon, hydrogen and nitrogen but only repulsive Gaussians for oxygen). The effect of including these Gaussian functions can be seen in Figure 2.20, which plots the difference in the MNDO and AM1 core–core terms for the carbon–hydrogen and oxygen–hydrogen interactions. The inclusion of these Gaussians significantly increased the number of parameters per atom, from seven in the MNDO to between 13 and 16 per atom in AM1. This, of course, made the parametrisation process considerably more difficult. Overall, AM1 was a significant improvement over MNDO and many of the deficiencies associated with the core repulsion were corrected.

2.9.8 PM3

PM3 is also based on MNDO (the name derives from the fact that it is the third parametrisation of MNDO, AM1 being considered the second) [Stewart 1989a, b]. The PM3 Hamiltonian contains essentially the same elements as that for AM1, but the parameters for the PM3 model were derived using an automated parametrisation procedure devised by JJP Stewart. By contrast, many of the parameters in AM1 were obtained by applying chemical knowledge and ‘intuition’. As a consequence, some of the parameters have significantly different values in AM1 and PM3, even though both methods use the same functional form and they both predict various thermodynamic and structural properties to approximately the same level of accuracy. Some problems do remain with PM3. One of the most important of these is the rotational barrier of the amide bond, which is much too low and in some cases almost non-existent. This problem can be corrected through the use of an

empirical torsional potential (see Section 4.5). There has been considerable debate over the relative merits of the AM1 and PM3 approaches to parametrisation.

2.9.9 SAM1

The final offering from the Dewar group* was SAM1, which stands for ‘Semi-Ab-initio Model 1’ [Dewar *et al.* 1993]. The name was chosen to reflect Dewar’s belief that methods like AM1 offer such a significant enhancement over the earlier semi-empirical methods like CNDO/2 that they should be given a different generic name. In SAM1 a standard STO-3G Gaussian basis set is used to evaluate the electron repulsion integrals; close inspection of the results from AM1 and MNDO suggested that steric effects were overestimated because of the way in which the electron repulsion integrals were calculated. The resulting integrals were then scaled, partly to enable some of the effects of electron correlation to be included and partly to compensate for the use of a minimal basis set. The Gaussian terms in the core–core repulsion were retained to fine-tune the model. The number of parameters in SAM1 is no greater than in AM1 and fewer than in PM3. It does take longer to run (by up to two orders of magnitude) though it was felt that with the improvements in computer hardware such an increase was acceptable.

2.9.10 Programs for Semi-empirical Quantum Mechanical Calculations

The popularity of the MNDO, AM1 and PM3 methods is due in large part to their implementation in the MOPAC and AMPAC programs. The programs are able to perform many kinds of calculation and to calculate many different properties.

The contributions of the Dewar group are rightly recognised as particularly significant in the development of semi-empirical methods, but other research groups have also made important contributions. The SINDO1 and ZINDO programs have been developed in the groups of Jug and Zerner, respectively, and both contain novel features. The ZINDO program of Zerner and co-workers can perform a wide variety of semi-empirical calculations and has been particularly useful for calculations on transition metal and lanthanide compounds and for predicting molecular electronic spectra.

2.10 Hückel Theory

Hückel theory can be considered the ‘grandfather’ of approximate molecular orbital methods, having been formulated in the early 1930s [Hückel 1931]. Hückel theory is limited to conjugated π systems and was originally devised to explain the non-additive nature of certain properties of aromatic compounds. For example, the properties of benzene are much different from those of the hypothetical ‘cyclohexatriene’ molecule. Although Hückel theory, as originally formulated, is relatively little used in research today, extensions

* Michael Dewar died in 1997

to it such as extended Hückel theory are still employed and can provide qualitative insights into the electronic structure of important classes of molecule. Hückel theory is also widely used for teaching purposes to introduce a 'real' theory that can be applied to relatively complex systems with little more than pencil and paper or a simple computer program.

Hückel theory separates the π system from the underlying σ framework and constructs molecular orbitals into which the π electrons are then fed in the usual way according to the Aufbau principle. The π electrons are thus considered to be moving in a field created by the nuclei and the 'core' of σ electrons. The molecular orbitals are constructed from linear combinations of atomic orbitals and so the theory is an LCAO method. For our purposes it is most appropriate to consider Hückel theory in terms of the CNDO approximation (in fact, Hückel theory was the first ZDO molecular orbital theory to be developed). Let us examine the three types of Fock matrix element in Equations (2.252)–(2.254). First, $F_{\mu\mu}$. In a neutral species, the net charge on each atom will be approximately zero, and so if we take Equation (2.258), from which penetration effects have been eliminated, then we are left with $U_{\mu\mu} + (P_{AA} - 0.5P_{\mu\mu})\gamma_{AA}$. Now if each nucleus (A) in the π system is the same (i.e. carbon) then this expression will be approximately constant for all nuclei being considered. The matrix elements $F_{\mu\mu}$ are often (confusingly) called Coulomb integrals in Hückel theory and are assigned the symbol α . All off-diagonal elements of the Fock matrix are assumed to be zero with the exception of elements $F_{\mu\nu}$, where μ and ν are π orbitals on two bonded atoms. These $F_{\mu\nu}$ are assumed to be constant, are assigned the symbol β and are known as resonance integrals. The Fock matrix in Hückel theory thus has as many rows and columns as the number of atoms in the π system with diagonal elements that are all set to α . All off-diagonal elements F_{ij} are zero unless there is a bond between the atoms i and j , in which case the element is β . For benzene the Fock matrix is of the following form (atom labelling as in Figure 2.21):

$$\begin{pmatrix} \alpha & \beta & 0 & 0 & 0 & \beta \\ \beta & \alpha & \beta & 0 & 0 & 0 \\ 0 & \beta & \alpha & \beta & 0 & 0 \\ 0 & 0 & \beta & \alpha & \beta & 0 \\ 0 & 0 & 0 & \beta & \alpha & \beta \\ \beta & 0 & 0 & 0 & \beta & \alpha \end{pmatrix} \quad (2.288)$$

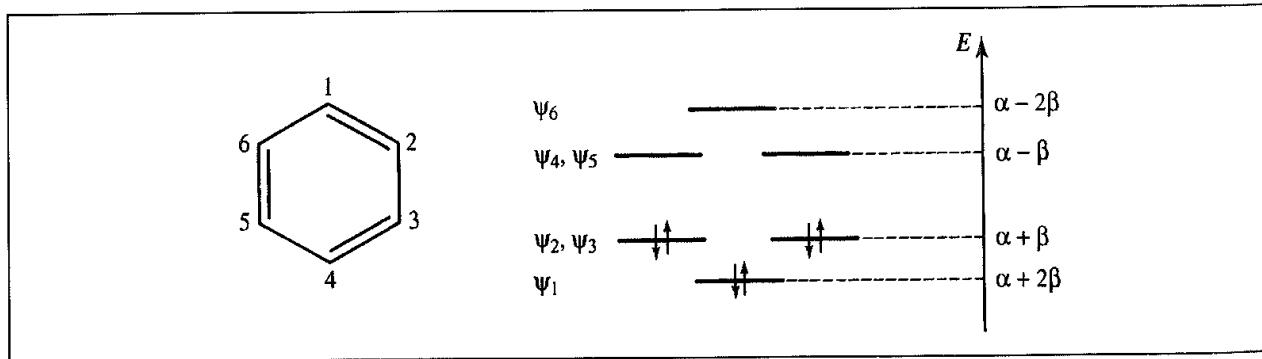


Fig 2.21: Benzene and its Hückel molecular orbitals.

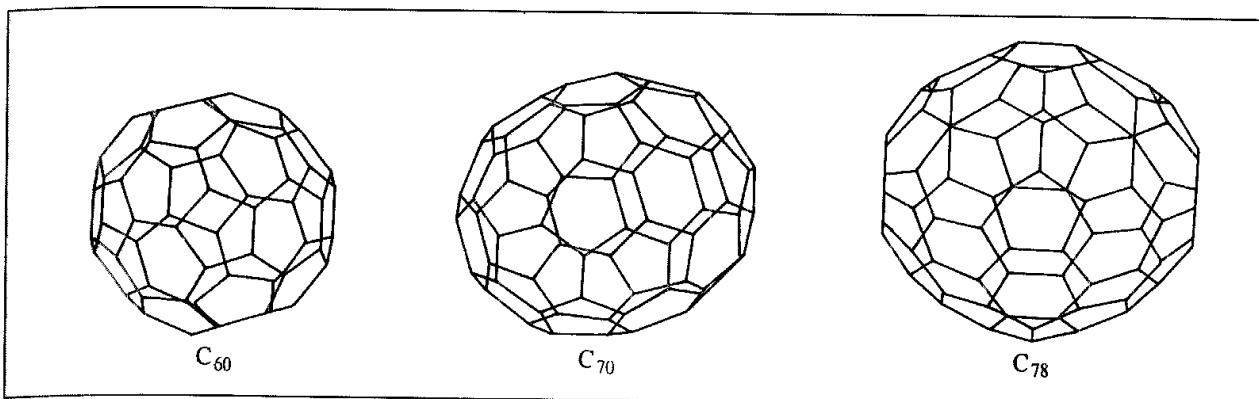


Fig 2.22 Three fullerenes, C₆₀, C₇₀ and C₇₈.

As with the other semi-empirical methods that we have considered so far, the overlap matrix is equal to the identity matrix. The following simple matrix equation must then be solved:

$$\mathbf{FC} = \mathbf{CE} \quad (2.289)$$

The equation can be solved by standard methods to give the basis set coefficients and the molecular orbital energies E. The orbital energies for benzene are $E_1 = \alpha + 2\beta$; $E_2, E_3 = \alpha + \beta$; $E_4, E_5 = \alpha - \beta$; $E_6 = \alpha - 2\beta$, and so the ground state places two electrons in ψ_1 and two each in the two degenerate orbitals ψ_2 and ψ_3 . The lowest-energy orbital ψ_1 is a linear combination of the six carbon p orbitals.

Hückel theory was extended to cover various other systems, including those with hetero-atoms, but it was not particularly successful and has largely been superseded by other semi-empirical methods. Nevertheless, for appropriate problems Hückel theory can be very useful. One example is the calculations of P W Fowler and colleagues, who studied the relationship between geometry and electronic structure for a range of buckminsterfullerenes (the parent molecule of which, C₆₀, was discovered in 1985) [Fowler 1993]. The fullerenes (or ‘buckyballs’) are excellent candidates for Hückel theory as they are composed of carbon and have extensive π systems; three examples are shown in Figure 2.22.

The results of their calculations were summarised in two rules. The first rule states that at least one isomer C_n with a properly closed p shell (i.e. bonding HOMO, antibonding LUMO) exists for all $n = 60 + 6k$ ($k = 0, 2, 3, \dots$, but not 1). Thus C₆₀, C₇₂, C₇₈, etc., are in this group. The second rule is for carbon cylinders and states that a closed-shell structure is found for $n = 2p(7 + 3k)$ (for all k). C₇₀ is the parent of this family. The calculations were extended to cover different types of structure and fullerenes doped with metals.

2.10.1 Extended Hückel Theory

Hückel theory is clearly limited, in part because it is restricted to π systems. The extended Hückel method is a molecular orbital theory that takes account of all the valence electrons in the molecule [Hoffmann 1963]. It is largely associated with R Hoffmann, who received the Nobel Prize for his contributions. The equation to be solved is $\mathbf{FC} = \mathbf{SCE}$, with the

Fock matrix elements taking the following simple forms:

$$F_{\mu\mu}^{\text{AA}} = H_{\mu\mu} = -I_\mu \quad (2.290)$$

$$F_{\mu\nu}^{\text{AB}} = H_{\mu\nu} = -\frac{1}{2}K(I_\mu + I_\nu)S_{\mu\nu} \quad (2.291)$$

In these equations, μ and ν are two atomic orbitals (e.g. Slater type orbitals), I_μ is the ionisation potential of the orbital and K is a constant, which was originally set to 1.75. The formula for the off-diagonal elements $H_{\mu\nu}$ (where μ and ν are on different atoms) was originally suggested by R S Mulliken. These off-diagonal matrix elements are calculated between all pairs of valence orbitals and so extended Hückel theory is not limited to π systems.

The extended Hückel approach has proved to be rather successful for such a simple theory; for example, the famous Woodward–Hoffmann rules (see Section 5.9.4) were based upon calculations using this model. Extended Hückel theory has found particular application in those areas where alternative theories cannot be used. This is largely due to the fact that the basis set requires no more than experimentally determined ionisation potentials. It is particularly useful for studying systems containing metals; these systems are problematic for many other methods due to the lack of suitable basis sets.

2.11 Performance of Semi-empirical Methods

Our discussion of the application of quantum mechanics calculations was not explicitly directed towards any particular quantum mechanical theory but was – implicitly at least – written with *ab initio* methods in mind. All of the properties we considered in Section 2.7 can also be determined using semi-empirical methods. Extensive tables detailing the performance of the popular semi-empirical methods have been published, both in the original papers and in review articles, some of which are listed at the end of this chapter. The parametrisation of the semi-empirical approaches typically includes geometrical variables, dipole moments, ionisation energies and heats of formation. In Table 2.7 we provide a summary of the performance of the MINDO/3, MNDO, AM1, PM3 and SAM1 semi-empirical methods from data supplied in the original publications. The performance of successive semi-empirical methods has gradually improved from one method to another, though one should always remember that anomalous results may be obtained for certain types of system. Some of these limitations were outlined in the discussion of the various semi-empirical methods. It is worth emphasising that some of the major drawbacks with the semi-empirical methods arise simply because one is trying to calculate properties that were not given a major consideration in the parametrisation process. For example, many of the molecules used for the parametrisation of the MNDO, AM1 and PM3 methods had little or no conformational flexibility and it is therefore not so surprising that some rotational barriers are not calculated with the same accuracy as (say) heats of formation. In addition, to achieve optimal performance for specific classes of molecules (e.g. the amino acids) or specific properties (e.g. conformational barriers) then it would be appropriate to include representative systems during the parametrisation procedure.

	MINDO/3	MINDO	AM1	PM3	SAM1	Reference
138 heats of formation (kcal/mol)	11.0	6.3				
228 bond lengths	0.022 Å	0.014 Å				[Dewar and Thel 1977b]
91 angles	5.6°	2.8°				
57 dipole moments	0.49 D	0.38 D				[Dewar et al. 1985]
58 heats of formation of hydrocarbons (kcal/mol)	9.7	5.87	5.07			
80 heats of formation for species with N and/or O (kcal/mol)	11.69	6.64	5.88			
46 dipole moments	0.54 D	0.32 D	0.26 D			
29 ionisation energies	0.31 eV	0.39 eV	0.29 eV			
406 heats of formation (kcal/mol)		8.82	7.12	5.21		
196 dipole moments		0.35 D	0.40 D	0.32 D		[Dewar et al. 1993]

Table 2.7 Comparison of quantities calculated with various semi-empirical methods.

Appendix 2.1 Some Common Acronyms Used in Computational Quantum Chemistry

AM1	Austin Model 1
AO	Atomic orbital
B3LYP	Scheme for hybrid Hartree–Fock/density functional theory introduced by Becke
BLYP	Becke–Lee–Yang–Parr gradient-corrected functional for use with density functional theory
BSSE	Basis set superposition error
CASSCF	Complete active space self-consistent field
CI	Configuration interaction
CIS	Configuration interaction singles
CISD	Configuration interaction singles and doubles
CNDO	Complete neglect of differential overlap
DFT	Density functional theory
DIIS	Direct inversion of iterative subspace
DVP	Double zeta with polarisation
DZ	Double zeta
EHT	Extended Hückel theory
GVB	Generalised valence bond model
HF	Hartree–Fock
HOMO	Highest occupied molecular orbital
INDO	Intermediate neglect of differential overlap
LCAO	Linear combination of atomic orbitals
LDA	Local density approximation
LSDFT	Local spin density functional theory
LUMO	Lowest unoccupied molecular orbital
MBPT	Many-body perturbation theory
MINDO/3	Modified INDO version 3
MNDO	Modified neglect of diatomic overlap
MO	Molecular orbital
MP	Møller–Plesset
MP2, MP3, etc.	Møller–Plesset theory at second order, third order, etc.
NDDO	Neglect of diatomic differential overlap
PM3	Parametrisation 3 of MNDO
QCISD	Quadratic configuration interaction singles and doubles
QCISD(T)	Configuration interation method involving single, double and quadratic excitations with an estimated triple excitation
RHF	Restricted Hartree–Fock
SAM1	Semi- <i>Ab initio</i> Model 1
SCF	Self-consistent field
STO	Slater type orbital
STO-3G, STO-4G, etc.	Minimal basis sets in which 3, 4 etc, Gaussian functions are used to represent the atomic orbitals on an atom

UHF	Unrestricted Hartree-Fock
WVN	Correlation functional due to Wilk, Vosko and Nusair
ZDO	Zero differential overlap

Further Reading

- Atkins P W 1991 *Quanta. A Handbook of Concepts*. Oxford, Oxford University Press
- Atkins P W 1998. *Physical Chemistry*. 6th Edition. Oxford, Oxford University Press.
- Atkins P W and R S Friedman 1996 *Molecular Quantum Mechanics*. Oxford, Oxford University Press
- Clark T 1985 *A Handbook of Computational Chemistry: A Practical Guide to Chemical Structure and Energy Calculations*. New York, Wiley-Interscience
- Dewar M J S 1969. *The Molecular Orbital Theory of Organic Chemistry*. New York, McGraw-Hill.
- Hinchliffe A 1988 *Computational Quantum Chemistry* Chichester, John Wiley & Sons.
- Hinchliffe A 1995. *Modelling Molecular Structures* Chichester, John Wiley & Sons.
- Hirst D M 1990 *A Computational Approach to Chemistry*. Oxford, Blackwell Scientific.
- Pople J A and D L Beveridge 1970. *Approximate Molecular Orbital Theory*. New York, McGraw-Hill
- Richards W G and D L Cooper 1983. *Ab initio Molecular Orbital Calculations for Chemists*. 2nd Edition Oxford, Clarendon Press.
- Schaeffer H F III (Editor) 1977. *Applications of Electronic Structure Theory* New York, Plenum Press
- Schaeffer H F III (Editor) 1977. *Methods of Electronic Structure Theory*. New York, Plenum Press
- Stewart J J P 1990. MOPAC. A Semi-Empirical Molecular Orbital Program. *Journal of Computer-Aided Molecular Design* 4:1-45
- Stewart J J P 1990. Semi-empirical Molecular Orbital Methods. In Lipkowitz K B and D B Boyd (Editors). *Reviews in Computational Chemistry* Volume 1. New York, VCH Publishers, pp 45-82
- Szabo A and N S Ostlund 1982. *Modern Quantum Chemistry Introduction to Advanced Electronic Structure Theory*. New York, McGraw-Hill.
- Zerner M C 1991. Semi-empirical Molecular Orbital Methods. In Lipkowitz K B and D B Boyd (Editors). *Reviews in Computational Chemistry* Volume 2. New York, VCH Publishers, pp 313-366

References

- Allinger N L, R S Grev, B F Yates and H F Schaeffer III 1990. The Syn Rotational Barrier in Butane *Journal of the American Chemical Society* 112:114-118.
- Bachrach S M 1994. Population Analysis and Electron Densities from Quantum Mechanics. In Lipkowitz K B and D B Boyd (Editors). *Reviews in Computational Chemistry* Volume 5. New York, VCH Publishers, pp 171-227.
- Bader R F W 1985 Atoms in Molecules. *Accounts of Chemistry Research* 18:9-15.
- Bingham R C, M J S Dewar and D H Lo 1975a. Ground States of Molecules. XXV. MNDO/3. An Improved Version of the MNDO Semi-empirical SCFMO Method. *Journal of the American Chemical Society* 97:1285-1293.
- Bingham R C, M J S Dewar and D H Lo 1975b. Ground States of Molecules XXVI. MNDO/3. Calculations for Hydrocarbons *Journal of the American Chemical Society* 97:1294-1301.
- Bingham R C, M J S Dewar and D H Lo 1975c. Ground States of Molecules XXVII MNDO/3 Calculations for CHON Species. *Journal of the American Chemical Society* 97:1302-1306.
- Bingham R C, M J S Dewar and D H Lo 1975d. Ground States of Molecules. XXVIII MNDO/3. Calculations for Compounds Containing Carbon, Hydrogen, Fluorine and Chlorine. *Journal of the American Chemical Society* 97:1307-1310.

- Boys S F 1950. Electronic Wave Functions. I. A General Method of Calculation for the Stationary States of Any Molecular System. *Proceedings of the Royal Society (London)* **A200**:542–554.
- Cusachs L C and Politzer 1968. On the Problem of Defining the Charge on an Atom in a Molecule *Chemical Physics Letters* **1**:529–531.
- Dewar M J S, C Jie and J Yu 1993. SAM1; The First of a New Series of General Purpose Quantum Mechanical Molecular Models. *Tetrahedron* **49**:5003–5038.
- Dewar M J S and Thiel W 1977a. Ground States of Molecules. 38. The MNDO Method Approximations and Parameters. *Journal of the American Chemical Society* **99**:4899–4907.
- Dewar M J S and Thiel W 1977b. Ground States of Molecules 39. MNDO Results for Molecules Containing Hydrogen, Carbon, Nitrogen and Oxygen. *Journal of the American Chemical Society* **99**:4907–4917
- Dewar M J S, E G Zoebisch, E F Healy and J J P Stewart 1985. AM1: A New General Purpose Quantum Mechanical Model. *Journal of the American Chemical Society* **107**:3902–3909
- Dunning T H Jr 1970 Gaussian Basis Functions for Use in Molecular Calculations. I. Contraction of (9s5p) Atomic Basis Sets for First-Row Atoms. *Journal of Chemical Physics* **53**:2823–2883.
- Fowler P W 1993 Systematics of Fullerenes and Related Clusters *Philosophical Transactions of the Royal Society (London)* **A343**:39–52.
- Hall G G 1951 The Molecular Orbital Theory of Chemical Valency VIII A Method for Calculating Ionisation Potentials. *Proceedings of the Royal Society (London)* **A205**:541–552
- Hehre W J, R F Stewart and J A Pople 1969 Self-Consistent Molecular-Orbital Methods I Use of Gaussian Expansions of Slater-Type Atomic Orbitals. *Journal of Chemical Physics* **51**:2657–2664
- Hehre W J, L Radom, P v R Schleyer and J A Pople 1986 *Ab initio Molecular Orbital Theory* New York, John Wiley & Sons.
- Hoffmann R 1963. An Extended Hückel Theory. I. Hydrocarbons. *Journal of Chemical Physics* **39**:1397–1412.
- Hückel Z 1931 Quanten theoretische Beiträge zum Benzolproblem. I Die Electron enkonfiguration des Benzols *Zeitschrift für Physik* **70**:203–286
- Huzinga S 1965. Gaussian-type Functions for Polyatomic Systems I. *Journal of Chemical Physics* **42**:1293–1302.
- Löwdin P-Q 1970. On the Orthogonality Problem. *Advances in Quantum Chemistry* **5**:185–199.
- Mayer I 1983. Charge, Bond Order and Valence in the *Ab initio* SCF Theory. *Chemical Physics Letters* **97**:270–274
- Mulliken R S 1955. Electronic Population Analysis on LCAO-MO Molecular Wave Functions. I *Journal of Chemical Physics* **23**:1833–1846.
- Politzer P and J S Murray 1991. Molecular Electrostatic Potentials and Chemical Reactivity. In Lipkowitz K B and D B Boyd (Editors). *Reviews in Computational Chemistry* Volume 2. New York, VCH Publishers, pp 273–312.
- Pople J A, D L Beveridge and P A Dobosh 1967 Approximate Self-Consistent Molecular Orbital Theory. V. Intermediate Neglect of Differential Overlap. *Journal of Chemical Physics* **47**:2026–2033.
- Pople J A, D P Santry and G A Segal 1965. Approximate Self-Consistent Molecular Orbital Theory. I. Invariant Procedures. *Journal of Chemical Physics* **43**:S129–S135
- Pople J A and G A Segal 1965. Approximate Self-Consistent Molecular Orbital Theory. II Calculations with Complete Neglect of Differential Overlap *The Journal of Chemical Physics* **43**:S136–S149
- Pople J A and G A Segal 1966. Approximate Self-Consistent Molecular Orbital Theory. III. CNDO Results for AB₂ and AB₃ systems. *Journal of Chemical Physics* **44**:3289–3296
- Reed A E, R B Weinstock and F Weinhold 1985. Natural Population Analysis. *Journal of Chemical Physics* **83**: 735–746.
- Rooshaan C C J 1951. New Developments in Molecular Orbital Theory *Reviews of Modern Physics* **23**:69–89
- Slater J C 1930. Atomic Shielding Constants. *Physical Review* **36**:57–64

- Smith G D and R L Jaffe 1996. Quantum Chemistry Study of Conformational Energies and Rotational Energy Barriers in *n*-Alkanes. *Journal of Physical Chemistry* **100**:18718–18724
- Stewart J J P 1989a. Optimisation of Parameters for Semi-empirical Methods I. Method. *Journal of Computational Chemistry* **10**:209–220
- Stewart J J P 1989b. Optimisation of Parameters for Semi-empirical Methods II. Applications. *Journal of Computational Chemistry* **10**:221–264
- Thiel W and A A Voityuk 1994. Extension of MNDO to d Orbitals: Parameters and Results for Silicon. *Journal of Molecular Structure (Theochem)* **313**:141–154.
- Wiberg K B and M A Murcko 1988. Rotational Barriers. 2. Energies of Alkane Rotamers. An Examination of Gauche Interactions. *Journal of the American Chemical Society* **110**:8029–8038.
- Wiberg K B and P R Rablen 1993. Comparison of Atomic Charges Derived via Different Procedures. *Journal of Computational Chemistry* **14**:1504–1518.

Advanced *ab initio* Methods, Density Functional Theory and Solid-state Quantum Mechanics

3.1 Introduction

In Chapter 2 we worked through the two most commonly used quantum mechanical models for performing calculations on ground-state ‘organic’-like molecules, the *ab initio* and semi-empirical approaches. We also considered some of the properties that can be calculated using these techniques. In this chapter we will consider various advanced features of the *ab initio* approach and also examine the use of density functional methods. Finally, we will examine the important topic of how quantum mechanics can be used to study the solid state.

3.2 Open-shell Systems

The Roothaan–Hall equations are not applicable to open-shell systems, which contain one or more unpaired electrons. Radicals are, by definition, open-shell systems as are some ground-state molecules such as NO and O₂. Two approaches have been devised to treat open-shell systems. The first of these is *spin-restricted* Hartree–Fock (RHF) theory, which uses combinations of singly and doubly occupied molecular orbitals. The closed-shell approach that we have developed thus far is a special case of RHF theory. The doubly occupied orbitals use the same spatial functions for electrons of both α and β spin. The orbital expansion Equation (2.144) is employed together with the variational method to derive the optimal values of the coefficients. The alternative approach is the *spin-unrestricted* Hartree–Fock (UHF) theory of Pople and Nesbet [Pople and Nesbet 1954], which uses two distinct sets of molecular orbitals: one for electrons of α spin and the other for electrons of β spin. Two Fock matrices are involved, one for each type of spin, with elements as follows:

$$F_{\mu\nu}^{\alpha} = H_{\mu\nu}^{\text{core}} + \sum_{\lambda=1}^K \sum_{\sigma=1}^K [[P_{\lambda\sigma}^{\alpha} + P_{\lambda\sigma}^{\beta}] (\mu\nu|\lambda\sigma) - P_{\lambda\alpha}^{\alpha} (\mu\lambda|\nu\sigma)] \quad (3.1)$$

$$P_{\mu\nu}^{\beta} = H_{\mu\nu}^{\text{core}} + \sum_{\lambda=1}^K \sum_{\sigma=1}^K [[P_{\lambda\sigma}^{\alpha} + P_{\lambda\sigma}^{\beta}](\mu\nu|\lambda\sigma) - P_{\lambda\alpha}^{\beta}(\mu\lambda|\nu\sigma)] \quad (3.2)$$

UHF theory also uses two density matrices, the full density matrix being the sum of these two:

$$P_{\mu\nu}^{\alpha} = \sum_{i=1}^{\alpha_{\text{occ}}} c_{\mu i}^{\alpha} c_{\nu i}^{\alpha} \quad P_{\mu\nu}^{\beta} = \sum_{i=1}^{\beta_{\text{occ}}} c_{\mu i}^{\beta} c_{\nu i}^{\beta} \quad (3.3)$$

$$P_{\mu\nu} = P_{\mu\nu}^{\alpha} + P_{\mu\nu}^{\beta} \quad (3.4)$$

The summations in Equations (3.3) and (3.4) are over the occupied orbitals with α and β spin as appropriate. Thus, $\alpha_{\text{occ}} + \beta_{\text{occ}}$ equals the total number of electrons in the system. In a closed-shell Hartree-Fock wavefunction the distribution of electron spin is zero everywhere because the electrons are paired. In an open-shell system, however, there is an excess of electron spin, which can be expressed as the spin density, analogous to the electron density. The spin density $\rho^{\text{spin}}(\mathbf{r})$ at a point \mathbf{r} is given by:

$$\rho^{\text{spin}}(\mathbf{r}) = \rho^{\alpha}(\mathbf{r}) - \rho^{\beta}(\mathbf{r}) = \sum_{\mu=1}^K \sum_{\nu=1}^K [P_{\mu\nu}^{\alpha} - P_{\mu\nu}^{\beta}] \phi_{\mu}(\mathbf{r}) \phi_{\nu}(\mathbf{r}) \quad (3.5)$$

Clearly, the UHF approach is more general and indeed the restricted Hartree-Fock approach is a special case of unrestricted Hartree-Fock. Figure 3.1 illustrates the conceptual difference between the RHF and the UHF models. Unrestricted wavefunctions are also the most appropriate way to deal with other problems such as molecules near the dissociation limit. The simplest example of this type of behaviour is the H_2 molecule, the ground state of which is a singlet with a bond length of approximately 0.75 Å. The restricted wavefunction is the appropriate Hartree-Fock wavefunction, with two paired electrons in a single spatial orbital. As the bond length increases towards the dissociation limit, this description is clearly inappropriate, for hydrogen is experimentally observed to dissociate to two

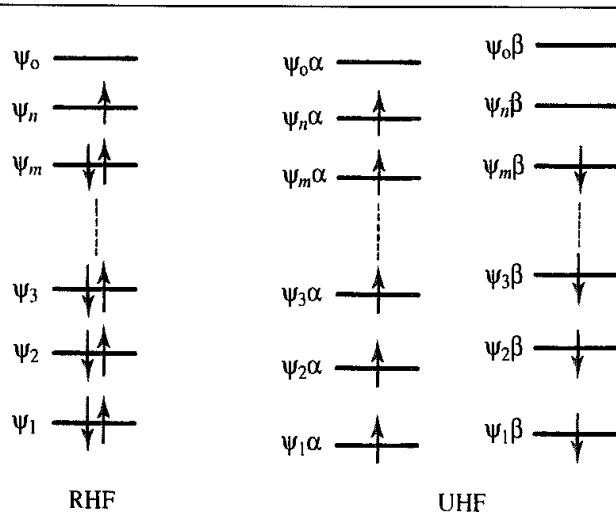


Fig. 3.1. The conceptual difference between the RHF and UHF models

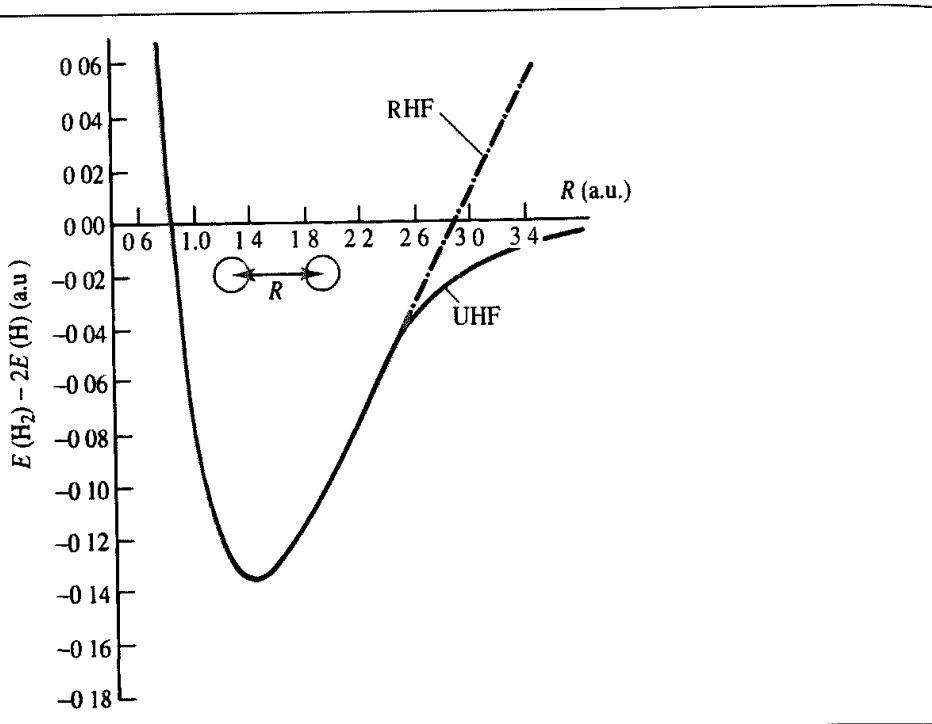


Fig. 3.2. UHF and RHF dissociation curves for H_2 . (Figure adapted from Szabo A, N S Ostlund 1982. Modern Quantum Chemistry Introduction to Advanced Electronic Structure Theory. New York, McGraw-Hill.)

hydrogen atoms. This behaviour cannot be achieved using a restricted Hartree–Fock wavefunction, which requires the two electrons to occupy the same spatial orbital and leads to H^+ and H^- , but it is appropriately described by a UHF wavefunction. Beyond about 1.2 \AA the ‘correct’ wavefunction for hydrogen must thus be obtained using UHF theory. The results obtained by calculating the potential energy curves of the hydrogen molecule using the RHF and UHF theories are shown in Figure 3.2. As can be seen, RHF theory gives a dissociation energy that is much too large, whereas the UHF theory shows the correct dissociation behaviour.

3.3 Electron Correlation

The most significant drawback of Hartree–Fock theory is that it fails to adequately represent electron correlation. In the self-consistent field method the electrons are assumed to be moving in an average potential of the other electrons, and so the instantaneous position of an electron is not influenced by the presence of a neighbouring electron. In fact, the motions of electrons are correlated and they tend to ‘avoid’ each other more than Hartree–Fock theory would suggest, giving rise to a lower energy. The correlation energy is defined as the difference between the Hartree–Fock energy and the exact energy. Neglecting electron correlation can lead to some clearly anomalous results, especially as the dissociation limit is approached. For example, an uncorrelated calculation would predict that the electrons in H_2 spend equal time on both nuclei, even when they are

infinitely separated. Hartree–Fock geometries and relative energies for equilibrium structures are often in good agreement with experiment and as many molecular modelling applications are concerned with species at equilibrium it might be considered that correlation effects are not so important. Nevertheless, there is increasing evidence that the inclusion of correlation effects is warranted, especially when quantitative information is required. Moreover, electron correlation is crucial in the study of dispersive effects (which we shall consider in Section 4.10.1), which play a major role in intermolecular interactions. Electron correlation is most frequently discussed in the context of *ab initio* calculations, but it should be noted that the effects of electron correlation are implicitly included in the semi-empirical methods because of the way in which they are parametrised. However, specific electron correlation methods have also been developed for use with the various levels of semi-empirical calculation; this in turn necessitates the modification of some parameters.

3.3.1 Configuration Interaction

There are a number of ways in which correlation effects can be incorporated into an *ab initio* molecular orbital calculation. A popular approach is configuration interaction (CI), in which excited states are included in the description of an electronic state. To illustrate the principle, let us consider a lithium atom. The ground state of lithium can be written $1s^22s^1$ (although we have used the conventional nomenclature here, we should remember that the wavefunction is really a Slater determinant). Excitation of the outer valence electron gives states such as $1s^23s^1$. A better description of the overall wavefunction is a linear combination of the ground and excited-state wavefunctions. If a Hartree–Fock calculation is performed with K basis functions then $2K$ spin orbitals are obtained. If these $2K$ spin orbitals are filled with N electrons ($N < 2K$) there will be $2K - N$ unoccupied, virtual orbitals. The wavefunction obtained from the single-determinant approach that we have considered thus far is expressed only in terms of the occupied orbitals. For example, a very simple calculation on H_2 , using as a basis set just the 1s orbitals on each hydrogen, results in two molecular orbitals ($1\sigma_g$ and $1\sigma_u$). In the ground state, the $1\sigma_g$ orbital is filled with two electrons. An excited state can be generated by replacing one or more of the occupied spin orbitals with a virtual spin orbital. Possible excited states for the hydrogen molecule might thus include $1\sigma_g^1\sigma_u^1$ and $1\sigma_u^2$ (in fact, the first of these two configurations cannot be combined with the ground state, as we shall see). In addition to the replacement of single spin orbitals by single virtual orbitals, two spin orbitals can be replaced by two virtual orbitals, three spin orbitals by three virtual orbitals, and so on. In general, the CI wavefunction can be written as:

$$\Psi = c_0 \Psi_0 + c_1 \Psi_1 + c_2 \Psi_2 + \dots \quad (3.6)$$

Ψ_0 is the single-determinant wavefunction obtained by solving the Hartree–Fock equations. Ψ_1 , Ψ_2 , etc. are wavefunctions (expressed as determinants) that represent configurations derived by replacing one or more of the occupied spin orbitals by a virtual spin orbital. The energy of the system is then minimised in order to determine the coefficients c_0 , c_1 , etc., using a linear variational approach, just as for a single-determinant calculation. A CI

calculation thus involves an additional level of complexity; each configuration is written in terms of molecular orbitals, which in turn are expressed as a linear combination of basis functions. The number of integrals can become extremely large. The total number of ways to permute N electrons and K orbitals is $(2K!)/[N!(2K - N)!]$. This is a very large number for all except small values of K and N , which explains why it is not usual to consider all possibilities (termed *full configuration interaction*) except for very small systems. However, full CI is important because it is the most complete treatment possible within the limitations imposed by the basis set. In the limit of a complete basis set full CI becomes complete CI and virtually exact – but is generally considered impractical as at large K the number of Slater determinants increases exponentially with N as $K^N/N!$. It is common practice to limit the excited states considered. For example, in configuration interaction singles (CIS) only wavefunctions that differ from the Hartree–Fock wavefunction by a single spin orbital are included. The next levels of the theory involve double substitutions (configuration interaction doubles, CID) or both singles and double substitutions (configuration interaction singles and doubles, CISD). Even at the CIS or CID levels, the number of excited states to be included can be very large, and it may be desirable (or necessary) to restrict the spin orbitals that are involved in the substitutions. For example, only excitations involving the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO) may be permitted. Alternatively, the orbitals corresponding to the inner electron core may be neglected (the ‘frozen core’ approximation). Some of these options are illustrated in Figure 3.3.

Not all excitations necessarily help to lower the energy; some determinants do not mix with the ground state. A consequence of *Brillouin’s theorem* is that single excitations do not mix

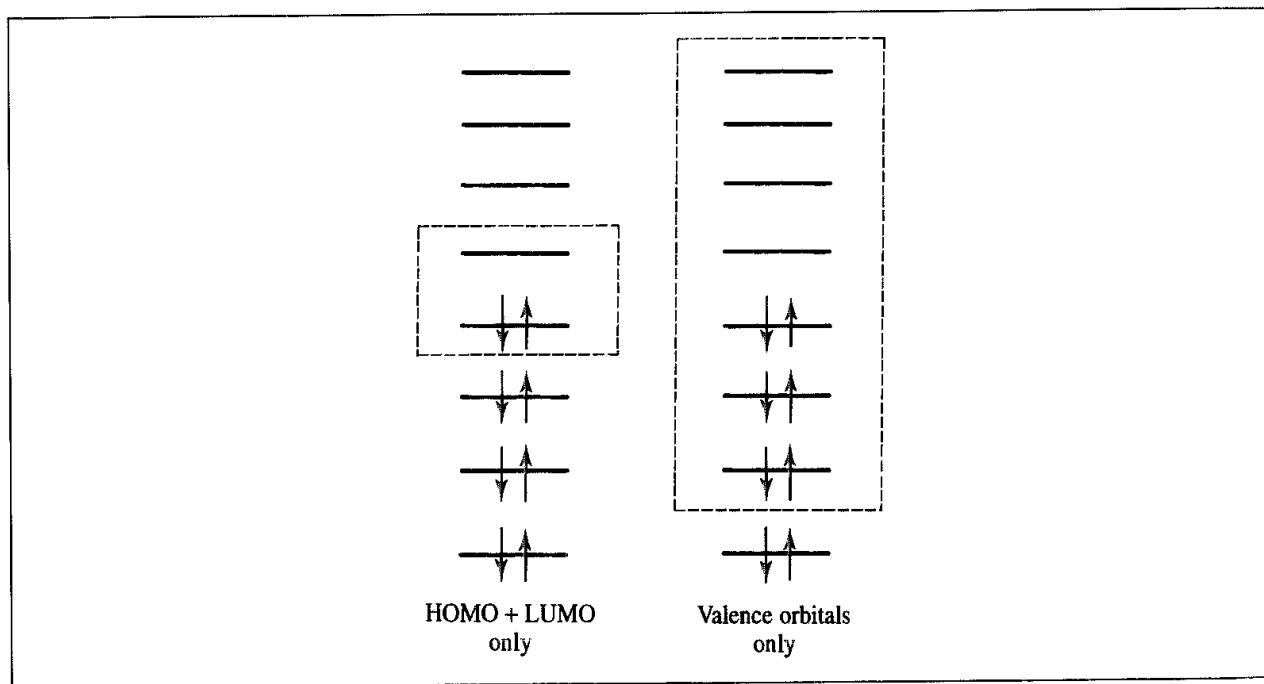


Fig. 3.3 Some of the ways in which excited-state wavefunctions can be included in a configuration interaction calculation (Figure adapted from Hehre W J, L Radom, P v R Schleyer and J A Hehre 1986 Ab initio Molecular Orbital Theory New York, Wiley)

directly with the single-determinant, ground-state wavefunction Ψ_0 . It would therefore be anticipated that double excitations would be most important and that single excitations would have no effect on the energy of the ground state. However, the single excitations can interact with the double excitations, which in turn interact with Ψ_0 , and so single excitations do have a small indirect effect on the energy. The determinants of triple and higher excitations also do not interact directly with Ψ_0 (though they may do indirectly via other levels of excitation). This is because the Hamiltonian contains elements involving at most interactions between pairs of electrons, and so if the Slater determinants differ by more than two electron functions, their integral over all space will be zero.

In a ‘traditional’ CI calculation the determinants in the expansion, Equation (3.6), are those obtained from a Hartree–Fock calculation; only the coefficients c_0 , c_1 , etc. are permitted to vary. Clearly, a better (i.e. lower-energy) wavefunction should be obtained if the coefficients of the basis functions themselves can vary as well as the coefficients of the determinants. This approach is known as the multiconfiguration self-consistent field method (MCSCF). MCSCF theory is considerably more complicated than the Roothaan–Hall equations and well beyond the scope of our discussion. One MCSCF technique that has attracted considerable attention is the complete active-space SCF method (CASSCF) of Roos [Roos *et al.* 1980]. CASSCF enables very large numbers of configurations to be included in the calculation by dividing the molecular orbitals into three sets: those which are doubly occupied in all configurations, those which are unoccupied in all configurations, and then all the remaining ‘active’ orbitals. The list of configurations is generated by considering all possible arrangements of the active electrons among the active orbitals.

A CI calculation is variational: the energy obtained is guaranteed to be greater than the ‘true’ energy. A drawback of CI calculations other than those performed at the full CI level is that they are not size consistent. Simply put, this means that the energy of a number N of non-interacting atoms or molecules is not equal to N times the energy of a single atom or molecule. Another consequence of size consistency is that, as the bond length in a diatomic molecule increases to infinity, so the energy of the system should become equal to the sum of the energies of the respective atoms. To illustrate why this lack of size consistency arises, consider CID calculations on Be_2 and on two beryllium atoms. The electronic configuration of Be is $1s^2 2s^2$ and so if we label the two atoms A and B, then the wavefunction for each of the two separated atoms will include the configuration $1s_A^2 2p_A^2 1s_B^2 2p_B^2$ ($\equiv 1s_A^2 1s_B^2 2p_A^2 2p_B^2$), in which two electrons have been promoted in each beryllium atom from the 2s to the 2p orbitals. This configuration represents a *quadruple* excitation for the beryllium dimer, which has the electronic configuration $1s_A^2 1s_B^2 2s_A^2 2s_B^2$. This quadruply excited configuration is not included in the CID wavefunction for the dimer, which is restricted to double excitations. In fact, the energy of a CI calculation including only doubly excited states is expected to scale in proportion to \sqrt{N} , where N is the number of non-interacting species present, rather than N . The Quadratic Configuration Interaction method (QCISD) was introduced to try to deal with this; it can be considered a size-consistent CISD theory [Pople *et al.* 1987]. The procedure involves the addition of higher excitation terms which are quadratic in their expansion coefficients. Higher still in theory is QCISD(T), in which an estimated contribution from the triple excitations can be incorporated, though with extra computational expense.

3.3.2 Many-body Perturbation Theory

Møller and Plesset proposed an alternative way to tackle the problem of electron correlation [Møller and Plesset 1934]. Their method is based upon Rayleigh–Schrödinger perturbation theory, in which the ‘true’ Hamiltonian operator \mathcal{H} is expressed as the sum of a ‘zeroth-order’ Hamiltonian \mathcal{H}_0 (for which a set of molecular orbitals can be obtained) and a perturbation, \mathcal{V} :

$$\mathcal{H} = \mathcal{H}_0 + \mathcal{V} \quad (3.7)$$

The eigenfunctions of the true Hamiltonian operator are Ψ_i with corresponding energies E_i . The eigenfunctions of the zeroth-order Hamiltonian are written $\Psi_i^{(0)}$ with energies $E_i^{(0)}$. The ground-state wavefunction is thus $\Psi_0^{(0)}$ with energy $E_0^{(0)}$. To devise a scheme by which it is possible to gradually improve the eigenfunctions and eigenvalues of \mathcal{H}_0 we can write the true Hamiltonian as follows:

$$\mathcal{H} = \mathcal{H}_0 + \lambda \mathcal{V} \quad (3.8)$$

λ is a parameter that can vary between 0 and 1; when λ is zero then \mathcal{H} is equal to the zeroth-order Hamiltonian, but when λ is 1 then \mathcal{H} equals its true value. The eigenfunctions Ψ_i and eigenvalues E_i of \mathcal{H} are then expressed in powers of λ :

$$\Psi_i = \Psi_i^{(0)} + \lambda \Psi_i^{(1)} + \lambda^2 \Psi_i^{(2)} + \dots = \sum_{n=0} \lambda^n \Psi_i^{(n)} \quad (3.9)$$

$$E_i = E_i^{(0)} + \lambda E_i^{(1)} + \lambda^2 E_i^{(2)} + \dots = \sum_{n=0} \lambda^n E_i^{(n)} \quad (3.10)$$

$E_i^{(1)}$ is the first-order correction to the energy, $E_i^{(2)}$ is the second-order correction, and so on. These energies can be calculated from the eigenfunctions as follows:

$$E_i^{(0)} = \int \Psi_i^{(0)} \mathcal{H}_0 \Psi_i^{(0)} d\tau \quad (3.11)$$

$$E_i^{(1)} = \int \Psi_i^{(0)} \mathcal{V} \Psi_i^{(0)} d\tau \quad (3.12)$$

$$E_i^{(2)} = \int \Psi_i^{(0)} \mathcal{V} \Psi_i^{(1)} d\tau \quad (3.13)$$

$$E_i^{(3)} = \int \Psi_i^{(0)} \mathcal{V} \Psi_i^{(2)} d\tau \quad (3.14)$$

To determine the corrections to the energy it is therefore necessary to determine the wavefunctions to a given order. In Møller–Plesset perturbation theory the unperturbed Hamiltonian \mathcal{H}_0 is the sum of the one-electron Fock operators for the N electrons:

$$\mathcal{H}_0 = \sum_{i=1}^N \mathcal{F}_i = \sum_{i=1}^N \left(\mathcal{H}^{\text{core}} + \sum_{j=1}^N (\mathcal{J}_i + \mathcal{H}_i) \right) \quad (3.15)$$

The Hartree–Fock wavefunction, $\Psi_0^{(0)}$, is an eigenfunction of \mathcal{H}_0 , and the corresponding zeroth-order energy $E_0^{(0)}$ is equal to the sum of orbital energies for the occupied molecular

orbitals:

$$E_0^{(0)} = \sum_{i=1}^{\text{occupied}} \varepsilon_i \quad (3.16)$$

In order to calculate higher-order wavefunctions we need to establish the form of the perturbation, \mathcal{V} . This is the difference between the ‘real’ Hamiltonian \mathcal{H} and the zeroth-order Hamiltonian, \mathcal{H}_0 . Remember that the Slater determinant description, based on an orbital picture of the molecule, is only an approximation. The true Hamiltonian is equal to the sum of the nuclear attraction terms and electron repulsion terms:

$$\mathcal{H}_0 = \sum_{i=1}^N (\mathcal{H}^{\text{core}}) + \sum_{i=1}^N \sum_{j=i+1}^N \frac{1}{r_{ij}} \quad (3.17)$$

Hence the perturbation \mathcal{V} is given by:

$$\mathcal{V} = \sum_{i=1}^N \sum_{j=i+1}^N \frac{1}{r_{ij}} - \sum_{j=1}^N (\mathcal{J}_j + \mathcal{K}_j) \quad (3.18)$$

The first-order energy $E_0^{(1)}$ is given by:

$$E_0^{(1)} = -\frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \frac{1}{r_{ij}} [(ii|jj) - (ij|ij)] \quad (3.19)$$

The sum of the zeroth-order and first-order energies thus corresponds to the Hartree–Fock energy (compare with Equation (2.110), which gives the equivalent result for a closed-shell system):

$$E_0^{(0)} + E_0^{(1)} = \sum_{i=1}^N \varepsilon_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N [(ii|jj) - (ij|ij)] \quad (3.20)$$

To obtain an improvement on the Hartree–Fock energy it is therefore necessary to use Møller–Plesset perturbation theory to at least second order. This level of theory is referred to as MP2 and involves the integral $\int \Psi_0^{(0)} \mathcal{V} \Psi_0^{(1)} d\tau$. The higher-order wavefunction $\Psi_0^{(1)}$ is expressed as linear combinations of solutions to the zeroth-order Hamiltonian:

$$\Psi_0^{(1)} = \sum_j c_j^{(1)} \Psi_j^{(0)} \quad (3.21)$$

The $\Psi_j^{(0)}$ in Equation (3.21) will include single, double, etc. excitations obtained by promoting electrons into the virtual orbitals obtained from a Hartree–Fock calculation. The second-order energy is given by:

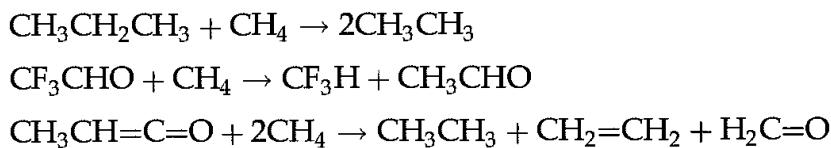
$$E_0^{(2)} = \sum_i^{\text{occupied}} \sum_{j>1}^{\text{virtual}} \sum_a^{\text{virtual}} \sum_{b>a}^{\text{virtual}} \frac{\iint d\tau_1 d\tau_2 \chi_i(1)\chi_j(2) \left(\frac{1}{r_{12}}\right) [\chi_a(1)\chi_b(2) - \chi_b(1)\chi_a(2)]}{\varepsilon_a + \varepsilon_b - \varepsilon_i - \varepsilon_j} \quad (3.22)$$

These integrals will be non-zero only for double excitations, according to the Brillouin theorem. Third- and fourth-order Møller–Plesset calculations (MP3 and MP4) are also

available as standard options in many *ab initio* packages. For the fourth-order calculations single, triple and quadruple excitations will also contribute. As the triple substitutions are most difficult to perform computationally a partial theory that involves just single, double and quadruple substitutions (MP4SDQ) is a popular alternative.

The advantage of many-body perturbation theory is that it is size-independent, unlike configuration interaction – even when a truncated expansion is used. However, Møller-Plesset perturbation theory is not variational and can sometimes give energies that are lower than the ‘true’ energy. Møller-Plesset calculations are computationally intensive and so their use is often restricted to ‘single-point’ calculations at a geometry obtained using a lower level of theory. They are at present the most popular way to incorporate electron correlation into molecular quantum mechanical calculations, especially at the MP2 level. A Møller-Plesset calculation is specified using the level of theory used (e.g. MP2, MP3) together with the basis set. Thus MP2/6-31G* indicates a second-order Møller-Plesset calculation with the 6-31G* basis set.

Certain properties benefit more from the use of correlation methods than others do. For example, a single-determinant Hartree-Fock method and a reasonable basis set give geometrical parameters often very close (bond lengths within 0.01–0.02 Å and angles within 1–2°) to the experimental values. This contrasts with the situation for processes which result in the unpairing of electrons. A simple example is the bond dissociation energy of H₂, for which the Hartree-Fock limit is 84 kcal/mol. MP2, MP3 and MP4 calculations using the 6-31G** basis set give results of 101, 105 and 106 kcal/mol, respectively, for this process, much closer to the experimental value of 109 kcal/mol. In these and similar situations, electron correlation is often advised, if the computational resources permit. However, one class of reactions can be well described using single-determinant Hartree-Fock theory. These are known as *isodesmic reactions*, which are transformations in which the number of electron pairs is constant and the chemical bond types are conserved. Such reactions would be expected to benefit from a judicious cancellation of errors as only the environment of the bonds has changed. Examples of isodesmic reactions are:



Even at the STO-3G level quite respectable results can often be obtained.

In an attempt to deal with some of the shortcomings of even the correlated methods a number of correction factors have been developed. The Gaussian-*n* procedures [Pople *et al.* 1989, Curtiss *et al.* 1991, 1998] represent an attempt to develop a protocol for the accurate calculation of various properties such as atomisation energies, ionisation potentials, electron affinities and proton affinities for atoms and molecules containing first-row and second-row elements. Currently, the most recent member of this series is Gaussian-3 (G3) theory [Curtiss *et al.* 1998]. The G3 method involves a defined sequence of calculations involving geometry optimisation first at the Hartree-Fock level with the 6-31G* basis set and then at the MP2/6-31G* level. A single-point calculation is next carried out using this geometry with the full MP4 method (singles, doubles, triples and quadruples). This energy is then

refined through a series of corrections, which deal with the need for higher polarisation functions, for correlation effects beyond fourth-order perturbation theory (i.e. QCISD(T)) and for larger basis set effects. These correction factors are combined, together with a zero-point energy derived from a series of scaled harmonic frequencies determined from the first, HF/6-31G⁺, geometry optimisation, to give the final G3 energy. When tested on 299 experimental energies the overall average absolute deviation from experiment was 1.02 kcal/mol, with the average deviations for the four different types of data being 0.94 kcal/mol for the enthalpies of formation (148 values), 1.13 kcal/mol for the ionisation energies (85 values), 1.00 kcal/mol for electron affinities (58 values) and 1.34 kcal/mol for proton affinities (eight values). Detailed examination of the results can help to identify systems requiring most attention in subsequent developments of the theory. For example, the enthalpy of formation of both SO₂ and PF₃ have large negative deviations from experiment, perhaps due to the need for a larger basis set to describe the bonding in these molecules. Likewise some of the strained hydrocarbon ring systems (cyclopropene, cyclobutene and bicyclobutane) also show relatively large deviations.

The G3 method is still rather computationally intensive and so some efforts have been made to reduce the computational requirements whilst retaining an acceptable level of error. The G3(MP2) variant [Curtiss *et al.* 1999] replaces the MP4 calculations (which are particularly time-consuming), with comparable calculations at the MP2 level. This leaves the QCISD(T) stage as the most demanding step. The average absolute deviation of the energies calculated using the G3(MP2) method was 1.89 kcal/kmol on the entire 299 test systems, a significantly less accurate result than that of the full G3 method, but still noteworthy.

3.4 Practical Considerations When Performing *ab initio* Calculations

Ab initio calculations can be extremely time-consuming, especially when using the higher levels of theory or when the nuclei are free to move, as in a minimisation calculation (see Chapter 5). Various ‘tricks’ have been developed which can significantly reduce the computational effort involved. Many of these options are routinely available in the major software packages and are invoked by the specification of simple keywords. One common tactic is to combine different levels of theory for the various stages of a calculation. For example, a lower level of theory can be used to provide the initial guess for the density matrix prior to the first SCF iteration. Lower levels of theory can also be used in other ways. Suppose we wish to determine some of the electronic properties of a molecule in a minimum energy structure. Energy minimisation requires that the nuclei move and is typically performed in a series of steps, at each of which the energy (and frequently the gradient of the energy) must be calculated. Minimisation is therefore a computationally expensive procedure, particularly when performed at the high level of theory. To reduce this computational burden a lower level of theory can be employed for the geometry optimisation. A ‘single-point’ calculation using a high level of theory is then performed at the geometry so obtained to give a wavefunction from which the properties are determined. The assumption here of course is that the geometry does not change much between the two levels of

theory. Such calculations are denoted by slashes (/). For example, a calculation that is described as '6-31G*/STO-3G' indicates that the geometry was determined using the STO-3G basis set and the wavefunction was obtained using the 6-31G* basis set. Two slashes are used when each calculation is itself described using a slash, such as when electron correlation methods are used. For example, 'MP2/6-31G*//HF/6-31G*' indicates a geometry optimisation using a Hartree-Fock calculation with a 6-31G* basis set followed by a single-point calculation using the MP2 method for incorporating electron correlation, again using a 6-31G* basis set.

3.4.1 Convergence of Self-consistent Field Calculations

In an SCF calculation the wavefunction is gradually refined until self-consistency is achieved. For closed-shell ground-state molecules this is usually quite straightforward and the energy converges after a few cycles. However, in some cases convergence is a problem, and the energy may oscillate from one iteration to the next or even diverge rapidly. Various methods have been proposed to deal with such situations. A simple strategy is to use an average set of orbital coefficients rather than the set obtained from the immediately preceding iteration. The coefficients in this average set can be weighted according to the energies of each iteration. This tends to weed out those coefficients that give rise to higher energies.

The initial guess of the density matrix may influence the convergence of the SCF calculation; a null matrix is the simplest approach, but better results may be obtained by using a density matrix from a calculation performed at a lower level of theory. For example, the density matrix from a semi-empirical calculation may be used as the starting point for an *ab initio* calculation. Conversely, such an approach may itself lead to problems if there is a significant difference between the density matrices for the lower and higher levels of theory.

A more sophisticated method that is often very successful is Pulay's direct inversion of the iterative subspace (DIIS) [Pulay 1980]. Here, the energy is assumed to vary as a quadratic function of the basis set coefficients. In DIIS the coefficients for the next iteration are calculated from their values in the previous steps. In essence, one is predicting where the minimum in the energy will lie from a knowledge of the points that have been visited and by assuming that the energy surface adopts a parabolic shape.

3.4.2 The Direct SCF Method

An *ab initio* calculation can be logically considered to involve two separate stages. First, the various one- and two-electron integrals are calculated. This is a computationally intensive task and considerable effort has been expended finding ways to make the calculation of the integrals as efficient as possible. In the second stage, the wavefunction is determined using the variation theorem. In a 'traditional' SCF calculation all of the integrals are first calculated and stored on disk, to be retrieved later during the SCF calculation as required. The number of integrals to be stored may run into millions and this inevitably leads to delays in accessing the data, particularly as the retrieval of information from a disk requires

physical movement of the read head and so is slow. Modern computers (both workstations and supercomputers) have much faster (and cheaper) processing units, and many of these machines also have a substantial amount of internal memory, which can be accessed in a fraction of the time it takes to read data from the disk. In a direct SCF calculation, the integrals are not stored on the disk but are kept in memory or recalculated when required [Almlöf *et al.* 1982].

A much-quoted ‘fact’ is that *ab initio* calculations scale as the fourth power of the number of basis functions for ground-state, closed-shell systems. This scaling factor arises because each two-electron integral ($\mu\nu|\lambda\sigma$) involves four basis functions, so the number of two-electron integrals would be expected to increase in proportion to the fourth power of the number of basis functions. In fact, the number of such integrals is not exactly equal to the fourth power of the number of basis functions because many of the integrals are related by symmetry. We can calculate exactly the number of two-electron integrals that are required in a Hartree–Fock *ab initio* calculation as follows. There are seven different types of two-electron integral:

1. $(ab|cd) \equiv (ab|dc) \equiv (ba|cd) \equiv (ba|dc) \equiv (cd|ab) \equiv (cd|ba) \equiv (dc|ab) \equiv (dc|ba)$
2. $(aa|bc) \equiv (aa|cb) \equiv (bc|aa) \equiv (cb|aa)$
3. $(ab|ac) \equiv (ab|ca) \equiv (ba|ac) \equiv (ba|ca) \equiv (ac|ab) \equiv (ac|ba) \equiv (ca|ab) \equiv (ca|ba)$
4. $(aa|bb) \equiv (bb|aa)$
5. $(ab|ab) \equiv (ab|ba) \equiv (ba|ab) \equiv (ba|ba)$
6. $(aa|ab) \equiv (aa|ba) \equiv (ab|aa) \equiv (ba|aa)$
7. $(aa|aa)$

For a basis set with K basis functions, there are $K(K - 1)(K - 2)(K - 3)$ integrals of type $(ab|cd)$, but due to symmetry only one-eighth of these are unique as shown. Similarly, there are $2K(K - 1)(K - 2)$ of type (2); $4K(K - 1)(K - 2)$ of type (3); $K(K - 1)$ of type (4); $2K(K - 1)$ of type (5); $4K(K - 1)$ of type (6) and K of type 7. Thus, a basis set with 200 functions has a total of 202 015 050 unique two-electron integrals. For all but the smallest of basis sets most integrals are of type (1) which is why an *ab initio* problem is often considered to scale as $K^4/8$ ($200^4/8 = 200\,000\,000$). Including electron correlation adds significantly to the computational cost; for example, MP2 calculations scale as the fifth power of the number of basis functions. Electron correlation methods may also require significantly more memory and disk than the comparable SCF calculation; the higher levels scale as the sixth power, and in QCISD(T), one part of the calculation is seventh order.

In practice, *ab initio* calculations often scale as a significantly smaller power than four. It is found that in favourable cases the computational cost of a direct SCF calculation on a large molecule scales as approximately the *square* of the number of basis functions used. This significant reduction (from four to two) is due to several factors. We have already noted some of the ways in which a carefully chosen basis set can reduce the computational effort, for example by making many of the integrals (particularly the two-electron integrals) identical by using the same Gaussian exponents for s and p orbitals in the same shell. Another way in which the calculation time can be significantly reduced is to exploit any symmetry of the system. Many isolated molecules contain symmetry elements such as centres of inversion and mirror planes, information which can be used to reduce the

computational effort required. In the case of an *ab initio* calculation that scales as the fourth power of the number of basis functions then a four-fold reduction in the number of atoms can (in principle at least) result in the computational time being reduced by about 250 times. The most effective way to reduce the computational effort is to identify integrals which are so small that ignoring them (i.e. setting them to zero) will not affect the results. The number of ‘important’ integrals is believed to scale as $K^2 \ln K$. The negligible integrals are determined by calculating an upper limit for each integral. This can be done rapidly and so those integrals that are guaranteed to be negligible can be identified and so ignored. The cutoff value which determines whether an integral is explicitly calculated or is set to zero can vary from one program to another, so it is always useful to check its value if different programs give different results for a given calculation.

3.4.3 Calculating Derivatives of the Energy

Considerable effort has been spent devising efficient ways of directly calculating the first and second derivatives of the energy with respect to the nuclear coordinates. Derivatives are primarily used during minimisation procedures for finding equilibrium structures (the first derivative of the energy with respect to its coordinates equals the force on an atom) and are also used by methods which locate transition structures and determine reaction pathways.

A self-consistent field wavefunction (and thus its energy) can be considered a complicated function of the nuclear coordinates, basis functions and basis function coefficients (and, for a CI calculation, the coefficients of single determinantal wavefunctions). In order to determine the first, second, etc. derivatives of the energy with respect to the nuclear coordinates [Pulay 1977] it is necessary to consider not only how the energy depends directly on the nuclear coordinates but also whether there is an indirect dependence via other parameters. Indeed, it is only the one-electron part of the Hamiltonian that depends directly upon the nuclear coordinates ($H^{\text{core}}(1)$, Equation (2.125)), to which is added an internuclear Coulomb repulsion term. For the other parameters the derivative with respect to the nuclear coordinates is generally determined via the chain rule (for first derivatives). For example, for a generic nuclear coordinate q_i and a generic parameter x_j we can write:

$$\frac{\partial E}{\partial q_i} = \frac{\partial E}{\partial x_j} \frac{\partial x_j}{\partial q_i} \quad (3.23)$$

In Equation (3.23) q_i would be the x , y or z coordinate of an atom and x_j would be a parameter such as a basis function coefficient or a basis function exponent. An important result is that the terms involving variationally determined parameters (such as basis function coefficients) are equal to zero; the energy is a minimum when $(\partial E / \partial c_j)$ is zero. This greatly reduces the computational effort. Most of the numerical work in calculating the gradient is due to the various basis set parameters (e.g. orbital centres and exponents) which require the derivatives of the various electron integrals. For Gaussian basis sets these derivatives can be obtained analytically and indeed it is relatively straightforward to obtain first derivatives for many levels of theory. The time taken to calculate the derivatives is comparable to that required for the calculation of the total energy. Second (and

higher) derivatives are more difficult and expensive to calculate, even at the lower levels of theory.

A possible alternative approach to the calculation of forces is via the use of the Hellmann-Feynman theorem. If Ψ is an exact wavefunction of a Hamiltonian H with energy E then this theorem states that the derivative of E with respect to some parameter P can be written:

$$\frac{\partial E}{\partial P} = \left\langle \frac{\partial H}{\partial P} \right\rangle \quad (3.24)$$

In the case of the derivative with respect to some nuclear coordinate q_i , we would consider the exact force and the Hellmann-Feynman force to be equal:

$$\frac{\partial}{\partial q_i} \langle \Psi | H | \Psi \rangle = \left\langle \Psi \left| \frac{\partial H}{\partial q_i} \right| \Psi \right\rangle \quad (3.25)$$

Unfortunately, this only holds for the exact wavefunction and certain other types of wavefunction (such as at the Hartree-Fock limit). Moreover, even though the Hellmann-Feynman forces are much easier to calculate they are very unreliable, even for accurate wavefunctions, giving rise to spurious forces (often referred to as 'Pulay forces' [Pulay 1987]).

3.4.4 Basis Set Superposition Error

Suppose we wish to calculate the energy of formation of a bimolecular complex, such as the energy of formation of a hydrogen-bonded water dimer. Such complexes are sometimes referred to as 'supermolecules'. One might expect that this energy value could be obtained by first calculating the energy of a single water molecule, then calculating the energy of the dimer, and finally subtracting the energy of the two isolated water molecules (the 'reactants') from that of the dimer (the 'products'). However, the energy difference obtained by such an approach will invariably be an overestimate of the true value. The discrepancy arises from a phenomenon known as *basis set superposition error* (BSSE). As the two water molecules approach each other, the energy of the system falls not only because of the favourable intermolecular interactions but also because the basis functions on each molecule provide a better description of the electronic structure around the other molecule. It is clear that the BSSE would be expected to be particularly significant when small, inadequate basis sets are used (e.g. the minimal basis STO- nG basis sets) which do not provide for an adequate representation of the electron distribution far from the nuclei, particularly in the region where non-covalent interactions are strongest. One way to estimate the basis set superposition error is via the counterpoise correction method of Boys and Bernardi, in which the entire basis set is included in all calculations [Boys and Bernardi 1970]. Thus, in the general case:



$$\Delta E = E(AB) - [E(A) + E(B)] \quad (3.27)$$

The calculation of the energy of the individual species A is performed in the presence of 'ghost' orbitals of B; that is, without the nuclei or electrons of B. A similar calculation is

performed for B using ghost orbitals on A. An alternative approach is to use a basis set in which the orbital exponents and contraction coefficients have been optimised for molecular calculations rather than for atoms. The relevance of the basis set superposition error and its dependence upon the basis set and the level of theory employed (i.e. SCF or with electron correlation) remains a subject of much research.

3.5 Energy Component Analysis

The interaction between atoms and molecules can vary from the weak attraction between a pair of closed-shell atoms (e.g. two rare gas atoms in a molecular beam) to the large energy associated with the formation of a chemical bond. Intermediate between these two extremes are interactions due to hydrogen bonding or electron donor–acceptor processes. In these intermediate cases it is often difficult to determine what factors are important in contributing to the interaction. For example, what can a hydrogen bond be ascribed to?

Morokuma analysis is a method for decomposing the energy change on formation of an intermolecular complex into five components: electrostatic, polarisation, exchange repulsion, charge transfer and mixing [Morokuma 1977]. Suppose we have performed *ab initio* SCF calculations on two molecules, X and Y, and on the intermolecular complex (or ‘supermolecule’) XY. The wavefunctions obtained can be written $A\Psi_X^0$, $A\Psi_Y^0$ and $A\Psi_{XY}^0$. ‘A’ indicates the use of an antisymmetrised wavefunction (e.g. a Slater determinant). The sum of the energies of the isolated molecules is E_0 and the energy of the supermolecule is E_4 (we follow the original notation of Morokuma). The interaction energy ΔE is thus given by $E_4 - E_0$. The five components are calculated as follows.

The electrostatic contribution equals the interaction between the unperturbed electron distributions of the two isolated species, A and B. It is identical to the classical Coulomb interaction and equals the difference $E_1 - E_0$, where E_1 is the energy associated with the product of the two individual wavefunctions, Ψ_1 :

$$\Psi_1 = A\Psi_A^0 A\Psi_B^0 \quad (3\ 28)$$

The electronic distributions of both X and Y will be changed by the presence of the other molecule. These polarisation effects cause a dipole to be induced in (say) molecule Y due to the charge distribution in molecule X and vice versa. Polarisation also affects the higher-order multipoles. To calculate the polarisation contribution we first calculate molecular wavefunctions Ψ_A and Ψ_B in the presence of the other molecule. The energy of the wavefunction Ψ_2 is determined as E_2 , where Ψ_2 is:

$$\Psi_2 = A\Psi_X A\Psi_Y \quad (3\ 29)$$

The polarisation contribution equals $E_2 - E_1$ and is always attractive.

In determining Ψ_1 and Ψ_2 , no electron exchange interactions are considered. The overlap between the electron distributions of X and Y at short range causes a repulsion because to bring together electrons with the same spin into the same region of space ultimately leads to a violation of the Pauli principle.

The exchange repulsion is calculated as $E_3 - E_1$, where E_3 is the energy of the wavefunction Ψ_3 :

$$\Psi_3 = A(\Psi_X^0 \cdot \Psi_Y^0) \quad (3.30)$$

Ψ_3 is derived from the undistorted wavefunctions of X and Y but the exchange of electrons is permitted. The exchange term is always repulsive.

The charge transfer term arises from the transfer of charge (i.e. electrons) from occupied molecular orbitals on one molecule to unoccupied orbitals on the other molecule. This contribution is calculated as the difference between the energy of the supermolecule XY when this charge transfer is specifically allowed to occur, and an analogous calculation in which it is not.

The Morokuma formalism also requires an additional, ‘mixing’ or ‘coupling’ term to be included. This equals the difference between the total SCF difference, ΔE , and the sum of the four contributions (electrostatic, polarisation, exchange repulsion and charge transfer). The mixing term has little physical significance and is used because the four components do not completely account for the entire interaction energy (it is a fudge factor!). Fortunately, it is often relatively small.

Morokuma studied a number of hydrogen-bonded complexes using this scheme in order to assess the contribution from each component. The systems studied were typically of intermolecular complexes involving small molecules such as H₂O, HF and NH₃. In addition, Morokuma and his colleagues also examined a series of electron donor–acceptor complexes such as H₃N-BF₃, OC-BH₃, HF-CIF and benzene-OC(CN)₂. He also studied the basis-set dependence of the results and observed that the energy components were more sensitive than the energy differences. For example, a minimal STO-3G basis set overestimates the charge transfer contribution, whereas double zeta basis sets tend to exaggerate the electrostatic interaction.

3.5.1 Morokuma Analysis of the Water Dimer

The water dimer (H₂O)₂ has been subject to perhaps the closest scrutiny of all hydrogen-bonded complexes. A variety of stable geometries are available to the water dimer, in which one or more hydrogen bonds are present. There has been considerable debate over the relative energies of these structures and even some dispute over which structures are actually at minimum points on the energy surface [Smith *et al.* 1990]. As might be expected, the results depend upon the basis set used. A linear geometry is observed experimentally and is also predicted to be the most stable structure by *ab initio* calculations with a wide variety of basis sets (see Figure 3.4). Using a 6-31G** basis set, Umeyama and Morokuma calculated that the –5.6 kcal/mol stabilisation energy was composed of –7.5 kcal/mol electrostatic stabilisation, 4.3 kcal/mol exchange repulsion, –0.5 kcal/mol polarisation and –1.8 kcal/mol charge transfer [Umeyama and Morokuma 1977]. The ‘mixing term’ contributed –0.1 kcal/mol. Thus the hydrogen bond in the water dimer was considered to arise primarily from electrostatic effects with a smaller charge transfer contribution. Morokuma and Umeyama also extended their analysis of charge transfer to investigate whether this was due to transfer from the proton donor to the acceptor, or from acceptor

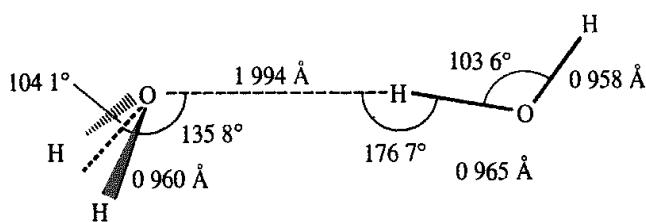


Fig. 3.4 The linear structure of the water dimer [Smith et al 1990].

to donor. The results showed that approximately 90% of the charge transfer resulted from proton acceptor to proton donor transfer.

Morokuma analysis was widely used in the years after its introduction; it is less popular now as some problems have been encountered when trying to interpret the results with the larger basis sets that are feasible with today's faster computers and improved algorithms. In particular, when diffuse basis sets are used then there is a substantial amount of intermolecular overlap even at relatively large distances, which can make it difficult to factor out the different components. Nevertheless, the approach is certainly a useful way to assess the major causes of a particular type of intermolecular interaction, if only to provide a qualitative picture.

3.6 Valence Bond Theories

An entirely different way to treat the electronic structure of molecules is provided by valence bond theory, which was developed at about the same time as the molecular orbital approach. However, valence bond theory was not so amenable to calculations on large molecules, and molecular orbital theory came to dominate electronic structure theory for such systems. Nevertheless, valence bond theories are often considered to be more appropriate for certain types of problem than molecular orbital theory, especially when dealing with processes that involve bonds being broken and/or formed. Recall from Figure 3.2 that a self-consistent field wavefunction gives a wholly inaccurate picture for the dissociation of H_2 ; by contrast, the correct dissociation behaviour is naturally built into valence bond theories.

Valence bond theory is usually introduced using the famous Heitler-London model of the hydrogen molecule [Heitler and London 1927]. This model considers two non-interacting hydrogen atoms (a and b) in their ground states that are separated by a long distance. The wavefunction for this system is:

$$\Psi = \phi_{1sa}(1)\phi_{1sb}(2) \quad (3.31)$$

As the two hydrogen atoms approach to form a hydrogen molecule, such a wavefunction is inappropriate as it implies that electron 1 remains confined to orbital 1sa and electron 2 to orbital 1sb. This clearly violates the indistinguishability principle, and so a linear combination is used

$$\Psi_{vb} \propto \phi_{1sa}(1)\phi_{1sb}(2) + \phi_{1sa}(2)\phi_{1sb}(1) \quad (3.32)$$

The corresponding molecular orbital function for this system is:

$$\Psi_{\text{mo}} \propto \phi_{1s_a}(1)\phi_{1s_b}(2) + \phi_{1s_a}(2)\phi_{1s_b}(1) + \phi_{1s_a}(1)\phi_{1s_a}(2) + \phi_{1s_b}(1)\phi_{1s_b}(2) \quad (3.33)$$

The additional terms in the molecular orbital wavefunction correspond to states with the two electrons in the same orbital, which endows ionic character to the bond (H^+H^-). The valence bond wavefunction does not include any ionic character and in fact it correctly describes the dissociation into two hydrogen atoms. The simple valence bond and molecular orbital pictures in Equations (3.32) and (3.33) are extremes, with the ‘true’ wavefunction being somewhere in the middle. The valence bond representation can be improved by including a degree of ionic character as follows:

$$\Psi_{\text{vb}} \propto \phi_{1s_a}(1)\phi_{1s_b}(2) + \phi_{1s_a}(2)\phi_{1s_b}(1) + \lambda[\phi_{1s_a}(1)\phi_{1s_a}(2) + \phi_{1s_b}(1)\phi_{1s_b}(2)] \quad (3.34)$$

λ is a parameter that can be varied to give the ‘correct’ amount of ionic character. Another way to view the valence bond picture is that the incorporation of ionic character corrects the overemphasis that the valence bond treatment places on electron correlation. The molecular orbital wavefunction underestimates electron correlation and requires methods such as configuration interaction to correct for it. Although the presence of ionic structures in species such as H_2 appears counterintuitive to many chemists, such species are widely used to explain certain other phenomena such as the ortho/para or meta directing properties of substituted benzene compounds under electrophilic attack. Moreover, it has been shown that the ionic structures correspond to the deformation of the atomic orbitals when they are involved in chemical bonds.

One widely used valence bond theory is the generalised valence bond (GVB) method of Goddard and co-workers [Bobrowicz and Goddard 1977]. In the simple Heitler-London treatment of the hydrogen molecule the two orbitals are the non-orthogonal atomic orbitals on the two hydrogen atoms. In the GVB theory the analogous wavefunction is written:

$$\Psi_{\text{GVB}} \propto u(1)v(2) + u(2)v(1) \quad (3.35)$$

u and v are non-orthogonal orbitals that are each expressed as a basis set expansion with the coefficients being variationally optimised to minimise the energy. The construction of the wavefunction from orbitals that are not necessarily orthogonal is characteristic of many valence bond theories and complicates the computational problem. The GVB approach is particularly successful for describing the electronic nature of systems as they approach dissociation.

Another approach is spin-coupled valence bond theory, which divides the electrons into two sets: ‘core’ electrons, which are described by doubly occupied orthogonal orbitals, and ‘active’ electrons, which occupy singly occupied non-orthogonal orbitals. Both types of orbital are expressed in the usual way as a linear combination of basis functions. The overall wavefunction is completed by two spin functions; one that describes the coupling of the spins of the core electrons and one that deals with the active electrons. The choice of spin function for these active electrons is a key component of the theory [Gerratt *et al.* 1997]. One of the distinctive features of this theory is that a considerable amount of chemically significant electronic correlation is incorporated into the wavefunction, giving an accuracy comparable to CASSCF. An additional benefit is that the orbitals tend to be

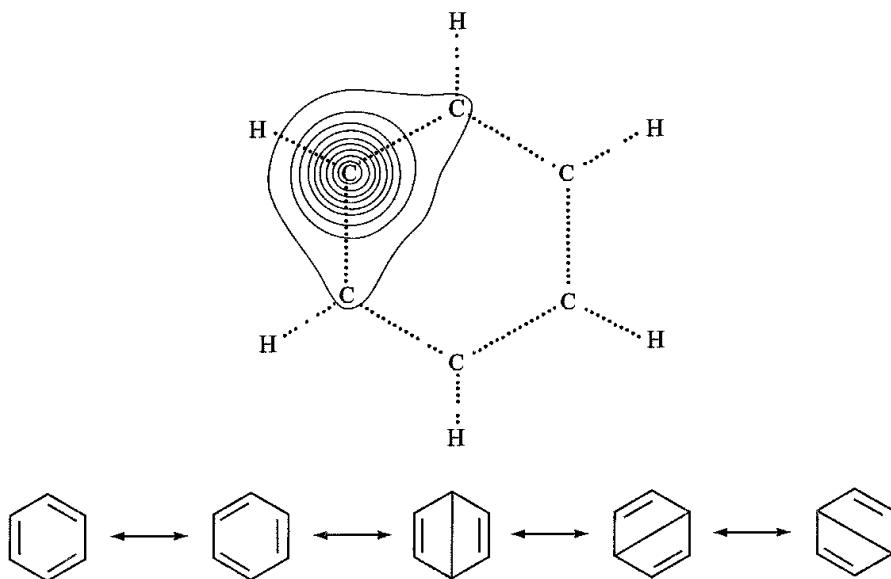


Fig 3.5: π orbital for benzene obtained from spin-coupled valence bond theory (Figure redrawn from Gerratt J, D L Cooper, P B Karadakov and M Raimondi 1997 Modern valence bond theory Chemical Society Reviews 87 100) The figure also shows the two Kekulé and three Dewar benzene forms which contribute to the overall wavefunction; each Kekulé form contributes approximately 40.5% and each Dewar form approximately 6.4%

localised, closely resembling atomic or hybrid atomic orbitals, and consequently very visual. Various chemical phenomena have been examined using this approach, including dissociation reactions and hypervalence. One particularly interesting study was of the π system of benzene [Cooper *et al.* 1986]. This calculation resulted in six orbitals, each localised on one of the carbon atoms in the ring, though with some deformations towards neighbouring atoms (Figure 3.5). Moreover, the spin-coupling patterns suggested that the bonding was more akin to the Kekulé picture of benzene (with alternating double and single bonds) together with small contributions from Dewar benzene rather than the completely delocalised representation from molecular orbital theory.

3.7 Density Functional Theory

Density functional theory (DFT) is an approach to the electronic structure of atoms and molecules which has enjoyed a dramatic surge of interest since the late 1980s and 1990s [Parr 1983; Wimmer 1997]. Our approach here will be to introduce the key elements of the theory and to identify the similarities and differences between DFT and the Hartree-Fock approach. In Hartree-Fock theory the multi-electron wavefunction is expressed as a Slater determinant which is constructed from a set of N single-electron wavefunctions (N being the number of electrons in the molecule). DFT also considers single-electron functions. However, whereas Hartree-Fock theory does indeed calculate the full N -electron wavefunction, density functional theory only attempts to calculate the total electronic energy and the overall electronic density distribution. The central idea underpinning DFT is that

there is a relationship between the total electronic energy and the overall electronic density. This is not a particularly new idea; indeed an approximate model developed in the late 1920s (the Thomas–Fermi model) contains some of the basic elements. However, the real breakthrough came with a paper by Hohenberg and Kohn in 1964 [Hohenberg and Kohn 1964], who showed that the ground-state energy and other properties of a system were uniquely defined by the electron density. This is sometimes expressed by stating that the energy, E , is a unique *functional* of $\rho(\mathbf{r})$. A functional enables a function to be mapped to a number and is usually written using square brackets. Thus:

$$Q[f(\mathbf{r})] = \int f(\mathbf{r}) d\mathbf{r} \quad (3.36)$$

The function $f(\mathbf{r})$ is usually dependent upon other well-defined functions. A simple example of a functional would be the area under a curve, which takes a function $f(\mathbf{r})$ defining the curve between two points and returns a number (the area, in this case). In the case of DFT the function depends upon the electron density, which would make Q a functional of $\rho(\mathbf{r})$, in the simplest case $f(\mathbf{r})$ would be equivalent to the density (i.e. $f(\mathbf{r}) \equiv \rho(\mathbf{r})$). If the function $f(\mathbf{r})$ were to depend in some way upon the gradients (or higher derivatives) of $\rho(\mathbf{r})$ then the functional is referred to as being ‘non-local’, or ‘gradient-corrected’. By contrast, a ‘local’ functional would only have a simple dependence upon $\rho(\mathbf{r})$. In DFT the energy functional is written as a sum of two terms:

$$E[\rho(\mathbf{r})] = \int V_{\text{ext}}(\mathbf{r})\rho(\mathbf{r}) d\mathbf{r} + F[\rho(\mathbf{r})] \quad (3.37)$$

The first term arises from the interaction of the electrons with an external potential $V_{\text{ext}}(\mathbf{r})$ (typically due to the Coulomb interaction with the nuclei). $F[\rho(\mathbf{r})]$ is the sum of the kinetic energy of the electrons and the contribution from interelectronic interactions. The minimum value in the energy corresponds to the exact ground-state electron density, so enabling a variational approach to be used (i.e. the ‘best’ solution corresponds to the minimum of energy and an incorrect density gives an energy above the true energy). There is a constraint on the electron density as the number of electrons (N) is fixed:

$$N = \int \rho(\mathbf{r}) d\mathbf{r} \quad (3.38)$$

In order to minimise the energy we introduce this constraint as a Lagrangian multiplier ($-\mu$), leading to:

$$\frac{\delta}{\delta\rho(\mathbf{r})} \left[E[\rho(\mathbf{r})] - \mu \int \rho(\mathbf{r}) d\mathbf{r} \right] = 0 \quad (3.39)$$

From this we can write:

$$\left(\frac{\delta E[\rho(\mathbf{r})]}{\delta\rho(\mathbf{r})} \right)_{V_{\text{ext}}} = \mu \quad (3.40)$$

Equation (3.40) is the DFT equivalent of the Schrodinger equation. The subscript V_{ext} indicates that this is under conditions of constant external potential (i.e. fixed nuclear positions). It is interesting to note that the Lagrange multiplier, μ , can be identified with the chemical potential of an electron cloud for its nuclei, which in turn is related to the

electronegativity, χ :

$$-\chi = \mu = \left(\frac{\partial E}{\partial N} \right)_{V_{\text{ext}}} \quad (3.41)$$

The second landmark paper in the development of density functional theory was by Kohn^{*} and Sham who suggested a practical way to solve the Hohenberg–Kohn theorem for a set of interacting electrons [Kohn and Sham 1965]. The difficulty with Equation (3.37) is that we do not know what the function $F[\rho(\mathbf{r})]$ is. Kohn and Sham suggested that $F[\rho(\mathbf{r})]$ should be approximated as the sum of three terms:

$$F[\rho(\mathbf{r})] = E_{\text{KE}}[\rho(\mathbf{r})] + E_{\text{H}}[\rho(\mathbf{r})] + E_{\text{XC}}[\rho(\mathbf{r})] \quad (3.42)$$

where $E_{\text{KE}}[\rho(\mathbf{r})]$ is the kinetic energy, $E_{\text{H}}[\rho(\mathbf{r})]$ is the electron–electron Coulombic energy, and $E_{\text{XC}}[\rho(\mathbf{r})]$ contains contributions from exchange and correlation. It is important to note that the first term in Equation (3.42), $E_{\text{KE}}[\rho(\mathbf{r})]$, is defined as the kinetic energy of a system of *non-interacting* electrons with the same density $\rho(\mathbf{r})$ as the real system:

$$E_{\text{KE}}[\rho(\mathbf{r})] = \sum_{i=1}^N \int \psi_i(\mathbf{r}) \left(-\frac{\nabla^2}{2} \right) \psi_i(\mathbf{r}) d\mathbf{r} \quad (3.43)$$

The second term, $E_{\text{H}}(\rho)$, is also known as the Hartree electrostatic energy. The Hartree approach to solving the Schrödinger equation was introduced briefly in Section 2.3.3 and almost immediately dismissed because it fails to recognise that electronic motions are correlated. In the Hartree approach this electrostatic energy arises from the classical interaction between two charge densities, which, when summed over all possible pairwise interactions, gives:

$$E_{\text{H}}[\rho(\mathbf{r})] = \frac{1}{2} \iint \frac{\rho(\mathbf{r}_1)\rho(\mathbf{r}_2)}{|\mathbf{r}_1 - \mathbf{r}_2|} d\mathbf{r}_1 d\mathbf{r}_2 \quad (3.44)$$

Combining these two and adding the electron–nuclear interaction leads to the full expression for the energy of an N -electron system within the Kohn–Sham scheme:

$$\begin{aligned} E[\rho(\mathbf{r})] = & \sum_{i=1}^N \int \psi_i(\mathbf{r}) \left(-\frac{\nabla^2}{2} \right) \psi_i(\mathbf{r}) d\mathbf{r} + \frac{1}{2} \iint \frac{\rho(\mathbf{r}_1)\rho(\mathbf{r}_2)}{|\mathbf{r}_1 - \mathbf{r}_2|} d\mathbf{r}_1 d\mathbf{r}_2 + E_{\text{XC}}[\rho(\mathbf{r})] \\ & - \sum_{A=1}^M \int \frac{Z_A}{|\mathbf{r} - \mathbf{R}_A|} \rho(\mathbf{r}) d\mathbf{r} \end{aligned} \quad (3.45)$$

This equation acts to *define* the exchange-correlation energy functional $E_{\text{XC}}[\rho(\mathbf{r})]$, which thus contains not only contributions due to exchange and correlation but also a contribution due to the difference between the true kinetic energy of the system and $E_{\text{KE}}[\rho(\mathbf{r})]$.

^{*} Walter Kohn, whose name appears on the two key papers which provided the impetus for the development of ‘modern’ density functional theory, was awarded the Nobel Prize for Chemistry in 1998, jointly with John Pople

Kohn and Sham wrote the density $\rho(\mathbf{r})$ of the system as the sum of the square moduli of a set of one-electron orthonormal orbitals:

$$\rho(\mathbf{r}) = \sum_{i=1}^N |\psi_i(\mathbf{r})|^2 \quad (3.46)$$

By introducing this expression for the electron density and applying the appropriate variational condition the following one-electron Kohn-Sham equations result:

$$\left\{ -\frac{\nabla_1^2}{2} - \left(\sum_{A=1}^M \frac{Z_A}{r_{1A}} \right) + \int \frac{\rho(\mathbf{r}_2)}{r_{12}} d\mathbf{r}_2 + V_{XC}[\mathbf{r}_1] \right\} \psi_i(\mathbf{r}_1) = \varepsilon_i \psi_i(\mathbf{r}_1) \quad (3.47)$$

In Equation (3.47) we have written the external potential in the form appropriate to the interaction with M nuclei. ε_i are the orbital energies and V_{XC} is known as the exchange-correlation functional, related to the exchange-correlation energy by:

$$V_{XC}[\mathbf{r}] = \left(\frac{\delta E_{XC}[\rho(\mathbf{r})]}{\delta \rho(\mathbf{r})} \right) \quad (3.48)$$

The total electronic energy is then calculated from Equation (3.45).

To solve the Kohn-Sham equations a self-consistent approach is taken. An initial guess of the density is fed into Equation (3.47) from which a set of orbitals can be derived, leading to an improved value for the density, which is then used in the second iteration, and so on until convergence is achieved.

3.7.1 Spin-polarised Density Functional Theory

Local spin density functional theory (LSDFT) is an extension of ‘regular’ DFT in the same way that restricted and unrestricted Hartree-Fock extensions were developed to deal with systems containing unpaired electrons. In this theory both the electron density and the spin density are fundamental quantities with the net spin density being the difference between the density of up-spin and down-spin electrons:

$$\sigma(\mathbf{r}) = \rho_\uparrow(\mathbf{r}) - \rho_\downarrow(\mathbf{r}) \quad (3.49)$$

The total electron density is just the sum of the densities for the two types of electron. The exchange-correlation functional is typically different for the two cases, leading to a set of spin-polarised Kohn-Sham equations:

$$\left\{ -\frac{\nabla_1^2}{2} - \left(\sum_{A=1}^M \frac{Z_A}{r_{1A}} \right) + \int \frac{\rho(\mathbf{r}_2)}{r_{12}} d\mathbf{r}_2 + V_{XC}[\mathbf{r}_1, \sigma] \right\} \psi_i^\sigma(\mathbf{r}_1) = \varepsilon_i^\sigma \psi_i^\sigma(\mathbf{r}_1) \quad \sigma = \alpha, \beta \quad (3.50)$$

This leads to two sets of wavefunctions, one for each spin, similar to UHF theory.

3.7.2 The Exchange-correlation Functional

The exchange-correlation functional is clearly key to the success (or otherwise) of the density functional approach. One reason why DFT is so appealing is that even relatively simple

approximations to the exchange-correlation functional can give favourable results. The simplest way to obtain this contribution uses the so-called *local density approximation* (LDA; the acronym LSDA is also used, for local spin density approximation), which is based upon a model called the uniform electron gas, in which the electron density is constant throughout all space. The total exchange-correlation energy, E_{XC} , for our system can then be obtained by integrating over all space:

$$E_{XC}[\rho(\mathbf{r})] = \int \rho(\mathbf{r}) \varepsilon_{XC}(\rho(\mathbf{r})) d\mathbf{r} \quad (3.51)$$

$\varepsilon_{XC}(\rho(\mathbf{r}))$ is the exchange-correlation energy per electron as a function of the density in the uniform electron gas. The exchange-correlation functional is obtained by differentiation of this expression:

$$V_{XC}[\mathbf{r}] = \rho(\mathbf{r}) \frac{d\varepsilon_{XC}(\rho(\mathbf{r}))}{d\rho(\mathbf{r})} + \varepsilon_{XC}(\rho(\mathbf{r})) \quad (3.52)$$

In the local density approximation it is assumed that at each point \mathbf{r} in the inhomogeneous electron distribution (i.e. in the system of interest) where the density is $\rho(\mathbf{r})$ then $V_{XC}[\rho(\mathbf{r})]$ and $\varepsilon_{XC}(\rho(\mathbf{r}))$ have the same values as in the homogeneous electron gas. In other words, the real electron density surrounding a volume element at position \mathbf{r} is replaced by a constant electron density with the same value as at \mathbf{r} . However, this ‘constant’ electron density is different for each point in space (Figure 3.6).

The exchange-correlation energy per electron (i.e. the energy density) of the uniform electron gas is known accurately for all densities of practical interest from various approaches such as quantum Monte Carlo methods [Ceperley and Alder 1980]. In order to be of practical use this exchange-correlation energy density is then expressed in an analytical form that makes it amenable to computation. It is usual to express $\varepsilon_{XC}[\rho(\mathbf{r})]$ as an analytical function of the electron density and to consider the exchange and correlation contributions separately. However, some analytical expressions for the combined exchange and correlation energy density do exist, such as the following expression of Gunnarsson and

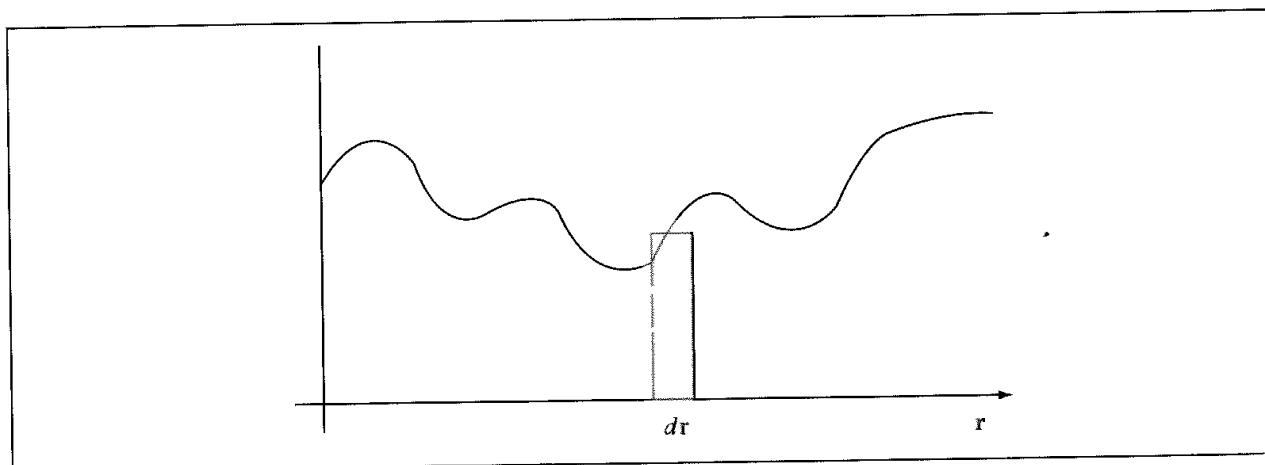


Fig. 3.6 Schematic representation of the way in which the local density approximation assumes that the electron density within a volume element dr surrounding a point \mathbf{r} is assumed to be constant.

Lundqvist [Gunnarsson and Lundqvist 1976]:

$$\begin{aligned}\varepsilon_{XC}(\rho(\mathbf{r})) &= -\frac{0.458}{r_s} - 0.0666G\left(\frac{r_s}{11.4}\right); \\ G(x) &= \frac{1}{2} \left[(1+x) \log(1+x^{-1}) - x^2 + \frac{x}{2} - \frac{1}{3} \right], \quad r_s^3 = \frac{3}{4\pi\rho(\mathbf{r})}\end{aligned}\quad (3.53)$$

The following relatively simple expression is commonly used for the exchange-only energy under the local density approximation [Slater 1974]:

$$E_X[\rho_\alpha(\mathbf{r}), \rho_\beta(\mathbf{r})] = -\frac{3}{2} \left(\frac{3}{4\pi} \right)^{1/3} \int (\rho_\alpha^{4/3}(\mathbf{r}) + \rho_\beta^{4/3}(\mathbf{r})) d\mathbf{r} \quad (3.54)$$

where α and β represent up and down spins. In general, more attention has been paid to the correlation contribution, for which there is no such simple functional form. Perdew and Zunger suggested the following parametric relationship for the correlation contribution [Perdew and Zunger 1981]:

$$\varepsilon_C(\rho(\mathbf{r})) = \begin{cases} -0.1423/(1 + 1.9529r_s^{1/2} + 0.3334r_s) & r_s \geq 1 \\ -0.0480 + 0.0311 \ln r_s - 0.0116r_s + 0.0020r_s \ln r_s & r_s < 1 \end{cases} \quad (3.55)$$

This result applies when the number of up spins equals the number of down spins and so is not applicable to systems with an odd number of electrons. The correlation energy functional was also considered by Vosko, Wilk and Nusair [Vosko *et al.* 1980], whose expression is:

$$\begin{aligned}\varepsilon_C(\rho(\mathbf{r})) &= \frac{A}{2} \left\{ \ln \frac{x^2}{X(x)} + \frac{2b}{Q} \tan^{-1} \frac{Q}{2x+b} - \frac{bx_0}{X(x_0)} \left[\ln \frac{(x-x_0)^2}{X(x)} + \frac{2(b+2x_0)}{Q} \tan^{-1} \frac{Q}{2x+b} \right] \right\} \\ x &= r_s^{1/2}, \quad X(x) = x^2 + bx + c, \quad Q = (4c - b^2)^{1/2}; \\ A &= 0.062\,1814, \quad x_0 = -0.409\,286, \quad b = 13.0720, \quad c = 42.7198\end{aligned}\quad (3.56)$$

In addition to the energy terms for the exchange-correlation contribution (which enables the total energy to be determined) it is necessary to have corresponding terms for the potential, $V_{XC}[\rho(\mathbf{r})]$, which are used to solve the Kohn-Sham equations. These are obtained as the appropriate first derivatives using Equation (3.52).

To solve the Kohn-Sham equations a number of different approaches and strategies have been proposed. One important way in which these can differ is in the choice of basis set for expanding the Kohn-Sham orbitals. In most (but not all) DFT programs for calculating the properties of molecular systems (rather than for solid-state materials) the Kohn-Sham orbitals are expressed as a linear combination of atomic-centred basis functions:

$$\psi_i(\mathbf{r}) = \sum_{\nu=1}^K c_{\nu i} \phi_{\nu} \quad (3.57)$$

Several functional forms have been investigated for the basis functions ϕ_{ν} . Given the vast experience of using Gaussian functions in Hartree-Fock theory it will come as no surprise to learn that such functions have also been employed in density functional theory. However, these are not the only possibility: Slater type orbitals are also used, as are *numerical*

basis functions. We encountered Slater type orbitals in Chapter 2, but the notion of a numerical basis function is new. A numerical basis function can be generated by solving the Kohn-Sham equations for isolated atoms. This gives a set of values on a spherical polar grid centred on each atom. The variation at each grid point can be stored as a cubic spline function so enabling analytical gradients to be calculated. One advantage of a numerical basis set (if properly derived) is that it has the correct nodal behaviour close to the nucleus together with an exponential decay.

More than one function may be used to represent a particular atomic orbital. This is obviously a well-understood tactic when using Gaussian functions, but the use of basis set contractions also applies to the Slater type orbitals and the numerical basis sets. For a numerical basis set the ‘contraction’ can be derived from two functions, one corresponding to the neutral atom and the other to a positive ion.

If the basis set expansion for the Kohn-Sham orbitals in Equation (3.57) is substituted into the Kohn-Sham equations then it is possible to express them in a matrix form, identical in form to the Roothaan–Hall equations:

$$\mathbf{H}\mathbf{C} = \mathbf{S}\mathbf{C}\mathbf{E} \quad (3.58)$$

In this matrix equation the elements of the Kohn-Sham matrix \mathbf{H} are given by:

$$H_{\mu\nu} = \int d\mathbf{r}_1 \phi_\mu(\mathbf{r}_1) \left\{ -\frac{\nabla_1^2}{2} - \left(\sum_{A=1}^M \frac{Z_A}{r_{1A}} \right) + \int \frac{\rho(\mathbf{r}_2)}{r_{12}} d\mathbf{r}_2 + V_{XC}[\mathbf{r}_1] \right\} \phi_\nu(\mathbf{r}_1) \quad (3.59)$$

The first two terms are straightforward and are equal to the core contribution, $H_{\mu\nu}^{\text{core}}$. The Coulomb repulsion contribution (the Hartree term) can be expanded in terms of the basis functions and the density matrix, \mathbf{P} :

$$\iint \frac{\phi_\mu(\mathbf{r}_1)\rho(\mathbf{r}_2)\phi_\nu(\mathbf{r}_1)}{|\mathbf{r}_1 - \mathbf{r}_2|} d\mathbf{r}_1 d\mathbf{r}_2 = \sum_{\lambda=1}^K \sum_{\sigma=1}^K P_{\lambda\sigma} \iint \frac{\phi_\mu(\mathbf{r}_1)\phi_\nu(\mathbf{r}_1)\phi_\lambda(\mathbf{r}_2)\phi_\sigma(\mathbf{r}_2)}{|\mathbf{r}_1 - \mathbf{r}_2|} d\mathbf{r}_1 d\mathbf{r}_2 \quad (3.60)$$

For a closed-shell system with N electrons the elements of the density matrix are given by:

$$P_{\mu\nu} = 2 \sum_{i=1}^{N/2} c_{\mu i} c_{\nu i} \quad (3.61)$$

This is just the same as for the Roothaan–Hall approach to Hartree–Fock theory. The overlap matrix, \mathbf{S} , is defined similarly:

$$S_{\mu\nu} = \int \phi_\mu(\mathbf{r})\phi_\nu(\mathbf{r}) d\mathbf{r} \quad (3.62)$$

The overall procedure to achieve self-consistency is very reminiscent of that used in Hartree–Fock theory, involving first an initial guess of the density by superimposing atomic densities, construction of the Kohn–Sham and overlap matrices, and diagonalisation to give the eigenfunctions and eigenvectors from which the Kohn–Sham orbitals* can be

* It is important to note that the Kohn–Sham orbitals used in density functional theory are a set of non-interacting orbitals designed to give the correct density and have no physical meaning beyond that, unlike the orbitals used in Hartree–Fock theory

constructed and thus the density for the next iteration. This cycle continues until convergence is achieved.

The appearance of the four-centre integrals in Equation (3.60) might lead one to question the advantage of the DFT approach, at least as far as computational efficiency is concerned. Whilst these integrals can certainly be tackled using the same techniques as in Hartree-Fock theory, it is also viable in density functional theory to avoid having to calculate them by considering the left-hand side of Equation (3.60). There are two basic ways to do this. First, one can approximate the charge density by another basis set expansion:

$$\rho(\mathbf{r}) \approx \sum_k c_k \phi'_k(\mathbf{r}) \quad (3.63)$$

These auxiliary basis functions ϕ' have the same functional form as the orbital expansion and the coefficients c_k are obtained by a least-squares fitting procedure. Substituting for the density in the four-centre integrals gives a computationally less demanding three-centre, two-electron integral:

$$\iint \frac{\phi_\mu(\mathbf{r}_1)\phi_\nu(\mathbf{r}_1)\phi_\lambda(\mathbf{r}_2)\phi_\sigma(\mathbf{r}_2)}{|\mathbf{r}_1 - \mathbf{r}_2|} d\mathbf{r}_1 d\mathbf{r}_2 = \iint \frac{\phi_\mu(\mathbf{r}_1)\phi_\nu(\mathbf{r}_1)\phi'_k(\mathbf{r}_2)}{|\mathbf{r}_1 - \mathbf{r}_2|} d\mathbf{r}_1 d\mathbf{r}_2 \quad (3.64)$$

The second approach focuses on the Coulomb integral and uses Poisson's equation. Let us introduce $V_{\text{el}}(\mathbf{r}_1)$:

$$V_{\text{el}}(\mathbf{r}_1) = \int \frac{\rho(\mathbf{r}_2)}{|\mathbf{r}_1 - \mathbf{r}_2|} d\mathbf{r}_2 \quad (3.65)$$

Poisson's equation relates the second derivative of the electric potential to the charge density:

$$\nabla^2 V(\mathbf{r}) = -4\pi\rho(\mathbf{r}) \quad (3.66)$$

We can thus write:

$$\nabla^2 \int \frac{\rho(\mathbf{r}_2)}{|\mathbf{r}_1 - \mathbf{r}_2|} d\mathbf{r}_2 = -4\pi\rho(\mathbf{r}_1) \quad (3.67)$$

This equation can be solved numerically on a grid to determine $V_{\text{el}}(\mathbf{r}_1)$. The same grid is then used to numerically integrate the four-centre, two-electron integral, Equation (3.60), as follows:

$$\iint \frac{\phi_\mu(\mathbf{r}_1)\rho(\mathbf{r}_2)\phi_\nu(\mathbf{r}_1)}{|\mathbf{r}_1 - \mathbf{r}_2|} d\mathbf{r}_1 d\mathbf{r}_2 \equiv \int \phi_\mu(\mathbf{r}_1) V_{\text{el}}(\mathbf{r}_1) \phi_\nu(\mathbf{r}_1) \approx \sum_{i=1}^P \phi_\mu(\mathbf{R}_i) V_{\text{el}}(\mathbf{R}_i) \phi_\nu(\mathbf{R}_i) W_i \quad (3.68)$$

In this equation the P points \mathbf{R}_i correspond to the grid used to solve the Poisson equation for V_{el} and W_i are weighting factors.

It might be wondered why these two simplifications for the four-centre, two-electron integrals can be used in density functional theory and not in Hartree-Fock theory. The reason is that the exchange contribution in Hartree-Fock theory is not a function that can be simplified (technically, it is a non-local functional), in contrast to the situation in

density functional theory. As the four-centre integrals must therefore still be determined for the exchange component in Hartree–Fock theory there is nothing to be gained from simplifying the corresponding Coulomb term.

The exchange-correlation contribution to the Kohn–Sham matrix elements (the final term in Equation (3.59)) is invariably evaluated using a grid of points. This is a consequence of the complexity of the functionals employed. The integration may then be performed using the grid directly or by fitting a further auxiliary basis set expansion with which analytical integration can be used. If a DFT program uses a basis set containing K functions and employs either a grid-based integration scheme with P points or an auxiliary basis set with P functions then the computational complexity of the calculation scales as K^2P . As P is often linearly related to K , density functional theory is often said to scale as the cube of the number of basis functions, K^3 . This contrasts with the fourth-power scaling for conventional Hartree–Fock calculations. However, many practical density functional calculations with a well-engineered computer program do not scale as the simple third power, just as practical Hartree–Fock calculations do not scale as the fourth power; these oft-quoted statements apply only to the most naïve implementations or for calculations on very small, test systems where integral neglect thresholds are not employed.

Whilst most of the programs which use density functional theory for molecular calculations employ one of the three types of basis set described thus far, there are two important alternatives to this approach. The first of these involves the solution of the Kohn–Sham equations numerically (on a grid) using what is sometimes referred to as a ‘basis-set free’ approach [Becke and Dickson 1990]. Such an approach is thus free from the limitations of a finite basis set expansion (provided, of course, that sufficient grid points are employed!) and can be used to evaluate different exchange-correlation functionals, as these represent the only remaining source of error. The second alternative is particularly important for the study of bulk systems such as metals and alloys and involves the use of *plane waves*. This approach will be discussed later in this chapter when we consider the general problem of using quantum mechanics to study the solid state.

3.7.3 Beyond the Local Density Approximation: Gradient-corrected Functionals

The most important feature of density functional theory is probably the way in which it directly incorporates exchange and correlation effects; the latter in particular are only truly considered in the more complex, post-Hartree–Fock approaches such as configuration interaction or many-body perturbation theory. Despite its simplicity the local density approximation performs surprisingly well. However, the local density approximation has been shown to be clearly inadequate for some problems and for this reason extensions have been developed. The most common method is to use gradient-corrected, ‘non-local’ functionals which depend upon the gradient of the density at each point in space and not just on its value. These gradient corrections are typically divided into separate exchange and correlation contributions. A variety of gradient corrections have been proposed in the literature. The gradient correction to the exchange functional proposed by Becke is popular [Becke 1988, 1992]; this corrects

the local spin density approximation result as follows:

$$E_X[\rho(\mathbf{r})] = E_X^{\text{LSDA}}[\rho(\mathbf{r})] - b \sum_{\sigma=\alpha,\beta} \int \rho_{\sigma}^{4/3} \frac{x_{\sigma}^2}{(1 + 6bx_{\sigma} \sinh^{-1} x_{\sigma})} d\mathbf{r}; \quad x_{\sigma} = \frac{|\nabla \rho_{\sigma}|}{\rho_{\sigma}^{4/3}} \quad (3.69)$$

$E_X^{\text{LSDA}}[\rho(\mathbf{r})]$ is the standard Slater form of the exchange energy, Equation (3.54). The form written in Equation (3.69) is for a spin-unrestricted system, from which the appropriate expression for a closed-shell system is easily derived. x_{σ} is a dimensionless parameter and b is constant with a value of 0.0042 a.u. The value of b was determined by fitting to exact exchange Hartree–Fock energies for the noble gas atoms helium to radon. Two particular features of this functional form are that in the limit $r \rightarrow \infty$ the limiting form of the exchange-correlation integral is correctly achieved and that it uses just a single parameter, b . The correlation functional of Lee, Yang and Parr is also widely used [Lee *et al.* 1988]; in its original form it was expressed as follows (for a closed-shell system):

$$E_C[\rho(\mathbf{r})] = -a \int \frac{1}{1 + d\rho^{-1/3}} \{ r + b\rho^{-2/3} [C_F \rho^{5/3} - 2t_W + (\frac{1}{9}t_W + \frac{1}{18}\nabla^2 \rho)e^{-cr^{-1/3}}] \} d\mathbf{r} \quad (3.70)$$

$$t_W(\mathbf{r}) = \sum_{i=1}^N \frac{|\nabla \rho_i(\mathbf{r})|^2}{\rho_i(\mathbf{r})} - \frac{1}{8}\nabla^2 \rho; \quad C_F = \frac{3}{10}(3\pi^2)^{2/3}$$

a , b , c and d are constants with values 0.049, 0.132, 0.2533 and 0.349, respectively. This expression provides both local and non-local components within a single expression and the gradient contribution to second order. A combination of the standard local spin density approximation exchange result (Equation (3.54)) with the Becke gradient-exchange correction and the Lee–Yang–Parr correlation functional is currently a popular choice, commonly abbreviated to BLYP (pronounced ‘blip’).

3.7.4 Hybrid Hartree–Fock/Density Functional Methods

As we stated earlier, a key feature of density functional theory is the way in which correlation effects are incorporated from the beginning, unlike Hartree–Fock theory. Moreover, the incorporation of correlation into the Hartree–Fock formalism often involves significant computational overhead, as we have considered in Section 3.3. However, it is important to recognise that Hartree–Fock theory does provide an essentially exact means of treating the exchange contribution. One potentially attractive option is thus to add a correlation energy derived from DFT (e.g. the local density approximation) to the Hartree–Fock energy. In such an approach the exchange-correlation energy is written as a sum of the exact exchange term together with the correlation component from the local density approximation. This ‘exact’ exchange energy is obtained from the Slater determinant of the Kohn–Sham orbitals.

Unfortunately, this simple approach does not work well, but Becke has proposed a strategy which does seem to have much promise [Becke 1993a, b]. In his approach the exchange-correlation energy E_{XC} is written in the following form:

$$E_{XC} = \int_0^1 U_{XC}^{\lambda} d\lambda \quad (3.71)$$

Equation (3.71) contains a coupling parameter λ , which takes values from 0 to 1. A value of zero corresponds to a system where there is no Coulomb repulsion U_{XC} between the electrons (i.e. the Kohn-Sham non-interacting reference state). As λ increases to 1 the interelectronic Coulomb repulsion is introduced until $\lambda = 1$, which corresponds to the 'real' system with full interactions. For all values of λ the electron density is the same and equal to the density of the real system. It is not practical to perform this integral analytically and so it must be approximated. The simplest approximation is a linear interpolation:

$$U_{XC} = \frac{1}{2}(U_{XC}^0 + U_{XC}^1) \quad (3.72)$$

When $\lambda = 0$ we have U_{XC}^0 , which is the exchange-correlation potential energy of the non-interacting reference system. As there are no electronic interactions in this system there is no correlation term and so U_{XC}^0 corresponds to the pure exchange energy of the Kohn-Sham determinant and can be determined exactly. U_{XC}^1 is the exchange-correlation potential energy of the full-interacting real system. Becke proposed that this should be calculated using the local spin-density approximation. This potential energy (note that it is not the total energy, E) is available from:

$$U_{XC}^1 \approx U_{XC}^{\text{LSDA}} = \int u_{XC}[\rho_\alpha(\mathbf{r}), \rho_\beta(\mathbf{r})] d\mathbf{r} \quad (3.73)$$

u_{XC} is the exchange-correlation potential energy density of an electron gas for which appropriate expressions are available.

This so-called 'half-and-half' theory proved to be significantly better than the alternative methods based upon mixing exact exchange and correlation energies. In a refinement of the scheme, Becke recognised that there were problems with the model when $\lambda = 0$. These problems arise because the electron gas model is not appropriate near this exchange-only limit for molecular bonds. Hence a key feature of Becke's modified model is to eliminate the term U_{XC}^0 and to write the exchange-correlation energy as the following linear combination:

$$E_{XC} = E_{XC}^{\text{LSDA}} + a_0(E_X^{\text{exact}} - E_X^{\text{LSDA}}) + a_X \Delta E_X^{\text{GC}} + a_C \Delta E_C^{\text{GC}} \quad (3.74)$$

In Equation (3.74) E_X^{exact} is the exact exchange energy (obtained from the Slater determinant of the Kohn-Sham orbitals), E_X^{LSDA} is the exchange energy under the local spin density approximation, ΔE_X^{GC} is the gradient correction for exchange and ΔE_C^{GC} is the gradient correction for correlation. a_0 , a_X and a_C are empirical coefficients obtained by least-squares fitting to experimental data (56 atomisation energies, 42 ionisation potentials, eight proton affinities and the total atomic energies of the ten first-row elements). Their values are $a_0 = 0.20$, $a_X = 0.72$ and $a_C = 0.81$. In Becke's original paper his own gradient correction for exchange was used together with a gradient correction for correlation suggested by Perdew and Wang. An alternative to this scheme is to employ the Lee-Yang-Parr correlation functional (with the gradient term) and the standard local correlation functional due to Vosko, Wilk and Nusair (VWN). This is the 'B3LYP' density functional:

$$E_{XC}^{\text{B3LYP}} = (1 - a_0)E_{XC}^{\text{LSDA}} + a_0 E_X^{\text{HF}} + a_X \Delta E_X^{\text{B88}} + a_C E_C^{\text{LYP}} + (1 - a_C)E_C^{\text{VWN}} \quad (3.75)$$

3.7.5 Performance and Applications of Density Functional Theory

The application of density functional theory to isolated, ‘organic’ molecules is still in relative infancy compared with the use of Hartree–Fock methods. There continues to be a steady stream of publications designed to assess the performance of the various approaches to DFT. As we have discussed there is a plethora of ways in which density functional theory can be implemented with different functional forms for the basis set (Gaussians, Slater type orbitals, or numerical), different expressions for the exchange and correlation contributions within the local density approximation, different expressions for the gradient corrections and different ways to solve the Kohn–Sham equations to achieve self-consistency. This contrasts with the situation for Hartree–Fock calculations, which mostly use one of a series of tried and tested Gaussian basis sets and where there is a substantial body of literature to help choose the most appropriate method for incorporating post-Hartree–Fock methods, should that be desired.

A clear conclusion from such comparative studies is that density functional methods using gradient-corrected functionals can give results for a wide variety of properties that are competitive with, and in some cases superior to, *ab initio* calculations using correlation (e.g. MP2). Gradient-corrected functionals are required for the calculation of relative conformational energies and the study of intermolecular systems, particularly those involving hydrogen bonding [Sim *et al.* 1992]. As is the case with the *ab initio* methods the choice of basis set is also important in determining the results. By keeping the basis set constant (6-31G* being a popular choice) it is possible to make objective comparisons. Four examples of such comparative studies are those of Johnson and colleagues, who considered small neutral molecules [Johnson *et al.* 1993]; St-Amant *et al.*, who examined small organic molecules [St-Amant *et al.* 1995]; Stephens *et al.*, who performed a detailed study of the absorption and circular dichroism spectra of 4-methyl-2-oxetanone [Stephens *et al.* 1994]; and Frisch *et al.*, who compared a variety of density functional methods with one another and to traditional *ab initio* approaches [Frisch *et al.* 1996]. The evolution of defined sets of data such as those associated with the Gaussian-*n* series of models has also acted as a spur to those involved in developing density functional methods. For example, much of Becke’s work on gradient corrections and on mixed Hartree–Fock/density function methods was evaluated using data sets originally collated for the Gaussian-1 and Gaussian-2 methods. A more recent example is a variant of the Gaussian-3 method which uses B3LYP to determine geometries and zero-point energies [Baboul *et al.* 1999].

One of the most important developments for the practical application of DFT were methods for calculating analytical gradients of the energy with respect to the nuclear coordinates. This enables molecular geometries to be optimised. Once more there are some differences between the way this is done with density functional theory compared with Gaussian-based Hartree–Fock methods. A potential problem is that the use of grid-based integration schemes makes it difficult to provide exact expressions for the gradients. However, the errors associated with the grid-based method are generally very small and do not cause problems during the optimisation.

3.8 Quantum Mechanical Methods for Studying the Solid State

3.8.1 Introduction

The quantum mechanical methods used to study the behaviour of solid-phase systems are often somewhat different to those traditionally employed for studies of individual molecules or isolated intermolecular complexes. A perfect crystalline system can be constructed by stacking copies of some repeating unit (the *unit cell*) in a systematic fashion without overlapping and without gaps. The structure of a crystal can be specified by defining the size and shape of the unit cell and the positions of the atoms within it. The unit cell is parallelepiped in shape and is characterised by three lattice vectors \mathbf{a} , \mathbf{b} and \mathbf{c} and the angles between them (Figure 3.7). It may be possible to conceive of more than one unit cell, with different unit cell parameters. In such cases a set of standard cell parameters can be obtained by the application of standardisation rules. The coordinates of the atoms in the unit cell may be expressed as fractional coordinates ($\alpha\mathbf{a}$, $\beta\mathbf{b}$, $\gamma\mathbf{c}$). Indeed, any general vector \mathbf{r} can be written in terms of these basis vectors:

$$\mathbf{r} = (\alpha\mathbf{a}, \beta\mathbf{b}, \gamma\mathbf{c}) \quad (3.76)$$

where α , β and γ are not necessarily restricted to values between 0 and 1. There are fourteen different types of basic unit cell; these are the *Bravais lattices*. Common Bravais lattices include the simple cubic, body-centred cubic and face-centred cubic (Figure 3.8). Another common structure also shown in Figure 3.8 is the hexagonal close-packed arrangement, for which the underlying Bravais lattice (called the simple hexagonal) is formed from an underlying triangular arrangement. In addition to the translational symmetry that the unit cell must possess there may be some symmetry to the arrangement of the atoms within the unit cell. The particular combination of symmetry elements in a crystal defines its *space group*. There are 230 different space groups. If there is symmetry within the unit cell then it is strictly only necessary to specify the asymmetric unit (the unique part of the structure); the positions of the other atoms can be generated using the appropriate symmetry operators.

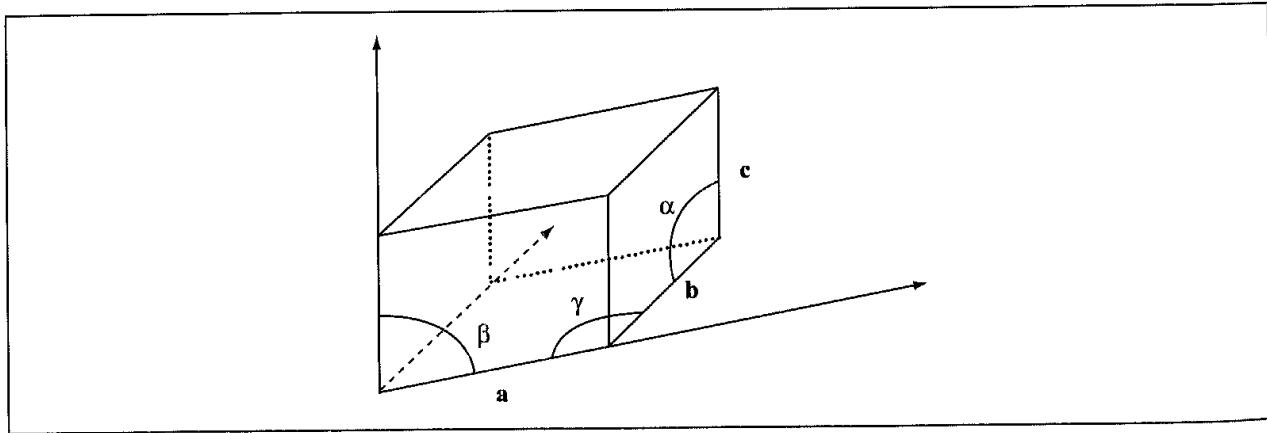


Fig. 3.7. The six parameters a , b , c , α , β , γ which characterise the unit cell.

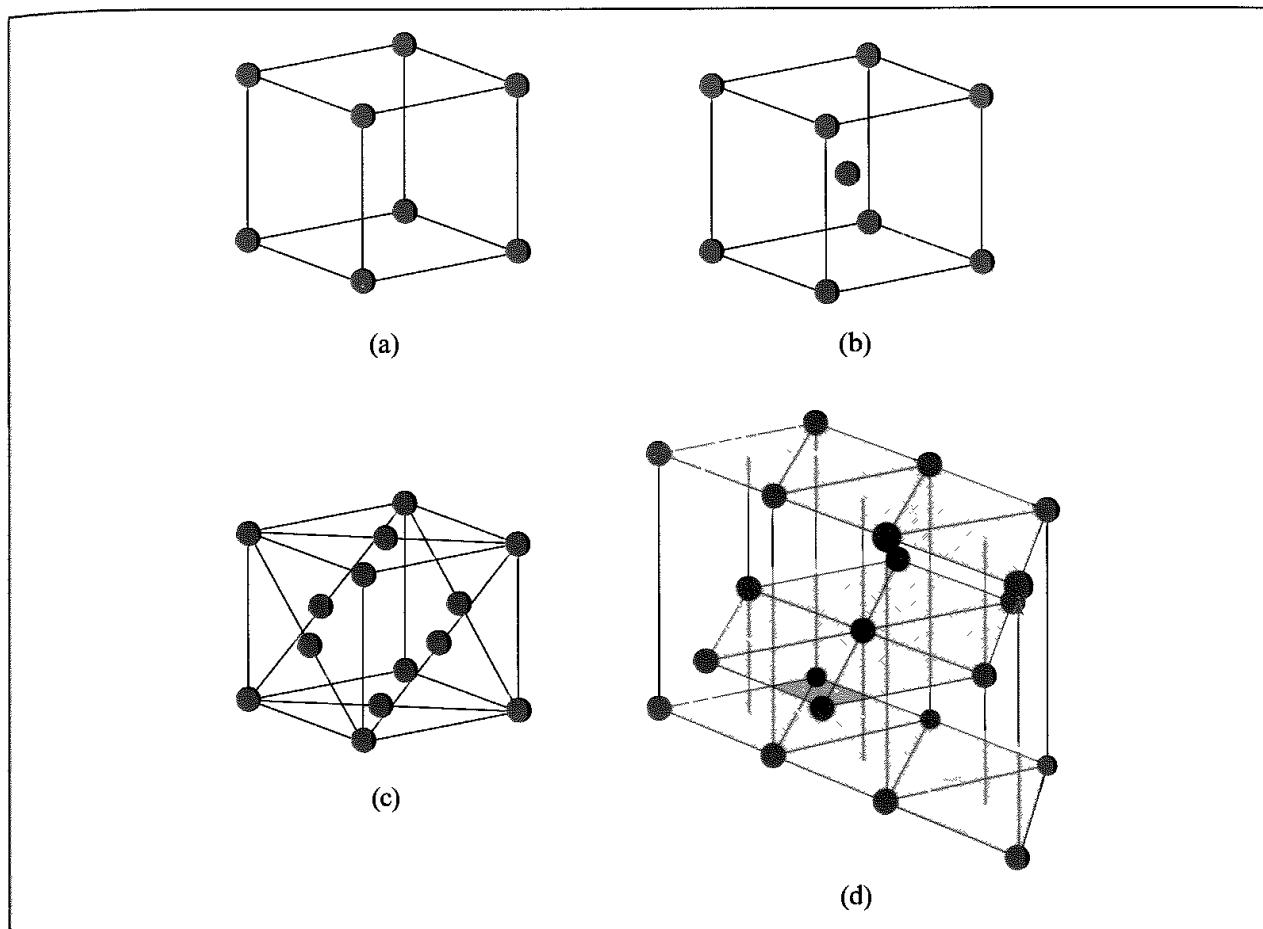


Fig. 3.8 Some basic Bravais lattices (a) simple cubic, (b) body-centred cubic, (c) face-centred cubic and (d) simple hexagonal close-packed (Figure adapted in part from Ashcroft N W and Mermin N D 1976. Solid State Physics New York, Holt, Rinehart and Winston.)

Another concept that is extremely powerful when considering lattice structures is the *reciprocal lattice*. X-ray crystallographers use a reciprocal lattice defined by three vectors \mathbf{a}^* , \mathbf{b}^* and \mathbf{c}^* in which \mathbf{a}^* is perpendicular to \mathbf{b} and \mathbf{c} and is scaled so that the scalar product of \mathbf{a}^* and \mathbf{a} equals 1. \mathbf{b}^* and \mathbf{c}^* are similarly defined. In three dimensions this leads to the following definitions:

$$\mathbf{a}^* = \frac{\mathbf{b} \times \mathbf{c}}{\mathbf{a} \cdot \mathbf{b} \times \mathbf{c}}; \quad \mathbf{b}^* = \frac{\mathbf{a} \times \mathbf{c}}{\mathbf{b} \cdot \mathbf{a} \times \mathbf{c}}; \quad \mathbf{c}^* = \frac{\mathbf{a} \times \mathbf{b}}{\mathbf{c} \cdot \mathbf{a} \times \mathbf{b}} \quad (3.77)$$

Note that the denominator in each case is equal to the volume of the unit cell. The fact that \mathbf{a}^* , \mathbf{b}^* and \mathbf{c}^* have the units of 1/length gives rise to the terms ‘reciprocal space’ and ‘reciprocal lattice’. It turns out to be convenient for our computations to work with an expanded reciprocal space that is defined by three closely related vectors $\mathbf{a}^\$$, $\mathbf{b}^\$$ and $\mathbf{c}^\$$, which are multiples by 2π of the X-ray crystallographic reciprocal lattice vectors:

$$\mathbf{a}^\$ = 2\pi\mathbf{a}^*; \quad \mathbf{b}^\$ = 2\pi\mathbf{b}^*; \quad \mathbf{c}^\$ = 2\pi\mathbf{c}^* \quad (3.78)$$

A simple illustrative example of reciprocal space is that of a 2D square lattice where the vectors \mathbf{a} and \mathbf{b} are orthogonal and of length equal to the lattice spacing, a . Here \mathbf{a}^* and \mathbf{b}^* are directed along the same directions as \mathbf{a} and \mathbf{b} respectively and have a length $1/a$

combined to give the equivalent of molecular orbitals. It is based on the assumption that the effect of orbital overlap is to modulate but not change completely the initial atomic levels. The approximation is traditionally considered most useful for describing the electronic structure of systems such as insulators and transition metals with partially filled d shells. The second approach is called the *nearly free-electron approximation*. This theory starts by considering the electrons as free particles whose motion is modulated by the presence of the lattice. The nearly free-electron approximation is traditionally considered the more suitable approach to systems such as metals where there is substantial overlap of the valence orbitals. We will outline both approaches in turn, making use of some of the fundamental principles and properties of lattices discussed earlier.

3.8.2 Band Theory and Orbital-based Approaches

Band theory is perhaps easier for chemists to understand, starting as it does from an orbital picture. We will therefore spend somewhat less space discussing this than the nearly free-electron approximation. We will start by considering the simplest problem, a 1D lattice. Initially we consider what happens if we bring together two atoms along the x axis until they are separated by a distance, a . If each atom has one s orbital, then the combined system has two molecular orbitals (one bonding and one anti-bonding). If we then add a third atom then three molecular orbitals are obtained (one bonding, one non-bonding and one anti-bonding). Four atoms give four energy levels, and so on. As more atoms are added the energy levels merge to give what is an essentially continuous *band* of energy levels (Figure 3.11). Each energy level can accommodate two electrons so if each atom contributes one electron the band will be half full. The presence of unoccupied energy levels near to the top of the filled level means that it is very easy to excite electrons from the filled to the unfilled levels. The electrons are consequently very mobile, giving rise to the special conduction and thermal properties of a metal. By contrast, if each atom contributes two electrons then the band will be completely filled. Such electrons would have to be excited to higher bands, which might, for example, be formed by the overlap of p orbitals. In an insulator the energy of this p band would typically be significantly higher than the lower s band and so excitation would require considerable energy. In a semiconductor the band gap is smaller and it may be possible to excite electrons from the top of the highest filled band (the *valence band*) to the lowest unoccupied band (the *conduction band*) at normal temperatures. These three difference scenarios are illustrated in Figure 3.11.

The periodicity of the lattice means that the values of a function (such as the electron density) will be identical at equivalent points on the lattice. Likewise there is a relationship between the wavefunction at a point (x in our 1D lattice) and at an equivalent point elsewhere on the lattice (for the 1D lattice this would be $x + na$, where n is an integer). *Bloch's theorem* provides the link; each allowed lattice wavefunction must satisfy the following relationship:

$$\psi^k(x + a) = e^{ika} \psi^k(x) \quad (3.81)$$

In this equation we have identified the wavefunction with a label, k , which for now can be considered an index; there are as many values of k as there are atoms in the 1D lattice.

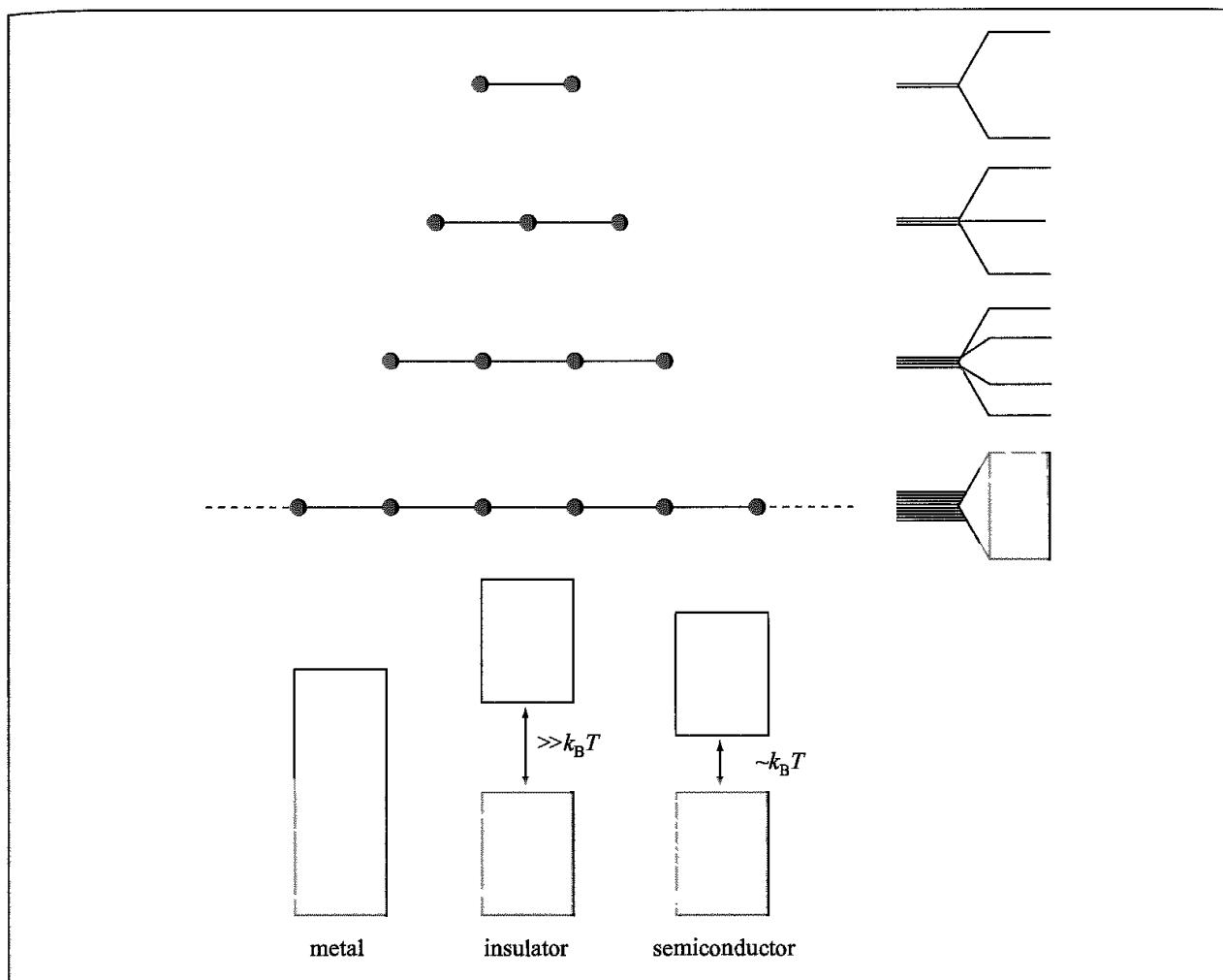


Fig 3.11: The creation of a band of energy levels from the overlap of two, three, four, etc atomic orbitals, which eventually gives rise to a continuum. Also shown are the conceptual differences between metals, insulators and semiconductors

We wish to construct linear combinations of the atomic orbitals such that the overall wavefunction meets the Bloch requirement. Suppose the s orbitals in our lattice are labelled χ_n , where the n th orbital is located at position $x = na$. An acceptable linear combination of these orbitals that satisfies the Bloch requirements is:

$$\psi^k = \sum_n e^{ikna} \chi_n \quad (3.82)$$

We now need to consider how the form of the wavefunction varies with k . The first situation we consider corresponds to $k = 0$, where the exponential terms are all equal to 1 and the overall wavefunction becomes a simple additive linear combination of the atomic orbitals:

$$\psi^{k=0} = \sum_n \chi_n = \chi_0 + \chi_1 + \chi_2 + \dots \quad (3.83)$$

The other situation we consider is $k = \pi/a$. Recall that $\exp(ix)$ can be written $\cos(x) + i\sin(x)$. If $k = \pi/a$ then the sine terms will all be zero, leaving just the cosine terms $\cos(n\pi)$, which can

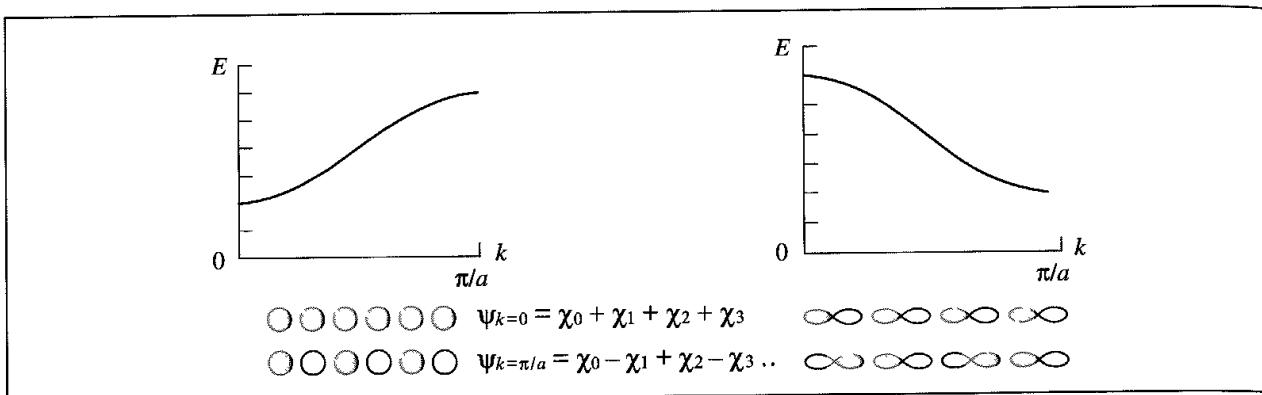


Fig 3.12 The variation in energy with k for a 1D lattice for a set of s orbitals (left) and for a set of p_x orbitals (right). Also shown are the corresponding arrangements of orbitals.

be expressed more generally as $(-1)^n$. Hence the wavefunction is:

$$\psi^{k=\pi/a} = \sum_n (-1)^n \chi_n = \chi_0 - \chi_1 + \chi_2 - \dots \quad (3.84)$$

Equations (3.83) and (3.84) correspond to the lowest- and highest-energy wavefunctions for our simple system over this range of k . Wavefunctions for values of k between 0 and π/a have intermediate energies. The energy varies in a cosine-like manner with k between $k = 0$ and $k = \pi/a$ (Figure 3.12). Note that k can adopt negative values and that $E(-k)$ equals $E(k)$. Also worthy of note is that p orbitals show different behaviour to the s orbitals. For a set of p_x orbitals it is the $k = 0$ state that is of highest energy and $k = \pi/a$ is of lowest energy, due to their nodal behaviour.

The graph of energy versus k is called the *band structure*; the *bandwidth* is the difference in energy between the lowest and highest levels in the band. For the one-dimensional lattice the bandwidth is determined by the lattice spacing; a smaller spacing a gives a greater bandwidth in much the same way that the difference between the bonding and antibonding orbitals in H_2 increases as the atoms get closer together. As we noted above there are as many values of k (and so as many energy levels) as there are atoms in the lattice and that each energy level can accommodate two electrons.

We now move on to consider a two-dimensional square lattice in the (x, y) plane, where the inter-lattice spacing is still a . The Bloch theorem is now written in the following more general form:

$$\psi^{\mathbf{k}}(\mathbf{r} + \mathbf{T}) = e^{i\mathbf{k}\cdot\mathbf{T}} \psi^{\mathbf{k}}(\mathbf{r}) \quad (3.85)$$

In Equation (3.85) \mathbf{T} is a translation vector that maps each position into an equivalent position in a neighbouring cell, \mathbf{r} is a general positional vector and \mathbf{k} is the *wavevector* which characterises the wavefunction. \mathbf{k} has components k_x and k_y in two dimensions and is equivalent to the parameter k in the one-dimensional system. For the two-dimensional square lattice the Schrödinger equation can be expressed in terms of separate wavefunctions along the x - and y -directions. This results in various combinations of the atomic 1s orbitals, some of which are shown in Figure 3.13. These combinations have different energies. The lowest-energy solution corresponds to $(k_x = 0, k_y = 0)$ and is a straightforward linear

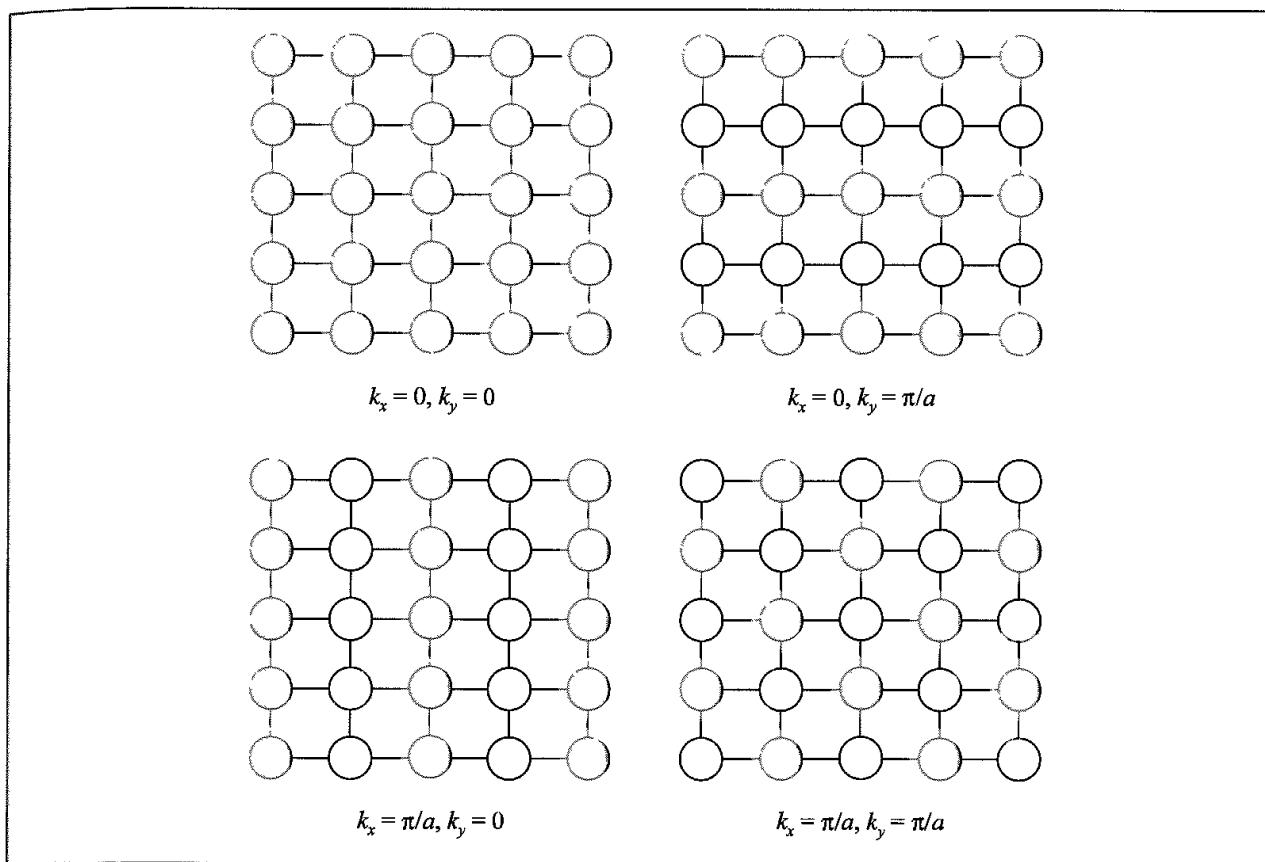


Fig 3.13. Some of the possible combinations of atomic 1s orbitals for a 2D square lattice corresponding to different values of k_x and k_y . A shaded circle indicates a positive coefficient; an open circle corresponds to a negative coefficient.

combination of the atomic orbitals. The highest-energy solution corresponds to the situation where both k_x and k_y have values of π/a . The wavefunction for this high-energy solution shows a rapid variation in sign. Another important feature evident in Figure 3.13 is the wave-like nature of the various linear combinations, particularly if one imagines the lattice extending infinitely in all directions over the (x, y) plane.

The reciprocal space and the reciprocal lattice are directly related to the wavevector, \mathbf{k} ; different values of \mathbf{k} can be considered as points within the reciprocal space defined by $\mathbf{a}^\$$, $\mathbf{b}^\$$ and $\mathbf{c}^\$$. It turns out that, when we are calculating the wavefunction and energy levels for a solid, we need to restrict \mathbf{k} to one cell in the reciprocal lattice (typically chosen to the cell containing $\mathbf{k} = 0$, or the first Brillouin zone), otherwise there is a danger of counting some states more than once. A very common way to represent the band structure for lattice structures is to plot how the energy changes as a function of \mathbf{k} along certain lines of symmetry within the first Brillouin zone. For example, to return to our square lattice (for which the reciprocal lattice is also square) one could imagine taking a 'tour' starting at the origin ($\mathbf{k} = (0, 0)$), moving along the x axis to $\mathbf{k} = (\pi/a, 0)$ up the y axis to $\mathbf{k} = (\pi/a, \pi/a)$, and finally returning to the origin. As we undertake this tour the energy changes as shown in Figure 3.14. In this diagram we have labelled certain values of \mathbf{k} which have particular symmetry with their conventional Roman or Greek capital letters, Γ , X and M [Bradley and Cracknell 1972].

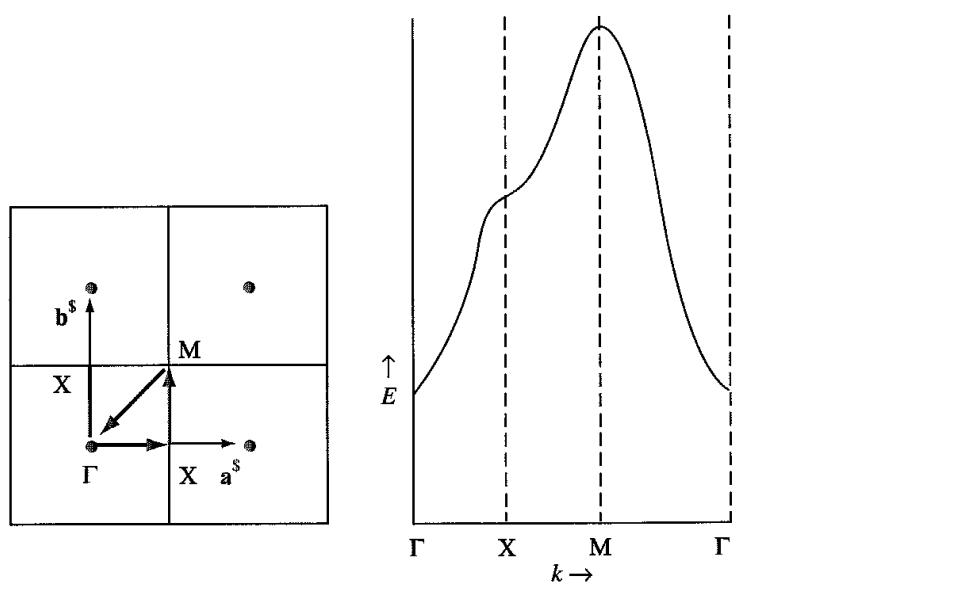


Fig. 3.14: Variation in energy for a 'tour' (Γ -X-M- Γ) of the reciprocal lattice for a 2D square lattice of hydrogen atoms. (Figure adapted in part from Hoffmann R 1988. Solids and Surfaces. A Chemist's View on Bonding in Extended Structures New York, VCH Publishers.)

3.8.3 The Periodic Hartree–Fock Approach to Studying the Solid State

In the *periodic Hartree–Fock* approach the elements of the Fock matrix are constructed from linear combinations of so-called Bloch functions:

$$\psi_i^{\mathbf{k}}(\mathbf{r}) = \sum_{\omega} a_{\omega i}(\mathbf{k}) \varphi_{\omega}^{\mathbf{k}}(\mathbf{r}) \quad (3.86)$$

Each Bloch function is itself a linear combination of atomic orbitals:

$$\varphi_{\omega}^{\mathbf{k}}(\mathbf{r}) = \sum_{\mathbf{T}} \chi_{\omega}^{\mathbf{T}}(\mathbf{r}) \exp(i\mathbf{k} \cdot \mathbf{T}) \quad (3.87)$$

$\chi_{\omega}^{\mathbf{T}}$ is the ω th atomic orbital in the crystal cell characterised by the lattice vector \mathbf{T} . As such, this method works in real space, which contrasts with the usual implementations of the alternative plane-wave methods that we will discuss below [Dovesi *et al.* 2000]. Each atomic orbital is expressed as a linear combination of (for example) Gaussian functions, as in molecular Hartree–Fock theory. The coefficients $a_{\omega i}(\mathbf{k})$ in Equation (3.86) are obtained by solving the following matrix equation for every value of \mathbf{k} to self-consistency:

$$\mathbf{F}_{\mathbf{k}} \mathbf{A}_{\mathbf{k}} = \mathbf{S}_{\mathbf{k}} \mathbf{A}_{\mathbf{k}} \mathbf{E}_{\mathbf{k}} \quad (3.88)$$

$\mathbf{S}_{\mathbf{k}}$ is the overlap matrix for the Bloch functions for the wavevector \mathbf{k} , with $\mathbf{E}_{\mathbf{k}}$ being the energy matrix and \mathbf{A} the matrix of coefficients. $\mathbf{F}_{\mathbf{k}}$ is the Fock matrix, which consists of a sum of one- and two-electron terms. The values of \mathbf{k} are typically selected to sample from the first Brillouin zone according to a special scheme as described in Section 3.8.6. When these terms are expanded they involve infinite sums over the nuclei and electrons in the lattice. As is usual in a Hartree–Fock approach the one-electron terms involve the sum of a kinetic energy term and one due to the Coulomb interaction between the nuclei and the

electrons; the two-electron terms involve Coulomb and exchange two-electron integrals. Unfortunately, if these sums were to be evaluated individually and to completion then they would not converge to a consistent value, but would diverge. However, effective ways to determine these infinite sums have been proposed [Pisani and Dovesi 1980; Dovesi *et al.* 1983]. These involve a variety of procedures. The Coulomb interactions are divided into a series of terms corresponding to interacting and non-interacting charge distributions. The latter can then be grouped together into ‘shells’ and the interaction calculated using multipole expansions (see Section 4.9.1). For the shorter-range exchange interaction it is possible to truncate the integral summation at an appropriate distance without loss of accuracy. The truncation distance can depend upon the three-dimensional structure of the material and so may vary from one calculation to the next.

Within the periodic Hartree–Fock approach it is possible to incorporate many of the variants that we have discussed, such as UHF or RHF. Density functional theory can also be used. This makes it possible to compare the results obtained from these variants. Whilst density functional theory is more widely used for solid-state applications, there are certain types of problem that are currently more amenable to the Hartree–Fock method. Of particular relevance here are systems containing unpaired electrons, two recent examples being the electronic and magnetic properties of nickel oxide and alkaline earth oxides doped with alkali metal ions (Li in CaO) [Dovesi *et al.* 2000].

3.8.4 The Nearly Free-electron Approximation

Whereas the tight-binding approximation works well for certain types of solid, for other systems it is often more useful to consider the valence electrons as free particles whose motion is modulated by the presence of the lattice. Our starting point here is the Schrödinger equation for a free particle in a one-dimensional, infinitely large box:

$$\left(\frac{d^2}{dx^2}\right)\psi = -\left(\frac{2mE}{\hbar^2}\right)\psi \quad (3.89)$$

The solutions to this equation are:

$$\psi = C \exp(ikx); \quad E = (\hbar^2 k^2)/2m \quad (3.90)$$

The energy for a free particle can be related to the momentum by $E = p^2/2m$ and so the wavefunction is related to the momentum p by:

$$\psi = C \exp(\pm ipx/\hbar) \quad (3.91)$$

The wavelength of this motion is \hbar/p and the parameter k is equal to $2\pi p/\hbar$. Thus k has units of 1/length (i.e. reciprocal length). The energy for a free particle varies in a quadratic fashion with k and in principle any value of the energy is possible.

In two dimensions we obtain the following wavefunction:

$$\psi_{x,y} = C_x \exp(ik_x x/\hbar) C_y \exp(ik_y y/\hbar) = C \exp(i\mathbf{k} \cdot \mathbf{r}/\hbar) \quad (3.92)$$

Note that in Equation (3.92) we have expressed the wavefunction in terms of a vector, \mathbf{k} (which has components in the x and y directions of k_x and k_y) and the Cartesian vector \mathbf{r} .

The energy varies as a quadratic function of both k_x and k_y :

$$E_{x,y} = \frac{\hbar^2}{2m} (k_x^2 + k_y^2) \quad (3.93)$$

An analogous expression is obtained in three dimensions. We now need to consider periodic systems. As we have discussed, the wavefunction for a particle on a periodic lattice must satisfy Bloch's theorem, Equation (3.85). The wavevector \mathbf{k} in Bloch's theorem plays the same role in the study of periodic systems as the vector \mathbf{k} does for a free particle. One important difference is that whereas the wavevector is directly related to the momentum for a free particle (i.e. $\mathbf{k} = \mathbf{p}/\hbar$) this is not the case for the Bloch particle due to the presence of the external potential (i.e. the nuclei). However, it is very convenient to consider $\hbar\mathbf{k}$ as analogous to the momentum and it is often referred to as the *crystal momentum* for this reason. The possible values that \mathbf{k} can adopt are given by:

$$\mathbf{k} = \left(\frac{m_\alpha}{N_\alpha} \mathbf{a}^\$, \frac{m_\beta}{N_\beta} \mathbf{b}^\$, \frac{m_\gamma}{N_\gamma} \mathbf{c}^\$ \right) \quad (3.94)$$

m_α , m_β and m_γ are integers and $N_\alpha N_\beta N_\gamma = N$, the number of unit cells in the crystal. For a macroscopic system where N is very large (of the order of Avogadro's number) \mathbf{k} thus varies continuously. As we have seen before, the wavevector \mathbf{k} in the Bloch theorem (Equation (3.85)) can be considered as a point within the reciprocal lattice defined by $\mathbf{a}^\$$, $\mathbf{b}^\$$ and $\mathbf{c}^\$$. It can also be shown (see Appendix 3.1) that a wavefunction that satisfies Bloch's theorem can be written in the following form:

$$\psi^\mathbf{k}(\mathbf{r}) = e^{i\mathbf{k}\cdot\mathbf{r}} u^\mathbf{k}(\mathbf{r}) \quad (3.95)$$

Here, $u^\mathbf{k}(\mathbf{r})$ is a function that is periodic on the lattice. Recall from our earlier discussions on reciprocal lattice vectors that one way to construct such a periodic function is as a Fourier series expansion of plane wavefunctions $\exp(i\mathbf{G}\cdot\mathbf{r})$:

$$u^\mathbf{k}(\mathbf{r}) = \sum_{\mathbf{G}} c_{\mathbf{G}}^\mathbf{k} \exp(i\mathbf{G}\cdot\mathbf{r}) \quad (3.96)$$

The sum runs over the reciprocal lattice vectors \mathbf{G} we considered above. A simple case is $\mathbf{G} = \mathbf{a}^\$$, for which $\exp(i\mathbf{G}\cdot\mathbf{r})$ corresponds to a wave travelling perpendicular to the real-space axes \mathbf{b} and \mathbf{c} and with a wavelength such that it fits exactly into the unit cell. If $\mathbf{G} = 2\mathbf{a}^\$$ then two wavelengths fit into the cell.

The external potential due to the nuclei is periodic in the lattice and it too can be written as a Fourier expansion of exponential functions of the reciprocal lattice:

$$U(\mathbf{r}) = \sum_{\mathbf{G}} U_{\mathbf{G}} \exp(i\mathbf{G}\cdot\mathbf{r}) \quad (3.97)$$

$U_{\mathbf{G}}$ is the Fourier coefficient. When this form of the potential is incorporated into the Schrödinger equation the following equation can be derived [Ashcroft and Mermin 1976]:

$$\left(\frac{\hbar^2}{2m} |\mathbf{k} + \mathbf{G}|^2 - E \right) c_{\mathbf{G}}^\mathbf{k} + \sum_{\mathbf{G}'} U_{\mathbf{G}' + \mathbf{G}} c_{\mathbf{G}'}^\mathbf{k} = 0 \quad (3.98)$$

We can recover the free-particle result (i.e. zero potential) from Equation (3.98) by setting all of the Fourier coefficients U_G to zero, in which case the equation reduces to:

$$\left(\frac{\hbar^2}{2m} |\mathbf{k} + \mathbf{G}|^2 - E \right) c_{\mathbf{G}}^k = 0 \quad (3.99)$$

The solution of this equation requires that $E = \hbar^2 |\mathbf{k} + \mathbf{G}|^2 / 2m$ with the wavefunctions being of the form $\psi(\mathbf{r}) \propto \exp[i(\mathbf{k} + \mathbf{G}) \cdot \mathbf{r}]$. Although cast in a slightly different form, this is equivalent to our earlier expression for the wavefunction of a free particle, Equation (3.92).

The summations in Equations (3.98) are over all reciprocal lattice vectors \mathbf{G} . As can be seen, for a given value of \mathbf{k} there are as many forms of this equation as there are reciprocal lattice vectors in the system. Each of these equations for the different values of \mathbf{G} gives rise to a solution which is labelled with the band index n . Thus there are as many values of n as there are reciprocal lattice vectors \mathbf{G} . Just as there are n solutions to this Schrödinger equation for a given value of \mathbf{k} , so it is also possible to consider how the energy varies with \mathbf{k} for a given value of n . To understand the entire band structure of a solid requires one to consider the variation of both \mathbf{k} and n . As we indicated above, when calculating the band structure it is usual to restrict \mathbf{k} to just the first Brillouin zone to avoid duplicate counting of states.

Let us now examine how these results can be applied to some simple one- and two-dimensional periodic systems. Initially we will consider the situation where there is no external potential and then discuss what happens when we introduce one. The first case is the one-dimensional lattice, which has reciprocal lattice vectors at $\pm 2\pi/a$, $\pm 4\pi/a$, etc. In order to derive the energy diagram we need to consider, for each reciprocal lattice vector \mathbf{G} , how the energy varies as we change \mathbf{k} over the first Brillouin zone (which in this case corresponds to varying \mathbf{k} from $-\pi/a$ to $+\pi/a$). The first reciprocal lattice vector is $\mathbf{G} = 0$, for which the energy simply varies quadratically with \mathbf{k} , from zero at $\mathbf{k} = 0$ to $\hbar^2(\pi/a)^2/2m$ at $\mathbf{k} = \pm 2\pi/a$. We next need to consider the two reciprocal vectors $\mathbf{G} = \pm 2\pi/a$. At the point $\mathbf{k} = 0$ the energy due to both of these reciprocal lattice vectors is $\hbar^2(2\pi/a)^2/2m$. As \mathbf{k} increases from 0 to $+\pi/a$ the value of $|\mathbf{k} + \mathbf{G}|^2$ increases for the reciprocal lattice vector $\mathbf{G} = 2\pi/a$ but it decreases for the reciprocal lattice vector $\mathbf{G} = -2\pi/a$. Conversely, as \mathbf{k} varies from 0 to $-\pi/a$ the energy increases for the reciprocal lattice vector $\mathbf{G} = -2\pi/a$ and decreases for $\mathbf{G} = 2\pi/a$. These variations in energy are shown in Figure 3.15. Two types of energy diagram are shown in this figure; one is the ‘reduced-zone’ scheme because the entire dependency of the energy on the wavevector is contained within the first Brillouin zone. The alternative representation is called an extended-zone scheme in which the energy levels are ‘folded out’ for values of \mathbf{k} beyond the first Brillouin zone.

We next need to introduce the weak potential, which acts to modulate the wavefunctions and the associated energy levels. The effects of the potential are found to be most acute where there is degeneracy of the energy levels. This arises even in the one-dimensional situation, where we have degenerate energy levels due to different reciprocal lattice vectors at $\mathbf{k} = 0$ and $\mathbf{k} = \pi/a$. The effect of the potential is to perturb these energy levels in such a way that lifts the degeneracy to create an energy gap. In the one-dimensional case the effect of the potential is to ‘flatten’ the energy levels in the region close to the edge of the

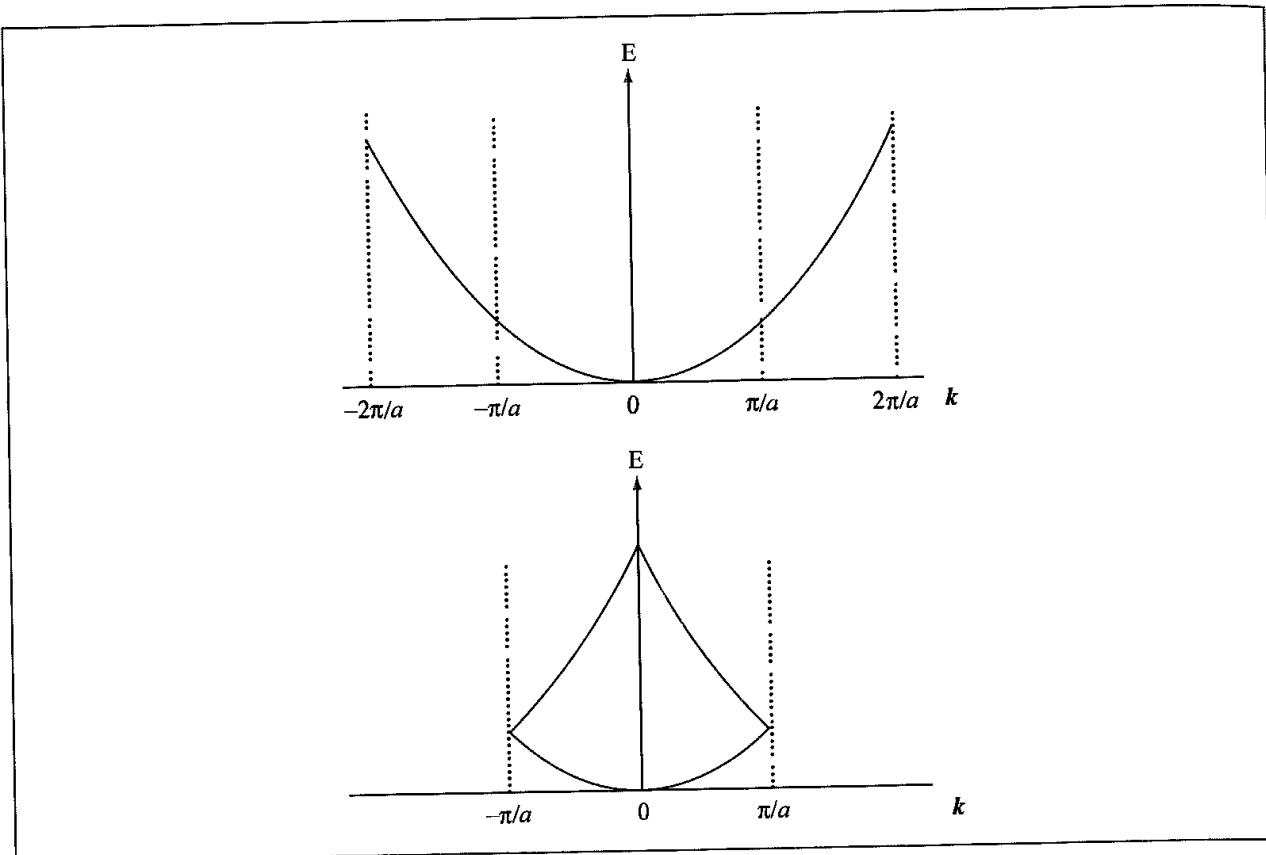


Fig. 3.15. Extended-zone and reduced-zone representations of band diagram for 1D lattice with no external potential.

Brillouin zone as shown in Figure 3.16. One way to explain the appearance of the energy gap at the edges of the Brillouin zone is to recognise that the states of a free electron are waves with a specific wavelength ($2\pi/k$ in the simple one-dimensional system). When the wavelength becomes comparable to the lattice spacing the lattice diffracts the wave and at the boundary of the Brillouin zone ($k = \pm\pi/a$) a standing wave is created. Two different standing waves are possible in a one-dimensional system, as shown in Figure 3.17. For one of the standing waves (A in Figure 3.17) the peak electron density occurs in the vicinity of the lattice points (the positive nuclei). This standing wave thus has a more favourable (i.e. lower) energy than the equivalent free travelling wave. By contrast, the peak electron density of the other standing wave (B in Figure 3.17) occurs between the nuclei and so its energy is higher. Further gaps arise at $k = \pm 2\pi/a$, and so on.

A somewhat more complex case is that of the 2D hexagonal lattice. As for the one-dimensional system we initially consider a free particle, restricting ourselves to wavevectors within the first Brillouin zone with higher-energy states being due to reciprocal lattice vectors beyond in the second, third, etc. Brillouin zones. We will consider how the energy varies as we undertake a 'tour' of the first Brillouin zone in reciprocal space starting at the origin ($\mathbf{k} = (0, 0)$), then moving to one of the vertices of the hexagon (the point $(\mathbf{k} = \cos \pi/6, \sin \pi/6)$), along to the mid-point of one of the edges ($\mathbf{k} = (0, \sin \pi/6)$), and finally back to the origin (Figure 3.18). The origin, the vertex and the mid-point are all points of symmetry and are identified by the symbols Γ , K and M , respectively. For a

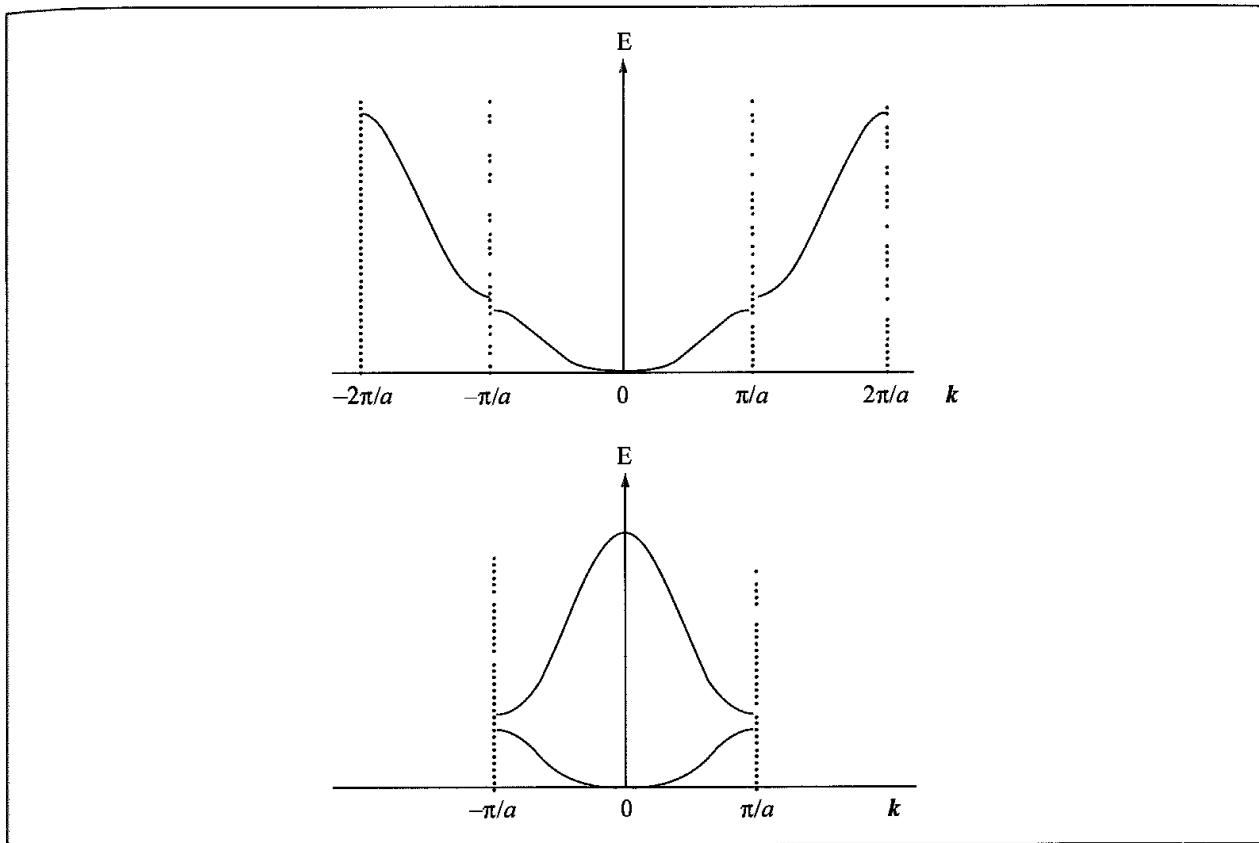


Fig 3.16: The effect of introducing a weak potential into the 1D lattice is to lift the degeneracy of the energy levels near to the edge of the Brillouin zone (shown in both extended-zone and reduced-zone representation)

given value of \mathbf{k} we compute the value of $|\mathbf{k} + \mathbf{G}|^2$ and thus the energy for the relevant reciprocal lattice vectors.

The simplest case is that corresponding to $\mathbf{G} = 0$. We still obtain a quadratic variation of energy with $|\mathbf{k}|$ wherever we move within the first Brillouin zone. The variation in energy for the three ‘legs’ of this tour can be represented in an energy band diagram as shown in Figure 3.18. As there are six nearest-neighbour cells in this system, there are six energy levels to monitor at the next stage. The distance from the origin to each of these six reciprocal lattice points is $2 \cos \pi/6$. At $\mathbf{k} = 0$ we therefore find that all six energy levels are degenerate

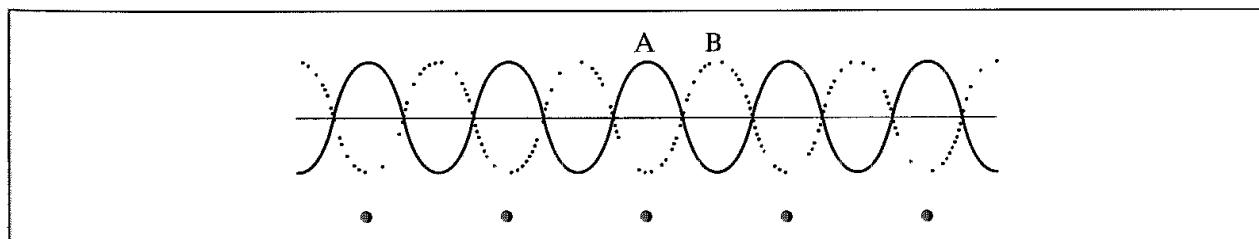


Fig 3.17 The two possible sets of standing waves at the Brillouin zone boundary. Standing wave A concentrates electron density at the nuclei, whereas wave B concentrates electron density between the nuclei. Wave A thus has a lower energy than wave B

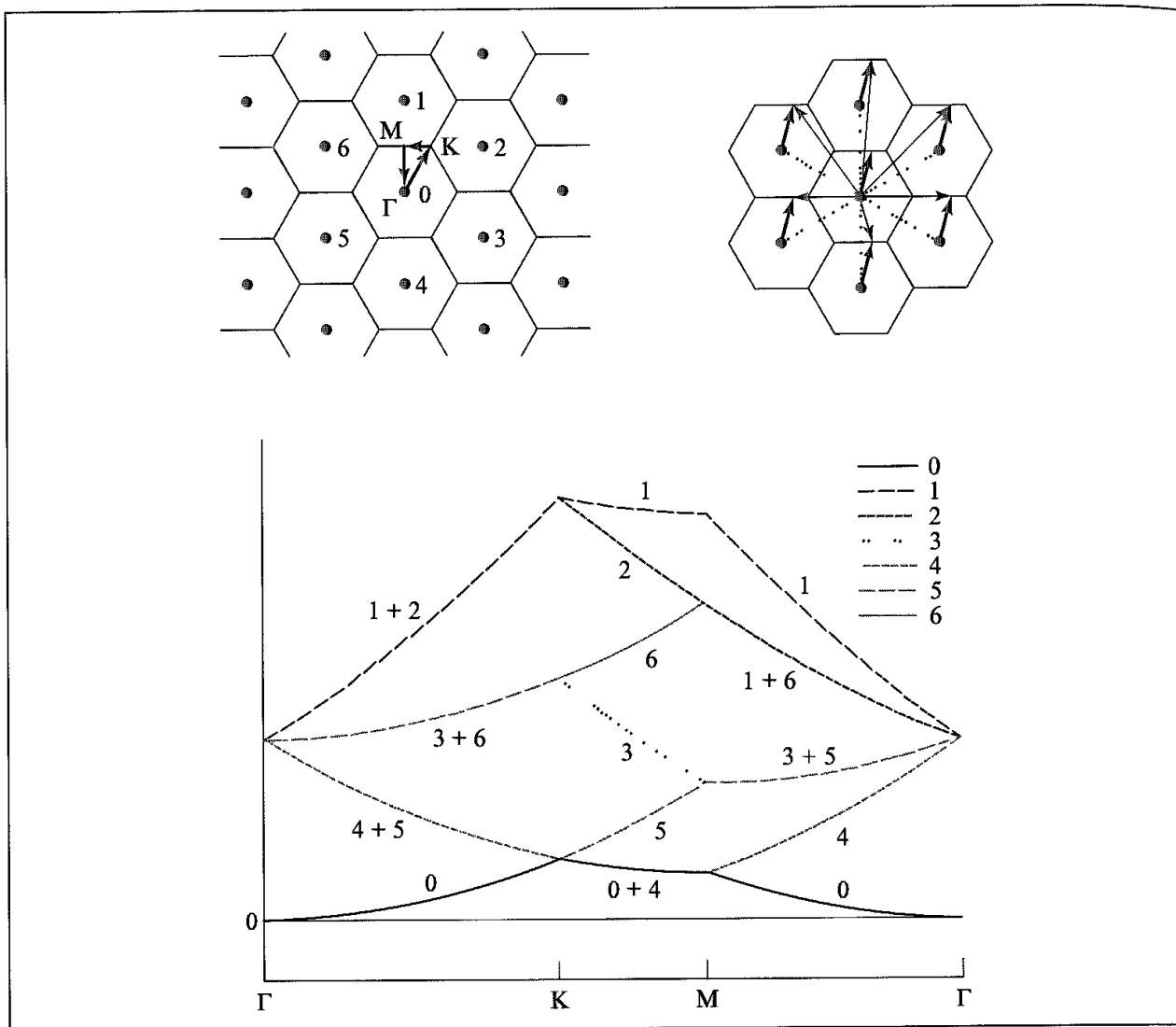


Fig. 3.18: Energy band diagram (bottom) for a free-particle 'tour' (Γ -K-M- Γ) of the reciprocal lattice for a 2D hexagonal structure (top left). A total of seven bands are shown, due to the central reciprocal lattice vector $\mathbf{G} = \mathbf{0}$ and the reciprocal lattice vectors from the six neighbouring cells. The energy varies as $|\mathbf{k} + \mathbf{G}|^2$, where the vector $\mathbf{k} + \mathbf{G}$ is computed as shown in the top right of the figure (\mathbf{k} : bold arrow; $\mathbf{k} + \mathbf{G}$: thin arrow).

and have a value of $3\hbar^2/2m$ ($(2\cos\pi/6)^2 \equiv 3$). Moving towards the point $(\cos\pi/6, \sin\pi/6)$ we find that the six vectors separate into three pairs of degenerate levels. These six reciprocal lattice points are labelled 1-6 in Figure 3.18, together with the corresponding energy levels. As the tour continues, the different energy bands show two-, three- and six-fold degeneracy, depending upon the value of \mathbf{k} . Another key feature is that along some legs of the tour certain pairs of bands are degenerate, though this degeneracy will often be lifted when a different leg is traversed. For example, the pairs 1-2, 3-6 and 4-5 are degenerate from Γ to K. Between K and M the pair 0-4 are degenerate; and on the final leg there is degeneracy between the pairs 2-6 and 3-4. When the periodic potential is introduced some, but not necessarily all, of this degeneracy will be lifted, giving rise to band gaps. The way in which this can occur is shown schematically in Figure 3.19.

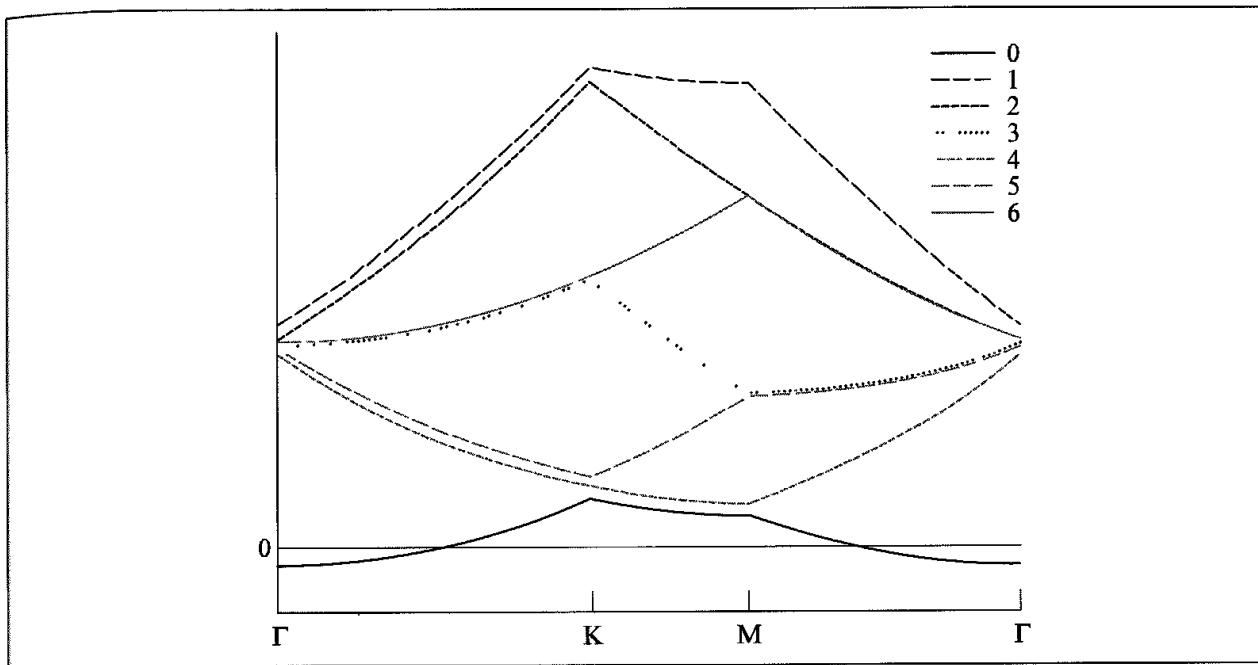


Fig. 3.19. The effect of a weak external potential is to lift degeneracy and create band gaps as illustrated for a 2D hexagonal lattice (compare with Figure 3.18).

3.8.5 The Fermi Surface and Density of States

To determine the ground state of a periodic system it is necessary to determine its band structure, by varying \mathbf{k} over the first Brillouin zone and computing at each value of \mathbf{k} the different energy bands resulting from the reciprocal lattice vectors. The number of energy levels in a band (i.e. the number of values permitted to \mathbf{k}) is equal to the number of primitive cells in the crystal, just as was the case for the orbital model in the tight-binding approximation. For each energy level corresponding to a particular value of \mathbf{k} the Pauli principle permits two electrons of opposite spin to be assigned. This process is repeated for the different bands until all the electrons have been allocated. The energy level of the highest occupied state is called the *Fermi energy* (for a metal; for an insulator, the Fermi energy is in the middle of a band gap). When all the electrons have been assigned then one of two different situations may result. In the first case all the occupied bands are completely filled. As we saw earlier, this gives rise to a band gap between the top of the highest occupied level and the bottom of the lowest empty level. The number of energy levels in each band is equal to the number of primitive cells in the crystal, so a band gap can only arise if there is an even number of electrons per primitive cell. The tight-binding approximation discussed in Section 3.8.2 may be an appropriate model to apply in this case. The second situation arises when one or more bands are partially filled. For each of these partially filled bands one can consider there to be a surface in the \mathbf{k} space that separates the occupied and the unoccupied levels, as defined by the Fermi energy. This set of surfaces is known as the *Fermi surface* and it defines a border between the occupied and unoccupied states. In many cases the Fermi surface is contained within a single band; if not, then the parts of the Fermi surface due to partially filled individual bands are known as the *branches* of the Fermi surface. The

Fermi surface will show the same underlying periodicity as the reciprocal lattice. A particularly attractive feature of the Fermi surface is that it can be measured experimentally, so providing a link between theory and experiment.

The *density of levels* is another useful way to describe the electronic structure of a solid. The density of levels indicates how many energy levels there are for a particular energy. It can thus be defined as the number of levels between E and $E + dE$. This is often normalised by volume, leading to the density of levels per unit volume $g(E)$, which is given by:

$$g(E) = \sum_n g_n(E) \quad (3.100)$$

The sum is over the bands n , with $g_n(E)$ being the density of levels in the band n :

$$g_n(E) = \frac{1}{4\pi^3} \int \delta(E - E_n(\mathbf{k})) d\mathbf{k} \quad (3.101)$$

The delta function $\delta(E - E_n(\mathbf{k}))$ has a value of 1 if $E_n(\mathbf{k})$ is in the range E to $E + dE$ and 0 otherwise. The density of states $D(E)$ is closely related to the density of levels; in the simple case where we have two electrons in each level then the density of states is just twice the density of levels. The integral of the density of states up to the Fermi level is equal to the number of electrons and the integral of the density of states multiplied by the energy is the total electronic energy:

$$N = \int D(E) dE \quad (3.102)$$

$$E_{\text{tot}} = \int D(E) E dE \quad (3.103)$$

The density of states can be usefully visualised by plotting the energy versus $D(E)$. For the simple one-dimensional situation where the energy varies in a cosine-like manner with \mathbf{k} and the levels are equally spaced, the density of states is greatest at the top and bottom of the band (Figure 3.20). The density of states is thus inversely proportional to

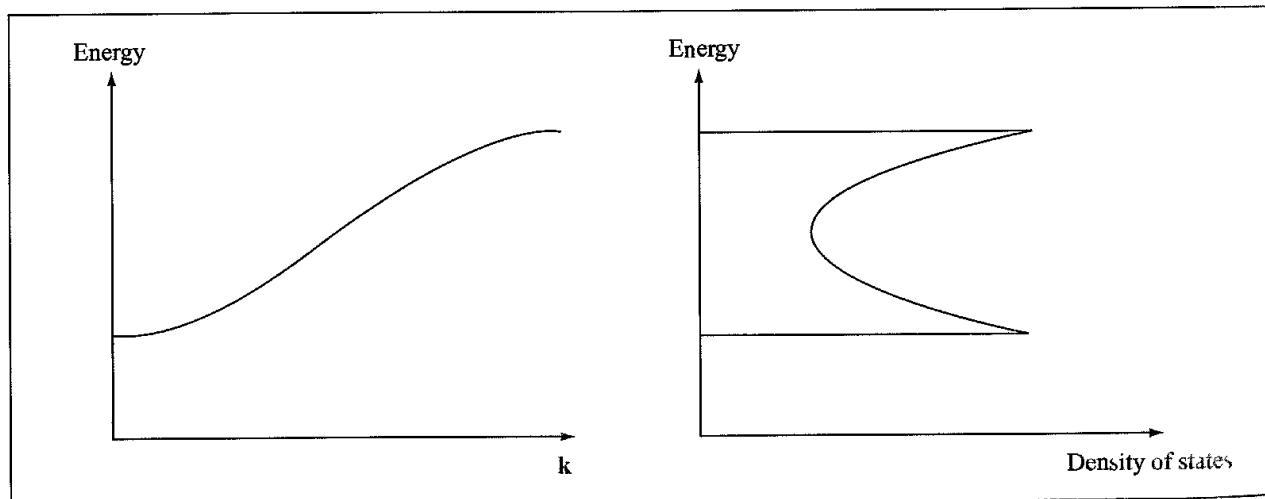


Fig 3.20. Variation of the density of states, $D(E)$, for the simple 1D lattice, shown with the corresponding energy diagram

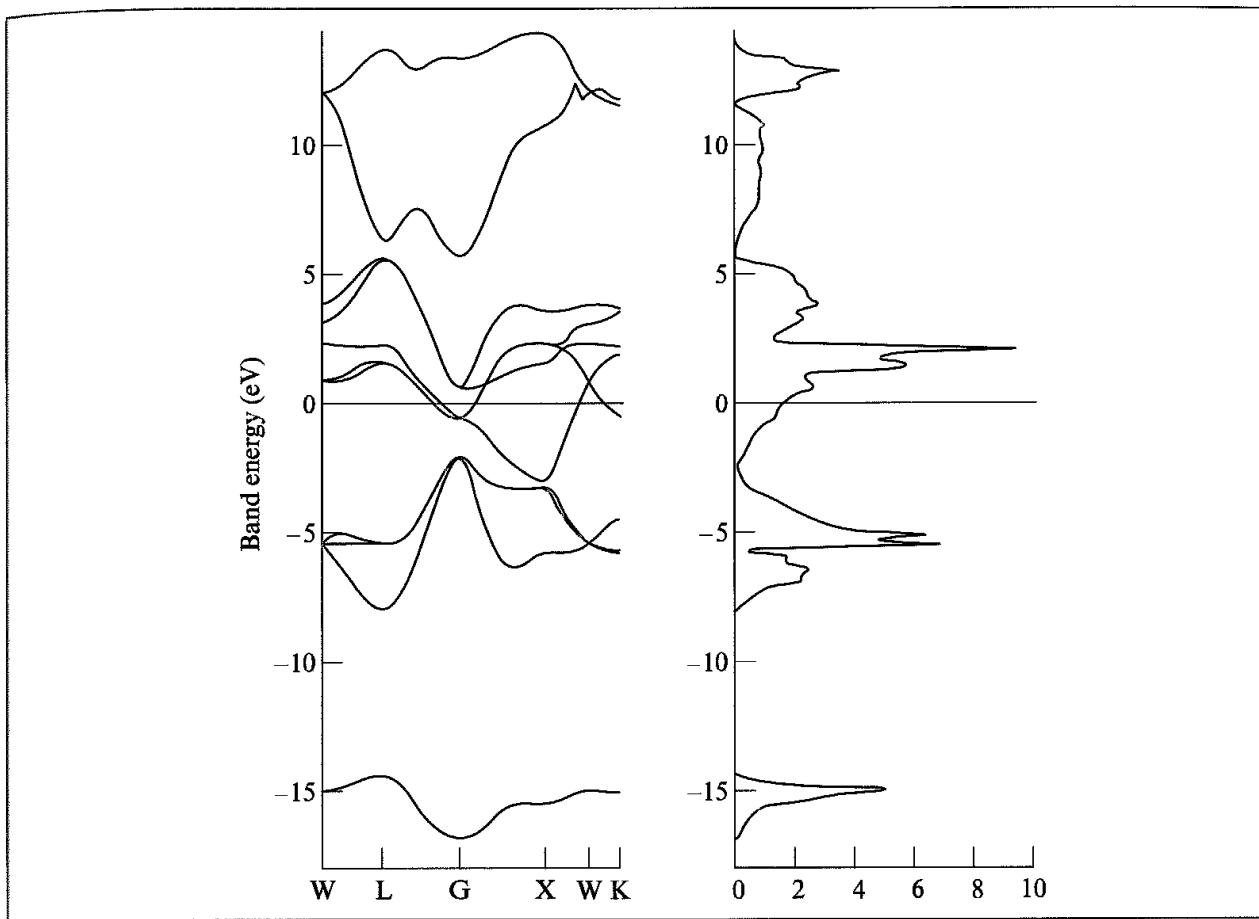


Fig. 3.21. Band structure and density of states for TiN.

the slope of the energy versus \mathbf{k} curve; the flatter the band the greater the density of states at that energy.

The density of states is somewhat like an orbital energy diagram, but unlike the latter does not contain well-defined individual energy levels. Nevertheless, in some situations it is possible to determine from which atomic orbitals a particular energy band is largely derived. Of course, most real systems have rather more complex electronic structures than the simple cases we have used to discuss the background, as illustrated in Figure 3.21, which shows the band structure and density of states diagram for TiN.

3.8.6 Density Functional Methods for Studying the Solid State: Plane Waves and Pseudopotentials

Plane waves are often considered the most obvious basis set to use for calculations on periodic systems, not least because this representation is equivalent to a Fourier series, which itself is the natural language of periodic functions. Each orbital wavefunction is expressed as a linear combination of plane waves which differ by reciprocal lattice vectors:

$$\psi_i^{\mathbf{k}}(\mathbf{r}) = \sum_{\mathbf{G}} a_{i,\mathbf{k}+\mathbf{G}} \exp(i(\mathbf{k} + \mathbf{G}) \cdot \mathbf{r}) \quad (3.104)$$

The Kohn–Sham equations of the density functional theory then take on the following form:

$$\sum_{\mathbf{G}'} \left\{ \frac{\hbar^2}{2m} |\mathbf{k} + \mathbf{G}| \delta_{\mathbf{G}\mathbf{G}'} + V_{\text{ion}}(\mathbf{G} - \mathbf{G}') + V_{\text{elec}}(\mathbf{G} - \mathbf{G}') + V_{\text{XC}}(\mathbf{G} - \mathbf{G}') \right\} a_{i,\mathbf{k}+\mathbf{G}'} = \varepsilon_i a_{i,\mathbf{k}+\mathbf{G}'} \quad (3.105)$$

V_{ion} , V_{elec} and V_{XC} represent the electron–nuclei, electron–electron and exchange–correlation functionals, respectively. The delta function $\delta_{\mathbf{G}\mathbf{G}'}$ is zero unless $\mathbf{G} = \mathbf{G}'$, in which case it has a value of 1. There are two potential problems with the practical use of this equation for a ‘macroscopic’ lattice. First, the summation over \mathbf{G}' (a Fourier series) is in theory over an infinite number of reciprocal lattice vectors. In addition, for a macroscopic lattice there are effectively an infinite number of \mathbf{k} points within the first Brillouin zone. Fortunately, there are practical solutions to both of these problems.

We are usually interested in the valence electrons of an atom, as these are largely responsible for the chemical bonding and most physical properties. The core electrons are little affected by the atomic environment. It is therefore common only to consider explicitly the valence electrons in the calculation and to subsume the core electrons into the nuclear core. One potential drawback to the representation of valence electron wavefunctions with a plane-wave basis set is that near to the atomic nuclei the wavefunctions of the valence electrons show rapid oscillations. This is because their wavefunctions must be orthogonal to those of the core electrons. These oscillations give rise to a large kinetic energy, and a very large number of plane waves would be required to properly model this behaviour. This corresponds to taking many terms in the plane-wave expansion of the orbital, Equation (3.104). This problem is compounded by the fact that the solid systems of interest often contain elements much later in the periodic table than are usually encountered in molecular Hartree–Fock calculations. Heavy elements have many more core electrons and so an even more pronounced oscillatory behaviour. However, in this inner region the kinetic energy is largely cancelled by the high electrostatic potential energy of interaction with the nucleus. A popular way to deal with these problems is to replace the ‘true’ potential in these core regions with a much weaker one called a *pseudopotential*. This represents the way in which the valence electrons interact with the combined nucleus plus core electrons [Heine 1970]. A pseudopotential is a potential function that gives wavefunctions with the same shape as the true wavefunction outside the core region but with fewer nodes inside the core region, as illustrated in Figure 3.22. This has the effect of reducing the number of terms required for the plane wave expansion of the wavefunction, which in turn drastically reduces the scale of the computational problem.

Pseudopotentials are usually derived from all-electron atomic calculations. The valence electron pseudopotential is then required to reproduce the behaviour and properties of the valence electrons in the full calculation. For example, the energy levels with the pseudopotential should be the same as for the all-electron calculation. In addition, the pseudopotential will often depend upon the orbital angular momentum of the wavefunction (i.e. for s, p, d, etc. orbitals) and it will be desired that the total valence electron density within the core radius equals that in the all-electron situation. Such pseudopotentials are

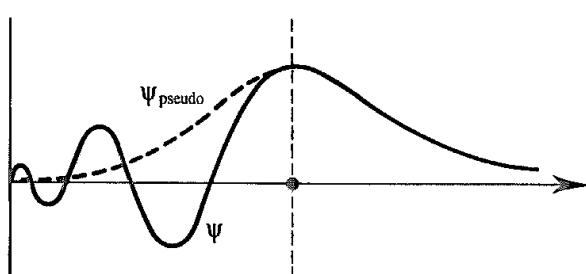


Fig. 3.22 Schematic representation of a pseudopotential (Figure adapted from Payne M C, M P Teter, D C Allan, R A Arias and D J Joannopoulos 1992 Iterative Minimisation Techniques for Ab initio Total-Energy Calculations: Molecular Dynamics and Conjugate Gradients. *Reviews of Modern Physics* 64 1045–1097.)

referred to as ‘non-local norm-conserving’. An additional advantage of the use of pseudopotentials for the heavy elements is that they enable some relativistic effects to be included in the model. A number of functional forms are possible for the pseudopotentials; it is usual to assume a specific functional form and then to vary the parameters. The various pseudopotentials differ in the number of plane waves that are required for their representation and in the degree to which they can be transferred between different atomic environments. So-called ‘soft’ pseudopotentials require fewer plane waves and are therefore computationally more attractive, though there is to some extent a trade-off between softness and transferability. Subsequently developed were the ‘ultrasoft’ or ‘supersoft’ pseudopotentials, which require even fewer plane waves.

In practice, therefore, a pseudopotential is invariably employed and only plane waves with a kinetic energy ($= (\hbar^2/2m)|\mathbf{k} + \mathbf{G}|^2$) less than some cutoff are included in the calculation. The cutoff used depends on the nature of the system under investigation. For example, in the first-row elements the 2p valence orbitals approach closer to the nucleus than the comparable 3p orbitals in the second-row elements (the latter are repelled by the lower 2p states). Thus elements such as silicon or sulphur usually have softer pseudopotentials than their first-row equivalents carbon and oxygen. Everything else being equal, a higher cutoff is consequently required for the latter and hence more plane waves in the expansion (i.e. more reciprocal lattice vectors, \mathbf{G}). Note that in the plane wave expansion the basis functions are not associated with particular atoms but are defined over the whole cell (this also removes the problem of basis-set superposition errors as an additional benefit). The coefficients $a_{i,\mathbf{k}+\mathbf{G}}$ are obtained by following the usual density functional scheme: an initial guess is made of the electron density variation $\rho(\mathbf{r})$, the Kohn-Sham and overlap matrices are constructed, diagonalisation gives the eigenfunctions and eigenvectors (and thus the coefficients a) from which the Kohn-Sham orbitals can be constructed and hence the density for the next iteration.

The second important practical consideration when calculating the band structure of a material is that, in principle, the calculation needs to be performed for all \mathbf{k} vectors in the Brillouin zone. This would seem to suggest that for a macroscopic solid an infinite number of vectors \mathbf{k} would be needed to generate the band structure. However, in practice a discrete sampling over the Brillouin zone is used. This is possible because the wavefunctions at points

that are close together in \mathbf{k} space will be almost identical and can be represented by a single representative point. Each of these discrete values is multiplied by a weight factor related to the volume of reciprocal space it represents. Obviously, the denser the set of \mathbf{k} vectors the smaller will be the error in the calculation. Various schemes have been suggested for selecting suitable sets of \mathbf{k} vectors which can give very accurate approximations to properties such as the charge density; the method of Monkhorst and Pack is particularly popular [Monkhorst and Pack 1976]. The selection of \mathbf{k} vectors is also influenced by the size and shape of the system; indeed, if the unit cell is large then it may only be necessary to consider just one vector. Typically, between ten and 100 vectors are sufficient to understand the structural and electronic properties of a solid, though for certain types of problem such as calculating the optical properties of a metal many more \mathbf{k} vectors may be required (several thousands). Ideally, one should ensure that the calculation converges both in terms of the number of wave-vectors \mathbf{k} considered and in terms of the number of reciprocal lattice vectors \mathbf{G} . An additional consideration is that the symmetry of the Brillouin zone itself may mean that it is not necessary for \mathbf{k} to vary over the entire zone but that only a smaller section need be considered. For example, in our two-dimensional hexagonal close-packed case we would only have to consider the small right-angled triangle over which we undertook our 'tour'. This has an area one-twelfth that of the entire zone. This is an example of the use of the point symmetry of the Brillouin zone rather than the translational symmetry of the lattice. The small section containing the explicit \mathbf{k} vectors required for the calculation is sometimes referred to as the *irreducible part* of the Brillouin zone.

3.8.7 Application of Solid-state Quantum Mechanics to the Group 14 Elements

The combination of density functional methods with pseudopotentials has been used extensively to study a wide variety of materials. Three systems that have been the subject of much interest are the group 14 elements carbon, silicon and germanium, reflecting their natural abundance, commercial importance (especially for silicon) and the large amount of experimental data available. Of particular interest is the problem of predicting the lowest-energy structure at a given volume [Cohen 1986; Mujica and Needs 1993, Needs and Mujica 1995]. In effect, this corresponds to predicting the most stable structure at a particular pressure. These elements all exist in the familiar diamond structure at normal pressures and temperatures but alternative structures can be formed by the application of pressure, at least for silicon and germanium. There has also been much speculation as to whether diamond itself could be transformed should a high enough pressure be applied. This last problem does have some practical interest as it would provide a theoretical upper limit to the pressures that could be achieved with ultra high-pressure diamond anvil cells.

There are many alternatives to the diamond structure, including body-centred cubic, face-centred cubic, hexagonal close-packed, simple hexagonal, simple cubic, β -tin, double-hexagonal close-packed and two complex tetrahedral structures. a body-centred cubic structure with eight atoms per unit cell and a simple tetragonal structure with twelve atoms per unit cell, not forgetting of course the many fullerene forms. Not all studies

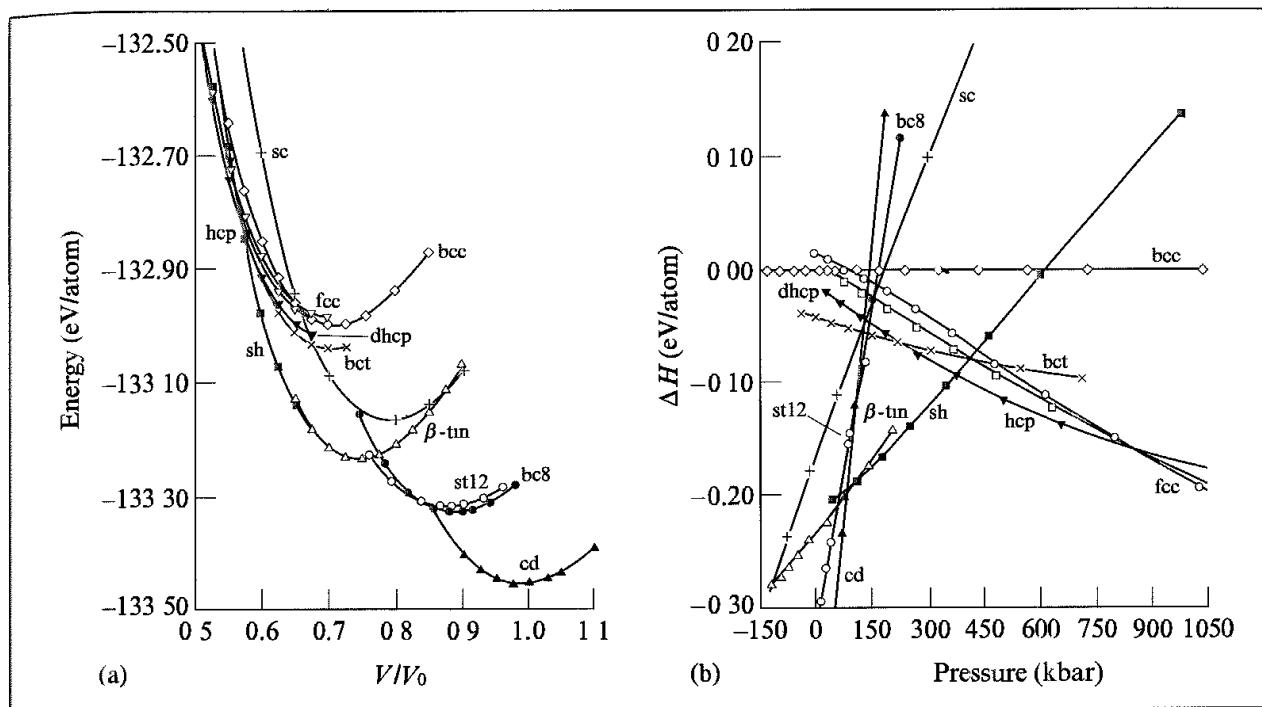


Fig. 3.23. (a) Graph of energy vs volume (scale normalised to the diamond structure) for eleven phases of silicon (b) Enthalpy-pressure plot for the same eleven phases relative to the body-centred cubic phase (Figures redrawn from Needs R J and A Mujica 1995. First-principles pseudopotential study of the structural phases of silicon. Physical Review B51 9652–9660.)

consider every one of these phases but by quoting the list in full we can appreciate the range of possibilities. The energy differences between many of these phases are often small and so it is particularly important to achieve an effective sampling of points in k space (recent studies suggest several thousands of such points are needed). The plane-wave cutoff can also have an effect on the results. The calculations involve minimising each structure at a number of different volumes and then fitting a polynomial to the data points. The results are usually displayed as a graph of the total energy versus the volume, as shown in Figure 3.23. Another way to display this type of data is an enthalpy-pressure plot, from which the most stable phase at any pressure is easily identified as that with the lowest enthalpy. Various bulk structural properties can also be calculated for comparison with experiment.

As we alluded, of the forms mentioned above only the diamond structure has been observed experimentally for carbon. For both silicon and germanium there is a transition to the β -tin phase around 100–130 kbar. Silicon further transforms into other structures such as the simple hexagonal with a relatively modest further increase in pressure, whereas for germanium this transition requires much more pressure. Why should this be, given that they are all in the same group? The electronic structure calculations provide some significant insights into this problem. Thus silicon has a strongly repulsive p-orbital pseudopotential due to the inner (2p) electrons, which carbon does not. This repulsion contributes to the formation of a single peak in the electron density along each Si-Si bond, whereas for carbon there are two peaks, each being near the position for the atomic p orbitals (Figure 3.24). The differences

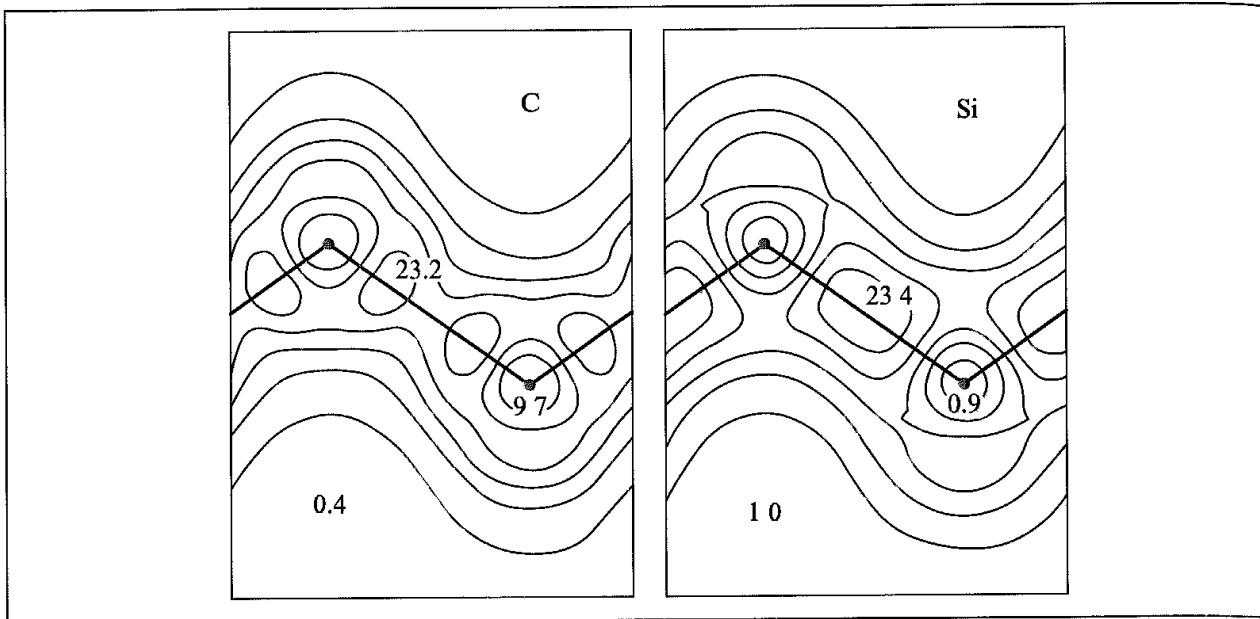


Fig. 3.24: Valence electron density for the diamond structures of carbon and silicon (Figure redrawn from Cohen M L 1986 Predicting New Solids and Superconductors. Science 234 549–553)

between silicon and germanium are ascribed to the d electron states; silicon does not have core d electrons, whereas germanium does. Certain transitions (e.g. carbon $\rightarrow \beta$ -tin) do not depend upon the d character of the electronic configuration in contrast to subsequent transitions.

3.9 The Future Role of Quantum Mechanics: Theory and Experiment Working Together

Of all the methods that we will discuss in this book, quantum mechanics is probably the most widely used and the most extensively developed. The importance of the subject can be gauged in many ways, from citation counts to the number of Nobel prizes awarded. The systems studied using quantum mechanics range from the simplest molecular species (e.g. H_2^+ , HD^+ , H_3^+) to some very large and complex molecules (e.g. DNA, proteins and complex solid-state materials). Some of the most productive situations occur when experiment and theory are used in combination to tackle a problem. The methylene molecule, CH_2 , is of particular historical interest. Despite its small size, this molecule and the controversy surrounding it played an important role in establishing the role of computational quantum mechanical methods in modern-day research and the relationship between theory and experiment [Schaeffer 1986]. The early debate concentrated on the ground state of the molecule and whether its geometry was linear or bent. Early *ab initio* calculations by Foster and Boys [Foster and Boys 1960] suggested an H-C-H angle of 129° but this was refuted by spectroscopic data from Herzberg's laboratory, which were interpreted to indicate a linear geometry. Unfortunately for Foster and Boys, empirical calculations favoured by their head of department, Longuet-Higgins, also gave a linear geometry. Events came to a head when Bender and Schaeffer calculated a geometry of 135.1° and concluded that

the energy barrier between the linear and bent geometries was so large that no further improvement in the theoretical model could remove it. Soon thereafter several other experiments were undertaken, showing a bent structure. Moreover, when Herzberg re-examined his original data it was found to be consistent with a bent model. As we shall see in the remaining chapters there are many kinds of problem that can be tackled using computational chemistry methods. By no means do they always work, but there is often a synergistic relationship between experiment and theory, which means that the two combined can be much more productive than either in isolation.

Appendix 3.1 Alternative Expression for a Wavefunction Satisfying Bloch's Function

We have Equation (3.81):

$$\psi^k(x + a) = e^{ika} \psi^k(x) \quad (3.106)$$

We write $\psi(x)$ as the product of the exponential and a function $u_k(x)$:

$$u_k(x) = \psi_k(x) / \exp(ikx) \quad (3.107)$$

If we perform the same manipulation for $\psi(x + a)$ we get:

$$u_k(x + a) = \frac{\psi_k(x + a)}{e^{ik(x+a)}} = \frac{\psi_k(x) e^{ika}}{e^{ikx} e^{ika}} = \frac{\psi_k(x)}{e^{ikx}} = u_k(x) \quad (3.108)$$

Thus $u_k(x)$ is a periodic function which can be used to formulate acceptable wavefunctions:

$$\psi_k(x) = e^{ikx} u_k(x) \quad (3.109)$$

Further Reading

- Ashcroft N W and N D Mermin 1976. *Solid State Physics* New York, Holt, Rinehart and Winston.
 Atkins P W 1991 *Quanta. A Handbook of Concepts* Oxford, Oxford University Press.
 Atkins P W and R S Friedman 1996. *Molecular Quantum Mechanics*, 3rd edition. Oxford, Oxford University Press.
 Catlow C R A 1997. Computer Modelling as a Technique in Materials Chemistry In Catlow C R A and A K Cheetham (Editors). *New Trends in Materials Chemistry*, NATO ASI Series C 498, Dordrecht, Kluwer.
 Catlow C R A 1998. Solids Computer Modelling. In Schleyer, P v R, N L Allinger, T Clark, J Gasteiger, P A Kollman, H F Schaeffer III and P R Schreiner (Editors) *The Encyclopedia of Computational Chemistry*, Chichester, John Wiley & Sons
 Gillan M J 1991 Calculating the Properties of Materials from Scratch In Meyer M and V Pontikis (Editors) *Computer Simulation*, NATO ASI Series E 205 (Computer Simulations in Materials Science) pp 257–281.
 Hehre W J, L Radom, P v R Schleyer and J A Pople 1986. *Ab initio Molecular Orbital Theory* New York, John Wiley & Sons

- Hoffmann R 1988. *Solids and Surfaces: A Chemist's View on Bonding in Extended Structures*. New York, VCH Publishers.
- Kohn W, A D Becke and R G Parr 1996. Density Functional Theory of Electronic Structure. *Journal of Chemical Physics* **100**:12974–12980.
- Kohn W and P Vashita 1983. General Density Functional Theory. In Lundquist S and N H March (Editors). *Theory of Inhomogeneous Electron Gas*, New York, Plenum, pp. 79–148.
- Kutzelnigg W and P von Herigone 2000. Electron Correlation at the Dawn of the 21st Century. *Advances in Quantum Chemistry* **36**:185–229.
- Pisani C, R Dovesi and C Roetti 1988 Hartree-Fock *Ab initio* Treatment of Crystalline Systems. *Lecture Notes in Chemistry* Vol. 48. Berlin, Springer-Verlag.
- Pisani C, R Dovesi, C Roetti, M Cansa, R Orlando, S Casass and V R Saunders 2000. CRYSTAL and EMBED, Two Computational Tools for the *ab initio* Study of Electronic Properties of Crystals. *International Journal of Quantum Chemistry* **77**: 1032–1048.
- Schaeffer H F III (Editor) 1977. *Applications of Electronic Structure Theory*. New York, Plenum Press.
- Schaeffer H F III (Editor) 1977 *Methods of Electronic Structure Theory*. New York, Plenum Press.
- Szabo A and N S Ostlund 1982. *Modern Quantum Chemistry Introduction to Advanced Electronic Structure Theory* New York, McGraw-Hill.
- Wimmer E 1991 Density Functional Theory for Solids, Surface and Molecules: from Energy Bands to Molecular Bonds In Labanowski J R and J W Andzelm (Editors) *Density Functional Methods in Chemistry*. Berlin, Springer-Verlag, pp 7–31

References

- Almlöf J, K Faegri Jr and K Korsell 1982. Principles for a Direct SCF Approach to LCAO-MO *Ab initio* Calculations. *Journal of Computational Chemistry* **3**:385–399
- Ashcroft N W and N D Mermin 1976 *Solid State Physics* New York, Holt, Rinehart & Winston.
- Baboul A G, L A Curtiss, P C Redfern and K Raghavachari 1999. Gaussian-3 Theory Using Density Functional Geometries and Zero-Point Energies. *Journal of Chemical Physics*, **110**:7650–7657.
- Becke A D 1988. Density-functional Exchange-energy Approximation with Correct Asymptotic Behaviour. *Physical Review A* **38** 3098–3100.
- Becke A D 1992 Density-functional Thermochemistry. I. The Effect of the Exchange-only Gradient Correction. *Journal of Chemical Physics* **96**:2155–2160
- Becke A D 1993a. A New Mixing of Hartree-Fock and Local Density-functional Theories. *Journal of Chemical Physics* **98** 1372–1377
- Becke A D 1993b. Density-functional Thermochemistry. III The Role of Exact Exchange. *Journal of Chemical Physics* **98**:5648–5652.
- Becke A D and R M Dickson 1990. Numerical Solution of the Schroedinger Equation in Polyatomic Molecules. *Journal of Chemical Physics* **92**:3610–3612
- Bobrowicz F W and W A Goddard III 1977 The Self-Consistent Field Equations for Generalized Valence Bond and Open-Shell Hartree-Fock Wave Functions In Schaeffer H F III (Editor). *Modern Theoretical Chemistry III*, New York, Plenum, pp. 79–127.
- Boys S F and F Bernardi 1970 The Calculation of Small Molecular Interactions by the Differences of Separate Total Energies Some Procedures with Reduced Errors. *Molecular Physics* **19**:553–566
- Bradley C J and A P Cracknell 1972 *The Mathematical Theory of Symmetry in Solids* Oxford, Clarendon Press.
- Ceperley D M and B J Alder 1980 Ground State of the Electron Gas by a Stochastic Method. *Physical Review Letters* **45**:566–569.

- Cohen M L 1986. Predicting New Solids and Superconductors. *Science* **234**:549–553
- Cooper D L, J Gerratt and M Raimondi 1986. The Electronic Structure of the Benzene Molecule. *Nature* **323**, 699–701.
- Curtiss L A, K Raghavachari, G W Trucks and J A Pople 1991 Gaussian-2 Theory for Molecular Energies of First- and Second-row Compounds. *Journal of Chemical Physics* **94**:7221–7230.
- Curtiss L A, K Raghavachari, P C Redfern, V Rassolov and J A Pople 1998. Gaussian-3 (G3) Theory for Molecules Containing First and Second-row Atoms. *Journal of Chemical Physics* **109**:7764–7776
- Curtiss L A, P C Redfern, K Raghavachari, V Rassolov and J A Pople 1999. Gaussian-3 Theory Using Reduced Møller-Plesset Order. *Journal of Chemical Physics* **110**:4703–4709
- Dovesi R, R Orlando, C Roetti, C Pisani and V R Saunders 2000. The Periodic Hartree-Fock Method and Its Implementation in the CRYSTAL Code. *Physica Status Solidi* **B217**.63–88.
- Dovesi R, C Pisani, C Roetti and V R Saunders 1983 Treatment of Coulomb Interactions in Hartree-Fock Calculations of Periodic-Systems *Physical Review* **B28**:5781–5792.
- Foster J M and S F Boys 1960. Quantum Variational Calculations for a Range of CH₂ Configurations *Reviews in Modern Physics* **32**:305–307.
- Frisch M J, G W Trucks and J R Cheeseman 1996. Systematic Model Chemistries Based on Density Functional Theory: Comparison with Traditional Models and with Experiment. *Theoretical and Computational Chemistry (Recent Developments and Applications of Modern Density Functional Theory)* **4**:679–707.
- Gerratt J, D L Cooper, P B Karadakov and M Raimondi 1997. Modern Valence Bond Theory. *Chemical Society Reviews* pp 87–100.
- Gunnarsson O and B I Lundqvist 1976. Exchange and Correlation in Atoms, Molecules, and Solids by the Spin-density-functional Formalism. *Physical Review* **B13**:4274–4298.
- Heine V 1970. The Pseudopotential Concept *Solid State Physics* **24**:1–36
- Heitler W and F London 1927 Wechselwirkung neutraler Atome und Homöopolare Bindung nach der Quantenmechanik *Zeitschrift für Physik* **44**:455–472.
- Hohenberg P and Kohn W 1964. Inhomogeneous Electron Gas. *Physical Review* **B136**:864–871.
- Johnson B G, P M W Gill and J A Pople 1993. The performance of a family of density functional methods. *Journal of Chemical Physics* **98** 5612–5626
- Kohn W and L J Sham 1965. Self-consistent Equations Including Exchange and Correlation Effects. *Physical Review* **A140**:1133–1138.
- Lee C, W Yang and R G Parr 1988 Development of the Colle-Salvetti Correlation Energy Formula into a Functional of the Electron Density *Physical Review* **B37**:785–789.
- Møller C and M S Plesset 1934. Note on an Approximate Treatment for Many-Electron Systems *Physical Review* **46**:618–622.
- Monkhorst H J and J D Pack 1976 Special Points for Brillouin-zone Integration. *Physical Review* **B13**:5188–5192.
- Morokuma K 1977. Why Do Molecules Interact? The Origin of Electron Donor-Acceptor Complexes, Hydrogen Bonding, and Proton Affinity. *Accounts of Chemical Research* **10** 294–300
- Mujica A and R J Needs 1993. First-principles Calculations of the Structural Properties, Stability, and Band Structure of Complex Tetrahedral Phases of Germanium. ST12 and BC8. *Physical Review* **B48**:17010–17017.
- Needs R J and Mujica 1995. First-principles Pseudopotential Study of the Structural Phases of Silicon. *Physical Review* **B51**:9652–9660.
- Parr R G 1983 Density Functional Theory. *Annual Review of Physical Chemistry* **34**:631–656
- Ferdew J P and A Zunger 1981. Self-Interaction Correction to Density-Functional Approximations for Many-Electron Systems. *Physical Review* **B23**:5048–5079.
- Pisani C and R Dovesi 1980. Exact-Exchange Hartree-Fock Calculations for Periodic Systems. I Illustration of the Method. *International Journal of Quantum Chemistry* **XVII**:501–516.

- Pople J A, M Head-Gordon and K Raghavachari 1987. Quadratic Configuration Interaction A General Technique for Determining Electron Correlation Energies. *Journal of Chemical Physics* **87**:5968–5975
- Pople J A, M Head-Gordon, D J Fox, K Raghavachari and L A Curtiss 1989 Gaussian-1 Theory: A General Procedure for Prediction of Molecular Energies *Journal of Chemical Physics* **90**:5622–5629
- Pople J A and R K Nesbet 1954. Self-consistent Orbitals for Radicals. *Journal of Chemical Physics* **22**:571–572
- Pulay P 1977 Direct Use of the Gradient for Investigating Molecular Energy Surfaces In Schaeffer H F III (Editor). *Applications of Electronic Structure Theory*, New York, Plenum, pp 153–185.
- Pulay P 1980. Convergence Acceleration of Iterative Sequences. The Case of SCF Iteration *Chemical Physics Letters* **73**:393–398.
- Pulay P 1987 Analytical Derivative Methods in quantum Chemistry In Lawley K P (Editor) *Ab initio Methods in Quantum Chemistry - II*, New York, John Wiley & Sons, pp 241–286.
- Roos B O, P R Taylor and E M Siegbahn 1980 A Complete Active Space SCF Method (CASSCF) Using a Density Matrix Formulated Super-CI Approach *Chemical Physics* **48** 157–173.
- Schaeffer H F III 1986. Methylenes: A Paradigm for Computational Quantum Chemistry *Science* **231**:1100–1107
- Sim F, St-Amant A, I Papai and D R Salahub 1992. Gaussian Density Functional Calculations on Hydrogen-Bonded Systems *Journal of the American Chemical Society* **114**:4391–4400
- Slater J C 1974. *Quantum Theory of Molecules and Solids Volume 4: The Self-Consistent Field for Molecules and Solids*. New York, McGraw-Hill
- Smith B J, D J Swanton, J A Pople, H F Schaeffer III and L Radom 1990. Transition Structures for the Interchange of Hydrogen Atoms within the Water Dimer. *Journal of Chemical Physics* **92**:1240–1247.
- St-Amant A, W D Cornell, P A Kollman and T A Halgren 1995. Calculation of Molecular Geometries, Relative Conformational Energies, Dipole Moments and Molecular Electrostatic Potential Fitted Charges of Small Organic Molecules of Biochemical Interest by Density Functional Theory. *Journal of Computational Chemistry* **16** 1483–1506
- Stephens P J, F J Devlin, C F Chabalowski and M J Frisch 1994. *Ab Initio* Calculation of Vibrational Absorption and Circular Dichroism Spectra Using Density Functional Force Fields. *Journal of Physical Chemistry* **98**:11623–11627
- Umeyama H and K Morokuma 1977. The Origin of Hydrogen Bonding An Energy Decomposition Study *Journal of the American Chemical Society* **99**:1316–1332
- Vosko S H, L Wilk and M Nusair 1980. Accurate Spin-dependent Electron Liquid Correlation Energies for Local Spin Density Calculations: A Critical Analysis *Canadian Journal of Physics* **58**:1200–1211.
- Wimmer E 1997. Electronic Structure Methods. In Catlow C R A and A K Cheetham (Editors). *New Trends in Materials Chemistry*, NATO ASI Series C **498** Dordrecht, Kluwer.

Empirical Force Field Models: Molecular Mechanics

4.1 Introduction

Many of the problems that we would like to tackle in molecular modelling are unfortunately too large to be considered by quantum mechanics. Quantum mechanical methods deal with the electrons in a system, so that even if some of the electrons are ignored (as in the semi-empirical schemes) a large number of particles must still be considered, and the calculations are time-consuming. Force field methods (also known as molecular mechanics) ignore the electronic motions and calculate the energy of a system as a function of the nuclear positions only. Molecular mechanics is thus invariably used to perform calculations on systems containing significant numbers of atoms. In some cases force fields can provide answers that are as accurate as even the highest-level quantum mechanical calculations, in a fraction of the computer time. However, molecular mechanics cannot of course provide properties that depend upon the electronic distribution in a molecule.

That molecular mechanics works at all is due to the validity of several assumptions. The first of these is the Born–Oppenheimer approximation, without which it would be impossible to contemplate writing the energy as a function of the nuclear coordinates at all. Molecular mechanics is based upon a rather simple model of the interactions within a system with contributions from processes such as the stretching of bonds, the opening and closing of angles and the rotations about single bonds. Even when simple functions (e.g. Hooke's law) are used to describe these contributions the force field can perform quite acceptably. Transferability is a key attribute of a force field, for it enables a set of parameters developed and tested on a relatively small number of cases to be applied to a much wider range of problems. Moreover, parameters developed from data on small molecules can be used to study much larger molecules such as polymers.

4.1.1 A Simple Molecular Mechanics Force Field

Many of the molecular modelling force fields in use today for molecular systems can be interpreted in terms of a relatively simple four-component picture of the intra- and inter-molecular forces within the system. Energetic penalties are associated with the deviation of bonds and angles away from their 'reference' or 'equilibrium' values, there is a function

that describes how the energy changes as bonds are rotated, and finally the force field contains terms that describe the interaction between non-bonded parts of the system. More sophisticated force fields may have additional terms, but they invariably contain these four components. An attractive feature of this representation is that the various terms can be ascribed to changes in specific internal coordinates such as bond lengths, angles, the rotation of bonds or movements of atoms relative to each other. This makes it easier to understand how changes in the force field parameters affect its performance, and also helps in the parametrisation process. One functional form for such a force field that can be used to model single molecules or assemblies of atoms and/or molecules is:

$$\begin{aligned} \mathcal{V}(\mathbf{r}^N) = & \sum_{\text{bonds}} \frac{k_i}{2} (l_i - l_{i,0})^2 + \sum_{\text{angles}} \frac{k_i}{2} (\theta_i - \theta_{i,0})^2 + \sum_{\text{torsions}} \frac{V_n}{2} (1 + \cos(n\omega - \gamma)) \\ & + \sum_{i=1}^N \sum_{j=i+1}^N \left(4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^6 \right] + \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \right) \end{aligned} \quad (4.1)$$

$\mathcal{V}(\mathbf{r}^N)$ denotes the potential energy, which is a function of the positions (\mathbf{r}) of N particles (usually atoms). The various contributions are schematically represented in Figure 4.1. The first term in Equation (4.1) models the interaction between pairs of bonded atoms, modelled here by a harmonic potential that gives the increase in energy as the bond length l_i deviates from the reference value $l_{i,0}$. The second term is a summation over all valence angles in the molecule, again modelled using a harmonic potential (a valence angle is the angle formed between three atoms A–B–C in which A and C are both bonded to B). The third term in Equation (4.1) is a torsional potential that models how the energy changes as a bond rotates. The fourth contribution is the non-bonded term. This is calculated between all pairs of atoms (i and j) that are in different molecules or that are in

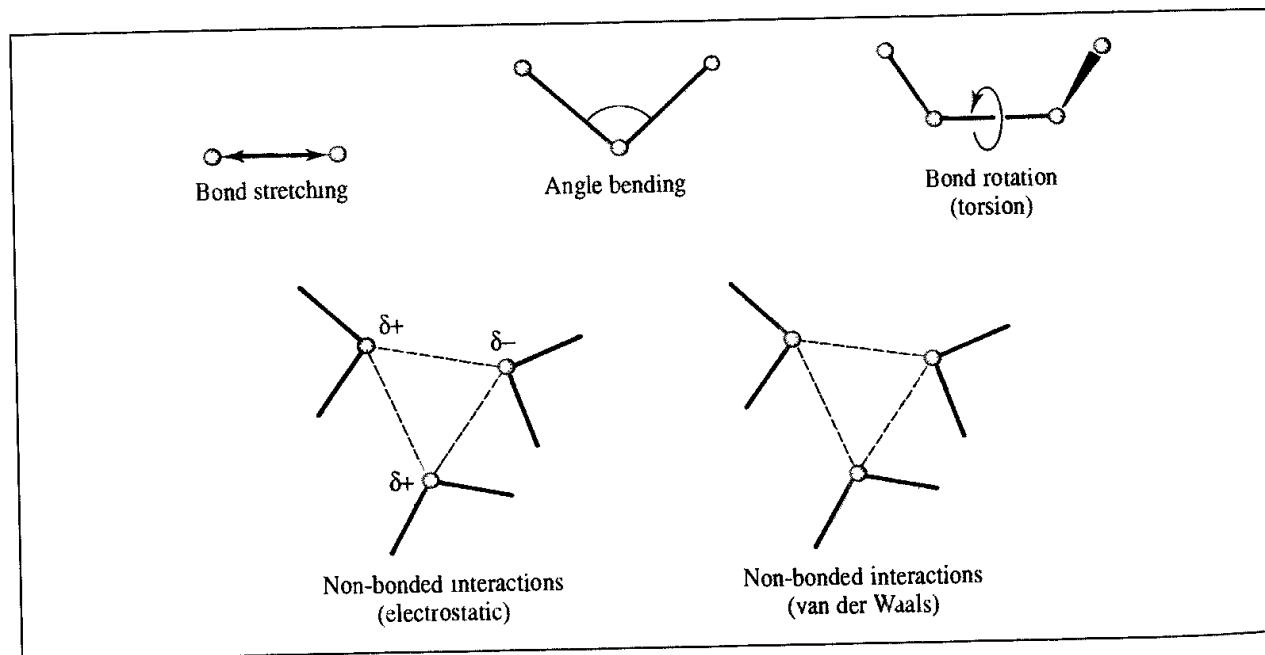


Fig 4.1 Schematic representation of the four key contributions to a molecular mechanics force field bond stretching, angle bending and torsional terms and non-bonded interactions

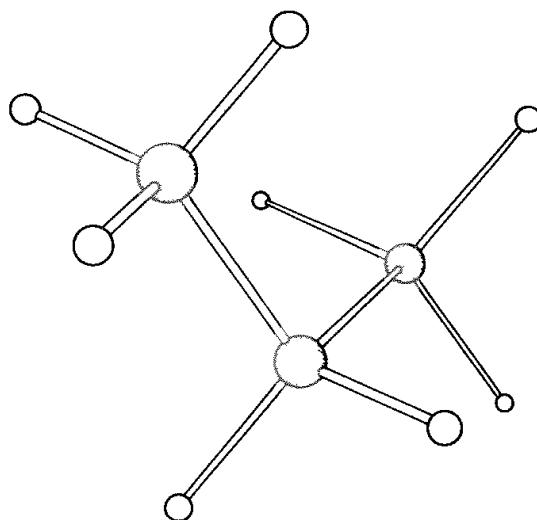


Fig. 4.2. A typical force field model for propane contains ten bond-stretching terms, eighteen angle-bending terms, eighteen torsional terms and 27 non-bonded interactions

the same molecule but separated by at least three bonds (i.e. have a $1, n$ relationship where $n \geq 4$). In a simple force field the non-bonded term is usually modelled using a Coulomb potential term for electrostatic interactions and a Lennard-Jones potential for van der Waals interactions.

We shall discuss the nature of these different contributions in more detail in Sections 4.3–4.10, but here we consider how the simple force field of Equation (4.1) would be used to calculate the energy of a conformation of propane (Figure 4.2). Propane has ten bonds: two C–C bonds and eight C–H bonds. The C–C bonds are symmetrically equivalent but the C–H bonds fall into two classes, one group corresponding to the two hydrogens bonded to the central methylene (CH_2) carbon and one group corresponding to the six hydrogens bonded to the methyl carbons. In some sophisticated force fields different parameters would be used for these two different types of C–H bond, but in most force fields the same bonding parameters (i.e. k_i and $l_{i,0}$) would be used for each of the eight C–H bonds. This is an example of the way in which the same parameters can be used for a wide variety of molecules. There are 18 different valence angles in propane, comprising one C–C–C angle, ten C–C–H angles and seven H–C–H angles. Note that all angles are included in the force field model even though some of them may not be independent of the others. There are 18 torsional terms: twelve H–C–C–H torsions and six H–C–C–C torsions. Each of these is modelled with a cosine series expansion that has minima at the *trans* and *gauche* conformations. Finally, there are 27 non-bonded terms to calculate, comprising 21 H–H interactions and six H–C interactions. The electrostatic contribution would be calculated using Coulomb's law from partial atomic charges associated with each atom and the van der Waals contribution as a Lennard-Jones potential with appropriate ϵ_{ij} and σ_{ij} parameters. A sizeable number of terms are thus included in the force field model, even for a molecule as simple as propane. Even so, the number of terms (73) is many fewer than the number of integrals that would be involved in an equivalent *ab initio* quantum mechanical calculation.

4.2 Some General Features of Molecular Mechanics Force Fields

To define a force field one must specify not only the functional form but also the parameters (i.e. the various constants such as k_i , V_n and σ_{ij} in Equation (4.1)); two force fields may use an identical functional form yet have very different parameters. Moreover, force fields with the same functional form but different parameters, and force fields with different functional forms, may give results of comparable accuracy. A force field should be considered as a single entity; it is not strictly correct to divide the energy into its individual components, let alone to take some of the parameters from one force field and mix them with parameters from another force field. Nevertheless, some of the terms in a force field are sufficiently independent of the others (particularly the bond and angle terms) to make this an acceptable approximation in certain cases.

The force fields used in molecular modelling are primarily designed to reproduce structural properties but they can also be used to predict other properties, such as molecular spectra. However, molecular mechanics force fields can rarely predict spectra with great accuracy (although the more recent molecular mechanics force fields are much better in this regard). A force field is generally designed to predict certain properties and will be parametrised accordingly. While it is useful to try to predict other quantities which have not been included in the parametrisation process it is not necessarily a failing if a force field is unable to do so.

Transferability of the functional form and parameters is an important feature of a force field. Transferability means that the same set of parameters can be used to model a series of related molecules, rather than having to define a new set of parameters for each individual molecule. For example, we would expect to be able to use the same set of parameters for all *n*-alkanes. Transferability is clearly important if we want to use the force field to make predictions. Only for some small systems, where particularly accurate work is required, may it be desirable to develop a model specific to that molecule.

One important point that we should bear in mind as we undertake a deeper analysis of molecular mechanics is that force fields are *empirical*; there is no 'correct' form for a force field. Of course, if one functional form is shown to perform better than another it is likely that form will be favoured. Most of the force fields in common use do have a very similar form, and it is tempting to assume that this must therefore be the optimal functional form. Certainly such models tend to conform to a useful picture of the interactions present in a system, but it should always be borne in mind that there may be better forms, particularly when developing a force field for new classes of molecule. The functional forms employed in molecular mechanics force fields are often a compromise between accuracy and computational efficiency; the most accurate functional form may often be unsatisfactory for efficient computation. As the performance of computers increases so it becomes possible to incorporate more sophisticated models. An additional consideration is that in order to use techniques such as energy minimisation and molecular dynamics, it is usually desirable to be able to calculate the first and second derivatives of the energy with respect to the atomic coordinates.

A concept that is common to most force fields is that of an *atom type*. When preparing the input for a quantum mechanics calculation it is usually necessary to specify the atomic numbers of the nuclei present, together with the geometry of the system and the overall charge and spin multiplicity. For a force field the overall charge and spin multiplicity are not explicitly required, but it is usually necessary to assign an atom type to each atom in the system. The atom type is more than just the atomic number of an atom; it usually contains information about its hybridisation state and sometimes the local environment. For example, it is necessary in most force fields to distinguish between sp^3 -hybridised carbon atoms (which adopt a tetrahedral geometry), sp^2 -hybridised carbons (which are trigonal) and sp -hybridised carbons (which are linear). Each force field parameter is expressed in terms of these atom types, so that the reference angle θ_0 for a tetrahedral carbon atom would be near 109.5° and that for a trigonal carbon would be near 120° . The atom types in some force fields reflect the neighbouring environment as well as the hybridisation and can be quite extensive for some atoms. For example, the MM2, MM3 and MM4 force fields of Allinger and co-workers that are widely used for calculations on 'small' molecules [Allinger 1977; Allinger *et al.* 1989, 1990a, b, 1996a, b; Lii and Allinger 1989; Nevins *et al.* 1996a, b, c] distinguish the following types of carbon atom: sp^3 , sp^2 , sp , carbonyl, cyclopropane, radical, cyclopropene and carbonium ion. In the AMBER force field of Kollman and co-workers [Weiner *et al.* 1984; Cornell *et al.* 1995] the carbon atom at the junction between a six- and a five-membered ring (e.g. in the amino acid tryptophan) is assigned

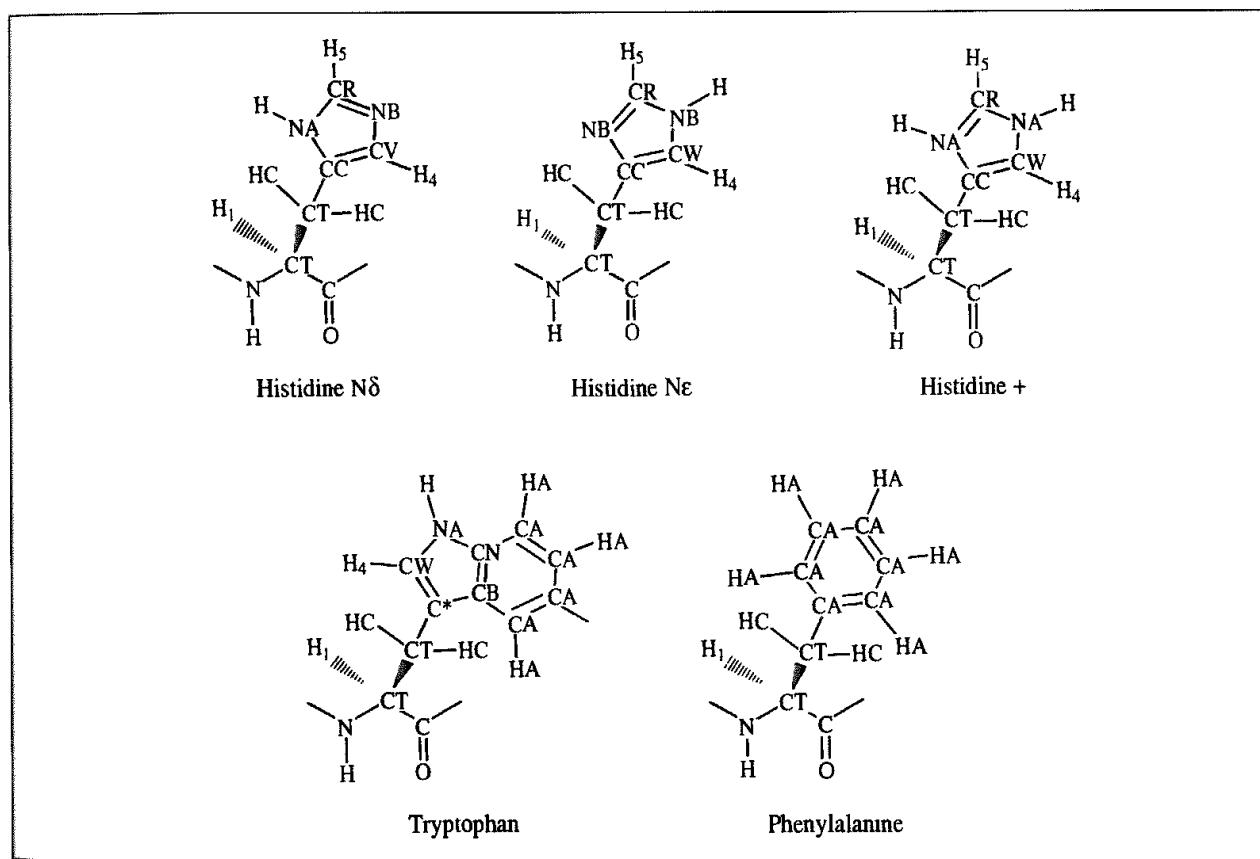


Fig. 4.3 AMBER atom types for the amino acids histidine, tryptophan and phenylalanine. There are three possible protonation states of histidine.

an atom type that is different from the carbon atom in an isolated five-membered ring such as histidine, which in turn is different from the atom type of a carbon atom in a benzene ring. Indeed, the AMBER force field uses different atom types for a histidine amino acid depending upon its protonation state (Figure 4.3). Other, more general, force fields would assign these atoms to the same generic 'sp² carbon' atom type. It is often found that force fields which are designed for modelling specific classes of molecule (such as proteins and nucleic acids, in the case of AMBER) use more specific atom types than force fields designed for general-purpose use.

We now discuss in some detail the individual contributions to a molecular mechanics force field, giving a selection of the various functional forms that are in common use. We shall then consider the important task of parametrisation, in which values for the many force constants are derived. Our discussion will be illuminated by examples chosen from contemporary force fields in widespread use and the MM2/MM3/MM4 and AMBER force fields in particular.

4.3 Bond Stretching

The potential energy curve for a typical bond has the form shown in Figure 4.4. Of the many functional forms used to model this curve, that suggested by Morse is particularly useful. The Morse potential has the form:

$$\nu(l) = D_e \{1 - \exp[-a(l - l_0)]\}^2 \quad (4.2)$$

D_e is the depth of the potential energy minimum and $a = \omega\sqrt{\mu/2D_e}$, where μ is the reduced mass and ω is the frequency of the bond vibration. ω is related to the stretching constant of the bond, k , by $\omega = \sqrt{k/\mu}$. l_0 is the reference value of the bond. The Morse potential is not usually used in molecular mechanics force fields. In part this is because it is not particularly amenable to efficient computation but also because it requires three parameters to be specified for each bond. Moreover, it is rare in molecular mechanics calculations for

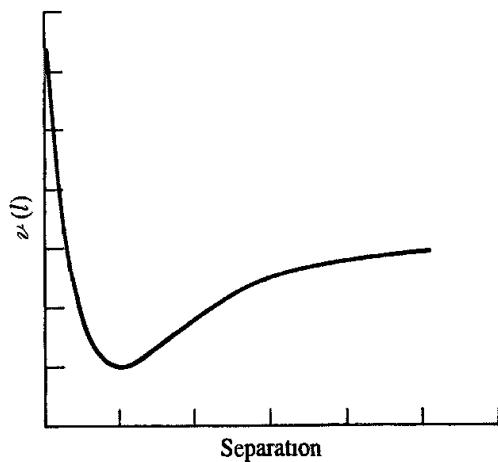


Fig. 4.4: Variation in bond energy with interatomic separation.

bonds to deviate significantly from their equilibrium values; the Morse curve describes a wide range of behaviour from the strong equilibrium behaviour to dissociation. Consequently, simpler expressions are often used. The most elementary approach is to use a Hooke's law formula in which the energy varies with the square of the displacement from the reference bond length l_0 :

$$\nu(l) = \frac{k}{2}(l - l_0)^2 \quad (4.3)$$

The astute reader will have noticed our use of the term 'reference bond length' (sometimes called the 'natural bond length') for the parameter l_0 . This parameter is commonly called the 'equilibrium' bond length, but to do so can be misleading. The reference bond length is the value that the bond adopts when all other terms in the force field are set to zero. The equilibrium bond length, by contrast, is the value that is adopted in a minimum energy structure, when all other terms in the force field contribute. The complex interplay between the various components in the force field means that the bond may well deviate slightly from its reference value in order to compensate for other contributions to the energy. It is also important to recognise that 'real' molecules undergo vibrational motion (even at absolute zero, there is a zero-point energy due to vibrational motion). A true bond-stretching potential is not harmonic but has a shape similar to that in Figure 4.4, which means that the 'average' length of the bond in a vibrating molecule will deviate from the equilibrium value for the hypothetical motionless state. The effects are usually small, but they are significant if one wishes to predict bond lengths to thousandths of an ångström. When comparing the results of calculations with experimental data, one must also remember that different experimental techniques measure different 'equilibrium' values, especially when the experiments are performed at different temperatures. The errors in experimentally determined bond lengths can be quite large; for example, libration of a molecule in a crystal means that the bond lengths determined by X-ray methods at room temperature may have errors as large as 0.015 Å. MM2 was parametrised to fit the values obtained by electron diffraction, which give the mean distances between atoms averaged over the vibrational motion at room temperature.

The forces between bonded atoms are very strong and considerable energy is required to cause a bond to deviate significantly from its equilibrium value. This is reflected in the magnitude of the force constants for bond stretching; some typical values from the MM2 force field are shown in Table 4.1, where it can be seen that those bonds one would

Bond	l_0 (Å)	k (kcal mol ⁻¹ Å ⁻²)
Csp ³ —Csp ³	1.523	317
Csp ³ —Csp ²	1.497	317
Csp ² =Csp ²	1.337	690
Csp ² =O	1.208	777
Csp ³ —Nsp ³	1.438	367
C—N (amide)	1.345	719

Table 4.1 Force constants and reference bond lengths for selected bonds [Allinger 1977]

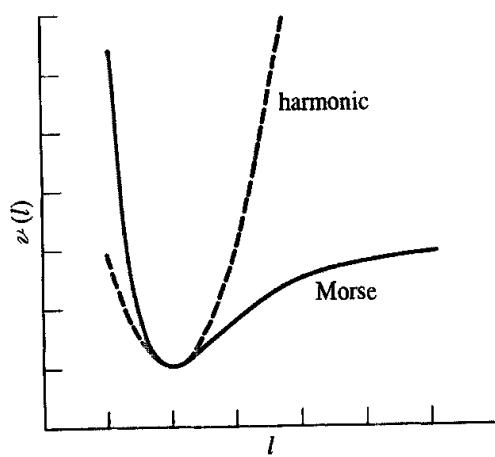


Fig. 4.5: Comparison of the simple harmonic potential (Hooke's law) with the Morse curve

intuitively expect to be stronger have large force constants (contrast C–C with C=C and N≡N). A deviation of just 0.2 Å from the reference value l_0 with a force constant of 300 kcal mol⁻¹ Å⁻² would cause the energy of the system to rise by 12 kcal/mol.

The Hooke's law functional form is a reasonable approximation to the shape of the potential energy curve at the bottom of the potential well, at distances that correspond to bonding in ground-state molecules. It is less accurate away from equilibrium (Figure 4.5). To model the Morse curve more accurately, cubic and higher terms can be included and the bond-stretching potential can be written as follows:

$$v(l) = \frac{k}{2}(l - l_0)^2[1 - k'(l - l_0) - k''(l - l_0)^2 - k'''(l - l_0)^3 \dots] \quad (4.4)$$

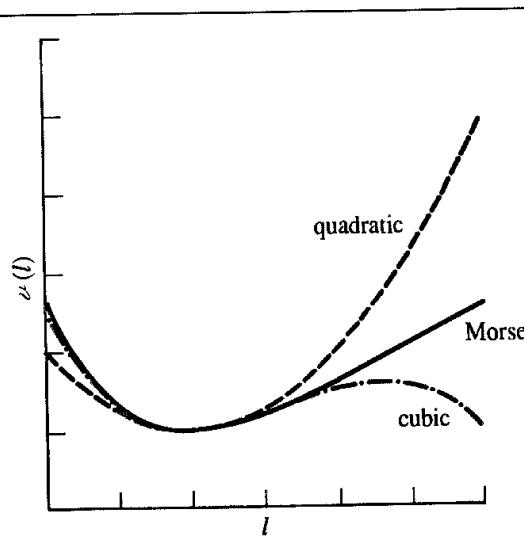


Fig. 4.6: A cubic bond-stretching potential passes through a maximum but gives a better approximation to the Morse curve close to the equilibrium structure than the quadratic form

An undesirable side-effect of an expansion that includes just a quadratic and a cubic term (as is employed in MM2) is that, far from the reference value, the cubic function passes through a maximum. This can lead to a catastrophic lengthening of bonds (Figure 4.6). One way to accommodate this problem is to use the cubic contribution only when the structure is sufficiently close to its equilibrium geometry and is well inside the ‘true’ potential well. MM3 also includes a quartic term; this eliminates the inversion problem and leads to an even better description of the Morse curve.

4.4 Angle Bending

The deviation of angles from their reference values is also frequently described using a Hooke’s law or harmonic potential:

$$\nu(\theta) = \frac{k}{2}(\theta - \theta_0)^2 \quad (4.5)$$

The contribution of each angle is characterised by a force constant and a reference value. Rather less energy is required to distort an angle away from equilibrium than to stretch or compress a bond, and the force constants are proportionately smaller, as can be observed in Table 4.2.

Angle	θ_0	k (kcal mol ⁻¹ deg ⁻¹)
Csp ³ —Csp ³ —Csp ³	109.47	0.0099
Csp ³ —Csp ³ —H	109.47	0.0079
H—Csp ³ —H	109.47	0.0070
Csp ³ —Csp ² —Csp ³	117.2	0.0099
Csp ³ —Csp ² —Csp ²	121.4	0.0121
Csp ³ —Csp ² —O	122.5	0.0101

Table 4.2 Force constants and reference angles for selected angles
[Allinger 1977].

As with the bond-stretching terms, the accuracy of the force field can be improved by the incorporation of higher-order terms. MM2 contains a quartic term in addition to the quadratic term. Higher-order terms have also been included to treat certain pathological cases such as very highly strained molecules. The general form of the angle-bending term then becomes:

$$\nu(\theta) = \frac{k}{2}(\theta - \theta_0)^2[1 - k'(\theta - \theta_0) - k''(\theta - \theta_0)^2 - k'''(\theta - \theta_0)^3 \dots] \quad (4.6)$$

4.5 Torsional Terms

The bond-stretching and angle-bending terms are often regarded as ‘hard’ degrees of freedom, in that quite substantial energies are required to cause significant deformations from

their reference values. Most of the variation in structure and relative energies is due to the complex interplay between the torsional and non-bonded contributions.

The existence of barriers to rotation about chemical bonds is fundamental to understanding the structural properties of molecules and conformational analysis. The three minimum-energy staggered conformations and three maximum-energy eclipsed structures of ethane are a classic example of the way in which the energy changes with a bond rotation. Quantum mechanical calculations suggest that this barrier to rotation can be considered to arise from antibonding interactions between the hydrogen atoms on opposite ends of the molecule; the antibonding interactions are minimised when the conformation is staggered and are at a maximum when the conformation is eclipsed. Many force fields are used for modelling flexible molecules where the major changes in conformation are due to rotations about bonds; in order to simulate this it is essential that the force field properly represents the energy profiles of such changes.

Not all molecular mechanics force fields use torsional potentials; it may be possible to rely upon non-bonded interactions between the atoms at the end of each torsion angle (the 1,4 atoms) to achieve the desired energy profile. However, most force fields for 'organic' molecules do use explicit torsional potentials with a contribution from each bonded quartet of atoms A–B–C–D in the system. Thus there would be nine individual torsional terms for ethane and 24 for benzene ($6 \times C-C-C-C$, $12 \times C-C-C-H$ and $6 \times H-C-C-H$). Torsional potentials are almost always expressed as a cosine series expansion. One functional form is:

$$\nu(\omega) = \sum_{n=0}^N \frac{V_n}{2} [1 + \cos(n\omega - \gamma)] \quad (4.7)$$

ω is the torsion angle.

An alternative but equivalent expression is:

$$\nu(\omega) = \sum_{n=0}^N C_n \cos(\omega)^n \quad (4.8)$$

V_n in Equation (4.7) is often referred to as the 'barrier' height, but to do so is misleading, obviously so when more than one term is present in the expansion. Moreover, other terms in the force field equation contribute to the barrier height as a bond is rotated, especially the non-bonded interactions between the 1,4 atoms. The value of V_n does, however, give a qualitative indication of the relative barriers to rotation; for example, V_n for an amide bond will be larger than for a bond between two sp^3 carbon atoms. n in Equation (4.7) is the *multiplicity*; its value gives the number of minimum points in the function as the bond is rotated through 360° . γ (the phase factor) determines where the torsion angle passes through its minimum value. For example, the energy profile for rotation about the single bond between two sp^3 carbon atoms could be represented by a single torsional term with $n = 3$ and $\gamma = 0^\circ$. This would give a threefold rotational profile with minima at torsion angles of $+60^\circ$, -60° and 180° and maxima at $\pm 120^\circ$ and 0° . A double bond between two sp^2 carbon atoms would have $n = 2$ and $\gamma = 180^\circ$, giving minima at 0° and 180° . The value of V_n would also be significantly larger for the double bond than for the single

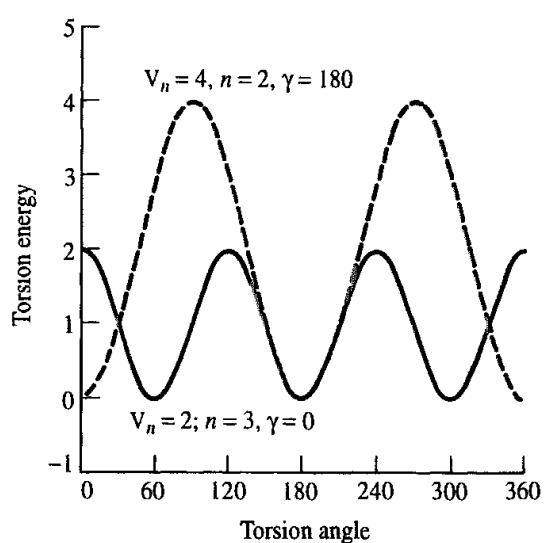


Fig 4.7. Torsional potential varies as shown for different values of V_n , n and γ .

bond. The effects of varying V_n , n and γ are illustrated in Figure 4.7 for commonly occurring torsional potentials.

Many of the torsional terms in the AMBER force field contain just one term from the cosine series expansion, but for some bonds it was found necessary to include more than one term. For example, to correctly model the tendency of O—C—C—O bonds to adopt a *gauche* conformation, a torsional potential with two terms was used for the O—C—C—O contribution:

$$\nu(\omega_{\text{C}-\text{O}-\text{O}-\text{C}}) = 0.25(1 + \cos 3\omega) + 0.25(1 + \cos 2\omega) \quad (4.9)$$

The torsional energy for a $\text{OCH}_2-\text{CH}_2\text{O}$ fragment (found in the sugars in DNA) varies with the torsion angle ω as shown in Figure 4.8. Another feature of the AMBER force field is its use of general torsional parameters. The energy profile for rotation about a bond that is described by a general torsional potential depends solely upon the atom types of the two

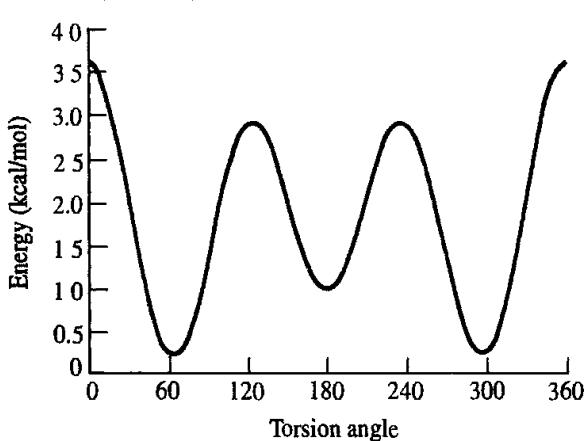


Fig 4.8: Variation in torsional energy (AMBER force field) with O—C—C—O torsion angle (ω) for $\text{OCH}_2-\text{CH}_2\text{O}$ fragment. The minimum energy conformations arise for $\omega = 60^\circ$ and 300°

atoms that comprise the central bond and not upon the atom types of the terminal atoms. For example, all torsion angles in which the central bond is between two sp^3 -hybridised carbon atoms (e.g. H–C–C–H, C–C–C–C, H–C–C–C) are assigned the same torsional parameters, unless the torsion is a special case such as O–C–C–O. In its treatment of the torsional contribution, AMBER takes a position intermediate between those force fields which only ever use a single term in the torsional expansion and those which consistently use more terms for all torsions. MM2 falls into the latter category; it uses three terms in the expansion:

$$\nu(\omega) = \frac{V_1}{2} (1 + \cos \omega) + \frac{V_2}{2} (1 - \cos 2\omega) + \frac{V_3}{2} (1 + \cos 3\omega) \quad (4.10)$$

A physical interpretation has been ascribed to each of the three terms in the MM2 torsional expansion from an analysis of *ab initio* calculations on simple fluorinated hydrocarbons. The first, onefold term corresponds to interactions between bond dipoles, which are due to differences in electronegativity between bonded atoms. The twofold term is due to the effects of hyperconjugation (in alkanes) and conjugation effects (in alkenes), which provide ‘double bond’ character to the bond. The threefold term corresponds to steric interactions between the 1,4 atoms. It was found that the additional terms in the torsional potential were especially important for systems containing heteroatoms, such as the halogenated hydrocarbons and molecules containing CCOC and CCNC fragments.

With careful parametrisation a force field which uses more than one term in the torsional expansion will be more successful than a force field that uses only a single term (and this is borne out by the MM2 force field). The major drawback is that many parameters are required to model even a modest range of molecules.

4.6 Improper Torsions and Out-of-plane Bending Motions

Let us consider how cyclobutanone would be modelled using a force field containing just standard bond-stretching and angle-bending terms of the type in Equation (4.1). The equilibrium structure obtained with such a force field would have the oxygen atom located out of the plane formed by the adjoining carbon atom and the two carbon atoms bonded to it, as shown in Figure 4.9. In this structure, the angles to the oxygen adopt values close to the reference value of 120° . Experimentally, it is found that the oxygen atom remains in the plane of the cyclobutane ring, even though the C–C=O angles are large (133°). This is because the π -bonding energy, which is maximised in the coplanar arrangement, would be much reduced if the oxygen were bent out of the plane. To achieve the desired geometry it is necessary to incorporate an additional term (or terms) in the force field that keeps the sp^2 carbon and the three atoms bonded to it in the same plane. The simplest way to achieve this is to use an *out-of-plane* bending term.

There are several ways in which out-of-plane bending terms can be incorporated into a force field. One approach is to treat the four atoms as an ‘improper’ torsion angle (i.e. a torsion angle in which the four atoms are not bonded in the sequence 1–2–3–4). One way to define an improper torsion for cyclobutane would involve the atoms 1–5–3–2 in Figure 4.9.

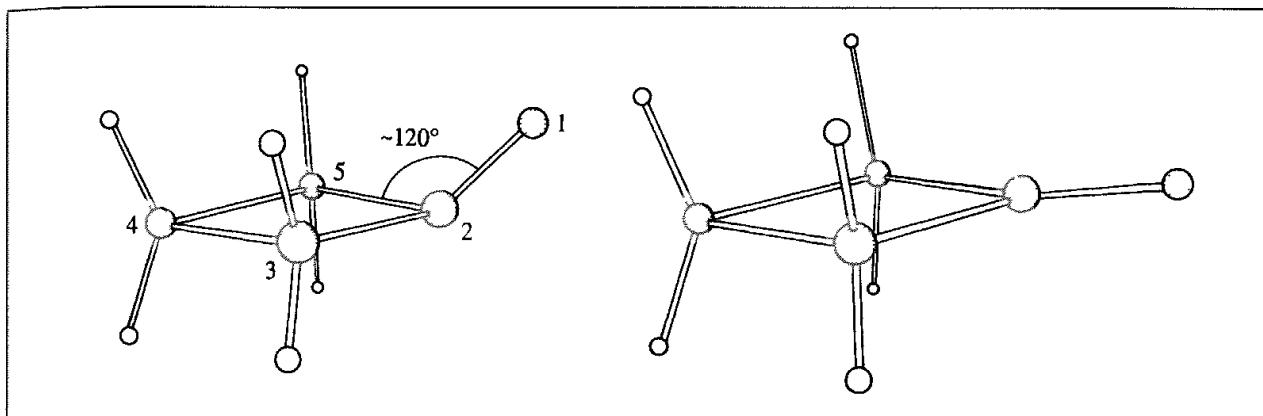


Fig. 4.9. Without an out-of-plane term, the oxygen atom in cyclobutane is predicted to lie out of the plane of the ring (left) rather than in the plane.

A torsional potential of the following form is then used to maintain the improper torsion angle at 0° or 180° :

$$\nu(\omega) = k(1 - \cos 2\omega) \quad (4.11)$$

Various other ways to incorporate the out-of-plane bending contribution are possible. For example, one definition that is closer to the notion of an 'out-of-plane bend' involves a calculation of the angle between a bond from the central atom and the plane defined by the central atom and the other two atoms (Figure 4.10). A value of 0° corresponds to all four atoms being coplanar. A third approach is to calculate the height of the central atom above a plane defined by the other three atoms (Figure 4.10). With these two definitions the deviation of the out-of-plane coordinate (be it an angle or a distance) can be modelled using a harmonic potential of the form

$$\nu(\theta) = \frac{k}{2}\theta^2; \quad \nu(h) = \frac{k}{2}h^2 \quad (4.12)$$

Of these three functional forms, the improper torsion definition is most widely used as it can then be easily included with the 'proper' torsional terms in the force field. However, the

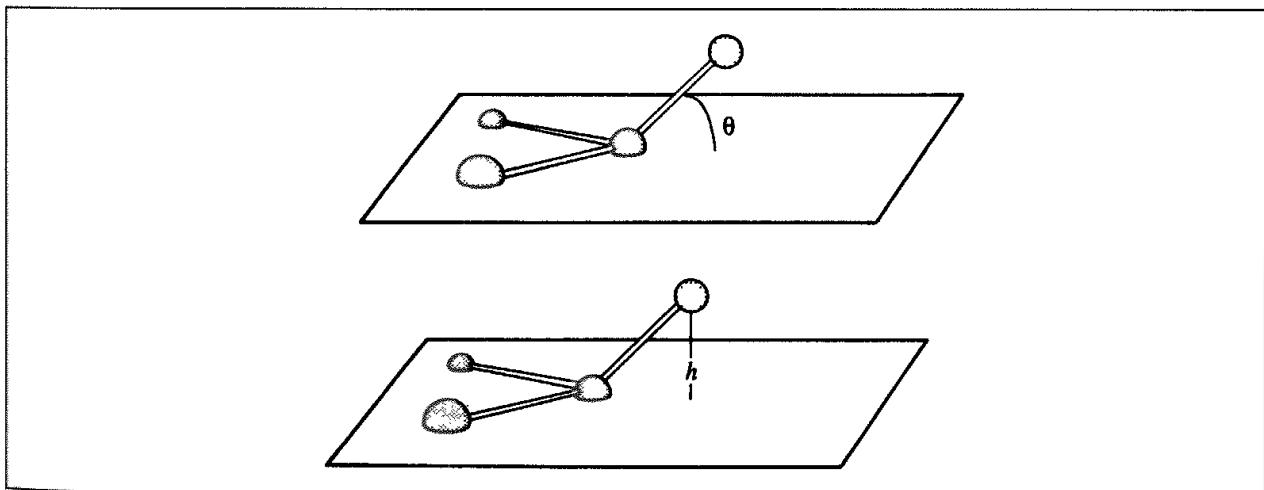


Fig. 4.10: Two ways to model the out-of-plane bending contributions.

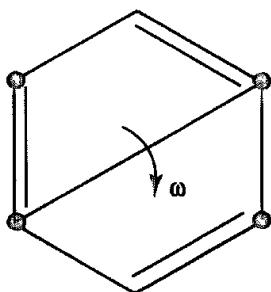


Fig. 4.11. Improper torsional terms can be used to keep a benzene ring planar

other two functional forms may be better ways to implement out-of-plane bending in the force field. Out-of-plane terms may also be used to achieve a particular geometry. For example, if it is desired to ensure that an aromatic ring such as benzene maintains an approximately planar structure then this can be achieved using a suitable set of out-of-plane bending terms involving atoms on opposite sides of the ring (Figure 4.11). Improper torsional terms are commonly used in the so-called united atom force fields to maintain stereochemistry at chiral centres (see Section 4.14). It is important to remember that out-of-plane terms may not always be necessary, and that to include such terms may have a deleterious effect on the performance of the force field. Vibrational frequencies in particular are often rather sensitive to the presence of out-of-plane terms.

4.7 Cross Terms: Class 1, 2 and 3 Force Fields

The presence of *cross terms* in a force field reflects coupling between the internal coordinates. For example, as a bond angle is decreased it is found that the adjacent bonds stretch to reduce the interaction between the 1,3 atoms, as illustrated in Figure 4.12. Cross terms were found to be important in force fields designed to predict vibrational spectra that were the forerunners of molecular mechanics force fields, and so it is not surprising that cross terms must often be included in a molecular mechanics force field to achieve optimal performance. One should in principle include cross terms between all contributions to a force field. However, only a few cross terms are generally found to be necessary in order to reproduce structural properties accurately; more may be needed to reproduce other properties such as vibrational frequencies, which are more sensitive to the presence of such terms. In general, any interactions involving motions that are far apart in a molecule

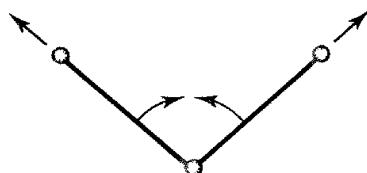


Fig. 4.12. Coupling between the stretching of the bonds as an angle closes.

can usually be set to zero. Most cross terms are functions of two internal coordinates, such as stretch–stretch, stretch–bend and stretch–torsion terms, but cross terms involving more than two internal coordinates such as the bend–bend–torsion have also been used. Various functional forms are possible for the cross terms. For example, the stretch–stretch cross term between two bonds 1 and 2 can be modelled as:

$$\nu(l_1, l_2) = \frac{k_{l_1, l_2}}{2} [(l_1 - l_{1,0})(l_2 - l_{2,0})] \quad (4.13)$$

The stretching of the two bonds adjoining an angle could be modelled using an equation of the following form (as in MM2, MM3 and MM4):

$$\nu(l_1, l_2, \theta) = \frac{k_{l_1, l_2, \theta}}{2} [(l_1 - l_{1,0}) + (l_2 - l_{2,0})](\theta - \theta_0) \quad (4.14)$$

In a *Urey–Bradley* force field, angle bending is achieved using 1,3 non-bonded interactions rather than an explicit angle-bending potential. The stretch-bond term in such a force field would be modelled by a harmonic function of the distance between the 1,3 atoms:

$$\nu(r_{1,3}) = \frac{k_{r_{1,3}}}{2} (r_{1,3} - r_{1,3}^0)^2 \quad (4.15)$$

A stretch-torsion cross term can be used to model the stretching of a bond that occurs in an eclipsed conformation. Two possible functional forms are:

$$\nu(l, \omega) = k(l - l_0) \cos n\omega \quad (4.16)$$

$$\nu(l, \omega) = k(l - l_0)[1 + \cos n\omega] \quad (4.17)$$

n is the periodicity of the rotation about the bond ($n = 3$ for sp^3 - sp^3 bonds).

Torsion-bend and torsion-bend-bend terms may also be included; the latter, for example, would couple two angles A–B–C and B–C–D to a torsion angle A–B–C–D. Maple, Dinur and Hagler used quantum mechanics calculations to investigate which of the cross terms are most important and suggested that the stretch–stretch, stretch–bend, bend–bend, stretch–torsion and bend–bend–torsion were most important [Dinur and Hagler 1991] (schematically illustrated in Figure 4.13).

It has been suggested that the presence of cross terms (together with some other features) can provide a general way to classify force fields [Hwang *et al.* 1994]. A class I force field was considered one which is restricted to harmonic terms (e.g. for bond stretching and angle bending) and which does not have any cross terms. A class II force field would have anharmonic terms (e.g. through the use of Morse potentials or quartic terms) and explicit cross terms to account for the coupling between coordinates. The presence of these higher and cross terms would tend to improve the ability of the force field to predict the properties of more unusual systems (such as those which are highly strained) and also to enhance its ability to reproduce vibrational spectra. Another characteristic of a class II force field was that it could be used without modification to model the properties of isolated small molecules, condensed phases and macromolecular systems. It was subsequently suggested by Allinger [Allinger *et al.* 1996b] that a class III force field would also take account of chemical effects and other features such as electronegativity and hyperconjugation. A classic

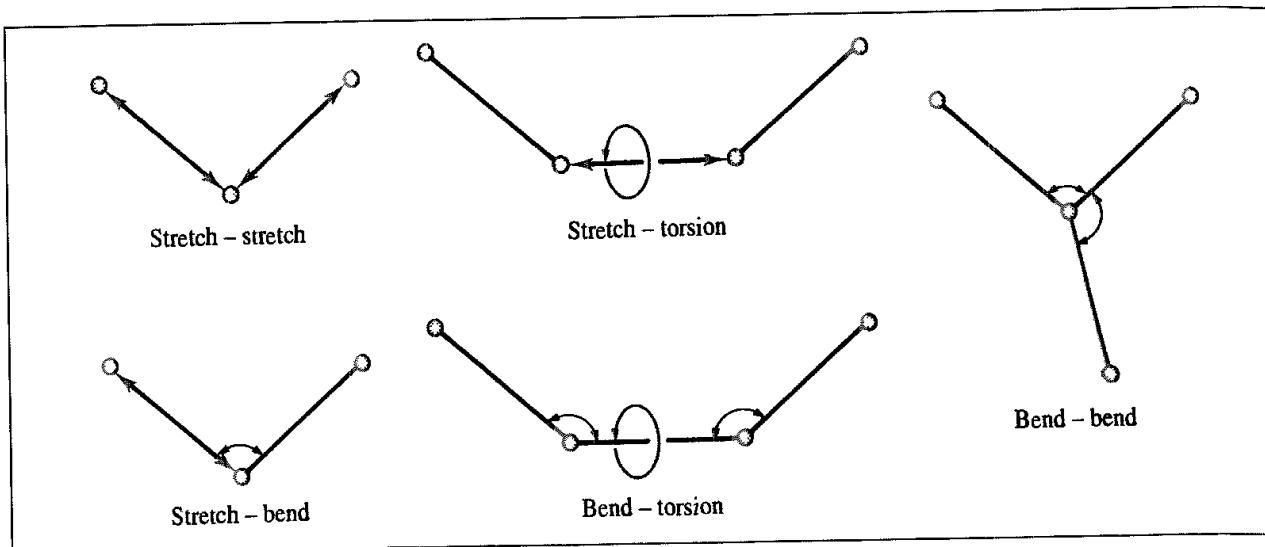


Fig. 4.13. Schematic illustration of the cross terms believed to be most important in force fields. (Adapted from Dinur U and A T Hagler 1991. New Approaches to Empirical Force Fields In Reviews in Computational Chemistry, Lipkowitz K B and D B Boyd (Editors) New York, VCH Publishers, pp 99–164)

example of the latter effect (hyperconjugation) is the change in the length of the C–H bond in acetaldehyde with rotation about the C–C bond. When the C–H bond is perpendicular to the plane of the carbonyl group there is maximum overlap between the σ orbital of the C–H bond and the π^* orbital of the carbonyl carbon. Donation of electron density from the C–H bond to this π^* orbital is accompanied by a lengthening of the bond and a greater contribution from the charged resonance structure (Figure 4.14). When the bond to the hydrogen atom is in the plane the overlap is minimal. *Ab initio* calculations suggested that the bond length changed by 0.006 Å between the two forms. This effect was incorporated within MM4 by a term of the following form:

$$\Delta l = k(1 - \cos 2\omega) \quad (4.18)$$

This is a kind of torsion–stretch cross term but different from the one where the central bond changes with torsion angle. There has been some considerable debate about the existence and origin of the hyperconjugative effects, but low-temperature X-ray crystallographic experiments on appropriate compounds together with *ab initio* calculations certainly reveal a detectable effect.

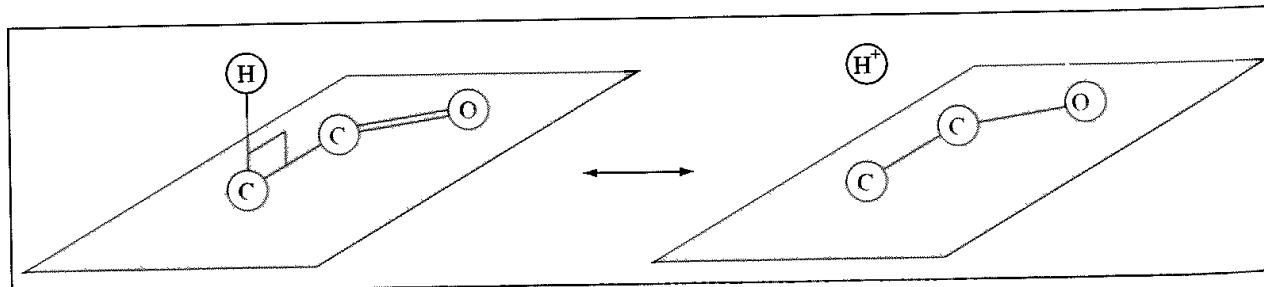


Fig. 4.14. Valence bond representation of the hyperconjugation effect which leads to a lengthening of the C–H bond in acetaldehyde

4.8 Introduction to Non-bonded Interactions

Independent molecules and atoms interact through non-bonded forces, which also play an important role in determining the structure of individual molecular species. The non-bonded interactions do not depend upon a specific bonding relationship between atoms. They are ‘through-space’ interactions and are usually modelled as a function of some inverse power of the distance. The non-bonded terms in a force field are usually considered in two groups, one comprising electrostatic interactions and the other van der Waals interactions.

4.9 Electrostatic Interactions

4.9.1 The Central Multipole Expansion

Electronegative elements attract electrons more than less electronegative elements, giving rise to an unequal distribution of charge in a molecule. This charge distribution can be represented in a number of ways, one common approach being an arrangement of fractional point charges throughout the molecule. These charges are designed to reproduce the electrostatic properties of the molecule. If the charges are restricted to the nuclear centres they are often referred to as *partial atomic charges* or *net atomic charges*. The electrostatic interaction between two molecules (or between different parts of the same molecule) is then calculated as a sum of interactions between pairs of point charges, using Coulomb’s law:

$$\mathcal{V} = \sum_{i=1}^{N_A} \sum_{j=1}^{N_B} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \quad (4.19)$$

N_A and N_B are the numbers of point charges in the two molecules. This approach to the representation and calculation of electrostatic interactions will be considered in more detail in Section 4.9.2. First, we shall consider an alternative approach to the calculation of electrostatic interactions which treats a molecule as a single entity and is (in principle at least) capable of providing a very efficient way to calculate electrostatic intermolecular interactions. This is the *central multipole expansion*, which is based upon the electric moments or multipoles: the charge, dipole, quadrupole, octopole, and so on introduced in Section 2.7.3. These moments are usually represented by the following symbols: q (charge), μ (dipole), Θ (quadrupole) and Φ (octopole). We are often interested in the lowest non-zero electric moment. Thus species such as Na^+ , Cl^- , NH_4^+ or CH_3CO_2^- have the charge as their lowest non-zero moment. For many uncharged molecules the dipole is the lowest non-zero moment. Molecules such as N_2 and CO_2 have the quadrupole as their lowest non-zero moment. The lowest non-zero moment for methane and tetrafluoromethane is the octopole. Each of these multipole moments can be represented by an appropriate distribution of charges. Thus a dipole can be represented using two charges placed an appropriate distance apart. A quadrupole can be represented using four charges and an octopole by eight charges. A complete description of the charge distribution around a molecule requires all of the non-zero electric moments to be specified. For some molecules,

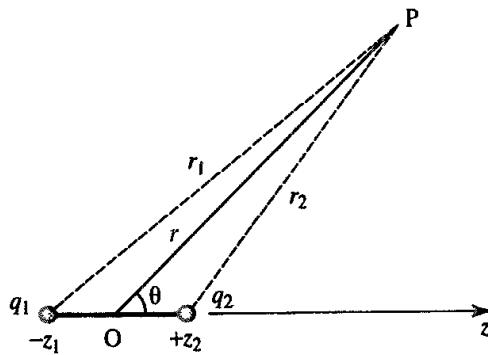


Fig. 4.15: The electrostatic potential due to two point charges.

the lowest non-zero moment may not be the most significant and it may therefore be unwise to ignore the higher-order terms in the expansion without first checking their values.

To illustrate how the multipolar expansion is related to a distribution of charges in a system, let us consider the simple case of a molecule with two charges q_1 and q_2 , positioned at $-z_1$ and z_2 , respectively (Figure 4.15). The electrostatic potential at point P (a distance r from the origin, r_1 from charge q_1 and r_2 from charge q_2) is then given by:

$$\phi(r) = \frac{1}{4\pi\epsilon_0} \left(\frac{q_1}{r_1} + \frac{q_2}{r_2} \right) \quad (4.20)$$

By applying the cosine rule this can be written as follows (see Figure 4.15):

$$\phi(r) = \frac{1}{4\pi\epsilon_0} \left(\frac{q_1}{\sqrt{r^2 + z_1^2 + 2rz_1 \cos\theta}} + \frac{q_2}{\sqrt{r^2 + z_2^2 - 2rz_1 \cos\theta}} \right) \quad (4.21)$$

If $r \gg z_1$ and $r \gg z_2$ then this expression can be expanded as follows:

$$\phi(r) = \frac{1}{4\pi\epsilon_0} \left(\frac{q_1 + q_2}{r} + \frac{(q_2 z_2 - q_1 z_1) \cos\theta}{r^2} + \frac{(q_1 z_1^2 + q_2 z_2^2)(3 \cos^2\theta - 1)}{2r^3} + \dots \right) \quad (4.22)$$

We can now associate the appropriate terms in the expansion with the various electric moments:

$$\phi(r) = \frac{1}{4\pi\epsilon_0} \left(\frac{q}{r} + \frac{\mu \cos\theta}{r^2} + \frac{\Theta(3 \cos^2\theta - 1)}{2r^3} + \dots \right) \quad (4.23)$$

Thus $(q_1 + q_2)$ is the charge; $(q_2 z_2 - q_1 z_1)$ is the dipole; $(q_1 z_1^2 + q_2 z_2^2)$ is the quadrupole, and so on. One interesting feature about a charge distribution is that only the first non-zero moment is independent of the choice of origin. Thus, if a molecule is electrically neutral (i.e. $q_1 + q_2 = 0$) then its dipole moment is independent of the choice of origin. This can be demonstrated for our two-charge system as follows. If the position of the origin is now moved to a point $-z'$, then the dipole moment relative to this new origin is given by:

$$\mu' = q_2(z_2 + z') - q_1(z_1 - z') = \mu + qz' \quad (4.24)$$

Only if the total charge on the system (q) equals zero will the dipole moment be unchanged. Similar arguments can be used to show that if both the charge and the dipole moment are zero then the quadrupole moment is independent of the choice of origin. For convenience, the origin is often taken to be the centre of mass of the charge distribution.

The electric moments are examples of *tensor properties*: the charge is a rank 0 tensor (which is the same as a scalar quantity); the dipole is a rank 1 tensor (which is the same as a vector, with three components along the x , y and z axes); the quadrupole is a rank 2 tensor with nine components, which can be represented as a 3×3 matrix. In general, a tensor of rank n has 3^n components.

For a distribution of charges (one not restricted to lie along one of the Cartesian axes), the dipole moment is given by:

$$\mu = \sum q_i r_i \quad (4.25)$$

The components of the dipole moment along the x , y and z axes are $\sum q_i x_i$, $\sum q_i y_i$ and $\sum q_i z_i$. The analogous way to define the quadrupole moment is as follows:

$$\Theta = \begin{pmatrix} \sum q_i x_i^2 & \sum q_i x_i y_i & \sum q_i x_i z_i \\ \sum q_i y_i x_i & \sum q_i y_i^2 & \sum q_i y_i z_i \\ \sum q_i z_i x_i & \sum q_i z_i y_i & \sum q_i z_i^2 \end{pmatrix} \quad (4.26)$$

This definition of the quadrupole is obviously dependent upon the orientation of the charge distribution within the coordinate frame. Transformation of the axes can lead to alternative definitions that may be more informative. Thus the quadrupole moment is commonly defined as follows:

$$\Theta = \frac{1}{2} \begin{pmatrix} \sum_i q_i (3x_i^2 - r_i^2) & 3 \sum_i q_i x_i y_i & 3 \sum_i q_i x_i z_i \\ 3 \sum_i q_i x_i z_i & \sum_i q_i (3y_i^2 - r_i^2) & 3 \sum_i q_i y_i z_i \\ 3 \sum_i q_i x_i z_i & 3 \sum_i q_i y_i z_i & \sum_i q_i (3z_i^2 - r_i^2) \end{pmatrix} \quad (4.27)$$

In Equation (4.27) $r_i^2 = x_i^2 + y_i^2 + z_i^2$. This definition enables one to assess the deviation from spherical symmetry as a spherically symmetric charge distribution will have

$$\sum_i q_i x_i^2 = \sum_i q_i y_i^2 = \sum_i q_i z_i^2 = \frac{1}{3} \sum_i q_i r_i^2 \quad (4.28)$$

and so the diagonal elements of the tensor will be zero. Quadrupoles are also reported in terms of the *principal axes*; these are three mutually perpendicular axes α , β and γ , which are linear combinations of x , y and z such that the quadrupole tensor is diagonal (i.e. off-diagonal elements are zero):

$$\Theta = \begin{pmatrix} \Theta_{\alpha\alpha} & 0 & 0 \\ 0 & \Theta_{\beta\beta} & 0 \\ 0 & 0 & \Theta_{\gamma\gamma} \end{pmatrix} \quad (4.29)$$

Let us now consider the effect of placing another molecule with a linear charge distribution (charges q'_1 and q'_2) with its centre of mass at the point P. The relative orientation of the two

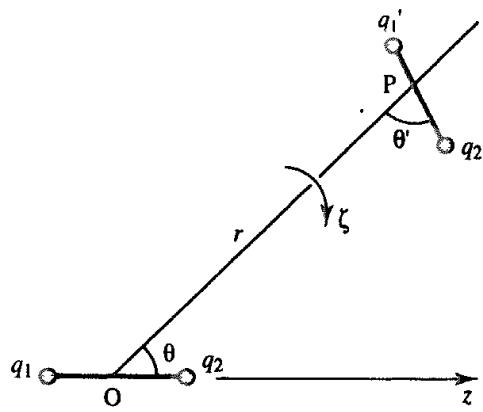


Fig. 4.16: The relative orientation of two dipoles

molecules can be described in terms of four parameters (the distance joining their centres of mass and three angles as shown in Figure 4.16). The electrostatic interaction between the two molecules is calculated by multiplying each charge by the potential at that point and adding the result for each charge. The following expression is the result [Buckingham 1959]:

$$\mathcal{V}(q, q') = \frac{1}{4\pi\epsilon_0} \left\{ \begin{aligned} & \frac{qq'}{r} \\ & + \frac{1}{r^2} (q\mu' \cos \theta + q'\mu \cos \theta') \\ & + \frac{\mu\mu'}{r^3} (2 \cos \theta \cos \theta' + \sin \theta \sin \theta' \cos \zeta) \\ & + \frac{1}{2r^3} [q\Theta'(3 \cos^2 \theta' - 1) + q'\Theta(3 \cos^2 \theta - 1)] \\ & + \frac{3}{2r^4} [\mu\Theta' \{ \cos \theta (3 \cos^2 \theta' - 1) + 2 \sin \theta \sin \theta' \cos \theta' \cos \zeta \} \\ & \quad + \mu' \Theta \{ \cos \theta' (3 \cos^2 \theta - 1) + 2 \sin \theta' \sin \theta \cos \theta \cos \zeta \}] \\ & + \frac{3\Theta\Theta'}{4r^5} [1 - 5 \cos^2 \theta - 5 \cos^2 \theta' + 17 \cos^2 \theta \cos^2 \theta' \\ & \quad + 2 \sin^2 \theta \sin^2 \theta' \cos^2 \zeta + 16 \sin \theta \sin \theta' \cos \theta \cos \theta' \cos \zeta] \\ & + \dots \end{aligned} \right\} \quad (4.30)$$

The energy of interaction between two charge distributions is thus an infinite series that includes charge-charge, charge-dipole, dipole-dipole, charge-quadrupole, dipole-quadrupole interactions, quadrupole-quadrupole terms, and so on. These terms depend on different inverse powers of the separation r . If the molecules are neutral (i.e. $q = q' = 0$) then the leading term in the expansion is that due to the dipole-dipole interaction, which varies as r^{-3} . This is a key result, for the range of the dipole-dipole interaction (r^{-3}) is much less than that of the Coulomb interaction (r^{-1}), Figure 4.17. This will be important in later chapters, where we shall collect atoms together into neutral groups. The electrostatic interaction

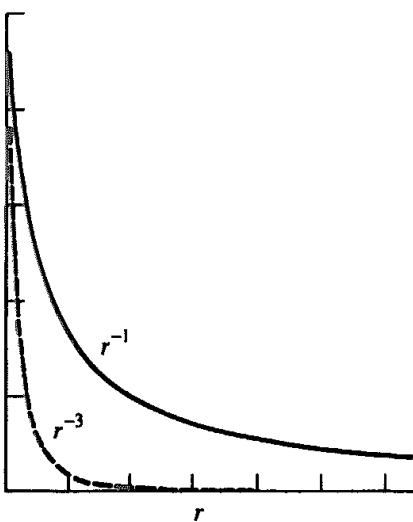


Fig. 4.17 The charge-charge energy decays much more slowly (αr^{-1}) than the dipole-dipole energy (αr^{-3})

between these groups then decays as r^{-3} rather than the r^{-1} dependence of each individual charge-charge interaction. This can be seen in Figure 4.17, in which the functions r^{-1} and r^{-3} have been plotted as a function of distance. Even when the dipole-dipole interaction energy has fallen off almost to zero the charge-charge interaction energy is still significant. In general, the interaction energy between two multipoles of order n and m decreases as $r^{-(n+m+1)}$. It should be emphasised again that these expressions are only valid when the separation of the two molecules, r , is much larger than the internal dimensions of the molecules. The favourable arrangements for the various multipoles are shown in Figure 4.18.

A central multipole expansion therefore provides a way to calculate the electrostatic interaction between two molecules. The multipole moments can be obtained from the wavefunction and can therefore be calculated using quantum mechanics (see Section 2.7.3) or can be determined from experiment. One example of the use of a multipole expansion is

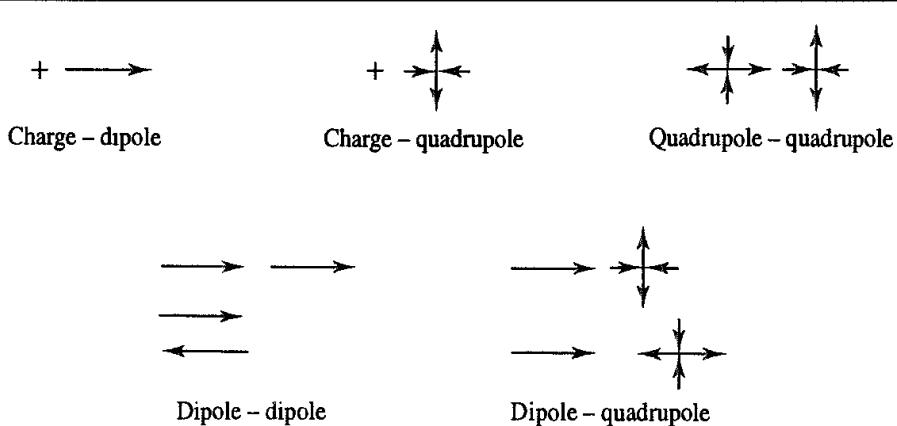


Fig. 4.18 The most favourable orientations of various multipoles (Figure adapted from Buckingham A D 1959 Molecular Quadrupole Moments. Quarterly Reviews of the Chemical Society 13:183–214.)

the benzene model of Claessens, Ferrario and Ryckaert [Claessens *et al.* 1983]. Benzene has no charge and no dipole moment, but it does have a sizeable quadrupole. The inclusion of the quadrupole was found to give clearly superior results in molecular dynamics simulations of the liquid state over models that lacked any electronic contribution.

The main advantage of the multipolar description for calculating the electrostatic interactions between molecules is its efficiency. For example, the charge-charge interaction energy between two benzene molecules would require 144 individual charge-charge interactions with a partial atomic charge model rather than the single quadrupole-quadrupole term. Unfortunately, the multipole expansion is not applicable when the molecules are separated by distances comparable with the molecular dimensions. The formal condition for convergence of the multipolar interaction energy is that the distance between two interacting molecules should be larger than the sum of the distances from the centre of each molecule to the furthest part of its charge distribution. If a sphere is constructed around each molecule, positioned on its centre of mass, with a radius that encompasses all of the charge distribution, then the multipole expansion for the interaction between two molecules will converge if these spheres do not intersect. Even if one requires the sphere to encompass just the nuclei in a molecule (i.e. ignoring the fact that the charge distribution around a molecule extends to infinity) there may still be problems. For example, the convergence sphere for a molecule such as butane would extend beyond the van der Waals radii in some directions, enabling other molecules to penetrate the convergence sphere, as illustrated in Figure 4.19. Another problem is that the multipolar expansion may be slow to converge. The multipolar expansion is often located at the centre of mass, but this may not be the best choice to achieve the most rapid convergence.

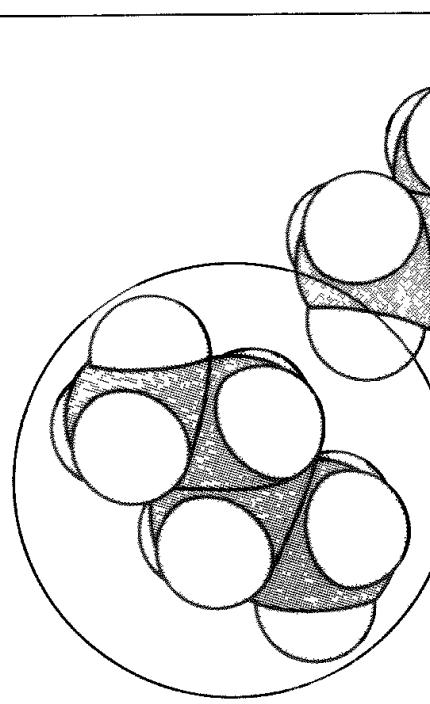


Fig. 4.19. The convergence sphere of the multipole expansion for a molecule such as butane may be penetrated by another molecule

There are other difficulties with the central multipole expansion. The multipole moments are properties of the entire molecule and so cannot be used to determine intramolecular interactions. The central multipole model thus tends to be restricted to calculations involving small molecules that are kept fixed in conformation during the calculation, and where the interactions between molecules act at their centres of mass. It can be a complicated procedure to calculate the forces acting on a molecule with a multipole model. The interaction between multipoles of zero order (i.e. charges) gives rise to a simple translational force. Multipoles of a higher order have directionality, and interactions between these produce a torque, or twisting force. Moreover, whereas the charge-charge forces are equal and opposite, the torque acting on molecule i due to another molecule j is not necessarily equal and opposite to the torque on molecule j due to molecule i .

4.9.2 Point-charge Electrostatic Models

We therefore return to the point-charge model for calculating electrostatic interactions. If sufficient point charges are used then all of the electric moments can be reproduced and the multipole interaction energy, Equation (4.30), is exactly equal to that calculated from the Coulomb summation, Equation (4.19).

An accurate representation of a molecule's electrostatic properties may require charges to be placed at locations other than at the atomic nuclei. A simple example of this is molecular nitrogen, which has a dipole moment of zero. The total charge on nitrogen is zero, and so an atomic partial charge model would put zero charge on each nucleus. However, nitrogen does have a quadrupole moment and this significantly affects its properties. The simplest way to model this is to place three partial charges along the bond: a charge of $-q$ at each nucleus and $+2q$ at the centre of mass. The quadrupole-quadrupole interaction between two nitrogen molecules can then be calculated by summing nine pairs of charge-charge interactions. The value of q can be calculated using the following relationship between the quadrupole moment and the partial charge:

$$\Theta = 2q(l/2)^2 \quad (4.31)$$

l is the bond length. The experimental quadrupole moment is consistent with a charge, q , of approximately $0.5e$. In fact, a better representation of the electrostatic potential around the nitrogen molecule is obtained using the five-charge model shown in Figure 4.20.

An alternative to the point charge model is to assign dipoles to the bonds in the molecule. The electrostatic energy is then given as a sum of dipole-dipole interaction energies. This approach (which is adopted in MM2/MM3/MM4) can be unwieldy for molecules that have a formal charge and which require charge-charge and charge-dipole terms to be included in the energy expression. Charged species are dealt with more naturally using the point charge model.

4.9.3 Calculating Partial Atomic Charges

Given the widespread use of the partial atomic charge model, it is important to consider how the charges are obtained. For simple species the atomic charges required to reproduce the

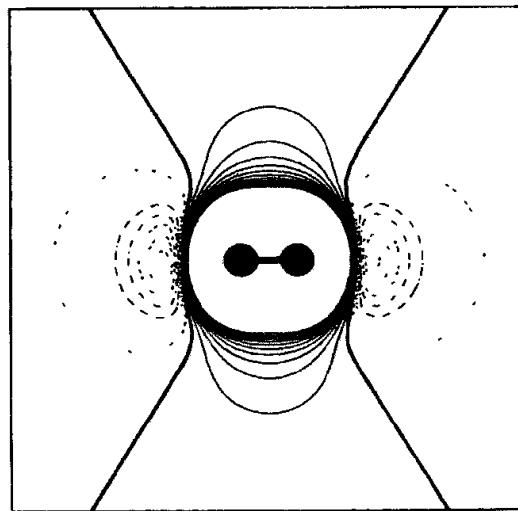
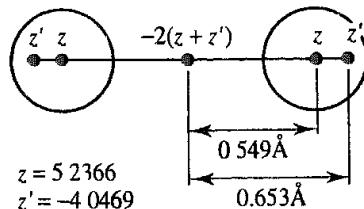
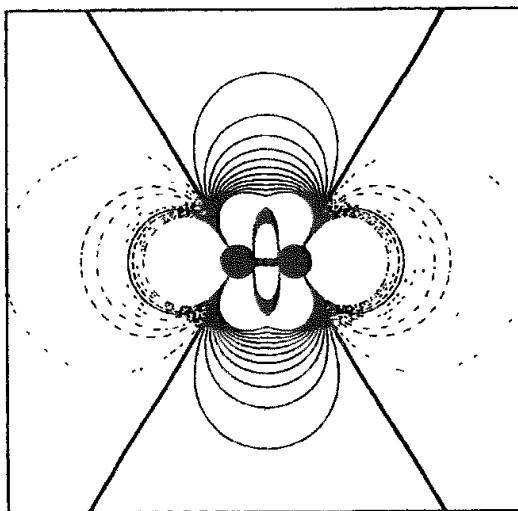
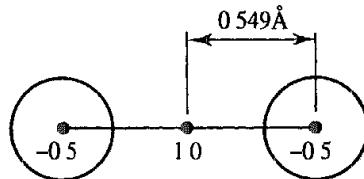
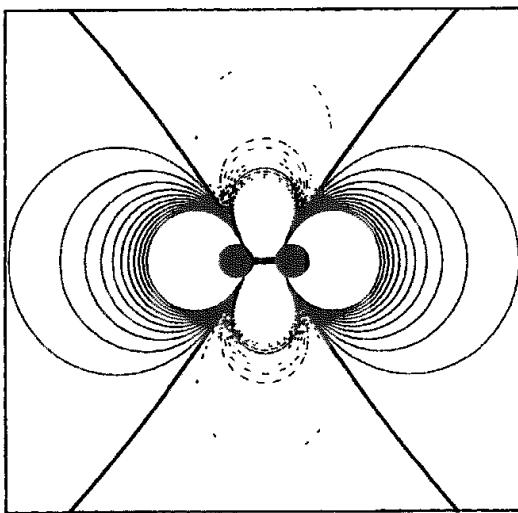


Fig. 4.20: Two charge models for N_2 with the electrostatic potentials that they generate. Also shown is the electrostatic potential calculated using ab initio quantum mechanics (6-31G^{*} basis set.) Negative contours are dashed and the zero contour is bold

electric moments can be calculated exactly if the geometry is known. For example, the experimentally determined dipole moment of HF (1.82 D) can be reproduced by placing equal but opposite charges of $0.413e$ on the two atomic nuclei (assuming a bond length of 0.917 \AA). The tetrahedral arrangement of the hydrogens about the carbon in methane means that each hydrogen atom has an identical charge equal to one quarter the charge on the carbon. The molecule is electrically neutral with zero dipole and quadrupole moments but a non-zero octopole moment, which can be reproduced using a hydrogen charge of approximately $0.14e$.

In some cases the atomic charges are chosen to reproduce thermodynamic properties calculated using a molecular dynamics or Monte Carlo simulation. A series of simulations is performed and the charge model is modified until satisfactory agreement with experiment is obtained. This approach can be quite powerful despite its apparent simplicity, but it is only really practical for small molecules or simple models.

The electrostatic properties of a molecule are a consequence of the distribution of the electrons and the nuclei and thus it is reasonable to assume that one should be able to obtain a set of partial atomic charges using quantum mechanics. Unfortunately, the partial atomic charge is not an experimentally observable quantity and cannot be unambiguously calculated from the wavefunction. This explains why numerous ways to determine partial atomic charges have been proposed, and why there is still considerable debate as to the 'best' method to derive them. Indirect comparisons of the various methods are possible, usually by calculating appropriate quantities from the charge model and then comparing the results with either experiment or quantum mechanics. For example, one might examine how well the charge model reproduces the experimental or quantum mechanical multipole moments or the electrostatic potential around the molecule.

We have already encountered in Section 2.7.5 the population analysis method for calculating partial atomic charges. Such sets of charges (commonly referred to as *Mulliken charges* when obtained from that particular partitioning scheme) are often considered to be inappropriate for accurately representing the interactions between molecules. This is because Mulliken charges are primarily dependent upon the constitution of the molecule – how the atoms are bonded together – rather than being designed to reproduce the properties that determine how molecules interact with each other, such as the electrostatic potential. The importance of the electrostatic potential in intermolecular interactions has resulted in much interest in schemes that calculate charges consistent with this particular property.

4.9.4 Charges Derived from the Molecular Electrostatic Potential

The electrostatic potential at a point is the force acting on a unit positive charge placed at that point. The nuclei give rise to a positive (i.e. repulsive) force, whereas the electrons give rise to a negative potential. The electrostatic potential is an observable quantity that can be determined from a wavefunction using Equations (2.222) and (2.223):

$$\phi(\mathbf{r}) = \phi_{\text{nucl}}(\mathbf{r}) + \phi_{\text{elec}}(\mathbf{r}) = \sum_{A=1}^M \frac{Z_A}{|\mathbf{r} - \mathbf{R}_A|} - \int \frac{d\mathbf{r}' \rho(\mathbf{r}')}{|\mathbf{r}' - \mathbf{r}|} \quad (4.32)$$

The electrostatic potential is a continuous property and is not easily represented by an analytical function. Consequently, it is necessary to derive a discrete representation for use in numerical analysis. The objective is to derive the set of partial charges (usually partial atomic charges) that best reproduces the quantum mechanical electrostatic potential at a series of points surrounding the molecule. A solution to this problem was suggested by Cox and Williams [Cox and Williams 1981]. The electrostatic potential at each of the chosen points is calculated from the wavefunction. A least-squares fitting procedure is then employed to determine the set of partial atomic charges that best reproduces the electrostatic potential at the points, subject to the constraint that the sum of the charges should be equal to the net charge on the molecule. Symmetry conditions may also be imposed to ensure that the charges on symmetrically equivalent atoms are equal. It is also possible to require the atomic charges to reproduce other electrostatic properties of the molecules such as the dipole moment. The fitting procedure minimises the sum of squares of the differences in the electrostatic potential. Thus, if the electrostatic potential at a point is ϕ_i^0 and if the value from the charge model is ϕ_i^{calc} , then the objective is to minimise the following function:

$$R = \sum_{i=1}^{N_{\text{points}}} w_i (\phi_i^0 - \phi_i^{\text{calc}})^2 \quad (4.33)$$

N_{points} is the number of points and w_i is a weighting factor that enables different points to be given different degrees of ‘importance’ in the fitting process. One of the charges is dependent on the values of the others (because the sum must equal Z , the molecular charge). This N th charge has a value given by:

$$q_N = Z - \sum_{j=1}^{N-1} q_j \quad (4.34)$$

The electrostatic potential due to the charges q_j at the point i is given by Coulomb’s law:

$$\phi_i^{\text{calc}} = \sum_{j=1}^{N-1} \frac{q_j}{4\pi\epsilon_0 r_{ij}} + \frac{Z - \sum_{j=1}^{N-1} q_j}{4\pi\epsilon_0 r_{iN}} \quad (4.35)$$

r_{ij} is the distance from the charge j to the point i . At a minimum value of the error function, R , the first derivative is equal to zero with respect to all charges q_k :

$$\frac{\partial R}{\partial q_k} = -2 \sum_{i=1}^{N_{\text{points}}} w_i (\phi_i^0 - \phi_i^{\text{calc}}) \left(\frac{\partial \phi_i^{\text{calc}}}{\partial q_k} \right) = 0 \quad (4.36)$$

This equation can be written in the following form:

$$\sum_{i=1}^{N_{\text{points}}} w_i \left(\phi_i^0 - \frac{Z}{r_{iN}} \right) \left(\frac{1}{r_{ik}} - \frac{1}{r_{iN}} \right) = \sum_{j=1}^{N-1} \left[\sum_{i=1}^{N_{\text{points}}} w_i \left(\frac{1}{r_{ik}} - \frac{1}{r_{iN}} \right) \left(\frac{1}{r_{ij}} - \frac{1}{r_{iN}} \right) \right] \frac{q_j}{4\pi\epsilon_0} \quad (4.37)$$

When expressed in this way, then the set of equations can be recast as a matrix equation of the form $\mathbf{A}\mathbf{q} = \mathbf{a}$. The charges \mathbf{q} are then determined using standard matrix methods via $\mathbf{q} = \mathbf{A}^{-1}\mathbf{a}$.

The points i ($1, 2, \dots, N_{\text{points}}$) where the potential is fitted can be chosen in a variety of ways but should be taken from the region where it is most important to model intermolecular interactions correctly. This region is just beyond the van der Waals radii of the atoms involved. Cox and Williams selected points from a regular grid in a shell defined by two surfaces, one corresponding to the union of the van der Waals radii plus 1.2 \AA and the others approximately 1 \AA beyond that. The CHELP procedure of Chirlian and Franci [Chirlian and Franci 1987] uses spherical shells, 1 \AA apart, centred on each atom with points symmetrically distributed on the surface. Any points within the van der Waals radius of any atom in the system are discarded and the shells extend to 3 \AA from the van der Waals surface of the molecule. The CHELP method employs a Lagrange multiplier method to find the atomic charges, rather than an iterative least-squares procedure. This minimises the error function R (Equation (4.33)) subject to the constraint that the charges sum to the total molecular charge. Such an analysis yields a set of $N + 1$ equations in $N + 1$ unknowns and can be solved using standard matrix methods. The CHELPG algorithm of Breneman and Wiberg [Breneman and Wiberg 1990] combines the regular grid of points of Cox and Williams with the Lagrange multiplier method of Chirlian and Franci as the results from CHELP were found to change if the molecule was reoriented in the coordinate system. In CHELPG a cubic grid of points (spaced $0.3\text{--}0.8 \text{ \AA}$ apart) is used and all grid points that lie within the van der Waals radius of any atom are discarded, together with all points that lie further than 2.8 \AA away from any atom.

The algorithm of Singh and Kollman used to derive the charges in the 1984 AMBER force field uses points on a series of molecular surfaces, constructed using gradually increasing van der Waals radii for the atoms [Singh and Kollman 1984]. The points at which the potential was fitted were located on these shells. For the 1995 AMBER force field a modified version of this electrostatic potential method was employed (termed ‘restrained electrostatic potential fit’, or RESP [Bayly *et al.* 1993]). The RESP algorithm uses hyperbolic restraints on non-hydrogen atoms. These restraints have the effect of reducing the charges on some atoms, particularly buried carbon atoms, which can be assigned artificially high charges in standard electrostatic potential fitting methods. The RESP charges also vary less with the molecular conformation.

4.9.5 Deriving Charge Models for Large Systems

Molecular mechanics is used to model systems containing thousands of atoms such as polymers. How then can charges be derived for such species? Clearly one cannot routinely perform quantum mechanical calculations on a molecule with so many atoms and so it must be broken into fragments of a suitable size. In some cases the fragments might appear relatively easy to define; for example, many polymeric systems are constructed by connecting together chemically defined monomeric units. The atomic charges for each monomer should be obtained from calculations on suitable fragments that recreate the immediate local environment of the fragment in the larger molecule. For example, partial atomic charges for amino acids are often obtained from calculations on a ‘dipeptide’ fragment (see Figure 4.21), which is more akin to the environment within a protein than in an isolated amino acid.

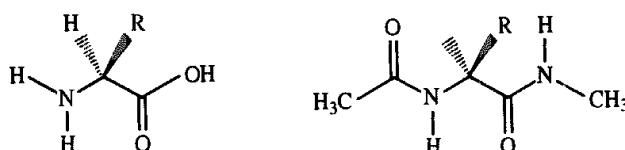


Fig 4.21 The charges used for calculations on proteins are best derived using a suitable fragment for each amino acid that reflects the environment within the protein (right), rather than the isolated amino acid (left)

The charge sets obtained from electrostatic potential fitting can be highly dependent upon the basis set used to derive the wavefunction. Moreover, the charges do not always improve if a larger basis set is used. It is generally considered that the 6-31G* basis set gives reasonable results for calculations relevant to condensed phases. In many cases it is possible to scale the results of a calculation using a small basis set or even a lower level of theory (such as a semi-empirical calculation) to obtain results comparable with those of a high-level calculation. Of the various semi-empirical methods available, MNDO appears to give the best correspondence with the charges derived from *ab initio* calculations, and scaling factors have been determined by several research groups [Ferenczy *et al.* 1990; Luque *et al.* 1990; Bezler *et al.* 1990]. An additional complicating factor is that the charges obtained from electrostatic potential fitting will often depend upon the conformation for which the quantum mechanical calculation was performed [Williams 1990]. One solution is to perform a series of charge calculations for different conformations and then use a charge model in which each charge is weighted according to the relative population of that particular conformation as calculated from the Boltzmann distribution [Reynolds *et al.* 1992]. In a few charge models the charges vary continuously with the conformation [Rappé and Goddard 1991; Dinur and Hagler 1995].

4.9.6 Rapid Methods for Calculating Atomic Charges

Some methods calculate atomic charges solely from information about the atoms present in the molecule and the way in which the atoms are connected. The great advantage of such methods is that they are very fast and can be used to calculate the charge distributions for large numbers of molecules (e.g. in a database). We will consider the Gasteiger and Marsili method [Gasteiger and Marsili 1980] as an example.

The Gasteiger–Marsili approach uses the concept of the *partial equalisation of orbital electronegativity*. Electronegativity is a concept well known to chemists, being defined by Pauling as ‘the power of an atom to attract electrons to itself’. Mulliken subsequently defined the electronegativity of an atom A as the average of its ionisation potential I_A and its electron affinity E_A :

$$\chi_A = \frac{1}{2}(I_A + E_A) \quad (4.38)$$

As Mulliken pointed out, the ionisation potential and electron affinity are specific to a given valence state of an atom, and therefore the electronegativities of an atom’s valence states would not be expected to be the same. This idea can be extended to the concept of orbital

electronegativity, which is the electronegativity of a specific orbital in a given valence state. For example, an sp orbital has a higher electronegativity than an sp^3 orbital. The orbital electronegativity will also depend on the occupancy of the orbital; an empty orbital will be better able to attract an electron than an orbital with a single electron, which in turn will be better than an orbital with two electrons. The electronegativity of an orbital will also be affected by the charges in other orbitals. Gasteiger and Marsili assumed a polynomial relationship between the orbital electronegativity $\chi_{\mu A}$ of an orbital ϕ_μ in atom A and the charge Q_A on the atom A:

$$\chi_{\mu A} = a_\mu + b_{\mu A} Q_A + c_{\mu A} Q_A^2 \quad (4.39)$$

Values of the coefficients a , b and c were derived for common elements in their usual valence states (for example, for carbon there are different values for sp^3 , $sp^2\pi$ and $sp\pi^2$ valence states).

Electrons flow from the less electronegative elements to the more electronegative ones. This flow of electrons results in a positive charge on the less electronegative atoms and a negative charge on the more electronegative atoms, and as such the flow acts to equalise the electronegativities. Total equalisation of electronegativity does not, however, lead to chemically sensible results. This effect is modelled in the Gasteiger and Marsili approach by an iterative procedure, in which less and less charge is transferred between bonded atoms at each step. The electron charge transferred from an atom A to an atom B (where B is more electronegative than A) in iteration k is given by:

$$Q^{(k)} = \frac{\chi_B^{(k)} - \chi_A^{(k)}}{\chi_A^+} \alpha^k \quad (4.40)$$

In Equation (4.40), $Q^{(k)}$ is the charge (in electrons) transferred; $\chi_A^{(k)}$ and $\chi_B^{(k)}$ are the electronegativities of the atoms A and B; χ_A^+ is the electronegativity of the cation of the less electronegative atom and α is a damping factor which is raised to the power k . Gasteiger and Marsili set α to $\frac{1}{2}$. The charge on each atom is initially assigned its formal charge. In each iteration, the electronegativities are calculated using Equation (4.39) and hence the charge to be transferred. The total charge on an atom at the end of each iteration is thus obtained by adding the charge transferred from all bonds to the atom to the value of the charge from the previous iteration. The damping factor α^k reduces the influence of the more electronegative atoms. This influence decreases with each iteration. With a damping factor of $\frac{1}{2}$ rapid convergence is achieved, usually within four or five steps.

A somewhat related method is the charge equilibration method of Rappé and Goddard [Rappé and Goddard 1991]. This is employed in the 'Universal Force Field' (UFF) [Rappé *et al.* 1992] as a general method for calculating charge distributions over a very wide range of molecules (in principle, the entire periodic table). An additional feature of the method is that the charges are dependent upon the molecular geometry and so can change during the course of a calculation such as a molecular dynamics simulation. The starting point for this approach is a series expansion of the energy of an isolated atom in terms of the charge:

$$\nu_A(q) = \nu_{A0} + q_A \left(\frac{\partial \nu}{\partial q} \right)_{A0} + \frac{1}{2} q_A^2 \left(\frac{\partial^2 \nu}{\partial q^2} \right)_{A0} + \dots \quad (4.41)$$

Truncating this expansion after second-order terms and considering three specific states (for charges of 0, +1 and -1) leads to:

$$\nu_A(0) = \nu_{A0} \quad (4.42)$$

$$\nu_A(+1) = \nu_{A0} + q_A \left(\frac{\partial \nu}{\partial q} \right)_{A0} + \frac{1}{2} q_A^2 \left(\frac{\partial^2 \nu}{\partial q^2} \right)_{A0} \quad (4.43)$$

$$\nu_A(-1) = \nu_{A0} - q_A \left(\frac{\partial \nu}{\partial q} \right)_{A0} + \frac{1}{2} q_A^2 \left(\frac{\partial^2 \nu}{\partial q^2} \right)_{A0} \quad (4.44)$$

Now the energy of the positive species is the ionisation potential (*IP*) and the energy of the negative species is minus the electron affinity (*EA*). Combining these results gives:

$$\left(\frac{\partial \nu}{\partial q} \right)_{A0} = \frac{1}{2}(IP + EA) = \chi_A^0 \quad (4.45)$$

$$\left(\frac{\partial^2 \nu}{\partial q^2} \right)_{A0} = IP - EA \quad (4.46)$$

As usual, χ_A is the electronegativity. Rappé and Goddard suggested that for a neutral atom with a singly occupied orbital the difference between the ionisation potential and the electron affinity would correspond to the Coulomb repulsion between two electrons placed in that orbital (the orbital would be unoccupied in the positive ion and doubly occupied in the negative species). Writing this difference as J_{AA}^0 (referred to as the *idempotential*) leads to:

$$\nu_A(q) = \nu_{A0} + \chi_A^0 q_A + \frac{1}{2} J_{AA}^0 q_A^2 \quad (4.47)$$

Both the electronegativity and the idempotential can be derived from atomic data, though such atomic data generally need to be corrected for use in molecular systems. In order to use these equations to derive a set of charges for a molecule we first consider the total electrostatic energy of the system:

$$\mathcal{V}(q_1 \cdots q_N) = \sum_{i=1}^N (\nu_{A0} + \chi_A^0 q_A + \frac{1}{2} q_A^2 J_{AA}^0) + \sum_{A=1}^N \sum_{B=A+1}^N q_A q_B J_{AB} \quad (4.48)$$

In this equation J_{AB} represents a formulation of the Coulomb energy between charges q_A and q_B . For well-separated atoms a simple $1/r$ dependency is used. However, this simple Coulomb law is not appropriate for atoms whose charge distributions overlap. In such circumstances (which particularly arise for bonded atoms) there is a significant shielding correction. This shielding correction is a Coulomb integral (Equation (2.107)), with the atomic density being described using a single Slater type orbital whose precise form depends on the nature (ns, np or nd) of the outer valence orbital together with the covalent radius.

In order to derive the actual charges we first incorporate the factors J_{AA}^0 (the limiting value of J_{AA} as the distance tends to zero) into the double summation in Equation (4.48):

$$\mathcal{V}(q_1 \cdots q_N) = \sum_{A=1}^N (\nu_{A0} + \chi_A^0 q_A) + \frac{1}{2} \sum_{A=1}^N \sum_{B=1}^N q_A q_B J_{AB} \quad (4.49)$$

We can then take the derivative of the energy with respect to q_A , which leads to:

$$\frac{\partial \mathcal{V}}{\partial q_A} = \chi_A^0 + \sum_{A=1}^N q_B J_{AB} = \chi_A^0 + J_{AA}^0 q_A + \sum_{B=1, B \neq A}^N q_B J_{AB} \quad (4.50)$$

The derivative of the energy with respect to the charge is an atomic chemical potential; at equilibrium these chemical potentials will all be equal. The electrons move from regions of low electronegativity (high electrochemical potential) to regions of high electronegativity (low electrochemical potential). A further constraint is that the sum of the atomic charges must sum to the total charge on the molecule. These conditions enable a set of simultaneous equations to be written (subject to per-element limits on the charge on any given atom).

The presence of the $q_A q_B$ term with its implied distance dependency means that the charges depend upon the molecular geometry. Thus, should the conformation of a molecule change the atomic charges will also change. Just three parameters are required for each atom in the system (the electronegativity, the idempotential and the covalent radius).

4.9.7 Beyond Partial Atomic Charge Models

Most of the charge models that we have considered so far place the charge on the nuclear centres. Atom-centred charges have many advantages. For example, the electrostatic forces due to charge-charge interactions then act directly on the nuclei. This is important if one wishes to calculate the forces on the nuclei as is required for energy minimisation or a molecular dynamics simulation. Nuclear-centred charges do nevertheless suffer from some drawbacks. In particular, they assume that the charge density about each atom is spherically symmetrical. However, an atom's valence electrons are often distributed in a far from spherical manner, especially in molecules that contain features such as lone pairs and π electron clouds above aromatic ring systems.

4.9.8 Distributed Multipole Models

One way to represent the anisotropy of a molecular charge distribution is to use *distributed multipoles*. In this model, point charges, dipoles, quadrupoles and higher multipoles are distributed throughout the molecule. These distributed multipoles can be determined in various ways but the distributed multipole analysis (DMA) model of A J Stone [Stone 1981; Stone and Alderton 1985] is probably the best-known example. The DMA method calculates the multipoles from a quantum mechanics wavefunction defined in terms of Gaussian basis functions. As we saw in Section 2.6, the overlap between two Gaussian functions can be represented by another Gaussian located at a point (P) along the line that connects them. Each product of basis functions $\phi_\mu \phi_\nu$ thus corresponds to a charge density at P. This density can be expressed as a multipole expansion about P. The highest multipole moment in the local expansion depends upon the basis set used; no multipole moment higher than the sum of the angular quantum numbers of the basis set is possible. Thus, when using a basis set that contains just s and p functions there will be local multipoles no higher than the quadrupole. The crucial feature is that the local multipole expansion

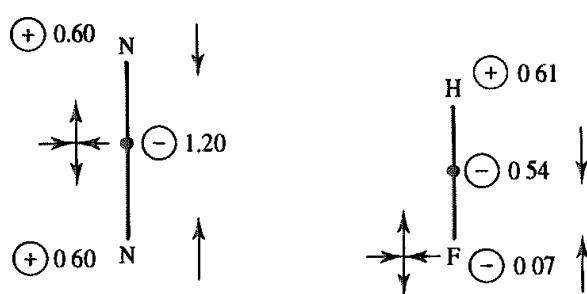


Fig 4.22: Distributed multipole models for N_2 and HF (Figure adapted from Stone A J and M Alderton 1985 *Distributed Multipole Analysis Methods and Applications*. Molecular Physics 56:1047–1064.)

about P can be represented as a multipole expansion about another nearby point S. In the distributed multipole approach, a set of site points is chosen and then the local multipole expansion for each pair of basis functions is ‘moved’ from the relevant point P to one of the sites S.

There are no limitations on the number or location of the multipole sites S; a natural set to use is obtained by placing a site point on each atomic nucleus. In some applications (especially for small molecules) additional sites are defined at the centres of bonds. For example, Stone derived a distributed multipole model for nitrogen from a Dunning [5s4p2d] basis set with two polarisation functions. This model contains charges of +0.60 on the nuclei and a charge of -1.20 at the centre of the bond, together with a dipole on each of the two nuclei and a quadrupole located at the centre of the bond (see Figure 4.22). For HF charges are placed on the two nuclei and at the centre of the bond with a dipole and a quadrupole on the fluorine and a small dipole at the centre of the bond (Figure 4.22). In larger molecules not every atom may be given a site, such as hydrogen atoms bonded to apolar atoms. It is also possible to restrict the order of the multipole expansion at a given atom so that, for example, only a charge component would be present on a polar hydrogen with the higher moments being represented by multipoles on the atom to which it is bonded. An important consideration when choosing the multipole sites is that, when a local multipole expansion is moved, the resulting multipole expansion is no longer a truncated series. However, the smaller the distance between P and the corresponding site point S, the quicker the series converges. In practice, therefore, each local multipole moment expansion is either moved to the nearest site point or is divided between the two nearest site points when they are equally close. With a basis set that contains just s and p functions and multipole sites at the atomic nuclei, it is usually found that the distributed multipole series converges rapidly after the quadrupole term. The multipoles themselves can vary considerably with the basis set used to perform the *ab initio* calculation, but the various electronic properties derived from them usually do not change much.

The distributed multipole model automatically includes non-spherical, anisotropic effects due to features such as lone pairs or π electrons. The original applications of the DMA approach were to small molecules such as diatomics and triatomics. The method has since been used to develop models for nucleic acids and for peptides and has even been applied to the undecapeptide cyclosporin [Price *et al.* 1989], which contains 199 atoms (the

quantum mechanical calculation on this molecule used 1000 basis functions). However, distributed multipole models have not yet been widely incorporated into force fields, not least because of the additional computational effort required. It can be complicated to calculate the atomic forces with the distributed multipole model; in particular, multipoles that are not located on atoms generate torques, which must be analysed further to determine the forces on the nuclei.

4.9.9 Using Charge Schemes to Study Aromatic–Aromatic Interactions

The attractive interactions between molecules containing π systems have long been studied by theoreticians and experimentalists. Such systems are involved in a variety of phenomena, including the stacking of the nucleic acid bases in DNA, the packing of aromatic molecules in crystals and interactions between amino acid side chains in proteins. A variety of orientations are observed for aromatic dimers, ranging from edge-on, T-shaped structures to face-to-face structures (Figure 4.23). Within these two families the molecules can move relative to each other, so that, for example, in a face-to-face arrangement the atoms are overlaid or are staggered. In the T-shaped structure the large quadrupole moments of the benzene molecules adopt their most favourable orientation.

One very simple model of the interactions in such systems was devised by Hunter and Saunders [Hunter and Saunders 1990], who wanted to explain the stacking behaviour of aromatic systems such as the porphyrins shown in Figure 4.24. It is experimentally observed that these molecules adopt a cofacial arrangement with their centres offset as shown. Hunter and Saunders placed point charges not only at the nuclei but also at locations above and below each atom, perpendicular to the plane of the ring. Thus in benzene each carbon atom was given a charge of +1 and also had two associated charges of $-\frac{1}{2}$ above and below the ring (Figure 4.25). The electrostatic interaction between two ring systems is

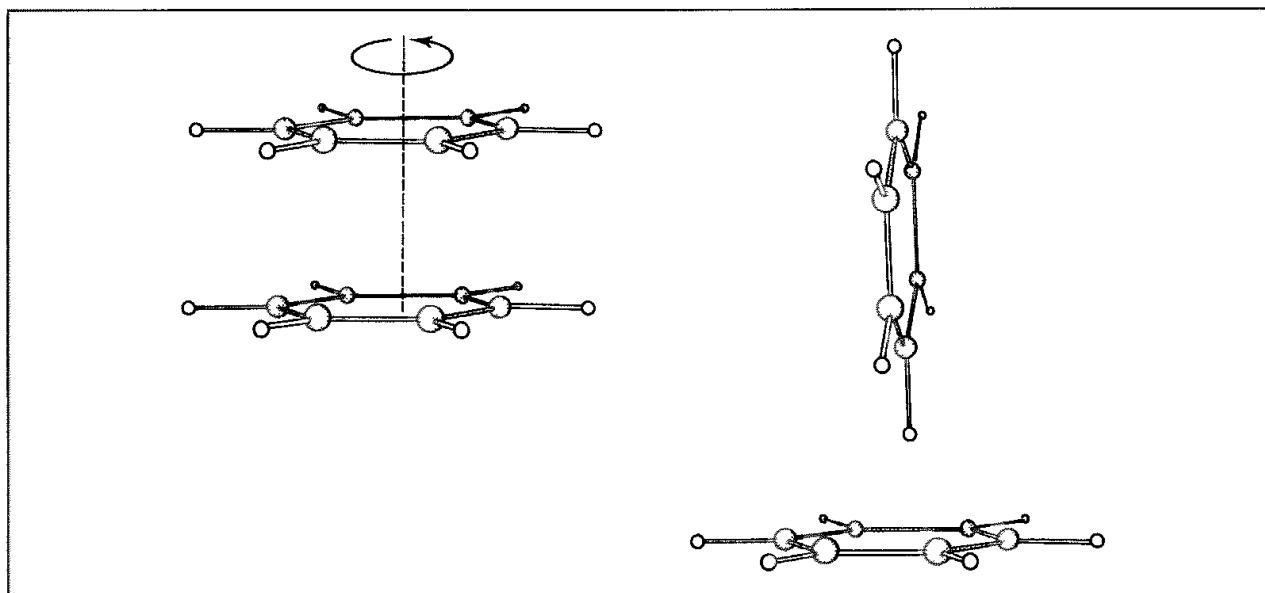


Fig. 4.23 Face-to-face (left) and T-shaped (right) orientations of the benzene dimer

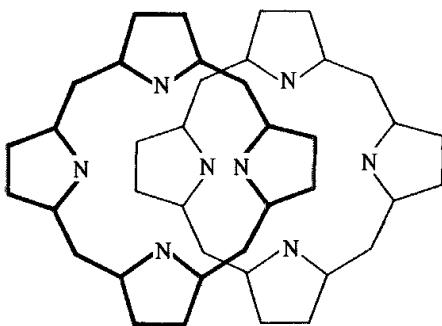


Fig. 4.24: Porphyrin system typical of those studied by Hunter and Saunders [Hunter and Saunders 1990]

calculated in the usual way by summing the charge-charge interactions using Coulomb's law. A major advantage of the Hunter-Saunders approach is its computational simplicity. Moreover, it can be extended to cover a wide range of atom types and so applied to many systems [Vinter 1994] with particular emphasis on simulating DNA [Hunter 1993, Packer *et al.* 2000]. Hunter and Saunders summarised the results of their investigations on porphyrins in three rules:

1. $\pi-\pi$ repulsion dominates in a face-to-face geometry;
2. $\pi-\sigma$ attraction dominates in an edge-on geometry;
3. $\pi-\sigma$ attraction dominates in an offset π -stacked geometry.

The interactions between aromatic systems have also been studied using point charge models, central multipoles and distributed multipoles. Fowler and Buckingham examined homodimers of *sym*-triazine and 1,3,5-trifluorobenzene (Figure 4.26) [Fowler and Buckingham 1991]. They were particularly keen to calculate how the electrostatic energy changed as the rings were twisted in the face-to-face geometry. All but one of the energy models suggested that the staggered orientations were the arrangements of minimum energy, but the energy difference between the eclipsed and staggered structures varied widely, depending upon the model. The central multipole model was found to be ineffective due to convergence problems. Three different point-charge models were considered, all of

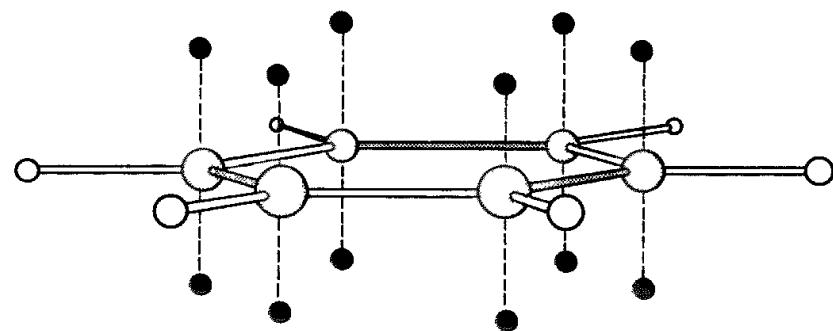


Fig. 4.25: Anisotropic model of benzene developed by Hunter and Saunders [Hunter and Saunders 1990]

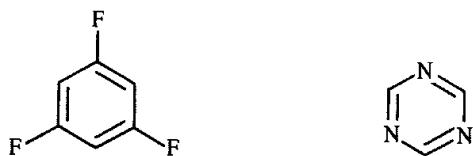


Fig 4.26. Sym-triazine and 1,3,5-trifluorobenzene.

which gave acceptable energy curves. The distributed multipole model also performed well, being comparable to the most accurate of the point-charge models.

4.9.10 Polarisation

Our discussion of electronic effects has concentrated so far on 'permanent' features of the charge distribution. Electrostatic interactions also arise from changes in the charge distribution of a molecule or atom caused by an external field, a process called *polarisation*. The primary effect of the external electric field (which in our case will be caused by neighbouring molecules) is to induce a dipole in the molecule. The magnitude of the induced dipole moment μ_{ind} is proportional to the electric field E , with the constant of proportionality being the polarisability α :

$$\mu_{\text{ind}} = \alpha E \quad (4.51)$$

The energy of interaction between a dipole μ_{ind} and an electric field E (the induction energy) is determined by calculating the work done in charging the field from zero to E , using the following integral:

$$v(\alpha, E) = - \int_0^E dE \mu_{\text{ind}} = - \int_0^E dE \alpha E = - \frac{1}{2} \alpha E^2 \quad (4.52)$$

In strong electric fields contributions to the induced dipole moment that are proportional to E^2 or E^3 can also be important, and higher-order moments such as quadrupoles can also be induced. We will not be concerned with such contributions.

For isolated atoms, the polarisability is isotropic – it does not depend on the orientation of the atom with respect to the applied field, and the induced dipole is in the direction of the electric field, as in Equation (4.51). However, the polarisability of a molecule is often anisotropic. This means that the orientation of the induced dipole is not necessarily in the same direction as the electric field. The polarisability of a molecule is often modelled as a collection of isotropically polarisable atoms. A small molecule may alternatively be modelled as a single isotropic polarisable centre.

Let us consider the electric field due to a dipole μ aligned along the z axis. The magnitude of the electric field at a point P due to the dipole (see Figure 4.27) is:

$$E(r, \theta) = \frac{\mu \sqrt{1 + 3 \cos^2 \theta}}{4\pi\epsilon_0 r^3} \quad (4.53)$$

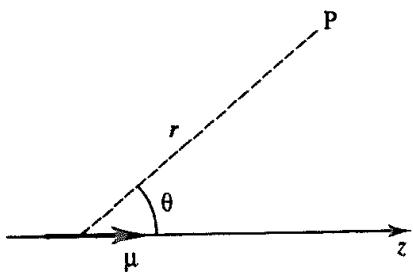


Fig. 4.27: Electric field at point P due to dipole at the origin

The induction energy with another molecule of polarisability α placed at P is therefore

$$\nu(r, \theta) = -\alpha\mu^2 \frac{1 + 3\cos^2\theta}{(4\pi\epsilon_0 r^3)^2} \quad (4.54)$$

The interaction between a dipole and an induced dipole is independent of the disorienting effect of thermal motion, whereas the dipole-dipole interaction between two permanent dipoles does vary with the relative orientation of the two dipoles. This is because the induced dipole follows the direction of the permanent dipole even as the molecules change their orientations as a consequence of molecular collisions.

An important consideration when modelling polarisation effects is that the dipole induced on a molecule (A) will affect the charge distribution of another molecule (B). The electric field at A due to the dipole(s) on B will in turn be affected. The presence of other molecules can also influence the interaction. Consider the polarisation interaction between a polar molecule and a neighbour (Figure 4.28). A third molecule may reduce the size of the electric field on the second molecule and so lower the induction energy. This type of three-body effect will be particularly significant when polarisable atoms are close to polar groups. Polarisation is a cooperative effect and, as such, is modelled using a set of coupled equations which are typically solved iteratively. Initially, the induced dipoles are set to zero. An initial approximation to each induced dipole is then calculated from the permanent charges (i.e. partial atomic charges). The electric field due to these induced dipoles is then added to the electric field from the permanent charges. This gives a refined value of the electric field from which a new induced dipole can be determined. The calculation continues until the induced dipoles do not change significantly between iterations.

A variety of schemes for including polarisation into molecular mechanics force fields have been devised. One approach is to model the polarisation effects at the atomic level, with

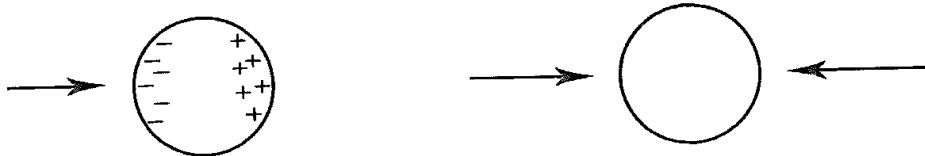


Fig. 4.28: The polarisation interaction between a dipole and a polarisable molecule can be affected by the presence of a second dipole (right) and is therefore a many-body effect

dipoles being induced on each atom [Dang *et al.* 1991]. The magnitude of the dipole induced on an atom i is given by:

$$\mu_{\text{ind},i} = \alpha_i E_i \quad (4.55)$$

α_i is the atomic polarisability, assumed to be isotropic. Appropriate values of α_i have been determined for various systems. The electric field, E_i , at atom i is the vector sum of the field due to the permanent and induced dipoles of the other atoms in the system:

$$E_i = \sum_{j \neq i} \frac{q_j r_{ij}}{r_{ij}^3} + \sum_{j \neq i} \frac{\mu_j}{r_{ij}^3} \left(3r_{ij} \frac{r_{ij}}{r_{ij}^2} - 1 \right) \quad (4.56)$$

r_i and r_j are the position vectors of the atoms i and j . Convergence of these equations in procedures such as molecular dynamics, where successive configurations are generated, can be accelerated if the induced dipoles obtained at each current step are used as the starting points for the next configuration.

An alternative way to model polarisation effects is exemplified by the water model of Sprik and Klein [Sprik and Klein 1988], where the polarisation centre is represented as a collection of closely spaced charges whose values are permitted to vary but whose total sums to zero. In the water model, shown in Figure 4.29, four tetrahedrally arranged charges are used to model the polarisation centre. These charges endow the molecule with an induced dipole moment of any magnitude and direction. The charges are determined iteratively for each configuration of the system. The isotropic polarisability of a simple ion can similarly be treated using two charges of equal magnitude but opposite sign placed either side of the ion. The direction of the 'bond' linking the two polarisation charges and the ion can reorient to change the direction of the induced dipole. In a subsequent refinement of this model Sprik and Klein replaced the point charges by Gaussian charge distributions at the polarisation sites; these were better at modelling features such as hydrogen bonding.

One appealing approach is the dynamically fluctuating charge model of Berne and colleagues [Rick *et al.* 1994]. This method has much in common with the charge equilibration scheme of Rappé and Goddard (see Section 4.9.6) in its use of the electronegativity equalisation approach, which ensures that the atomic chemical potentials are equal in the molecule. The charges are considered as dynamically fluctuating variables, along with the atomic nuclei in a molecular dynamics simulation. This means that the charges evolve in a natural

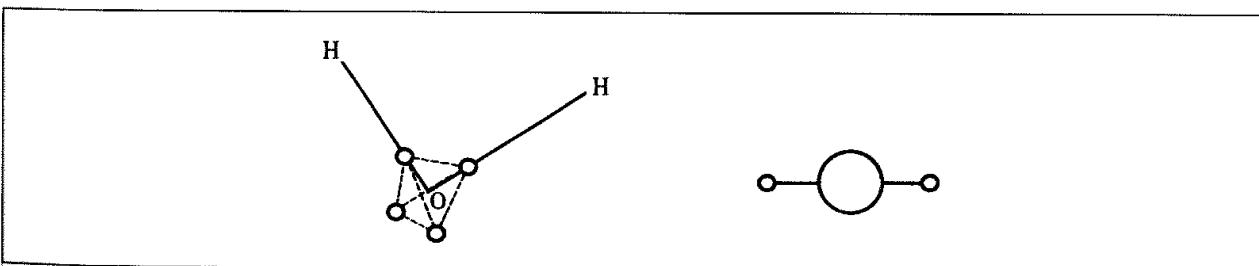


Fig. 4.29. Polarisable models of water and ions developed by Sprik and Klein. (Figure adapted from Sprik M 1993 Effective Pair Potentials and Beyond In Computer Simulation in Chemical Physics, Allen M P, D J Tildesley (Editors) Dordrecht, Kluwer)

manner during the simulation rather than having to determine a new set of charges at each iteration of the procedure. This fluctuating charge model includes intramolecular interactions and so the traditional Coulombic $1/r$ expression is not appropriate. Rather, the charges are replaced by charge distributions (formulated as Slater's orbitals) whose interaction is calculated using a Coulomb integral expression. This interaction is effectively identical to the standard Coulomb expression for intermolecular interactions, only differing for the intramolecular contribution.

One feature of this oscillating charge model is that it requires rather less computational effort than traditional polarisation models. It also implicitly preserves the higher-order multipole terms, which need to be explicitly incorporated in some of the alternative approaches. Ions are represented by two partial charges (which sum to the required integral ionic charge) which are connected by a harmonic spring. The mass of one of these two species is made much greater than the other so that the heavier site remains near the centre of mass as the spring oscillates. This particular model has been used for simulations of pure liquid water [Rick *et al.* 1994], the solvation of amides [Rick and Berne 1996] and to investigate the effects of polarisability on the hydration of the chloride ion in water clusters [Stuart and Berne 1996]. These calculations predicted that the chloride ions were located on the outside of the clusters, even when they contained more than 100 water molecules. This was in contrast to equivalent calculations using a non-polarisable model, the difference being attributed to the presence of fluctuations in the dipole strengths of the water molecules in the cluster, which are, as a consequence, more mobile.

Due to the computational expense, polarisation effects are often included in a calculation only when their effect is likely to be significant, such as simulations of ionic solutions. These systems usually contain atoms or ions and small molecules only. It is important to be aware of the following problem when using atomic polarisabilities. Consider a diatomic molecule. The application of an external field will induce dipoles on both atoms. The dipole on one atom will also contribute to the electric field at the other atom, and thereby influence its induced dipole, but the model takes no account of the fact that the charge distributions on the two atoms are inherently linked. For this reason (and for reasons of computational efficiency) it is common to treat small molecules such as water as single polarisable centres when calculating polarisation effects.

4.9.11 Solvent Dielectric Models

All of the formulae that we have written for electrostatic energies, potentials and forces have included the permittivity of free space, ϵ_0 . This is as one would expect for species acting in a vacuum. However, under some circumstances a different dielectric model is used in the equations for the electrostatic interactions. This is often done when it is desired to mimic solvent effects, without actually including any explicit solvent molecules. One effect of a solvent is to dampen the electrostatic interactions. A very simple way to model this damping effect is to increase the permittivity, most easily by using an appropriate value for the relative permittivity in the Coulomb's law equation (i.e. $\epsilon = \epsilon_0 \epsilon_r$). An alternative approach is to make the dielectric dependent upon the separation of the charged species; this gives rise

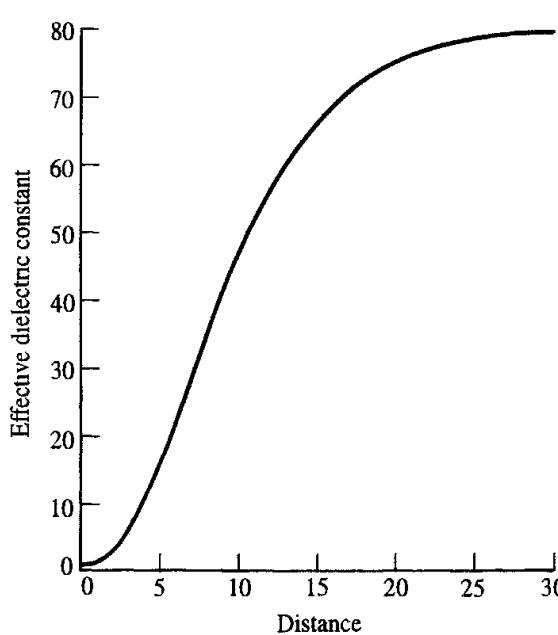


Fig. 4.30 A sigmoidal dielectric model smoothly varies the effective permittivity from 80 to 1 as shown

to the so-called distance-dependent dielectric models. The simplest implementation of a distance-dependent dielectric is to make the relative permittivity proportional to the distance. The interaction energy between two charges q_i and q_j then becomes:

$$\nu(r) = \frac{1}{4\pi\epsilon_0} \frac{q_i q_j}{r^2} \quad (4.57)$$

The simple distance-dependent dielectric has no physical basis and so it is not generally recommended, except when no alternative is possible. More sophisticated distance-dependent functions can also be employed. Many of these have an approximately sigmoidal shape in which the relative permittivity is low at short distances and then rises towards the bulk value at long distances. One example of such a function is [Smith and Pettit 1994]:

$$\epsilon_{\text{eff}}(r) = \epsilon_r - \frac{\epsilon_r - 1}{2} [(rS)^2 + 2rS + 2] e^{-rS} \quad (4.58)$$

The value of ϵ_{eff} varies from a value of 1 at zero separation to ϵ_r (the bulk permittivity of the solvent) at large distances, in a manner determined by the parameter S (which is typically given a value between 0.15 \AA^{-1} and 0.3 \AA^{-1} ; Figure 4.30). Sigmoidal functions give better behaviour than the simple distance-dependent dielectric model. However, it may be difficult to choose the appropriate value for the bulk dielectric ϵ_r when performing calculations on large solutes, as the shortest distance between two charges may be through the solute molecule rather than through the solvent (Figure 4.31).

The polarisation term can be a major contributor to the free energy of solvation of a solute, and a variety of schemes have been devised to incorporate such effects where the solvent is modelled as a continuum. We shall discuss these methods in more detail in Sections 11.9–11.12.

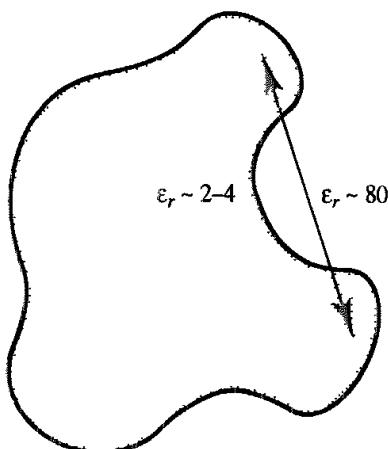


Fig. 4.31 A line joining two points may pass through regions of different permittivity

4.10 Van der Waals Interactions

Electrostatic interactions cannot account for all of the non-bonded interactions in a system. The rare gas atoms are an obvious example; all of the multipole moments of a rare gas atom are zero and so there can be no dipole-dipole or dipole-induced dipole interactions. But there clearly must be interactions between the atoms, how else could rare gases have liquid and solid phases or show deviations from ideal gas behaviour? Deviations from ideal gas behaviour were famously quantitated by van der Waals, thus the forces that give rise to such deviations are often referred to as van der Waals forces.

If we were to study the interaction between two isolated argon atoms using a molecular beam experiment then we would find that the interaction energy varies with the separation in a manner as shown in Figure 4.32. The other rare gases show a similar behaviour. The essential features of this curve are as follows. The interaction energy is zero at infinite distance (and indeed is negligible even at relatively short distances). As the separation is reduced, the energy decreases, passing through a minimum at a distance of approximately 3.8 Å for argon. The energy then rapidly increases as the separation decreases further. The force between the atoms, which equals minus the first derivative of the potential energy with respect to distance, is also shown in Figure 4.32. A variety of experiments have been used to provide evidence for the nature of the van der Waals interactions, including gas imperfections, molecular beams, spectroscopic studies and measurements of transport properties.

4.10.1 Dispersive Interactions

The curve in Figure 4.32 is usually considered to arise from a balance between attractive and repulsive forces. The attractive forces are long-range, whereas the repulsive forces act at short distances. The attractive contribution is due to *dispersive forces*. London first showed how the dispersive force could be explained using quantum mechanics [London 1930] and so this interaction is sometimes referred to as the London force. The dispersive force

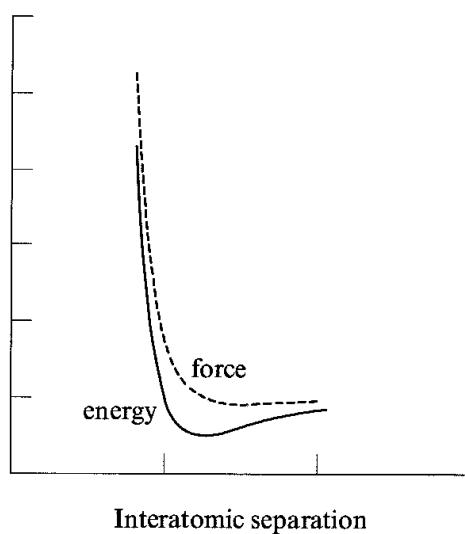


Fig. 4.32: The interaction energy and the force between two argon atoms

is due to instantaneous dipoles which arise during the fluctuations in the electron clouds. An instantaneous dipole in a molecule can in turn induce a dipole in neighbouring atoms, giving rise to an attractive inductive effect.

A simple model to explain the dispersive interaction was proposed by Drude. This model consists of 'molecules' with two charges, $+q$ and $-q$, separated by a distance r . The negative charge performs simple harmonic motion with angular frequency ω along the z axis about the stationary positive charge (Figure 4.33). If the force constant for the oscillator is k and if the mass of the oscillating charge is m , then the potential energy of an isolated Drude molecule is $\frac{1}{2}kz^2$, where z is the separation of the two charges. ω is related to the force constant by $\omega = \sqrt{k/m}$. The Schrödinger equation for a Drude molecule is:

$$-\frac{\hbar^2}{2m} \frac{\partial^2 \psi}{\partial z^2} + \frac{1}{2}kz^2\psi = E\psi \quad (4.59)$$

This is the Schrödinger equation for a simple harmonic oscillator. The energies of the system are given by $E_\nu = (\nu + \frac{1}{2}) \times \hbar\omega$ and the zero-point energy is $\frac{1}{2}\hbar\omega$.

We now introduce a second Drude molecule, identical to the first, with the positive charge also located on the z axis and an oscillating negative charge (Figure 4.33). When the two molecules are infinitely separated, they do not interact and the total ground-state energy of the system is

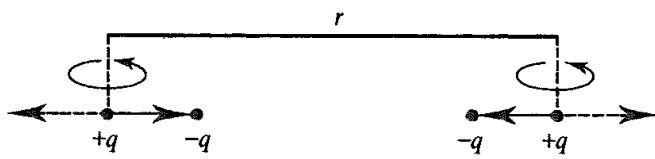


Fig. 4.33 The Drude model for dispersive interactions (Figure adapted from Rigby M, E B Smith, W A Wakeham and G C Maitland 1986 The Forces Between Molecules Oxford, Clarendon Press)

just twice the zero-point energy of a single molecule, $\hbar\omega/2\pi$. As the molecules approach (along the z axis) there are interactions between the two dipoles, and the interaction energy between the two ‘molecules’ can be shown to be approximately given by (see Appendix 4.1):

$$\nu(r) = -\frac{\alpha^4 \hbar\omega}{2(4\pi\epsilon_0)^2 r^6} \quad (4.60)$$

The Drude model thus predicts that the dispersion interaction varies as $1/r^6$.

The two-dimensional Drude model can be extended to three dimensions, the result being:

$$\nu(r) = -\frac{3\alpha^4 \hbar\omega}{4(4\pi\epsilon_0)^2 r^6} \quad (4.61)$$

The Drude model only considers the dipole–dipole interaction; if higher-order terms, due to dipole–quadrupole, quadrupole–quadrupole, etc., interactions are included as well as other terms in the binomial expansion, then the energy of the Drude model is more properly written as a series expansion:

$$\nu(r) = \frac{C_6}{r^6} + \frac{C_8}{r^8} + \frac{C_{10}}{r^{10}} + \dots \quad (4.62)$$

All of the coefficients C_n are negative, implying an attractive interaction. Despite its simplicity, the Drude model gives quite reasonable results; if just the C_6 term is included then for argon the resulting dispersion energy is only about 25% too small.

4.10.2 The Repulsive Contribution

Below about 3 Å, even a small decrease in the separation between a pair of argon atoms causes a large increase in the energy. This increase has a quantum mechanical origin and can be understood in terms of the Pauli principle, which formally prohibits any two electrons in a system from having the same set of quantum numbers. The interaction is due to electrons with the same spin, therefore the short-range repulsive forces are often referred to as *exchange forces*. They are also known as overlap forces. The effect of exchange is to reduce the electrostatic repulsion between pairs of electrons by forbidding them to occupy the same region of space (i.e. the internuclear region). The reduced electron density in the internuclear region leads to repulsion between the incompletely shielded nuclei. At very short internuclear separations, the interaction energy varies as $1/r$ due to this nuclear repulsion, but at larger separations the energy decays exponentially, as $\exp(-2r/a_0)$, where a_0 is the Bohr radius.

4.10.3 Modelling Van der Waals Interactions

The dispersive and exchange-repulsive interactions between atoms and molecules can be calculated using quantum mechanics, though such calculations are far from trivial, requiring electron correlation and large basis sets. For a force field we require a means to model the interatomic potential curve accurately (Figure 4.32), using a simple empirical

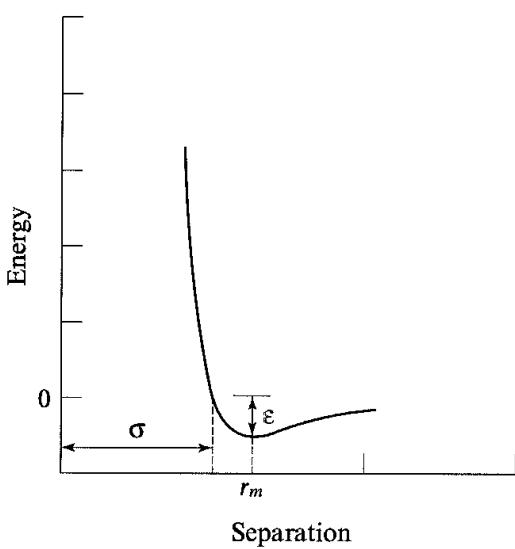


Fig 4.34. The Lennard-Jones potential.

expression that can be rapidly calculated. The need for a function that can be rapidly evaluated is a consequence of the large number of van der Waals interactions that must be determined in many of the systems that we would like to model. The best known of the van der Waals potential functions is the *Lennard-Jones 12-6 function*, which takes the following form for the interaction between two atoms:

$$\nu(r) = 4\varepsilon \left[\left(\frac{\sigma}{r} \right)^{12} - \left(\frac{\sigma}{r} \right)^6 \right] \quad (4.63)$$

The Lennard-Jones 12-6 potential contains just two adjustable parameters: the collision diameter σ (the separation for which the energy is zero) and the well depth ε . These parameters are graphically illustrated in Figure 4.34. The Lennard-Jones equation may also be expressed in terms of the separation at which the energy passes through a minimum, r_m (also written r^*). At this separation, the first derivative of the energy with respect to the internuclear distance is zero (i.e. $\partial\nu/\partial r = 0$), from which it can easily be shown that $r_m = 2^{1/6}\sigma$. We can thus also write the Lennard-Jones 12-6 potential function as follows:

$$\nu(r) = \varepsilon \left\{ (r_m/r)^{12} - 2(r_m/r)^6 \right\} \quad (4.64)$$

or

$$\nu(r) = A/r^{12} - C/r^6 \quad (4.65)$$

A is equal to εr_m^{12} (or $4\varepsilon\sigma^{12}$) and C is equal to $2\varepsilon r_m^6$ (or $4\varepsilon\sigma^6$).

The Lennard-Jones potential is characterised by an attractive part that varies as r^{-6} and a repulsive part that varies as r^{-12} . These two components are drawn in Figure 4.35. The r^{-6} variation is of course the same power-law relationship found for the leading term in theoretical treatments of the dispersion energy such as the Drude model. There are no strong theoretical arguments in favour of the repulsive r^{-12} , especially as quantum

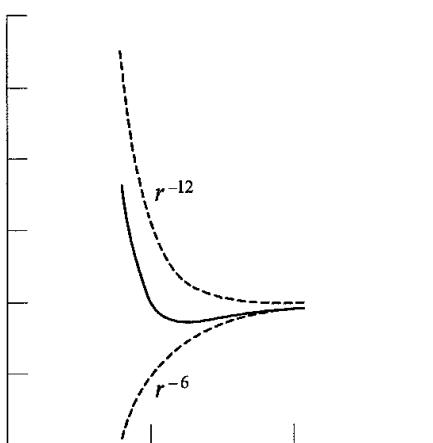


Fig. 4.35 The Lennard-Jones potential is constructed from a repulsive component (αr^{-12}) and an attractive component (αr^{-6})

mechanics calculations suggest an exponential form. The twelfth power term is found to be quite reasonable for rare gases but is rather too steep for other systems such as hydrocarbons. However, the 6–12 potential is widely used, particularly for calculations on large systems, as r^{-12} can be rapidly calculated by squaring the r^{-6} term. The r^{-6} term can also be calculated from the square of the distance without having to perform a computationally expensive square root calculation. Different powers have also been used for the repulsive part of the potential; values of 9 or 10 give a less steep curve and are used in some force fields. Lennard-Jones' original potential has been written in the following general form:

$$\nu(r) = k\epsilon \left[\left(\frac{\sigma}{r} \right)^n - \left(\frac{\sigma}{r} \right)^m \right]; \quad k = \frac{n}{n-m} \left(\frac{n}{m} \right)^{m/(n-m)} \quad (4.66)$$

Equation (4.66) returns the Lennard-Jones potential for $n = 12$ and $m = 6$.

Halgren has proposed an alternative functional form designed to be simple enough to be easily incorporated into molecular mechanics calculations whilst also improving the ability to reproduce experimental data [Halgren 1992, 1996a, b]. In this sense it is an attempt to improve on the Lennard-Jones potential without introducing the complexity of some of the potentials employed by spectroscopists. This potential has the general form:

$$\nu(r) = \epsilon_{ij} \left(\frac{1+\delta}{\rho_{ij} + \delta} \right)^{(n-m)} \left(\frac{1+\gamma}{\rho_{ij}^m + \gamma} - 2 \right) \quad (4.67)$$

In this equation $\rho_{ij} = r_{ij}/r_{ij}^*$. The constants δ and γ apply to all interactions between the atoms i and j . This potential reduces to the standard Lennard-Jones 12–6 potential if the following choice of parameters is used: $n = 12$, $m = 6$, $\delta = \gamma = 0$. Halgren proposed a 'buffered 14–7' potential in which $n = 14$, $m = 7$, $\delta = 0.07$ and $\gamma = 0.12$, giving the following equation:

$$\nu(r) = \epsilon_{ij} \left(\frac{1.07r_{ij}^*}{r_{ij} + 0.07r_{ij}^*} \right)^7 \left(\frac{1.12r_{ij}^7}{r_{ij}^7 + 0.12r_{ij}^{*7}} \right) \quad (4.68)$$

There were several reasons for developing this functional form. First was the desire to keep the potential finite as the interatomic potential approaches zero (unlike the Lennard-Jones function, which becomes infinite). Second, it gives a more accurate reproduction of the series expansion for the dispersion interaction, Equation (4.62). Third, if a larger value of d is used then the repulsive component is greatly reduced without significantly changing the distance at which the potential crosses zero or the depth of the energy minimum. This feature is useful for optimising structures with crude initial geometries; other functional forms can have significant problems with such situations.

In the buffered 14-7 potential the minimum-energy separation r_{ii}^* for an atom i depends on its atomic polarisability:

$$r_{ii}^* = A_i \alpha_i^{1/4} \quad (4.69)$$

Several formulations in which the r^{-12} term in the standard Lennard-Jones formulation is replaced by a theoretically more realistic exponential expression have been proposed. These include the *Buckingham potential*:

$$\nu(r) = \varepsilon \left[\frac{6}{\alpha - 6} \exp[-\alpha(r/r_m - 1)] - \frac{\alpha}{\alpha - 6} \left(\frac{r_m}{r} \right)^6 \right] \quad (4.70)$$

There are three adjustable parameters in the Buckingham potential (ε , r_m and α). A value of α between approximately 14 and 15 gives a potential that closely corresponds to the Lennard-Jones 12-6 potential in the minimum-energy region. When using the Buckingham potential it is important to remember that at very short distances the potential becomes strongly attractive, as shown in Figure 4.36. This could lead to nuclei being fused together during a calculation, and so the program must check that atoms are not becoming too close. The

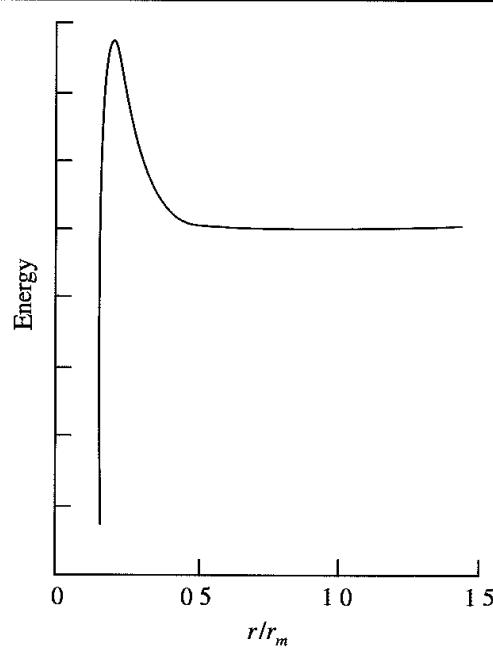


Fig 4.36 A drawback of the Buckingham potential is that it becomes steeply attractive at short distances.

Hill potential is an exponential-6 potential with just two parameters: the minimum energy radius r_m and the well depth ε [Hill 1948]:

$$\nu(r) = -2.25\varepsilon(r_m/r)^6 + 8.28 \times 10^5 \varepsilon \exp(-r/0.0736r_m) \quad (4.71)$$

The Hill potential was originally developed to enable the more realistic exponential term to be written in terms of Lennard-Jones parameters. The coefficients 2.25, 8.25×10^5 and 0.0736 in Equation (4.71) were determined by fitting to data for the rare gases and were assumed to be applicable to other non-polar gases. A Morse potential may also be used to model the van der Waals interactions in a force field, with appropriate parameters.

4.10.4 Van der Waals Interactions in Polyatomic Systems

The interaction energy between molecules depends not only upon their separation but also on their relative orientations and, where appropriate, their conformations. It is usual to calculate the van der Waals interaction energy between two molecules using a site model in which the interaction is determined as the sum of the interactions between all pairs of sites on the two molecules. The sites are often identified with the nuclear positions, but this need not necessarily be the case.

Polyatomic systems invariably involve the calculation of van der Waals interactions between different types of atoms. For example, to calculate the Lennard-Jones interaction energy between two carbon monoxide molecules using a two-site model would require not only van der Waals parameters for the carbon–carbon interactions and the oxygen–oxygen interactions but also for the carbon–oxygen interactions. A system containing N different types of atom would require $N(N - 1)/2$ sets of parameters for the interaction between unlike atoms. The determination of van der Waals parameters can be a difficult and time-consuming process and so it is common to assume that parameters for the cross interactions can be obtained from the parameters of the pure atoms using *mixing rules*. In the commonly used Lorentz–Berthelot mixing rules, the collision diameter σ_{AB} for the A–B interaction equals the arithmetic mean of the values for the two pure species, and the well depth ε_{AB} is given as the geometric mean:

$$\sigma_{AB} = \frac{1}{2}(\sigma_{AA} + \sigma_{BB}) \quad (4.72)$$

$$\varepsilon_{AB} = \sqrt{\varepsilon_{AA}\varepsilon_{BB}} \quad (4.73)$$

When written in terms of the separation of minimum energy (r^* or r_m), the following notation may be encountered:

$$r_{AB}^* = R_{AA}^* + R_{BB}^* \quad (4.74)$$

R_{AA}^* and R_{BB}^* are atomic parameters, equal to one half of r_{AA}^* and r_{BB}^* , respectively.

The Lorentz–Berthelot combining rules are most successful when applied to similar species. Their major failing is that the well depth can be overestimated by the geometric mean rule. Some force fields calculate the collision diameter for mixed interactions as the geometric mean of the values for the two component atoms. Jorgensen's OPLS force field falls into this category [Jorgensen and Tirado-Reeves 1988].

For the buffered 14–7 functional form more elaborate combination rules are employed:

$$r_{ij}^* = \frac{(r_{ii}^{*3} + r_{jj}^{*3})}{(r_{ii}^{*2} + r_{jj}^{*2})} \quad (4.75)$$

This is similar in spirit to the arithmetic-mean rule but with each individual r_{ii}^* being weighted according to the square of its value. The well depth in this function starts with a formula proposed by Slater and Kirkwood for the C_6 coefficient of the dispersion series expansion:

$$C_{6ij} = \frac{3}{2} \frac{\alpha_i \alpha_j}{(\alpha_i/N_i)^{1/2} + (\alpha_j/N_j)^{1/2}} = \frac{2\alpha_i \alpha_j}{\alpha_i^2 C_{ijj} + \alpha_j^2 C_{6ii}} \quad (4.76)$$

In this equation N represents the effective number of electrons and α are atomic polarisabilities; the second formulation in Equation (4.76) is derived using the relationship:

$$N_i = 16C_{6ii}^2/9\alpha_i^3 \quad (4.77)$$

From this the well depths ε are then obtained as follows:

$$\varepsilon_{ij} = \frac{1}{2} \frac{kG_i G_j C_{6ij}}{r_{ij}^{*6}} = \frac{181.16 G_i G_j \alpha_i \alpha_j}{(\alpha_i/N_i)^{1/2} + (\alpha_j/N_j)^{1/2}} \frac{1}{r_{ij}^{*6}} \quad (4.78)$$

Here, k is a factor which converts to units (kcal/mol in this case where the distances are in Å and the polarisabilities in Å³). G_i and G_j are constants chosen to reproduce the well depths for like-with-like interactions. The atomic polarisability values are obtained from an examination of appropriate molecular experimental data (such as measurements of molar refractivity).

In some force fields the interaction sites are not all situated on the atomic nuclei. For example, in the MM2, MM3 and MM4 programs, the van der Waals centres of hydrogen atoms bonded to carbon are placed not at the nuclei but are approximately 10% along the bond towards the attached atom. The rationale for this is that the electron distribution about small atoms such as oxygen, fluorine and particularly hydrogen is distinctly non-spherical. The single electron from the hydrogen is involved in the bond to the adjacent atom and there are no other electrons that can contribute to the van der Waals interactions. Some force fields also require lone pairs to be defined on particular atoms; these have their own van der Waals and electrostatic parameters.

The van der Waals and electrostatic interactions between atoms separated by three bonds (i.e. the 1,4 atoms) are often treated differently from other non-bonded interactions. The interaction between such atoms contributes to the rotational barrier about the central bond, in conjunction with the torsional potential. These 1,4 non-bonded interactions are often scaled down by an empirical factor; for example, a factor of 2.0 is suggested for both the electrostatic and van der Waals terms in the 1984 AMBER force field (a scale factor of 1/1.2 is used for the electrostatic terms in the 1995 AMBER force field). There are several reasons why one would wish to scale the 1,4 interactions. The error associated with the use of an r^{-12} repulsion term (which is too steep compared with the more correct exponential term) would be most significant for 1,4 atoms. In addition, when two 1,4

atoms come close together some redistribution of the charge along the connecting bonds would be expected that would act to reduce the interaction. Such a charge redistribution would not be possible for two atoms at a similar distance apart if they were in different molecules.

The parameters for the van der Waals interactions can be obtained in a variety of ways. In the early force fields, such parameters were often determined from an analysis of crystal packing. The objective of such studies was to produce a set of van der Waals parameters which enabled the experimental geometries and thermodynamic properties such as the heat of sublimation to be reproduced as accurately as possible. More recent force fields derive their van der Waals parameters using liquid simulations in which the parameters are optimised to reproduce a range of thermodynamic properties such as the densities and enthalpies of vaporisation for appropriate liquids.

4.10.5 Reduced Units

The Lennard-Jones potential is completely specified by the two parameters ε and σ . This means that the results of a calculation performed on (say) liquid argon can be easily converted to give equivalent results for another noble gas. For this reason it is common to simulate the rare gases in terms of reduced units with ε and σ both set to 1. The results can then be converted to any system as appropriate. For example, the reduced density ρ^* is related to the real density by $\rho^* = \rho\sigma^3$; the reduced energy E^* is given by $E^* = E/\varepsilon$, and so on. Electrostatic interactions given by Coulomb's law are also often written in terms of a reduced unit of charge, which corresponds to each charge being divided by $\sqrt{4\pi\varepsilon_0}$. This means that Coulomb's law takes the less cumbersome form:

$$\nu(q_1, q_2) = q_1 q_2 / r_{12} \quad \text{or} \quad \nu(q_1, q_2) = q_1 q_2 / \varepsilon_r r_{12} \quad (4.79)$$

4.11 Many-body Effects in Empirical Potentials

The electrostatic and van der Waals energies that we have considered so far are calculated between pairs of interaction sites. The total non-bonded interaction energy is thus determined by adding together the interactions between all pairs of sites in the system. However, the interaction between two molecules can be affected by the presence of a third, fourth or more molecules. For example, the interaction energy between three molecules A, B and C is not in general given by the sum of the pairwise interaction energies: $\nu(A, B, C) \neq \nu(A, B) + \nu(A, C) + \nu(B, C)$. We have already seen an example of a non-pairwise contribution, namely the polarisation interaction, which is determined using a self-consistent procedure.

Three-body effects can significantly affect the dispersion interaction. For example, it is believed that three-body interactions account for approximately 10% of the lattice energy of crystalline argon. For very precise work, interactions involving more than three atoms may have to be taken into account, but they are usually small enough to be ignored. A potential that includes both two- and three-body interactions would be written in the following

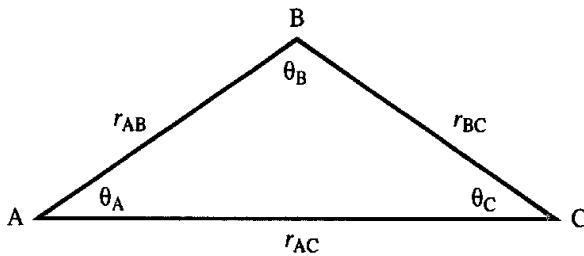


Fig. 4.37: Calculating the three-body Axilrod-Teller contribution

general form:

$$\mathcal{V}(\mathbf{r}^N) = \sum_{i=1}^N \sum_{j=i+1}^N \nu^{(2)}(r_{ij}) + \sum_{i=1}^N \sum_{j=i+1}^N \sum_{k=j+1}^N \nu^{(3)}(r_{ij}, r_{ik}, r_{jk}) \quad (4.80)$$

Axilrod and Teller investigated the three-body dispersion contribution and showed that the leading term is:

$$\nu^{(3)}(r_{AB}, r_{AC}, r_{BC}) = \nu_{A,B,C} \frac{3 \cos \theta_A \cos \theta_B \cos \theta_C}{(r_{AB} r_{AC} r_{BC})^3} \quad (4.81)$$

θ_A , θ_B and θ_C are the internal angles of the triangle with sides of length r_{AB} , r_{AC} and r_{BC} (Figure 4.37). $\nu_{A,B,C}$ is a constant characteristic of the three species A, B and C. If A, B and C are identical then $\nu_{A,B,C}$ is approximately related to the Lennard-Jones coefficient C_6 and the polarisability by

$$\nu_{A,B,C} = -\frac{3\alpha C_6}{4(4\pi\epsilon_0)} \quad (4.82)$$

The effect of the Axilrod-Teller term (also known as the triple-dipole correction) is to make the interaction energy more negative when three molecules are linear but to weaken it when the molecules form an equilateral triangle. This is because the linear arrangement enhances the correlations of the motions of the electrons, whereas the equilateral arrangement reduces it.

The three-body contribution may also be modelled using a term of the form $\nu^{(3)}(r_{AB}, r_{AC}, r_{BC}) = K_{A,B,C} \{ \exp(-\alpha r_{AB}) \exp(-\beta r_{AC}) \exp(-\gamma r_{BC}) \}$ where K , α , β and γ are constants describing the interaction between the atoms A, B and C. Such a functional form has been used in simulations of ion-water systems, where polarisation alone does not exactly model configurations when there are two water molecules close to an ion [Lybrand and Kollman 1985]. The three-body exchange repulsion term is thus only calculated for ion-water-water trimers when the species are close together.

The computational effort is significantly increased if three-body terms are included in the model. Even with a simple pairwise model, the non-bonded interactions usually require by far the greatest amount of computational effort. The number of bond, angle and torsional terms increases approximately with the number of atoms (N) in the system, but the number of non-bonded interactions increases with N^2 . There are $N(N - 1)/2$ distinct pairs of

interactions to evaluate for a pairwise potential. If three-body effects are included then there are $N(N - 1)(N - 2)/6$ unique three-body interactions. A system with 1000 atoms has 499 500 pairwise interactions and 166 167 000 three-body interactions. In general, there are approximately $N/3$ times more three-body terms than two-body terms and so it is clear why it is often considered preferable to avoid calculating the three-body interactions.

4.12 Effective Pair Potentials

Fortunately, it is found that a significant proportion of the many-body effects can be incorporated into a pairwise model, if properly parametrised. The pair potentials most commonly used in molecular modelling are thus ‘effective’ pairwise potentials; they do not represent the true interaction energy between two isolated particles but are parametrised to include many-body effects in the pairwise energy. Similarly, polarisation effects can be implicitly included in a force field by the simple expedient of enhancing the electrostatic interaction. This can be done by using larger partial charges than those for an isolated molecule. This is most obviously manifested in larger multipole moments; the dipole moment of a single water molecule is 1.85 D, whereas the dipole moment of many simple water models designed to simulate liquid water are significantly larger (closer to the experimental value for liquid water of 2.6 D).

A notable example of a potential that does include many-body terms is the Barker–Fisher–Watts potential for argon, which combines a pairwise potential with an Axilrod–Teller triple

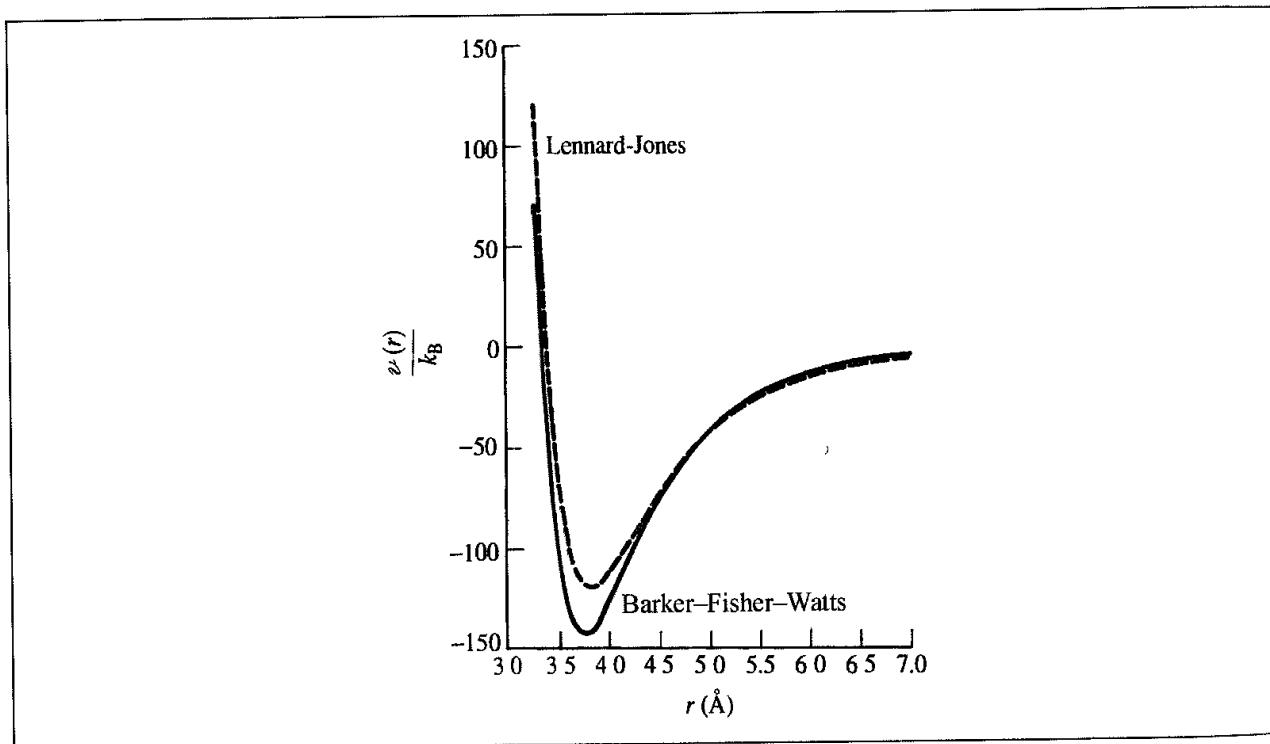


Fig. 4.38. Comparison of the Lennard-Jones potential for argon with the Barker–Fisher–Watts pair potential; k_B is Boltzmann’s constant

potential [Barker *et al.* 1971]. The pair potential is a linear combination of two potentials that each take the following form:

$$\begin{aligned} \nu^*(r) = & e^{\alpha(1-r^*)}[A_0 + A_1(r^* - 1) + A_2(r^* - 1)^2 + A_3(r^* - 1)^3 + A_4(r^* - 1)^4 + A_5(r^* - 1)^5] \\ & + \frac{C_6}{\delta + R^{*6}} + \frac{C_8}{\delta + R^{*8}} + \frac{C_{10}}{\delta + R^{*10}} \end{aligned} \quad (4.83)$$

This potential function contains eleven constants: α , $A_0 \dots A_5$, C_6 , C_8 , C_{10} and δ . The function is expressed in terms of r^* , which is given by $r^* = r/r_m$, where r_m is the separation at the minimum in the potential. The 'true' interaction energy as a function of the separation, r , is then obtained by multiplying $\nu^*(r^*)$ by the depth of the potential well, ε :

$$\nu(r) = \varepsilon \nu^*(r^*) \quad (4.84)$$

A comparison of the pairwise contribution to the Barker–Fisher–Watts potential with the Lennard–Jones potential for argon is shown in Figure 4.38.

4.13 Hydrogen Bonding in Molecular Mechanics

Some force fields replace the Lennard–Jones 6–12 term between hydrogen-bonding atoms by an explicit hydrogen-bonding term, which is often described using a 10–12 Lennard–Jones potential:

$$\nu(r) = \frac{A}{r^{12}} - \frac{C}{r^{10}} \quad (4.85)$$

This function is used to model the interaction between the donor hydrogen atom and the heteroatom acceptor atom. Its use is intended to improve the accuracy with which the geometry of hydrogen-bonding systems is predicted. Other force fields incorporate a more complicated hydrogen-bonding function that takes into account deviations from the geometry of the hydrogen bond and is thus dependent upon the coordinates of the donor and acceptor atoms as well as the hydrogen atom. For example, the YETI force field [Vedani 1988] uses the following form for its hydrogen bonding term:

$$\nu_{\text{HB}} = \left(\frac{A}{r_{\text{H Acc}}^{12}} - \frac{C}{r_{\text{H Acc}}^{10}} \right) \cos^2 \theta_{\text{Don H Acc}} \cos^4 \omega_{\text{H Acc-LP}} \quad (4.86)$$

The energy in Equation (4.86) depends upon the distance from the hydrogen to the acceptor, the angle subtended at the hydrogen by the bonds to the donor and the acceptor, and the deviation of the hydrogen bond from the closest lone-pair direction at the acceptor atom ($\omega_{\text{H Acc-LP}}$ in Equation (4.86), Figure 4.39).

The GRID program [Goodford 1985] that is used for finding energetically favourable regions in protein binding sites uses a direction-dependent 6–4 function:

$$\nu_{\text{HB}} = \left(\frac{C}{d^6} - \frac{D}{d^4} \right) \cos^m \theta \quad (4.87)$$

θ is the angle subtended at the hydrogen and m is usually set to 4.

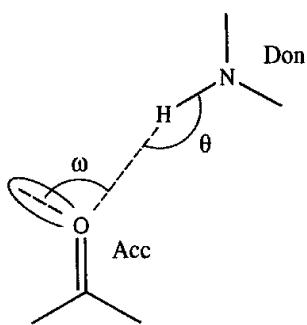


Fig 4.39: Definition of hydrogen-bond geometry used in YETI force field

By no means do all force fields contain explicit hydrogen-bonding terms; most rely upon electrostatic and van der Waals interactions to reproduce hydrogen bonding.

4.14 Force Field Models for the Simulation of Liquid Water

Many of the concepts that we have considered so far can be illustrated by examining some of the empirical models that have been developed to study water. Despite its small size, water acts as a paradigm for the different force field models that we have discussed. Moreover, many of its properties can be easily determined using computer simulation methods and so readily compared with experiment. It is also one of the most challenging systems to model accurately. A wide range of water models have been proposed. The computational efficiency with which the energy can be calculated using a given model is often an important factor as there may be a very large number of water molecules present, together with a solute; most of the force fields used to simulate liquid water thus use effective pairwise potentials with no explicit three-body terms or polarisation effects.

Water models can be conveniently divided into three types. In the simple interaction-site models each water molecule is maintained in a rigid geometry and the interaction between molecules is described using pairwise Coulombic and Lennard-Jones expressions. Flexible models permit internal changes in conformation of the molecule. Finally, models have been developed that explicitly include the effects of polarisation and many-body effects.

4.14.1 Simple Water Models

The ‘simple’ water models use between three and five interaction sites and a rigid water geometry. The TIP3P [Jorgensen *et al.* 1983] and SPC [Berendsen *et al.* 1981] models use a total of three sites for the electrostatic interactions; the partial positive charges on the hydrogen atoms are exactly balanced by an appropriate negative charge located on the oxygen atom. The van der Waals interaction between two water molecules is computed using a Lennard-Jones function with just a single interaction point per molecule centred on the oxygen atom; no van der Waals interactions involving the hydrogen atoms are calculated. The TIP3P and SPC models differ slightly in the geometry of each water molecule, in the

	SPC	SPC/E	TIP3P	BF	TIP4P	ST2
$r(\text{OH})$, Å	1.0	1.0	0.9572	0.96	0.9572	1.0
HOH, deg	109.47	109.47	104.52	105.7	104.52	109.47
$A \times 10^{-3}$, kcal Å ¹² /mol	629.4	629.4	582.0	560.4	600.0	238.7
C , kcal Å ⁶ /mol	625.5	625.5	595.0	837.0	610.0	268.9
$q(\text{O})$	-0.82	-0.8472	-0.834	0.0	0.0	0.0
$q(\text{H})$	0.41	0.4238	0.417	0.49	0.52	0.2375
$q(\text{M})$	0.0	0.0	0.0	-0.98	-1.04	-0.2375
$r(\text{OM})$, Å	0.0	0.0	0.0	0.15	0.15	0.8

Table 4.3 A comparison of various water models [Jorgensen et al. 1983]. For the ST2 potential, $q(\text{M})$ is the charge on the 'lone pairs', which are a distance 0.8 Å from the oxygen atom (see Figure 4.40)

hydrogen charges and in the Lennard-Jones parameters. These differences are indicated in Table 4.3, which also includes data for the SPC/E model [Berendsen et al. 1987], which is an updated version of the SPC model. The four-site models such as that of Bernal and Fowler [Bernal and Fowler 1933] (which is now relatively little used but is important for historical reasons as it dates from 1933) and Jorgensen's TIP4P model [Jorgensen et al. 1983] shift the negative charge from the oxygen atom to a point along the bisector of the HOH angle towards the hydrogens (Figure 4.40). The parameters for these two models are also given in the table. The most commonly used five-site model is the ST2 potential of Stillinger and Rahman [Stillinger and Rahman 1974]. Here, charges are placed on the hydrogen atoms and on two lone-pair sites on the oxygen. The electrostatic contribution is modulated so that for oxygen-oxygen distances below 2.016 Å it is zero and for distances greater than 3.1287 Å it takes its full value. Between these two distances the electrostatic contribution is modulated using a function that smoothly varies from 0.0 at the shorter distance to 1.0 at the longer distance (see Section 6.7.3).

The experimentally determined dipole moment of a water molecule in the gas phase is 1.85 D. The dipole moment of an individual water molecule calculated with any of these simple models is significantly higher; for example, the SPC dipole moment is 2.27 D and that for TIP4P is 2.18 D. These values are much closer to the effective dipole moment of liquid water, which is approximately 2.6 D. These models are thus all effective pairwise models. The simple water models are usually parametrised by calculating various properties using molecular dynamics or Monte Carlo simulations and then modifying the

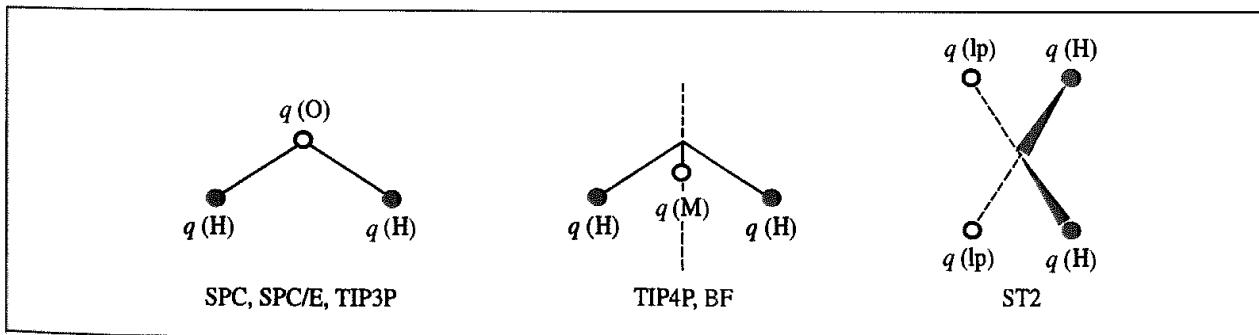


Fig. 4.40. Some 'simple' water models (Table 4.3) [Jorgensen et al. 1983].

parameters until the desired level of agreement between experiment and theory is achieved. Thermodynamic and structural properties are usually used in the parametrisation, such as the density, radial distribution function, enthalpy of vaporisation, heat capacity, diffusion coefficient and dielectric constant.* It is found that some properties such as the density and the enthalpy of vaporisation are predicted rather well by all of the models, but there is significant variation in the values for other properties such as the dielectric constant [Jorgensen *et al.* 1983]. When comparing the different models, it is also important to take account of the computational effort each requires. Thus, nine site-site distances must be calculated for each water dimer using a three-site model; ten are required for a four-site model, and seventeen for the ST2 model.

The use of a rigid model for water is obviously an approximation, and it means that some properties cannot be determined at all. For example, only when internal flexibility is included can the vibrational spectrum be calculated and compared with experiment. Flexibility is most easily incorporated by 'grafting' bond-stretching and angle-bending terms onto the potential function for a rigid model. Such an approach needs to be done with care. For example, Ferguson has developed a flexible model for water that is based upon the SPC model [Ferguson 1995]. The partial charges and van der Waals parameters in this model were slightly different from those in the rigid model, and flexibility was achieved using cubic and harmonic bond-stretching terms and a harmonic angle-bending term. The calculated values compared well with experimental results for a wide range of thermodynamic and structural properties, including the dielectric constant and self-diffusion coefficient.

4.14.2 Polarisable Water Models

The simple models give very good results for a wide range of properties of pure liquid water. However, there is some concern that they are not appropriate models to use for the most accurate work. This is especially the case for inhomogeneous systems where there are strong electric field gradients due to the presence of ions, and at the solute-solvent interface. Under such circumstances models that explicitly include polarisation effects and three-body terms are considered to be more appropriate. The inclusion of an explicit polarisation term should also enhance the ability of the model to reproduce the behaviour of water in other phases (e.g. solid and vapour) and at the interface between different phases. The dipole moment of an isolated water molecule in such a model should thus be closer to the gas-phase value rather than to the 'effective' value in liquid water. The simplest way to include polarisation is to use an isotropic molecular polarisability contribution, an alternative is to use atom-centred polarisabilities or the variable charge method. The incorporation of polarisability may significantly increase the computational effort required for a liquid simulation, and even then only the best polarisable models currently compete with the well-established models that use effective pairwise potentials. We have already considered some of the polarisable water models in our discussion of polarisation effects. One early attempt to incorporate such effects into a water model was made by Barnes,

* A discussion of the calculation of these properties from computer simulation is given in Section 6.2

Finney, Nicholas and Quinn [Barnes *et al.* 1979]. Their polarisable electropole water model represented the charge distribution by a multipole expansion comprising a dipole of 1 855 D and a quadrupole moment that was determined from quantum mechanical calculations on an isolated molecule. Polarisation effects were calculated using an isotropic molecular polarisability from the electric fields being produced by the dipoles and quadrupoles of surrounding molecules. The model also used a spherically symmetric Lennard-Jones function. A more recent study used the fluctuating charge model with both the TIP4P and SPC geometries [Rick *et al.* 1994]. The charges were assigned to reproduce the correct dipole moment of the gas-phase molecule (in contrast to the equivalent non-polarisable models). Of the two geometries, the TIP4P model gave the better results for various properties. The dielectric properties were considered particularly well reproduced, including features in the dielectric spectrum arising from the translational motion of a water molecule in the cage of its neighbours. This feature is not present in fixed-charge models. Moreover, the computational cost with this particular model was only about 1.1 times that of the fixed-charge equivalent.

4.14.3 *Ab initio* Potentials for Water

The final category of water model that we shall consider are the '*ab initio*' potentials. These are based upon *ab initio* quantum mechanical calculations on small clusters of water molecules. One example of this type is the NCC model of Nieser, Corongiu and Clementi, which combines a two-molecule potential with a polarisation term [Niesar *et al.* 1990]. They had previously tried to explicitly include both three- and four-body effects but found this model computationally too expensive. The two-body model uses partial charges on the hydrogen atoms and a compensating negative charge on a site located along the bisector of the HOH angle, as in the TIP4P model. The equation used is:

$$\begin{aligned} \mathcal{V}_{\text{two-body}} = & q^2 \left(\frac{1}{R_{13}} + \frac{1}{R_{14}} + \frac{1}{R_{23}} + \frac{1}{R_{24}} \right) \\ & + \frac{4q^2}{R_{78}} - 2q^2 \left(\frac{1}{R_{81}} + \frac{1}{R_{82}} + \frac{1}{R_{73}} + \frac{1}{R_{74}} \right) \\ & + A_{OO} e^{-B_{OO}R_{56}} + A_{HH} (e^{-B_{HH}R_{13}} + e^{-B_{HH}R_{14}} + e^{-B_{HH}R_{23}} + e^{-B_{HH}R_{24}}) \\ & + A_{OH} (e^{-B_{OH}R_{53}} + e^{-B_{OH}R_{54}} + e^{-B_{OH}R_{61}} + e^{-B_{OH}R_{62}}) \\ & - A'_{OH} (e^{-B'_{OH}R_{53}} + e^{-B'_{OH}R_{54}} + e^{-B'_{OH}R_{61}} + e^{-B'_{OH}R_{62}}) \\ & + A_{PH} (e^{-B_{PH}R_{73}} + e^{-B_{PH}R_{74}} + e^{-B_{PH}R_{81}} + e^{-B_{PH}R_{82}}) \\ & + A_{PO} (e^{-B_{PO}R_{76}} + e^{-B_{PO}R_{85}}) \end{aligned} \quad (4.88)$$

The points P are the locations where the negative charge is placed (numbered 7 and 8 in Figure 4.41) and the terms A_{PH} and A_{PO} are used to enhance the performance of the model at short distances. q is the charge on each hydrogen. The polarisation term is calculated in an iterative manner using induced dipoles along each O–H bond. The NCC model was parametrised by fitting to the energies and other properties of 250 trimer and

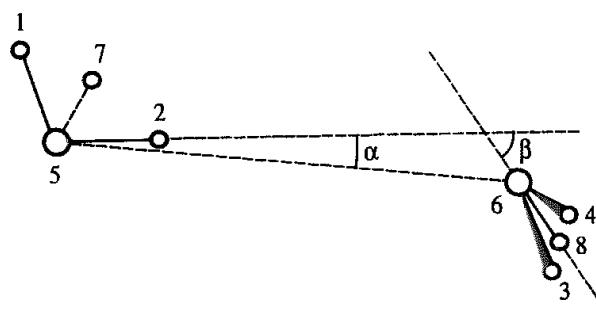


Fig 4.41 The NCC water model (After Corongiu G 1992 Molecular Dynamics Simulation for Liquid Water Using a Polarisable and Flexible Potential International Journal of Quantum Chemistry 42:1209–1235)

350 dimer configurations determined with high-level *ab initio* methods and large basis sets. The water trimer data was used to fit the many-body parameters (i.e. the locations of the induced dipole moments and the point charges, together with the polarisability and the value of the hydrogen charge). The dimer data were then used to fit the remaining terms in the potential.

The original NCC potential was designed as a rigid water model and performed well in tests of its ability to reproduce experimental data for both water dimers and liquid water. A flexible version has also been developed [Corongiu 1992], with the energy being expressed as a function of the three internal coordinates (two bond lengths and one angle) with terms up to quartics:

$$\begin{aligned}
 V_{\text{intra}} = & \frac{1}{2} f_{RR}(\delta_1^2 + \delta_2^2) + \frac{1}{2} f_{\theta\theta}(\delta_3^2) + f_{RR'}\delta_1\delta_2 + f_{R\theta}(\delta_1 + \delta_2)\delta_3 \\
 & + \frac{1}{R_e} [f_{RRR}(\delta_1^3 + \delta_2^3) + f_{\theta\theta\theta}\delta_3^3 + f_{RRR'}(\delta_1 + \delta_2)\delta_1\delta_2 \\
 & \quad + f_{RR\theta}(\delta_1^2 + \delta_2^2)\delta_3 + f_{RR'\theta}\delta_1\delta_2\delta_3 + f_{R\theta\theta}(\delta_1 + \delta_2)\delta_3^2] \\
 & + \frac{1}{R_e^2} [f_{RRRR}(\delta_1^4 + \delta_2^4) + f_{\theta\theta\theta\theta}\delta_3^4 + f_{RRR'R'}(\delta_1^2 + \delta_2^2)\delta_1\delta_2 \\
 & \quad + f_{RRR'R'}\delta_1^2\delta_2^2 + f_{RRR\theta}(\delta_1^3 + \delta_2^3)\delta_3] \\
 & + \frac{1}{R_e^2} [f_{RRR'\theta}(\delta_1 + \delta_2)\delta_1\delta_2\delta_3 + f_{RR\theta\theta}(\delta_1^2 + \delta_2^2)\delta_3^2 \\
 & \quad + f_{RR'\theta\theta}\delta_1\delta_2\delta_3 + f_{R\theta\theta\theta}(\delta_1 + \delta_2)\delta_3^2]
 \end{aligned} \tag{4.89}$$

where $\delta_1 = R_1 - R_e$, $\delta_2 = R_2 - R_e$ and $\delta_3 = R_e(\theta - \theta_e)$.

The functional form of the NCC model demonstrates the complexity of some empirical models (and this for a molecule that contains only three atoms!). We should also note that the development of empirical models from *ab initio* quantum mechanical data is an approach that is already well established and looks likely to be a method that is more widely used in the future.

4.15 United Atom Force Fields and Reduced Representations

In our discussion so far, we have assumed that all of the atoms in the system are explicitly represented in the model. However, as the number of non-bonded interactions scales with the square of the number of interaction sites present, there are clear advantages if the number of interaction sites can be reduced. The simplest way to do this is to subsume some or all of the atoms (usually just the hydrogen atoms) into the atoms to which they are bonded. A methyl group would then be modelled as a single 'pseudo-atom' or 'united atom'. The van der Waals and electrostatic parameters would be modified to take account of the adjoining hydrogen atoms. Considerable computational savings are possible; for example, if butane is modelled as a four-site model rather than one with twelve atoms then the van der Waals interaction between two butane molecules involves the calculation of sixteen terms rather than 144. Other hydrocarbons are often represented using united atom models. Many of the earliest calculations on proteins used united atom representations. In this case, not all of the hydrogen atoms in the protein are subsumed into their adjacent atoms, but just those that are bonded to carbon atoms. Hydrogen atoms bonded to polar atoms such as nitrogen and oxygen are able to participate in hydrogen-bonding interactions, which are modelled much better if these hydrogens are explicitly represented.

One drawback with a united atom force field is that chiral centres may be able to invert during a calculation. This was found to be a problem with the united atom force fields for proteins. The alpha carbon in the peptide unit (C_α in Figure 4.42) is bonded to a hydrogen atom and to the side chain (glycine and proline are slightly different; see Section 10.1). A united atom force field model would not explicitly include the alpha hydrogen. Unfortunately, the stereochemistry at the alpha carbon can then invert during a calculation. This should be avoided as the naturally occurring amino acids have a defined stereochemistry (as shown in Figure 4.42). This inversion may be prevented through the use of an improper torsion term (e.g. $N-C-C_\alpha-R$) to keep the side chain in the correct relative position.

In a united atom force field the van der Waals centre of the united atom is usually associated with the position of the heavy (i.e. non-hydrogen) atom. Thus, for a united CH_3 or CH_2 group the van der Waals centre would be located at the carbon atom. It would be more accurate to associate the van der Waals centre with a position that was offset slightly from the carbon position, in order to reflect the presence of the hydrogen atoms. Toxvaerd has developed such a model that gives superior performance for alkanes than do the simple united atom models, particularly for simulations at high pressures [Toxvaerd 1990]. In



Fig 4.42 Representations of the naturally occurring amino acids

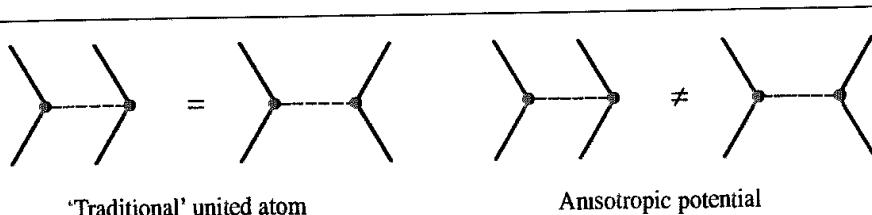


Fig. 4.43: The interaction energy between the two arrangements shown is equal in a 'traditional' united atom force field but different in the Toxvaerd anisotropic model (Figure adapted from Toxvaerd S 1990 Molecular Dynamics Calculations of the Equation of State of Alkanes The Journal of Chemical Physics 93:4290-4295)

In Toxvaerd's model the interaction sites are located at the geometrical centres of the CH₂ or CH₃ groups. The forces between these sites act on the united atom mass centre, which remains located on the carbon atom (with a mass of 14 for a CH₂ group and 15 for a CH₃ group). As the interaction site is no longer located at an atomic nucleus the forces acting on the masses are more complicated to calculate, but little additional computational expense is required. The effect of using such an anisotropic potential is nicely illustrated by the two arrangements of methylene units shown schematically in Figure 4.43. In the united atom model both arrangements would have the same energies and forces, but this is not so with the Toxvaerd anisotropic potential.

4.15.1 Other Simplified Models

In some force field models, even simpler representations are used than the united atom approach, with entire groups of atoms being modelled as single interaction points. For example, a benzene ring might be modelled as a single site with appropriately chosen parameters.

Yet other models have no obvious relationship to any ‘real’ molecule but are useful because their simplicity enables larger or more extensive calculations to be performed than would otherwise be possible. The polymer field is full of such models, as we shall discuss in Section 8.6. Another area where such models have been widely applied is in the study of liquid crystals. Liquid crystals are able to form phases that are characterised by a long-range order of the molecular orientations in at least one dimension. Many of the molecules that exhibit liquid crystalline behaviour are rod-shaped, but disc-like molecules can also form liquid crystalline phases. Some typical examples of molecules that can show such behaviour are shown in Figure 4.44. In the liquid crystalline state the rod-shaped molecules are aligned with their long axes pointing in approximately the same direction. Some very simple computer models have been used to investigate the behaviour of liquid crystals. These simple models enable large simulations to be performed on assemblies of many ‘molecules’. One example of such a simplified model is the Gay–Berne potential [Gay and Berne 1981], which models the anisotropic interaction between two particles as:

$$\nu(r_{ij}) = 4\varepsilon(\hat{\mathbf{u}}_i, \hat{\mathbf{u}}_j, \hat{\mathbf{r}}) \left\{ \left[\frac{\sigma_0}{r_{ij} - \sigma(\hat{\mathbf{u}}_i, \hat{\mathbf{u}}_j, \hat{\mathbf{r}}) + \sigma_0} \right]^{12} - \left[\frac{\sigma_s}{r_{ij} - \sigma(\hat{\mathbf{u}}_i, \hat{\mathbf{u}}_j, \hat{\mathbf{r}}) + \sigma_s} \right]^6 \right\} \quad (4.90)$$

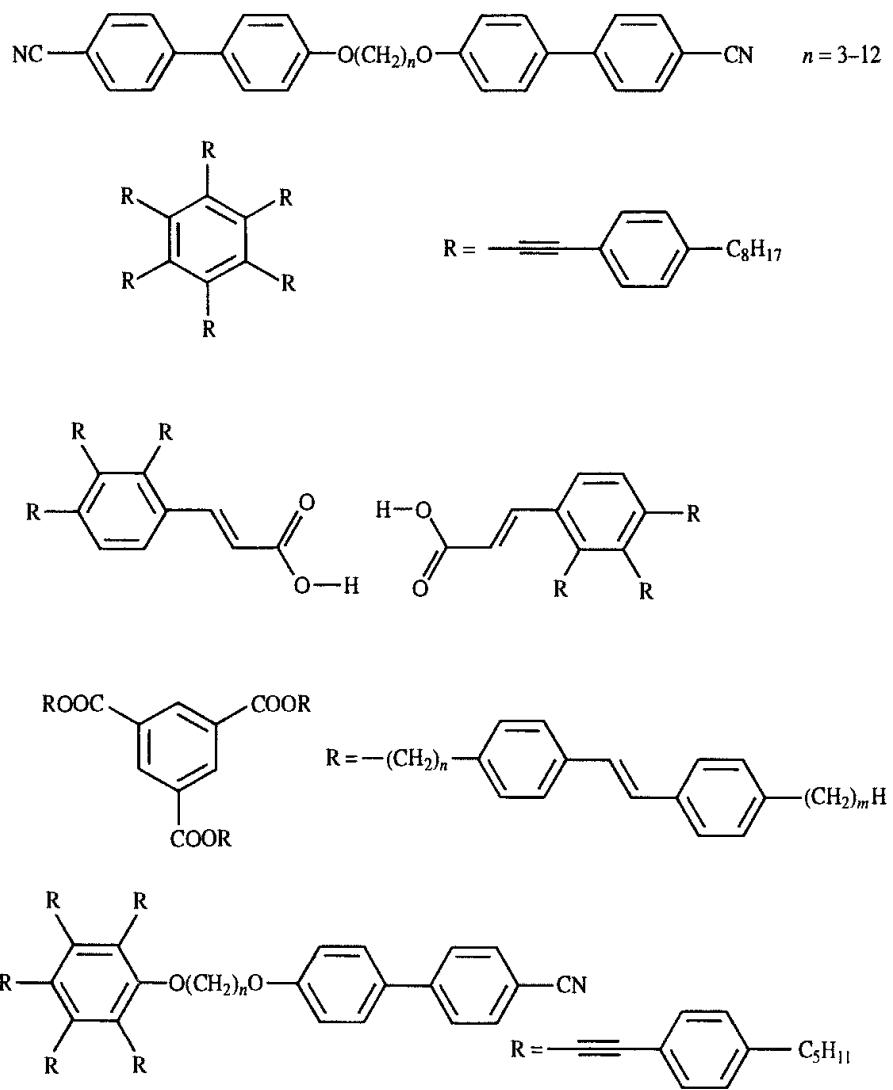


Fig. 4.44 Some typical liquid crystal molecules.

$\hat{\mathbf{u}}_i$ and $\hat{\mathbf{u}}_j$ are unit vectors that describe the orientations of the two molecules i and j and $\hat{\mathbf{r}}$ is a unit vector along the line connecting their centres (Figure 4.45). The molecules can be considered as ellipsoids which have a shape that is reflected in two size parameters, σ_s and σ_e , which are the separations at which the attractive and repulsive terms in the potential cancel for end-to-end and side-by-side arrangements respectively. These are incorporated into the potential via the parameter σ :

$$\sigma(\hat{\mathbf{u}}_i, \hat{\mathbf{u}}_j, \hat{\mathbf{r}}) = \sigma_0 \left\{ 1 - \frac{\chi}{2} \left[\frac{(\hat{\mathbf{u}}_i \cdot \hat{\mathbf{r}} + \hat{\mathbf{u}}_j \cdot \hat{\mathbf{r}})^2}{1 + \chi(\hat{\mathbf{u}}_i \cdot \hat{\mathbf{u}}_j)} + \frac{(\hat{\mathbf{u}}_i \cdot \hat{\mathbf{r}} - \hat{\mathbf{u}}_j \cdot \hat{\mathbf{r}})^2}{1 - \chi(\hat{\mathbf{u}}_i \cdot \hat{\mathbf{u}}_j)} \right] \right\}^{-1/2} \quad (4.91)$$

where

$$\chi = \frac{(\sigma_e/\sigma_s)^2 - 1}{(\sigma_e/\sigma_s)^2 + 1} \quad (4.92)$$

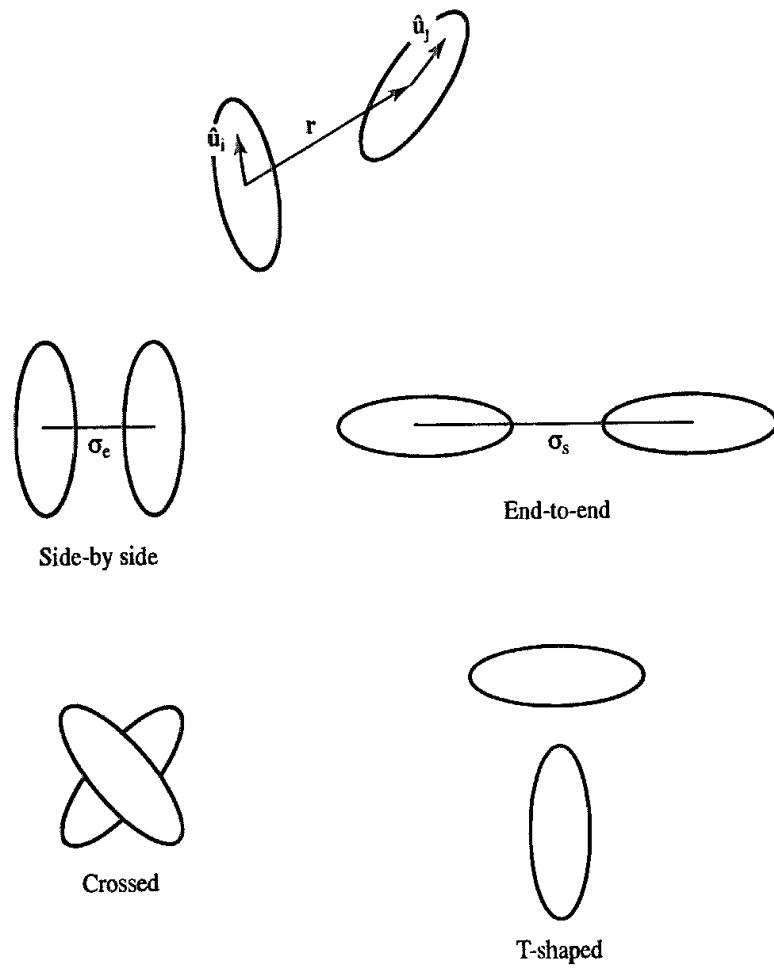


Fig. 4.45 The Gay-Berne model for liquid crystal systems and some typical arrangements.

χ is the *shape anisotropy parameter*; it is zero for spherical particles and is equal to 1 for infinitely long rods and -1 for infinitely thin discs; σ_0 is typically set equal to σ_s .

The energy term is also orientation-dependent and is written as follows:

$$\varepsilon(\hat{u}_i, \hat{u}_j, \hat{r}) = \varepsilon_0 \varepsilon'^\mu(\hat{u}_i, \hat{u}_j, \hat{r}) \varepsilon^\nu(\hat{u}_i, \hat{u}_j) \quad (4.93)$$

where

$$\begin{aligned} \varepsilon(\hat{u}_i, \hat{u}_j) &= [1 - \chi^2(\hat{u}_i \cdot \hat{u}_j)^2]^{-1/2} \\ \varepsilon'(\hat{u}_i, \hat{u}_j, \hat{r}) &= \left\{ 1 - \frac{\chi'}{2} \left[\frac{(\hat{u}_i \cdot \hat{r} + \hat{u}_j \cdot \hat{r})^2}{1 + \chi'(\hat{u}_i \cdot \hat{u}_j)} + \frac{(\hat{u}_i \cdot \hat{r} - \hat{u}_j \cdot \hat{r})^2}{1 - \chi'(\hat{u}_i \cdot \hat{u}_j)} \right] \right\} \end{aligned} \quad (4.94)$$

χ' measures the anisotropy of the attractive forces:

$$\chi' = \frac{1 - (\varepsilon_e / \varepsilon_s)^{1/\mu}}{(\varepsilon_e / \varepsilon_s)^{1/\mu} + 1} \quad (4.95)$$

ε_e is the well depth for an end-to-end arrangement of the ellipsoids when the attractive and repulsive contributions cancel, and ε_s is the corresponding well depth for the side-by-side arrangement (Figure 4.45).

The Gay-Berne potential is rather complex but is governed by a relatively small number of parameters, some of which have readily interpretable meanings. The effect of changing the parameters can be most clearly understood by considering certain orientations, such as the side-by-side, end-to-end, crossed and T-shaped structures (Figure 4.45). In the crossed structure the well depth $\varepsilon(\hat{\mathbf{u}}_i, \hat{\mathbf{u}}_j, \hat{\mathbf{r}})$ and the separation $\sigma(\hat{\mathbf{u}}_i, \hat{\mathbf{u}}_j, \hat{\mathbf{r}})$ are independent of χ and χ' . The ratio of the well depths for the end-to-end and side-by-side arrangements is $\varepsilon_e/\varepsilon_s$. The exponents μ and ν are considered adjustable parameters. One way to obtain values for these is to fit the Gay-Berne function to arrangements of Lennard-Jones particles. For example, Luckhurst, Stevens and Phippen determined a value of 1 for ν and a value of 2 for μ by fitting to a linear array of four Lennard-Jones centres [Luckhurst *et al.* 1990].

Depending upon the parameters chosen, simulations performed using the Gay-Berne potential show behaviour typical of liquid crystalline materials. Moreover, by modifying the potential one can determine what contributions affect the liquid crystalline properties and so help to suggest what types of molecule should be made in order to attain certain properties.

4.16 Derivatives of the Molecular Mechanics Energy Function

Many molecular modelling techniques that use force-field models require the derivatives of the energy (i.e. the force) to be calculated with respect to the coordinates. It is preferable that analytical expressions for these derivatives are available because they are more accurate and faster than numerical derivatives. A molecular mechanics energy is usually expressed in terms of a combination of internal coordinates of the system (bonds, angles, torsions, etc.) and interatomic distances (for the non-bonded interactions). The atomic positions in molecular mechanics are invariably expressed in terms of Cartesian coordinates (unlike quantum mechanics, where internal coordinates are often used). The calculation of derivatives with respect to the atomic coordinates usually requires the chain rule to be applied. For example, for an energy function that depends upon the separation between two atoms (such as the Lennard-Jones potential, Coulomb electrostatic interaction or bond-stretching term) we can write:

$$r_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2} \quad (4.96)$$

$$\frac{\partial \nu}{\partial x_i} = \frac{\partial \nu}{\partial r_{ij}} \frac{\partial r_{ij}}{\partial x_i} \quad (4.97)$$

$$\frac{\partial r_{ij}}{\partial x_i} = \frac{(x_i - x_j)}{r_{ij}} \quad (4.98)$$

Thus, for the Lennard-Jones potential:

$$\frac{\partial \nu}{\partial r_{ij}} = \frac{24\varepsilon}{r_{ij}} \left[-2\left(\frac{\sigma}{r_{ij}}\right)^{12} + \left(\frac{\sigma}{r_{ij}}\right)^6 \right] \quad (4.99)$$

The force in the x direction acting on atom i due to its interaction with atom j is given by:

$$\mathbf{f}_{xi} = (\mathbf{x}_i - \mathbf{x}_j) \frac{24\epsilon}{r_{ij}^2} \left[2\left(\frac{\sigma}{r_{ij}}\right)^{12} - \left(\frac{\sigma}{r_{ij}}\right)^6 \right] \quad (4.100)$$

Analytical expressions for the derivatives of the other terms that are commonly found in force fields are also available [Niketic and Rasmussen 1977]. Similar expressions must be derived from scratch when new functional forms are developed.

4.17 Calculating Thermodynamic Properties Using a Force Field

A molecular mechanics program will return an ‘energy value’ for any configuration or conformation of the system. This value is properly described as a ‘steric energy’ and is the energy of the system relative to a zero point that corresponds to a hypothetical molecule in which all of the bond lengths, valence angles, torsions and non-bonded separations are set to their strainless values. It is not necessary to know the actual value of the zero point to calculate the *relative* energies of different configurations or different conformations of the system.

Molecular mechanics can be used to calculate heats of formation. To do so requires the energy to form the bonds in the molecule to be added to the steric energy. These bond energies are typically obtained by fitting to experimentally determined heats of formation and are stored as empirical parameters within the force field. The accuracy with which heats of formation can be predicted with molecular mechanics is, in appropriate cases, comparable with experiment. Thus, the steric energy of a given structure may vary considerably from one force field to another, but its heat of formation should be much closer (if the force fields have been properly parametrised).

A third type of ‘energy’ that can be obtained from a molecular mechanics calculation is the ‘strain energy’. Differences in steric energy are only valid for different conformations or configurations of the same system. Strain energies enable different molecules to be compared. To determine the strain energy it is usual to define some ‘strainless’ reference point. The reference points can be chosen in many ways and so many different definitions of strain energy have been proposed in the literature. For example, Allinger and co-workers defined the reference point using a set of ‘strainless’ compounds such as the all-*trans* conformations of the straight-chain alkanes from methane to hexane. From this set of compounds it was possible to derive a set of strainless energy parameters for constituent parts of the molecules. The inherent strain energy of a hydrocarbon is then obtained by subtracting the reference ‘strainless’ energy from the actual steric energy calculated using the force field. One interesting conclusion of this study was that chair cyclohexane has an inherent strain energy due to the presence of 1,4 van der Waals interactions between the carbon atoms within the ring.

The sources of strain are often quantified by examining the different components (bonds, angles, etc.) of the force field. Such analyses can provide useful information, especially for cases such as highly strained rings. However, in many molecules the strain is distributed

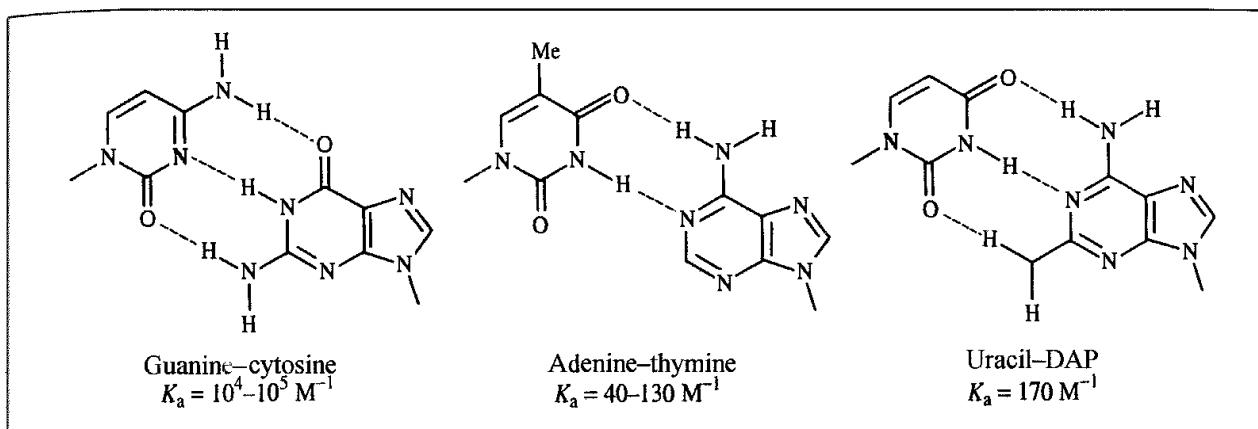


Fig. 4.46 The DNA base pairs guanine (G), cytosine (C), adenine (A) and thymine (T). The uracil-2,6-diaminopyridine pair can also form three hydrogen bonds but has a much lower association constant than G-C

among a variety of internal parameters (and in any case is force-field-dependent). For intermolecular interactions the interpretation can be easier, for the ‘interaction energy’ is simply equal to the difference between the energies of the two isolated species and the energy of the intermolecular complex. A good example of this type of calculation and the conclusions that can be drawn from it is the study by Jorgensen and Pranata [Jorgensen and Pranata 1990] of the interaction between analogues of the DNA base pairs. In the double helical structure of DNA the bases pair up adenine (A) with thymine (T) and guanine (G) with cytosine (C) (Figure 4.46).

The association constant of the G-C base pair in chloroform is between 10^4 M^{-1} and 10^5 M^{-1} whereas the association between the A-T base pair is significantly weaker, at $40\text{--}130 \text{ M}^{-1}$. One obvious reason for this difference is that there are three hydrogen bonds in the G-C base pair and only two in the A-T base pair. However, a simple hydrogen-bond count

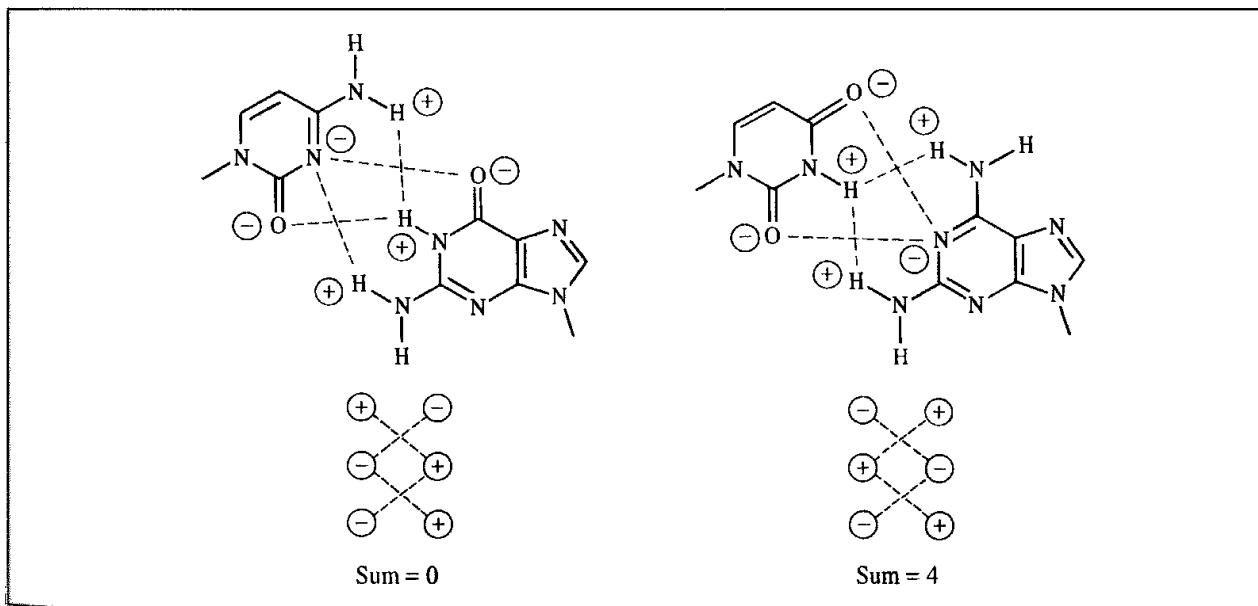


Fig. 4.47: Secondary interactions in guanine-cytosine and uracil-DAP.

does not explain all of the data, for synthetic analogues show a significant variation in their association constants, despite having three hydrogen bonds. The weak binding of the uracil-2,6-diaminopyridine (DAP) system (Figure 4.46) could be considered especially anomalous as it contains the same types of hydrogen bond as in G-C ($\text{NH}_2 \cdots \text{O}$, $\text{NH} \cdots \text{N}$, $\text{NH}_2 \cdots \text{O}$). A qualitative explanation for this phenomenon was proposed by Jorgensen and Pranata who examined the secondary interactions in these complexes. As shown in Figure 4.47, the G-C system contains two unfavourable secondary interactions and two favourable ones, an overall sum of zero. In the uracil-DAP system, all four secondary interactions are unfavourable.

4.18 Force Field Parametrisation

A force field can contain a large number of parameters, even if it is intended for calculations on only a small set of molecules. Parametrisation of a force field is not a trivial task. A significant amount of effort is required to create a new force field entirely from scratch, and even the addition of a few parameters to an existing force field in order to model a new class of molecules can be a complicated and time-consuming procedure. The performance of a force field is often particularly sensitive to just a few of the parameters (usually the non-bonded and torsional terms), so it is often sensible to spend more time optimising these parameters rather than others (such as the bond-stretching and angle-bending terms), the values of which do not greatly affect the results.

The first step is to select the data that are going to be used to guide the parametrisation process. Molecular mechanics force fields may be used to determine a variety of structurally related properties and the parametrisation data should be chosen accordingly. The geometries and relative conformational energies of certain key molecules are usually included in the data set. It is increasingly common to include vibrational frequencies in the parametrisation; these are usually more difficult to reproduce but the incorporation of appropriate cross terms can often help. Some force fields are parametrised to reproduce thermodynamic properties using computer simulation techniques. The OPLS (optimised parameters for liquid simulations [Jorgensen and Tirado-Reeves 1988]) parameters have been obtained in this way.

Unfortunately, experimental data may be non-existent or difficult to obtain for particular classes of molecules. Quantum mechanics calculations are thus increasingly used to provide the data for the parametrisation of molecular mechanics force fields. This is an important development because it greatly extends the range of chemical systems that can be treated using the force-field approach. *Ab initio* calculations are able to reproduce experimental results for small representative systems. Clearly, one should be careful to properly validate a force field derived in such a way by testing against experimental data if at all possible.

Once a functional form for the force field has been chosen and the data to be used in the parametrisation identified, there are then two basic methods that can be used to actually obtain the parameters. The first approach is ‘parametrisation by trial and error’, in which

the parameters are gradually refined to give better and better fits to the data. It is difficult to simultaneously modify a large number of parameters in such a strategy and so it is usual to perform the parametrisation in stages. It is important to remember that there is some coupling between all of the degrees of freedom and so for the most sensitive work none of the parameters can truly be taken in isolation. Parameters for the hard degrees of freedom (bond stretching and angle bending) can, however, often be treated separately from the others (indeed the bond and angle parameters are often transferred from one force field to another without modification). By contrast, the soft degrees of freedom (non-bonded and torsional contributions) are closely coupled and can significantly influence each other. One protocol that can be quite successful is to first establish a series of van der Waals parameters. The electrostatic model is then determined (e.g. by electrostatic potential fitting). Finally, the torsional potentials are determined by ensuring that the torsional barriers are reproduced together with the relative energies of the different conformations. Of course, it may be necessary to modify any of the parameters at any stage should the results be inadequate and so parametrisation is invariably an iterative procedure.

As experimental information on torsional barriers is often sparse or non-existent, quantum mechanical calculations are widely used to determine torsional potentials. The general strategy is as follows. First, a molecular fragment that adequately represents the rotatable bond of interest and its immediate environment is chosen. A series of structures are then generated by rotating about the bond and their energies determined using quantum mechanics. The torsional potential is then fitted to reproduce the energy curve, in conjunction with the van der Waals potential and partial charges. This procedure can be illustrated using the study of Pranata and Jorgensen who wanted to perform some calculations on FK506, a potent immunosuppressant (Figure 4.48) [Pranata and Jorgensen 1991]. FK506 contains a ketoamide functionality that has a *trans* conformation when the molecule is bound to its receptor but which is *cis* in the crystal structure of isolated FK506. NMR experiments suggested that the molecule adopts both *cis* and *trans* conformations in solution. This part of the molecule is clearly implicated in its function and so it was considered important

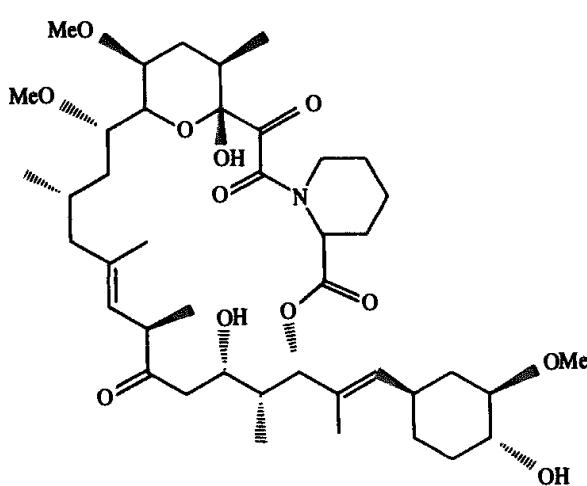


Fig. 4.48 The immunosuppressant FK506

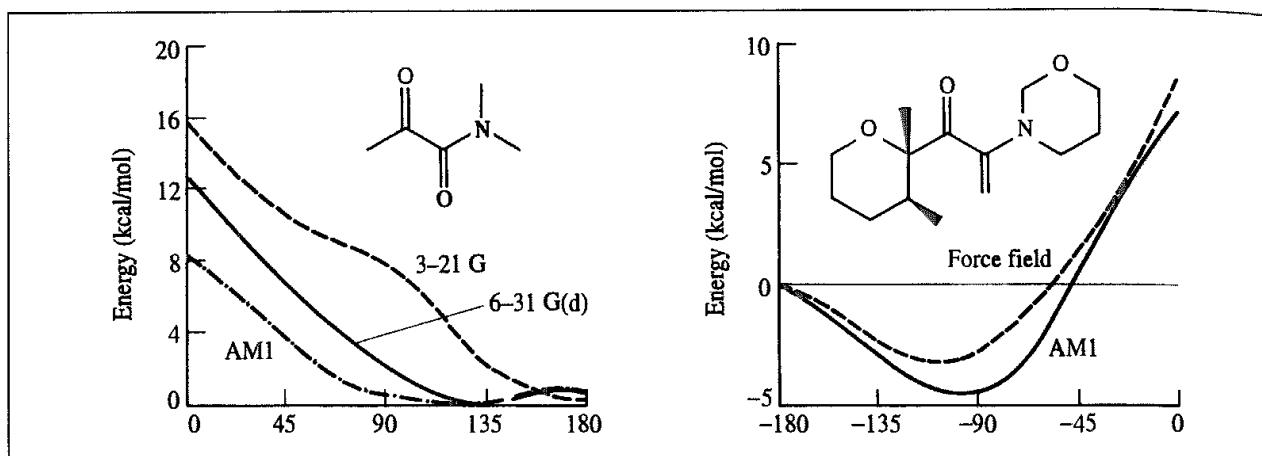


Fig 4.49 Fragments used to derive and evaluate parameters for the ketoamide functionality in FK506 (Figure redrawn from J Pranata and W L Jorgensen 1991 Computational Studies on FK506. Conformational Search and Molecular Dynamics Simulations in Water The Journal of the American Chemical Society 113:9483–9493)

to correctly model the torsional potential about this bond. Pranata and Jorgensen intended to use the AMBER force field for their calculations but the force field contained no parameters for this link.

Molecular orbital calculations were performed on *N,N*-dimethyl- α -ketopropanamide (Figure 4.49, left), which was chosen as an appropriate model system. Semi-empirical calculations using AM1 and *ab initio* calculations using a 6-31G(d) basis set suggested that the minimum energy conformation corresponded to a torsion angle of 124° and 135°, respectively, with the *anti* conformation being slightly higher in energy (~0.7 kcal/mol). However, an analogous calculation using the 3-21G basis set did predict that the *anti* conformation was at a minimum (Figure 4.49). Crystal structures of compounds containing this fragment revealed that an orthogonal structure was commonly encountered. Torsional parameters were then fitted to the 6-31G(d) potential and evaluated by calculating an energetic profile for rotation in a larger fragment of the FK506 molecule using the force field and comparing it with that obtained using AM1 (Figure 4.49, right).

An alternative approach to parametrisation, pioneered by Lifson and co-workers in the development of their ‘consistent’ force fields, is to use least-squares fitting to determine the set of parameters that gives the optimal fit to the data [Lifson and Warshel 1968]. Again, the first step is to choose a set of experimental data that one wishes the force field to reproduce (or calculate using quantum mechanics, if appropriate). Warshel and Lifson used thermodynamic data, equilibrium conformations and vibrational frequencies. The ‘error’ for a given set of parameters equals the sum of squares of the differences between the observed and calculated values for the set of properties. The objective is to change the force field parameters to minimise the error. This is done by assuming that the properties can be related to the force field by a Taylor series expansion:

$$\Delta\mathbf{y}(\mathbf{x} + \delta\mathbf{x}) = \Delta\mathbf{y}(\mathbf{x}) + \mathbf{Z}\delta\mathbf{x} + \dots \quad (4.101)$$

$\Delta\mathbf{y}$ is a vector of the differences between the calculated and experimental data and is a vector whose components are the force field parameters. \mathbf{Z} is a matrix whose elements are the

derivatives of each property with respect to each of the parameters, $\partial x/\partial y$. An iterative procedure is used to minimise the sum of squares of the differences, Δy^2 . The method is easily modified to enable various weighting factors to be assigned to the different pieces of experimental data, so that (for example) the thermodynamic data could be given greater importance than the vibrational frequencies.

A well-known application of the least-squares approach to the optimisation of a force field was performed by Hagler, Huler and Lifson, who derived a force field for peptides by fitting to crystal data of a variety of appropriate compounds [Hagler *et al.* 1977; Hagler and Lifson 1974]. A key result of their work was that no explicit hydrogen bond term was required to model the hydrogen-bonding interactions, but that a combination of appropriate electrostatic and van der Waals models was sufficient. A group led by Hagler more recently developed a force field based upon the results of *ab initio* quantum mechanics calculations on small molecules, again using least-squares fitting [Maple *et al.* 1988]. The quantum mechanics calculations were performed not only on small molecules at equilibrium geometries but also on structures that were distorted from equilibrium. For each geometry the energy was calculated together with the first and second derivatives of the energy. This provided a wealth of data for the subsequent fitting procedure. This research has resulted in many new algorithms for the derivation of force-field parameters and has also challenged some of the assumptions about the development and functional form of force fields. One feature of the resulting force field, named CFF (standing for consistent force field), is that it contains rather more cross terms than other force fields. This can be ascribed to the objective of accurately reproducing vibrational spectra.

4.19 Transferability of Force Field Parameters

The range of systems that have been studied by force field methods is extremely varied. Some force fields have been developed to study just one atomic or molecular species under a wider range of conditions. For example, the chlorine model of Rodger, Stone and Tildesley [Rodger *et al.* 1988] can be used to study the solid, liquid and gaseous phases. This is an anisotropic site model, in which the interaction between a pair of sites on two molecules depends not only upon the separation between the sites (as in an isotropic model such as the Lennard-Jones model) but also upon the orientation of the site-site vector with respect to the bond vectors of the two molecules. The model includes an electrostatic component which contains dipole-dipole, dipole-quadrupole and quadrupole-quadrupole terms, and the van der Waals contribution is modelled using a Buckingham-like function.

Other force fields are designed for use with specific classes of molecules; we have already encountered the AMBER force field, which is designed for calculations on proteins and nucleic acids. Yet other force fields are intended to be applied to a wide range of molecules, and indeed some force fields are designed to model the entire periodic table. Intuitively, one might expect a 'specialised' force field to perform better than a 'general' force field, and while this is certainly true for the best of the specialised force fields, a good general force field can often outperform a poor specific force field.

The ability to transfer parameters from one molecule to another is crucial for any force field. Without it, the task of parametrisation would be impossible, because so many parameters would be required, and the force field would have no predictive ability. Transferability has a number of important consequences for the development and application of force fields. The problem of transferability is often first encountered when a molecular mechanics program fails to run because parameters are missing for the molecule being studied. One must somehow find values for the missing parameters. Some programs automatically 'guess' force field parameters; it is wise to check these assignments as they may be suspect. For the developer of a force field, a compromise must often be found between a complex functional form and a large number of atom types. It is also important to try to ensure that the errors in the force field are balanced, in the sense that it would be silly to spend a lot of time getting (say) the bond-stretching terms just right, if the van der Waals parameters give rise to large errors.

An alternative to 'guessing' parameters (which, if done properly, can sometimes give quite reasonable results) is to construct the force field in such a way that the parameters can be derived from atomic properties. This is particularly pertinent to those force fields which are designed to be used on a very wide range of elements and atom types, such as the Universal Force Field [Rappé *et al.* 1992]. This force field is claimed to model the entire periodic table and as such it would probably be impossible to derive individual parameters for each of the terms; indeed, the data required for such an exercise does not exist for many cases. Thus the UFF has a set of atom types which are characterised by atomic number, hybridisation and formal oxidation state. Reference bond lengths are initially set equal to the sum of the two relevant atomic bond radii and then corrected for bond order and the relative electronegativities of the two atoms. Bond force constants are obtained from Badger's rules, under which the force constant is proportional to the product of the 'effective atomic charges' for the two atoms and inversely proportional to the cube of the interatomic distance:

$$k_{ij} \propto \frac{q_i^* q_j^*}{r_{ij}^3} \quad (4.102)$$

The effective atomic charges are either obtained by fitting to data on diatomic molecules (where it exists) or by interpolation or extrapolation from this fit.

Transferability can be helped by using the same parameters for as wide a range of situations as possible. The non-bonded terms are particularly problematic in this regard; it would, in principle, be necessary to have parameters for the non-bonded interactions between all possible pairs of atom types. This would give rise to a very large number of parameters. It is therefore commonly assumed that the same set of van der Waals parameters can be used for most, if not all, atoms of the same element. For example, all carbon atoms (sp^3 , sp^2 , sp , etc.) would be treated with the same set of van der Waals parameters, all nitrogens by a common set, and so on. The torsional terms may also be generalised, so that the torsional parameters depend solely upon the atom types of the two atoms that form the central bond, rather than on all four atoms that comprise the torsion angle, as described in Section 4.5 for the AMBER force field.

4.20 The Treatment of Delocalised π Systems

The bonds in conjugated π systems are often of different lengths. For example, the central bond in butadiene is approximately 1.47 Å long, but the two terminal CH=CH₂ bonds are approximately 1.34 Å. If butadiene is modelled using a force field in which all four carbon atoms are assigned the same atom type (e.g. 'carbon sp²') then each bond will be assigned the same bonding parameters and in the equilibrium structure all carbon–carbon bonds will be almost identical in length. A similar situation arises for aromatic systems. For example, not all the bonds in naphthalene are of equal length (unlike benzene). The bond lengths in a delocalised π system depend upon the bond orders; the higher the bond order, the shorter the bond.

In some cases it may be possible to circumvent this problem by creating a model specific to the conjugated system. For butadiene the central carbon–carbon bond of the π system could be treated in a different manner to the two terminal bonds, for example by using one atom type for the –CH= carbon atoms and one for the =CH₂ carbon atoms in butadiene. This approach might be acceptable if we wanted to perform an extensive series of calculations on substituted butadienes, but it does compromise the transferability of the force field parameters. An alternative is to incorporate a molecular orbital calculation into the force field. Two variants on this theme have been developed. In one approach, the π and σ systems are treated separately [Warshel and Karplus 1972; Warshel and Lappicirella 1981]. For a given geometry, a self-consistent field quantum mechanical calculation is performed on the π system, typically with an appropriate semi-empirical theory. Molecular mechanics is simultaneously applied to the σ system. The energies of the quantum mechanical and molecular mechanical calculations are added together, and the geometry is modified to minimise this combined energy. A obvious assumption inherent in this approach is that the π and σ systems can be separated, which may be difficult to justify when deviations from planarity are present. Nevertheless, the approach has been extended to include those containing conjugated nitrogen and oxygen atoms, which has enabled the study of the properties of not only the ground states of some important biological chromophores (such as porphyrins) but also their excited states [Warshel and Lappicirella 1981].

An alternative approach is exemplified by the MM2/MM3/MM4 family of programs. First, a molecular orbital calculation is performed on the π system. If the initial conformation of the system is non-planar the calculation is performed on the equivalent planar system. The force field parameters are then modified according to the quantum mechanical bond orders. In MMP2 (the name given to the special version of MM2 which incorporated these features) these parameters are the force constant for the bonds in the π system, the reference bond lengths and the torsional barriers [Sprague *et al* 1987; Allinger and Sprague 1973]. The system is then subjected to the usual molecular mechanics treatment using the new force field parameters. A linear relationship between the stretching constants and the bond orders, and between the reference bond lengths and the bond orders was found to give good results. Initially, the torsional barriers were assumed to be proportional to the square of the bond orders, but this relationship was modified slightly in subsequent versions

of the program. Thus in MM4 the V_2 and V_3 terms become:

$$V_2 = [A + p_{ij}^{\omega=0} \beta_{ij}] V_2^0 \quad (4.103)$$

$$V_3 = K_{V_3} [1 - p_{ij}(\omega)] V_3^0 \quad (4.104)$$

In Equation (4.103) p_{ij} is the bond order about the central bond $i-j$ of the torsion angle calculated for a torsion angle of zero and β_{ij} is the resonance integral from the molecular orbital calculation. The parameter A has a value of -0.09 and so the V_2 term is lower for those conjugated bonds with a lower bond order. In Equation (4.104) p_{ij} is now the bond order for the bond $i-j$ calculated for the torsion angle ω . K_{V_3} equals 1.25 and so V_3 increases with decreasing bond order. A bond with a lower bond order (and so a lower V_2 and a higher V_3) is thus more likely to deviate from planarity.

4.21 Force Fields for Inorganic Molecules

It may come as a surprise to many readers to learn that the earliest force field calculations on inorganic molecules were reported at much the same time as the first calculations on organic systems. For example, Corey and Bailar described the use of empirical force field calculations on octahedral complexes of cobalt in 1959 [Corey and Bailar 1959]. The range of metal-containing systems that can be considered by force field methods has steadily expanded since then. Moreover, many systems of commercial interest contain metals or other elements not usually found in 'organic' or 'biochemical' systems.

Some inorganic systems (such as certain coordination complexes) are little different to organic systems from a force field point of view; the bonding can be represented in a similar way and many of the force field parameters originally developed for organic systems can be transferred without modification. However, inorganic molecules do have certain properties which makes them more difficult to model than their organic counterparts. Perhaps the two most striking properties are the much wider range of geometries and the presence of highly delocalised bonds. Thus inorganic molecules include square planar and sawhorse (e.g. SF_4) shapes for four coordination and T-shaped for three coordination. Coordination numbers higher than four are also possible, with five (square pyramidal, trigonal bipyramidal) and six (octahedral and trigonal prismatic) being particularly common. To model such systems using conventional organic force fields would often be problematic because their geometries do not have a high degree of symmetry. For example, in a trigonal bipyramidal there are in principle three different types of bond angle subtended at the central atom (90° , 120° and 180°). Moreover, in such systems the atoms are often equivalent (interchanging them gives the same structure back). However, if these atoms are assigned different force field parameters then this equivalence is not reproduced by the calculation. At least in these cases there is an obvious localised bonding scheme that can be applied; this is often not possible with organometallic molecules. For example, how should the bonding in ferrocene be represented in a force field calculation? Is there a bond between the iron and each of the carbon atoms in the two cyclopentadienyl rings? Is there a 'bond' from the iron to the centre of each of the rings? A yet further complication is that significant deviations from ideal geometries are often observed due to electronic effects such as the Jahn-Teller effect.

Whilst there is no universal solution to these problems within the context of a single force field similar to those used in organic chemistry, for certain situations it is possible to use an organic-like force field with only relatively small modifications. For obvious reasons those complexes with a high degree of symmetry are most amenable to such a treatment. Thus octahedral and square planar complexes are the simplest to model because of their symmetry (in addition to the geometries common in organic chemistry). However, even these have two types of equilibrium angle (180° and 90°). The situation can be much more complicated for the other geometries or for structures where the geometry about the metal is a distortion of a regular arrangement. A Urey-Bradley treatment of the bonding about the metal can often be quite successful in achieving the correct geometries. Here, there are no angle-bending terms at the metal but terms due to pairs of atoms bonded to the metal.

It is much more difficult to use such a force field to model metal π systems, where the bonding between the metal and the ligand is not easily represented by a conventional bonding picture. As we have discussed, metal atoms can adopt a wide range of geometries in π complexes, which are often significantly distorted from regular structures. Nevertheless, force fields have been developed which can cope with such systems, as well as being able to model more traditional systems such as organic compounds. These force fields often use a rather different functional form from Equation (4.1) and the parameters are obtained in a different way. One distinctive feature of both the Universal Force Field and the SHAPES force field developed by Landis and co-workers [Allured *et al.* 1991; Cleveland and Landis 1996] is the way in which angle bending is treated. The harmonic potential that is commonly employed in standard force fields is inappropriate to model the distortion of systems as the angle approaches 180° . UFF [Rappé *et al.* 1993] uses a cosine Fourier series for each angle ABC:

$$\nu(\theta) = K_{ABC} \sum_{n=0}^m C_n \cos n\theta \quad (4.105)$$

The coefficients C_n are chosen to ensure that the function has a minimum at the appropriate reference bond angle. For linear, trigonal, square planar and octahedral coordination, Fourier series with just two terms are used with a C_0 term and a term for $n = 1, 2, 3$ or 4, respectively:

$$\nu(\theta) = K_{ABC}[1 - \cos(n\theta)] \quad (4.106)$$

Thus, for example, if $n = 4$ then the function has minima at both 90° and 180° as required for octahedral geometries. The general case is exemplified by the H-O-H angle in water, where it is desired to have a minimum in the energy at an angle of 104.5° . Moreover, at this angle (θ_0) the second derivative of the energy equals the force constant. If in addition it is required that the energy is a maximum at 180° the following expression results:

$$\nu(\theta) = K_{ABC}[C_0 + C_1 \cos(\theta) + C_2(\cos 2\theta)] \quad (4.107)$$

The three coefficients are defined as:

$$C_2 = \frac{1}{4 \sin^2(\theta_0)}; \quad C_1 = -4C_2 \cos(\theta_0); \quad C_0 = C_2[2 \cos^2(\theta_0) + 1] \quad (4.108)$$

The SHAPES angle-bending term is very similar:

$$\nu(\theta) = K_{ABC} \sum_{n=0}^m [1 + \cos(n\theta - \delta)] \quad (4.109)$$

δ is the phase shift. Landis subsequently developed a formulation (called VALBOND) for the angle-bending term that is based on valence bond theory and which can produce results that compare well with *ab initio* calculations [Landis *et al.* 1995, 1998]. For example, using just one set of C–H parameters the H–C–H bond angles in ethene, formaldehyde and both singlet and triplet carbene match closely those found experimentally. One key practical advantage of this method is that it is not necessary to define equilibrium bond angles.

4.22 Force Fields for Solid-state Systems

Empirical potential models are widely used to study the solid state, complementing the quantum mechanical approaches we discussed in Chapter 3. One important difference between solid-state materials and ‘organic’ molecules (and indeed, some inorganic complexes) is that whilst the latter can generally be described using a localised bond model this is not always the case for the former. As a consequence, molecular mechanics approaches of the kind we have discussed so far in this chapter can be applied successfully only to certain types of material. Ionic and metallic systems especially require an alternative approach. Perhaps the key difference between solid-state materials and isolated molecules is the way in which the electrostatic terms are considered. As we shall see in Sections 6.7 and 6.8 it is common to truncate such interactions at some cutoff distance. However, solid-state modelling is concerned with materials that have long-range order; moreover, they often contain highly charged species. This means that the use of cutoffs can have a particularly detrimental effect, necessitating the use of special techniques such as the Ewald summation that enable more accurate interaction energies to be calculated. First, however, we shall consider the treatment of covalent systems which are amenable to the ‘organic’ style of molecular mechanics force field treatment, as exemplified by the study of zeolites.

4.22.1 Covalent Solids: Zeolites

Zeolites are materials generally composed of silicon, aluminium, oxygen and a metal cation or proton. They have a multitude of commercial uses including catalysis and separation (e.g. they are used in oil refining to separate linear and branched alkanes). Many of these important properties are a consequence of the presence within the zeolite of channels of molecular dimensions. It is therefore natural that molecular modelling techniques should be used to investigate the intrinsic properties of such materials and the way in which they interact with adsorbates.

The size of many zeolite systems means that considerable computational resources may be required for the calculation. In some cases therefore, such as the study of adsorption

processes, the zeolite is kept rigid and attention is concentrated on the intermolecular interactions between the zeolite and the adsorbate. This is often done using a combination of van der Waals and electrostatic terms; a Lennard-Jones potential may be used for the van der Waals component, but a Buckingham-like potential is often preferred. Electrostatic interactions can be very important for zeolites. However, the partial charges used in the various published force fields can vary enormously (from $0.4e$ to as much as $1.9e$ for the silicon atoms in silicates).

It is obviously an approximation to keep the zeolite rigid, and in more complex models the structure can vary. Many of the force fields that have been developed to model zeolites are very similar to the valence force fields used for organic and biological molecules, typically containing bond-stretching, angle-bending and torsional terms in addition to the non-bonded interactions. One important consideration when modelling zeolites is that very little energy is required to deform the Si–O–Si bond over an extremely wide range (at least 120° to 180°). This is shown in Figure 4.50, which shows the results of *ab initio* calculations using a 3-21G* basis set for $\text{H}_3\text{SiOSiH}_3$. The Fourier series expansions used by the UFF and SHAPES force fields for the angle-bending terms are designed to cope with such angular variation; Nicholas, Hopfinger, Trouw and Iton suggested the following quartic potential as an alternative specifically for the Si–O–Si angle [Nicholas *et al.* 1991]:

$$\nu(\theta) = \frac{k_1}{2}(\theta - \theta_0)^2 + \frac{k_2}{2}(\theta - \theta_0)^3 + \frac{k_3}{2}(\theta - \theta_0)^4 \quad (4.110)$$

With the correct choice of the parameters k_i and θ_0 the *ab initio* data in Figure 4.50 could be reproduced very well. In this force field a Urey-Bradley term was also included between the silicon atoms in such angles to model the lengthening of the Si–O bond as the angle decreased.

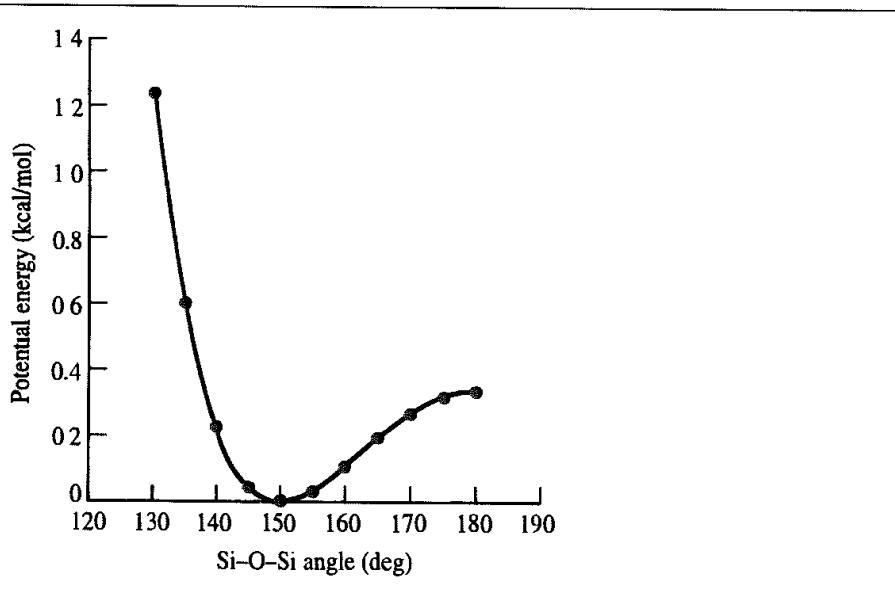


Fig 4.50. Variation in energy with the Si–O–Si angle (Figure redrawn from Grigoras S and T H Lane 1988 Molecular Parameters for Organosilicon Compounds Calculated from Ab Initio Computations Journal of Computational Chemistry 9,25–39.)

4.22.2 Ionic Solids

The covalent approach is rarely appropriate for ionic and polar solids such as oxides and halides. The usual starting point for studying such systems is to write the potential as a series expansion of pairwise, three-body, etc., terms:

$$\mathcal{V} = \mathcal{V}_0 + \sum_{i=1}^N \sum_{j=i+1}^N v_{ij}(r) + \sum_{i=1}^N \sum_{j=i+1}^N \sum_{k=j+1}^N v_{ijk}(r) + \dots \quad (4.111)$$

One of the oldest of such models is due to Born [Born 1920], who restricted the series to pairwise terms, which were in turn divided into long-range Coulomb interactions and short-range repulsive forces. If an inverse power law is used for the repulsive term the potential energy is thus:

$$\mathcal{V} = \sum_{i=1}^N \sum_{j=i+1}^N \left(\frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} + \frac{A}{r_{ij}^n} \right) \quad (4.112)$$

The simplest way to apply such an equation is to assume that the charges q are equal to the oxidation states of the relevant species and that the repulsive potential only acts between nearest neighbours (though in common with many solid-state calculations the long-range ionic interaction is generally calculated for all possible interactions using an approach such as the Ewald sum, Section 6.8). This only leaves the two parameters A and n whose determination in principle requires only two pieces of experimental data (though the values obtained may vary quite considerably depending upon which data is chosen). An obvious extension of the simple form of Equation (4.112) is to model the short-range interactions by an alternative functional form; the Buckingham potential is commonly employed.

For a simple material such as sodium chloride the oxidation state assumption is a reasonable one. However, for other systems this is not necessarily the case. Various methods have been proposed for determining appropriate sets of non-integral charges. One strategy is to examine the distribution of charge within the material, as can be obtained from high-resolution X-ray experiments. However, there is no unique way to partition the charge unless there is zero bonding overlap between the ions. The atoms-in-molecules approach (see Section 2.7.7) may be a good way to do this but this is not the only option. It is worth mentioning that one advantage of the formal charge approach is that it can facilitate the transferability of potentials from one material to another whilst still maintaining charge neutrality.

The Born model with integral or partial charges assumes that the ions have zero polarisability. This is reasonable for small cations such as Li^+ or Mg^{2+} but can introduce significant errors for other systems. One property that clearly demonstrates this is the high-frequency dielectric constant. At a suitably high frequency only the electrons can keep up with the external field and the dielectric constant is given by the Clausius–Mosotti relationship:

$$\frac{(\epsilon_r - 1)}{(\epsilon_r + 2)} = \frac{4\pi}{3V_m} \sum_{i=1}^N \alpha_i \quad (4.113)$$

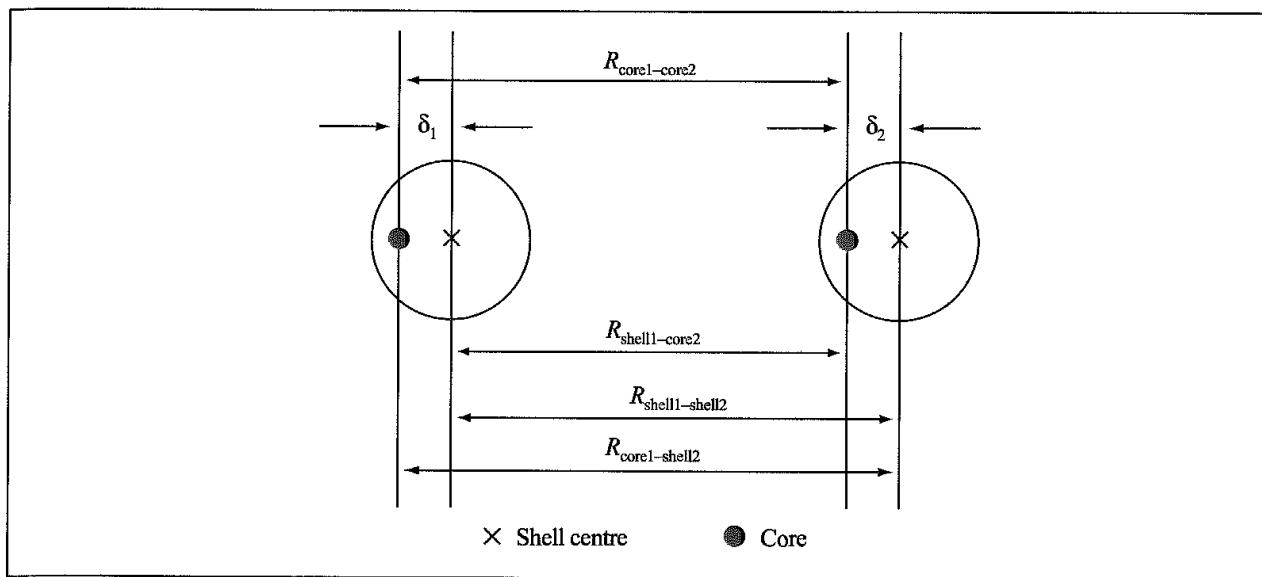


Fig 4.51 The Dick-Overhauser shell model

ϵ_r is the relative permittivity, V_m is the molar volume and α_i is the polarisability of the i th ion with the sum being over the N ions. If the ions were not polarisable then ϵ_r would have a value of 1. As we have seen, one way to incorporate polarisation is to assign a point polarisability to each ion. However, this model does not often give good results, at least for certain properties. This is because it fails to account for the coupling between polarisation and short-range repulsion effects. Thus polarisation causes distortions in the distribution of the valence electrons, and short-range repulsion is itself a consequence of the overlap between such electrons. The overall effect of short-range repulsion is to reduce polarisation effects. One model that can take this coupling into account is the shell model of Dick and Overhauser [Dick and Overhauser 1958] (Figure 4.51). In this model the ion is represented by a massive core linked to a massless shell by a harmonic spring. Both the core and the shell have charges associated with them. In an electric field the shell retains its charge but moves with respect to the core. The polarisability of an isolated ion in this model is proportional to Y^2/k where k is the spring constant of the harmonic spring and Y is the charge on the shell. The electrostatic interaction energy equals the sum over all ions and shells, not counting any interaction between an ion and its own shell. Although it is appealing to assume that the shells somehow play the role of the valence electrons this is probably an over-interpretation if only due to the fact that the shell charges, Y , do not necessarily assume small negative values.

Three-body and higher terms are sometimes incorporated into solid-state potentials. The Axilrod-Teller term is the most obvious way to achieve this. For systems such as the alkali halides this makes a small contribution to the total energy. Other approaches involve the use of terms equivalent to the harmonic angle-bending terms in valence force fields; these have the advantage of simplicity but, as we have already discussed, are only really appropriate for small deviations from the equilibrium bond angle. Nevertheless, it can make a significant difference to the quality of the results in some cases.

As for molecular systems, the parameters used to study the solid state can be derived using both experimental and theoretical data. There is a long tradition of using quantum mechanical calculations to extract such potentials. Whereas it is now common for the sophisticated Hartree–Fock and density functional theory approaches to be used for such parameter derivations, an approach called electron gas theory (a crude version of density functional theory) played a significant historical role and is still used [Allan and Mackrodt 1994]. One example of the way in which *ab initio* quantum mechanical calculations can play a role in this process is provided by the derivation of a potential model for $\alpha\text{-Al}_2\text{O}_3$ [Gale *et al.* 1992]. Previous attempts to derive empirical potentials for this material (using a shell model combined with a Buckingham potential) were not entirely successful; in particular these did not correctly predict that the corundum structure should have the lowest energy. One interesting feature of these earlier parameterisations was the great variation in the core and shell charges; for example, in one of the models the aluminium core and shell charges were 1.617 and 1.383 respectively; in another they were 10.6063 and –8.0563. A feature of the periodic Hartree–Fock calculations (see Section 3.8.3) was the use of distorted structures to provide more information on the nature of the energy surface, which was found to give better results.

4.23 Empirical Potentials for Metals and Semiconductors

Perhaps the most important consideration when discussing the development and use of empirical potentials for studying atomic solids is that pairwise potential models are often not very suitable. The performance of pairwise potential models can be bad for transition metals and even worse for semiconductors! There are a number of reasons why this is so, many of which are due to the fundamental behaviour of pairwise potentials for certain experimental properties. The most oft-quoted properties are as follows:

1. The ratio between the cohesive energy and the melting temperature, $E_c/k_B T$. The cohesive energy is the energy cost of removing an atom from within the solid matrix. This ratio is observed to be approximately 30 in metals but about 10 in pairwise systems.
2. The ratio between the vacancy formation energy and the cohesive energy, E_v/E_c . This ratio is between $\frac{1}{4}$ and $\frac{1}{3}$ in metals but closer to unity in two-body systems (exactly 1 if the structure is not permitted to relax). This can be understood as follows. Suppose each atom in a solid has Z neighbours. If one of the atoms is removed then the coordination of the surrounding Z atoms will fall to $Z - 1$. Using a pairwise energy model the vacancy formation energy is thus Z times the atom–atom bond energy. The cohesive energy is the energy to reduce the coordination of an atom from Z to zero and so would also equal Z times the atom–atom bond energy. The energy change for both of these processes is thus equal for the pairwise model.
3. The ratio between the elastic constants C_{12}/C_{44} . Elastic constants will be discussed in Section 5.10; for a cubic solid there are three distinct values, which are labelled C_{11} , C_{12} and C_{44} . For a two-body system the ratio is exactly 1 (this is known as the Cauchy relationship). For metals and oxides deviation from unity is common, gold has a particularly high value, which is indicative of its high malleability.

4. The surface properties of metals are such that the surface tends to relax inwards but systems described by two-body interactions tend to relax outwards.

The main reason for the failure of pairwise potentials is that they are unable to deal simultaneously with both surface and bulk environments. Thus on the surface there are generally fewer bonds, but these tend to be stronger than in the bulk, where there are more, but weaker, bonds. Several many-body potentials have been devised to try to address this problem. Many of these potentials have a similar, sometimes mathematically equivalent, functional form. This reflects their common origins in some form of quantum mechanical description of bonding. However, they differ in their underlying approach, the degree to which they conform to these quantum mechanical origins and the way in which they are parametrised. Here we will outline various models: the Finnis-Sinclair model (and the Sutton-Chen extension), the embedded-atom model, the Stillinger-Weber model and the Tersoff model.

The origins of the Finnis-Sinclair potential [Finnis and Sinclair 1984] lie in the density of states and the *moments theorem*. Recall that the density of states $D(E)$ (see Section 3.8.5) describes the distribution of electronic states in the system. $D(E)$ gives the number of states between E and $E + \delta E$. Such a distribution can be described in terms of its *moments*. The moments are usually defined relative to the energy of the atomic orbital from which the molecular orbitals are formed. The m th moment, μ^m , is given by:

$$\mu^m = \sum_n (E - E_{\text{atomic}})^m D(E) \quad (4.114)$$

The summation runs over the molecular orbitals or bonds. The first moment is the mean of the distribution. If the moments are defined relative to the atomic orbital energy then this first moment will be zero. The second moment (the sum of the squares of the deviations) is the width of the distribution (the variance). The third moment describes how skewed the distribution is about the mean. If all the moments are known then the distribution can be completely characterised. Of these various moments one would expect the second to be most related to the binding energy, as this indicates how much the energy levels in the solid differ from those in the atom. Indeed, a high correlation is found to exist between the binding energy and the square root of the second moment. Armed with this relationship it would be possible to predict the binding energy for perfect lattices where the atomic environments were identical. However, a more useful model is one based on a local atomic environment ('real' materials contain features such as surfaces and defects). This requires a local density of states to be defined for each atom, $d_i(E)$, where the contribution of each molecular orbital is weighted by the amount of the orbital on the atom. In a linear combination of atomic orbitals (LCAO) model this weight is the sum of the squares of the basis set coefficients for those atomic orbitals centred on the atom. The global density of states is equal to the sum of the local densities of states over all atoms and the electronic binding energy for each atom equals the integral of $d_i(E)E$:

$$E_i^{\text{el}} = \int d_i(E) E dE \quad (4.115)$$

Thus, if we knew the second moment of the local density of states we should be able to determine the atomic binding energy via the square root relationship. However, as quantum

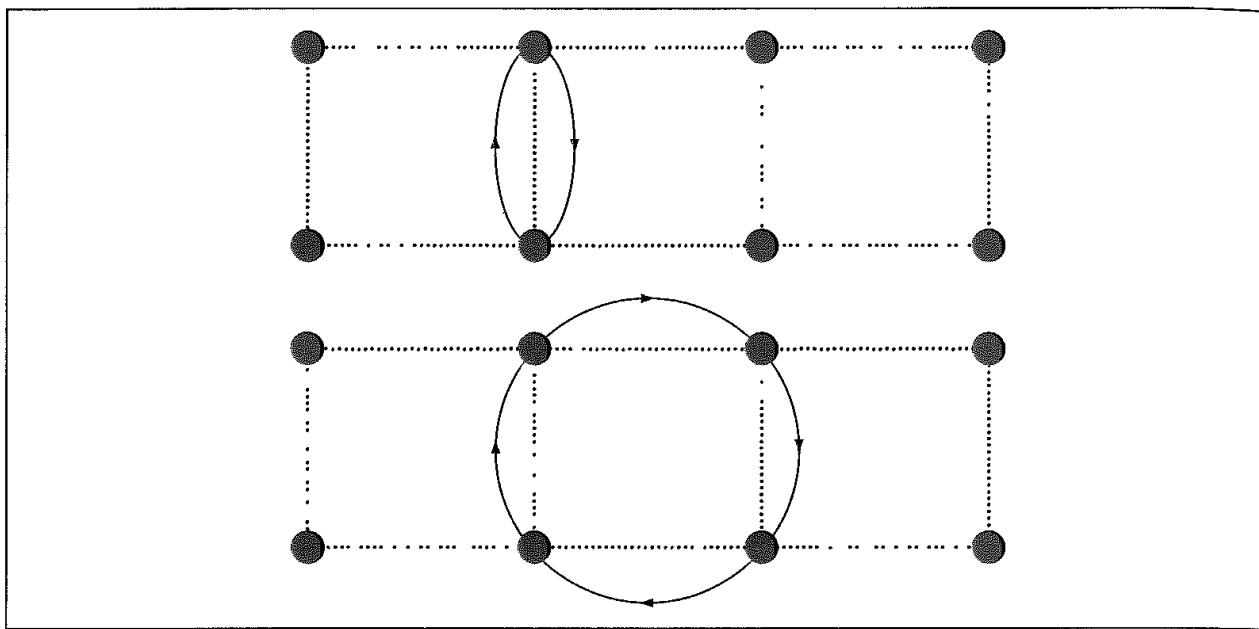


Fig 4.52. Calculating paths using the moments theorem. Illustrated are paths of lengths 2 and 4

mechanics is the only way we currently know of to determine the density of states, this might seem rather self-defeating. This is the role of the *moments theorem*, which relates the bonding topology to the moments of the local density of states without requiring an explicit calculation of the electronic energy levels.

The moments theorem states that the m th moment of the local density of states on an atom i is determined by the sum of all paths of length m over neighbouring atoms that start and end at i . For the second moment these paths involve just two ‘hops’, from the atom in question to a neighbour and back again (Figure 4.52). For the higher moments, the number of possible paths increases dramatically and becomes a challenging calculation. However, for the second moment the number of paths of length 2 is simply equal to the number of nearest neighbours, Z . Consequently, the local electronic binding energy for each atom is approximately equal to the square root of the number of neighbours. This is the *second-moment approximation*:

$$E_i^{\text{el}} \propto \sqrt{Z_i} \quad (4.116)$$

As an aside, we can easily show how this satisfies the ratio E_v/E_c (property 2, page 240) The energy E_v associated with Z atoms having their coordination reduced from Z to $Z - 1$ will be $Z[\sqrt{Z} - \sqrt{Z - 1}]$. The cohesive energy E_c is proportional to \sqrt{Z} For typical values of Z this gives E_v/E_c as approximately $\frac{1}{2}$.

In the Finnis–Sinclair potential a pairwise contribution is added to the many-body term to give the following form:

$$\mathcal{V} = \sum_{i=1}^N \sum_{j=i+1}^N P(r_{ij}) + \sum_{i=1}^N A\sqrt{\rho_i} \quad (4.117)$$

$P(r_{ij})$ is the pairwise potential, which, depending upon the model, can be considered to include electrostatic and repulsive contributions. The second term is a function of the electron density, ρ_i , and varies with the square root, in keeping with the second-moment approximation. The electron density for an atom includes contributions from the neighbouring atoms as follows:

$$\rho_i = \sum_{j=1, j \neq i}^N \phi_{ij}(r_{ij}) \quad (4.118)$$

$\phi_{ij}(r_{ij})$ is a short-range, decreasing function of the distance between the two atoms i and j . In the original Finnis-Sinclair model the function $\phi_{ij}(r_{ij})$ was written as a parabolic function of the interatomic distance, $(r_{ij} - r_c)^2$, where r_c is a cutoff distance chosen to lie between the second and third neighbouring shells. ϕ_{ij} is zero beyond this cutoff distance. The pairwise potential was expressed as a quartic polynomial up to some cutoff and zero beyond.

The Finnis-Sinclair potential can be written in a more general form by replacing the number of neighbouring atoms by an exponential function of the distance between atoms. This is necessary because the number of neighbours is not always straightforward to define, especially in disordered systems and near defects. An exponential function also reflects the fact that electron densities decay exponentially from the nucleus. Moreover, the pairwise potential can also be written as an exponential function of distance to give the following general equation:

$$\mathcal{V} = \sum_{i=1}^N \left\{ \sum_{j=1, j \neq i}^N A e^{-\alpha r_{ij}} - B \left[\sum_{j=1, j \neq i}^N e^{-\beta r_{ij}} \right]^{1/2} \right\} \quad (4.119)$$

Sutton and Chen extended the potential to longer range to enable the study of certain problems such as the interactions between clusters of atoms [Sutton and Chen 1990]. Their objective was to combine the superior Finnis-Sinclair description of short-range interactions with a van der Waals tail to model the long-range interactions. The form of the Sutton-Chen potential is:

$$\mathcal{V} = \varepsilon \left\{ \sum_{i=1}^N \sum_{j=i+1}^N \left(\frac{a}{r_{ij}} \right)^n - c \sum_{i=1}^N \left[\sum_{j=1, j \neq i}^N \left(\frac{a}{r_{ij}} \right)^m \right]^{1/2} \right\} \quad (4.120)$$

In this equation, ε and a are parameters with dimensions of energy and length respectively, c is a dimensionless (positive) parameter, and m and n are integers such that n is greater than m . The use of power-law relationships in the Sutton-Chen potential has a number of useful consequences, analogous to the scaling properties of the Lennard-Jones potential. For example, for a given crystal structure (e.g. hexagonal close-packed, face-centred cubic, body-centred cubic, etc.) the value of c is fixed. Moreover, if two metals are described by the same values of m and n then the results for one system may be converted directly to the other by rescaling the energy and length parameters ε and a . Typical values for m are between 6 and 8 and for n between 9 and 12.

The embedded-atom method [Daw and Baskes 1984] is an empirical embodiment of a simplified quantum mechanical model for bonding in solids called *effective medium*

theory. The key feature of effective medium theory is the replacement of the complex environment around each atom by a simplified model known as jellium. The jellium environment corresponds to a homogeneous electron gas with a positive background. Each atom is considered to be surrounded by a sphere with a radius such that the electronic charge within each sphere due to the background jellium is equal and opposite to the charge on the atom. In the embedded-atom method the background electron density is replaced by a sum of electron densities from the neighbouring atoms. The many-body term is known as an *embedding function*; this gives the energy of each atom as a function of the electron density, ρ_i . In the embedded-atom method the electron density ρ_i equals the sum of the electron densities ϕ_{ij} from neighbouring atoms (Equation (4.118)). In the Daw and Baskes model a Coulomb potential was used for the pairwise potential but with an effective charge $Z(r)$ that decreases gradually with internuclear distance. The embedding function was represented with a cubic spline equation that has a single minimum and goes to zero at vanishing density. The densities were obtained from quantum mechanical calculations.

Both the Finnis-Sinclair and the embedded-atom potentials (together with others that we have not considered here) can be represented using a very similar functional form. However, it is important to realise that they differ in the way that they connect to the first-principles, quantum mechanical model of bonding. They also differ in the procedures used to parametrise the models, so that different parametrisations may be reported for the same material.

The construction of empirical potentials for semiconductors is considered to be an even greater challenge than for metals. In our earlier discussion of the use of density functional methods to determine the electronic structure of the group 14 elements carbon, silicon and germanium we referred to the fact that, whilst the most stable form of silicon is the diamond structure, as pressure is applied so new structures can be obtained. That such a variety of structures can be achieved indicates that they are rather close in energy. Another interesting property of silicon is that in the liquid form it is a metal and the liquid is more dense than the solid. Two of the potentials that have been applied to these systems are the Stillinger-Weber and the Tersoff potentials. The Stillinger-Weber potential [Stillinger and Weber 1985] uses a two-body and three-body term:

$$\mathcal{V} = \sum_{i=1}^N \sum_{j=i+1}^N f_2(r_{ij}) + \sum_{i=1}^N \sum_{j=i+1}^N \sum_{k=j+1}^N [h(r_{ij}, r_{ik}, \theta_{jik}) + h(r_{ji}, r_{jk}, \theta_{ijk}) + h(r_{ki}, r_{kj}, \theta_{ikj})] \quad (4.121)$$

$$f_2(r_{ij}) = A(Br_{ij}^{-p} - r_{ij}^{-q}) \exp[(r_i - a)^{-1}] \quad (4.122)$$

$$h(r_{ij}, r_{ik}, \theta_{jik}) = \lambda \exp[\gamma(r_{ij} - a)^{-1} + \gamma(r_{ik} - a)^{-1}] (\cos \theta_{jik} + \frac{1}{3})^2 \quad (4.123)$$

These equations all use distances and energies in reduced units and the functional form is designed to go to zero without discontinuities at the cutoff distance $r = a$. There are seven parameters ($A, B, p, q, a, \lambda, \gamma$), which were determined by a search procedure, with care being taken to ensure that the diamond structure was the most stable periodic arrangement and that the melting point and liquid structure (as determined by molecular dynamics simulations) were in reasonable agreement with experiment. The three-body term is

designed to favour the tetrahedral geometry found in the diamond structure, which is why it works reasonably well for this form of crystalline silicon. However, it does not perform so well for the other solid forms, which have a different atomic geometry, or for other properties such as the liquid structure.

The Tersoff potential [Tersoff 1988] is based on a model known as the *empirical bond-order potential*. This potential can be written in a form very similar to the Finnis–Sinclair potential:

$$\mathcal{V} = \sum_{i=1}^N \left\{ \sum_{j=1, j \neq i}^N A e^{\alpha r_{ij}} - b_{ij} B e^{-\beta r_{ij}} \right\} \quad (4.124)$$

The key term is b_{ij} , which is the bond order between the atoms i and j . This parameter depends upon the number of bonds to the atom i ; the strength of the ‘bond’ between i and j decreases as the number of bonds to the atom i increases. The original bond-order potential [Abell 1985] is mathematically equivalent to the Finnis–Sinclair model if the bond order b_{ij} is given by:

$$b_{ij} = \left(1 + \sum_{k=1; k \neq i, k \neq j}^N e^{-\beta(r_{ik} - r_{ij})} \right)^{-1/2} \quad (4.125)$$

It can be readily confirmed that b_{ij} decreases as the number of bonds N increases and/or their length (r_{ik}) decreases. This relationship between the bond strength and the number of neighbours provides a useful way to rationalise the structure of solids. Thus the high coordination of metals suggests that it is more effective for them to form more bonds, even though each individual bond is weakened as a consequence. Materials such as silicon achieve the balance for an intermediate number of neighbours and molecular solids have the smallest atomic coordination numbers.

The Tersoff potential was designed specifically for the group 14 elements and extends the basic empirical bond-order model by including an angular term. The interaction energy between two atoms i and j using this potential is:

$$\nu_{ij} = f_C(r_{ij})[A e^{-\lambda_1 r_{ij}} - b_{ij} B e^{-\lambda_2 r_{ij}}]$$

where

$$b_{ij} = (1 + \beta^n \zeta_{ij}^n)^{-1/2n}; \quad \zeta_{ij} = \sum_{k \neq i, j} f_C(r_{ik}) g(\theta_{ijk}) \exp[\lambda_3^3 (r_{ij} - r_{ik})^3] \quad (4.126)$$

$$g(\theta) = 1 + \frac{c^2}{d^2} - \frac{c^2}{[d^2 + (h - \cos \theta)^2]}$$

The function f_C is a smoothing function with the value 1 up to some distance r_{ij} (typically chosen to include just the first neighbour shell) and then smoothly tapers to zero at the cutoff distance. b_{ij} is the bond-order term, which incorporates an angular term dependent upon the bond angle θ_{ijk} . The Tersoff potential is more broadly applicable than the Stillinger–Weber potential, but does contain more parameters.

Appendix 4.1 The Interaction Between Two Drude Molecules

In the system comprising two Drude molecules (see Section 4.9.1), an additional term must be included in the Hamiltonian [Rigby *et al.* 1986]. This additional term arises from the interactions between the two dipoles. The instantaneous dipole of each molecule is $qz(t)$, where $z(t)$ is the separation of the charges. Thus, if we label the molecules 1 and 2, we can write the dipole-dipole interaction energy as:

$$\nu(\mu_1, \mu_2) = -\frac{2\mu_1\mu_2}{4\pi\varepsilon_0 r^3} = -\frac{2z_1 z_2 q^2}{4\pi\varepsilon_0 r^3} \quad (4.127)$$

r is the separation of the two molecules. The Schrödinger equation for this system is thus:

$$-\frac{\hbar^2}{2m} \frac{\partial^2 \psi}{\partial z_1^2} - \frac{\hbar^2}{2m} \frac{\partial^2 \psi}{\partial z_2^2} + \left[\frac{1}{2} k z_1^2 + \frac{1}{2} k z_2^2 - \frac{2z_1 z_2 q^2}{4\pi\varepsilon_0 r^3} \right] \psi = E\psi \quad (4.128)$$

This equation can be solved by making the following substitutions:

$$a_1 = \frac{z_1 + z_2}{\sqrt{2}}; \quad a_2 = \frac{z_1 - z_2}{\sqrt{2}}; \quad k_1 = k - \frac{2q^2}{4\pi\varepsilon_0 r^3}; \quad k_2 = k + \frac{2q^2}{4\pi\varepsilon_0 r^3} \quad (4.129)$$

These reduce Equation (4.128) to

$$-\frac{\hbar^2}{2m} \frac{\partial^2 \psi}{\partial a_1^2} - \frac{\hbar^2}{2m} \frac{\partial^2 \psi}{\partial a_2^2} + [\frac{1}{2} k_1 a_1^2 + \frac{1}{2} k_2 a_2^2] \psi = E\psi \quad (4.130)$$

This is the Schrödinger equation for two independent (i.e. non-interacting) oscillators with frequencies given as follows:

$$\omega_1 = \omega \sqrt{1 - \frac{2q^2}{4\pi\varepsilon_0 r^3 k}}; \quad \omega_2 = \omega \sqrt{1 + \frac{2q^2}{4\pi\varepsilon_0 r^3 k}} \quad (4.131)$$

$\omega/2\pi$ is the frequency of an isolated Drude molecule. The ground state energy of the system is therefore just the sum of the zero-point energies of the two oscillators: $E_0 = \frac{1}{2}\hbar(\omega_1 + \omega_2)$

If we now substitute for ω_1 and ω_2 and expand the square roots using the binomial theorem, then we obtain the following:

$$E_0(r) = \hbar\omega - \frac{q^4 \hbar \omega}{2(4\pi\varepsilon_0)^2 r^6 k^2} - \dots \quad (4.132)$$

The interaction energy of the two oscillators is the difference between this zero-point energy and the energy of the system when the oscillators are infinitely separated and so:

$$\nu(r) = -\frac{q^4 \hbar \omega}{2(4\pi\varepsilon_0)^2 r^6 k^2} \quad (4.133)$$

The force constant, k , is related to the polarisability of the molecule, α as follows. Suppose a single Drude molecule is exposed to an external electric field \mathbf{E} . In the electric field, a force $q\mathbf{E}$ acts on each charge (in opposite directions as the charges are of opposite sign). This force causes the charges to separate and equilibrium is reached when the restoring force due to the stretching of the bond (kz) is equal to the electrostatic force: $qE = kz$. This separation

of the charges is equivalent to a static dipole given by $\mu_{ind} = qz = q^2E/k$. However, the induced dipole is also related to the polarisability by $\mu_{ind} = \alpha E$. Thus the polarisability can be written in terms of the force constant k : $\alpha = q^2/k$. With this substitution the result for the Drude model in two dimensions is:

$$\nu(r) = -\frac{\alpha^4 \hbar \omega}{2(4\pi\epsilon_0)^2 r^6} \quad (4.134)$$

In three dimensions the equivalent result is:

$$\nu(r) = -\frac{3\alpha^4 \hbar \omega}{4(4\pi\epsilon_0)^2 r^6} \quad (4.135)$$

Further Reading

- Bowen J P and N L Allinger 1991. Molecular Mechanics. The Art and Science of Parameterisation. In Lipkowitz K B and D B Boyd (Editors). *Reviews in Computational Chemistry* Volume 2. New York, VCH Publishers, pp 81-97
- Brenner D W, O A Shendreova and D A Areshkin 1998 Quantum-Based Analytic Interatomic Forces and Materials Simulation. In Lipkowitz K B and D B Boyd (Editors) *Reviews in Computational Chemistry* Volume 12 New York, VCH Publishers, pp. 207-239.
- Burkert U and N L Allinger 1982. *Molecular Mechanics*. ACS Monograph 177. Washington D.C., American Chemical Society.
- Dykstra C E 1993. Electrostatic Interaction Potentials in Molecular Force Fields *Chemical Reviews* **93**:2339-2353
- Landis C R, D M Root and T Cleveland 1995 Molecular Mechanics Force Fields for Modeling Inorganic and Organometallic Compounds. In Lipkowitz K B and D B Boyd (Editors) *Reviews in Computational Chemistry* Volume 6 New York, VCH Publishers, pp. 73-148
- Niketic S R and K Rasmussen 1977. *The Consistent Force Field. A Documentation*. Berlin, Springer-Verlag
- Price S L 2000 Towards More Accurate Model Intermolecular Potentials for Organic Molecules. In Lipkowitz K B and D B Boyd (Editors) *Reviews in Computational Chemistry* Volume 14 New York, VCH Publishers, pp 225-289
- Rigby M, E B Smith, W A Wakeham and G C Maitland 1981 *Intermolecular Forces: Their Origin and Determination*. Oxford, Clarendon Press.
- Rigby M, E B Smith, W A Wakeham and G C Maitland 1986. *The Forces Between Molecules* Oxford, Clarendon Press.
- Van der Graaf B, S L Njo and K S Smirnov 2000. Introduction to Zeolite Modeling In Lipkowitz K B and D B Boyd (Editors) *Reviews in Computational Chemistry* Volume 14 New York, VCH Publishers, pp. 137-223
- Williams D E 1991. Net Atomic Charge and Multipole Models for the *Ab Initio* Molecular Electric Potential In Lipkowitz K B and D B Boyd (Editors) *Reviews in Computational Chemistry* Volume 2 New York, VCH Publishers, pp. 219-271.

References

- Abell G C 1985. Empirical Chemical Pseudopotential Theory of Molecular and Metallic Bonding *Physical Review B* **31**:6184-6196.

- Allan, N L and W C Mackrodt 1994 Density Functional Theory and Interionic Potentials *Philosophical Magazine* **B69**:871–878
- Allinger N L 1977. Conformational Analysis 130. MM2. A Hydrocarbon Force Field Utilizing V₁ and V₂ Torsional Terms *Journal of the American Chemical Society* **99**:8127–8134.
- Allinger N L, K Chen and J-H Lii 1996a. An Improved Force Field (MM4) for Saturated Hydrocarbons *Journal of Computational Chemistry* **17**:642–668
- Allinger N L, K Chen, J A Katzenelenbogen, S R Wilson and G M Anstead 1996b. Hyperconjugative Effects on Carbon–Carbon Bond Lengths in Molecular Mechanics (MM4) *Journal of Computational Chemistry* **17**:747–755.
- Allinger N L, F Li and L Yan 1990a Molecular Mechanics The MM3 Force Field for Alkenes. *Journal of Computational Chemistry* **11**:848–867.
- Allinger N L, F Li, L Yan and J C Tai 1990b. Molecular Mechanics (MM3) Calculations on Conjugated Hydrocarbons. *Journal of Computational Chemistry* **11**:868–895
- Allinger N L and J T Sprague 1973 Calculation of the Structures of Hydrocarbons Containing Delocalised Electronic Systems by the Molecular Mechanics Method. *Journal of the American Chemical Society* **95** 3893–3907
- Allinger N L, Y H Yuh and J-J Lii 1989. Molecular Mechanics The MM3 Force Field for Hydrocarbons I. *Journal of the American Chemical Society* **111**:8551–9556
- Allured V S, C M Kelly and C R Landis 1991 SHAPES Empirical Force-Field – New Treatment of Angular Potentials and Its Application to Square-Planar Transition-Metal Complexes. *Journal of the American Chemical Society* **113**:1–12
- Barker J A, R A Fisher and R O Watts 1971. Liquid Argon: Monte Carlo and Molecular Dynamics Calculations *Molecular Physics* **21**:657–673
- Barnes P, J L Finney, J D Nicholas and J E Quinn 1979 Cooperative Effects in Simulated Water *Nature* **282**:459–464.
- Bayly C I, P Cieplak, W D Cornell and P A Kollman 1993. A Well-Behaved Electrostatic Potential Based Method for Deriving Atomic Charges – The RESP Model. *Journal of Physical Chemistry* **97**:10269–10280.
- Berendsen H C, J P M Postma, W F van Gunsteren and J Hermans 1981 Interaction Models for Water in Relation to Protein Hydration. In Pullman B (Editor) *Intermolecular Forces* Dordrecht, Reidel, pp 331–342.
- Berendsen H J C, J R Grigera and T P Straatsma 1987. The Missing Term in Effective Pair Potentials *Journal of Physical Chemistry* **91**:6269–6271.
- Bernal J D and R H Fowler 1933. A Theory of Water and Ionic Solution, with Particular Reference to Hydrogen and Hydroxyl Ions *Journal of Chemical Physics* **1**:515–548.
- Bezler B H, K M Merz Jr and P A Kollman 1990. Atomic Charges Derived from Semi-Empirical Methods *Journal of Computational Chemistry* **11**:431–439
- Born M 1920. Volumen und Hydratationswärme der Ionen *Zeitschrift für Physik* **1**:45–48.
- Breneman C M and K B Wiberg 1990 Determining Atom-Centred Monopoles from Molecular Electrostatic Potentials. The Need for High Sampling Density in Formamide Conformational Analysis *Journal of Computational Chemistry* **11**:361–373.
- Buckingham A D 1959 Molecular Quadrupole Moments. *Quarterly Reviews of the Chemical Society* **13**:183–214.
- Churlan L E and M M Franci 1987. Atomic Charges Derived from Electrostatic Potentials: A Detailed Study *Journal of Computational Chemistry* **8**:894–905.
- Claessens M, M Ferrario and J-P Ryckaert 1983 The Structure of Liquid Benzene. *Molecular Physics* **50**:217–227.
- Cleveland T and C R Landis 1996. Valence Bond Concepts Applied to the Molecular Mechanics Description of Molecular Shapes. 2 Applications to Hypervalent Molecules of the P-Block *Journal of the American Chemical Society* **118** 6020–6030.

- Corey E J and J C Bailar Jr 1959 The Stereochemistry of Complex Inorganic Compounds XXII Stereospecific Effects in Complex Ions. *Journal of the American Chemical Society* **81**:2620–2629
- Cornell W D, P Cieplak, C I Bayly, I R Gould, K M Merz Jr, D M Ferguson, D C Spellmeyer, T Fox, J W Caldwell and P A Kollman 1995. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids and Organic Molecules *Journal of the American Chemical Society* **117**:5179–5197
- Corongiu G 1992. Molecular Dynamics Simulation for Liquid Water Using a Polarisable and Flexible Potential. *International Journal of Quantum Chemistry* **42** 1209–1235.
- Cox S R and D E Williams 1981. Representation of the Molecular Electrostatic Potential by a New Atomic Charge Model. *Journal of Computational Chemistry* **2**:304–323
- Dang L X, J E Rice, J Caldwell and P A Kollman 1991. Ion Solvation in Polarisable Water: Molecular Dynamics Simulations. *Journal of the American Chemical Society* **113** 2481–2486
- Daw M S and M I Baskes 1984. Embedded-atom Method: Derivation and Application to Impurities, Surfaces, and Other Defects in Metals *Physical Review* **B29**:6443–6453.
- Dick B G and A W Overhauser 1958 Theory of the Dielectric Constants of Alkali Halide Crystals *Physical Review* **112** 90–103.
- Dinur U and A T Hagler 1991. New Approaches to Empirical Force Fields. In K B Lipkowitz and D B Boyd (Editors). *Reviews in Computational Chemistry* Volume 2. New York, VCH Publishers, pp. 99–164.
- Dinur U and A T Hagler 1995. Geometry-Dependent Atomic Charges Methodology and Application to Alkanes, Aldehydes, Ketones and Amides *Journal of Computational Chemistry* **16**:154–170.
- Ferenczy G G, C A Reynolds and W G Richards 1990 Semi-Empirical AM1 Electrostatic Potentials and AM1 Electrostatic Potential Derived Charges – A Comparison with *Ab Initio* Values. *Journal of Computational Chemistry* **11**:159–169.
- Ferguson D M 1995 Parameterisation and Evaluation of a Flexible Water Model. *Journal of Computational Chemistry* **16**.501–511.
- Finnis M W and J E Sinclair 1984. A Simple Empirical *N*-body Potential for Transition Metals. *Philosophical Magazine* **A50** 45–55.
- Fowler P W and A D Buckingham 1991 Central or Distributed Multipole Moments? Electrostatic Models of Aromatic Dimers *Chemical Physics Letters* **176**:11–18.
- Gale J D, C R A Catlow and W C Mackrodt 1992. Periodic *Ab Initio* Determination of Interatomic Potentials for Alumina. *Modelling and Simulation in Materials Science and Engineering* **1** 73–81
- Gasteiger J and M Marsili 1980 Iterative Partial Equalization of Orbital Electronegativity – Rapid Access to Atomic Charges. *Tetrahedron* **36**.3219–3288.
- Gay J G and B J Berne 1981. Modification of the Overlap Potential to Mimic a Linear Site–Site Potential. *Journal of Chemical Physics* **74**:3316–3319.
- Goodford P J 1985 A Computational Procedure for Determining Energetically Favorable Binding Sites on Biologically Important Macromolecules. *Journal of Medicinal Chemistry* **28**.849–857
- Hagler A T, E Huler and S Lifson 1977. Energy Functions for Peptides and Proteins. I. Derivation of a Consistent Force Field Including the Hydrogen Bond from Amide Crystals *Journal of the American Chemical Society* **96** 5319–5327.
- Hagler A T and S Lifson 1974 Energy Functions for Peptides and Proteins II. The Amide Hydrogen Bond and Calculation of Amide Crystal Properties. *Journal of the American Chemical Society* **96**:5327–5335.
- Halgren T A 1992 Representation of van der Waals (vdW) Interactions in Molecular Mechanics Force Fields. Potential Form, Combination Rules, and vdW Parameters. *Journal of the American Chemical Society* **114**:7827–7843.
- Halgren T A 1996a. Merck Molecular Force Field I. Basis, Form, Scope, Parameterisation and Performance of MMFF94 *Journal of Computational Chemistry* **17**:490–519.
- Halgren T A 1996b. Merck Molecular Force Field II. MMFF94 van der Waals and Electrostatic Parameters for Intermolecular Interactions. *Journal of Computational Chemistry* **17**:520–552

- Hill T L 1948. Steric Effects I. Van der Waals Potential Energy Curves *Journal of Chemical Physics* **16**:399–404.
- Hunter C A 1993. Sequence-dependent DNA structure The role of base stacking interactions. *Journal of Molecular Biology* **230**:1024–1054
- Hunter C A and J K M Saunders 1990. The Nature of π - π Interactions *The Journal of the American Chemical Society* **112**:5525–5534.
- Hwang M J, T P Stockfisch and A T Hagler 1994. Derivation of Class II Force Fields 2 Derivation and Characterisation of a Class II Force Field, CFF93, for the Alkyl Functional Group and Alkane Molecules. *Journal of the American Chemical Society* **116**:2515–2525.
- Jorgensen W L, J Chandrasekhar, J D Madura, R W Impey and M L Klein 1983 Comparison of Simple Potential Functions for Simulating Liquid Water. *Journal of Chemical Physics* **79**:926–935.
- Jorgensen W L and J Pranata 1990 Importance of Secondary Interactions in Triply Hydrogen Bonded Complexes: Guanine–Cytosine vs Uracil–2,6-Diaminopyridine. *Journal of the American Chemical Society* **112**:2008–2010.
- Jorgensen W L and J Tirado-Rives 1988 The OPLS Potential Functions for Proteins – Energy Minimizations for Crystals of Cyclic-Peptides and Crambin. *Journal of the American Chemical Society* **110**:1666–1671.
- Landis C R, T Cleveland and T K Firman 1995. Making Sense of the Shapes of Simple Metal Hydrides *Journal of the American Chemical Society* **117**:1859–1860
- Landis C R, T K Firman, D M Root and T Cleveland 1998. A Valence Bond Perspective on the Molecular Shapes of Simple Metal Alkyls and Hydrides *Journal of the American Chemical Society* **120**:1842–1854
- Lifson S and A Warshel 1968. Consistent Force Field for Calculations of Conformations, Vibrational Spectra and Enthalpies of Cycloalkane and *n*-Alkane Molecules *Journal of Chemical Physics* **49**:5116–5129.
- Lii J-H and N L Allinger 1989. Molecular Mechanics. The MM3 Force Field for Hydrocarbons 2 Vibrational Frequencies and Thermodynamics *Journal of the American Chemical Society* **111**:8566–8582
- London F 1930 Zur Theorie und Systematik der Molekularkräfte *Zeitschrift für Physik* **63**:245–279
- Luckhurst G R, R A Stephens and R W Phippen 1990 Computer Simulation Studies of Anisotropic Systems XIX Mesophases Formed by the Gay–Berne Model Mesogen *Liquid Crystals* **8**:451–464
- Luque F J, F Ilas and M Orozco 1990 Comparative Study of the Molecular Electrostatic Potential Obtained from Different Wavefunctions – Reliability of the Semi-Empirical MNDO Wavefunction *Journal of Computational Chemistry* **11**:416–430.
- Lybrand T P and P A Kollman 1985 Water-Water and Water-Ion Potential Functions Including Terms for Many Body Effects. *Journal of Chemical Physics* **83**:2923–2933
- Maple J R, U Dinur and A T Hagler 1988. Derivation of Force Fields for Molecular Mechanics and Molecular Dynamics from *Ab Initio* Energy Surfaces *Proceedings of the National Academy of Sciences USA* **85**:5350–5354
- Nevins N, K Chen and N L Allinger 1996a. Molecular Mechanics (MM4) Calculations on Alkenes. *Journal of Computational Chemistry* **17**:669–694.
- Nevins N, K Chen and N L Allinger 1996b. Molecular Mechanics (MM4) Calculations on Conjugated Hydrocarbons. *Journal of Computational Chemistry* **17**:695–729.
- Nevins N, K Chen and N L Allinger 1996c. Molecular Mechanics (MM4) Vibrational Frequency Calculations for Alkenes and Conjugated Hydrocarbons. *Journal of Computational Chemistry* **17**:730–746
- Nicholas J B, A J Hopfinger, F R Trouw and L E Iton 1991 Molecular Modelling of Zeolite Structure. 2. Structure and Dynamics of Silica Sodalite and Silicate Force Field. *The Journal of the American Chemical Society* **113**:4792–4800.
- Niesar U, G Corongiu, E Clementi, G R Keller and D K Bhattacharya 1990. Molecular Dynamics Simulations of Liquid Water Using the NCC *Ab Initio* Potential *Journal of Physical Chemistry* **94**:7949–7956.

- Niketic S R and K Rasmussen 1977 *The Consistent Force Field: A Documentation*. Berlin, Springer-Verlag
- Packer M J, M P Dauncey and C A Hunter 2000 Sequence-dependent DNA Structure Dinucleotide Conformational Maps *Journal of Molecular Biology* **295**:71–83.
- Pranata J and W L Jorgensen 1991. Computational Studies on FK506: Computational Search and Molecular Dynamics Simulations in Water. *Journal of the American Chemical Society* **113**:9483–9493.
- Price S L, R J Harrison and M F Guest 1989. An *Ab Initio* Distributed Multipole Study of the Electrostatic Potential Around an Undecapeptide Cyclosporin Derivative and a Comparison with Point Charge Electrostatic Models *Journal of Computational Chemistry* **10**:552–567.
- Rappé A K, C J Casewit, K S Colwell, W A Goddard III and W M Skiff 1992 UFF, a Full Periodic Table Force Field for Molecular Mechanics and Molecular Dynamics Simulations *Journal of the American Chemical Society* **114** 10024–10035.
- Rappé A K, K S Colwell and C J Casewit 1993 Application of a Universal Force Field to Metal Complexes. *Inorganic Chemistry* **32**:3438–3450.
- Rappé A K and W A Goddard III 1991. Charge Equilibration for Molecular Dynamics Simulations. *Journal of Physical Chemistry* **95**:3358–3363
- Reynolds C A, J W Essex and W G Richards 1992. Atomic Charges for Variable Molecular Conformations *Journal of the American Chemical Society* **114**:9075–9079.
- Rick S W and B J Berne 1996. Dynamical Fluctuating Charge Force Fields: The Aqueous Solvation of Amides. *Journal of the American Chemical Society* **118**:672–679.
- Rick S W, S J Stuart and B J Berne 1994. Dynamical Fluctuating Charge Force Fields: Application to Liquid Water. *Journal of Chemical Physics* **101**:6141–6156.
- Rigby M, E B Smith, W A Wakeham and G C Maitland 1986 *The Forces Between Molecules*. Oxford, Clarendon Press.
- Rodger P M, A J Stone and D J Tildesley 1988 The Intermolecular Potential of Chlorine. A Three Phase Study. *Molecular Physics* **63**:173–188
- Singh U C and P A Kollman 1984. An Approach to Computing Electrostatic Charges for Molecules. *Journal of Computational Chemistry* **5**:129–145
- Smith P E and B M Pettitt 1994. Modelling Solvent in Biomolecular Systems. *Journal of Physical Chemistry* **98**:9700–9711.
- Sprague J T, J C Tai, Y Yuh and N L Allinger 1987 The MMP2 Calculational Method *Journal of Computational Chemistry* **8**:581–603
- Sprik M and M L Klein 1988. A Polarisable Model for Water Using Distributed Charge Sites. *Journal of Chemical Physics* **89**:7556–7560
- Stillinger F H and A Rahman 1974 Improved Simulation of Liquid Water by Molecular Dynamics. *Journal of Chemical Physics* **60**:1545–1557.
- Stillinger F H and T A Weber 1985. Computer Simulation of Local Order in Condensed Phases of Silicon. *Physical Review B* **31**:5262–5271
- Stone A J 1981. Distributed Multipole Analysis, or How to Describe a Molecular Charge Distribution *Chemical Physics Letters* **83**:233–239
- Stone A J and M Alderton 1985 Distributed Multipole Analysis Methods and Applications *Molecular Physics* **56**:1047–1064.
- Stuart S J and B J Berne 1996. Effects of Polarisability on the Hydration of the Chloride Ion. *Journal of Physical Chemistry* **100**:11934–11943.
- Sutton A P and J Chen 1990. Long-range Finnis-Sinclair Potentials. *Philosophical Magazine Letters* **61**:139–146.
- Tersoff J 1988. New Empirical Approach for the Structure and Energy of Covalent Systems. *Physical Review B* **37**:6991–7000
- Toxvaerd S 1990. Molecular Dynamics Calculation of the Equation of State of Alkanes *Journal of Chemical Physics* **93**:4290–4295

- Vedani A 1988. YETI: An Interactive Molecular Mechanics Program for Small-Molecular Protein Complexes *Journal of Computational Chemistry* **9**:269–280.
- Vinter J G 1994. Extended Electron Distributions Applied to the Molecular Mechanics of Some Intermolecular Interactions *Journal of Computer-Aided Molecular Design* **8**:653–668.
- Warshel A and M Karplus 1972. Calculation of Ground and Excited State Potential Surfaces of Conjugated Molecules. I Formulation and Parameterisation *Journal of the American Chemical Society* **94**.5612-5622.
- Warshel A and A Lappicirella 1981. Calculations for Ground- and Excited-State Potential Surfaces for Conjugated Heteroatomic Molecules. *Journal of the American Chemical Society* **103**.4664-4673
- Weiner S J, P A Kollman, D A Case, U C Singh, C Ghio, G Alagona, S Profeta and P Weiner 1984 A New Force Field for Molecular Mechanical Simulation of Nucleic Acids and Proteins. *Journal of the American Chemical Society* **106**.765–784.
- Williams D E 1990. Alanyl Dipeptide Potential-Derived Net Atomic Charges and Bond Dipoles, and Their Variation with Molecular Conformation. *Biopolymers* **29**:1367–1386

Energy Minimisation and Related Methods for Exploring the Energy Surface

5.1 Introduction

For all except the very simplest systems the potential energy is a complicated, multi-dimensional function of the coordinates. For example, the energy of a conformation of ethane is a function of the 18 internal coordinates or 24 Cartesian coordinates that are required to completely specify the structure. As we discussed in Section 1.3, the way in which the energy varies with the coordinates is usually referred to as the *potential energy surface* (sometimes called the *hypersurface*). In the interests of brevity all references to ‘energy’ should be taken to mean ‘potential energy’ for the rest of this chapter, except where explicitly stated otherwise. For a system with N atoms the energy is thus a function of $3N - 6$ internal or $3N$ Cartesian coordinates. It is therefore impossible to visualise the entire energy surface except for some simple cases where the energy is a function of just one or two coordinates. For example, the van der Waals energy of two argon atoms (as might be modelled using the Lennard-Jones potential function) depends upon just one coordinate: the interatomic distance. Sometimes we may wish to visualise just a part of the energy surface. For example, suppose we take an extended conformation of pentane and rotate the two central carbon–carbon bonds so that the torsion angles vary from 0° to 360° , calculating the energy of each structure generated. The energy in this case is a function of just two variables and can be plotted as a contour diagram or as an isometric plot, as shown in Figure 5.1.

We will use the term ‘energy surface’ to refer not only to systems in which the bonding remains unchanged, as in these two examples, but also where bonds are broken and/or formed. Our discussion will be appropriate to both quantum mechanics and molecular mechanics, except where otherwise stated.

In molecular modelling we are especially interested in minimum points on the energy surface. Minimum energy arrangements of the atoms correspond to stable states of the system; any movement away from a minimum gives a configuration with a higher energy. There may be a very large number of minima on the energy surface. The minimum with the very lowest energy is known as the *global energy minimum*. To identify those geometries of

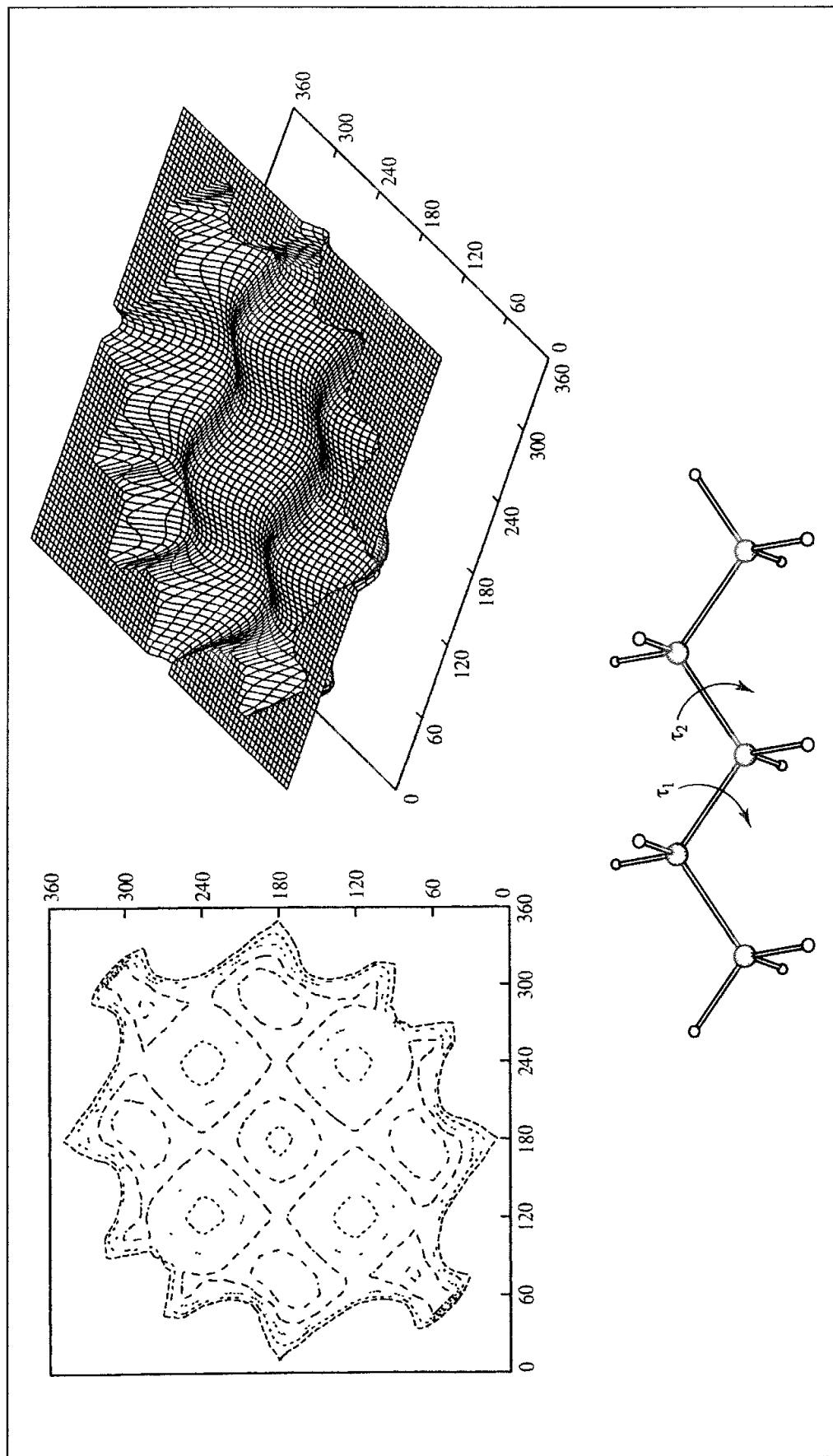


Fig. 5.1 Variation in the energy of pentane with the two torsion angles indicated and represented as a contour diagram and isometric plot. Only the lowest-energy regions are shown.

the system that correspond to minimum points on the energy surface we use a *minimisation algorithm*. There is a vast literature on such methods and so we will concentrate on those approaches that are most commonly used in molecular modelling. We may also be interested to know how the system changes from one minimum energy structure to another. For example, how do the relative positions of the atoms vary during a reaction? What structural changes occur as a molecule changes its conformation? The highest point on the pathway between two minima is of especial interest and is known as the *saddle point*, with the arrangement of the atoms being the *transition structure*. Both minima and saddle points are stationary points on the energy surface, where the first derivative of the energy function is zero with respect to all the coordinates.

A geographical analogy can be a helpful way to illustrate many of the concepts we shall encounter in this chapter. In this analogy minimum points correspond to the bottom of valleys. A minimum may be described as being in a ‘long and narrow valley’ or ‘a flat and featureless plain’. Saddle points correspond to mountain passes. We refer to algorithms taking steps ‘uphill’ or ‘downhill’.

5.1.1 Energy Minimisation: Statement of the Problem

The minimisation problem can be formally stated as follows: given a function f which depends on one or more independent variables x_1, x_2, \dots, x_i , find the values of those variables where f has a minimum value. At a minimum point the first derivative of the function with respect to each of the variables is zero and the second derivatives are all positive:

$$\frac{\partial f}{\partial x_i} = 0; \quad \frac{\partial^2 f}{\partial x_i^2} > 0 \quad (5.1)$$

The functions of most interest to us will be the quantum mechanics or molecular mechanics energy with the variables x_i being the Cartesian or the internal coordinates of the atoms. Molecular mechanics minimisations are nearly always performed in Cartesian coordinates, where the energy is a function of $3N$ variables; it is more common to use internal coordinates (as defined in the Z-matrix) with quantum mechanics. For analytical functions, the minimum of a function can be found using standard calculus methods. However, this is not generally possible for molecular systems due to the complicated way in which the energy varies with the coordinates. Rather, minima are located using numerical methods, which gradually change the coordinates to produce configurations with lower and lower energies until the minimum is reached. To illustrate how the various minimisation algorithms operate, we shall consider a simple function of two variables: $f(x, y) = x^2 + 2y^2$. This function is represented as a contour diagram in Figure 5.2. The function has one minimum point, located at the origin. In our examples we will attempt to locate the minimum from the point (9.0, 9.0). Although this is a function of just two variables for the purposes of illustration, all of the methods that we shall consider can be applied to functions of many more variables.

We can classify minimisation algorithms into two groups: those which use derivatives of the energy with respect to the coordinates and those which do not. Derivatives can be useful

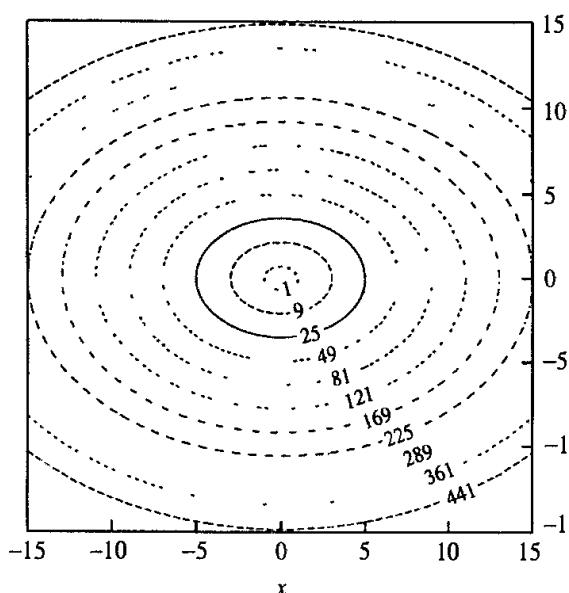


Fig. 5.2. The function $x^2 + 2y^2$.

because they provide information about the shape of the energy surface, and, if used properly, they can significantly enhance the efficiency with which the minimum is located. There are many factors that must be taken into account when choosing the most appropriate algorithm (or combination of algorithms) for a given problem; the ideal minimisation algorithm is the one that provides the answer as quickly as possible, using the least amount of memory. No single minimisation method has yet proved to be the best for all molecular modelling problems and so most software packages offer a choice of methods. In particular, a method that works well with quantum mechanics may not be the most suitable for use with molecular mechanics. This is partly because quantum mechanics is usually used to model systems with fewer atoms than molecular mechanics; some operations that are integral to certain minimisation procedures (such as matrix inversion) are trivial for small systems but formidable for systems containing thousands of atoms. Quantum mechanics and molecular mechanics also require different amounts of computational effort to calculate the energies and the derivatives of the various configurations. Thus an algorithm that takes many steps may be appropriate for molecular mechanics but inappropriate for quantum mechanics.

Most minimisation algorithms can only go downhill on the energy surface and so they can only locate the minimum that is nearest (in a downhill sense) to the starting point. Thus, Figure 5.3 shows a schematic energy surface and the minima that would be obtained starting from three points A, B and C. The minima can be considered to correspond to the locations where a ball rolling on the energy surface under the influence of gravity would come to rest. To locate more than one minimum or to locate the global energy minimum we therefore usually require a means of generating different starting points, each of which is then minimised. Some specialised minimisation methods can make uphill moves to seek out minima lower in energy than the nearest one, but no algorithm has yet proved capable of locating the

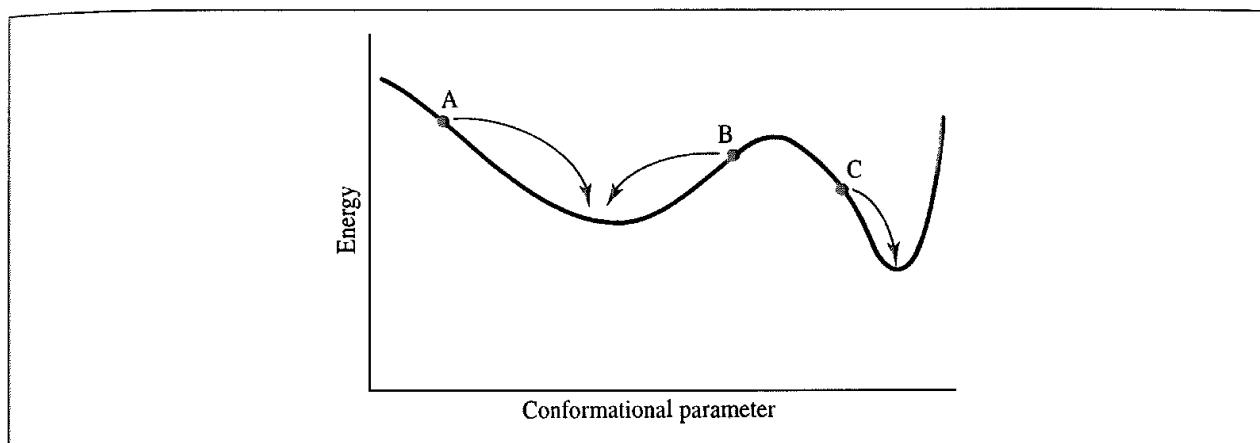


Fig 5.3: A schematic one-dimensional energy surface. Minimisation methods move downhill to the nearest minimum. The statistical weight of the narrow, deep minimum may be less than a broad minimum which is higher in energy.

global energy minimum from an arbitrary starting position. The shape of the energy surface may be important if one wishes to calculate the relative populations of the various minimum energy structures. For example, a deep and narrow minimum may be less highly populated than a broad minimum that is higher in energy as the vibrational energy levels will be more widely spaced in the deeper minimum and so less accessible. For this reason the global energy minimum may not be the most highly populated minimum. In any case, the 'active' structure (e.g. the biologically active conformation of a drug molecule) may not correspond to the global minimum, or to the most highly populated conformation, or even to a minimum energy structure at all.

The input to a minimisation program consists of a set of initial coordinates for the system. The initial coordinates may come from a variety of sources. They may be obtained from an experimental technique, such as X-ray crystallography or NMR. In other cases a theoretical method is employed, such as a conformational search algorithm. A combination of experimental and theoretical approaches may also be used. For example, to study the behaviour of a protein in water one may take an X-ray structure of the protein and immerse it in a solvent 'bath', where the coordinates of the solvent molecules have been obtained from a Monte Carlo or molecular dynamics simulation.

5.1.2 Derivatives

In order to use a derivative minimisation method it is obviously necessary to be able to calculate the derivatives of the energy with respect to the variables (i.e. the Cartesian or internal coordinates, as appropriate). Derivatives may be obtained either analytically or numerically. The use of analytical derivatives is preferable as they are exact, and because they can be calculated more quickly; if only numerical derivatives are available then it may be more effective to use a non-derivative minimisation algorithm. The problems of calculating analytical derivatives with quantum mechanics and molecular mechanics were discussed in Sections 3.4.3 and 4.16, respectively.

Nevertheless, under some circumstances it is necessary to use numerical derivatives. These can be calculated as follows. If one of the coordinates x_i is changed by a small change (δx_i) and the energy for the new arrangement is computed then the derivative $\partial E / \partial x_i$ is obtained by dividing the change in energy (δE) by the change in coordinate ($\delta E / \delta x_i$). This strictly gives the derivative at the mid-point between the two points x_i and $x_i + \delta x_i$. A more accurate value of the derivative at the point x_i may be obtained (at the cost of an additional energy calculation) by evaluating the energy at two points, $x_i + \delta x_i$ and $x_i - \delta x_i$. The derivative is then obtained by dividing the difference in the energies by $2\delta x_i$.

5.2 Non-derivative Minimisation Methods

5.2.1 The Simplex Method

A *simplex* is a geometrical figure with $M + 1$ interconnected vertices, where M is the dimensionality of the energy function. For a function of two variables the simplex is thus triangular in shape. A tetrahedral simplex is used for a function of three variables and so for an energy function of $3N$ Cartesian coordinates the simplex will have $3N + 1$ vertices; if internal coordinates are used then the simplex will have $3N - 5$ vertices. Each vertex corresponds to a specific set of coordinates for which an energy can be calculated. For our function $f(x, y) = x^2 + 2y^2$ the simplex method would use a triangular simplex.

The simplex algorithm locates a minimum by moving around on the potential energy surface in a fashion that has been likened to the motion of an amoeba. Three basic kinds of move are possible. The most common type of move is a reflection of the vertex with the highest value through the opposite face of the simplex, in an attempt to generate a new point that has a lower value. If this new point is lower in energy than any of the other points in the simplex then a ‘reflection and expansion’ move may be applied. When a ‘valley floor’ is reached then a reflection move will fail to produce a better point. Under such circumstances the simplex contracts along one dimension from the highest point. If this fails to reduce the energy then a third type of move is possible, in which the simplex contracts in all directions, pulling around the lowest point. These three moves are illustrated in Figure 5.4.

To implement the simplex algorithm it is first necessary to generate the vertices of the initial simplex. The initial configuration of the system corresponds to just one of these vertices. The remaining points can be obtained in a variety of ways, but one simple method is to add a constant increment to each coordinate in turn. The energy of the system is calculated at the new point, giving the function value for the relevant vertex.

The simplex method is most useful where the initial configuration of the system is very high in energy, because it rarely fails to find a better solution. However, it can be rather expensive in terms of computer time due to the large number of energy evaluations which are required (merely to generate the initial simplex requires $3N + 1$ energy evaluations). For this reason the simplex method is often used in combination with a different minimisation algorithm: a few steps of the simplex method are used to refine the initial structure and then a more efficient method can take over.

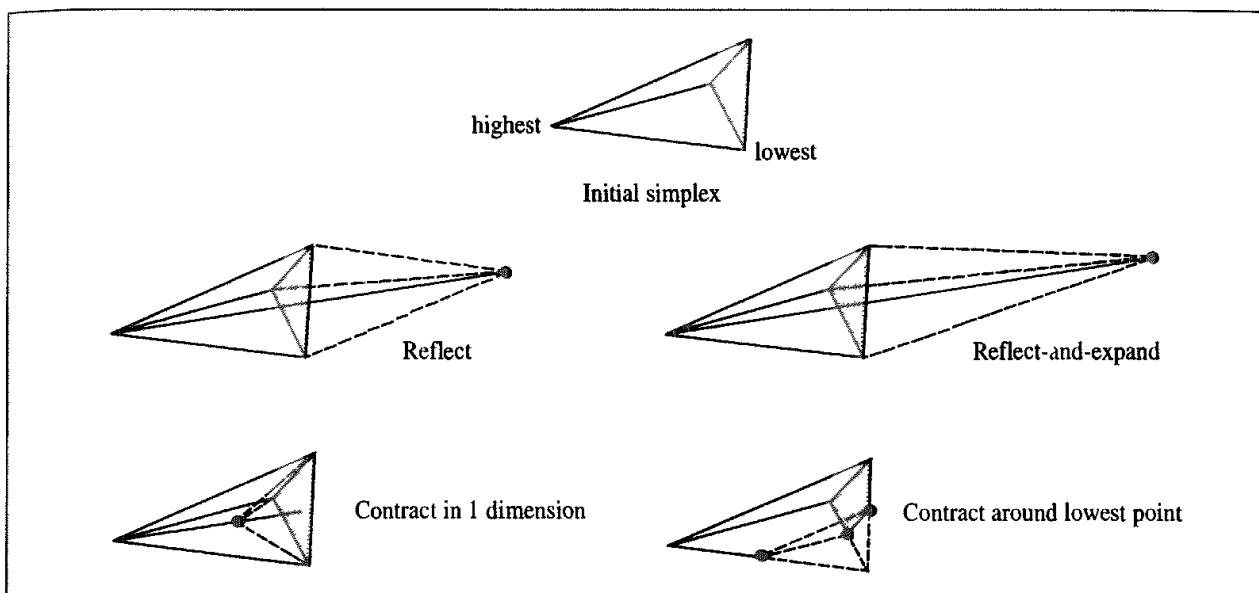


Fig 5.4: The three basic moves permitted to the simplex algorithm (reflection, and its close relation reflect-and-expand; contract in one dimension and contract around the lowest point) (Figure adapted from Press W H, B P Flannery, S A Teukolsky and W T Vetterling 1992. Numerical Recipes in Fortran. Cambridge, Cambridge University Press.)

Let us consider the application of the simplex method to our quadratic function, $f = x^2 + 2y^2$ (Figure 5.5). Suppose our initial simplex contains vertices located at the points (9, 9), (11, 9) and (9, 11), which have been generated by adding a constant factor 2 to each of the variables in turn. The values of the function at these points are 243, 283 and 323, respectively. The vertex with the highest function value is at (9, 11) and so in the first iteration this point is reflected through the opposite face of the triangle to generate a point with coordinates (11, 7) and a function value of 219 (we do not use the reflect-and-expand move in our

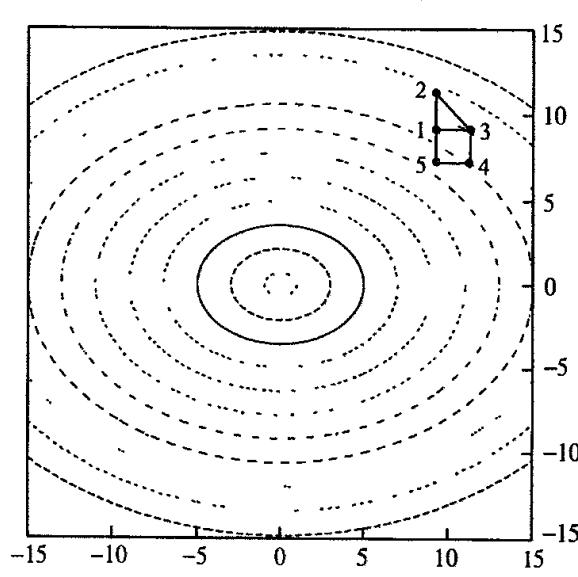


Fig 5.5: The first few steps of the simplex algorithm with the function $x^2 + 2y^2$. The initial simplex corresponds to the triangle 123. Point 2 has the largest value of the function and the next simplex is the triangle 134. The simplex for the third step is 145.

illustration). The highest vertex is now at (11, 9), which is reflected through the opposite face of the simplex to give the point (9, 7), where the function has a value of 179. In fact, for this admittedly artificial problem the simplex algorithm takes more than 30 steps to find a point where the function has a value less than 0.1.

Why does the simplex contain one more vertex than the number of degrees of freedom? The reason is that with fewer than $M + 1$ vertices the algorithm cannot explore the whole energy surface. Suppose we use only a two-vertex simplex to explore our quadratic energy surface. A simplex with just two vertices is a straight line. The only moves that would be possible in this case would be to other points that lie on this line; none of the energy surface away from the line would be explored. Similarly, if we have a function of three variables and restrict the simplex to a triangle then we will only be able to explore the region of space that lies in the same plane as the triangle, whereas the minimum may not lie in this plane.

5.2.2 The Sequential Univariate Method

The simplex method is rarely considered suitable for quantum mechanical calculations, due to the number of energy evaluations that must be performed. The sequential univariate method is a non-derivative method that is considered more appropriate in this case. This method systematically cycles through the coordinates in turn. For each coordinate, two new structures are generated by changing the current coordinate (i.e. $x_i + \delta x_i$ and $x_i + 2\delta x_i$). The energies of these two structures are calculated. A parabola is then fitted through the three points corresponding to the two distorted structures and the original structure. The minimum point in this quadratic function is determined and the coordinate is then changed to the position of the minimum. The procedure is illustrated in Figure 5.6. When the changes in all the coordinates are

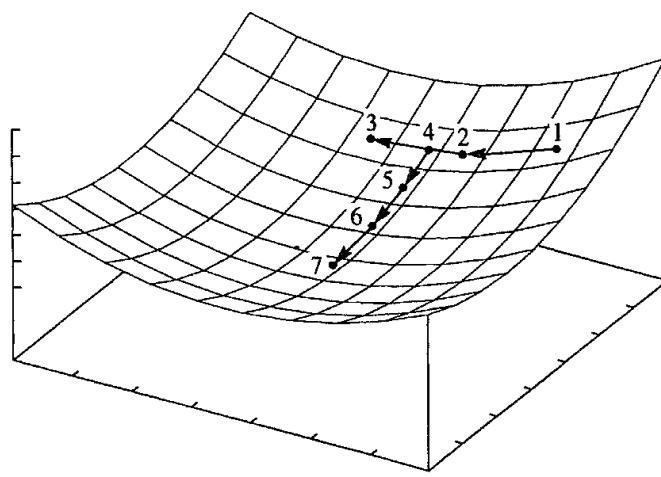


Fig. 5.6. The sequential univariate method. Starting at the point labelled 1 two steps are made along one of the coordinates to give points 2 and 3. A parabola is fitted to these three points and the minimum located (point 4). The same procedure is then repeated along the next coordinate (points 5, 6 and 7). (Figure adapted from Schlegel H B 1987. Optimization of Equilibrium Geometries and Transition Structures In Lawley K P (Editor) *Ab Initio Methods in Quantum Chemistry - I* New York, John Wiley, pp. 249–286)

sufficiently small then the minimum is deemed to have been reached, otherwise a new iteration is performed. The sequential invariate method usually requires fewer function evaluations than the simplex method but it can be slow to converge especially if there is strong coupling between two or more of the coordinates or when the energy surface is analogous to a long narrow valley.

5.3 Introduction to Derivative Minimisation Methods

Derivatives provide information that can be very useful in energy minimisation, and derivatives are used by most popular minimisation methods. The direction of the first derivative of the energy (the gradient) indicates where the minimum lies, and the magnitude of the gradient indicates the steepness of the local slope. The energy of the system can be lowered by moving each atom in response to the force acting on it; the force is equal to minus the gradient. Second derivatives indicate the curvature of the function, information that can be used to predict where the function will change direction (i.e. pass through a minimum or some other stationary point).

When discussing derivative methods it is useful to write the function as a Taylor series expansion about the point x_k :

$$\mathcal{V}(x) = \mathcal{V}(x_k) + (x - x_k)\mathcal{V}'(x_k) + (x - x_k)^2\mathcal{V}''(x_k)/2 + \dots \quad (5.2)$$

For a multidimensional function, the variable x is replaced by the vector \mathbf{x} and matrices are used for the various derivatives. Thus if the potential energy $\mathcal{V}(\mathbf{x})$ is a function of $3N$ Cartesian coordinates, the vector \mathbf{x} will have $3N$ components and \mathbf{x}_k corresponds to the current configuration of the system. $\mathcal{V}'(\mathbf{x}_k)$ is a $3N \times 1$ matrix (i.e. a vector), each element of which is the partial derivative of \mathcal{V} with respect to the appropriate coordinate, $\partial\mathcal{V}/\partial x_i$. We will also write the gradient at the point k as \mathbf{g}_k . Each element (i,j) of the matrix $\mathcal{V}''(\mathbf{x}_k)$ is the partial second derivative of the energy function with respect to the two coordinates x_i and x_j , $\partial^2\mathcal{V}/\partial x_i \partial x_j$. $\mathcal{V}''(\mathbf{x}_k)$ is thus of dimension $3N \times 3N$ and is known as the *Hessian* matrix or the *force constant* matrix. The Taylor series expansion can be written in the following form for the multidimensional case:

$$\mathcal{V}(\mathbf{x}) = \mathcal{V}(\mathbf{x}_k) + (\mathbf{x} - \mathbf{x}_k)\mathcal{V}'(\mathbf{x}_k) + (\mathbf{x} - \mathbf{x}_k)^T \cdot \mathcal{V}''(\mathbf{x}_k) \cdot (\mathbf{x} - \mathbf{x}_k)/2 + \dots \quad (5.3)$$

The energy functions used in molecular modelling are rarely quadratic and so the Taylor series expansion, Equation (5.3), can only be considered an approximation. There are two important consequences of this. The first consequence is that the performance of a given minimisation method will not be as good for a molecular mechanics or quantum mechanics energy surface as it is for a pure quadratic function. As we shall see, a second derivative method such as the Newton-Raphson algorithm can locate the minimum in a single step for a purely quadratic function, but several iterations are usually required for a typical molecular modelling energy function. The second consequence is that, far from the minimum, the harmonic approximation is a poor one and some of the less robust methods will fail, even though they may work very well close to a minimum, where the harmonic approximation is more valid. For this reason it is important to choose the minimisation

protocol with care, possibly using a robust (but perhaps inefficient) method at first, and then a less robust but more efficient method.

The derivative methods can be classified according to the highest-order derivative used. First-order methods use the first derivatives (i.e. the gradients) whereas second-order methods use both first and second derivatives. The simplex method can thus be considered a zeroth-order method as it does not use any derivatives.

5.4 First-order Minimisation Methods

Two first-order minimisation algorithms that are frequently used in molecular modelling are the method of *steepest descents* and the *conjugate gradient* method. These gradually change the coordinates of the atoms as they move the system closer and closer to the minimum point. The starting point for each iteration (k) is the molecular configuration obtained from the previous step, which is represented by the multidimensional vector \mathbf{x}_{k-1} . For the first iteration the starting point is the initial configuration of the system provided by the user, the vector \mathbf{x}_1 .

5.4.1 The Steepest Descents Method

The steepest descents method moves in the direction parallel to the net force, which in our geographical analogy corresponds to walking straight downhill. For $3N$ Cartesian coordinates this direction is most conveniently represented by a $3N$ -dimensional unit vector, \mathbf{s}_k . Thus:

$$\mathbf{s}_k = -\mathbf{g}_k / |\mathbf{g}_k| \quad (5.4)$$

Having defined the direction along which to move it is then necessary to decide how far to move along the gradient. Consider the two-dimensional energy surface of Figure 5.7. The gradient direction from the starting point is along the line indicated. If we imagine a cross-section through the surface along the line, the function will pass through a minimum and then increase, as shown in the figure. We can choose to locate the minimum point by performing a *line search* or we can take a step of arbitrary size along the direction of the force

5.4.2 Line Search in One Dimension

The purpose of a line search is to locate the minimum along a specified direction (i.e. along a line through the multidimensional space). The first stage of the line search is to *bracket* the minimum. This entails finding three points along the line such that the energy of the middle point is lower than the energy of the two outer points. If three such points can be found, then at least one minimum must lie between the two outer points. An iterative procedure can then be used to decrease the distance between the three points, gradually restricting the minimum to an even smaller region. This is conceptually an easy process but it may require a considerable number of function evaluations, making it computationally expensive. An alternative is to fit a function such as a quadratic to the three points. Differentiation of the

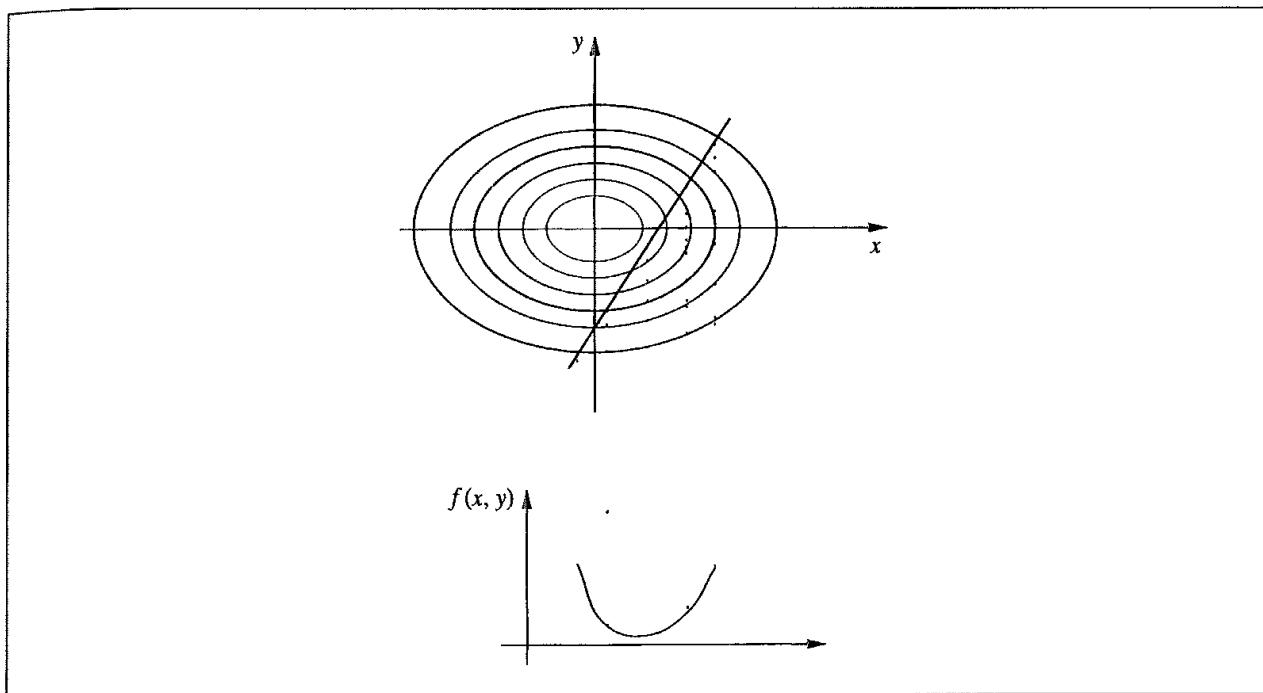


Fig 5.7 A line search is used to locate the minimum in the function in the direction of the gradient.

fitted function enables an approximation to the minimum along the line to be identified analytically. A new function can then be fitted to give a better estimate, as shown in Figure 5.8. Higher-order polynomials may give a better fit to the bracketing points but these can give incorrect interpolations when used with functions that change sharply in the bracketed region.

The gradient at the minimum point obtained from the line search will be perpendicular to the previous direction. Thus, when the line search method is used to locate the minimum along the gradient then the next direction in the steepest descents algorithm will be orthogonal to the previous direction (i.e. $\mathbf{g}_k \cdot \mathbf{g}_{k-1} = 0$).

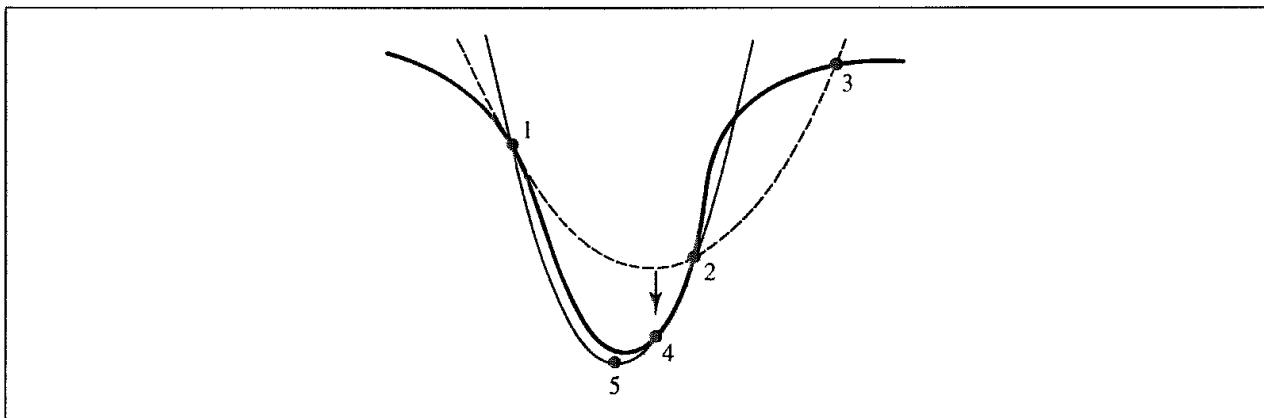


Fig 5.8: The minimum in a line search may be found more effectively by fitting an analytical function such as a quadratic to the initial set of three points (1, 2 and 3). A better estimate of the minimum can then be found by fitting a new function to the points 1, 2 and 4 and finding its minimum. (Figure adapted from Press W H, B P Flannery, S A Teukolsky and W T Vetterling 1992 Numerical Recipes in Fortran Cambridge, Cambridge University Press)

5.4.3 Arbitrary Step Approach

As the line search may itself be computationally demanding we could obtain the new coordinates by taking a step of arbitrary length along the gradient unit vector \mathbf{s}_k . The new set of coordinates after step k would then be given by the equation:

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \lambda_k \mathbf{s}_k \quad (5.5)$$

λ_k is the *step size*. In most applications of the steepest descents algorithm in molecular modelling the step size initially has a predetermined default value. If the first iteration leads to a reduction in energy, the step size is increased by a multiplicative factor (e.g. 1.2) for the second iteration. This process is repeated so long as each iteration reduces the energy. When a step produces an increase in energy, it is assumed that the algorithm has leapt across the valley which contains the minimum and up the slope on the opposite face. The step size is then reduced by a multiplicative factor (e.g. 0.5). The step size depends upon the nature of the energy surface; for a flat surface large step sizes would be appropriate but for a narrow, twisting gully a much smaller step would be more suitable. The arbitrary step method may require more steps to reach the minimum but it can often require fewer function evaluations (and thus less computer time) than the more rigorous line search approach.

The steepest descents method works as follows for our trial function, $f(x, y) = x^2 + 2y^2$. Differentiating the function gives $df = 2x \, dx + 4y \, dy$ and so the gradient at any point (x, y) equals $4y/2x$. The direction of the first move from the point $(9, 0, 9, 0)$ is $(-18, 0, -36, 0)$ and the equation of the line along which the search is performed is $y = 2x - 9$. The minimum of the function along this line can be obtained using Lagrange multipliers (see Section 1.10.5) and is at $(4.0, -1.0)$. The direction of the next move is the vector $(-8, 4)$ and the next line search is performed along the line $y = -0.5x + 1$. The minimum point along this line is $(2/3, 2/3)$ where the function has the value $4/3$. The third point found by the steepest descents method is at $(0.296, -0.074)$ where the function has the value 0.099. These moves are illustrated in Figure 5.9.

The direction of the gradient is determined by the largest interatomic forces and so steepest descents is a good method for relieving the highest-energy features in an initial configuration. The method is generally robust even when the starting point is far from a minimum, where the harmonic approximation to the energy surface is often a poor assumption. However, it suffers from the problem that many small steps will be performed when proceeding down a long narrow valley. The steepest descents method is forced to make a right-angled turn at each point, even though that might not be the best route to the minimum. The path oscillates and continually overcorrects itself, as illustrated in Figure 5.10; later steps reintroduce errors that were corrected by earlier moves.

5.4.4 Conjugate Gradients Minimisation

The conjugate gradients method produces a set of directions which does not show the oscillatory behaviour of the steepest descents method in narrow valleys. In the steepest descents method both the gradients and the direction of successive steps are orthogonal.

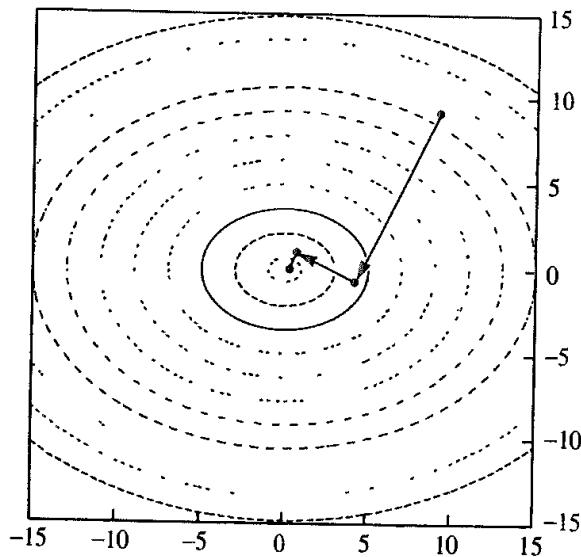


Fig 5.9. Application of steepest descents to the function $x^2 + 2y^2$.

In conjugate gradients, the gradients at each point are orthogonal but the directions are *conjugate* (indeed, the method is more properly called the conjugate directions method). A set of conjugate directions has the property that for a quadratic function of M variables, the minimum will be reached in M steps. The conjugate gradients method moves in a direction \mathbf{v}_k from point \mathbf{x}_k where \mathbf{v}_k is computed from the gradient at the point and the previous direction vector \mathbf{v}_{k-1} :

$$\mathbf{v}_k = -\mathbf{g}_k + \gamma_k \mathbf{v}_{k-1} \quad (5.6)$$

γ_k is a scalar constant given by

$$\gamma_k = \frac{\mathbf{g}_k \cdot \mathbf{g}_k}{\mathbf{g}_{k-1} \cdot \mathbf{g}_{k-1}} \quad (5.7)$$

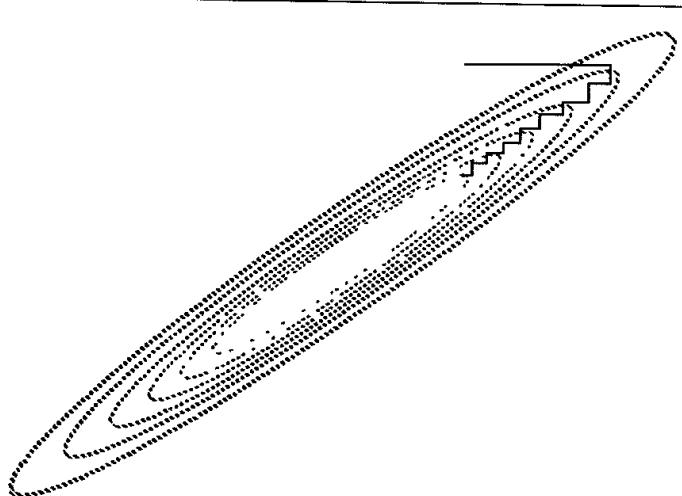


Fig 5.10 The steepest descents method can give undesirable behaviour in a long narrow valley

In the conjugate gradients method all of the directions and gradients satisfy the following relationships:

$$\mathbf{g}_i \cdot \mathbf{g}_j = 0 \quad (5.8)$$

$$\mathbf{v}_i \cdot \nabla''_{ij} \cdot \mathbf{v}_j = 0 \quad (5.9)$$

$$\mathbf{g}_i \cdot \mathbf{v}_j = 0 \quad (5.10)$$

Clearly Equation (5.6) can only be used from the second step onwards and so the first step in the conjugate gradients method is the same as the steepest descents (i.e. in the direction of the gradient). The line search method should ideally be used to locate the one-dimensional minimum in each direction to ensure that each gradient is orthogonal to all previous gradients and that each direction is conjugate to all previous directions. However, an arbitrary step method is also possible.

The conjugate gradients method deals with our simple quadratic function $f(x, y) = x^2 + 2y^2$ as follows. From the initial point $(9, 9)$ we move to the same point as in steepest descents, $(4, -1)$. To find the direction of the next move, we first determine the negative gradient at the current point. This is the vector $(-8, 4)$. This is then combined with the vector corresponding to minus the gradient at the initial point, $(-18, -36)$ multiplied by γ :

$$\mathbf{v}_k = \begin{pmatrix} -8 \\ 4 \end{pmatrix} + \frac{(-8)^2 + (4)^2}{(-18)^2 + (-36)^2} \begin{pmatrix} -18 \\ -36 \end{pmatrix} = \begin{pmatrix} -80/9 \\ +20/9 \end{pmatrix} \quad (5.11)$$

To locate the second point we therefore need to perform a line search along the line with gradient $-1/4$ that passes through the point $(4, -1)$. The minimum along this line is at the origin, at the true minimum of the function. The conjugate gradients method thus locates the exact minimum of the function exactly in just two moves, as illustrated in Figure 5.11.

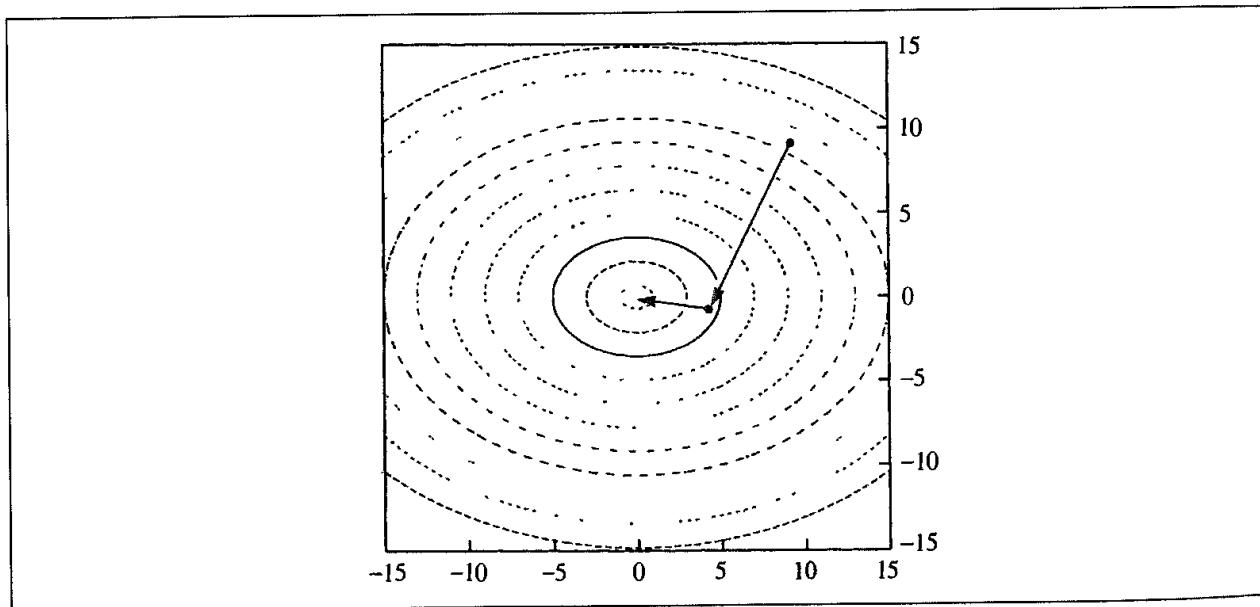


Fig 5.11 Application of conjugate gradients method to the function $x^2 + 2y^2$.

Several variants of the conjugate gradients method have been proposed. The formulation given in Equation (5.7) is the original Fletcher–Reeves algorithm. Polak and Ribiere proposed an alternative form for the scalar constant γ_k :

$$\gamma_k = \frac{(\mathbf{g}_k - \mathbf{g}_{k-1}) \cdot \mathbf{g}_k}{\mathbf{g}_{k-1} \cdot \mathbf{g}_{k-1}} \quad (5.12)$$

For a purely quadratic function the Polak–Ribiere method is identical to the Fletcher–Reeves algorithm as all gradients will be orthogonal. However, most functions of interest, including those used in molecular modelling, are at best only approximately quadratic. Polak and Ribiere claimed that their method performed better than the original Fletcher–Reeves algorithm, at least for the functions that they examined.

5.5 Second Derivative Methods: The Newton–Raphson Method

Second-order methods use not only the first derivatives (i.e. the gradients) but also the second derivatives to locate a minimum. Second derivatives provide information about the curvature of the function. The *Newton–Raphson* method is the simplest second-order method. Recall our Taylor series expansion about the point x_k , Equation (5.2):

$$\mathcal{V}(x) = \mathcal{V}(x_k) + (x - x_k)\mathcal{V}'(x_k) + (x - x_k)^2\mathcal{V}''(x_k)/2 + \dots \quad (5.13)$$

The first derivative of $\mathcal{V}(x)$ is:

$$\mathcal{V}'(x) = x\mathcal{V}'(x_k) + (x - x_k)\mathcal{V}''(x_k) \quad (5.14)$$

If the function is purely quadratic, the second derivative is the same everywhere, and so $\mathcal{V}''(x) = \mathcal{V}''(x_k)$.

At the minimum ($x = x^*$) $\mathcal{V}'(x^*) = 0$ and so

$$x^* = x_k - \mathcal{V}'(x_k)/\mathcal{V}''(x_k) \quad (5.15)$$

For a multidimensional function: $\mathbf{x}^* = \mathbf{x}_k - \mathcal{V}'(\mathbf{x}_k)\mathcal{V}''^{-1}(\mathbf{x}_k)$.

$\mathcal{V}''^{-1}(\mathbf{x}_k)$ is the inverse Hessian matrix of second derivatives, which, in the Newton–Raphson method, must therefore be inverted. This can be computationally demanding for systems with many atoms and can also require a significant amount of storage. The Newton–Raphson method is thus more suited to small molecules (usually less than 100 atoms or so). For a purely quadratic function the Newton–Raphson method finds the minimum in one step from any point on the surface, as we will now show for our function $f(x, y) = x^2 + 2y^2$.

The Hessian matrix for this function is:

$$\mathbf{f}'' = \begin{pmatrix} 2 & 0 \\ 0 & 4 \end{pmatrix} \quad (5.16)$$

The inverse of this matrix is:

$$\mathbf{f}''^{-1} = \begin{pmatrix} 1/2 & 0 \\ 0 & 1/4 \end{pmatrix} \quad (5.17)$$

The minimum is obtained using Equation (5.15):

$$\mathbf{x}^* = \begin{pmatrix} 9 \\ 9 \end{pmatrix} - \begin{pmatrix} 1/2 & 0 \\ 0 & 1/4 \end{pmatrix} \begin{pmatrix} 18 \\ 36 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad (5.18)$$

In practice, of course, the surface is only quadratic to a first approximation and so a number of steps will be required, at each of which the Hessian matrix must be calculated and inverted. The Hessian matrix of second derivatives must be *positive definite* in a Newton-Raphson minimisation. A positive definite matrix is one for which all the eigenvalues are positive. When the Hessian matrix is not positive definite then the Newton-Raphson method moves to points (e.g. saddle points) where the energy increases. In addition, far from a minimum the harmonic approximation is not appropriate and the minimisation can become unstable. One solution to this problem is to use a more robust method to get near to the minimum (i.e. where the Hessian is positive definite) before applying the Newton-Raphson method.

5.5.1 Variants on the Newton–Raphson Method

There are a number of variations on the Newton–Raphson method, many of which aim to eliminate the need to calculate the full matrix of second derivatives. In addition, a family of methods called the quasi-Newton methods require only first derivatives and gradually construct the inverse Hessian matrix as the calculation proceeds. One simple way in which it may be possible to speed up the Newton–Raphson method is to use the same Hessian matrix for several successive steps of the Newton–Raphson algorithm with only the gradients being recalculated at each iteration.

A widely used algorithm is the *block-diagonal Newton–Raphson* method in which just one atom is moved at each iteration. Consequently all terms of the form $\partial^2\mathcal{V}/\partial x_i \partial x_j$, where i and j refer to the Cartesian coordinates of atoms other than the atom being moved, will be zero. This only leaves those terms which involve the coordinates of the atom being moved and so reduces the problem to the trivial one of inverting a 3×3 matrix. However, the block-diagonal approach can be less efficient when the motions of some atoms are closely coupled, such as the concerted movements of connected atoms in a phenyl ring.

5.6 Quasi-Newton Methods

Calculation of the inverse Hessian matrix can be a potentially time-consuming operation that represents a significant drawback to the ‘pure’ second derivative methods such as Newton–Raphson. Moreover, one may not be able to calculate analytical second derivatives, which are preferable. The quasi-Newton methods (also known as variable metric methods) gradually build up the inverse Hessian matrix in successive iterations. That is, a sequence of

matrices \mathbf{H}_k is constructed that has the property

$$\lim_{k \rightarrow \infty} \mathbf{H}_k = \mathcal{V}^{\prime\prime-1} \quad (5.19)$$

At each iteration k , the new positions \mathbf{x}_{k+1} are obtained from the current positions \mathbf{x}_k , the gradient \mathbf{g}_k and the current approximation to the inverse Hessian matrix \mathbf{H}_k :

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \mathbf{H}_k \mathbf{g}_k \quad (5.20)$$

This formula is exact for a quadratic function, but for 'real' problems a line search may be desirable. This line search is performed along the vector $\mathbf{x}_{k+1} - \mathbf{x}_k$. It may not be necessary to locate the minimum in the direction of the line search very accurately, at the expense of a few more steps of the quasi-Newton algorithm. For quantum mechanics calculations the additional energy evaluations required by the line search may prove more expensive than using the more approximate approach. An effective compromise is to fit a function to the energy and gradient at the current point \mathbf{x}_k and at the point \mathbf{x}_{k+1} and determine the minimum in the fitted function.

Having moved to the new positions \mathbf{x}_{k+1} , \mathbf{H} is updated from its value at the previous step according to a formula depending upon the specific method being used. The methods of Davidon-Fletcher-Powell (DFP), Broyden-Fletcher-Goldfarb-Shanno (BFGS) and Murtaugh-Sargent (MS) are commonly encountered, but there are many others. These methods converge to the minimum, for a quadratic function of M variables, in M steps. The DFP formula is:

$$\mathbf{H}_{k+1} = \mathbf{H}_k + \frac{(\mathbf{x}_{k+1} - \mathbf{x}_k) \otimes (\mathbf{x}_{k+1} - \mathbf{x}_k)}{(\mathbf{x}_{k+1} - \mathbf{x}_k) \cdot (\mathbf{g}_{k+1} - \mathbf{g}_k)} - \frac{[\mathbf{H}_k \cdot (\mathbf{g}_{k+1} - \mathbf{g}_k)] \otimes [\mathbf{H}_k \cdot (\mathbf{g}_{k+1} - \mathbf{g}_k)]}{(\mathbf{g}_{k+1} - \mathbf{g}_k) \cdot \mathbf{H}_k \cdot (\mathbf{g}_{k+1} - \mathbf{g}_k)} \quad (5.21)$$

The symbol \otimes when interposed between two vectors means that a matrix is to be formed. The ij th element of the matrix $\mathbf{u} \otimes \mathbf{v}$ is obtained by multiplying \mathbf{u}_i by \mathbf{v}_j .

The BFGS formula differs from the DFP equation by an additional term:

$$\begin{aligned} \mathbf{H}_{k+1} = \mathbf{H}_k &+ \frac{(\mathbf{x}_{k+1} - \mathbf{x}_k) \otimes (\mathbf{x}_{k+1} - \mathbf{x}_k)}{(\mathbf{x}_{k+1} - \mathbf{x}_k) \cdot (\mathbf{g}_{k+1} - \mathbf{g}_k)} - \frac{[\mathbf{H}_k \cdot (\mathbf{g}_{k+1} - \mathbf{g}_k)] \otimes [\mathbf{H}_k \cdot (\mathbf{g}_{k+1} - \mathbf{g}_k)]}{(\mathbf{g}_{k+1} - \mathbf{g}_k) \cdot \mathbf{H}_k \cdot (\mathbf{g}_{k+1} - \mathbf{g}_k)} \\ &+ [(\mathbf{g}_{k+1} - \mathbf{g}_k) \cdot \mathbf{H}_k \cdot (\mathbf{g}_{k+1} - \mathbf{g}_k)] \mathbf{u} \otimes \mathbf{u} \end{aligned} \quad (5.22)$$

where

$$\mathbf{u} = \frac{(\mathbf{x}_{k+1} - \mathbf{x}_k)}{(\mathbf{x}_{k+1} - \mathbf{x}_k) \cdot (\mathbf{g}_{k+1} - \mathbf{g}_k)} - \frac{[\mathbf{H}_k \cdot (\mathbf{g}_{k+1} - \mathbf{g}_k)]}{(\mathbf{g}_{k+1} - \mathbf{g}_k) \cdot \mathbf{H}_k \cdot (\mathbf{g}_{k+1} - \mathbf{g}_k)} \quad (5.23)$$

The MS formula is:

$$\mathbf{H}_{k+1} = \mathbf{H}_k + \frac{[(\mathbf{x}_{k+1} - \mathbf{x}_k) - \mathbf{H}_k(\mathbf{g}_{k+1} - \mathbf{g}_k)] \otimes [(\mathbf{x}_{k+1} - \mathbf{x}_k) - \mathbf{H}_k(\mathbf{g}_{k+1} - \mathbf{g}_k)]}{[(\mathbf{x}_{k+1} - \mathbf{x}_k) - \mathbf{H}_k(\mathbf{g}_{k+1} - \mathbf{g}_k)] \cdot (\mathbf{g}_{k+1} - \mathbf{g}_k)} \quad (5.24)$$

All of these methods use just the new and current points to update the inverse Hessian. The default algorithm used in the Gaussian series of molecular orbital programs [Schlegel 1982] makes use of more of the previous points to construct the Hessian (and thence the inverse Hessian), giving better convergence properties. Another feature of this method is its use

of a quartic polynomial that is guaranteed to have just one local minimum in the line search. The DFP, BFGS and MS methods can also be used with numerical derivatives, but alternative approaches may be more effective under such circumstances.

The matrix \mathbf{H} is often initialised to the unit matrix \mathbf{I} . The performance of the quasi-Newton algorithms can be improved by using a better estimate of the inverse Hessian than just the unit matrix. The unit matrix gives no information about the bonding in the system, nor does it identify any coupling between the various degrees of freedom. For example, a molecular mechanics calculation can be used to provide an initial guess to \mathbf{H} prior to a quantum mechanical calculation. Alternatively the matrix can be obtained from a quantum mechanical calculation at a lower level of theory (e.g. semi-empirical or with a smaller basis set).

5.7 Which Minimisation Method Should I Use?

The choice of minimisation algorithm is dictated by a number of factors, including the storage and computational requirements, the relative speeds with which the various parts of the calculation can be performed, the availability of analytical derivatives and the robustness of the method. Thus, any method that requires the Hessian matrix to be stored (let alone its inverse calculated) may present memory problems when applied to systems containing thousands of atoms. Calculations on systems of this size are invariably performed using molecular mechanics, and so the steepest descents and the conjugate gradients methods are very popular here. For molecular mechanics calculations on small molecules, the Newton-Raphson method may be used, although this algorithm can have problems with structures that are far from a minimum. For this reason it is usual to perform a few steps of minimisation using a more robust method such as the simplex or steepest descents before applying the Newton-Raphson algorithm. Analytical expressions for both first and second derivatives are available for most of the terms found in common force fields.

The performance of the steepest descents and conjugate gradients methods is contrasted in the following example. A model of the antibiotic netropsin (Figure 5.12) bound to DNA was constructed using an automated docking program. This initial model was then subjected to two stages of minimisation. In the first stage, the aim was to produce a structure that did not

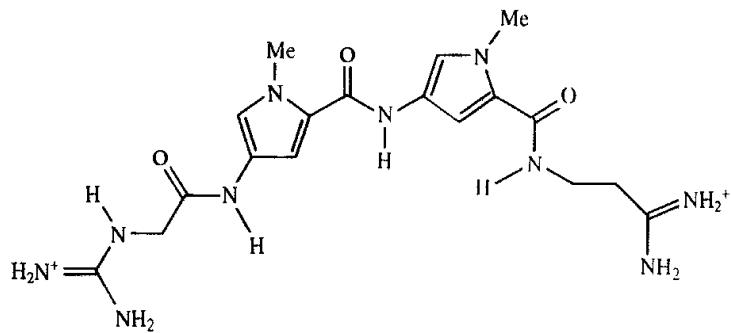


Fig 5.12: The DNA inhibitor netropsin

Method	Initial refinement (Av. gradient <1 kcal Å⁻²)		Stringent minimisation (Av. gradient <0.1 kcal Å⁻²)	
	CPU time (s)	Number of iterations	CPU time (s)	Number of iterations
Steepest descents	67	98	1405	1893
Conjugate gradients	149	213	257	367

Table 5.1 A comparison of the steepest descents and conjugate gradients methods for an initial refinement and a stringent minimisation.

have any significant high-energy interactions. The structure was then further minimised to give a structure much closer to the minimum. The results are shown in Table 5.1.

This study shows that the steepest descent method can actually be superior to conjugate gradients when the starting structure is some way from the minimum. However, conjugate gradients is much better once the initial strain has been removed.

Quantum mechanical calculations are restricted to systems with relatively small numbers of atoms, and so storing the Hessian matrix is not a problem. As the energy calculation is often the most time-consuming part of the calculation, it is desirable that the minimisation method chosen takes as few steps as possible to reach the minimum. For many levels of quantum mechanics theory analytical first derivatives are available. However, analytical second derivatives are only available for a few levels of theory and can be expensive to compute. The quasi-Newton methods are thus particularly popular for quantum mechanical calculations.

When using internal coordinates in a quantum mechanical minimisation it can be important to use an appropriate Z-matrix as input. For many systems the Z-matrix can often be written in many different ways as there are many combinations of internal coordinates. There should be no strong coupling between the coordinates. *Dummy atoms* can often help in the construction of an appropriate Z-matrix. A dummy atom is used solely to define the geometry and has no nuclear charge and no basis functions. A simple example of the use of dummy atoms is for a linear molecule such as HN_3 , where the angle of 180° would cause problems. The geometry of this molecule can be defined using a dummy atom as illustrated in Figure 5.13; the associated Z-matrix for this system would be:

1	N						
2	N	1	RN1N2				
3	X	1	1.0	2	90.0		
4	N	1	RN1N4	3	AN4N1X	2	180.0
5	H	4	RN4H	1	AHN4N1	3	180.0

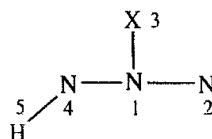


Fig 5.13 Internal coordinates of HN_3 molecule defined using dummy atom X

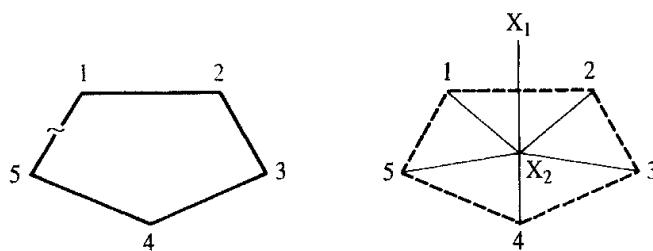


Fig. 5.14: The ring closure bond between atoms 1 and 5 would be strongly coupled to the other internal coordinates (left) unless dummy atoms are used to define the Z-matrix (right)

Strong coupling between coordinates can give long ‘valleys’ in the energy surface, which may also present problems. Care must be taken when defining the Z-matrix for cyclic systems in particular. The natural way to define a cyclic compound would be to number the atoms sequentially around the ring. However, this would then mean that the ring closure bond will be very strongly coupled to all of the other bonds, angles and torsion angles (Figure 5.14). A better definition uses a dummy atom placed at the centre of the ring (Figure 5.14). Some quantum mechanics programs are able to convert the input coordinates (be they Cartesian or internal) into the most efficient set for minimisation so removing from the user the problems of trying to decide what is an appropriate set of internal coordinates. For energy minimisations redundant internal coordinates have been shown to give significant improvements in efficiency compared with Cartesian coordinates or non-redundant internal coordinates, especially for flexible and polycyclic systems [Peng *et al.* 1996]. The redundant internal coordinates employed generally comprise the bond lengths, angles and torsion angles in the system. These methods obviously also require the means to interconvert between the internal coordinate representation and the Cartesian coordinates that are often used as input and desired as output. Of particular importance is the need to transform energy derivatives and the Hessian matrices (if appropriate).

5.7.1 Distinguishing Between Minima, Maxima and Saddle Points

A configuration at which all the first derivatives are zero need not necessarily be a minimum point; this condition holds at both maxima and saddle points as well. From simple calculus we know that the second derivative of a function of one variable, $f'(x)$ is positive at a minimum and negative at a maximum. It is necessary to calculate the eigenvalues of the Hessian matrix to distinguish between minima, maxima and saddle points. At a minimum point there will be six zero and $3N - 6$ positive eigenvalues if $3N$ Cartesian coordinates are used. The six zero eigenvalues correspond to the translational and rotational degrees of freedom of the molecule (thus these six zero eigenvalues are not obtained when internal coordinates are used). At a maximum point all eigenvalues are negative and at a saddle point one or more eigenvalues are negative. We will consider the uses of the eigenvalue and eigenvector information in Sections 5.8 and 5.9.

5.7.2 Convergence Criteria

In contrast to the simple analytical functions that we have used to illustrate the operation of the various minimisation methods, in ‘real’ molecular modelling applications it is rarely possible to identify the ‘exact’ location of minima and saddle points. We can only ever hope to find an approximation to the true minimum or saddle point. Unless instructed otherwise, most minimisation methods would keep going forever, moving ever closer to the minimum. It is therefore necessary to have some means to decide when the minimisation calculation is sufficiently close to the minimum and so can be terminated. Any calculation is of course limited by the precision with which numbers can be stored on the computer, but in most instances it is usual to stop well before this limit is reached. A simple strategy is to monitor the energy from one iteration to the next and to stop when the difference in energy between successive steps falls below a specified threshold. An alternative is to monitor the change in coordinates and to stop when the difference between successive configurations is sufficiently small. A third method is to calculate the root-mean-square gradient. This is obtained by adding the squares of the gradients of the energy with respect to the coordinates, dividing by the number of coordinates and taking the square root:

$$\text{RMS} = \sqrt{\frac{\mathbf{g}^T \mathbf{g}}{3N}} \quad (5.25)$$

It is also useful to monitor the maximum value of the gradient to ensure that the minimisation has properly relaxed all the degrees of freedom and has not left a large amount of strain in one or two coordinates.

5.8 Applications of Energy Minimisation

Energy minimisation is very widely used in molecular modelling and is an integral part of techniques such as conformational search procedures (Chapter 9). Energy minimisation is also used to prepare a system for other types of calculation. For example, energy minimisation may be used prior to a molecular dynamics or Monte Carlo simulation in order to relieve any unfavourable interactions in the initial configuration of the system. This is especially recommended for simulations of complex systems such as macromolecules or large molecular assemblies. In the following sections we will discuss some techniques that are specifically associated with energy minimisation methods.

5.8.1 Normal Mode Analysis

The molecular mechanics or quantum mechanics energy at an energy minimum corresponds to a hypothetical, motionless state at 0 K. Experimental measurements are made on molecules at a finite temperature when the molecules undergo translational, rotational and vibration motion. To compare the theoretical and experimental results it is

necessary to make appropriate corrections to allow for these motions. These corrections are calculated using standard statistical mechanics formulae. The internal energy $U(T)$ at a temperature T is given by:

$$U(T) = U_{\text{trans}}(T) + U_{\text{rot}}(T) + U_{\text{vib}}(T) + U_{\text{vib}}(0) \quad (5.26)$$

If all translational and rotational modes are fully accessible in accordance with the equipartition theorem, then $U_{\text{trans}}(T)$ and $U_{\text{rot}}(T)$ are both equal to $\frac{3}{2}k_B T$ per molecule (except that $U_{\text{rot}}(T)$ equals $k_B T$ for a linear molecule); k_B is Boltzmann's constant. However, the vibrational energy levels are often only partially excited at room temperature. The vibrational contribution to the internal energy at a temperature T thus requires knowledge of the actual vibrational frequencies. The vibrational contribution equals the difference in the vibrational enthalpy at the temperature T and at 0 K and is given by:

$$U_{\text{vib}}(T) = \sum_{i=1}^{N_{\text{nm}}} \left(\frac{h\nu_i}{2} + \frac{h\nu_i}{\exp[h\nu_i/k_B T] + 1} \right) \quad (5.27)$$

N_{nm} is the number of *normal vibrational modes* for the system. Even the zero-point energy ($U_{\text{vib}}(0)$, obtained by summing $\frac{1}{2}h\nu_i$ for each normal mode) can be quite substantial, amounting to about 100 kcal/mol for a six-carbon alkane. Other thermodynamic quantities such as entropies and free energies may also be calculated from the vibrational frequencies using the relevant statistical mechanics expressions.

Normal modes are useful because they correspond to collective motions of the atoms in a coupled system that can be individually excited. The three normal modes of water are schematically illustrated in Figure 5.15; a non-linear molecule with N atoms has $3N - 6$ normal modes. The frequencies of the normal modes together with the displacements of the individual atoms may be calculated from a molecular mechanics force field or from the wavefunction using the Hessian matrix of second derivatives (\mathcal{V}''). Of course, if we have used an appropriate minimisation algorithm then we already know the Hessian. The Hessian must first be converted to the equivalent force-constant matrix in *mass-weighted coordinates* (F), as follows.

$$F = M^{-1/2} \mathcal{V}'' M^{-1/2} \quad (5.28)$$

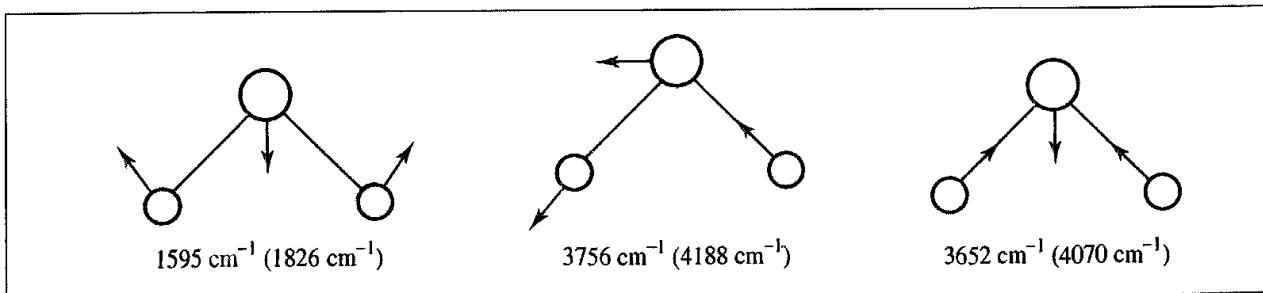


Fig. 5.15 Normal modes of water. Experimental and (calculated) frequencies are shown. Theoretical frequencies calculated using a 6-31G* basis set

M is a diagonal matrix of dimension $3N \times 3N$, containing the atomic masses. All elements of M are zero except those on the diagonal; $M_{1,1} = m_1$, $M_{2,2} = m_1$, $M_{3,3} = m_1$, $M_{4,4} = m_2, \dots, M_{3N-2,3N-2} = m_N$, $M_{3N-1,3N-1} = m_N$, $M_{3N,3N} = m_N$. Each non-zero element of $M^{-1/2}$ is thus the inverse square root of the mass of the appropriate atom. The masses of the atoms must be taken into account because a force of a given magnitude will have a different effect upon a larger mass than a smaller one. For example, the force constant for a bond to a deuterium atom is, to a good approximation, the same as to a proton, yet the different mass of the deuteron gives a different motion and a different zero-point energy. The use of mass-weighted coordinates takes care of these problems.

We next solve the secular equation $|\mathbf{F} - \mathbf{I}| = 0$ to obtain the eigenvalues and eigenvectors of the matrix \mathbf{F} . This step is usually performed using matrix diagonalisation, as outlined in Section 1.10.3. If the Hessian is defined in terms of Cartesian coordinates then six of these eigenvalues will be zero as they correspond to translational and rotational motion of the entire system. The frequency of each normal mode is then calculated from the eigenvalues using the relationship:

$$\nu_i = \frac{\sqrt{\lambda_i}}{2\pi} \quad (5.29)$$

As a simple example of a normal mode calculation consider the linear triatomic system in Figure 5.16. We shall just consider motion along the long axis of the molecule. The displacements of the atoms from their equilibrium positions along this axis are denoted by ξ_i . It is assumed that the displacements are small compared with the equilibrium values l_0 and the system obeys Hooke's law with bond force constants k . The potential energy is given by:

$$\mathcal{V} = \frac{1}{2}k(\xi_1 - \xi_2)^2 + \frac{1}{2}k(\xi_2 - \xi_3)^2 \quad (5.30)$$

We next calculate the first and then the second derivatives of the potential energy with respect to the three coordinates ξ_1 , ξ_2 and ξ_3 :

$$\frac{\partial \mathcal{V}}{\partial \xi_1} = k(\xi_1 - \xi_2); \quad \frac{\partial \mathcal{V}}{\partial \xi_2} = -k(\xi_1 - \xi_2) + k(\xi_2 - \xi_3); \quad \frac{\partial \mathcal{V}}{\partial \xi_3} = -k(\xi_2 - \xi_3) \quad (5.31)$$

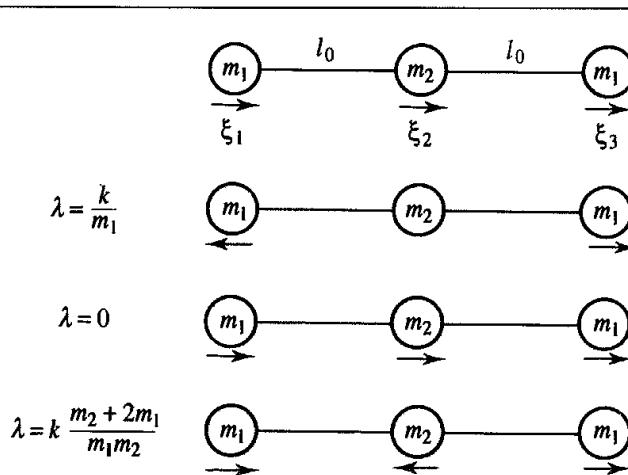


Fig 5.16. Linear three-atom system with results of normal mode calculation

The second derivatives are conveniently represented as a 3×3 matrix:

$$\begin{vmatrix} k & -k & 0 \\ -k & 2k & -k \\ 0 & -k & k \end{vmatrix} \quad (5.32)$$

The mass-weighted matrix is

$$\begin{vmatrix} m_1 & 0 & 0 \\ 0 & m_2 & 0 \\ 0 & 0 & m_3 \end{vmatrix} \quad (5.33)$$

The secular equation to be solved is thus:

$$\begin{vmatrix} \frac{k}{m_1} - \lambda & -\frac{k}{\sqrt{m_1}\sqrt{m_2}} & 0 \\ -\frac{k}{\sqrt{m_1}\sqrt{m_2}} & \frac{2k}{m_2} - \lambda & -\frac{k}{\sqrt{m_1}\sqrt{m_2}} \\ 0 & -\frac{k}{\sqrt{m_1}\sqrt{m_2}} & \frac{k}{m_1} - \lambda \end{vmatrix} = 0 \quad (5.34)$$

This determinant leads to a cubic in λ which has three roots (λ_k), each corresponding to a different mode of motion:

$$\lambda = \frac{k}{m_1}, \quad \lambda = 0, \quad \lambda = k \frac{m_2 + 2m_1}{m_1 m_2} \quad (5.35)$$

The corresponding frequencies can be obtained from Equation (5.29). The amplitudes (A) of each normal mode are given by the eigenvector solutions of the secular equation $\mathbf{F}\mathbf{A} = \lambda\mathbf{A}$. If A_1 , A_2 and A_3 are the amplitudes of each atom then the amplitudes obtained for each eigenvalue are:

$$\lambda = \frac{k}{m_1} : \quad A_1 = -A_3; \quad A_2 = 0 \quad (5.36)$$

$$\lambda = 0: \quad A_1 = A_3; \quad A_2 = \sqrt{\frac{m_2}{m_1}} A_1 \quad (5.37)$$

$$\lambda = k \frac{m_2 + 2m_1}{m_1 m_2} : \quad A_1 = A_3; \quad A_2 = -2\sqrt{\frac{m_1}{m_2}} A_1 \quad (5.38)$$

These normal modes are schematically illustrated in Figure 5.16. They correspond to a symmetric stretch, a translation and an asymmetric stretch respectively.

We have already seen how the results of normal mode calculations can be used to calculate thermodynamic quantities. The frequencies themselves can also be compared with the results of spectroscopic experiments, information which can be used in the parametrisation of a force field. For example, the experimental frequencies for the normal modes of water are shown in Figure 5.15, together with the frequencies determined using a 6-31G* *ab initio* calculation. The calculated values clearly deviate from those obtained experimentally, but the ratio of the experimental and theoretical frequencies is

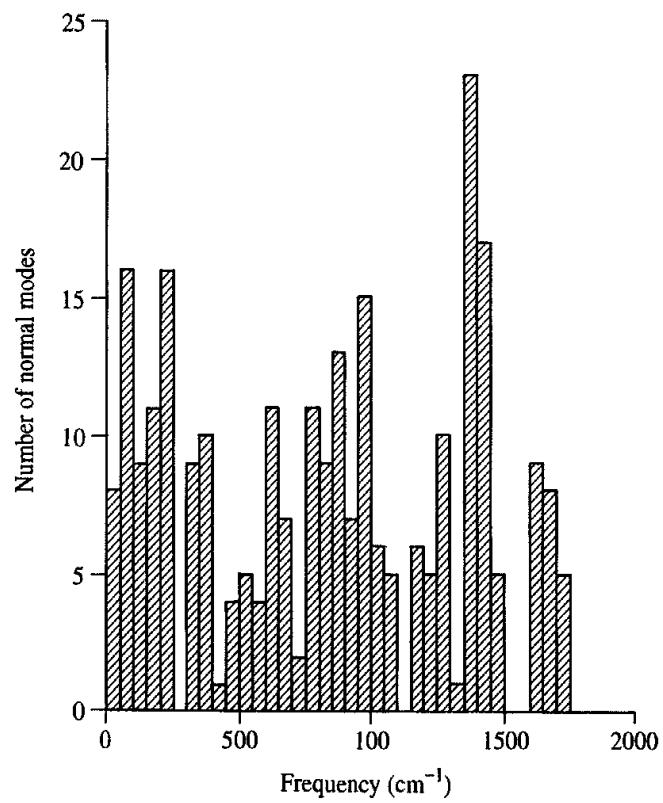
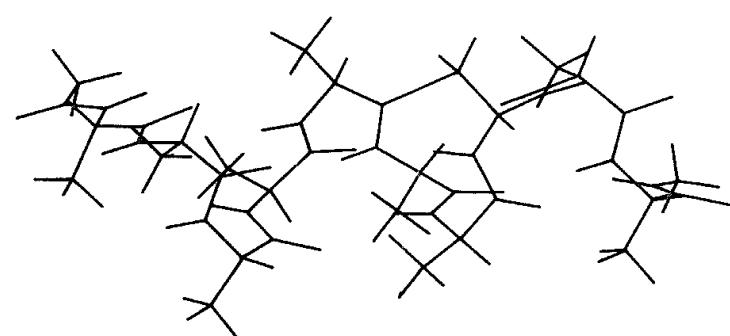


Fig. 5.17. Histogram of the normal modes calculated for a polyalanine polypeptide in an α -helical conformation. The height of each bar indicates the number of normal modes in each 50 cm^{-1} section.

remarkably consistent (at about 1.1). Such empirical scaling factors have been derived which enable frequencies obtained using a given level of theory to be converted to values for experiment or a higher level of theory [Pople *et al.* 1993]. The normal modes of much larger molecules can be calculated using molecular mechanics. For example, the vibrations of a helical polypeptide constructed from a sequence of ten alanine residues (112 atoms) are shown in Figure 5.17. In such cases it is usually the low-frequency vibrations that are of most interest as these correspond to the large-scale conformational motions of the molecule. The results of such analyses can be compared with molecular dynamics simulations from which vibrational contributions can also be extracted [Brooks and Karplus 1983].

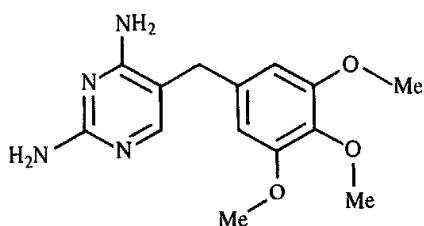


Fig. 5.18: Trimethoprim.

A normal mode calculation is based upon the assumption that the energy surface is quadratic in the vicinity of the energy minimum (the harmonic approximation). Deviations from the harmonic model can require corrections to calculated thermodynamic properties. One way to estimate anharmonic corrections is to calculate a force constant matrix using the atomic motions obtained from a molecular dynamics simulation; such simulations are not restricted to movements on a harmonic energy surface. The eigenvalues and eigenvectors are then calculated for this quasi-harmonic force-constant matrix in the normal way, giving a model which implicitly incorporates the anharmonic effects.

The harmonic approximation to the energy surface is found to be appropriate for well-defined energy minima such as the intramolecular degrees of freedom of small molecules and for some small intermolecular complexes. For larger systems such as liquids and large, 'floppy' molecules, the harmonic approximation breaks down. Such systems also have an extraordinarily large number of 'minima' on the energy surface. In such cases it is not possible to calculate accurately thermodynamic properties using energy minimisation and normal mode calculations. Rather, molecular dynamics or Monte Carlo simulations must be used to sample the energy surface from which properties can be derived, as we will discuss in Chapters 6–8.

5.8.2 The Study of Intermolecular Processes

One example of the use of minimisation methods and normal-mode analysis is the study by Hagler and co-workers of the binding of the antibacterial drug trimethoprim (Figure 5.18) to the enzyme dihydrofolate reductase (DHFR) [Dauber-Osguthorpe *et al.* 1988; Fisher *et al.* 1991]. DHFR catalyses the reduction of folic acid and dihydrofolic acid to tetrahydrofolic acid (Figure 5.19) and plays a vital metabolic role in the biosynthesis of nucleic acids in bacteria, protozoa, plants and animals. Trimethoprim exploits the structural differences between bacterial and vertebrate DHFR, binding much more strongly to the former, and is clinically used as an antibacterial agent. Inhibitors of human DHFR are used in cancer therapy. Hagler and colleagues applied energy minimisation to an isolated trimethoprim molecule, to the crystal structure of trimethoprim, to trimethoprim in the presence of water molecules, and to trimethoprim in intermolecular complexes with DHFR from both bacterial and vertebrate sources. An important observation was that the conformation of the trimethoprim, when bound to the enzyme, was significantly different from that obtained for the isolated molecule. This reinforces the view that the use of

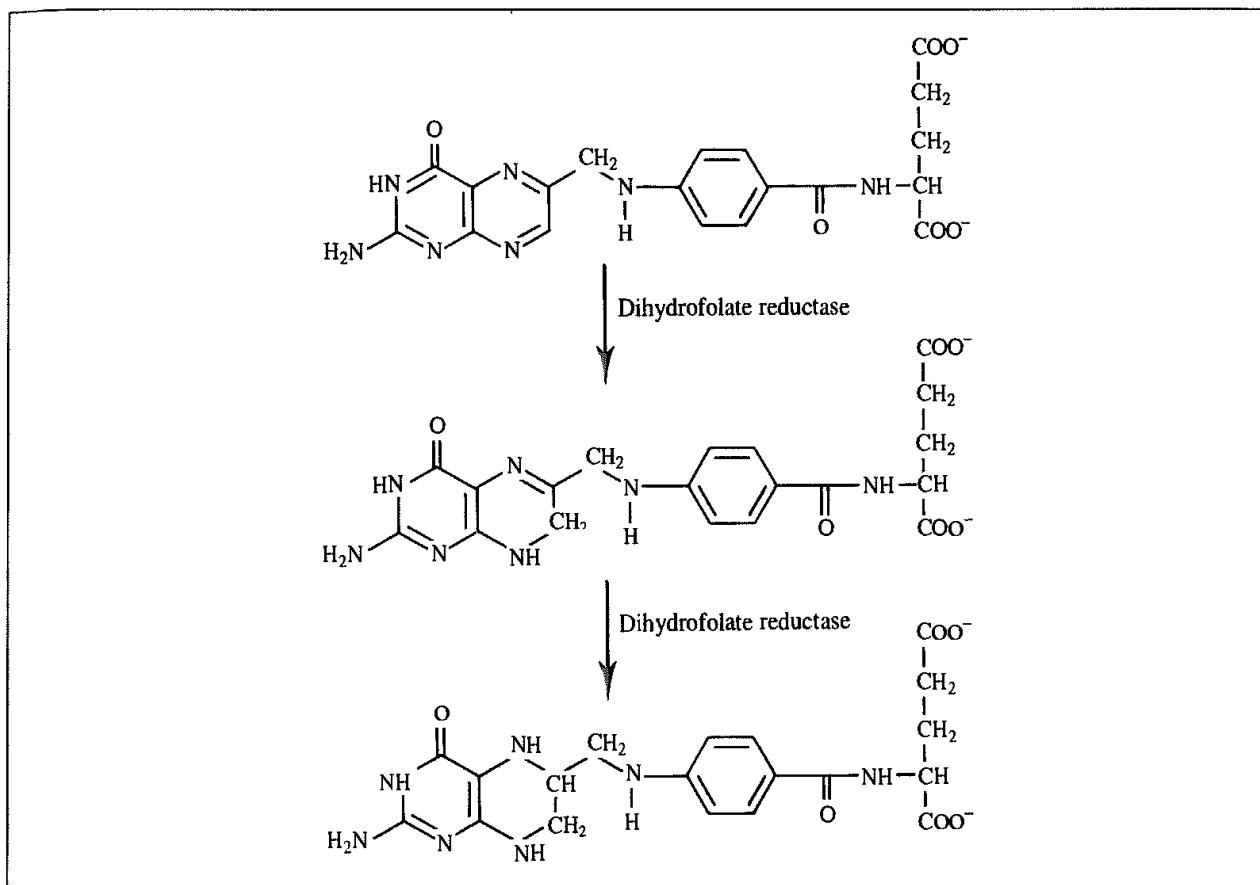


Fig 5.19: DHFR catalyses the reduction of folic acid to tetrahydrofolic acid.

structures obtained from energy minimisation calculations on isolated molecules can lead to misleading conclusions. Intermolecular interactions with the receptor enable the ligand to adopt a conformation whose intramolecular energy is significantly higher than any of its minimum energy structures.

A normal mode analysis on the isolated and bound trimethoprim molecules enabled an estimate to be made of the entropic contribution to binding. Low-frequency modes for the isolated ligand were found to be shifted to higher frequencies for the ligand in the enzyme complex, reflecting a restriction of the motion of the ligand by the protein. This entropic contribution to the free energy of binding was predicted to be quite significant, indicating that conclusions based solely upon energies may be misleading.

5.9 Determination of Transition Structures and Reaction Pathways

Chemists are interested not only in the thermodynamics of a process (the relative stability of the various species) but also in its kinetics (the rate of conversion from one structure to another). Knowledge of the minimum points on an energy surface enables thermodynamic data to be interpreted, but for the kinetics it is necessary to investigate the nature of the

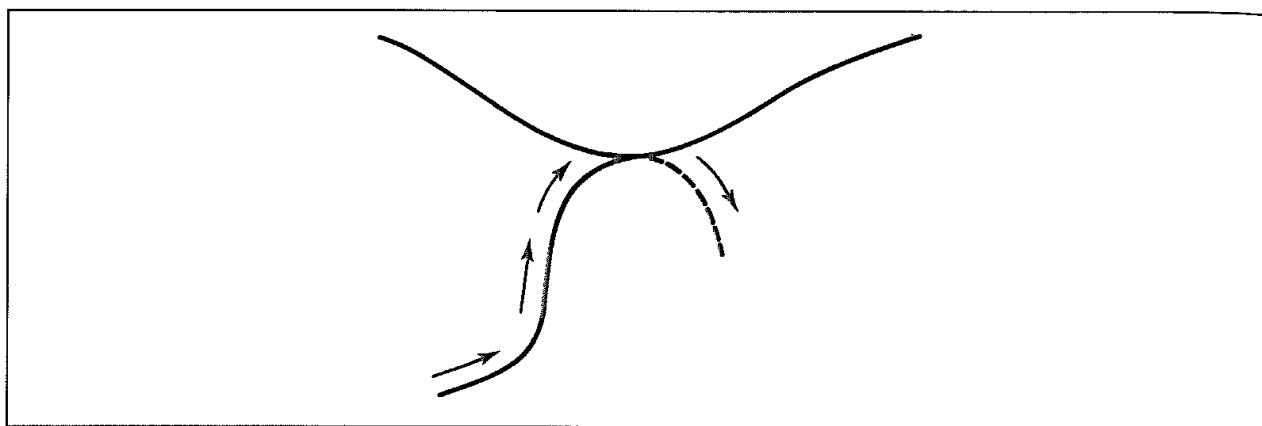


Fig. 5.20: The lowest-energy path from one minimum to another passes through a saddle point

energy surface away from the minimum points. In particular, we would like to know how the system changes from one minimum to another. What changes in geometry are involved, and how does the energy vary during the transition? The minimum points on the energy surface may be the reactants and products of a chemical reaction, two conformations of a molecule, or two molecules that associate to form a non-covalently bound bimolecular complex. We shall use the term 'reaction pathway' to describe the path between two minima, but our use of the word 'reaction' does not necessarily mean that bond making and/or breaking is involved. Many methods have been proposed for finding transition structures and elucidating reaction pathways. We do not have space to cover all of the methods, and so we shall restrict our discussion to some of the more common approaches.

As a system moves from one minimum to another, the energy increases to a maximum at the transition structure and then falls. At a saddle point the first derivatives of the potential function with respect to the coordinates are all zero (just as they are at a minimum point). The number of negative eigenvalues in the Hessian matrix is used to distinguish different types of saddle point; an n th-order transition or saddle point has n negative eigenvalues. We are usually most interested in first-order saddle points, where the energy passes through a maximum for movement along the pathway that connects the two minima, but is a minimum for displacements in all other directions perpendicular to the path. This is shown schematically for a two-dimensional energy surface in Figure 5.20.

These negative eigenvalues of the Hessian matrix are often referred to as the 'imaginary' frequencies for motion of the system over the saddle point. We can illustrate this concept using the gas-phase S_N2 reaction between Cl⁻ and CH₃Cl. As the chloride ion approaches the methyl chloride along the line of the C–Cl bond the energy passes through an ion-dipole complex which is at an energy minimum. The energy then rises to a maximum at the pentagonal transition state. The energy profile is drawn in Figure 5.21. The geometries of the minimum and the pentagonal transition state, as determined by an *ab initio* HF/SCF calculation with the 6-31G* basis set are shown in Figure 5.22. The lowest-frequency eigenvalues and a representation of the corresponding eigenvectors for the two geometries are also given in Figure 5.22. There are three frequencies in the ion-dipole minimum that are of particularly low energy; two of these correspond to degenerate 'wagging' motions of the

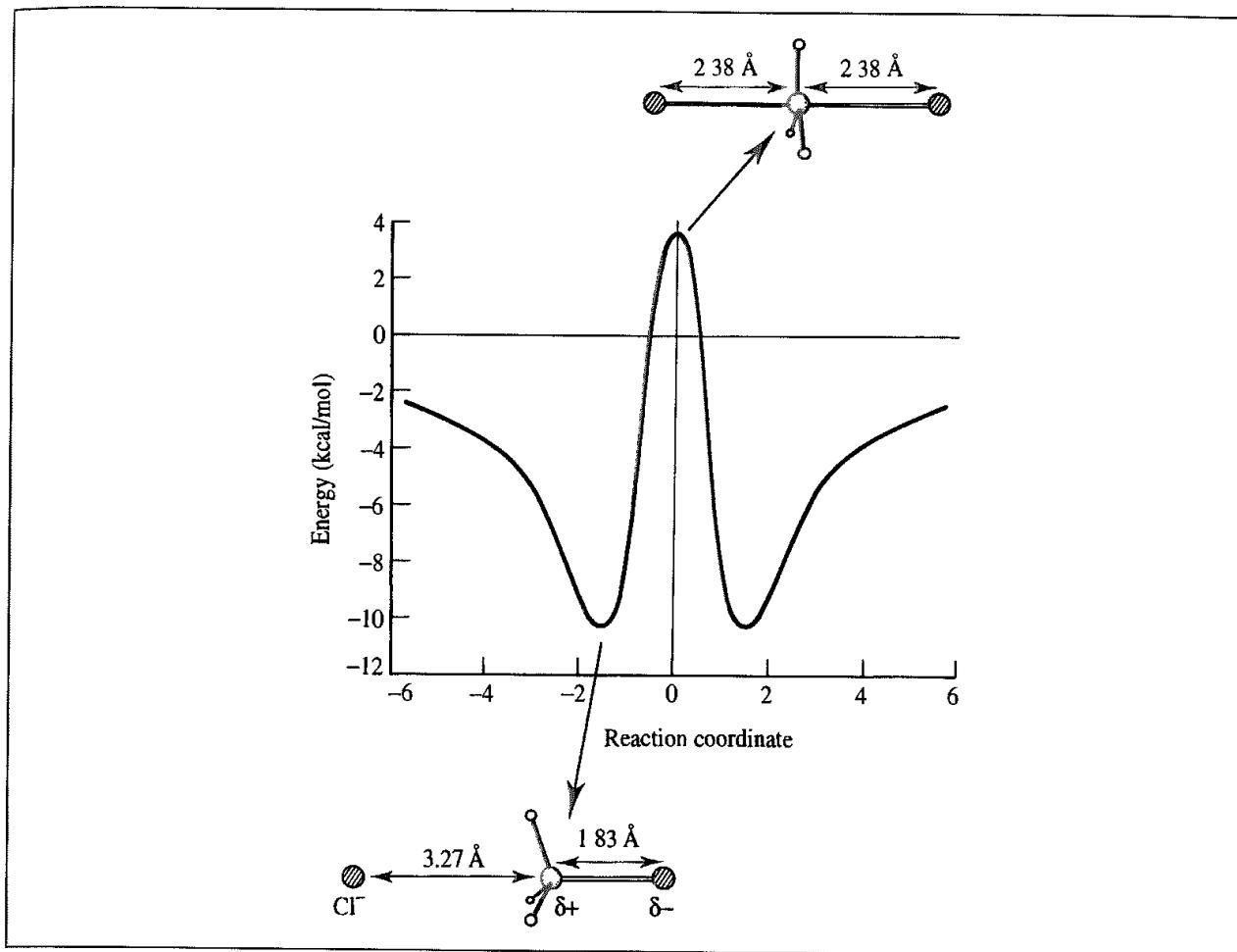


Fig 5.21. The energy profile for the gas-phase $\text{Cl}^- + \text{MeCl}$ reaction (Adapted in part from Chandrasekhar J, S F Smith and W L Jorgensen 1985 Theoretical Examination of the S_N2 Reaction Involving Chloride Ion and Methyl Chloride in the Gas Phase and Aqueous Solution Journal of the American Chemical Society 107 154–163.)

system (at 71.3 cm^{-1}). The vibration at 101.0 cm^{-1} is the normal mode that corresponds to motion towards the transition state. At the saddle point there is a single negative eigenvalue (with an imaginary ‘frequency’ of -415.0 cm^{-1}) that corresponds to vibration along the $\text{Cl}-\text{C}-\text{Cl}$ axis (i.e. motion along the reaction pathway). The other normal modes at the saddle point all have positive frequencies; the two lowest (at 204.2 cm^{-1}) correspond to wagging motions perpendicular to the $\text{Cl}-\text{C}-\text{Cl}$ axis and the third is a symmetric stretch of the two chlorine atoms along the symmetry axis.

It is important to distinguish the transition *structure* from the transition *state*. The transition structure is the point of highest potential energy along the pathway. By contrast, the transition state is the geometry at the peak in the free energy profile. In many cases the geometry at the transition state is very similar to that of the transition structure. However, the transition state may be different as the free energy of activation includes contributions from sources other than just the potential energy. If the transition state geometry is temperature-dependent then entropic factors may be important. An example is the following radical

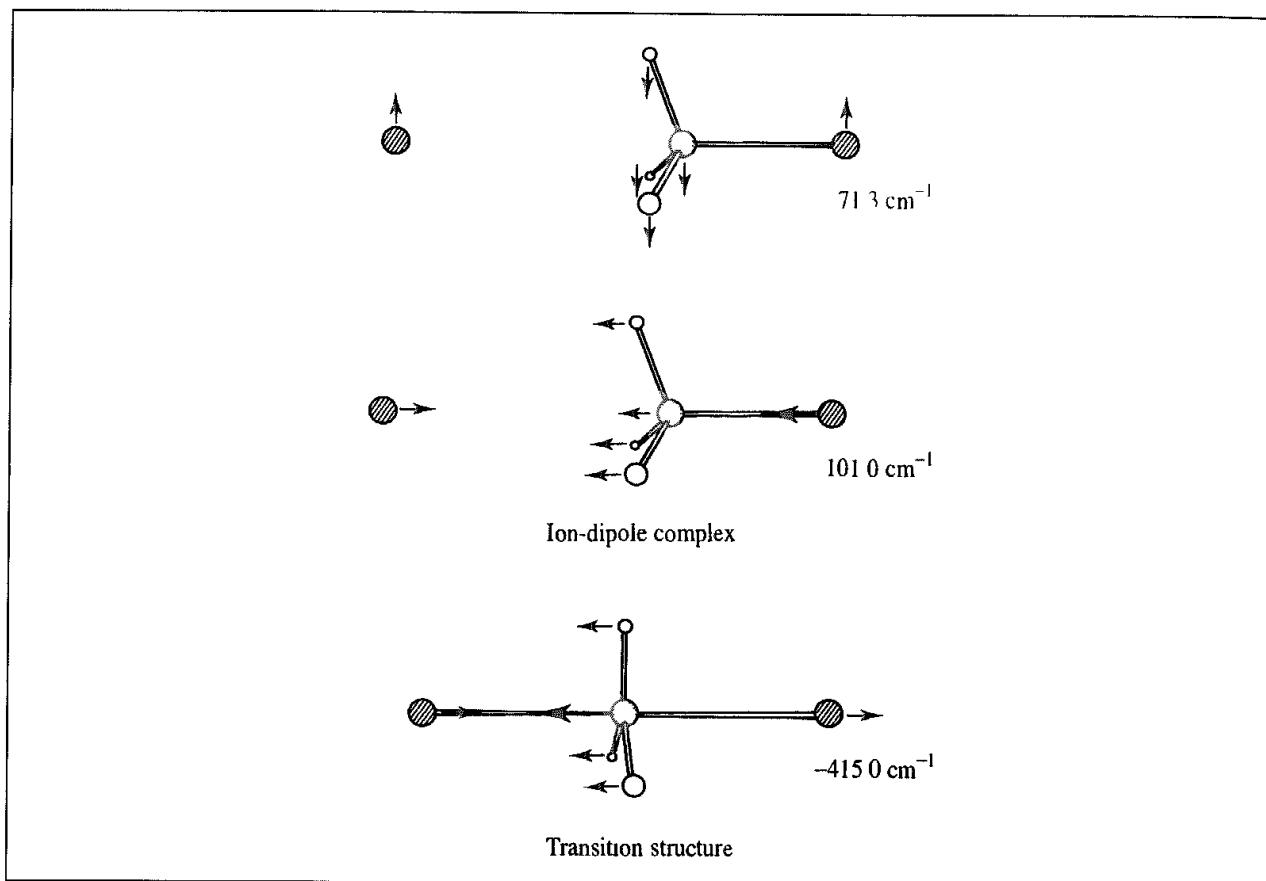
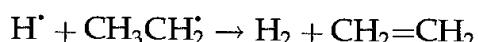


Fig 5.22 Schematic representation of some of the lower frequencies in the ion-dipole complex for the $\text{Cl}^- + \text{MeCl}$ reaction and the imaginary frequency of the transition structure, calculated using a 6-31G* basis set.

reaction:



The calculated geometry of the transition structure resembles the ethyl radical (Figure 5.23) [Doubleday *et al.* 1985]. The entropy change for this reaction is negative and so, as the temperature is increased, the maximum in the free energy profile shifts more towards the products, in the direction of lower entropy.

Methods for finding transition structures and reaction pathways are often closely related. Thus, some methods for finding the reaction pathway start from the transition structure and move down towards a minimum. Such methods must be supplied with the transition structure geometry as the starting point. Conversely, some methods for locating transition structures do so by searching along the reaction pathway, or an approximation to it. Yet other methods require neither the transition structure nor the pathway, but can determine both simultaneously from the two minima. In general, it is more difficult to locate transition structures and determine reaction pathways than to find minimum points. It is therefore crucial to check that the Hessian matrix at any proposed saddle point has the required single negative eigenvalue. Methods for locating saddle points are usually most effective when given as input a geometry that is as close as possible to the transition structure. It

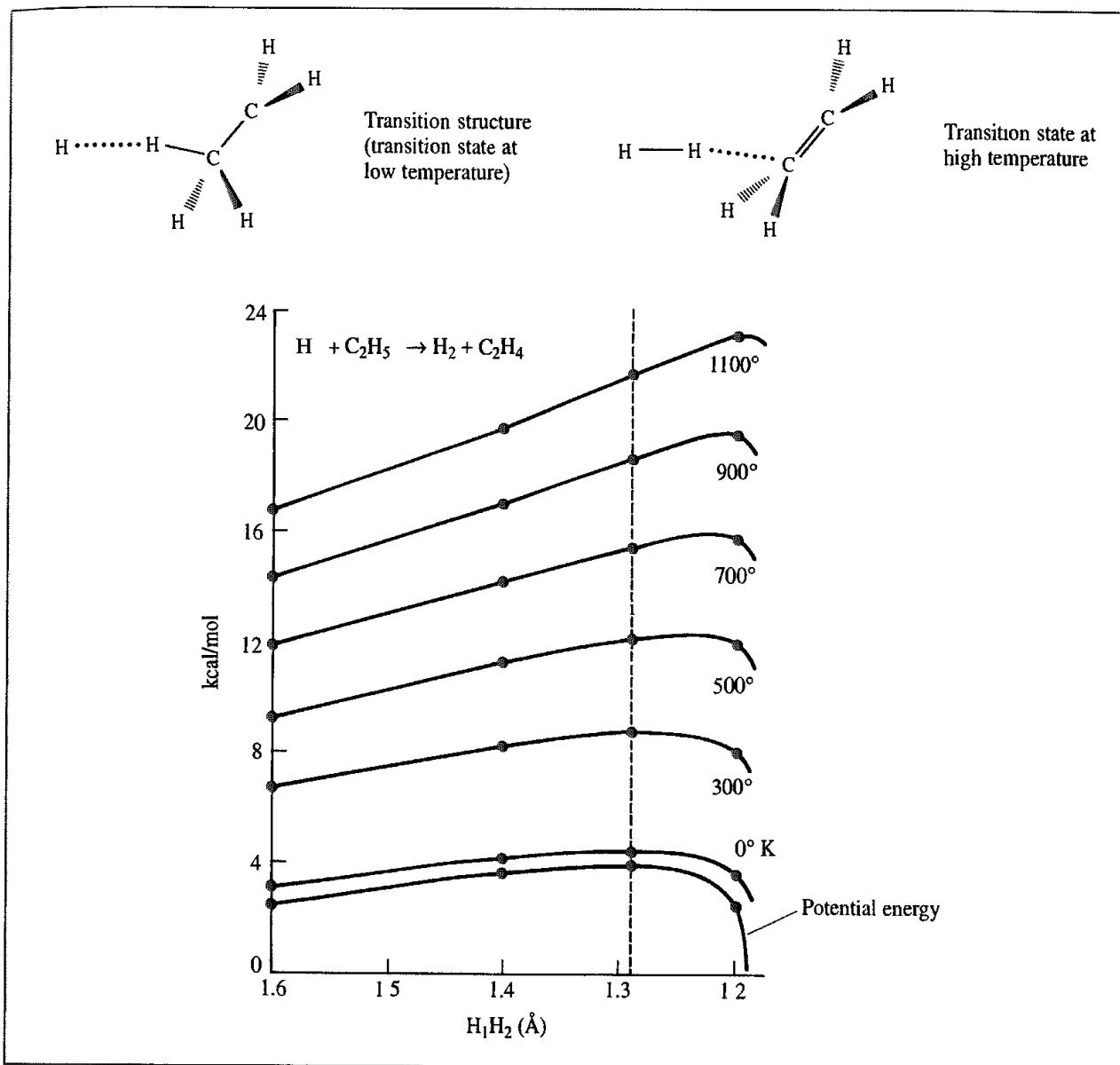


Fig 5.23 The transition structure for the $H\cdot + CH_3CH_2\cdot \rightarrow H_2 + CH_2=CH_2$ reaction. At low temperature the transition structure corresponds to the transition state (maximum of free energy). At high temperature the transition state moves closer to the products, as can be seen from the graph. (Redrawn from Doubleday C, J McIver, M Page and T Zielinski 1985. Temperature Dependence of the Transition-State Structure for the Disproportionation of Hydrogen Atom with Ethyl Radical. Journal of the American Chemical Society 107:5800–5801)

can also be helpful to examine the atomic displacements that correspond to the negative eigenvector, to ensure that it corresponds to the correct motion over the saddle point as for the $Cl^- + CH_3Cl$ reaction.

As one approaches the saddle point from a minimum, the Hessian matrix will change from having all positive eigenvalues to including one negative value. The *quadratic region* of a saddle point is that portion of the energy surface surrounding the point where the Hessian contains one negative eigenvalue. Similarly the quadratic region of a minimum is the

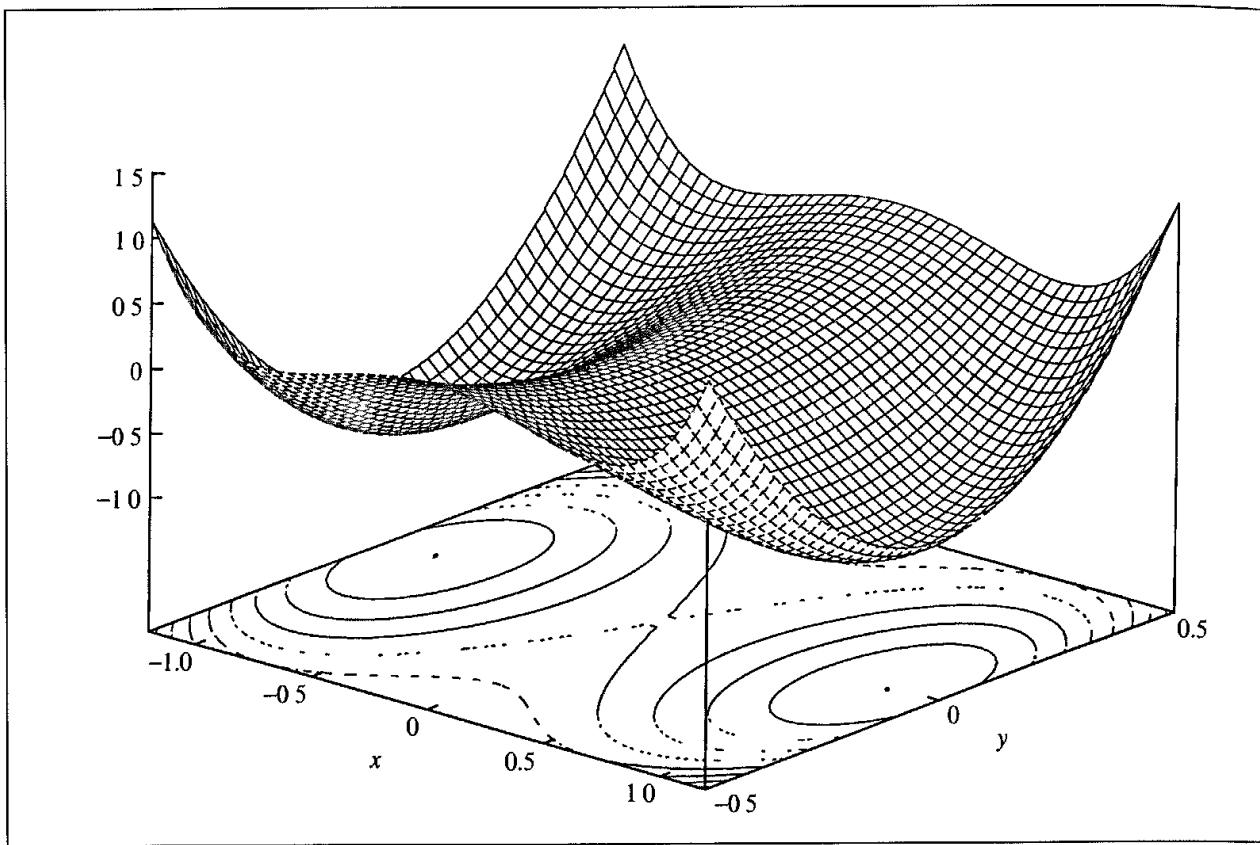


Fig. 5.24. The function $f(x,y) = x^4 + 4x^2y^2 - 2x^2 + 2y^2$ has a saddle point at $(0,0)$ and minima at $(1,0)$ and $(-1,0)$

region where all eigenvalues are positive and the Hessian is positive definite. Some algorithms for finding saddle points require a starting geometry within the quadratic region. We can illustrate the concept of a quadratic region by considering the function $f(x,y) = x^4 + 4x^2y^2 - 2x^2 + 2y^2$, which is drawn in Figure 5.24. This function has two minima at $(1,0)$ and $(-1,0)$ and one saddle point at $(0,0)$. In this case it is possible to derive and characterise the stationary points analytically. The Hessian matrix of second derivatives for this function is:

$$\begin{pmatrix} 12x^2 + 8y^2 - 4 & 16xy \\ 16xy & 8x^2 + 4 \end{pmatrix} \quad (5.39)$$

At the point $(1,0)$ the Hessian matrix is thus

$$\begin{pmatrix} 8 & 0 \\ 0 & 4 \end{pmatrix} \quad (5.40)$$

The eigenvalues of this matrix are obtained by setting the secular determinant to zero:

$$\begin{vmatrix} 8 - \lambda & 0 \\ 0 & 4 - \lambda \end{vmatrix} = 0 \quad (5.41)$$

The eigenvalues are $\lambda = 4$ and $\lambda = 8$. Thus both eigenvalues are positive and the point is a minimum. At the point $(0, 0)$ the Hessian matrix is

$$\begin{pmatrix} -4 & 0 \\ 0 & 4 \end{pmatrix} \quad (5.42)$$

with one negative and one positive eigenvalue (-4 and $+4$). The normalised eigenvectors corresponding to these eigenvalues are $(0, 1)$ for the eigenvalue $\lambda = 4$ and $(1, 0)$ for the eigenvalue $\lambda = -4$. These eigenvectors indicate the directions in which the gradient of the function changes sign. Thus along the line $x = 0$ the function passes through a minimum, as can be seen from Figure 5.24. By contrast, if one progresses from $(-1, 0)$ to $(1, 0)$ through the origin then the function passes through a maximum. As one progresses through a transition structure the eigenvector of the negative eigenvalue corresponds to the concerted motions of the atoms that give rise to motion through the saddle point. If we move along the x axis from the minimum at $(1, 0)$ to the saddle point at the origin, both eigenvalues will be positive so long as $12x^2 + 8y^2 - 4 > 0$. Thus, so long as x is larger than $1/\sqrt{3}$ the eigenvalues of the Hessian matrix will be positive. When x becomes smaller than $1/\sqrt{3}$ there will be one negative and one positive eigenvalue. In this case the quadratic region would correspond to all points where the absolute value of x was less than $1/\sqrt{3}$.

5.9.1 Methods to Locate Saddle Points

In some simple cases such as the chloride/methyl chloride reaction the geometry of the transition structure can be predicted by inspection. In other cases a *grid search* can be used to scan the energy surface in order to locate the approximate position of the transition state. In a grid search, the coordinates are systematically varied to generate a set of structures, for each of which the energy is calculated. It may then be possible to fit an analytical expression to these points, from which the saddle point can be predicted by standard calculus methods. The grid search method is widely used for constructing potential energy surfaces but is restricted to systems with a very small number of atoms or where only a limited number of degrees of freedom are being explored such as the $\text{H} + \text{H}_2 \rightarrow \text{H}_2 + \text{H}$ reaction. An advantage of the grid search is that it does provide information about the energy surface away from the pathway, which can be important if one wishes to investigate the dynamics of a reaction and the interconversion of energy between different modes. The grid search method is not the method of choice for all but the smallest systems due to the number of energy evaluations that are required. In any case it does not directly provide the transition structure.

The conversion of one minimum-energy structure into another may sometimes occur primarily along just one or two coordinates. In such cases, an approximation to the reaction pathway can be obtained by gradually changing the coordinate(s), allowing the system to relax at each stage using minimisation while keeping the chosen coordinate(s) fixed. The point of highest energy on the path is an approximation to the saddle point and the structures generated during the course of the calculation can be considered to represent a sequence of points on the interconversion pathway. When such coordinate driving methods are applied to conformational changes that occur primarily via rotation about bonds, the

the Hessian matrix of second derivatives is available then the appropriate direction to take is uphill along the eigenvector of the smallest eigenvalue when all eigenvalues are positive and downhill along the eigenvector corresponding to the negative eigenvalue when within the quadratic region of the saddle point [Baker 1986].

As we have stated frequently, at a saddle point the gradient is zero (as it is for a minimum). It might therefore be imagined that a minimisation algorithm (or some variant) could be used to locate saddle points. Some minimisation algorithms can occasionally incorrectly converge to a saddle point, especially if the starting structure is close to the transition structure. A simple example is the Newton-Raphson method, which will converge to a transition structure when giving a starting position that is within the quadratic region. Other minimisation algorithms can also be modified so that they consistently locate saddle points when provided with an initial structure within the quadratic region [Schlegel 1982].

5.9.2 Reaction Path Following

The traditional way to elucidate the reaction path is to move downhill from a saddle point to the two associated minima. There may be many different paths that could be followed from the saddle point to the associated minima. The *intrinsic reaction coordinate* (IRC) is the path that would be followed by a particle moving along the steepest descents path with an infinitely small step from the transition structure down to each minimum when the system is described using mass-weighted coordinates (as in a normal mode calculation) [Fukui 1981]. The initial directions towards each minimum can be obtained directly from the eigenvector that corresponds to the imaginary frequency at the transition structure. A simple steepest descents algorithm with a reasonable step size will usually give a path that oscillates about the true minimum energy path, as illustrated in Figure 5.28. This is perfectly acceptable in a minimisation, where the objective is to locate the minimum as efficiently as possible and where we are not interested in the intermediate structures. To determine the true reaction pathway (or a better approximation to it) it is necessary to 'correct' the path taken by the steepest descents algorithm. These corrective methods are especially useful when the path is curved.

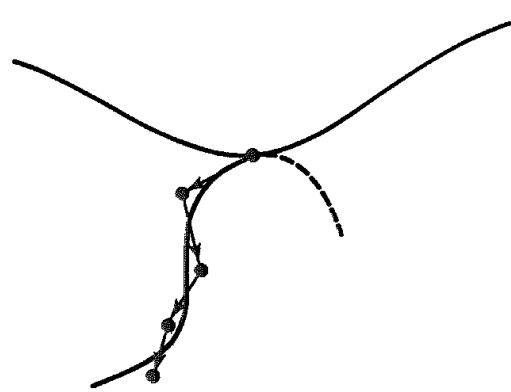


Fig 5.28 A steepest descents minimisation algorithm produces a path that oscillates about the true reaction pathway from the transition structure to a minimum

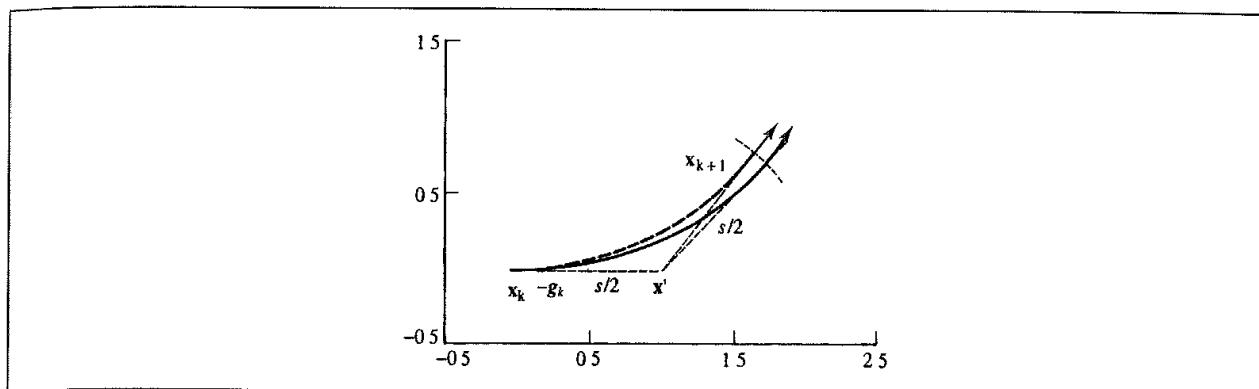


Fig 5.29: Method for correcting the path followed by a steepest descents algorithm to generate the intrinsic reaction coordinate. The solid line shows the real path and the dotted line shows the algorithmic approximation to it (Figure redrawn from Gonzalez C and H B Schlegel 1988 An Improved Algorithm for Reaction Path Following Journal of Chemical Physics **90** 2154-2161)

Many different algorithms have been suggested for determining reaction paths. The real challenge is to find an approach that is sufficiently general to work well in many (if not all) situations and with relatively little computational expense. One widely used method was devised by Gonzalez and Schlegel [Gonzalez and Schlegel 1988] and is illustrated in Figure 5.29. First it calculates the gradient at the current point, x_k . A step of length $s/2$ is taken along the direction of this gradient to give a new point (x'). The next point on the reaction path is obtained by minimising the energy subject to the constraint that the distance between x' and the new point on the reaction path (x_{k+1}) is $s/2$. The reaction path is then approximated by a circle that passes through both x_k and x_{k+1} and whose tangents at those two points are in the directions of the gradients. A refined version of this path-following algorithm has been incorporated into an efficient combined procedure which can determine reaction paths, minima and transition state geometries [Ayala and Schlegel 1997] without the need for second derivatives to be calculated.

5.9.3 Transition Structures and Reaction Pathways for Large Systems

Most of the algorithms we have discussed so far, with the possible exception of adiabatic mapping, were originally designed to be used with quantum mechanics where relatively small numbers of atoms are involved. It is often difficult to apply these methods to the study of conformational transitions. There are several reasons for this, but one important feature is that it is assumed that there is only one saddle point between the initial and final states. There may be a number of transition structures along the pathway between two conformations of a complex molecule. Here we will discuss two related methods that were originally designed to tackle this problem using molecular mechanics.

In the self-penalty walk (SPW) method of Czerninski and Elber [Czerninski and Elber 1990; Nowak *et al.* 1991] a ‘polymer’ is constructed that consists of a series of $M + 2$ ‘monomers’. Each monomer is a complete copy of the actual system and so there are $(M + 2)N$ atoms present in the calculation. The two ends of the polymer correspond to the two minima between which we are trying to elucidate the pathway (the ‘reactant’ and the ‘product’).

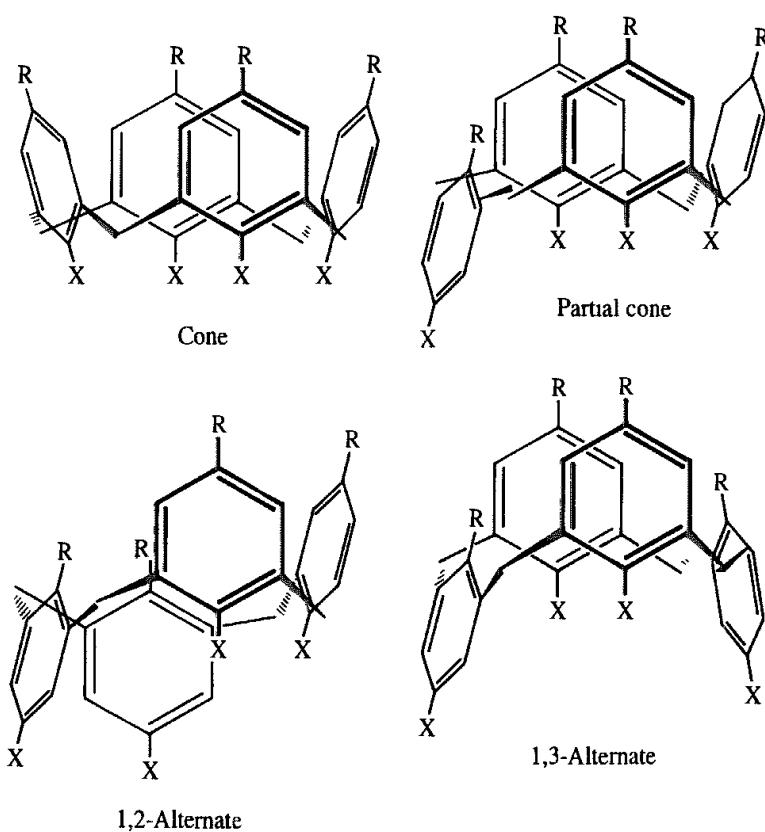


Fig. 5.31 Possible conformations of the calix[4]arene systems. (Figure adapted from Fischer S, P D J Groothuis, L C Groenen, W P van Hoorn, F C J M van Geggel, D N Reinhoudt and M Karplus 1995 Pathways for Conformational Interconversion of Calix[4]arenes Journal of the American Chemical Society 117:1611–1620)

involved, giving the energy diagram in Figure 5.32. The predicted activation barrier of 14.5 kcal/mol for the cone → inverted cone transition was in very good agreement with the experimentally determined value of 14.2 kcal/mol. Much of the barrier (9.1 kcal/mol) was due to the need to break two hydrogen bonds; the remainder was due to the need to deform some bond angles such as those of the bridging methylene carbons.

5.9.4 The Transition Structures of Pericyclic Reactions

One of the most celebrated examples of the use of quantum mechanical methods in understanding chemical reactivity is the work of Woodward and Hoffmann [Woodward and Hoffmann 1969] who were able to explain the experimentally observed nature of certain types of concerted reaction. The reactions which they studied include cycloadditions, sigma-tropic rearrangements, cheletropic reactions, electrocyclic reactions and the ene reaction (Figure 5.33) and are collectively known as pericyclic reactions. The products obtained from such reactions can be understood in terms of simple mechanistic arguments, but such arguments cannot explain some aspects. In particular, the reactions are often highly stereospecific with the reaction rates and the stereoselectivity changes dramatically with

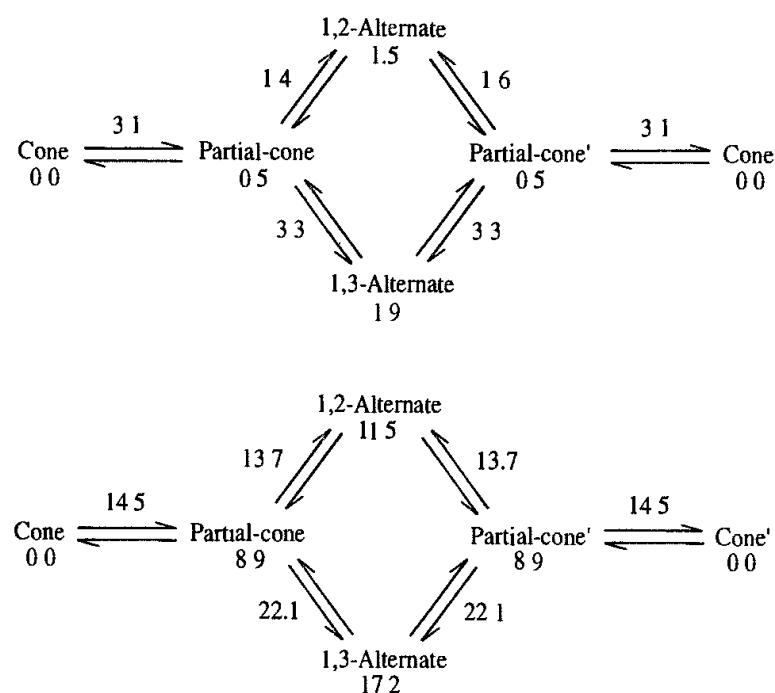


Fig 5.32 Interconversion between various conformations of calix[4]arenes $X = H, R = H$ (top); $X = OH, R = H$ (bottom). Energies in kcal/mol.

the reaction conditions. Woodward and Hoffmann successfully employed molecular orbital theory to rationalise the existing data and their theory has also been very successful in predicting the outcome of similar reactions. The basic principle applied by Woodward and Hoffmann was that of the conservation of orbital symmetry and as a consequence of their work a series of rules (often called the Woodward–Hoffmann rules) were developed. The Woodward–Hoffmann rules apply only to concerted reactions and are based upon the principle that maximum bonding is maintained throughout the course of a reaction. Fukui also discovered the importance of orbital symmetry and suggested that the majority of chemical reactions should take place at the position of, and in the direction of, maximum overlap between the highest occupied molecular orbital (HOMO) of one species and the lowest unoccupied molecular orbital (LUMO) of the other component [Fukui 1971]. These orbitals are collectively known as the *frontier orbitals*.

The HOMO–LUMO interaction depends on various factors, including the geometry of approach (which affects the amount of overlap), the phase relationship of the orbitals and their energy separation. For example, the HOMO and LUMO of ethene are illustrated pictorially in Figure 5.34. The most obvious mode of interaction between the two molecules involves suprafacial attack shown in Figure 5.34 to give cyclobutane. However, the symmetries of the overlapping orbitals must have the same phase for a favourable interaction to occur and this is not possible for ethene unless an energetically unfavourable antarafacial approach is adopted. By contrast, the interaction between ethene and the butadiene does occur in a suprafacial sense with both HOMO/LUMO pairs of orbitals having the appropriate phase relationship (Figure 5.34).

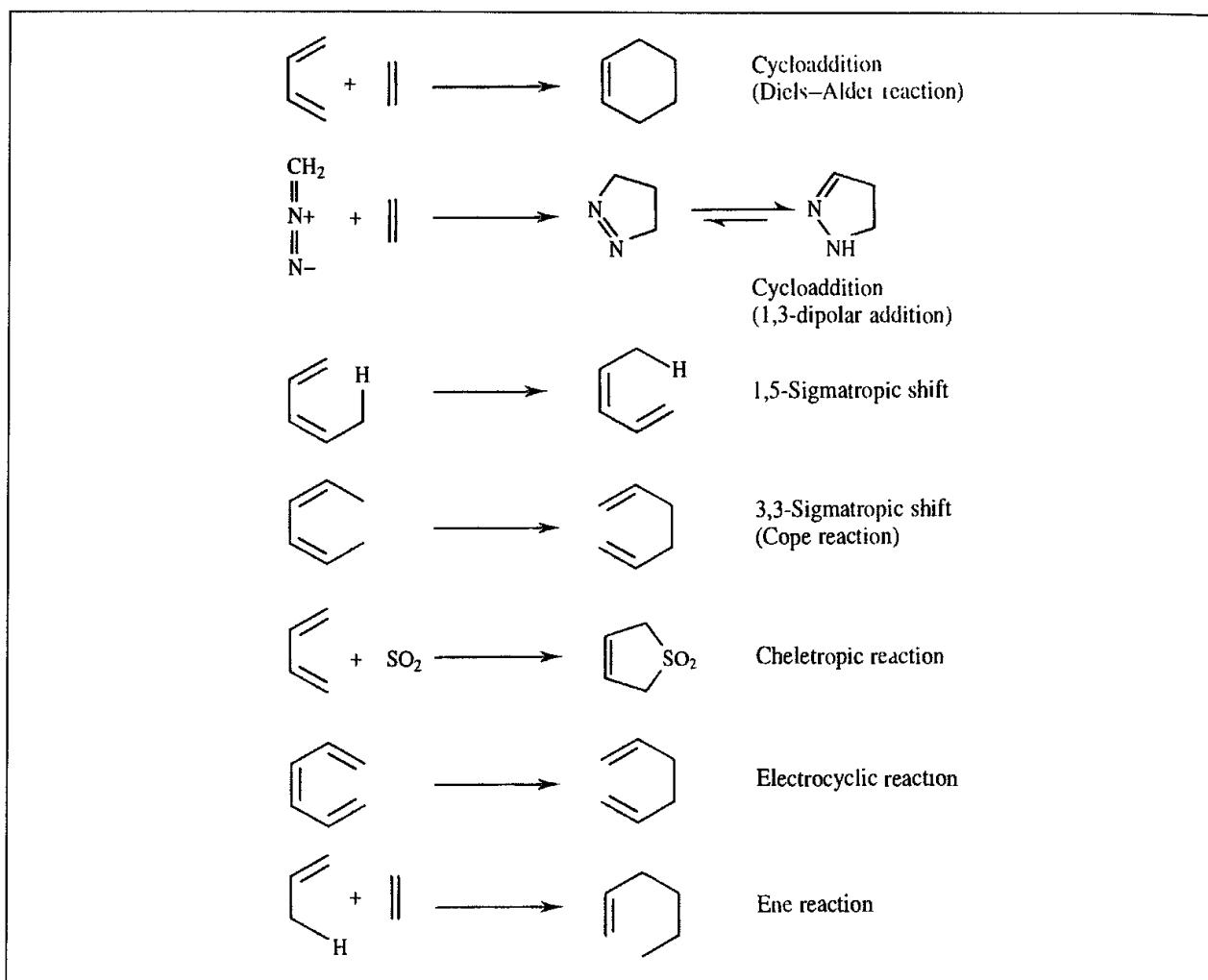


Fig. 5.33: Typical pericyclic reactions.

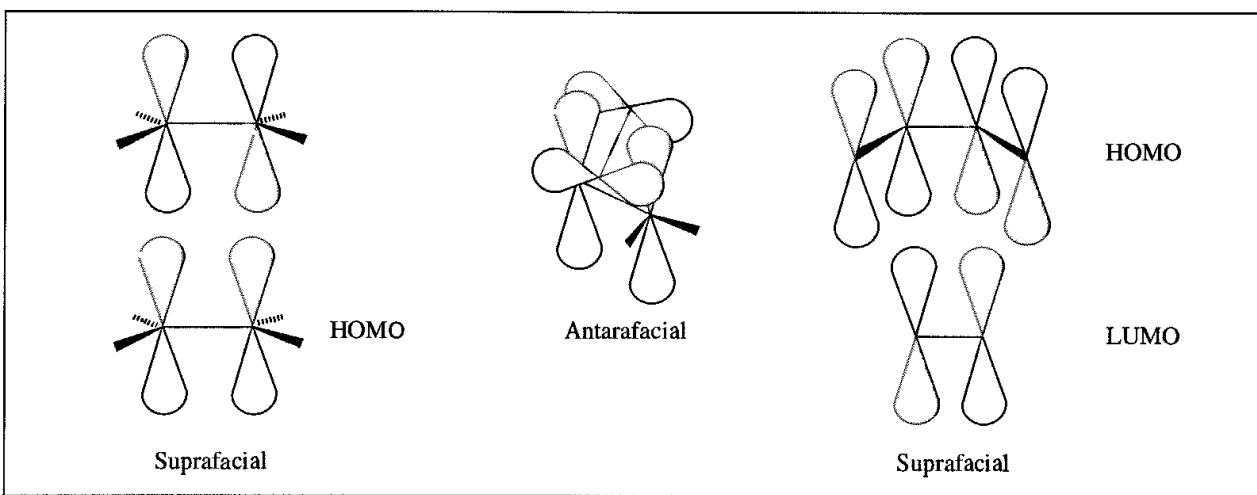
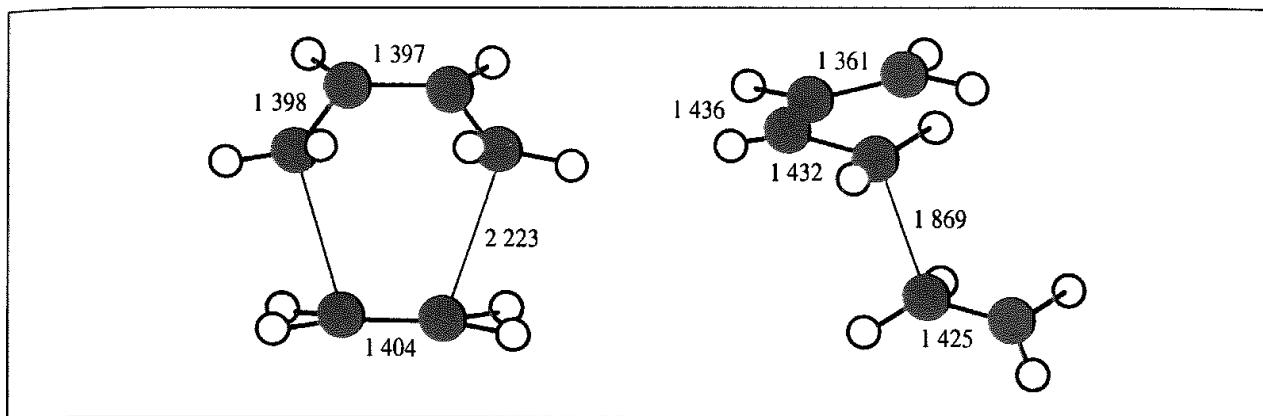


Fig. 5.34: Suprafacial attack of one ethene molecule on another (left) is not permitted by the Woodward–Hoffmann rules and the alternative antarafacial mode of attack is sterically unfavourable. Suprafacial attack is however permitted for the Diels–Alder reaction between butadiene and ethene (right).



*Fig 5.35 Geometry predicted by CASSCF ab initio calculations of the two possible transition structure geometries for the Diels–Alder reaction between ethene and butadiene (Figure adapted from Houk K N, J González and Y Li 1995 *Pericyclic Reaction Transition States: Passions and Punctilios 1935–1995 Accounts of Chemical Research* 28 81–90.)*

The Woodward–Hoffmann rules state what the outcome of a pericyclic reaction will be, but they do not define the mechanism by which the reaction occurs. Many theoretical techniques have been applied to the study of these problems over the years [Houk *et al.* 1992] and a passionate debate has ensued on the nature of the transition structures involved in these reactions. The debate has been fuelled by the fact that different theoretical treatments (especially semi-empirical methods) give different results. For example, at one extreme the Diels–Alder reaction between butadiene and ethene would proceed via a two-step mechanism involving a biradical transition structure. At the other extreme the reaction would involve a symmetrical transition state formed in a concerted, synchronous reaction. *Ab initio* calculations at various levels of theory suggest the concerted transition structure. The geometry obtained for the prototypical Diels–Alder reaction between butadiene/ethene using a CASSCF calculation and a 6-31G* basis set is shown in Figure 5.35 [Houk *et al.* 1995]. The alternative biradical structure is also shown in Figure 5.35; this is predicted to be 6 kcal/mol higher in energy than the symmetrical transition structure.

5.10 Solid-state Systems: Lattice Statics and Lattice Dynamics

Energy minimisation and normal mode analysis have an important role to play in the study of the solid state. Algorithms similar to those discussed above are employed but an extra feature of such systems, at least when they form a perfect lattice, is that it is possible to exploit the space group symmetry of the lattice to speed up the calculations. It is also important to properly take the interactions with atoms in neighbouring cells into account.

The most straightforward type of lattice minimisation is performed at constant volume, where the dimensions of the basic unit cell do not change. A more advanced type of calculation is one performed at constant pressure, in which case there are forces on both the atoms and the unit cell as a whole. The lattice vectors are considered as additional variables along with the atomic coordinates. The laws of elasticity describe the behaviour of a material when

subjected to a *stress* (defined as the force per unit area). One obvious source of stress is any external pressure, but stress may also arise from other sources, especially from interatomic forces within the cell, which give rise to 'internal stress'. The concept of *strain* is also key to this subject; the strain is the fractional change in the dimension (for example, the change per unit length when a steel rod is stretched). In the general case we consider a situation where a point \mathbf{r} in the unstrained material moves to a new point \mathbf{r}' under the effect of some strain:

$$\mathbf{u} = \mathbf{r}' - \mathbf{r} \quad (5.48)$$

If we apply the strain uniformly in one dimension (e.g. the x axis) then the x coordinate of a point that was initially at x will change by an amount proportional to x . This is written:

$$u_x = \varepsilon_{xx}x \quad (5.49)$$

In the general case the constant of proportionality is written as the first derivative:

$$\varepsilon_{xx} = \partial u_x / \partial x \quad (5.50)$$

Deformation in the y and z directions is described in an analogous manner. In order to cater for shear-type strains additional elements are defined.

$$\varepsilon_{xy} = \varepsilon_{yx} = \frac{1}{2}(\partial u_y / \partial x + \partial u_x / \partial y) \quad (5.51)$$

These values ε give rise to a *strain tensor* (see Section 4.9.1 for more discussion on tensors), which is symmetric and is often written in the following form:

$$\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 & \frac{1}{2}\varepsilon_6 & \frac{1}{2}\varepsilon_5 \\ \frac{1}{2}\varepsilon_6 & \varepsilon_2 & \frac{1}{2}\varepsilon_4 \\ \frac{1}{2}\varepsilon_5 & \frac{1}{2}\varepsilon_4 & \varepsilon_3 \end{bmatrix} \quad \begin{aligned} \varepsilon_1 &\equiv \varepsilon_{xx}, & \varepsilon_2 &\equiv \varepsilon_{yy}, & \varepsilon_3 &\equiv \varepsilon_{zz} \\ \varepsilon_4 &\equiv \varepsilon_{yz}, & \varepsilon_5 &\equiv \varepsilon_{xz}, & \varepsilon_6 &\equiv \varepsilon_{xy} \end{aligned} \quad (5.52)$$

There are thus six different numbers present in the strain tensor. The symmetric form of the strain tensor prevents rotation of the unit cell with respect to the Cartesian axis system. It is possible to use this matrix to relate how a vector \mathbf{r} in the unstrained matrix is related to one \mathbf{r}' in the strained structure as follows:

$$\mathbf{r}' = (\mathbf{I} + \boldsymbol{\varepsilon})\mathbf{r} \quad (5.53)$$

\mathbf{I} is the identity matrix. The six first derivatives of the energy with respect to the strain components ε_i measure the forces acting on the unit cell. When combined with the atomic coordinates we get a matrix with $3N + 6$ dimensions. At a minimum not only should there be no force on any of the atoms but the forces on the unit cell should also be zero. Application of a standard iterative minimisation procedure such as the Davidon-Fletcher-Powell method will optimise all these degrees of freedom to give a strain-free final structure. In such procedures a reasonably accurate estimate of the initial inverse Hessian matrix is usually required to ensure that the changes in the atomic positions and in the cell dimensions are matched.

Two common properties which can be calculated from the minimum-energy structure are the elastic and dielectric constants. The elastic constant matrix is used to relate the strains of a material to the internal forces, or stresses. It is defined as the second derivative of the energy with respect to the strain, normalised by the cell volume. The inverse of the elastic

constant matrix gives the constant of proportionality between the stress and the strain. The elastic constant matrix has dimensions 6×6 and is given by the following expression:

$$\mathbf{C} = \frac{1}{V} [\boldsymbol{\gamma}_{\varepsilon\varepsilon}'' - (\boldsymbol{\gamma}_{\varepsilon r}'' \cdot \boldsymbol{\gamma}_{rr}''^{-1} \cdot \boldsymbol{\gamma}_{rz}'')] \quad (5.54)$$

In this equation $\boldsymbol{\gamma}_{\varepsilon\varepsilon}''$ is the 6×6 matrix of second derivatives (elements $\partial^2 \mathcal{V} / \partial \varepsilon_{ij}^2$), $\boldsymbol{\gamma}_{\varepsilon i}''$ and $\boldsymbol{\gamma}_{rz}''$ are the corresponding $3N \times 6$ and $6 \times 3N$ mixed coordinate/strain matrices, $\boldsymbol{\gamma}_{rr}''$ is the $3N \times 3N$ second-derivative coordinate matrix and V is the unit cell volume. It is the second term in Equation (5.54) that accounts for internal atomic relaxations as the cell distorts.

The strains on the lattice are equal to the stress divided by the elastic constant matrix:

$$\boldsymbol{\varepsilon} = (P_{\text{static}} + P_{\text{applied}}) \cdot \mathbf{C}^{-1} \quad (5.55)$$

Here we have expressed the stress as the sum of the (external) applied pressure P_{applied} together with a static pressure P_{static} , which arises from the internal forces acting on the unit cell.

The dielectric constant is concerned with the electrical properties of a material. The dielectric constant for a solid is a 3×3 matrix with different components according to the Cartesian axes. These elements are given by:

$$D_{ij} = \delta_{ij} + \frac{4\pi}{V} \mathbf{q}^T \cdot \boldsymbol{\gamma}_{rr}''^{-1} \cdot \mathbf{q} \quad (5.56)$$

In Equation (5.56) i and j are one of x , y or z ; δ_{ij} is the delta function (i.e. equal to one when $i \equiv j$ and zero otherwise) and \mathbf{q} is a vector containing the charges of each species. It is well known that the effect of a dielectric changes in an oscillating electric field (at high enough frequencies the permanent dipoles in the material are unable to keep up with the rapidly changing field). Thus one usually calculates two sets of dielectric constant matrices, corresponding to the low- and high-frequency regimes. If polarisation is included via a shell model (see Section 4.22.2) then both the cores and the shells are used to determine the low-frequency dielectric matrix; at high frequency only the shells are considered.

Comparison of the relative energies following a minimisation calculation can enable predictions to be made of the likely structure for a given material. In the same way that an organic molecule may be able to exist in more than one three-dimensional structure (or conformation – see Chapter 9) so a solid may (in principle) be able to adopt more than one three-dimensional arrangement of its atoms whilst still maintaining a periodic lattice structure. Silica, SiO_2 , has been the subject of considerable attention using these methods. The lowest-energy form is $\alpha\text{-SiO}_2$, or quartz. However, it can also form more open structures. A number of such microporous structures are in principle available, three being silicalite, mordenite and faujasite. In one study the energies of these structures relative to the quartz structure were found to be approximately 2.6, 4.9 and 5.1 kcal/mol, respectively [Ooms *et al.* 1988]. Indeed, the silicalite structure is the only one which can be prepared as the pure silicon oxide; the other forms usually require a high aluminium content and are more traditional zeolites. In an extension of this work two slightly different forms of the silicalite structure were simulated. The normal form at room temperature has orthorhombic

symmetry but at low temperatures this changes to monoclinic. These two forms are very closely related, differing only by the distortion of a key angle by 0.64° . Nevertheless, an energy-minimisation calculation starting from the orthorhombic structure did indeed change to the monoclinic, in agreement with the experimental data [Bell *et al.* 1990]. The orthorhombic \rightarrow monoclinic transition could only be observed using a force field which included polarisation effects (i.e. the shell model). Lattice minimisation methods can sometimes be very useful in helping to solve the structure of materials, a noteworthy example being the determination of the structure of a zeolite NU-87 [Shannon *et al.* 1991]. This synthetic material is of particular interest as a catalyst as it contains a multidimensional channel system. Multidimensional systems permit more complex catalytic reactions to occur and are also less prone to deactivation than one-dimensional systems. In this case, there are rings containing ten and twelve oxygen atoms (Figure 5.36 (colour plate section)). NU-87 also has a high silica content, which confers improved stability to heat. A number of experimental techniques were used to try to determine the structure, including electron diffraction and powder synchrotron X-ray diffraction, as a result of which an approximate structure was deduced but there remained some features in the powder diffraction spectrum that could not be accounted for. These were initially believed to be due to impurities but after energy-minimisation studies some subtle changes in the structure occurred to give a related structure that was a better match to the experimental data. A key feature of this particular minimisation was that the structure was not forced to adopt any specific symmetry but rather each atom was able to move independently of the others.

The calculation of vibrational frequencies (called *phonons*) is important to the study of the solid state. Indeed, the calculation of and study of phonons is often given a special name, *lattice dynamics*. To calculate the vibrational frequencies for a solid one follows a very similar approach to that described earlier for molecules, with the exception that when a shell model is being used* then their effect must be incorporated into the mass-weighted matrix of second derivatives (though not directly as they have no mass):

$$\boldsymbol{\mathcal{V}}'' = \boldsymbol{\mathcal{V}}''_{\text{core-core}} - \boldsymbol{\mathcal{V}}''_{\text{core-shell}} \cdot \boldsymbol{\mathcal{V}}'^{-1}_{\text{shell-shell}} \cdot \boldsymbol{\mathcal{V}}''_{\text{core-shell}} \quad (5.57)$$

Of additional importance is that the vibrational modes are dependent upon the reciprocal lattice vector \mathbf{k} . As with calculations of the electronic structure of periodic lattices these calculations are usually performed by selecting a suitable set of points from within the Brillouin zone. For periodic solids it is necessary to take this periodicity into account; the effect on the second-derivative matrix is that each element ij needs to be multiplied by the phase factor $\exp(i\mathbf{k} \cdot \mathbf{r}_{ij})$. A *phonon dispersion curve* indicates how the phonon frequencies vary over the Brillouin zone, an example being shown in Figure 5.37. The phonon density of states is the variation in the number of frequencies as a function of frequency. A purely transverse vibration is one where the displacement of the atoms is perpendicular to the direction of motion of the wave; in a purely longitudinal vibration the atomic displacements are in the direction of the wave motion. Such motions can be observed in simple systems (e.g. those that contain just one or two atoms per unit cell) but for general three-dimensional lattices most of the vibrations are a mixture of transverse and longitudinal motions, the exceptions

* The use of a shell model is generally recommended otherwise the resulting frequencies are too high.

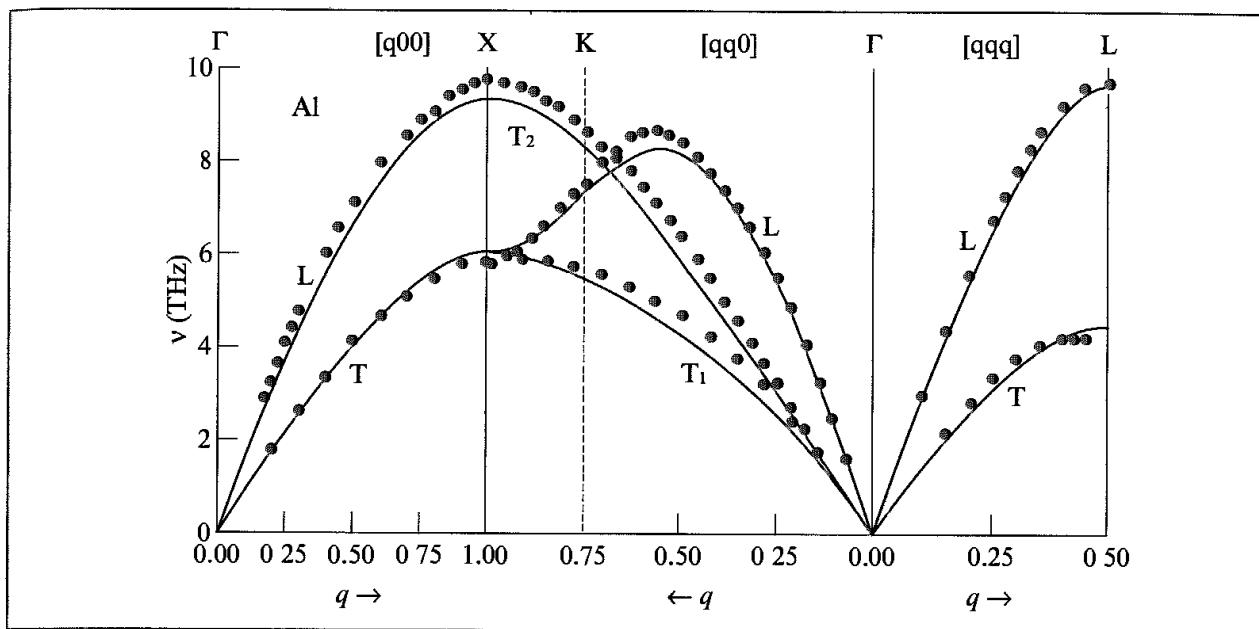


Fig 5.37. Comparison of the calculated phonon dispersion curve for Al with the experimental values measured using neutron diffraction (Figure redrawn from Michin Y, D Farkas, M J Mehl and D A Papaconstantopoulos 1999 Interatomic Potentials for Monomatomic Metals from Experimental Data and ab initio Calculations Physical Review B59 3393–3407)

being those along directions of high symmetry. The phonons are additionally classified as acoustic or optical; the former are typically of longer wavelength (lower-frequency oscillations) where the atoms move as a unit. The name arises from the fact that these are often measured as sound waves. At the point $\mathbf{k} = 0$ (the gamma point) the first three vibrational frequencies correspond to translation of the entire lattice. The optical phonons are typically higher in frequency. Various experimental techniques can be used to investigate lattice vibrations and to determine the phonon dispersion curves, the most powerful of which is inelastic scattering using thermal neutrons. These often allow the entire range of \mathbf{k} to be sampled, in contrast to some of the alternative types of radiation.

Once the phonon frequencies are known it becomes possible to determine various thermodynamic quantities using statistical mechanics (see Appendix 6.1). Here again some slight modifications are required to the standard formulae. These modifications are usually a consequence of the need to sum over the points sampled in the Brillouin zone. For example, the zero-point energy is:

$$U_{\text{vib}}(0) = \sum_{q=1}^p w_q \sum_{i=1}^{N_{\text{nm}}} \frac{\hbar\nu_i}{2} \quad (5.58)$$

In Equation (5.58) the outer summation is over the p points q which are used to sample the Brillouin zone, w_q is the fractional weight associated with each point (related to the volume of Brillouin zone space surrounding q) and ν_i are the phonon frequencies. In addition to the internal energy due to the vibrational modes it is also possible to calculate the vibrational entropy, and hence the free energy. The Helmholtz free energy at a temperature

T is thus given in the quasi-harmonic approximation by the sum of static and vibrational contributions:

$$A = \mathcal{V} + \sum_{q=1}^p w_q \sum_{i=1}^{N_{\text{hm}}} \left(\frac{\hbar\nu_i}{2} + k_B T \ln \left[1 - \exp \left(-\frac{\hbar\nu_i}{k_B T} \right) \right] \right) \quad (5.59)$$

Here, \mathcal{V} is the internal energy calculated from the potential energy model. The heat capacity at constant volume is another useful thermodynamic quantity that can be determined directly from the frequencies as it equals the derivative of the vibrational internal energy with respect to temperature.

An extension of these ideas involves minimisation of the free energy as a function of the coordinates and the temperature. The function to be minimised is sometimes referred to as an *availability*, given by $G^* = A + P_{\text{ext}} V$ where P_{ext} is the external pressure and V is the volume. Such a free energy minimisation requires derivatives of the free energy with respect to the coordinates. Early implementations used approximations such as separating the changes in the external coordinates (i.e. the dimensions of the unit cell) from the internal coordinates (i.e. the locations of the ions within the unit cell). In addition, true free energy derivatives might be calculated for the external coordinates only (due to the computational cost) with the internal coordinates being changed using the static potential energy. It is now possible to calculate a full set of analytical free energy first derivatives and hence to perform a full minimisation of the free energy with respect to external and internal coordinates simultaneously [Taylor *et al.* 1998].

Free energy minimisation provides information that is in many ways complementary to molecular dynamics simulations [Allan *et al.* 2000]. The former is particularly useful for investigating materials at lower temperatures where the harmonic assumption is valid; moreover it includes zero-point energy and quantisation effects, which are ignored by molecular dynamics. In addition, free energy minimisation provides free energies directly, rather than free energy differences, and is computationally significantly less expensive. Conversely anharmonic effects become important at higher temperatures, making molecular dynamics and Monte Carlo more suitable. One property that can be calculated via free energy minimisation is the thermal expansivity. This involves a series of free energy minimisations at different temperatures. It is also possible to calculate the free energy of disordered solids, and thus enthalpies and entropies of mixing.

Further Reading

Catlow C R A 1998. Solids: Computer Modelling. In Schleyer, P v R, N L Allinger, T Clark, J Gasteiger, P A Kollman, H F Schaefer III and P R Schreiner (Editors) *The Encyclopedia of Computational Chemistry*, John Wiley & Sons, Chichester.

Gill P E and W Murray 1981. *Practical Optimization*. London, Academic Press.

McKee M L and M Page 1993. Computing Reaction Pathways on Molecular Potential Energy Surfaces. In Lipkowitz K B and D B Boyd (Editors) *Reviews in Computational Chemistry Volume 4*. New York, VCH Publishers, pp. 35–65

- Press W H, B P Flannery, S A Teukolsky and W T Vetterling 1992 *Numerical Recipes in Fortran*. Cambridge, Cambridge University Press.
- Schlegel H B 1987. Optimization of Equilibrium Geometries and Transition Structures In Lawley K P (Editor) *Ab Initio Methods in Quantum Chemistry - I* New York, John Wiley & Sons, pp 249-286.
- Schlegel H B 1989 Some Practical Suggestions for Optimizing Geometries and Locating Transition States In Bertrán J and J G Csizmadia (Editors). *New Theoretical Concepts for Understanding Organic Reactions* Dordrecht, Kluwer, pp. 33-53
- Schlick T 1992 Optimization Methods in Computational Chemistry. In Lipkowitz K B and D B Boyd (Editors) *Reviews in Computational Chemistry* Volume 3. New York, VCH Publishers, pp 1-71
- Stassis C 19 Lattice Dynamics. In Sköld and D L Price (Editors) *Methods of Experimental Physics Volume 23. Neutron Scattering Part A*. Orlando, Academic Press, pp. 369-440.
- Watson G W, P Tschaufeser, A Wall, R A Jackson and S C Parker 1997. Lattice Energy and Free Energy Minimisation Techniques. *Computer Modelling in Inorganic Crystallography*. San Diego, Academic Press, pp 55-81.
- Williams I H 1993. Interplay of Theory and Experiment in the Determination of Transition-State Structures. *Chemical Society Reviews* 1:277-283.

References

- Allan N L, G D Barrera, J A Purton, C E Sims and M B Taylor 2000. Ionic Solids at High Temperatures and Pressures: *Ab initio*, Lattice Dynamics and Monte Carlo Studies. *Physical Chemistry Chemical Physics* **2**:1099-1111.
- Ayala P Y and H B Schlegel 1997 A Combined Method for Determining Reaction Paths, Minima and Transition State Geometries *Journal of Chemical Physics* **107** 375-384
- Baker J 1986. An Algorithm for the Location of Transition States. *Journal of Computational Chemistry* **7**:385-395.
- Bell R G, R A Jackson and C R A Catlow 1990. Computer Simulation of the Monoclinic Distortion in Silicalite. *Journal of the Chemical Society Chemical Communications* **10**:782-783.
- Brooks B and M Karplus 1983 Harmonic Dynamics of Proteins: Normal Modes and Fluctuations in Bovine Pancreatic Trypsin Inhibitor. *Proceedings of the National Academy of Sciences USA* **80**:6571-6575
- Czerninski R and R Elber 1990. Self-Avoiding Walk Between 2 Fixed-Points as a Tool to Calculate Reaction Paths in Large Molecular Systems. *International Journal of Quantum Chemistry* **S24**:167-186.
- Dauber-Osguthorpe P, V A Roberts, D J Osguthorpe, J Wolff, M Genest and A T Hagler 1988 Structure and Energetics of Ligand Binding to Proteins: *Escherichia coli* Dihydrofolate Reductase-Tri-methoprim, A Drug-Receptor System. *Proteins Structure, Function and Genetics* **4**:31-47
- Doubleday C, J McIver, M Page and T Zielinski 1985 Temperature Dependence of the Transition-State Structure for the Disproportionation of Hydrogen Atom with Ethyl Radical. *Journal of the American Chemical Society* **107**:5800-5801
- Elber R and M Karplus 1987. A Method for Determining Reaction Paths in Large Molecules: Application to Myoglobin. *Chemical Physics Letters* **139**:375-380.
- Fischer S, P D J Groothuis, L C Groenen, W P van Hoorn, F C J M van Geggel, D N Reinhoudt and M Karplus 1995 Pathways for Conformational Interconversion of Calix[4]arenes. *Journal of the American Chemical Society* **117**:1611-1620.
- Fischer S and M Karplus 1992 Conjugate Peak Refinement: An Algorithm for Finding Reaction Paths and Accurate Transition States in Systems with Many Degrees of Freedom. *Chemical Physics Letters* **194**:252-261.

- Fisher C L, V A Roberts and A T Hagler 1991. Influence of Environment on the Antifolate Drug Trimethoprim: Energy Minimization Studies *Biochemistry* **30**:3518–3526
- Fukui K 1971. Recognition of Stereochemical Paths by Orbital Interaction. *Accounts of Chemical Research* **4**:57–64
- Fukui K 1981. The Path of Chemical Reactions – The IRC Approach *Accounts of Chemical Research* **14**:368–375.
- Gelin B R and M Karplus 1975. Sidechain Torsional Potential and Motion of Amino Acids in Proteins: Bovine Pancreatic Trypsin Inhibitor *Proceedings of the National Academy of Sciences USA* **72** 2002–2006.
- Gonzalez C and H B Schlegel 1988. An Improved Algorithm for Reaction Path Following *Journal of Chemical Physics* **90**:2154–2161.
- Houk K N, J González and Y Li 1995 Pericyclic Reaction Transition States: Passions and Punctilios 1935–1995 *Accounts of Chemical Research* **28**:81–90
- Houk K N, Y Li and J D Evanseck 1992 Transition Structures of Hydrocarbon Pericyclic Reactions. *Angewandte Chemie International Edition in English* **31**:682–708.
- Nowak W, R Czerminski and R Elber 1991. Reaction Path Study of Ligand Diffusion in Proteins: Application of the Self Penalty Walk (SPW) Method to Calculate Reaction Coordinates for the Motion of CO through Leghemoglobin *Journal of the American Chemical Society* **113** 5627–5737.
- Ooms G, R A van Santen, C J J Den Ouden, R A Jackson and C R A Catlow 1988. Relative Stabilities of Zeolitic Aluminosilicates. *Journal of Physical Chemistry* **92** 4462–4465.
- Peng C, P Y Ayala, H B Schlegel and M J Frisch 1996. Using Redundant Internal Coordinates to Optimise Equilibrium Geometries and Transition States. *Journal of Computational Chemistry* **17**:49–56.
- Pople J A, A P Scott, M W Wong and L Radom 1993. Scaling Factors for Obtaining Fundamental Vibrational Frequencies and Zero-Point Energies from HF/6-31G* and MP2/6-31G* Harmonic Frequencies *Israel Journal of Chemistry* **33**:345–350.
- Schlegel H B 1982. Optimisation of Equilibrium Geometries and Transition Structures *Journal of Computational Chemistry* **3**:214–218
- Shannon M D, J L Casci, P A Cox and S J Andrews 1991. Structure of the Two-dimensional Medium-pore High-silica Zeolite NU-87. *Nature* **353**:417–420
- Taylor M B, G D Barrera, N L Allan, T H K Barron and W C Mackrodt 1998. Shell: A Code for Lattice Dynamics and Structure Optimisation of Ionic Crystals. *Computer Physics Communications* **109**: 135–143.
- Woodward R B and R Hoffmann 1969. The Conservation of Orbital Symmetry. *Angewandte Chemie International Edition in English* **8**:781–853.

Computer Simulation Methods

6.1 Introduction

Energy minimisation generates individual minimum energy configurations of a system. In some cases the information provided by energy minimisation can be sufficient to predict accurately the properties of a system. If all minimum configurations on an energy surface can be identified then statistical mechanical formulae can be used to derive a partition function from which thermodynamic properties can be calculated. However, this is possible only for relatively small molecules or small molecular assemblies in the gas phase. The molecular modeller more often wants to understand and to predict the properties of liquids, solutions and solids, to study complex processes such as the adsorption of molecules onto surfaces and into solids and to investigate the behaviour of macromolecules which have many closely separated minima. In such systems the experimental measurements are made on macroscopic samples that contain extremely large numbers of atoms or molecules, with an enormous number of minima on their energy surfaces. A full quantification of the energy surfaces of such systems is not possible, nor is it ever likely to be. Computer simulation methods enable us to study such systems and predict their properties through the use of techniques that consider small replications of the macroscopic system with manageable numbers of atoms or molecules. A simulation generates representative configurations of these small replications in such a way that accurate values of structural and thermodynamic properties can be obtained with a feasible amount of computation. Simulation techniques also enable the time-dependent behaviour of atomic and molecular systems to be determined, providing a detailed picture of the way in which a system changes from one conformation or configuration to another. Simulation techniques are also widely used in some experimental procedures, such as the determination of protein structures from X-ray crystallography.

In this chapter we shall discuss some of the general principles involved in the two most common simulation techniques used in molecular modelling: the molecular dynamics and the Monte Carlo methods. We shall also discuss several concepts that are common to both of these methods. A more detailed discussion of the two simulation methods can be found in Chapters 7 and 8.

6.1.1 Time Averages, Ensemble Averages and Some Historical Background

Suppose we wish to determine experimentally the value of a property of a system such as the pressure or the heat capacity. In general, such properties will depend upon the positions and

momenta of the N particles that comprise the system. The instantaneous value of the property A can thus be written as $A(\mathbf{p}^N(t), \mathbf{r}^N(t))$, where $\mathbf{p}^N(t)$ and $\mathbf{r}^N(t)$ represent the N momenta and positions respectively at time t (i.e. $A(\mathbf{p}^N(t), \mathbf{r}^N(t)) \equiv A(p_{1x}, p_{1y}, p_{1z}, p_{2x}, \dots, x_1, y_1, z_1, x_2, \dots, t)$ where p_{1x} is the momentum of particle 1 in the x direction and x_1 is its x coordinate). Over time, the instantaneous value of the property A fluctuates as a result of interactions between the particles. The value that we measure experimentally is an average of A over the time of the measurement and is therefore known as a *time average*. As the time over which the measurement is made increases to infinity, so the value of the following integral approaches the 'true' average value of the property:

$$A_{\text{ave}} = \lim_{\tau \rightarrow \infty} \frac{1}{\tau} \int_{t=0}^{\tau} A(\mathbf{p}^N(t), \mathbf{r}^N(t)) dt \quad (6.1)$$

To calculate average values of the properties of the system, it would therefore appear to be necessary to simulate the dynamic behaviour of the system (i.e. to determine values of $A(\mathbf{p}^N(t), \mathbf{r}^N(t))$, based upon a model of the intra- and intermolecular interactions present). In principle, this is relatively straightforward to do. For any arrangement of the atoms in the system, the force acting on each atom due to interactions with other atoms can be calculated by differentiating the energy function. From the force on each atom it is possible to determine its acceleration via Newton's second law. Integration of the equations of motion should then yield a trajectory that describes how the positions, velocities and accelerations of the particles vary with time, and from which the average values of properties can be determined using the numerical equivalent of Equation (6.1). The difficulty is that for 'macroscopic' numbers of atoms or molecules (of the order of 10^{23}) it is not even feasible to determine an initial configuration of the system, let alone integrate the equations of motion and calculate a trajectory. Recognising this problem, Boltzmann and Gibbs developed statistical mechanics, in which a single system evolving in time is replaced by a large number of replications of the system that are considered simultaneously. The time average is then replaced by an *ensemble average*:

$$\langle A \rangle = \iint d\mathbf{p}^N d\mathbf{r}^N A(\mathbf{p}^N, \mathbf{r}^N) \rho(\mathbf{r}^N, \mathbf{r}^N) \quad (6.2)$$

The angle brackets $\langle \rangle$ indicate an ensemble average, or *expectation value*; that is, the average value of the property A over all replications of the ensemble generated by the simulation. Equation (6.2) is written as a double integral for convenience but in fact there should be $6N$ integral signs on the integral for the $6N$ positions and momenta of all the particles. $\rho(\mathbf{p}^N, \mathbf{r}^N)$ is the *probability density* of the ensemble; that is, the probability of finding a configuration with momenta \mathbf{p}^N and positions \mathbf{r}^N . The ensemble average of the property A is then determined by integrating over all possible configurations of the system. In accordance with the *ergodic hypothesis*, which is one of the fundamental axioms of statistical mechanics, the ensemble average is equal to the time average. Under conditions of constant number of particles, volume and temperature, the probability density is the familiar Boltzmann distribution:

$$\rho(\mathbf{p}^N, \mathbf{r}^N) = \exp(-E(\mathbf{p}^N, \mathbf{r}^N)/k_B T)/Q \quad (6.3)$$

In Equation (6.3), $E(\mathbf{p}^N, \mathbf{r}^N)$ is the energy, Q is the partition function, k_B is Boltzmann's constant and T is the temperature. The partition function is more generally written in terms of the Hamiltonian, \mathcal{H} ; for a system of N identical particles the partition function

for the canonical ensemble is as follows:

$$Q_{NVT} = \frac{1}{N!} \frac{1}{h^{3N}} \iint d\mathbf{p}^N d\mathbf{r}^N \exp \left[-\frac{\mathcal{H}(\mathbf{p}^N, \mathbf{r}^N)}{k_B T} \right] \quad (6.4)$$

The canonical ensemble is the name given to an ensemble for constant temperature, number of particles and volume. For our purposes \mathcal{H} can be considered the same as the total energy, $E(\mathbf{p}^N, \mathbf{r}^N)$, which equals the sum of the kinetic energy ($\mathcal{K}(\mathbf{p}^N)$) of the system, which depends upon the momenta of the particles, and the potential energy ($\mathcal{V}(\mathbf{r}^N)$), which depends upon the positions. The factor $N!$ arises from the indistinguishability of the particles and the factor $1/h^{3N}$ is required to ensure that the partition function is equal to the quantum mechanical result for a particle in a box. A short discussion of some of the key results of statistical mechanics is provided in Appendix 6.1 and further details can be found in standard textbooks.

The first computer simulations of fluids were performed in 1952 by Metropolis, Rosenbluth, Rosenbluth, Teller and Teller, who developed a scheme for sampling from the Boltzmann distribution to give ensemble averages. This gave rise to the Monte Carlo simulation method. Not long afterwards (in 1957) Alder recognised that it was, in fact, possible to integrate the equations of motion for a relatively small number of particles, and to mimic the behaviour of a real system using periodic boundary conditions. This led to the first molecular dynamics simulations of molecular systems.

6.1.2 A Brief Description of the Molecular Dynamics Method

Molecular dynamics calculates the ‘real’ dynamics of the system, from which time averages of properties can be calculated. Sets of atomic positions are derived in sequence by applying Newton’s equations of motion. Molecular dynamics is a *deterministic* method, by which we mean that the state of the system at any future time can be predicted from its current state. The first molecular dynamics simulations were performed using very simple potentials such as the hard-sphere potential. The behaviour of the particles in this potential is similar to that of billiard or snooker balls. the particles move in straight lines at constant velocity between collisions. The collisions are perfectly elastic and occur when the separation between a pair of spheres equals the sum of their radii. After a collision, the new velocities of the colliding spheres are calculated using the principle of conservation of linear momentum. The hard-sphere model has provided many useful results but is obviously not ideal for simulating atomic or molecular systems. In potentials such as the Lennard-Jones potential the force between two atoms or molecules changes continuously with their separation. By contrast, in the hard-sphere model there is no force between particles until they collide. The continuous nature of the more realistic potentials requires the equations of motion to be integrated by breaking the calculation into a series of very short time steps (typically between 1 femtosecond and 10 femtoseconds; 10^{-15} s to 10^{-14} s). At each step, the forces on the atoms are computed and combined with the current positions and velocities to generate new positions and velocities a short time ahead. The force acting on each atom is assumed to be constant during the time interval. The atoms are then moved to the new positions, an updated set of forces is computed, and so on. In this way a molecular dynamics simulation generates a

trajectory that describes how the dynamic variables change with time. Molecular dynamics simulations are typically run for tens or hundreds of picoseconds (a 100 ps simulation using a 1 fs time step requires 100 000 steps) Thermodynamic averages are obtained from molecular dynamics as time averages using numerical integration of Equation (6.2):

$$\langle A \rangle = \frac{1}{M} \sum_{i=1}^M A(\mathbf{p}^N, \mathbf{r}^N) \quad (6.5)$$

M is the number of time steps. Molecular dynamics is also extensively used to investigate the conformational properties of flexible molecules as will be discussed in Chapters 7 and 9.

6.1.3 The Basic Elements of the Monte Carlo Method

In a molecular dynamics simulation the successive configurations of the system are connected in time. This is not the case in a Monte Carlo simulation, where each configuration depends only upon its predecessor and not upon any other of the configurations previously visited. The Monte Carlo method generates configurations randomly and uses a special set of criteria to decide whether or not to accept each new configuration. These criteria ensure that the probability of obtaining a given configuration is equal to its Boltzmann factor, $\exp\{-\mathcal{V}(\mathbf{r}^N)/k_B T\}$, where $\mathcal{V}(\mathbf{r}^N)$ is calculated using the potential energy function. States with a low energy are thus generated with a higher probability than configurations with a higher energy. For each configuration that is accepted the values of the desired properties are calculated and at the end of the calculation the average of these properties is obtained by simply averaging over the number of values calculated, M :

$$\langle A \rangle = \frac{1}{M} \sum_{i=1}^M A(\mathbf{r}^N) \quad (6.6)$$

Most Monte Carlo simulations of molecular systems are more properly referred to as Metropolis Monte Carlo calculations after Metropolis and his colleagues, who reported the first such calculation. The distinction can be important because there are other ways in which an ensemble of configurations can be generated. As we shall see in Chapter 7, the Metropolis scheme is only one of a number of possibilities, though it is by far the most popular.

In a Monte Carlo simulation each new configuration of the system may be generated by randomly moving a single atom or molecule. In some cases new configurations may also be obtained by moving several atoms or molecules or by rotating about one or more bonds. The energy of the new configuration is then calculated using the potential energy function. If the energy of the new configuration is lower than the energy of its predecessor then the new configuration is accepted. If the energy of the new configuration is higher than the energy of its predecessor then the *Boltzmann factor* of the energy difference is calculated: $\exp[-(\mathcal{V}_{\text{new}}(\mathbf{r}^N) - \mathcal{V}_{\text{old}}(\mathbf{r}^N))/k_B T]$. A random number is then generated between 0 and 1 and compared with this Boltzmann factor. If the random number is higher than the Boltzmann factor then the move is rejected and the original configuration is retained for the next iteration; if the random number is lower then the move is accepted and the new

configuration becomes the next state. This procedure has the effect of permitting moves to states of higher energy. The smaller the uphill move (i.e. the smaller the value of $\mathcal{V}_{\text{new}}(\mathbf{r}^N) - \mathcal{V}_{\text{old}}(\mathbf{r}^N)$) the greater is the probability that the move will be accepted.

6.1.4 Differences Between the Molecular Dynamics and Monte Carlo Methods

The molecular dynamics and Monte Carlo simulation methods differ in a variety of ways. The most obvious difference is that molecular dynamics provides information about the time dependence of the properties of the system whereas there is no temporal relationship between successive Monte Carlo configurations. In a Monte Carlo simulation the outcome of each trial move depends only upon its immediate predecessor, whereas in molecular dynamics it is possible to predict the configuration of the system at any time in the future - or indeed at any time in the past. Molecular dynamics has a kinetic energy contribution to the total energy whereas in a Monte Carlo simulation the total energy is determined directly from the potential energy function. The two simulation methods also sample from different ensembles. Molecular dynamics is traditionally performed under conditions of constant number of particles (N), volume (V) and energy (E) (the microcanonical or constant NVE ensemble) whereas a traditional Monte Carlo simulation samples from the canonical ensemble (constant N , V and temperature, T). Both the molecular dynamics and Monte Carlo techniques can be modified to sample from other ensembles; for example, molecular dynamics can be adapted to simulate from the canonical ensemble. Two other ensembles are common:

isothermal-isobaric: fixed N, T, P (pressure)

grand canonical: fixed μ (chemical potential), V, T

In the canonical, microcanonical and isothermal-isobaric ensembles the number of particles is constant but in a grand canonical simulation the composition can change (i.e. the number of particles can increase or decrease). The equilibrium states of each of these ensembles are characterised as follows:

canonical ensemble: minimum Helmholtz free energy (A)

microcanonical ensemble: maximum entropy (S)

isothermal-isobaric ensemble: minimum Gibbs function (G)

grand canonical ensemble: maximum pressure \times volume (PV)

6.2 Calculation of Simple Thermodynamic Properties

A wide variety of thermodynamic properties can be calculated from computer simulations; a comparison of experimental and calculated values for such properties is an important way in which the accuracy of the simulation and the underlying energy model can be quantified. Simulation methods also enable predictions to be made of the thermodynamic properties of systems for which there is no experimental data, or for which experimental data is difficult or impossible to obtain. Simulations can also provide structural information about the

conformational changes in molecules and the distributions of molecules in a system. The emphasis in our discussion will be on those properties that are routinely calculated in computer simulations and on the way in which they are obtained. It is important to recognise that the results we derive are for the canonical ensemble. Sometimes the equivalent expressions in other ensembles are provided. The result obtained from one ensemble may also be transformed to another ensemble, though this is strictly only possible in the limit of an infinitely large system. The expressions follow from standard statistical mechanical relationships, which are given in standard texts and summarised in Appendix 6.1.

6.2.1 Energy

The internal energy is easily obtained from a simulation as the ensemble average of the energies of the states that are examined during the course of the simulation:

$$U = \langle E \rangle = \frac{1}{M} \sum_{i=1}^M E_i \quad (6.7)$$

6.2.2 Heat Capacity

At a phase transition the heat capacity will often show a characteristic dependence upon the temperature (a first-order phase transition is characterised by an infinite heat capacity at the transition but in a second-order phase transition the heat capacity changes discontinuously). Monitoring the heat capacity as a function of temperature may therefore enable phase transitions to be detected. Calculations of the heat capacity can also be compared with experimental results and so be used to check the energy model or the simulation protocol.

The heat capacity is formally defined as the partial derivative of the internal energy with respect to temperature:

$$C_V = \left(\frac{\partial U}{\partial T} \right)_V \quad (6.8)$$

The heat capacity can therefore be calculated by performing a series of simulations at different temperatures, and then differentiating the energy with respect to the temperature. The differentiation can be done numerically or by fitting a polynomial to the data and then analytically differentiating the fitted function. The heat capacity may also be calculated from a single simulation by considering the instantaneous fluctuations in the energy as follows:

$$C_V = \{ \langle E^2 \rangle - \langle E \rangle^2 \} / k_B T^2 \quad (6.9)$$

An alternative way to write this expression uses the relationship:

$$\langle (E - \langle E \rangle)^2 \rangle = \langle E^2 \rangle - \langle E \rangle^2 \quad (6.10)$$

giving:

$$C_V = \langle (E - \langle E \rangle)^2 \rangle / k_B T^2 \quad (6.11)$$

A derivation of this result is provided in Appendix 6.2.

The heat capacity can therefore be obtained by keeping a running count of E^2 and E during the simulation, from which their expectation values $\langle E^2 \rangle$ and $\langle E \rangle$ can be calculated at the end of the calculation. Alternatively, if the energies are stored during the simulation then the value of $\langle (E - \langle E \rangle)^2 \rangle$ can be calculated once the simulation has finished. This second approach may be more accurate due to round-off errors; $\langle E^2 \rangle$ and $\langle E \rangle^2$ are usually both large numbers and so there may be a large uncertainty in their difference.

6.2.3 Pressure

The pressure is usually calculated in a computer simulation via the virial theorem of Clausius. The *virial* is defined as the expectation value of the sum of the products of the coordinates of the particles and the forces acting on them. This is usually written $W = \sum x_i \dot{p}_{x_i}$ where x_i is a coordinate (e.g. the x or y coordinate of an atom) and \dot{p}_{x_i} is the first derivative of the momentum along that coordinate (\dot{p}_i is the force, by Newton's second law). The virial theorem states that the virial is equal to $-3Nk_B T$.

In an ideal gas, the only forces are those due to interactions between the gas and the container and it can be shown that the virial in this case equals $-3PV$. This result can also be obtained directly from $PV = Nk_B T$.

Forces between the particles in a real gas or liquid affect the virial, and thence the pressure. The total virial for a real system equals the sum of an ideal gas part ($-3PV$) and a contribution due to interactions between the particles. The result obtained is:

$$W = -3PV + \sum_{i=1}^N \sum_{j=i+1}^N r_{ij} \frac{d\nu(r_{ij})}{dr_{ij}} = -3Nk_B T \quad (6.12)$$

The real gas part is derived in Appendix 6.3. If $d\nu(r_{ij})/dr_{ij}$ is written as f_{ij} , the force acting between atoms i and j , then we have the following expression for the pressure:

$$P = \frac{1}{V} \left[Nk_B T - \frac{1}{3} \sum_{i=1}^N \sum_{j=i+1}^N r_{ij} f_{ij} \right] \quad (6.13)$$

The forces are calculated as part of a molecular dynamics simulation, and so little additional effort is required to calculate the virial and thus the pressure. The forces are not routinely calculated during a Monte Carlo simulation, and so some additional effort is required to determine the pressure by this route. When calculating the pressure it is also important to check that the components of the pressure in all three directions are equal.

6.2.4 Temperature

In a canonical ensemble the total temperature is constant. In the microcanonical ensemble, however, the temperature will fluctuate. The temperature is directly related to the kinetic energy of the system as follows:

$$\mathcal{K} = \sum_{i=1}^N \frac{|\mathbf{p}_i|^2}{2m_i} = \frac{k_B T}{2} (3N - N_c) \quad (6.14)$$

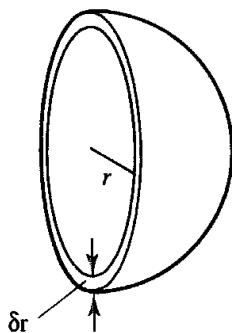


Fig. 6.1. Radial distribution functions use a spherical shell of thickness δr .

In this equation, \mathbf{p}_i is the total momentum of particle i and m_i is its mass. According to the theorem of the equipartition of energy each degree of freedom contributes $k_B T/2$. If there are N particles, each with three degrees of freedom, then the kinetic energy should equal $3Nk_B T/2$. N_c in Equation (6.14) is the number of constraints on the system. In a molecular dynamics simulation the total linear momentum of the system is often constrained to a value of zero, which has the effect of removing three degrees of freedom from the system and so N_c would be equal to 3. Other types of constraint are also possible as we shall discuss in Section 7.5.

6.2.5 Radial Distribution Functions

Radial distribution functions are a useful way to describe the structure of a system, particularly of liquids. Consider a spherical shell of thickness δr at a distance r from a chosen atom (Figure 6.1). The volume of the shell is given by:

$$\begin{aligned} V &= \frac{4}{3}\pi(r + \delta r)^3 - \frac{4}{3}\pi r^3 \\ &= 4\pi r^2 \delta r + 4\pi r \delta r^2 + \frac{4}{3}\pi \delta r^3 \approx 4\pi r^2 \delta r \end{aligned} \quad (6.15)$$

If the number of particles per unit volume is ρ , then the total number in the shell is $4\pi\rho r^2 \delta r$ and so the number of atoms in the volume element varies as r^2 .

The pair distribution function, $g(r)$, gives the probability of finding an atom (or molecule, if simulating a molecular fluid) a distance r from another atom (or molecule) compared to the ideal gas distribution. $g(r)$ is thus dimensionless. Higher radial distribution functions (e.g. the triplet radial distribution function) can also be defined but are rarely calculated and so references to the ‘radial distribution function’ are usually taken to mean the pairwise version. In a crystal, the radial distribution function has an infinite number of sharp peaks whose separations and heights are characteristic of the lattice structure.

The radial distribution function of a liquid is intermediate between the solid and the gas, with a small number of peaks at short distances, superimposed on a steady decay to a constant value at longer distances. The radial distribution function calculated from a molecular dynamics simulation of liquid argon (shown in Figure 6.2) is typical. For short distances (less

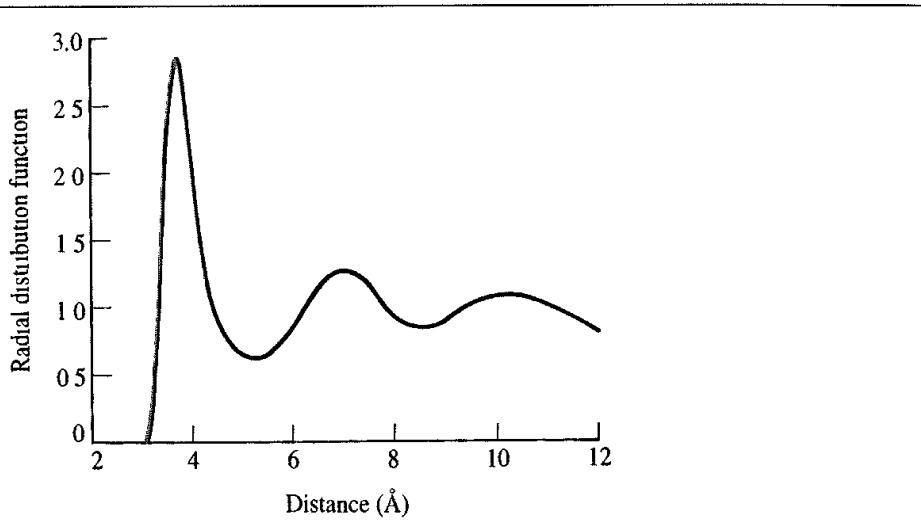


Fig 6.2 Radial distribution function determined from a 100 ps molecular dynamics simulation of liquid argon at a temperature of 100 K and a density of $1\,396 \text{ g cm}^{-3}$

than the atomic diameter) $g(r)$ is zero. This is due to the strong repulsive forces. The first (and largest) peak occurs at $r \approx 3.7 \text{ \AA}$, with $g(r)$ having a value of about 3. This means that it is three times more likely that two molecules would have this separation than in the ideal gas. The radial distribution function then falls and passes through a minimum value around $r \approx 5.4 \text{ \AA}$. The chances of finding two atoms with this separation are less than for the ideal gas. At long distances, $g(r)$ tends to the ideal gas value, indicating that there is no long-range order.

To calculate the pair distribution function from a simulation, the neighbours around each atom or molecule are sorted into distance ‘bins’, or histograms. The number of neighbours in each bin is then averaged over the entire simulation. For example, a count is made of the number of neighbours between (say) 2.5 \AA and 2.75 \AA , 2.75 \AA and 3.0 \AA and so on for every atom or molecule in the simulation. This count can be performed during the simulation itself or by analysing the configurations that are generated.

Radial distribution functions can be measured experimentally using X-ray diffraction. The regular arrangement of the atoms in a crystal gives the characteristic X-ray diffraction pattern with bright, sharp spots. For liquids, the diffraction pattern has regions of high and low intensity but no sharp spots. The X-ray diffraction pattern can be analysed to calculate an experimental distribution function, which can then be compared with that obtained from the simulation.

Thermodynamic properties can be calculated using the radial distribution function, if pairwise additivity of the forces is assumed. These properties are usually given as an ideal gas part plus a real gas part. For example, to calculate the energy of a real gas, we consider the spherical shell of volume $4\pi r^2 \delta r$ that contains $4\pi r^2 \rho g(r) \delta r$ particles. If the pair potential at a distance r has a value $\psi(r)$ then the energy of interaction between the particles in the shell and the central particle is $4\pi r^2 \rho g(r) \psi(r) \delta r$. The total potential energy of the real gas is obtained by integrating this between 0 and ∞ and multiplying the result

by $N/2$ (the factor $1/2$ ensures that we only count each interaction once). The total energy is then given by:

$$E = \frac{3}{2} N k_B T + 2\pi N \rho \int_0^\infty r^2 v(r) g(r) dr \quad (6.16)$$

In a similar way the following expression for the pressure can be derived:

$$PV = N k_B T - \frac{2\pi N \rho}{3k_B T} \int_0^\infty r^2 r \frac{dv(r)}{dr} g(r) dr \quad (6.17)$$

It is usually more accurate to calculate such properties directly, partly because the radial distribution function is not obtained as a continuous function but is derived by dividing the space into small but discrete bins.

For molecules, the orientation must be taken into account if the true nature of the distribution is to be determined. The radial distribution function for molecules is usually measured between two fixed points, such as between the centres of mass. This may then be supplemented by an orientational distribution function. For linear molecules, the orientational distribution function may be calculated as the angle between the axes of the molecule, with values ranging from -180° to $+180^\circ$. For more complex molecules it is usual to calculate a number of site-site distribution functions. For example, for a three-site model of water, three functions can be defined ($g(O-O)$, $g(O-H)$ and $g(H-H)$). An advantage of the site-site models is that they can be directly related to information obtained from the X-ray scattering experiments. The O–O, O–H and H–H radial distribution functions have been particularly useful for refining the various potential models for simulating liquid water.

6.3 Phase Space

An important concept in computer simulation is that of the *phase space*. For a system containing N atoms, $6N$ values are required to define the state of the system (three coordinates per atom and three components of the momentum). Each combination of $3N$ positions and $3N$ momenta (usually denoted by Γ_N) defines a point in the $6N$ -dimensional phase space; an ensemble can thus be considered to be a collection of points in phase space. The way in which the system moves through phase space is governed by Hamiltonian's equations:

$$\frac{d\mathbf{r}_i}{dt} = \frac{\partial \mathcal{H}}{\partial \mathbf{p}_i} \quad (6.18)$$

$$\frac{d\mathbf{p}_i}{dt} = -\frac{\partial \mathcal{H}}{\partial \mathbf{r}_i} \quad (6.19)$$

where i varies from 1 to N . Molecular dynamics generates a sequence of points in phase space that are connected in time. These points correspond to the successive configurations of the system generated by the simulation. A molecular dynamics simulation performed in the microcanonical (constant NVE) ensemble will sample phase space along a contour of constant energy. There is no momentum component in a Monte Carlo simulation and such simulations sample from the $3N$ -dimensional space corresponding to the positions of

the atoms. It might seem odd that thermodynamic properties can be obtained from Monte Carlo simulations, given that there is no momentum contribution and so $3N$ degrees of freedom are not explored. In fact, all of the deviations from ideal gas behaviour are a consequence of interactions between the atoms and are encapsulated in the potential function, $\psi(\mathbf{r}^N)$, which only depends upon the positions of the atoms. A Monte Carlo simulation does sample from the positional degrees of freedom and so can be used to provide the deviations of thermodynamic properties from ideal gas behaviour, which is what we want to calculate. We shall return to this point in Chapter 8.

If it were possible to visit all the points in phase space then the partition function could be calculated by summing the values of $\exp(-E/k_B T)$. The phase-space trajectory in such a case would be termed *ergodic* and the results would be independent of the initial configuration. For the systems that are typical of those studied using simulation methods the phase space is immense, and an ergodic trajectory is not achievable (indeed, even for relatively small systems with only a few tens of atoms the time that would be required to cycle round all of the points in phase space is longer than the age of the universe). A simulation can thus only ever provide an estimate of the 'true' energies and other thermodynamic properties and so a sequence of simulations using different starting conditions would be expected to give similar, but different, results.

The thermodynamic properties that we have considered so far, such as the internal energy, the pressure and the heat capacity are collectively known as the mechanical properties and can be routinely obtained from a Monte Carlo or molecular dynamics simulation. Other thermodynamic properties are difficult to determine accurately without resorting to special techniques. These are the so-called entropic or thermal properties: the free energy, the chemical potential and the entropy itself. The difference between the mechanical and thermal properties is that the mechanical properties are related to the derivative of the partition function whereas the thermal properties are directly related to the partition function itself. To illustrate the difference between these two classes of properties, let us consider the internal energy, U , and the Helmholtz free energy, A . These are related to the partition function by:

$$U = \frac{k_B T^2}{Q} \frac{\partial Q}{\partial T} \quad (6.20)$$

$$A = -k_B T \ln Q \quad (6.21)$$

Q is given by Equation (6.4) for a system of identical particles. We shall ignore any normalisation constants in our treatment here to enable us to concentrate on the basics, and so it does not matter whether the system consists of identical or distinguishable particles. We also replace the Hamiltonian by the energy, E . The internal energy is obtained via Equation (6.20):

$$\begin{aligned} U &= k_B T^2 \frac{1}{Q} \iint d\mathbf{p}^N d\mathbf{r}^N \frac{E(\mathbf{p}^N, \mathbf{r}^N)}{k_B T^2} \exp(-E(\mathbf{p}^N, \mathbf{r}^N)/k_B T) \\ &= \iint d\mathbf{p}^N d\mathbf{r}^N E(\mathbf{p}^N, \mathbf{r}^N) \frac{\exp(-E(\mathbf{p}^N, \mathbf{r}^N)/k_B T)}{Q} \end{aligned} \quad (6.22)$$

Now consider the probability of the state with energy $E(\mathbf{p}^N, \mathbf{r}^N)$:

$$\frac{\exp(-E(\mathbf{p}^N, \mathbf{r}^N)/k_B T)}{Q} \quad (6.23)$$

This probability is written $\rho(\mathbf{p}^N, \mathbf{r}^N)$; the internal energy is thus given by

$$U = \iint d\mathbf{p}^N d\mathbf{r}^N E(\mathbf{p}^N, \mathbf{r}^N) \rho(\mathbf{p}^N, \mathbf{r}^N) \quad (6.24)$$

The crucial point about Equation (6.24) is that high values of $E(\mathbf{p}^N, \mathbf{r}^N)$ have a very low probability and make an insignificant contribution to the integral. The Monte Carlo and molecular dynamics methods preferentially generate states of low energy, which are the states that make a significant contribution to the integral in Equation (6.24). These methods sample from phase space in a way that is representative of the equilibrium state and are able to generate accurate estimates of properties such as the internal energy, heat capacity, and so on.

Let us now consider the problem of calculating the Helmholtz free energy of a molecular liquid. Our aim is to express the free energy in the same functional form as the internal energy, that is as an integral which incorporates the probability of a given state. First, we substitute for the partition function in Equation (6.21):

$$A = -k_B T \ln Q = k_B T \ln \left(\frac{N! h^{3N}}{\iint d\mathbf{p}^N d\mathbf{r}^N \exp(-E(\mathbf{p}^N, \mathbf{r}^N)/k_B T)} \right) \quad (6.25)$$

Next we recognise that the following integral is equal to 1:

$$1 = \frac{1}{(8\pi^2 V)^N} \iint d\mathbf{p}^N d\mathbf{r}^N \exp \left(-\frac{E(\mathbf{p}^N, \mathbf{r}^N)}{k_B T} \right) \exp \left(\frac{E(\mathbf{p}^N, \mathbf{r}^N)}{k_B T} \right) \quad (6.26)$$

Inserting this into the expression for the free energy and ignoring the constants (which act to change the zero point from which the free energy is calculated) gives:

$$A = k_B T \ln \left(\frac{\iint d\mathbf{p}^N d\mathbf{r}^N \exp \left(-\frac{E(\mathbf{p}^N, \mathbf{r}^N)}{k_B T} \right) \exp \left(+\frac{E(\mathbf{p}^N, \mathbf{r}^N)}{k_B T} \right)}{\iint d\mathbf{p}^N d\mathbf{r}^N \exp(-E(\mathbf{p}^N, \mathbf{r}^N)/k_B T)} \right) \quad (6.27)$$

We can now substitute for the probability density, $\rho(\mathbf{p}^N, \mathbf{r}^N)$ in this equation, leading to the final result (in which we have again ignored the normalisation factors):

$$A = k_B T \ln \left(\iint d\mathbf{p}^N d\mathbf{r}^N \exp \left(+\frac{E(\mathbf{p}^N, \mathbf{r}^N)}{k_B T} \right) \rho(\mathbf{p}^N, \mathbf{r}^N) \right) \quad (6.28)$$

The important feature of this result is that the configurations with a high energy make a significant contribution to the integral due to the presence of the exponential term $\exp(+E(\mathbf{p}^N, \mathbf{r}^N)/k_B T)$. A Monte Carlo or molecular dynamics simulation preferentially samples the *lower-energy* regions of phase space. An ergodic trajectory would, of course, visit all of these high-energy regions, but in practice these will never be adequately sampled

by a real simulation. The results for the free energy and other entropic properties will as a consequence be poorly converged and inaccurate.

To reiterate a point that we made earlier, these problems of accurately calculating the free energy and entropy do not arise for isolated molecules that have a small number of well-characterised minima which can all be enumerated. The partition function for such systems can be obtained by standard statistical mechanical methods involving a summation over the minimum energy states, taking care to include contributions from internal vibrational motion.

6.4 Practical Aspects of Computer Simulation

6.4.1 Setting Up and Running a Simulation

There are significant differences between the molecular dynamics and Monte Carlo simulation methods, but the same general strategies are used to set up and run either type of simulation. The first task is to decide which energy model is to be used to describe the interactions within the system. Simulations are usually performed with relatively large numbers of atoms over many iterations or time steps. The intra- and intermolecular interactions are therefore almost always described using an empirical (i.e. molecular mechanics) energy model. Faster computers and new theoretical techniques do now enable simulations to be performed using models based only on quantum mechanics or mixed models based on molecular mechanics/quantum mechanics as discussed in Section 11.13. Having chosen an energy model, the simulation itself can be broken into four distinct stages. First, an initial configuration for the system must be established. An *equilibration phase* is then performed, during which the system evolves from the initial configuration. Thermodynamic and structural properties are monitored during the equilibration until stability is achieved. Several distinct steps may be required during the equilibration, particularly for inhomogeneous systems. At the end of the equilibration the *production phase* commences. It is during the production phase that simple properties of the system are calculated. At regular intervals the configuration of the system (i.e. the atomic coordinates) is output to a disk file. Finally, the simulation is analysed; properties not calculated during the simulation are determined and the configurations are examined, not only to discover how the structure of the system changed but also to check for any unusual behaviour that might indicate a problem with the simulation.

6.4.2 Choosing the Initial Configuration

Before a simulation can be performed it is obviously necessary to select an initial configuration of the system. This should be done with some care, as the initial arrangement can often determine the success or failure of a simulation. For simulations of systems at equilibrium (the most common sort) it is wise to choose an initial configuration that is close to the state which it is desired to simulate. For example, it would be unwise to initiate a simulation of a face-centred cubic solid from a body-centred cubic starting point. It is also good practice to ensure that the initial configuration does not contain any high-energy

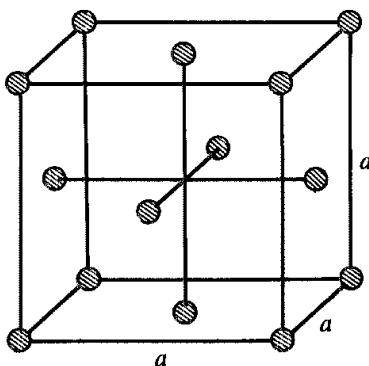


Fig. 6.3 The face-centred cubic cell.

interactions as these may cause instabilities in the simulation. Such 'hot spots' can often be eradicated by performing energy minimisation prior to the simulation itself.

To simulate homogeneous liquids which contain large numbers of the same molecule, a standard lattice structure is often chosen as the starting configuration. If an experimentally determined arrangement is available (e.g. an X-ray structure) then this could be used, provided that it was appropriate to the simulation being performed. When no experimental structure is available the initial configuration can be chosen from one of the common crystallographic lattices (simply placing molecules at random can often give rise to high-energy overlaps and instabilities). The most common lattice is the face-centred cubic lattice (fcc), shown in Figure 6.3. This structure contains $4M^3$ points ($M = 2, 3, 4, \dots$) For this reason, simulations are often performed using 108, 256, 525, 784, ..., atoms or molecules. The lattice size is chosen so that the density is appropriate to that of the system under study. For simulations of molecules it is also necessary to assign an orientation to each molecule. For small linear molecules, the solid structure of CO₂ is often chosen as the initial configuration. This is a face-centred cubic lattice with the molecules oriented in a regular fashion along the four diagonals of the unit cell. Alternatively, the orientations may be chosen completely at random or by making small random changes from the orientation in a regular lattice. At high densities non-physical overlaps may result, particularly if the molecules are large; in such cases it is more important to use an initial configuration that is close to the expected equilibrium distribution. For example, simulations of rod-shaped molecules such as liquid crystals are usually initiated from a configuration in which the molecules are all aligned approximately in the same direction.

For simulations of inhomogeneous systems comprising a solute molecule or intermolecular complex immersed in a solvent, the starting conformation of the solute may be obtained from an experimental technique such as X-ray crystallography or NMR, or may be generated by theoretical modelling. The coordinates of some solvent molecules may be known if the structure is obtained from X-ray crystallography, but it is usually necessary to add other solvent molecules to give the appropriate solvent density. A typical approach is to use the coordinates obtained from a previous simulation of the pure solvent. The solute is immersed in the solvent 'bath' and any solvent molecules that are too close to the solute are then discarded before the calculation proceeds.

6.5 Boundaries

The correct treatment of boundaries and boundary effects is crucial to simulation methods because it enables 'macroscopic' properties to be calculated from simulations using relatively small numbers of particles. The importance of boundary effects can be illustrated by considering the following simple example. Suppose we have a cube of volume 1 litre which is filled with water at room temperature. The cube contains approximately 3.3×10^{25} molecules. Interactions with the walls can extend up to 10 molecular diameters into the fluid. The diameter of the water molecule is approximately 2.8 Å and so the number of water molecules that are interacting with the boundary is about 2×10^{19} . So only about one in 1.5 million water molecules is influenced by interactions with the walls of the container. The number of particles in a Monte Carlo or molecular dynamics simulation is far fewer than 10^{25} – 10^{26} and is frequently less than 1000. In a system of 1000 water molecules most, if not all of them, would be within the influence of the walls of the boundary. Clearly, a simulation of 1000 water molecules in a vessel would not be an appropriate way to derive 'bulk' properties. The alternative is to dispense with the container altogether. Now, approximately three-quarters of the molecules would be at the surface of the sample rather than being in the bulk. Such a situation would be relevant to studies of liquid drops, but not to studies of bulk phenomena.

6.5.1 Periodic Boundary Conditions

Periodic boundary conditions enable a simulation to be performed using a relatively small number of particles, in such a way that the particles experience forces as if they were in bulk fluid. Imagine a cubic box of particles which is replicated in all directions to give a periodic array. A two-dimensional box is shown in Figure 6.4. In the two-dimensional example each box is surrounded by eight neighbours; in three dimensions each box would have 26 nearest neighbours. The coordinates of the particles in the image boxes can be computed simply by adding or subtracting integral multiples of the box sides. Should a particle leave the box during the simulation then it is replaced by an image particle that enters from the opposite side, as illustrated in Figure 6.4. The number of particles within the central box thus remains constant.

The cubic cell is the simplest periodic system to visualise and to program. However, a cell of a different shape might be more appropriate for a given simulation. This may be particularly important for simulations of systems which comprise a single molecule or intermolecular complex surrounded by solvent molecules. In such systems it is usually the behaviour of the central solute molecule that is of most interest and so it is desirable that as little of the computer time as possible is spent simulating the solvent far from the solute. In principle, any cell shape can be used provided it fills all of space by translation operations of the central box in three dimensions. Five shapes satisfy this condition: the cube (and its close relation, the parallelepiped), the hexagonal prism, the truncated octahedron, the rhombic dodecahedron and the 'elongated' dodecahedron (Figure 6.5) [Adams 1983]. It is often sensible to choose a periodic cell that reflects the underlying geometry of the system. For example, a rectangular cell is not the ideal choice to simulate an approximately spherical molecule.

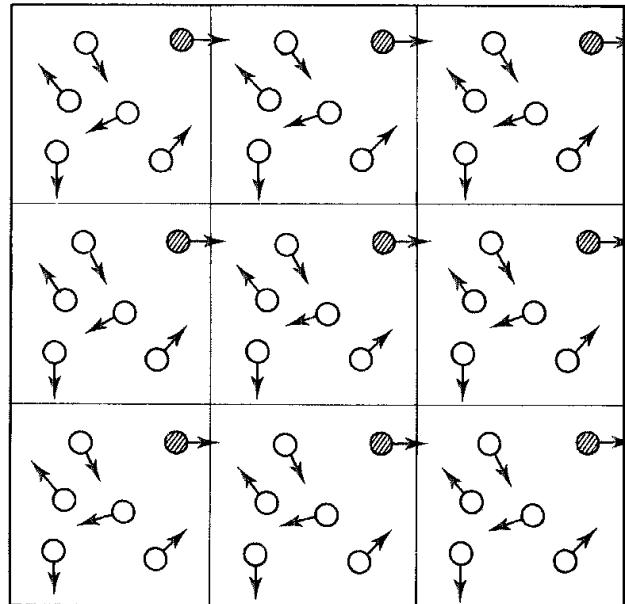


Fig. 6 4· Periodic boundary conditions in two dimensions

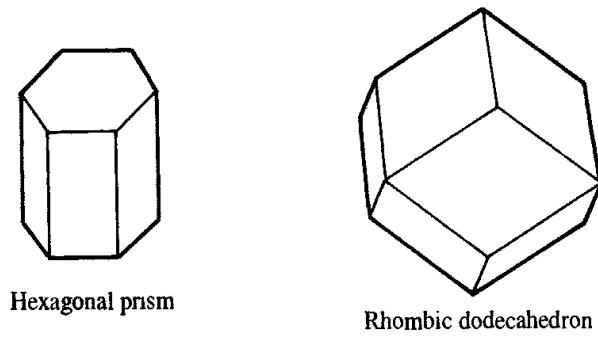
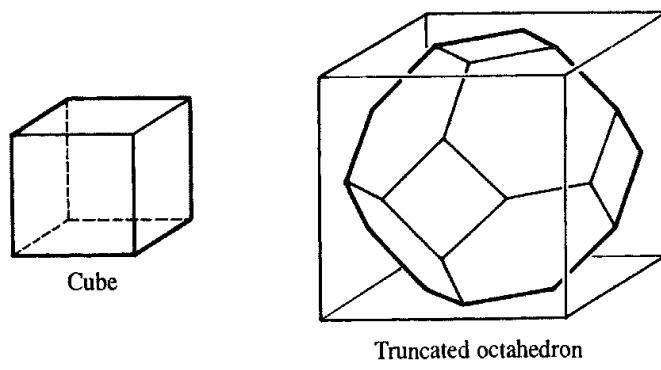


Fig. 6 5. Periodic cells used in computer simulations. the cube, truncated octahedron, hexagonal prism and rhombic dodecahedron

The truncated octahedron and the rhombic dodecahedron provide periodic cells that are approximately spherical and so may be more appropriate for simulations of spherical molecules. The distance between adjacent cells in the truncated octahedron or the rhombic dodecahedron is larger than the conventional cube for a system with a given number of particles and so a simulation using one of the spherical cells will require fewer particles than a comparable simulation using a cubic cell. Of the two approximately spherical cells, the truncated octahedron is often preferred as it is somewhat easier to program. The hexagonal prism can be used to simulate molecules with a cylindrical shape such as DNA.

Of the five possible shapes, the cube/parallelepiped and the truncated octahedron have been most widely used, with some simulations in the hexagonal prism. The formulae used to translate a particle back into the central simulation box for these three shapes are given in Appendix 6.4. It may be preferable to use one of the more common periodic cells even if there are aesthetic reasons for using an alternative. This is because the expressions for calculating the images may be difficult and inefficient to implement, even though the simulation would use fewer atoms.

For some simulations it is inappropriate to use standard periodic boundary conditions in all directions. For example, when studying the adsorption of molecules onto a surface, it is clearly inappropriate to use the usual periodic boundary conditions for motion perpendicular to the surface. Rather, the surface is modelled as a true boundary, for example by explicitly including the atoms in the surface. The opposite side of the box must still be treated; when a molecule strays out of the top side of the box it is reflected back into the simulation cell, as indicated in Figure 6.6. Usual periodic boundary conditions apply to motion parallel to the surface.

Periodic boundaries are widely used in computer simulations, but they do have some drawbacks. A clear limitation of the periodic cell is that it is not possible to achieve fluctuations that have a wavelength greater than the length of the cell. This can cause problems in certain situations, such as near the liquid-gas critical point. The range of the interactions present in the system is also important; if the cell size is large compared with the range over which the interactions act then there should be no problems. For example, for the relatively short-range Lennard-Jones potential the cell should have a side greater than approximately 6σ , which corresponds to about 20 Å for argon. For longer-range electrostatic interactions the situation is more difficult and it is often necessary to accept that some long-range order will be imposed upon the system. The effects of imposing a periodic boundary can be evaluated empirically by comparing the results of simulations performed using a variety of cell shapes and sizes.

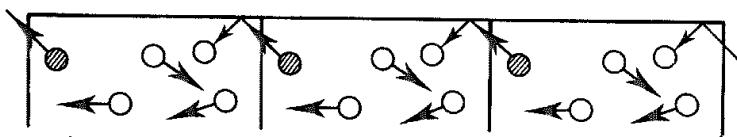


Fig 6.6 Periodic boundary conditions for surface simulations (Figure adapted from Allen M P and D J Tildesley 1987 Computer Simulation of Liquids. Oxford, Oxford University Press)

6.5.2 Non-periodic Boundary Methods

Periodic boundary conditions are not always used in computer simulations. Some systems, such as liquid droplets or van der Waals clusters, inherently contain a boundary. Periodic boundary conditions may also cause difficulties when simulating inhomogeneous systems or systems that are not at equilibrium. In other cases the use of periodic boundary conditions would require a prohibitive number of atoms to be included in the simulation. This particularly arises in the study of the structural and conformational behaviour of macromolecules such as proteins and protein-ligand complexes. The first simulations of such systems ignored all solvent molecules due to the limited computational resources then available. This corresponds to the unrealistic situation of simulating an isolated protein *in vacuo* and then comparing the results with experimental data obtained in solution. Vacuum calculations can lead to significant problems. A vacuum boundary tends to minimise the surface area and so may distort the shape of the system if it is non-spherical. Small molecules may adopt more compact conformations when simulated *in vacuo* due to favourable intramolecular electrostatic and van der Waals interactions, which would be damped in the presence of a solvent.

As computer power has increased it has become possible to incorporate explicitly some solvent molecules and thereby simulate a more realistic system. The simplest way to do this is to surround the molecule with a 'skin' of solvent molecules. If the skin is sufficiently deep then the system is equivalent to a solute molecule inside a 'drop' of solvent. The number of solvent molecules in such cases is usually significantly fewer than would be required in the analogous periodic boundary simulation, where the solute molecule is positioned at the centre of the cell and the empty space is filled with solvent. Boundary effects should be transferred from the molecule–vacuum interface to the solvent–vacuum interface and so might be expected to result in a more realistic treatment of the solute. To illustrate these three situations, we can consider dihydrofolate reductase, which is a small enzyme that contains approximately 2500 atoms. If this enzyme is surrounded by water molecules in a cubic periodic system such that the surface of the protein is at least 10 Å from any side of the box, then the number of atoms rises to almost 20 000. If a shell 10 Å thick is used then the number of atoms falls to 14 700, and with a 5 Å shell the system contains 8900 atoms.

Sometimes we are only interested in a specific part of the solute, such as the active site of an enzyme. It has been common practice in such cases to divide the system into two regions (Figure 6.7). One region, often called the *reaction zone*, contains all atoms or groups within a given radius R_1 of the site of interest. The atoms in the reaction zone are subjected to the full simulation method. The second region (the reservoir region) contains all atoms outside the reaction zone but within a distance R_2 of the active site. The atoms in the reservoir region may be kept fixed in their initial positions, or may be restrained so that they stay within the shell defined by R_1 and R_2 . Alternatively, they may be restrained to their initial positions using a harmonic potential. Any atoms further away from the active site than R_2 are discarded or may be kept fixed in their initial positions. It is important to be aware that restraining or fixing atoms in this way may prevent natural changes occurring and so lead to artificial behaviour. A variety of schemes for performing simulations using such *stochastic boundary conditions* have been proposed. However, such methods can be rather complicated to implement and if not used properly can give anomalous results.

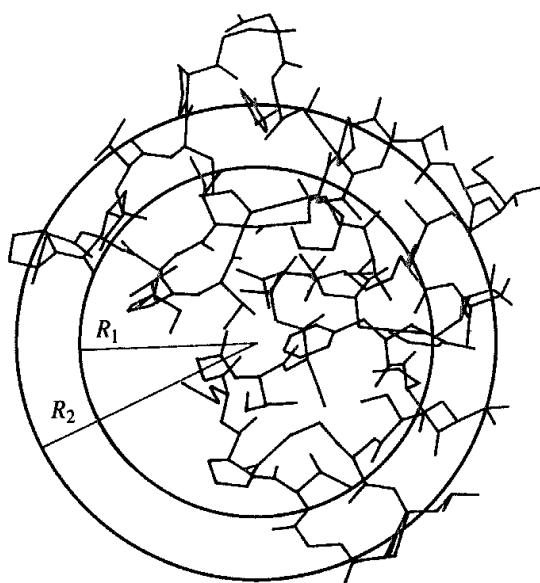


Fig. 6.7 Division into reaction zone and reservoir regions in a simulation using stochastic boundary conditions

If at all possible, a periodic boundary is the ‘safest’ way to ensure that boundary effects are minimised, but sometimes an alternative may be the only practical course.

6.6 Monitoring the Equilibration

The purpose of the equilibration phase is to enable the system to evolve from the starting configuration to reach equilibrium. Equilibration should continue until the values of a set of monitored properties become stable. The properties to be monitored usually include thermodynamic quantities such as the energy, temperature and pressure and also structural properties. Many simulations of the liquid state involve a starting configuration that corresponds to a solid lattice. It is therefore important to establish that the lattice has ‘melted’ before the production phase begins. *Order parameters* can be used to determine that the liquid state has been reached. An order parameter is a measure of the degree of order (or, equivalently, disorder) in the system. During a simulation of a crystal lattice the atoms would be expected to remain in approximately the same positions throughout and thereby maintain a high degree of order. In a liquid, however, we would expect considerable mobility of the species present, giving rise to translational disorder. One way to measure translational order in a system initially in a face-centred cubic lattice was suggested by Verlet, whose order parameter λ is:

$$\lambda = \frac{1}{3}[\lambda_x + \lambda_y + \lambda_z] \quad (6.29)$$

$$\lambda_x = \frac{1}{N} \sum_{i=1}^N \cos\left(\frac{4\pi x_i}{a}\right) \quad (6.30)$$

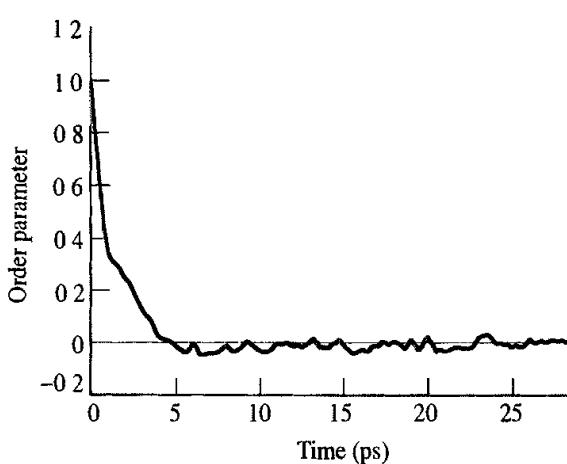


Fig. 6.8. Variation in Verlet order parameter during the equilibration phase of a molecular dynamics simulation of argon

where a is the length of one edge of the unit cell. Initially, all of the coordinates x_i , y_i and z_i are multiples of $a/2$ and so the order parameter has a value of 1. As the simulation proceeds the order parameter should gradually decrease to a value of zero, indicating that the atoms are distributed randomly. When equilibrium has been reached the fluctuations in the order parameter should be proportional to $1/\sqrt{N}$, where N is the size of the system. A typical result is shown in Figure 6.8 for an argon simulation.

For molecules, it is also necessary to consider their orientations, which can be monitored using a rotational order parameter. For some systems, such as carbon monoxide or water, complete disorder would be expected in the liquid state at equilibrium. However, if we were simulating a dense fluid of rod-shaped molecules which form a liquid crystalline phase then we might expect that, on average, the molecules would tend to line up in a common direction. The Viellard-Baron rotational order parameter for linear molecules is calculated using the following formula:

$$P_1 = \frac{1}{N} \sum_{i=1}^N \cos \gamma_i \quad (6.31)$$

where γ_i is the angle between the current and original direction of the molecular axis of molecule i . A value of 1 indicates that the molecules are perfectly aligned. Rotational disorder is indicated by a value of zero. The fluctuations about the average value should again be proportional to $1/\sqrt{N}$. For non-linear molecules, a number of rotational order parameters can be defined and each monitored.

The *mean squared displacement* also provides a means to establish whether a solid lattice has melted. The mean squared displacement is given by:

$$\Delta r^2(t) = \frac{1}{N} \sum_{i=1}^N [\mathbf{r}_i(t) - \mathbf{r}_i(0)]^2 \quad (6.32)$$

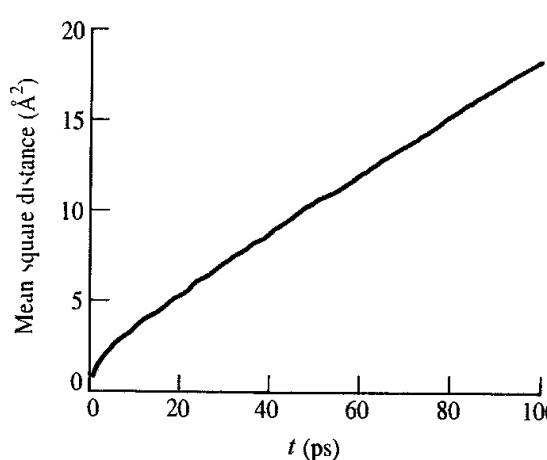


Fig 6.9: Variation in mean squared displacement during the initial steps of a molecular dynamics simulation of argon

For a fluid, with no underlying regular structure, the mean squared displacement gradually increases with time (Figure 6.9). For a solid, however, the mean squared displacement typically oscillates about a mean value. However, if there is diffusion within a solid then this can be detected from the mean squared displacement and may be restricted to fewer than three dimensions. For example, Figure 6.10 shows the mean squared displacement calculated for Li^+ ions in Li_3N at 400 K [Wolf *et al.* 1984]. This material contains layers of Li_2N ; mobility of the Li^+ ions is much greater within these planes than perpendicular to them.

The radial distribution function can also be used to monitor the progress of the equilibration. This function is particularly useful for detecting the presence of two phases. Such a situation is characterised by a larger than expected first peak and by the fact that $g(r)$ does not decay towards a value of 1 at long distances. If two-phase behaviour is inappropriate then the simulation should probably be terminated and examined. If, however, a two-phase system is desired, then a long equilibration phase is usually required.

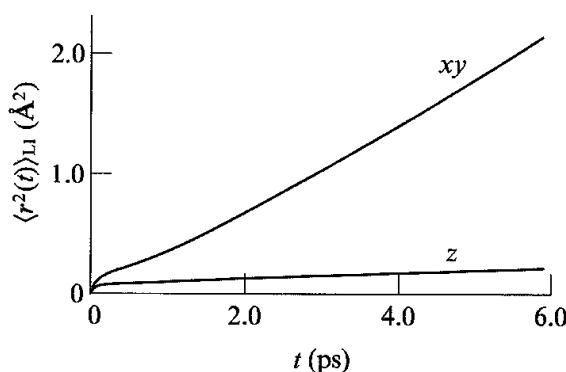


Fig 6.10: Mean squared displacement for Li^+ ions in Li_3N for motion parallel (xy) and perpendicular (z) to the Li_2N layers [Wolf *et al.* 1984].

6.7 Truncating the Potential and the Minimum Image Convention

The most time-consuming part of a Monte Carlo or molecular dynamics simulation (or, indeed, of an energy minimisation) is the calculation of the non-bonded energies and/or forces. The numbers of bond-stretching, angle-bending and torsional terms in a force field model are all proportional to the number of atoms but the number of non-bonded terms that need to be evaluated increases as the square of the number of atoms (for a pairwise model) and is thus of order N^2 . In principle, the non-bonded interactions are calculated between every pair of atoms in the system. However, for many interaction models this is not justified. The Lennard-Jones potential falls off very rapidly with distance: at 2.5σ the Lennard-Jones potential has just 1% of its value at σ . This reflects the r^{-6} distance dependence of the dispersion interaction. The most popular way to deal with the non-bonded interactions is to use a *non-bonded cutoff* and to apply the *minimum image convention*. In the minimum image convention, each atom 'sees' at most just one image of every other atom in the system (which is repeated infinitely via the periodic boundary method). The energy and/or force is calculated with the closest atom or image, as illustrated in Figure 6.11. When a cutoff is employed, the interactions between all pairs of atoms that are further apart than the cutoff value are set to zero, taking into account the closest image. When periodic boundary conditions are being used, the cutoff should not be so large that a particle sees its own image, or indeed the same molecule twice. This has the effect of limiting the cutoff to no more than half the length of the cell when simulating atomic fluids in a cubic cell. For rectangular cells the cutoff should be no greater than half the length of the shortest side. For simulations of molecules the upper limit on the cutoff may also be affected by the size of the molecules, as we shall see below in Section 6.7.2. In simulations where the Lennard-Jones potential is the only non-bonded interaction, a cutoff of 2.5σ gives rise to a relatively small error. However, when long-range electrostatic

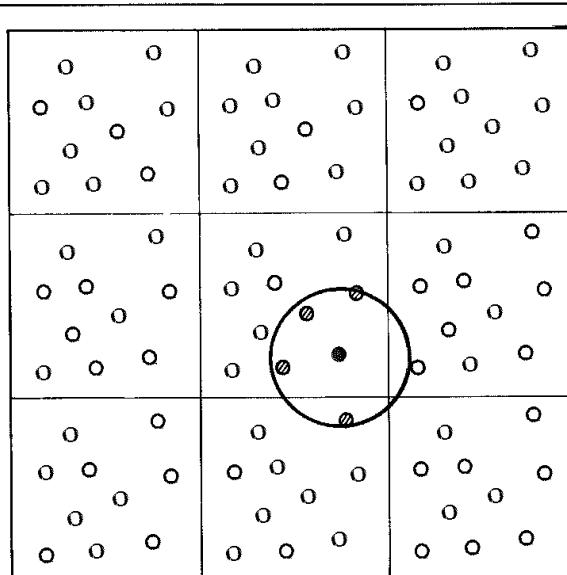


Fig. 6.11: The spherical cutoff and the minimum image convention

interactions are involved, the cutoff should be much greater and indeed there is evidence to suggest that using any cutoff leads to errors. A value of at least 10 Å is generally recommended but even this may be insufficient. More comprehensive methods have been devised for dealing with the electrostatic interactions, which are considered in Section 6.8.

6.7.1 Non-bonded Neighbour Lists

By itself, the use of a cutoff may not dramatically reduce the time taken to compute the number of non-bonded interactions. This is because we would still have to calculate the distance between every pair of atoms in the system simply to decide whether they are close enough to calculate their interaction energy. Calculating all the $N(N - 1)$ distances takes almost as much time as calculating the energy itself.

In simulations of fluids, an atom's neighbours (i.e. those atoms that are within the cutoff distance) do not change significantly over 10 or 20 molecular dynamics time steps or Monte Carlo iterations. If we 'knew' which atoms to include in the non-bonded calculation (for example, by storing them in an array), then it would be possible to identify directly each atom's neighbours without having to calculate the distances to all the other atoms in the system. The *non-bonded neighbour* list (first proposed by Verlet) is just such a device. The Verlet neighbour list [Verlet 1967] stores all atoms within the cutoff distance, together with all atoms that are slightly further away than the cutoff distance. This is most efficiently done using a large neighbour list array, L , and a pointer array, P . The pointer array indicates where in the neighbour list the first neighbour for that atom is located. The last neighbour of atom i is stored in element $P[i + 1] - 1$ of the neighbour list as shown in Figure 6.12. Thus the neighbours of atom i are stored in elements $L[P[i]]$ through $L[P[i + 1] - 1]$ of the array L . The neighbour list is updated at regular intervals throughout the simulation. Between updates, the neighbour and pointer lists are used to directly identify the nearest neighbours of each atom i . The distance used to calculate each atom's neighbours should be larger than the actual non-bonded cutoff distance so that no atom, initially outside the neighbour cutoff, approaches closer than the non-bonded cutoff distance before the neighbour list is updated again.

It is important to update the neighbour list at the correct frequency. If the update frequency is too high the procedure is inefficient; too low and the energies and forces may be calculated incorrectly due to atoms moving within the non-bonded cutoff region. An update frequency between 10 and 20 steps is common. An algorithm that can automatically update the neighbour list and so circumvent these problems is as follows [Thompson 1983]. An array element is set to zero for each atom whenever the neighbour list is updated. This array is used to store the displacement of each atom or molecule in subsequent steps. When the sum of the maximum displacements of any two atoms exceeds the difference between the non-bonded cutoff distance and the neighbour list distance, then it is time to update the neighbour list again.

There are no fixed rules that determine how much larger the neighbour list cutoff should be than the non-bonded cutoff. Clearly there will be a trade-off between the size of the cutoff and the frequency at which the neighbour list must be updated: the larger the difference, the less frequently will the neighbour list have to be updated. There may also be storage implications if the list is too large.

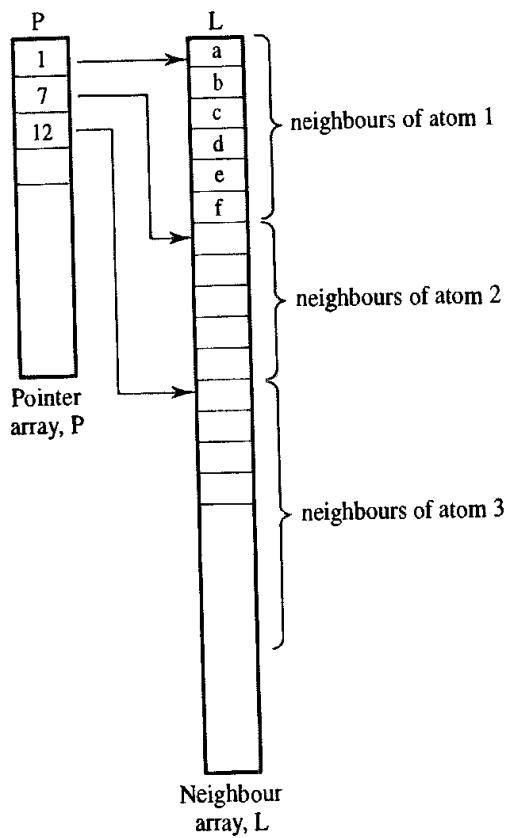


Fig. 6.12 Pointer and neighbour arrays can be used to implement the Verlet neighbour list.

When the number of molecules in the simulation is very large, it can require a significant computational effort just to update the neighbour list. This is because the standard way to update the neighbour list requires the distance between all pairs of atoms in the system to be calculated. When the size of the system is much larger than the cutoff distance, a *cell index method* can be used to make the updating procedure more efficient. In the cell index method, the simulation box is divided into a number of cells. The length of each cell is longer than the non-bonded cutoff distance. All of the neighbours of an atom will then be found either in the cell containing the atom or in one of the surrounding cells. If the entire system is divided into M^3 cells, there will be an average of N/M^3 molecules in each cell. To determine the neighbours of a given atom or molecule, it is then necessary to consider just $27N/M^3$ atoms rather than N . The cell index method requires a mechanism for identifying the atoms or molecules in each cell. Two arrays can be used to do this: a linked list array L and a pointer array P . The pointer array indicates the location of one of the atoms or molecules in a given cell. Thus, $P[1]$ would indicate the number of the 'first' atom or molecule in cell 1 and $P[2]$ is the number of the 'first' atom in cell 2. Each element of the linked list array then gives the number of the 'next' atom or molecule in the cell. Thus the value stored in $L[1]$ is the number of the second atom in the first cell. Suppose $P[1]$ is atom 10. Then the value stored in $L[10]$ is the second atom in the cell. If this second atom is number 15, then $L[15]$ contains the third atom in the cell. The last molecule in the sequence is identified by the fact that its array element is zero. The cell index method clearly

requires a mechanism for updating the pointer and linked list arrays when atoms or molecules move from one cell to another, which can add to the complexity.

When simulating species with a significant electrostatic contribution, it may be desirable to use different cutoffs for the electrostatic and van der Waals interactions. This is because the electrostatic interaction has a much longer range. Using a longer cutoff for the electrostatic interactions will, of course, significantly increase the number of pairs that must be calculated. A compromise is to use a *twin-range method*, in which two cutoffs are specified. All interactions below the lower cutoff are calculated as normal at each step. Interactions due to atoms between the lower and upper cutoffs are evaluated only when the neighbour list is updated and are kept constant between these updates. The rationale here is that the contribution of the atoms that are further away will not vary much between updates.

The use of a cutoff is amply justified in many cases, if only on the grounds of expediency, but its use will always lead to some fraction of the potential energy being ignored. This lost energy can be easily captured at the end of the simulation if it is assumed that the radial distribution function takes the value of 1 at distances greater than the cutoff. The calculation is analogous to that used to determine the total energy from the radial distribution function, Equation (6.16), but the integration is now performed between the cutoff distance r_c and infinity and $g(r)$ is now taken to be 1 in this range. For N particles, the correction is:

$$E_{\text{correction}} = 2\pi\rho N \int_{r_c}^{\infty} r^2 v(r) dr \quad (6.33)$$

For the Lennard-Jones potential the long-range contribution can be determined analytically:

$$E_{\text{correction}} = 8\pi\rho N \varepsilon \left[\frac{\sigma^{12}}{9r^9} - \frac{\sigma^6}{3r^3} \right] \quad (6.34)$$

6.7.2 Group-based Cutoffs

When simulating large molecular systems, it is often advantageous to use a group-based cutoff (sometimes called a residue-based cutoff). Here, the large molecules are divided into ‘groups’, each of which contains a relatively small number of connected atoms. If the calculation involves small solvent molecules then each solvent molecule is also conveniently regarded as a single unconnected group. Why are groups useful? Let us consider the electrostatic interaction between two water molecules. In the popular TIP3P model there is a charge of $-0.834e$ on the oxygen and $0.417e$ on each hydrogen. The electrostatic interaction between two water molecules is calculated as the sum of nine distinct site-site interactions. If we start from the minimum energy arrangement for the water dimer shown in Figure 6.13 and

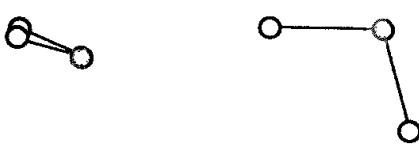


Fig. 6.13 Minimum energy structure for water dimer with TIP3P model

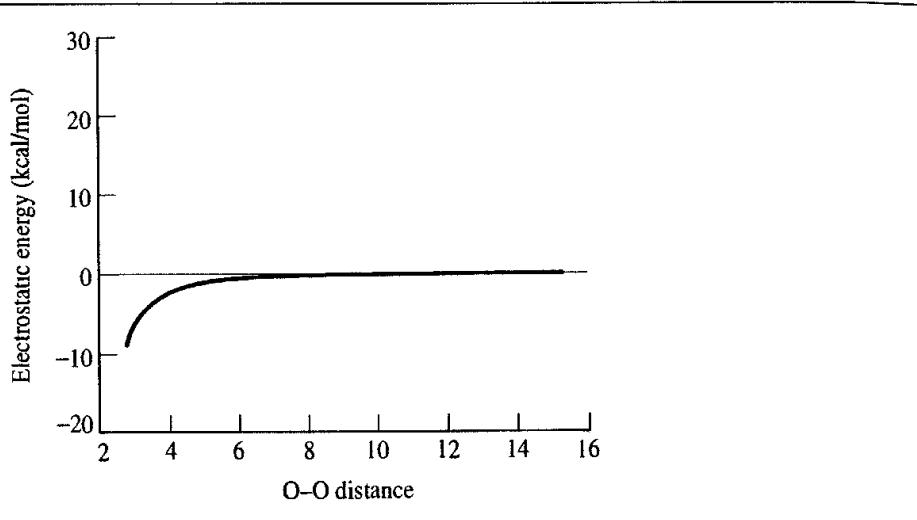


Fig. 6.14: The variation in the electrostatic interaction energy of the water dimer as a function of the O-O distance without a cutoff

gradually move one water molecule relative to the other as indicated then the electrostatic energy varies as shown in the graph in Figure 6.14.

Although the overall interaction energy is relatively small beyond 6 Å or so, each of these energies is the sum of several rather large terms; for example, at an O-O separation of 8 Å, the overall interaction energy is about -0.27 kcal/mol but this comprises an oxygen-oxygen interaction of approximately 29 kcal/mol, oxygen-hydrogen interactions of -59.4 kcal/mol and hydrogen-hydrogen interactions of 29.2 kcal/mol. Suppose that a simple atom-based non-bonded cutoff is applied to the water dimer. The interaction energy then fluctuates violently near the cutoff distance, as shown in Figure 6.15 for a cutoff of 8 Å. This is because only some of the pairwise interactions are included in this case. Clearly such a model would almost certainly lead to serious problems for any

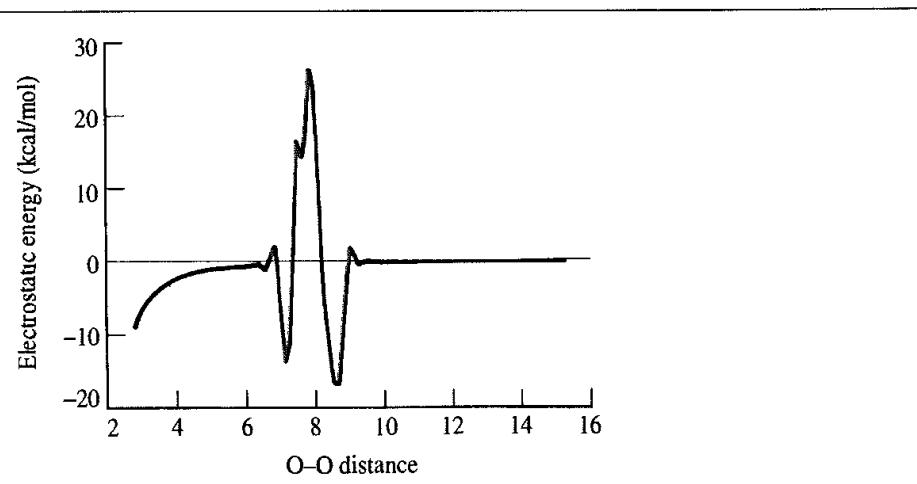


Fig. 6.15: The variation in interaction of the water dimer as a function of the O-O distance with an 8 Å atom-based cutoff

simulation. This problem can be avoided by collecting all of the atoms from each water molecule into a single group, and by calculating the interactions on a group-group basis, even though some of the atom pairs may have a separation larger than the cutoff.

How should a molecule be divided into groups? In some cases there may appear to be a chemically obvious way to define the groups, especially when the molecule is a polymer that is constructed from distinct chemical residues. A particularly desirable feature is that each group should, if possible, be of zero charge. The reason for this can be understood if we recall how the different electrostatic interactions vary with distance from Section 4.9.1:

- charge-charge $\sim 1/r$
- charge-dipole $\sim 1/r^2$
- dipole-dipole $\sim 1/r^3$
- dipole-quadrupole $\sim 1/r^4$
- charge-induced dipole $\sim 1/r^4$
- dipole-induced dipole $\sim 1/r^6$

If the groups are electrically neutral, then the leading term in the electrostatic interaction between a pair of groups is the dipole-dipole interaction, which is dependent upon $1/r^3$. By comparison, the charge-charge terms vary as $1/r$. Of course, it is not always possible to arrange atoms in neutral groups as occurs when some of the species are charged.

A further question with the group-based scheme is: how do we decide whether a particular group-group interaction needs to be considered? In other words, how are cutoffs included in the group scheme? One strategy is to include a particular group-group interaction if any pair of atoms in the two groups is closer than the cutoff distance. Alternatively, a 'marker' atom may be nominated within the group; when the marker atoms come closer than the cutoff then the appropriate group-group interaction is included. When using marker atoms, it is important that the groups are not too large; thus the groups used by some simulation programs are much smaller than the 'chemically obvious' groupings. For example, the most obvious choice for proteins and peptides is to define each entire amino acid residue as a single group. However, this is not necessarily the most appropriate strategy. Consider the situation in which two arginine residues are spatially close together (Figure 6.16). Arginine has a long side chain that is comparable in length to the non-bonded cutoff distances often employed. Suppose the alpha-carbon atom (marked C_α in Figure 6.16)

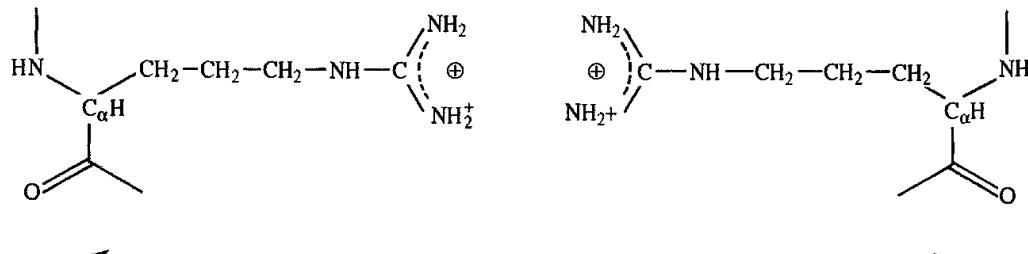


Fig 6.16 The use of a marker atom on the alpha-carbon in an arginine residue may lead to a significant electrostatic interaction being neglected because the distance between the marker atoms exceeds the cutoff

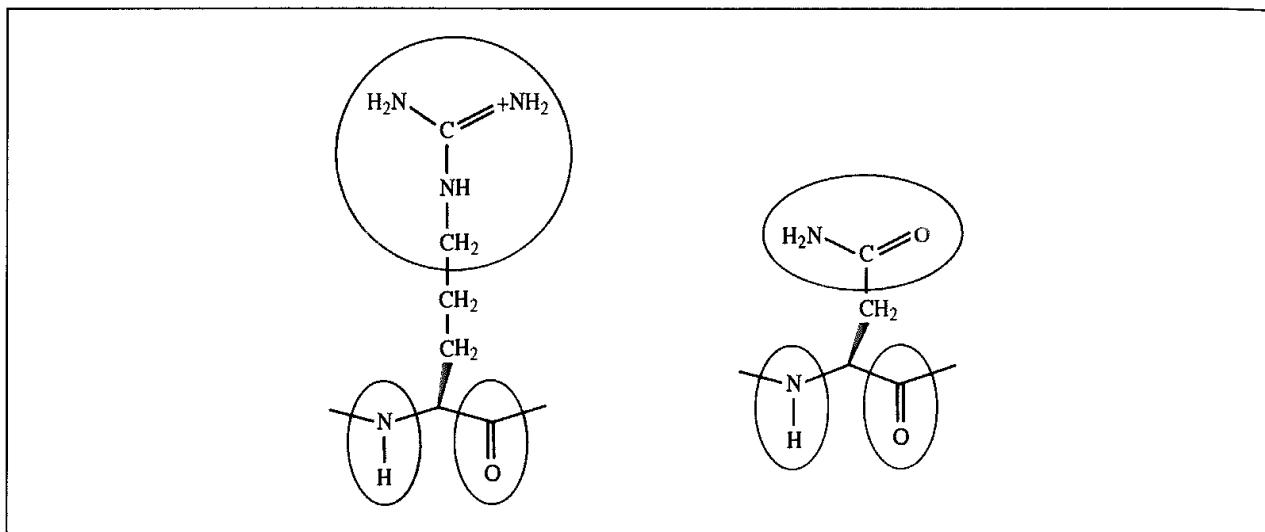


Fig 6.17: The charge groups used in the GROMOS simulation program for simulating proteins [van Gunsteren and Berendsen 1986], illustrated using the amino acids arginine and asparagine. The CH_2 groups have zero charge.

in each arginine residue is chosen as the marker atom. If the distance between the alpha-carbons of two arginine residues is greater than the cutoff, no interactions between any atoms in the arginine residues would be calculated, despite the fact that the positively charged ends of the residues could be very close, as shown in Figure 6.16. Were the alpha-carbons to approach closer than the cutoff, there would then be a dramatic increase in the energy due to the unfavourable interaction between the two side chains, inevitably leading to an unstable simulation. It may therefore be appropriate to define ‘charge groups’ that contain smaller numbers of atoms than are in the chemically obvious scheme. For example, the groups that are used by the GROMOS simulation program for the amino acids arginine and asparagine are shown in Figure 6.17.

6.7.3 Problems with Cutoffs and How to Avoid Them

A cutoff introduces a discontinuity in both the potential energy and the force near the cutoff value. This creates problems, especially in molecular dynamics simulations where energy conservation is required. There are several ways that the effects of this discontinuity can be counteracted. One approach is to use a shifted potential, in which a constant term is subtracted from the potential at all values (Figure 6.18):

$$\nu'(r) = \nu(r) - \nu_c \quad r \leq r_c \quad (6.35)$$

$$\nu'(r) = 0 \quad r > r_c \quad (6.36)$$

where r_c is the cutoff distance and ν_c is equal to the value of the potential at the cutoff distance. As the additional term is constant, it disappears when the potential is differentiated and so does not affect the force calculation in molecular dynamics. Use of the shifted potential does improve energy conservation, though as the number of atom pairs separated by a distance smaller than the cutoff varies, so the contribution of the shifted potential to the total energy will change. An additional problem is that there is a discontinuity in the force

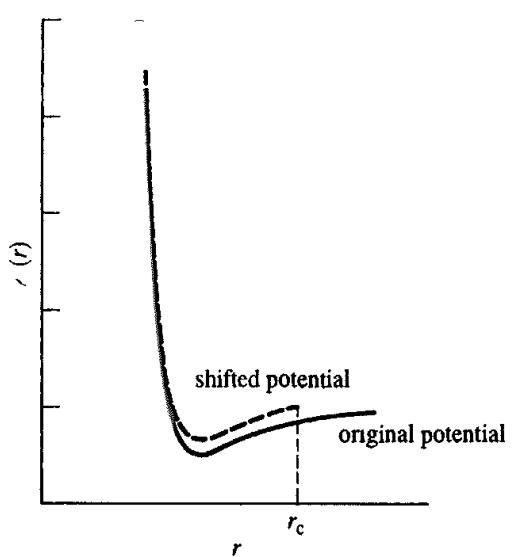


Fig 6.18 A shifted Lennard-Jones potential.

with the shifted potential; at the cutoff distance, the force will have a finite value which drops suddenly to zero just beyond the cutoff. This can also give instabilities in a simulation. To avoid this, a linear term can be added to the potential, making the derivative zero at the cutoff:

$$\nu'(r) = \nu(r) - \nu_c - \left(\frac{d\nu(r)}{dr} \right)_{r=r_c} (r - r_c) \quad r \leq r_c \quad (6.37)$$

$$\nu'(r) = 0 \quad r > r_c \quad (6.38)$$

The shift makes the potential deviate from the 'true' potential, and so any calculated thermodynamic properties will be changed. The 'true' values can be retrieved but it is difficult to do so, and the shifted potential is thus rarely used in 'real' simulations. Moreover, while it is relatively straightforward to implement for a homogeneous system under the influence of a simple potential such as the Lennard-Jones potential, it is not easy for inhomogeneous systems containing many different types of atom.

An alternative way to eliminate discontinuities in the energy and force equations is to use a *switching function*. A switching function is a polynomial function of the distance by which the potential energy function is multiplied. Thus the switched potential $\nu'(r)$ is related to the true potential $\nu(r)$ by $\nu'(r) = \nu(r)S(r)$. Some switching functions are applied to the entire range of the potential up to the cutoff point. One such function is:

$$\nu'(r) = \nu(r) \left[1 - 2\left(\frac{r}{r_c}\right)^2 + \left(\frac{r}{r_c}\right)^4 \right] \quad (6.39)$$

The switching function has a value of 1 at $r = 0$ and a value of 0 at $r = r_c$, the cutoff distance. Between these two values it varies as shown in Figure 6.19, which also shows how the potential function is affected.

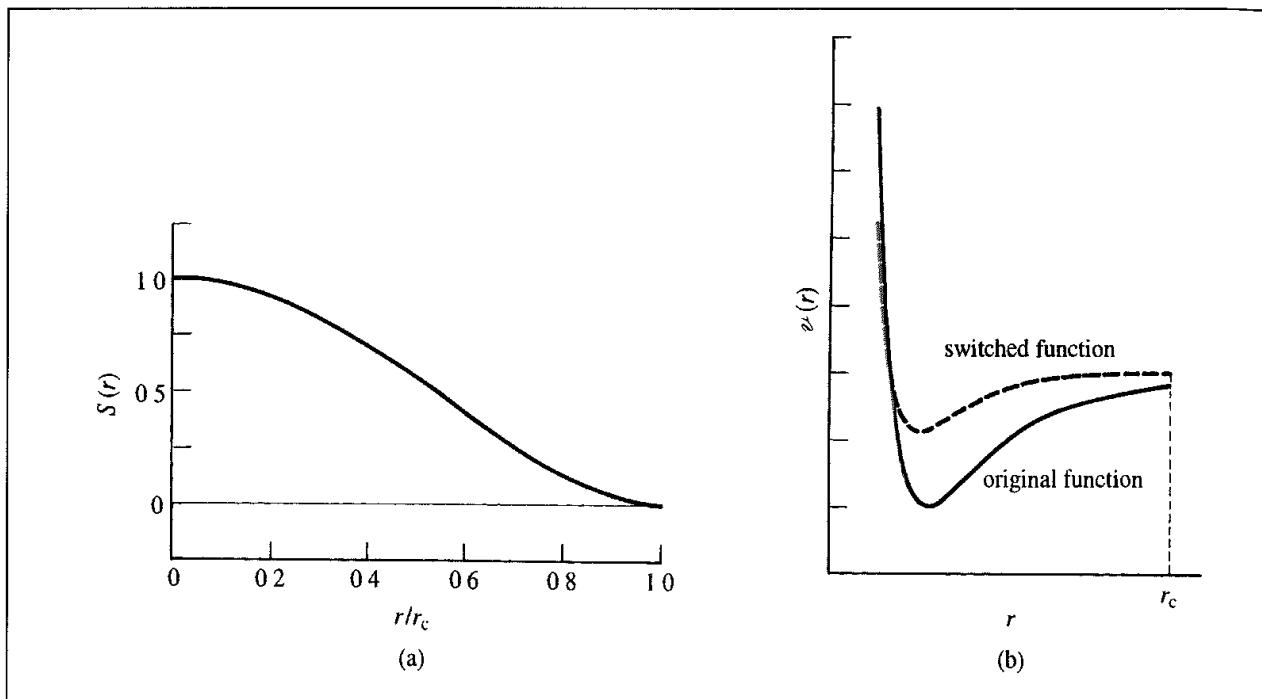


Fig. 6.19 (a) The effect of a switching function that applies over the entire range and (b) its effect on the Lennard-Jones potential

A switching function applied to the potential function over the entire range does have drawbacks; for example, equilibrium structures are affected (the minimum energy separation for the argon dimer decreases slightly). A more acceptable alternative is to gradually taper the potential between two cutoff values. The potential takes its usual value until the lower cutoff distance. Between the lower (r_l) and upper cutoff distance (r_u) the potential is multiplied by the switching function, which takes the value 1 at the lower cutoff distance and 0 at the upper cutoff distance. The lower cutoff distance is typically relative close to the upper cutoff distance (for example, r_l might be 9 Å and r_u 10 Å). A simple switching function has the following linear form:

$$S = 1.0 \quad r_{ij} < r_l \quad (6.40)$$

$$S = (r_u - r_{ij}) / (r_u - r_l) \quad r_l \leq r_{ij} \leq r_u \quad (6.41)$$

$$S = 0.0 \quad r_u < r_{ij} \quad (6.42)$$

This suffers from discontinuities in both the energy and the force at the two cutoff values. A more acceptable switching function smoothly changes from a value of 1 to a value of 0 (Figure 6.20) between r_l and r_u and satisfies the following requirements:

$$S_{r=r_l} = 1; \quad \left(\frac{dS}{dr} \right)_{r=r_l} = 0; \quad \left(\frac{d^2S}{dr^2} \right)_{r=r_l} = 0 \quad (6.43)$$

$$S_{r=r_u} = 0; \quad \left(\frac{dS}{dr} \right)_{r=r_u} = 0; \quad \left(\frac{d^2S}{dr^2} \right)_{r=r_u} = 0 \quad (6.44)$$

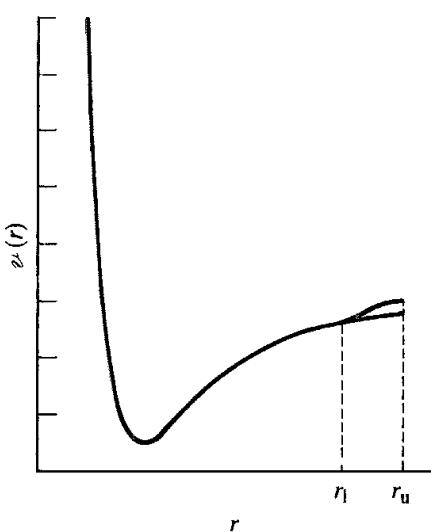


Fig. 6.20 A switching function that applies over a narrow range near the cutoff and its effect on the Lennard-Jones potential.

By ensuring that the first derivative is zero at the endpoints the force also approaches zero smoothly. A continuous second derivative is required to ensure that the integration algorithm works properly. If the switch function is assumed to take the following form:

$$S(r) = c_0 + c_1 \left[\frac{r - r_l}{r_u - r_l} \right] + c_2 \left[\frac{r - r_l}{r_u - r_l} \right]^2 + c_3 \left[\frac{r - r_l}{r_u - r_l} \right]^3 + c_4 \left[\frac{r - r_l}{r_u - r_l} \right]^4 + c_5 \left[\frac{r - r_l}{r_u - r_l} \right]^5 \quad (6.45)$$

then the following values of the coefficients $c_0 \dots c_5$ satisfy the six requirements in equations (6.43) and (6.44):

$$c_0 = 1; \quad c_1 = 0; \quad c_2 = 0; \quad c_3 = -10; \quad c_4 = 15; \quad c_5 = -6 \quad (6.46)$$

When using a switching function in a molecular simulation with a residue-based cutoff it is important that the function has the same value for all pairs of atoms in the two interacting groups. Otherwise, severe fluctuations in the energy can arise when the separation is within the cutoff region. These two contrasting situations can be formally expressed as follows:

$$\text{atom based: } \mathcal{V}_{AB} = \sum_{i=1}^{N_A} \sum_{j=1}^{N_B} S_{ij}(r_{ij}) v_{ij}(r_{ij}) \quad (6.47)$$

$$\text{residue or molecule based: } \mathcal{V}_{AB} = S_{AB}(|\mathbf{r}_A - \mathbf{r}_B|) \sum_{i=1}^{N_A} \sum_{j=1}^{N_B} v_{ij}(r_{ij}) \quad (6.48)$$

N_A and N_B are the numbers of atoms in the two groups A and B and S is the switching function. With the group-based switching function, it is necessary to define the 'distance' between the two groups (i.e. the two points \mathbf{r}_A and \mathbf{r}_B). There is no definitive way to do this. As with cutoffs, a special marker atom can be nominated within each residue, or the centre of mass, centre of geometry or centre of charge may be used.

Group-based switching functions have several advantages. Better energy conservation can be achieved, and there are advantages when performing energy minimisation, since the potential is defined analytically at all points. However, it is important to beware of possible problems with group-based switching functions when the groups are large. We have already seen how this can arise when an ordinary group-based cutoff is used. Let us re-examine our arginine problem (Figure 6.16) when a switching function is employed. When the two marker atoms have a separation only slightly less than the upper switch cutoff, the switching function will be close to zero, and so there will not be the dramatic increase in energy that is observed with the simple cutoff. Nevertheless, although the switching function does help to prevent the simulation from ‘blowing up’, the representation of the energy and the forces in the system is still unsatisfactory. The only real alternative is to make the groups smaller or dispense with cutoffs altogether.

6.8 Long-range Forces

Those interactions that decay no faster than r^{-n} , where n is the dimensionality of the system, can be a problem as their range is often greater than half the box length. The charge-charge interaction, which decays as r^{-1} , is particularly problematic in molecular simulations. There is much evidence that it is important to properly model these long-range forces, which are particularly acute when simulating charged species such as molten salts (when it is not possible to construct neutral groups). A proper treatment of long-range forces can also be important when calculating certain properties, such as the dielectric constant. One way to tackle the errors introduced by an inadequate treatment of long-range forces would be to use a much larger simulation cell, but this is usually impractical. Nevertheless, increasing computer power does mean that more rigorous ways of dealing with long-range forces can be considered, even in simulations of large systems. A variety of methods have been developed to handle long-range forces. The methods that we will discuss in detail are the Ewald summation, the reaction field method and the cell multiple method.

6.8.1 The Ewald Summation Method

The Ewald sum was first devised by Ewald [Ewald 1921] to study the energetics of ionic crystals. In this method, a particle interacts with all the other particles in the simulation box and with all of their images in an infinite array of periodic cells. Figure 6.21 illustrates how the array of simulation cells is constructed; in the limit, the cell array is considered to have a spherical shape. The position of each image box (assumed for simplicity to be a cube of side L containing N charges) can be related to the central box by specifying a vector, each of whose components is an integral multiple of the length of the box, $(\pm iL, \pm jL, \pm kL)$; $i, j, k = 0, 1, 2, 3$, etc. The charge-charge contribution to the potential energy due to all pairs of charges in the central simulation box can be written:

$$\mathcal{V} = \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \quad (6.49)$$

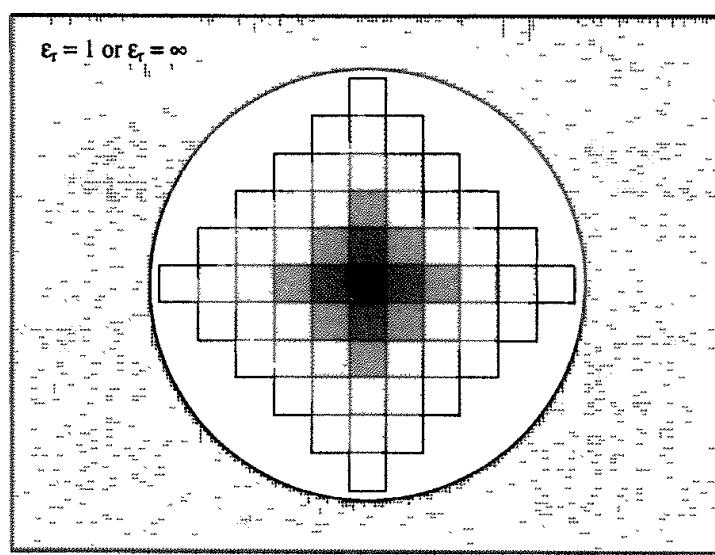


Fig. 6.21: The construction of a system of periodic cells in the Ewald method. (Figure adapted from Allen M P and D J Tildesley 1987 Computer Simulation of Liquids Oxford, Oxford University Press.)

where r_{ij} is the minimum distance between the charges i and j . There are six boxes at a distance L from the central box with coordinates $(\mathbf{r}_{\text{box}})$ given by $(0, 0, L)$, $(0, 0, -L)$, $(0, L, 0)$, $(0, -L, 0)$, $(L, 0, 0)$ and $(-L, 0, 0)$ (only four of these are shown in the two-dimensional picture in Figure 6.21). The contribution of the charge-charge interaction between the charges in the central box and all images of all particles in these six surrounding boxes is given by:

$$\mathcal{V} = \frac{1}{2} \sum_{n \text{ box}=1}^6 \sum_{i=1}^N \sum_{j=1}^N \frac{q_i q_j}{4\pi\epsilon_0 |\mathbf{r}_{ij} + \mathbf{r}_{\text{box}}|} \quad (6.50)$$

In general, for a box which is positioned at a cubic lattice point \mathbf{n} ($= (n_x L, n_y L, n_z L)$ with n_x, n_y, n_z being integers):

$$\mathcal{V} = \frac{1}{2} \sum_{\mathbf{n}} \sum_{i=1}^N \sum_{j=1}^N \frac{q_i q_j}{4\pi\epsilon_0 |\mathbf{r}_{ij} + \mathbf{n}|} \quad (6.51)$$

$|\mathbf{n}|$ thus takes the values $1, \sqrt{2}, \dots$. This expression is often written in such a way to incorporate the interactions between pairs of charges in the central box (for which $|\mathbf{n}| = 0$):

$$\mathcal{V} = \frac{1}{2} \sum_{|\mathbf{n}|=0}' \sum_{i=1}^N \sum_{j=1}^N \frac{q_i q_j}{4\pi\epsilon_0 |\mathbf{r}_{ij} + \mathbf{n}|} \quad (6.52)$$

The prime on the first summation indicates that the series does not include the interaction $i = j$ for $\mathbf{n} = 0$.

There is thus a contribution to the total energy from the interactions in the central box together with the interactions between the central box and all image boxes. There is also a contribution from the interaction between the spherical array of boxes and the surrounding

medium. The problem is that the summation in Equation (6.52) converges extremely slowly and in fact is *conditionally convergent*. A conditionally convergent series contains a mixture of positive and negative terms such that the positive terms alone form a divergent series (i.e. a series which does not have a finite sum) as do the negative terms when taken alone. The sum of a conditionally convergent series depends on the order in which its terms are considered. An additional problem with the Coulomb interaction is that it can vary rapidly at small distances.

The trick when calculating the Ewald sum is to convert the summation into two series, each of which converges much more rapidly. The mathematical foundation for this is the following identity:

$$\frac{1}{r} = \frac{f(r)}{r} + \frac{1-f(r)}{r} \quad (6.53)$$

The aim is thus to choose an appropriate function $f(r)$ which will deal with the rapid variation of $1/r$ at small r and the slow decay at long r . In the Ewald method each charge is considered to be surrounded by a neutralising charge distribution of equal magnitude but of opposite sign, as shown in Figure 6.22. A Gaussian charge distribution of the following functional form is commonly used:

$$\rho_i(\mathbf{r}) = \frac{q_i \alpha^3}{\pi^{3/2}} \exp(-\alpha^2 r^2) \quad (6.54)$$

The sum over point charges is now converted to a sum of the interactions between the charges *plus* the neutralising distributions. This dual summation (the 'real space'

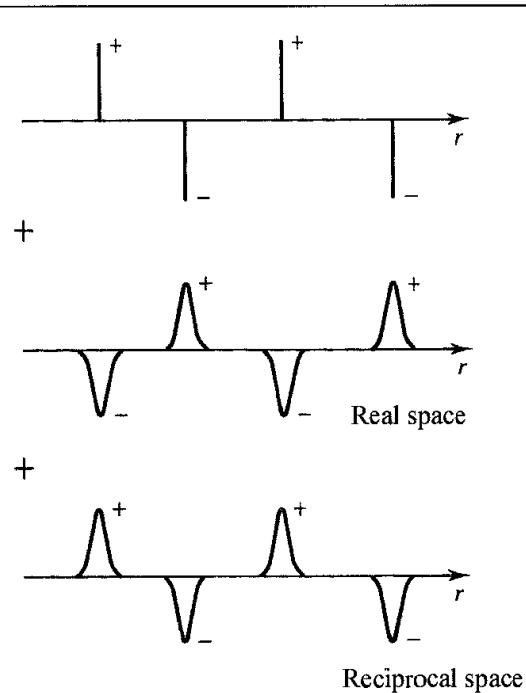


Fig 6.22 In the Ewald summation method the initial set of charges are surrounded by a Gaussian distribution (calculated in real space) to which a cancelling charge distribution must be added (calculated in reciprocal space)

summation) is given by:

$$\mathcal{V} = \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \sum'_{|\mathbf{n}|=0} \frac{q_i q_j}{4\pi\epsilon_0} \frac{\operatorname{erfc}(\alpha|\mathbf{r}_{ij} + \mathbf{n}|)}{|\mathbf{r}_{ij} + \mathbf{n}|} \quad (6.55)$$

erfc is the complementary error function, which is:

$$\operatorname{erfc}(x) = \frac{2}{\sqrt{\pi}} \int_x^\infty \exp(-t^2) dt \quad (6.56)$$

The Ewald method thus uses $\operatorname{erfc}(r)$ for the function $f(r)$ in Equation (6.53). The crucial point is that this new summation involving the error function converges very rapidly and beyond some cutoff distance its value can be considered negligible. The rate of convergence depends upon the width of the cancelling Gaussian distributions; the wider the Gaussian, the faster the series converges. Specifically, α should be chosen so that the only terms in the series (6.55) are those for which $|\mathbf{n}| = 0$ (i.e. only pairwise interactions involving charges in the central box, or if a cutoff is used α is chosen so that only interactions with other charges within the cutoff are included). A second charge distribution is now added to the system which exactly counteracts the first neutralising distribution (Figure 6.22). The contribution from this second charge distribution is:

$$\mathcal{V} = \frac{1}{2} \sum_{k \neq 0} \sum_{i=1}^N \sum_{j=1}^N \frac{1}{\pi L^3} \frac{q_i q_j}{4\pi\epsilon_0} \frac{4\pi^2}{k^2} \exp\left(-\frac{k^2}{4\alpha^2}\right) \cos(\mathbf{k} \cdot \mathbf{r}_{ij}) \quad (6.57)$$

This summation is performed in *reciprocal space*, the details of which need not concern us here. The vectors \mathbf{k} are reciprocal vectors and are given by $\mathbf{k} = 2\pi\mathbf{n}/L$. This reciprocal sum also converges much more rapidly than the original point-charge sum. However, the number of terms that must be included increases with the width of the Gaussians. There is thus a clear need to balance the real-space and reciprocal-space summations; the former converges more rapidly for large α , whereas the latter converges more rapidly for small α . A value for α of $5/L$ and 100–200 reciprocal vectors \mathbf{k} have been suggested as providing acceptable results. This reciprocal space summation corresponds to the second term $([1 - f(r)]/r)$ in Equation (6.53); the requirement for this term is that it is a slowly varying function for all r . As such, its Fourier transform (which is what the summation is) can be represented by a small number of reciprocal vectors. The sum of Gaussian functions in real space includes the interaction of each Gaussian with itself. A third self-term must therefore be subtracted:

$$\mathcal{V} = -\frac{\alpha}{\sqrt{\pi}} \sum_{k=1}^N \frac{q_k^2}{4\pi\epsilon_0} \quad (6.58)$$

A fourth correction term may also be required, depending upon the medium that surrounds the sphere of simulation boxes. If the surrounding medium has an infinite relative permittivity (e.g. if it is a conductor) then no correction term is required. However, if the surrounding medium is a vacuum (with a relative permittivity of 1) then the following energy must be added:

$$\mathcal{V}_{\text{correction}} = \frac{2\pi}{3L^3} \left| \sum_{i=1}^N \frac{q_i}{4\pi\epsilon_0} \mathbf{r}_i \right|^2 \quad (6.59)$$

The final expression is thus:

$$\mathcal{V} = \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \left\{ \begin{array}{l} \sum_{|\mathbf{n}|=0}^{\infty} \frac{q_i q_j}{4\pi\epsilon_0} \frac{\operatorname{erfc}(\alpha|\mathbf{r}_{ij} + \mathbf{n}|)}{|\mathbf{r}_{ij} + \mathbf{n}|} \\ + \sum_{k \neq 0} \frac{1}{\pi L^3} \frac{q_i q_j}{4\pi\epsilon_0} \frac{4\pi^2}{k^2} \exp\left(-\frac{k^2}{4\alpha^2}\right) \cos(\mathbf{k} \cdot \mathbf{r}_{ij}) \\ - \frac{\alpha}{\sqrt{\pi}} \sum_{k=1}^N \frac{q_k^2}{4\pi\epsilon_0} + \frac{2\pi}{3L^3} \left| \sum_{k=1}^N \frac{q_k}{4\pi\epsilon_0} r_k \right|^2 \end{array} \right\} \quad (6.60)$$

The Ewald sum is the most ‘correct’ way yet devised to accurately include all the effects of long-range forces in a computer simulation. It has been extensively used in simulations involving highly charged systems (such as ionic melts and in studies of processes in and on solids) and is increasingly being applied to other systems where electrostatic effects are important, such as lipid bilayers, proteins and DNA. Nevertheless, the Ewald method is not without problems and it tends to reinforce artefacts that arise from imposing periodic boundary conditions. For example, the method artificially results in each charge-charge interaction being minimised at a separation of half the box length. Instantaneous fluctuations in the simulation cell tend to be replicated throughout the infinite system rather than being damped out.

The Ewald summation is computationally quite expensive to implement. Under conditions of constant α (which will give the same density of reciprocal vectors, \mathbf{k}) then it scales as the square of the number of particles in the central simulation cell. If α is allowed to vary then the algorithm can be made to scale as $N^{3/2}$ though the consequent value of α might make the range of the Coulomb potential incompatible with the range of the van der Waals interactions. Several methods have been proposed to speed up the computationally demanding reciprocal space part of the calculation, such as the use of polynomial approximations but these do not solve the unfavourable N^2 scaling. The most promising way to tackle this difficulty is to modify the problem so that the fast Fourier transform (FFT) can be used to compute the reciprocal space summation. The fast Fourier algorithm scales as $N \ln N$, which gives considerable advantages over the N^2 alternative. If, in addition, a sufficiently large value of α is chosen such that the interatomic interaction is negligible for r_{ij} greater than a cutoff (e.g. 9 Å) then the real-space summation is reduced to order N and the order of the entire algorithm becomes $N \ln N$.

As outlined in Section 1.10.8, the FFT method requires that the data are not continuous but are discrete values. In order to employ the fast Fourier transform in the Ewald summation the point charges with their continuous coordinates must be replaced by a grid-based charge distribution. Each of the atomic point charges must thus be distributed among the surrounding grid points in some fashion so as to reproduce the potential of the charge at the original location. As usual an element of compromise is required; the more surrounding points that are used the more accurately the potential of the charge at the original location can be approximated but the greater the computational cost per particle. A popular approach is the particle-mesh method of Hockney and Eastwood [Hockney and Eastwood 1988], which uses the nearest 27 points in three dimensions. From this gridded charge density it is possible

to calculate (through use of the FFT algorithm) the potential due to the Gaussian distributions at the grid points, which by interpolation gives rise to the desired potential at (and thus the forces on) each of the particles. A number of variants on this general theme have been described, all of which use the fast Fourier transform algorithm but which differ in other aspects of their implementation. These include the particle-mesh Ewald method [Darden *et al.* 1993] and the particle-particle-particle-mesh approach [Hockney and Eastwood 1988; Luty *et al.* 1994, 1995]. Deserno and Holm presented a unification of these methods and also demonstrated that although very similar in spirit they could have very different accuracies [Deserno and Holm 1998a, b]. The particle-particle-particle-mesh approach was generally preferred as it was believed to be more flexible.

The Ewald method has been widely used to study highly polar or charged systems. Its use is considered routine for many types of solid-state materials. It is increasingly used for calculations on much larger molecular systems, such as proteins and DNA, due both to the increases in computer performance and to the new methodological advances we have just discussed [Darden *et al.* 1999]. For example, an early application of the particle-mesh Ewald method was the molecular dynamics simulation of a crystal of the protein bovine pancreatic trypsin inhibitor [York *et al.* 1994]. The full crystal environment was reproduced, with four protein molecules in the unit cell, together with associated water molecules and chloride counterions. Over the course of the 1 ns simulation the deviation of the simulated structures from the initial crystallographic structure was monitored. Once equilibrium was achieved this deviation (measured as the root-mean-square positional deviation) settled down to a value of 0.63 Å for all non-hydrogen atoms and 0.52 Å for the backbone atoms alone. By contrast, an equivalent simulation run with a 9 Å residue-based cutoff showed a deviation of more than 1.8 Å. In addition, the atomic fluctuations calculated from the Ewald simulation were in close agreement with those derived from the crystallographic temperature factors, unlike the non-Ewald simulation, which was significantly overestimated due to the use of the electrostatic cutoff. The highly charged nature of DNA makes it particularly important to deal properly with the electrostatic interactions and simulations using the particle-mesh Ewald approach are often much more stable, with the trajectories remaining much closer to the experimental structures [Cheatham *et al.* 1995].

6.8.2 The Reaction Field and Image Charge Methods

In the reaction field method, a sphere is constructed around the molecule with a radius equal to the cutoff distance. The interaction with molecules that are within the sphere is calculated explicitly. To this is added the energy of interaction with the medium beyond the sphere, which is modelled as a homogeneous medium of dielectric constant ϵ_s (Figure 6.23). The electrostatic field due to the surrounding dielectric is given by:

$$\mathbf{E}_i = \frac{2(\epsilon_s - 1)}{\epsilon_s + 1} \left(\frac{1}{r_c^3} \right) \sum_{j: r_j \leq r_c} \boldsymbol{\mu}_j \quad (6.61)$$

where $\boldsymbol{\mu}_j$ are the dipoles of the neighbouring molecules that are within the cutoff distance (r_c) of the molecule i . The interaction between the molecule i and the reaction field equals $\mathbf{E}_i \cdot \boldsymbol{\mu}_i$,

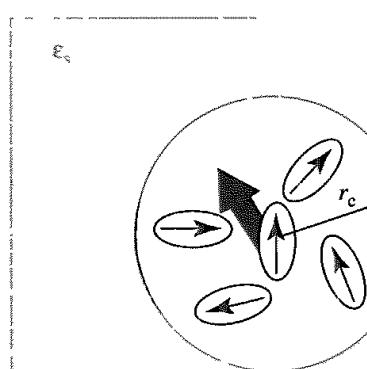


Fig. 6.23 The reaction field method. The shaded arrow represents the sum of the dipoles of the other molecules within the cutoff sphere.

which is added to the short-range molecule–molecule interaction. Problems with the reaction field method may arise from discontinuities in the energy and/or force when the number of molecules j within the cavity of the molecule i changes. These problems can be avoided by employing a switching function for molecules that are near the reaction field boundary.

Similar approaches employ a single boundary for the entire system. This boundary may be spherical or may have a more complicated shape that better approximates the true molecular surface of the molecule. In the *image charge method*, a spherical boundary is employed and the reaction field due to a charge inside the boundary is considered to arise from a so-called image charge situated in the continuous dielectric beyond the sphere (Figure 6.24) [Friedman 1975]. If the position of the charge is \mathbf{r}_i , then the image charge is located at $(R/r_i)^2 \mathbf{r}_i$ (where R is the radius of the bounding sphere) and has magnitude:

$$q_{im} = -\frac{(\epsilon_s - \epsilon_r)}{(\epsilon_s + \epsilon_r)} \frac{q_i R}{r_i} \quad (6.62)$$

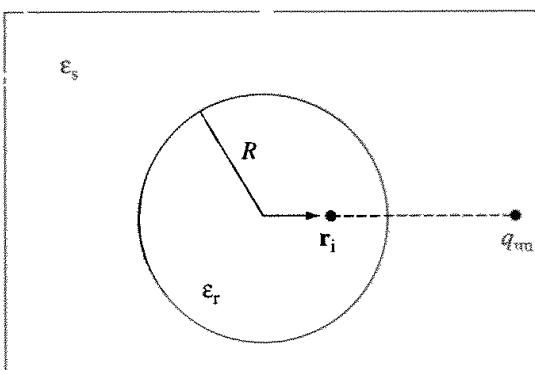


Fig. 6.24: The image charge method

where ϵ_r and ϵ_s are the dielectric constants inside and outside the boundary, respectively. This expression holds if the dielectric constant beyond the boundary is much greater than that inside ($\epsilon_s \gg \epsilon_r$). A drawback with this method is that as a charge approaches the boundary, so too does its image and the method breaks down.

The reaction field and image charge methods have the advantages of being conceptually simple, relatively easy to implement and computationally efficient. However, they do rely upon the assumption that molecules beyond the cutoff can be modelled as a continuous dielectric. This is not necessarily the case but is often a reasonable assumption for homogeneous fluids. A value for the dielectric of the surrounding continuum must also be specified. This can be taken from experimental data, but the dielectric constant may be the property that one is trying to calculate! In practice, it is often necessary only to ensure that $\epsilon_r \leq \epsilon_s \leq \infty$. There are several ways in which the dielectric constant can be calculated from a computer simulation. A common approach is via the average of the square of the total dipole moment of the system, $\langle \mathbf{M}^2 \rangle$. With a reaction field boundary the dielectric constant ϵ_r is given by:

$$\frac{4\pi}{9} \frac{\langle \mathbf{M}^2 \rangle}{V k_B T} = \frac{(\epsilon_r - 1)}{3} \frac{(2\epsilon_s + 1)}{(2\epsilon_s + \epsilon_r)} \quad (6.63)$$

where V is the volume of the simulation system. Even though the value of $\langle \mathbf{M}^2 \rangle$ can vary quite considerably with the reaction field dielectric ϵ_s , almost identical values of ϵ_r are obtained. An alternative approach is to determine the polarisation response of the liquid to an electric field E_0 . If the average dipole moment per unit volume along the direction of the applied field is $\langle \mathbf{P} \rangle$ then the dielectric constant is given by:

$$\frac{4\pi}{3} \frac{\langle \mathbf{P} \rangle}{E_0} = \frac{(\epsilon_r - 1)}{3} \frac{(2\epsilon_s + 1)}{(2\epsilon_s + \epsilon_r)} \quad (6.64)$$

This perturbation method is claimed to be more efficient than the fluctuating dipole method, at least for certain water models [Alper and Levy 1989], but it is important to ensure that the polarisation $\langle \mathbf{P} \rangle$ is linear in the electric field strength to avoid problems with dielectric saturation.

6.8.3 The Cell Multipole Method for Non-bonded Interactions

The cell multipole method (also called the fast multipole method) is an algorithm that enables *all* $N(N - 1)$ pairwise non-bonded interactions to be enumerated in a time that scales linearly with N , rather than N^2 , as in the standard Ewald approach [Greengard and Roklin 1987; Ding *et al.* 1992a, b; Greengard 1994]. The cell multipole method can be used to evaluate interactions that can be expressed in the following general form:

$$\sum_i \sum_{j>i} \frac{q_i q_j}{|\mathbf{r}_i - \mathbf{r}_j|^p} \quad (6.65)$$

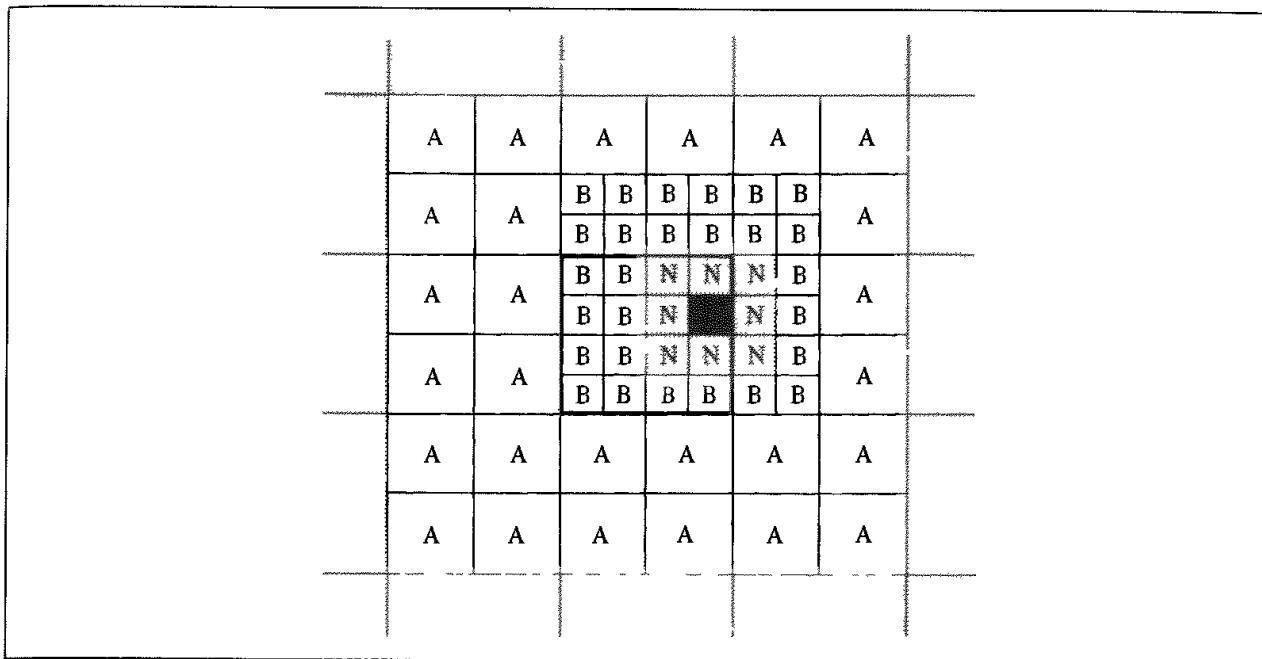
Both the Coulomb and Lennard-Jones potentials can be considered examples of this type. In the cell multipole method the simulation space is divided into uniform cubic

cells. The multipole moments (charge, dipole, quadrupole) of each cell are then calculated by summing over the atoms contained within the cell. The interaction between all of the atoms in the cell and another atom outside the cell (or indeed another cell) can then be calculated using an appropriate multipole expansion (see Section 4.9.1).

This multipole expansion is only valid if the separation between the interacting particles (be they atoms, molecules or cells) is larger than the sum of the radii of convergence of the multipoles. In the cell multipole method, the multipole expansion is used for interactions that are more than one cell distance away. For interactions that are within one cell distance the usual atomic pairwise interaction method is employed.

Consider an atom in a cell, C_0 . The interactions with atoms in nearby cells are calculated using the usual pairwise formulae. There are 27 such cells (i.e. the cell in which the atom is positioned and the surrounding 26 cells). The interaction between the atom and all of the atoms in each of the faraway cells is then calculated using the multipole expansion. The potential due to a faraway cell will be approximately constant for all atoms in the cell of current interest, C_0 (the cells are usually small, containing on average four atoms). Thus the potential due to each faraway cell can be represented as a Taylor series expansion about the centre of C_0 . If there are M cells in total then there are $M - 27$ faraway cells; then the calculation of these cell-cell interactions for the entire system will be of order $M(M - 27)$. As the number of cells is approximately equal to the number of atoms, this still leaves us with a quadratic dependency upon the number of atoms present (though it does now vary as about $N^2/16$, if there is an average of four atoms per cell).

The algorithm can be converted to one which shows linear dependency by recognising that in the method we have just described, the interactions due to very faraway cells are calculated with the same accuracy as interactions with cells that are much closer. This level of accuracy can be considered unnecessary as any error is largely due to the closer cells. The small cells are thus grouped into larger cells, with the cell size increasing with the distance from the cell of interest, C_0 . The accuracy of the calculation remains approximately constant if the ratio of the cell size to the distance remains constant. This grouping scheme is illustrated in Figure 6.25. The multipoles for each of the larger cells are calculated by translating and adding the moments of its constituent smaller cells. The use of multipole expansions and Taylor series approximations does mean that there will be a degree of truncation error, though this can be reduced by simply including more terms in the multipole expansion. The cell multipole algorithm requires an amount of bookkeeping to keep track of the hierarchy of the cells, which means that up to a certain size of problem the exact N^2 algorithm is faster. The algorithm then suddenly switches to a linear dependence. There is some debate over the break-even point at which the cell multipole method is equally as fast as an N^2 method, with estimates ranging from 300 particles to 100 000. Another complication to the debate is the introduction of the fast Fourier transform Ewald methods with their $N \ln N$ scaling. Nevertheless, for calculations on systems with thousands, if not millions, of atoms, the cell multipole methods appear promising, especially as enhanced versions are developed [Petersen *et al.* 1994].



*Fig 6.25: The hierarchy of cells used in the cell multipole method. For an atom in the black cell, the interactions with atoms in the 26 nearby cells (N) are calculated explicitly. Interactions with the atoms in cells labelled A and B are calculated using a Taylor series multipole expansion. (Figure adapted from Ding H-Q, N Karasawa and W A Goddard III, 1992b *The Reduced Cell Multipole Method for Coulomb Interactions in Periodic Systems with Million-Atom Unit Cells* Chemical Physics Letters 196:6-10)*

6.9 Analysing the Results of a Simulation and Estimating Errors

A simulation can generate an enormous amount of data, which should be properly analysed to extract relevant properties and to check that the calculation has behaved properly. The primary reason for undertaking a particular simulation may be to calculate just a single physical or thermodynamic property or to investigate the conformational properties of a molecule. However, it is also advisable to check that other aspects of the simulation have performed as expected. Some properties can be calculated during the simulation itself (such as the energy and the virial), but it is often sensible not to impose too severe a burden on the simulation program itself. In part this is because many properties do not vary significantly from one step to another but can be calculated at less frequent intervals. The configurations (i.e. positions of each atom or molecule in the system) do not change much from one step to another in either a molecular dynamics or Monte Carlo simulation, and so it is usual to store configurations every 5–25 steps, depending upon the nature of the system and the disk space available. It is good practice to visually examine configurations selected from throughout the simulation to ensure that no strange or unexpected behaviour is present. In many simulations of molecular systems the major objective is to investigate the structural behaviour of the system rather than to calculate thermodynamic properties, and so the focus of the analysis will change accordingly.

A computer simulation is subject to error, and this error should be properly calculated and assessed. Of course, computers only do what they are told to by the programmer, and so a program will always give the same results for the same set of initial conditions (if not, some serious fault should be suspected!). The results of a computer simulation may be subject to two kinds of error, just as any other scientific experiment. These are systematic errors and statistical errors. Systematic error results in a constant bias from the ‘proper’ behaviour. The most obvious effect of a systematic error is to displace the average property from its proper value. Systematic errors are sometimes due to a fault in the simulation algorithm or in the energy model and may be relatively easy to spot, especially if they have an obvious or even catastrophic effect on the simulation. Systematic errors may also arise from approximations inherent in the algorithm, such as truncation (all finite difference methods used in molecular dynamics generate only an approximation to the true integral of the equations of motion) and round-off errors (due to the limited precision with which numbers can be stored in a computer). Such errors can be more difficult to detect. One way to detect systematic error is to compare the distribution of the values of simple thermodynamic properties about their average values. The distribution of such properties about their average values should be normal (i.e. Gaussian), such that the probability of finding a particular value for the property A is given by:

$$p(A) = \frac{1}{\sigma\sqrt{2\pi}} \exp[-(A - \langle A \rangle)^2/2\sigma^2] \quad (6.66)$$

where σ^2 is the variance, given by $\sigma^2 = \langle (A - \langle A \rangle)^2 \rangle$. The standard deviation is the square root of the variance. More information on these statistical terms can be found in Section 1.10.7

The chi-squared test can be used to provide a quantitative estimate of the deviation of a calculated distribution from that expected. Suppose that the value of some property (A) has been calculated from the simulation at regular intervals to give a total of M values. The average value of the property A is determined together with the standard deviation. The data, comprising all of the A values from the simulation, are then divided into bins such that the number of values in each bin (M_i) is approximately the same. The number of values that would be expected in the i th bin is:

$$n_i = \frac{M}{\sigma\sqrt{2\pi}} \int_{A_i - \Delta A/2}^{A_i + \Delta A/2} \exp\left[\frac{-(A_i - \langle A \rangle)^2}{2\sigma^2}\right] \quad (6.67)$$

where A_i is the value of the property in the i th bin and ΔA is the width of each bin. The number of values that would be expected in each bin, n_i , does not have to be integral, though the actual number as determined from the simulation (M_i) will of course be an integer. The chi-squared function is given by:

$$\chi^2 = \sum_i \frac{(M_i - n_i)^2}{n_i} \quad (6.68)$$

If χ^2 is large (bigger than unity) then it is unlikely that the two distributions are the same. Any significant deviations from the expected behaviour should be investigated further to try to eliminate as much of the systematic error as possible. It is good practice to vary as

many of the parameters as possible: using different computers, different compilers, different algorithms and different ways of implementing a given algorithm, and different simulation methods (Monte Carlo and molecular dynamics) not only to test the component parts of the simulation but also the software used to perform the calculation.

If all sources of systematic error can be eliminated, there will still remain statistical errors. These errors are often reported as standard deviations. What we would particularly like to estimate is the error in the average value, $\langle A \rangle$. The standard deviation of the average value is calculated as follows:

$$\sigma_{\langle A \rangle} = \frac{\sigma_A}{\sqrt{M}} = \frac{\sqrt{\sum_{i=1}^M (A(i) - \langle A \rangle)^2}}{\sqrt{M}} \quad (6.69)$$

where $\sigma_{\langle A \rangle}$ is the standard deviation of the average value $\langle A \rangle$ obtained from M data values with respect to the run average, σ_A . Thus the standard deviation of the calculated average is inversely proportional to the square root of the number of data values, and so a longer simulation gives rise to a more accurate value. An important feature of Equation (6.69) is that it applies to *independent* (i.e. random) samples. Thus the number M in the denominator is not simply equal to the number of steps in the simulation. This is because there is a high degree of correlation between successive configurations in either a Monte Carlo or molecular dynamics simulation. What we need to know is the correlation or relaxation 'time' of the simulation; this is the number of steps required for the system to lose its 'memory' of previous configurations. In molecular dynamics, where successive steps are related in a temporal fashion, the correlation 'time' is a true time and will be discussed in more detail in Section 7.6. Usually, the correlation time will be unknown prior to the simulation but it can be estimated as follows. First, the configurations are broken down into a series of blocks. Suppose each block contains t_b successive steps and that there are n_b blocks (so the total simulation contains $t_b n_b$ steps, as shown in Figure 6.26). The average value of the property is calculated for each block:

$$\langle A \rangle_b = \frac{1}{t_b} \sum_{i=1}^{t_b} A_i \quad (6.70)$$

As the number of steps t_b in each block increases, so it would be expected that the block averages become uncorrelated. When this is the case, then the variance of the block averages, $\sigma^2(\langle A \rangle_b)$, will become inversely proportional to t_b . $\sigma^2(\langle A \rangle_b)$ is calculated as follows:

$$\sigma^2(\langle A \rangle_b) = \frac{1}{n_b} \sum_{b=1}^{n_b} (\langle A \rangle_b - \langle A \rangle_{\text{total}})^2 \quad (6.71)$$

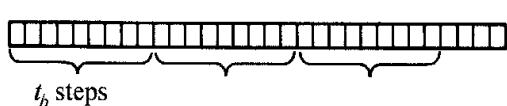


Fig 6.26 Blocking a simulation to calculate the statistical error.

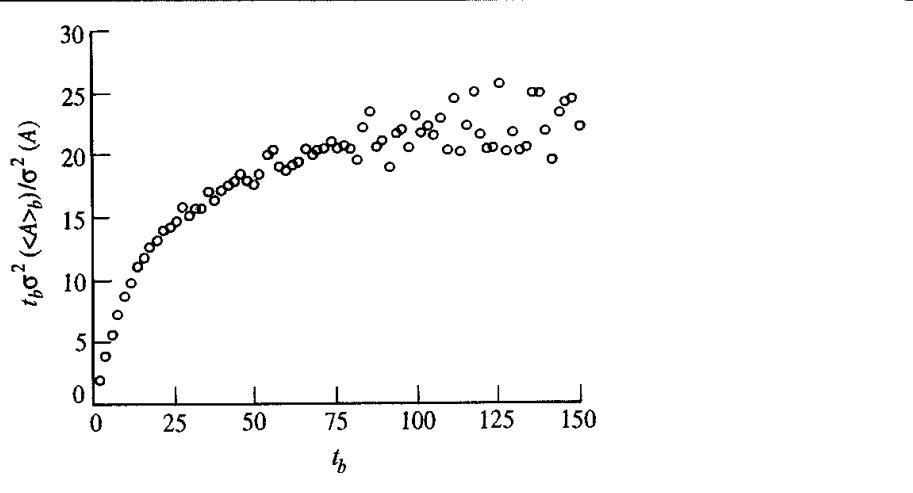


Fig 6.27: Calculating the statistical efficiency, σ . A plot of $t_b \sigma^2(\langle A \rangle_b) / \sigma^2(A)$ against t_b shows a steep rise before levelling off. Here the property A corresponds to the pressure calculated from the molecular dynamics simulation of argon.

where $\langle A \rangle_{\text{total}}$ is the average over the entire simulation. The limiting number of steps to obtain uncorrelated configurations (the statistical inefficiency, s) can be calculated using:

$$s = \lim_{t_b \rightarrow \infty} \frac{t_b \sigma^2(\langle A \rangle_b)}{\sigma^2(A)} \quad (6.72)$$

To determine s , $t_b \sigma^2(\langle A \rangle_b) / \sigma^2(A)$ is plotted against t_b or $\sqrt{t_b}$. The graph should show a steep rise for low t_b and then level off to give a plateau, as shown in Figure 6.27. The plateau value is the limiting value that gives the correlation time ($s \approx 23$ in this case).

Having determined the value of s , the ‘true’ standard deviation of the average value is related to the ‘true’ error for an infinite simulation by:

$$\sigma_{\langle A \rangle} \approx \sigma \sqrt{\frac{s}{M}} \quad (6.73)$$

M here is the actual number of steps or iterations in the simulation. If the value of s can be reduced, then a more accurate average value can be calculated for a given length of simulation. This should be an important consideration when deciding what simulation protocol to use. For example, it may be more appropriate to use a complex simulation algorithm than a simpler one if the statistical inefficiency is significantly reduced.

If the relaxation time is known, then sample averages are often best calculated using the block method (Figure 6.26). Each block should contain more steps than the relaxation time. The sample average for the whole run can be obtained in a variety of ways:

1. Stratified systematic sampling, in which a single value of the property is taken from each block;
2. Stratified random sampling, in which a single value is taken at random from each block;
3. Coarse graining, in which the average value for each block is determined and then the average for the run is calculated by averaging the coarse-grain averages.

The coarse-graining approach is commonly used for thermodynamic properties whereas the systematic or random sampling methods are appropriate for static structural properties such as the radial distribution function.

Another way to improve the error in a simulation, at least for properties such as the energy and the heat capacity that depend on the size of the system (the extensive properties), is to increase the number of atoms or molecules in the calculation. The standard deviation of the average of such a property is proportional to $1/\sqrt{N}$. Thus, more accurate values can be obtained by running longer simulations on larger systems. In computer simulation it is unfortunately the case that the more effort that is expended the better the results that are obtained. Such is life!

Appendix 6.1 Basic Statistical Mechanics

The Boltzmann distribution is fundamental to statistical mechanics. The Boltzmann distribution is derived by maximising the entropy of the system (in accordance with the second law of thermodynamics) subject to the constraints on the system. Let us consider a system containing N particles (atoms or molecules) such that the energy levels of the particles are $\varepsilon_1, \varepsilon_2, \dots$. If there are n_1 particles in the energy level ε_1 , n_2 particles in ε_2 and so on, then there are W ways in which this distribution can be achieved:

$$W(n_1, n_2, \dots) = N! / n_1! n_2! \dots \quad (6.74)$$

The most favourable distribution is the one with the highest weight, and this corresponds to the configuration with just one particle in each energy level ($W = N!$). However, there are two important constraints on the system. First, the total energy is fixed:

$$\sum_i n_i \varepsilon_i = E \quad (6.75)$$

The second constraint arises from the fact that the total number of particles is fixed:

$$\sum_i n_i = N \quad (6.76)$$

The Boltzmann distribution gives the number of particles n_i in each energy level ε_i as:

$$\frac{n_i}{N} = \frac{\exp(-\varepsilon_i/k_B T)}{\sum_i \exp(-\varepsilon_i/k_B T)} \quad (6.77)$$

The denominator in this expression is the molecular partition function:

$$q = \sum_i \exp(-\varepsilon_i/k_B T) \quad (6.78)$$

For translational, rotational and vibrational motion the partition function can be calculated using standard results obtained by solving the Schrödinger equation:

$$\text{translation: } q^t = \left(\frac{2\pi m k_B T}{h^2} \right)^{3/2} V \quad (6.79)$$

where V is the volume

$$\text{rotation: } q^r \approx \left(\frac{\pi^{1/2}}{\sigma} \right) \left(\frac{2I_A k_B T}{\hbar^2} \right) \left(\frac{2I_B k_B T}{\hbar^2} \right) \left(\frac{2I_C k_B T}{\hbar^2} \right) \quad (6.80)$$

where I_A, I_B, I_C are the moments of inertia and σ is the symmetry number (2 for H_2O , 3 for NH_3 , 12 for benzene)

$$\text{vibration: } r^v = \frac{1}{1 - \exp(-\hbar\omega/k_B T)} \quad (6.81)$$

ω is the angular frequency: $\omega = \sqrt{k/\mu}$, where μ is the reduced mass. This form of the vibrational partition function is measured relative to the zero-point energy.

In computer simulations, we are particularly interested in the properties of a system comprising a number of particles. An ensemble is a collection of such systems, as might be generated using a molecular dynamics or a Monte Carlo simulation. Each member of the ensemble has an energy, and the distribution of the system within the ensemble follows the Boltzmann distribution. This leads to the concept of the ensemble partition function, Q .

Various thermodynamic properties can be calculated from the partition function. Here we simply state some of the most common:

$$\text{internal energy: } U = \frac{k_B T^2}{Q} \left(\frac{\partial Q}{\partial T} \right)_V = k_B T^2 \left(\frac{\partial \ln Q}{\partial T} \right)_V \quad (6.82)$$

$$\text{enthalpy: } H = k_B T^2 \left(\frac{\partial \ln Q}{\partial T} \right)_V + k_B T V \left(\frac{\partial \ln Q}{\partial V} \right)_T \quad (6.83)$$

$$\text{Helmholtz free energy: } A = -k_B T \ln Q \quad (6.84)$$

$$\text{Gibbs free energy: } G = -k_B T \ln Q + k_B T V \left(\frac{\partial \ln Q}{\partial V} \right)_T \quad (6.85)$$

Appendix 6.2 Heat Capacity and Energy Fluctuations

The heat capacity is related to the internal energy U by

$$C_V = \left(\frac{\partial U}{\partial T} \right)_V \quad (6.86)$$

If we differentiate the expression for the internal energy, Equation (6.20), we can obtain the heat capacity in terms of the partition function:

$$C_V = \frac{\partial}{\partial T} \left(\frac{k_B T^2}{Q} \frac{\partial Q}{\partial T} \right)_V = \frac{k_B T^2}{Q} \frac{\partial^2 Q^2}{\partial T^2} + \frac{2k_B T}{Q} \frac{\partial Q}{\partial T} - \frac{k_B T^2}{Q^2} \left(\frac{\partial Q}{\partial T} \right)^2 \quad (6.87)$$

The desired expression is obtained by writing each of these three terms as a function of the average energy, $\langle E \rangle$. The internal energy is just the expectation value of the energy, $\langle E \rangle$, and so

$$\langle E \rangle = \frac{k_B T^2}{Q} \frac{\partial Q}{\partial T} \quad (6.88)$$

Thus for the second term in Equation (6.87) we have

$$\frac{2k_B T}{Q} \frac{\partial Q}{\partial T} = \frac{2\langle E \rangle}{T} \quad (6.89)$$

We can also rewrite the third term in Equation (6.87):

$$k_B T \left(\frac{1}{Q} \frac{\partial Q}{\partial T} \right)^2 = \frac{\langle E \rangle^2}{k_B T} \quad (6.90)$$

For the first term, we need to do a little more work. The starting point is:

$$\frac{\partial}{\partial T} \left(\frac{\langle E \rangle}{k_B T^2} \right) = \frac{\partial}{\partial T} \left\{ \frac{1}{Q} \left(\frac{\partial Q}{\partial T} \right) \right\} \quad (6.91)$$

or

$$-2 \frac{\langle E \rangle}{k_B T^3} = \frac{1}{Q} \frac{\partial^2 Q}{\partial T^2} + \frac{\partial Q}{\partial T} \frac{\partial}{\partial T} \left(\frac{1}{Q} \right) \quad (6.92)$$

We can use the chain rule as follows:

$$\frac{\partial Q}{\partial T} \frac{\partial}{\partial T} \left(\frac{1}{Q} \right) = \frac{\partial Q}{\partial T} \frac{\partial Q}{\partial T} \frac{\partial}{\partial T} \left(\frac{1}{Q} \right) = - \left(\frac{\partial Q}{\partial T} \right)^2 \left(\frac{1}{Q} \right)^2 \quad (6.93)$$

Thus

$$\frac{k_B T^2}{Q} \frac{\partial^2 Q}{\partial T^2} = -2 \frac{\langle E \rangle}{k_B T^3} + \frac{\langle E^2 \rangle}{k_B^2 T^4} \quad (6.94)$$

So

$$C_V = k_B T^2 \left\{ -2 \frac{\langle E \rangle}{k_B T^3} + \frac{\langle E^2 \rangle}{k_B^2 T^4} \right\} + 2 \frac{\langle E \rangle}{T} - \frac{\langle E \rangle^2}{k_B T^2} \quad (6.95)$$

or

$$C_V = \frac{(\langle E^2 \rangle - \langle E \rangle^2)}{k_B T^2} \quad (6.96)$$

Appendix 6.3 The Real Gas Contribution to the Virial

If the gas particles interact through a pairwise potential, then the contribution to the virial from the intermolecular forces can be derived as follows. Consider two atoms i and j separated by a distance r_{ij} .

$$r_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2} \quad (6.97)$$

The contribution to the virial from the interaction $\nu(r_{ij})$ between atoms i and j is given by:

$$W_{\text{real}} = \left[x_i \frac{\partial}{\partial x_i} + x_j \frac{\partial}{\partial x_j} + y_i \frac{\partial}{\partial y_i} + y_j \frac{\partial}{\partial y_j} + z_i \frac{\partial}{\partial z_i} + z_j \frac{\partial}{\partial z_j} \right] \nu(r_{ij}) \quad (6.98)$$

Since

$$x_i \frac{\partial r_{ij}}{\partial x_i} = x_i \frac{(x_i - x_j)}{r_{ij}} \quad \text{and} \quad x_j \frac{\partial r_{ij}}{\partial x_i} = -x_j \frac{(x_i - x_j)}{r_{ij}} \quad (6.99)$$

and similarly for the y and z coordinates, we can apply the chain rule, $\partial/\partial x_i = (\partial/\partial r_{ij})(\partial r_{ij}/\partial x_i)$, as follows:

$$W_{\text{real}} = \left[\frac{(x_i - x_j)^2}{r_{ij}} + \frac{(y_i - y_j)^2}{r_{ij}} + \frac{(z_i - z_j)^2}{r_{ij}} \right] \frac{d\nu(r_{ij})}{dr_{ij}} = r_{ij} \frac{d\nu(r_{ij})}{dr_{ij}} \quad (6.100)$$

When we include the contributions from all pairs of atoms, we obtain:

$$W_{\text{real}} = \sum_{i=1}^N \sum_{j=i+1}^N r_{ij} \frac{d\nu(r_{ij})}{dr_{ij}} \quad (6.101)$$

Appendix 6.4 Translating Particle Back into Central Box for Three Box Shapes

From Smith W 1983. The Periodic Boundary Condition in Non-Cubic MD Cells: Wigner-Seitz Cells with Reflection Symmetry. CCP5 Quarterly 10:37–42. This table is expressed using references to several built-in FORTRAN functions. The AINT function returns the integral part of its argument, e.g. AINT(3.4) = 3.0; AINT(4.7) = 4.0; AINT(-0.5) = 0.0 and AINT(-1.7) = -1.0. ANINT returns the nearest integer, so ANINT(0.49) = 0.0 and ANINT(0.51) = 1.0. SIGN(x, y) returns $|x|$ if $y \geq 0$ and $-|x|$ if $y < 0$. ABS(x) returns the absolute value of x , $|x|$. Equivalent functions also exist in most other programming languages.

Rectangular box, side $2a$ (x) by $2b$ (y) by $2c$ (z)

$x = x - 2 \times a \times \text{AIN}(x/a)$
 $y = y - 2 \times b \times \text{AIN}(y/b)$
 $z = z - 2 \times c \times \text{AIN}(z/c)$
A common alternative is
 $x = x - a \times \text{ANINT}(x/a)$
 $y = y - b \times \text{ANINT}(y/b)$
 $z = z - c \times \text{ANINT}(z/c)$

Truncated octahedron derived from cube of side $2a$

$x = x - 2 \times a \times \text{AIN}(x/a)$
 $y = y - 2 \times b \times \text{AIN}(y/b)$
 $z = z - 2 \times c \times \text{AIN}(z/c)$
if $(\text{ABS}(x) + \text{ABS}(y) + \text{ABS}(z)) \geq 1.5 \times A$
then
 $x = x - \text{SIGN}(a, x)$
 $y = y - \text{SIGN}(a, y)$
 $z = z - \text{SIGN}(a, z)$
endif

Hexagonal prism of length $2a$ (in z direction) and distance between opposite faces of the hexagon $2b$

$z = z - 2 \times a \times \text{AIN}(z/a)$
 $x = x - 2 \times b \times \text{AIN}(x/b)$
if $(\text{ABS}(x) + \sqrt{3} \times \text{ABS}(y)) \geq 2 \times B$ then
 $x = x - \text{SIGN}(b, x)$
 $y = y - \text{SIGN}(\sqrt{3} \times b, y)$
endif

Further Reading

- Allen M P and D J Tildesley 1987. *Computer Simulation of Liquids* Oxford, Oxford University Press
Bradbury T C 1968. *Theoretical Mechanics*. Malabar, FL, Krieger.
Chandler D 1987 *Introduction to Modern Statistical Mechanics*. New York, Oxford University Press.
Hansen J P and I R McDonald 1976 *Theory of Simple Liquids*. London, Academic Press.
Smith P E and van Gunsteren W F 1993. Methods for the Evaluation of Long Range Electrostatic Forces.
In van Gunsteren W F, P K Weiner and A J Wilkinson (Editors). *Computer Simulation of Biomolecular Systems*. Leiden, ESCOM.
van Gunsteren W F and H J C Berendsen 1990. Computer Simulation of Molecular Dynamics Methodology, Applications and Perspectives in Chemistry. *Angewandte Chemie International Edition in English* **29**:992–1023

References

- Adams D J 1983 Alternatives to the Periodic Cube in Computer Simulation. *CCP5 Quarterly* **10**:30–36.
Alper H E and R M Levy 1989 Computer Simulations of the Dielectric Properties of Water – Studies of the Simple Point-Charge and Transferable Intermolecular Potential Models. *Journal of Chemical Physics* **91**:1242–1251
Cheatham T E III, J L Miller, T Fox, T A Darden and P A Kollman 1995 Molecular Dynamics Simulations on Solvated Biomolecular Systems: The Particle Mesh Ewald Method Leads to Stable Trajectories of DNA, RNA and Proteins. *Journal of the American Chemical Society* **117** 4193–4194
Darden T A, L Perera, L Li and L Pedersen 1999 New Tricks for Modelers from the Crystallography Toolkit: The Particle Mesh Ewald Algorithm and Its Use in Nucleic Acid Simulations. *Structure with Folding and Design* **7** R55–R60.
Darden T A, D York and L Pedersen 1993 Particle-mesh Ewald: An $N \log(N)$ method for Ewald sums in large systems. *Journal of Chemical Physics* **98** 10089–10092
Deserno M and C Holm 1998a How to Mesh Up Ewald Sums. I A Theoretical and Numerical Comparison of Various Particle Mesh Routines. *Journal of Chemical Physics* **109** 7678–7693
Deserno M and C Holm 1998b. How to Mesh Up Ewald Sums. II An Accurate Error Estimate for the Particle-Particle-Particle-Mesh Algorithm. *Journal of Chemical Physics* **109**:7694–7701.
Ding H-Q, N Karasawa and W A Goddard III 1992a. Atomic Level Simulations on a Million Particles: The Cell Multipole Method for Coulomb and London Nonbonding Interactions. *Journal of Chemical Physics* **97**:4309–4315.
Ding H-Q, N Karasawa and W A Goddard III 1992b. The Reduced Cell Multipole Method for Coulomb Interactions in Periodic Systems with Million-Atom Unit Cells. *Chemical Physics Letters* **196**:6–10.
Ewald P 1921 Due Berechnung optischer und elektrostatischer Gitterpotentiale. *Annalen der Physik* **64** 253–287.
Friedman H L 1975. Image Approximation to the Reaction Field. *Molecular Physics* **29** 1533–1543
Greengard L 1994. Fast Algorithms for Classical Physics. *Science* **265**:909–914
Greengard L and V I Roklin 1987. A Fast Algorithm for Particle Simulations. *Journal of Computational Physics* **73**:325–348
Hockney R W and J W Eastwood 1988 *Computer Simulation using Particles* Bristol, Adam Hilger.
Luty B A, M E David, I G Tironi and W F van Gunsteren 1994 A Comparison of Particle-Particle, Particle-Mesh and Ewald Methods for Calculating Electrostatics Interactions in Periodic Molecular Systems. *Molecular Simulation* **14**:11–20

- Luty B A, I G Tironi and W F van Gunsteren 1995. Lattice-sum methods for calculating electrostatic interactions in molecular simulations. *Journal of Chemical Physics* **103**:3014–3021.
- Petersen H G, D Soelvaso, J W Perram and E R Smith 1994. The Very Fast Multipole Method. *Journal of Chemical Physics* **101**:8870–8876.
- Thompson S M 1983. Use of Neighbour Lists in Molecular Dynamics. *CCP5 Quarterly* **8** 20–28.
- van Gunsteren W F and H J C Berendsen 1986. *GROMOS User Guide*.
- Verlet L 1967. Computer ‘Experiments’ on Classical Fluids. II. Equilibrium Correlation Functions. *Physical Review* **165**:201–204.
- Wolf M L, J R Walker and C R A Catlow 1984. A Molecular Dynamics Simulation Study of the Superionic Conductor Lithium Nitride. I. *Journal of Physical Chemistry* **17**:6623–34.
- York D M, A Wlodawer, L G Pedersen and T A Darden 1994. Atomic-level Accuracy in Simulations of Large Protein Crystals. *Proceedings of the National Academy of Sciences USA* **91**:8715–8718.

Molecular Dynamics Simulation Methods

7.1 Introduction

In molecular dynamics, successive configurations of the system are generated by integrating Newton's laws of motion. The result is a trajectory that specifies how the positions and velocities of the particles in the system vary with time. Newton's laws of motion can be stated as follows:

1. A body continues to move in a straight line at constant velocity unless a force acts upon it.
2. Force equals the rate of change of momentum.
3. To every action there is an equal and opposite reaction.

The trajectory is obtained by solving the differential equations embodied in Newton's second law ($F = ma$):

$$\frac{d^2x_i}{dt^2} = \frac{F_{x_i}}{m_i} \quad (7\ 1)$$

This equation describes the motion of a particle of mass m_i along one coordinate (x_i) with F_{x_i} being the force on the particle in that direction.

It is helpful to distinguish three different types of problem to which Newton's laws of motion may be applied. In the simplest case, no force acts on each particle between collisions. From one collision to the next, the position of the particle thus changes by $v_i\delta t$, where v_i is the (constant) velocity and δt is the time between collisions. In the second situation, the particle experiences a constant force between collisions. An example of this type of motion would be that of a charged particle moving in a uniform electric field. In the third case, the force on the particle depends on its position relative to the other particles. Here the motion is often very difficult, if not impossible, to describe analytically, due to the coupled nature of the particles' motions

7.2 Molecular Dynamics Using Simple Models

The first molecular dynamics simulation of a condensed phase system was performed by Alder and Wainwright in 1957 using a hard-sphere model [Alder and Wainwright 1957]. In this model, the spheres move at constant velocity in straight lines between collisions. All collisions are perfectly elastic and occur when the separation between the centres of

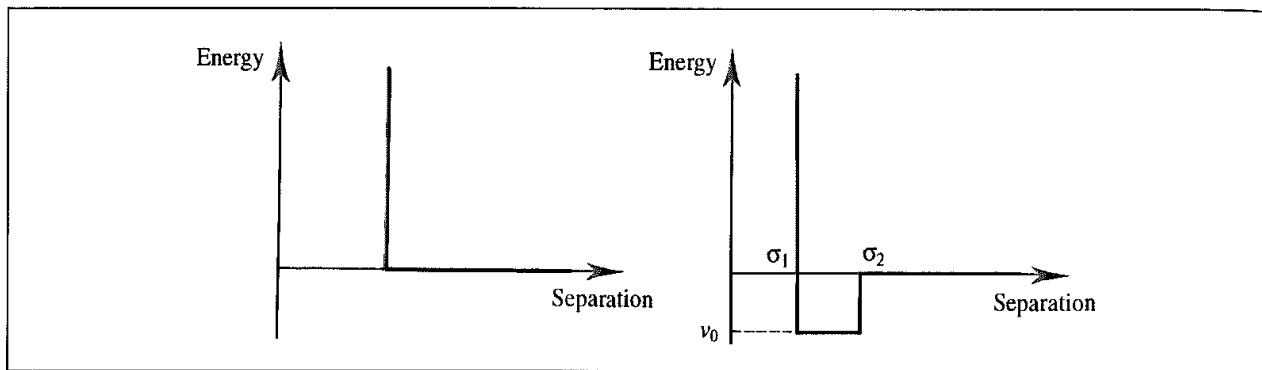


Fig. 7.1 The hard-sphere and square-well potentials

the spheres equals the sphere diameter. The pair potential thus has the form shown in Figure 7.1. Some early simulations also used the square-well potential, where the interaction energy between two particles is zero beyond a cutoff distance σ_2 ; infinite below a smaller cutoff distance σ_1 ; and equal to v_0 between the two cutoff values (Figure 7.1). The steps involved in the hard-sphere calculation are as follows:

1. Identify the next pair of spheres to collide and calculate when the collision will occur.
2. Calculate the positions of all the spheres at the collision time.
3. Determine the new velocities of the two colliding spheres after the collision.
4. Repeat from 1 until finished.

The new velocities of the colliding spheres are calculated by applying the principle of conservation of linear momentum.

Simple interaction models such as the hard-sphere potential obviously suffer from many deficiencies but have nevertheless provided many useful insights into the microscopic

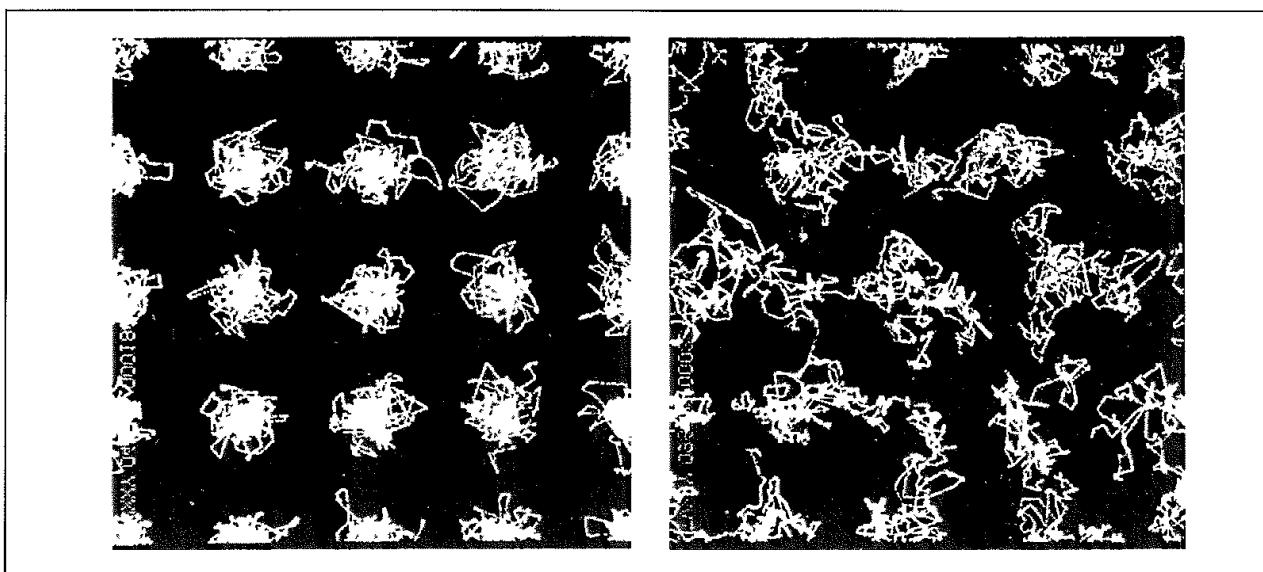


Fig. 7.2 Molecular graphics representation of the paths generated by 32 hard spherical particles in the solid (left) and fluid (right) phase (Reproduced from Alder B J and T E Wainwright 1959 Studies in Molecular Dynamics I General Method Journal of Chemical Physics 31 459–466)

nature of fluids. The early workers were particularly keen to quantify the differences between the solid and fluid phases; it is interesting to note that such investigations were facilitated by early molecular graphics systems, which enabled the trajectories of the particles to be represented simultaneously (Figure 7.2).

7.3 Molecular Dynamics with Continuous Potentials

In more realistic models of intermolecular interactions, the force on each particle will change whenever the particle changes its position, or whenever any of the other particles with which it interacts changes position. The first simulation using continuous potentials was of argon by Rahman [Rahman 1964], who also performed the first simulation of a molecular liquid (water) [Rahman and Stillinger 1971] and made many other important methodological contributions in molecular dynamics. Under the influence of a continuous potential the motions of all the particles are coupled together, giving rise to a many-body problem that cannot be solved analytically. Under such circumstances the equations of motion are integrated using a *finite difference method*.

7.3.1 Finite Difference Methods

Finite difference techniques are used to generate molecular dynamics trajectories with continuous potential models, which we will assume to be pairwise additive. The essential idea is that the integration is broken down into many small stages, each separated in time by a fixed time δt . The total force on each particle in the configuration at a time t is calculated as the vector sum of its interactions with other particles. From the force we can determine the accelerations of the particles, which are then combined with the positions and velocities at a time t to calculate the positions and velocities at a time $t + \delta t$. The force is assumed to be constant during the time step. The forces on the particles in their new positions are then determined, leading to new positions and velocities at time $t + 2\delta t$, and so on.

There are many algorithms for integrating the equations of motion using finite difference methods, several of which are commonly used in molecular dynamics calculations. All algorithms assume that the positions and dynamic properties (velocities, accelerations, etc.) can be approximated as Taylor series expansions:

$$\mathbf{r}(t + \delta t) = \mathbf{r}(t) + \delta t \mathbf{v}(t) + \frac{1}{2} \delta t^2 \mathbf{a}(t) + \frac{1}{6} \delta t^3 \mathbf{b}(t) + \frac{1}{24} \delta t^4 \mathbf{c}(t) + \dots \quad (7.2)$$

$$\mathbf{v}(t + \delta t) = \mathbf{v}(t) + \delta t \mathbf{a}(t) + \frac{1}{2} \delta t^2 \mathbf{b}(t) + \frac{1}{6} \delta t^3 \mathbf{c}(t) + \dots \quad (7.3)$$

$$\mathbf{a}(t + \delta t) = \mathbf{a}(t) + \delta t \mathbf{b}(t) + \frac{1}{2} \delta t^2 \mathbf{c}(t) \dots \quad (7.4)$$

$$\mathbf{b}(t + \delta t) = \mathbf{b}(t) + \delta t \mathbf{c}(t) + \dots \quad (7.5)$$

where \mathbf{v} is the velocity (the first derivative of the positions with respect to time), \mathbf{a} is the acceleration (the second derivative), \mathbf{b} is the third derivative, and so on. The *Verlet algorithm* [Verlet 1967] is probably the most widely used method for integrating the equations of

motion in a molecular dynamics simulation. The Verlet algorithm uses the positions and accelerations at time t , and the positions from the previous step, $\mathbf{r}(t - \delta t)$, to calculate the new positions at $t + \delta t$, $\mathbf{r}(t + \delta t)$. We can write down the following relationships between these quantities and the velocities at time t

$$\mathbf{r}(t + \delta t) = \mathbf{r}(t) + \delta t \mathbf{v}(t) + \frac{1}{2} \delta t^2 \mathbf{a}(t) + \dots \quad (7.6)$$

$$\mathbf{r}(t - \delta t) = \mathbf{r}(t) - \delta t \mathbf{v}(t) + \frac{1}{2} \delta t^2 \mathbf{a}(t) - \dots \quad (7.7)$$

Adding these two equations gives

$$\mathbf{r}(t + \delta t) = 2\mathbf{r}(t) - \mathbf{r}(t - \delta t) + \delta t^2 \mathbf{a}(t) \quad (7.8)$$

The velocities do not explicitly appear in the Verlet integration algorithm. The velocities can be calculated in a variety of ways; a simple approach is to divide the difference in positions at times $t + \delta t$ and $t - \delta t$ by $2\delta t$:

$$\mathbf{v}(t) = [\mathbf{r}(t + \delta t) - \mathbf{r}(t - \delta t)]/2\delta t \quad (7.9)$$

Alternatively, the velocities can be estimated at the half-step, $t + \frac{1}{2}\delta t$:

$$\mathbf{v}(t + \frac{1}{2}\delta t) = [\mathbf{r}(t + \delta t) - \mathbf{r}(t)]/\delta t \quad (7.10)$$

Implementation of the Verlet algorithm is straightforward and the storage requirements are modest, comprising two sets of positions ($\mathbf{r}(t)$ and $\mathbf{r}(t - \delta t)$) and the accelerations $\mathbf{a}(t)$. One of its drawbacks is that the positions $\mathbf{r}(t + \delta t)$ are obtained by adding a small term ($\delta t^2 \mathbf{a}(t)$) to the difference of two much larger terms, $2\mathbf{r}(t)$ and $\mathbf{r}(t - \delta t)$. This may lead to a loss of precision. The Verlet algorithm has some other disadvantages. The lack of an explicit velocity term in the equations makes it difficult to obtain the velocities, and indeed the velocities are not available until the positions have been computed at the next step. In addition, it is not a self-starting algorithm; the new positions are obtained from the current positions $\mathbf{r}(t)$ and the positions from the previous time step, $\mathbf{r}(t - \delta t)$. At $t = 0$ there is obviously only one set of positions and so it is necessary to employ some other means to obtain positions at $t - \delta t$. One way to obtain $\mathbf{r}(t - \delta t)$ is to use the Taylor series, Equation (7.2), truncated after the first term. Thus, $\mathbf{r}(-\delta t) = \mathbf{r}(0) - \delta t \mathbf{v}(0) - \delta t^2 \mathbf{a}(0)$.

Several variations on the Verlet algorithm have been developed. The *leap-frog* algorithm [Hockney 1970] uses the following relationships:

$$\mathbf{r}(t + \delta t) = \mathbf{r}(t) + \delta t \mathbf{v}(t + \frac{1}{2}\delta t) \quad (7.11)$$

$$\mathbf{v}(t + \frac{1}{2}\delta t) = \mathbf{v}(t - \frac{1}{2}\delta t) + \delta t \mathbf{a}(t) \quad (7.12)$$

To implement the leap-frog algorithm, the velocities $\mathbf{v}(t + \frac{1}{2}\delta t)$ are first calculated from the velocities at time $t - \frac{1}{2}\delta t$ and the accelerations at time t . The positions $\mathbf{r}(t + \delta t)$ are then deduced from the velocities just calculated together with the positions at time $\mathbf{r}(t)$ using Equation (7.11). The velocities at time t can be calculated from

$$\mathbf{v}(t) = \frac{1}{2}[\mathbf{v}(t + \frac{1}{2}\delta t) + \mathbf{v}(t - \frac{1}{2}\delta t)] \quad (7.13)$$

The velocities thus 'leap-frog' over the positions to give their values at $t + \frac{1}{2}\delta t$ (hence the name). The positions then leap over the velocities to give their new values at $t + \delta t$, ready for the velocities at $t + \frac{3}{2}\delta t$, and so on. The leap-frog method has two advantages over the

standard Verlet algorithm: it explicitly includes the velocity and also does not require the calculation of the differences of large numbers. However, it has the obvious disadvantage that the positions and velocities are not synchronised. This means that it is not possible to calculate the kinetic energy contribution to the total energy at the same time as the positions are defined (from which the potential energy is determined).

The *velocity Verlet* method [Swope *et al.* 1982] gives positions, velocities and accelerations at the same time and does not compromise precision:

$$\mathbf{r}(t + \delta t) = \mathbf{r}(t) + \delta t \mathbf{v}(t) + \frac{1}{2} \delta t^2 \mathbf{a}(t) \quad (7.14)$$

$$\mathbf{v}(t + \delta t) = \mathbf{v}(t) + \frac{1}{2} \delta t [\mathbf{a}(t) + \mathbf{a}(t + \delta t)] \quad (7.15)$$

The velocity Verlet method is actually implemented as a three-stage procedure because, as can be seen from Equation (7.15), to calculate the new velocities requires the accelerations at both t and $t + \delta t$. Thus in the first step the positions at $t + \delta t$ are calculated according to Equation (7.14) using the velocities and the accelerations at time t . The velocities at time $t + \frac{1}{2} \delta t$ are then determined using:

$$\mathbf{v}(t + \frac{1}{2} \delta t) = \mathbf{v}(t) + \frac{1}{2} \delta t \mathbf{a}(t) \quad (7.16)$$

New forces are next computed from the current positions, thus giving $\mathbf{a}(t + \delta t)$. In the final step, the velocities at time $t + \delta t$ are determined using:

$$\mathbf{v}(t + \delta t) = \mathbf{v}(t + \frac{1}{2} \delta t) + \frac{1}{2} \delta t \mathbf{a}(t + \delta t) \quad (7.17)$$

Beeman's algorithm [Beeman 1976] is also related to the Verlet method:

$$\mathbf{r}(t + \delta t) = \mathbf{r}(t) + \delta t \mathbf{v}(t) + \frac{2}{3} \delta t^2 \mathbf{a}(t) - \frac{1}{6} \delta t^2 \mathbf{a}(t - \delta t) \quad (7.18)$$

$$\mathbf{v}(t + \delta t) = \mathbf{v}(t) + \frac{1}{3} \delta t \mathbf{a}(t) + \frac{5}{6} \delta t \mathbf{a}(t) - \frac{1}{6} \delta t \mathbf{a}(t - \delta t) \quad (7.19)$$

The Beeman integration scheme uses a more accurate expression for the velocity. As a consequence it often gives better energy conservation, because the kinetic energy is calculated directly from the velocities. However, the expressions used are more complex than those of the Verlet algorithm and so it is computationally more expensive.

We have already encountered four different integration methods, with more to come! Why should we use one method in preference to another? What features characterise a 'good' integration method? As with any other computer algorithm, an ideal integration scheme should be fast, require minimal memory and be easy to program. However, for most molecular dynamics simulations these issues are of secondary importance; most calculations do not make significant memory demands of even a modest workstation, and the time required for the integration is usually trivial compared to the other parts of the calculation. The most demanding part of a molecular dynamics simulation is invariably the calculation of the force on each particle in the system. More important considerations are that the integration algorithm should conserve energy and momentum, be time-reversible, and should permit a long time step, δt , to be used. The size of the time step is particularly relevant to the computational demands as a simulation using a long time step will require fewer iterations to cover a given amount of phase space. A less important requirement is that the integration algorithm should give the same results as an exact, analytical trajectory

(this can be tested using simple problems for which an analytical solution can be derived). We would, in any case, expect the calculated trajectory to deviate from the exact trajectory because the computer can only store numbers to a given precision.

The *order* of an integration method is the degree to which the Taylor series expansion, Equation (7.2), is truncated: it is the lowest term that is not present in the expansion. The order may not always be apparent from the formulae used. For example, the highest-order derivative that appears in the Verlet formulae is the second, $\mathbf{a}(t)$, yet the Verlet algorithm is, in fact, a fourth-order method. This is because the third-order terms, which cancel when Equation (7.6) is added to Equation (7.7), are still implied in the expansion:

$$\mathbf{r}(t + \delta t) = \mathbf{r}(t) + \delta t \mathbf{v}(t) + \frac{1}{2} \delta t^2 \mathbf{a}(t) + \frac{1}{6} \delta t^3 \mathbf{b}(t) + \frac{1}{24} \delta t^4 \mathbf{c}(t) \quad (7.20)$$

$$\mathbf{r}(t - \delta t) = \mathbf{r}(t) - \delta t \mathbf{v}(t) + \frac{1}{2} \delta t^2 \mathbf{a}(t) - \frac{1}{6} \delta t^3 \mathbf{b}(t) + \frac{1}{24} \delta t^4 \mathbf{c}(t) \quad (7.21)$$

7.3.2 Predictor–Corrector Integration Methods

The predictor–corrector methods [Gear 1971] form a general family of integration algorithms from which one can select a scheme that is correct to a given order. These methods have three basic steps. First, new positions, velocities, accelerations and higher-order terms are predicted according to the Taylor expansion, Equations (7.2)–(7.4). In the second stage, the forces are evaluated at the new positions to give accelerations $\mathbf{a}(t + \delta t)$. These accelerations are then compared with the accelerations that are predicted from the Taylor series expansion, $\mathbf{a}^c(t + \delta t)$. The difference between the predicted and calculated accelerations is then used to ‘correct’ the positions, velocities, etc., in the correction step:

$$\Delta \mathbf{a}(t + \delta t) = \mathbf{a}^c(t + \delta t) - \mathbf{a}(t + \delta t) \quad (7.22)$$

Then

$$\mathbf{r}^c(t + \delta t) = \mathbf{r}(t + \delta t) + c_0 \Delta \mathbf{a}(t + \delta t) \quad (7.23)$$

$$\mathbf{v}^c(t + \delta t) = \mathbf{v}(t + \delta t) + c_1 \Delta \mathbf{a}(t + \delta t) \quad (7.24)$$

$$\mathbf{a}^c(t + \delta t)/2 = \mathbf{a}(t + \delta t)/2 + c_2 \Delta \mathbf{a}(t + \delta t) \quad (7.25)$$

$$\mathbf{b}^c(t + \delta t)/6 = \mathbf{b}(t + \delta t)/6 + c_3 \Delta \mathbf{a}(t + \delta t) \quad (7.26)$$

Gear has suggested ‘best’ values of the coefficients c_0, c_1, \dots . The set of coefficients to use depends upon the order of the Taylor series expansion. In Equations (7.23)–(7.26) the expansion has been truncated after the third derivative of the positions (i.e. $\mathbf{b}(t)$). The appropriate set of coefficients to use in this case is $c_0 = \frac{1}{6}$, $c_1 = \frac{5}{6}$, $c_2 = 1$ and $c_3 = \frac{1}{3}$.

The storage required for the Gear predictor–corrector algorithm is $3 \times (O + 1)N$, where O is the highest-order differential used in the Taylor series expansion and N is the number of atoms. Thus the storage required for our example is $15N$, which is rather more than for the Verlet algorithm, which uses $9N$. More importantly, the Gear algorithm requires two time-consuming force evaluations per time step, though this is not necessarily a disadvantage as it may permit a time step more than twice as long as an alternative algorithm.

There are many variants of the ‘predictor–corrector’ theme; of these, we will only mention the algorithm used by Rahman in the first molecular dynamics simulations with continuous potentials [Rahman 1964]. In this method, the first step is to predict new positions as follows:

$$\mathbf{r}(t + \delta t) = \mathbf{r}(t - \delta t) + 2\delta t \mathbf{v}(t) \quad (7.27)$$

New accelerations are calculated at these new positions in the usual way. These accelerations are then used to generate a set of new velocities, and then corrected positions:

$$\mathbf{v}(t + \delta t) = \mathbf{v}(t) + \frac{1}{2}\delta t(\mathbf{a}(t + \delta t) + \mathbf{a}(t)) \quad (7.28)$$

$$\mathbf{r}^c(t + \delta t) = \mathbf{r}(t) + \frac{1}{2}\delta t(\mathbf{v}(t) + \mathbf{v}(t + \delta t)) \quad (7.29)$$

The acceleration can then be recalculated at the new corrected positions to give new velocities. The method then iterates over the two Equations (7.28) and (7.29). Two or three passes are usually required to achieve consistency, with a force evaluation at each step. The computational demands of this scheme mean that it is now rarely used, though it does give accurate solutions of the equations of motion.

7.3.3 Which Integration Algorithm is Most Appropriate?

The wide variety of integration schemes available can make it difficult to decide which is the most appropriate one to use. Various factors may need to be taken into account when deciding which is most appropriate. Clearly, the computational effort required is a major consideration. As we have already indicated, an algorithm that is nominally more expensive (for example, because it requires more than one force evaluation per iteration) may permit a significantly longer time step to be used and so, in fact, be more cost-effective. One of the most important considerations is energy conservation; this can be calculated as the root-mean-square fluctuation and is often plotted against the time step, as shown in Figure 7.3. In Appendix 7.1 we show why energy conservation would be expected in a molecular dynamics simulation. The kinetic and potential energy components would be expected to fluctuate in equal and opposite directions; this is also shown in Figure 7.3.

As the time step increases, so the RMS energy fluctuation also increases. For the argon simulation reported in Figure 7.3, the RMS fluctuation in the total energy is approximately 0.006 kcal/mol and the RMS fluctuations in the kinetic and potential energies are approximately 2.5 kcal/mol. With a time step of 25 fs the RMS fluctuation rises to 0.04 kcal/mol and with a time step of 5 fs the value is 0.002 kcal/mol. A variation of one part in 10^4 is generally considered acceptable. The different algorithms may vary in the rate at which the error varies with the time step. For example, it has been shown that for short time steps the predictor–corrector methods may be more accurate, but for longer time steps the Verlet algorithm may be better [Fincham and Heyes 1982]. Other factors that may be important when choosing an integration algorithm include the memory required; the synchronisation of positions and velocities; whether they are self-starting (some methods require properties at $t - \delta t$, which obviously do not exist); and

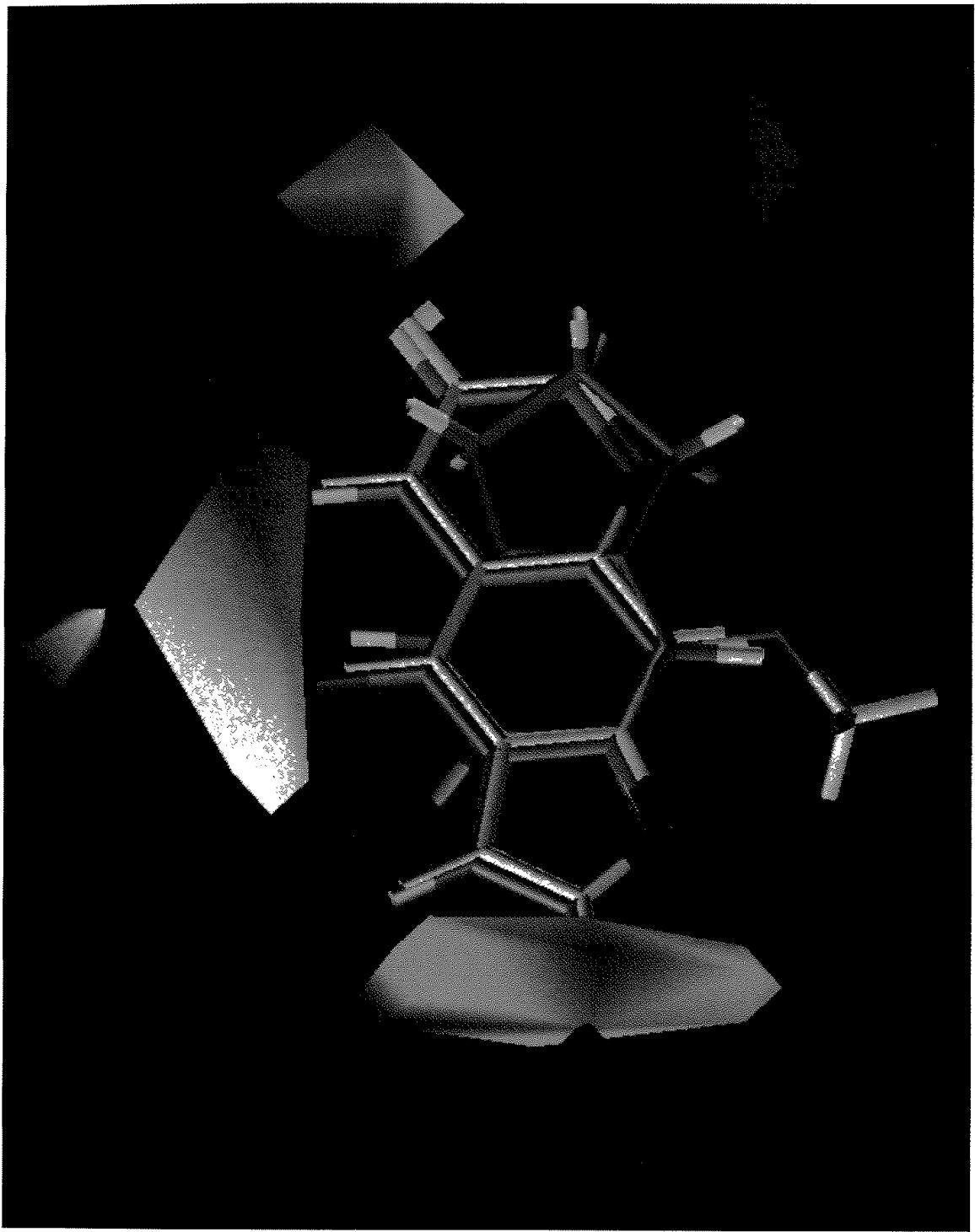


Fig. 12.41. Contour representation of key features from a CoMFA analysis of a series of coumarin substrates and inhibitors of cytochrome P₄₅₀2A5 [Poso et al. 1995]. The red and blue regions indicate positions where it would be favorable and unfavorable, respectively, to place a substituent on the coumarin ring. The red regions are located in the para position, while the blue regions are located in the meta position.

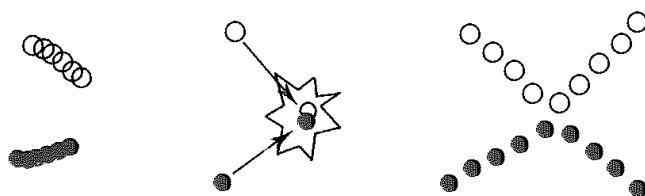


Fig. 7.4: With a very small time step (left) phase space is covered very slowly; a large time step (middle) gives instabilities. With an appropriate time step (right) phase space is covered efficiently and collisions occur smoothly.

consisting of two argon atoms interacting under the Lennard-Jones potential. The behaviour of this system can be determined analytically and so compared with the numerical integration. Suppose the argon atoms are moving towards each other along the x axis with initial velocities of 353 m s^{-1} (this corresponds to the most probable speed of argon at 300 K). We can then plot how the interatomic distance varies with time and compare it to the analytical potential. The result obtained using two time steps (10 fs and 50 fs) are shown in Figure 7.5. In both cases the numerical trajectory initially lags behind the analytical one, but then as the atoms pass through their minimum energy separation and move up the repulsive barrier the atoms 'jump through' the energy barrier. This leads to a gain in energy and the atoms then move apart with velocities that are slightly too high. In both numerical trajectories the total energy rises after the collision. Unfortunately, the atoms move most quickly and take the largest steps in the very region (i.e. near the energy minimum) where it would be best to take the smallest steps. The total error is correlated with the time step, with the largest errors arising for the largest time steps. Of course, with a small time step much more computer time will be required for a given length of calculation; the aim is to find the correct balance between simulating the 'correct' trajectory and covering the phase space. If the time

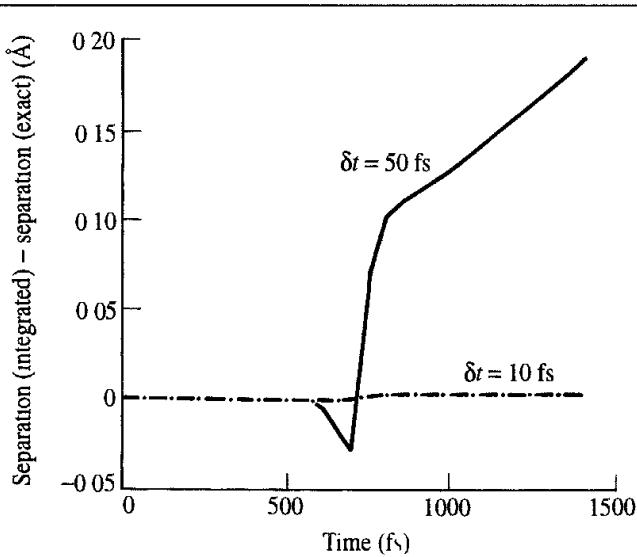


Fig. 7.5. Difference between the exact and numerical trajectories for the approach of two argon atoms with time steps of 10 fs and 50 fs.

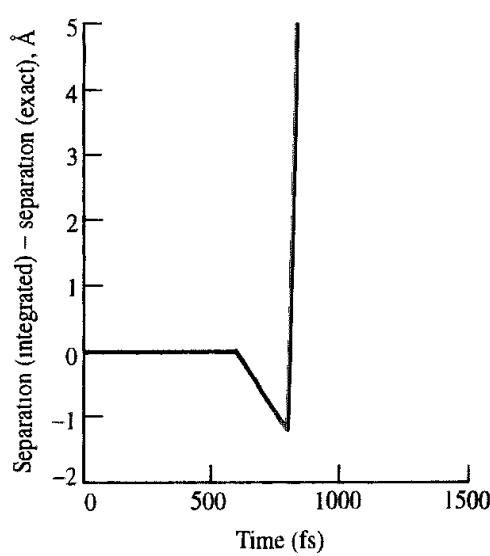


Fig. 7.6 Difference between exact and numerical trajectory for the approach of two argon atoms for a time step of 100 fs. The simulation ‘blows up’

step is too large, then the trajectory will ‘blow up’, as can be seen for the argon dimer system with a time step of 100 fs (Figure 7.6)

When simulating an atomic fluid the time step should be small compared to the mean time between collisions. When simulating flexible molecules a useful guide is that the time step should be approximately one-tenth the time of the shortest period of motion. In flexible molecules, the highest-frequency vibrations are due to bond stretches, especially those of bonds to hydrogen atoms. A C–H bond vibrates with a repeat period of approximately 10 fs. The timescales of some typical motions together with appropriate time steps are shown in Table 7.1, which can be used to choose an appropriate time step.

The requirement that the time step is approximately one order of magnitude smaller than the shortest motion is clearly a severe restriction, particularly as these high-frequency motions are usually of relatively little interest and have a minimal effect on the overall behaviour of the system. One solution to this problem is to ‘freeze out’ such vibrations by constraining the appropriate bonds to their equilibrium values while still permitting the rest of the degrees of freedom to vary under the intramolecular and intermolecular forces present. This enables a longer time step to be used. We will consider such constraint dynamics methods in Section 7.5.

System	Types of motion present	Suggested time step (s)
Atoms	Translation	10^{-14}
Rigid molecules	Translation and rotation	5×10^{-15}
Flexible molecules, rigid bonds	Translation, rotation, torsion	2×10^{-15}
Flexible molecules, flexible bonds	Translation, rotation, torsion, vibration	10^{-15} or 5×10^{-16}

Table 7.1 The different types of motion present in various systems together with suggested time steps.

7.3.5 Multiple Time Step Dynamics

Table 7.1 presents us with something of a dilemma. We would obviously desire to explore as much of the phase space as possible but this may be compromised by the need for a small time step. One possible approach is to use a multiple time step method. The underlying rationale is that certain interactions evolve more rapidly with time than other interactions. The twin-range method (Section 6.7.1) is a crude type of multiple time step approach, in that interactions involving atoms between the lower and upper cutoff distance remain constant and change only when the neighbour list is updated. However, this approach can lead to an accumulation of numerical errors in calculated properties. A more sophisticated approach is to approximate the forces due to these atoms using a Taylor series expansion [Streett *et al.* 1978]:

$$\mathbf{f}(t + \tau\delta t) = \mathbf{f}(t) + (\tau\delta t)\mathbf{d}\mathbf{f}(t)/dt + \frac{1}{2}(\tau\delta t)^2\mathbf{d}^2\mathbf{f}(t)/dt^2 + \dots \quad (7.30)$$

This series expansion is truncated at a specified order and is probably most easily implemented within a predictor–corrector type of algorithm, where the higher-order terms are already computed. This method has been applied to relatively simple systems such as molecular fluids [Streett *et al.* 1978] and alkane chain liquids [Swindoll and Haile 1984].

An alternative formulation of a multiple time step method is the ‘reversible reference system propagation algorithm’ (r-RESPA) method [Tuckerman *et al.* 1992]. In this method, the forces within a system are classified into a number of groups according to how rapidly the force varies over time. Each group then has its own time step while maintaining accuracy and numerical stability. The starting point for this algorithm is the Liouville equation, which defines how the state of the system, $\Gamma(t)$, evolves over time:

$$\Gamma(t) = e^{iLt}\Gamma \quad (t = 0) \quad (7.31)$$

The exponential $\exp(iLt)$ in Equation (7.31) involves the so-called *Liouville operator*, L , which in the case of a molecular system containing N atoms (and so $3N$ coordinates) can be expressed:

$$iL = \sum_{i=1}^{3N} \left[\frac{\partial x_i}{\partial t} \frac{\partial}{\partial x_i} + F_i(x) \frac{\partial}{\partial p_i} \right] \quad (7.32)$$

In the r-RESPA method this operator is decomposed into two or more parts, for example:

$$L = L_1 + L_2 + L_3 + L_4 \quad (7.33)$$

Each of these parts is then associated with specific terms in the force equation. For example, L_1 may correspond to the bond-stretching terms, L_2 to the angle-bending and torsional terms, L_3 to the short-range non-bonded interactions and L_4 to the long-range non-bonded interactions. Suppose the time step with which we evaluate the bond-stretching terms is δt_1 . Integers n_1 , n_2 and n_3 then define the time steps for the three other forces as follows:

$$\delta t_2 = n_1 \delta t_1; \quad \delta t_3 = n_1 n_2 \delta t_1; \quad \delta t_4 = n_1 n_2 n_3 \delta t_1 \quad (7.34)$$

The underlying theory of r-RESPA is somewhat involved, but the final result and consequent implementation is actually rather straightforward, being very closely related to the velocity Verlet integration scheme. For our four-way decomposition the algorithm would

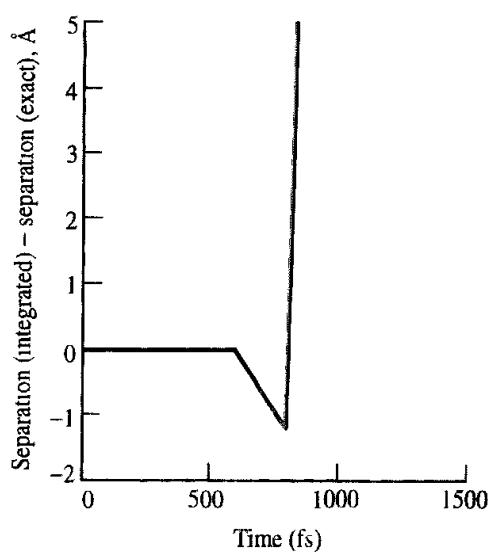


Fig. 7.6. Difference between exact and numerical trajectory for the approach of two argon atoms for a time step of 100 fs. The simulation ‘blows up’

step is too large, then the trajectory will ‘blow up’, as can be seen for the argon dimer system with a time step of 100 fs (Figure 7.6)

When simulating an atomic fluid the time step should be small compared to the mean time between collisions. When simulating flexible molecules a useful guide is that the time step should be approximately one-tenth the time of the shortest period of motion. In flexible molecules, the highest-frequency vibrations are due to bond stretches, especially those of bonds to hydrogen atoms. A C–H bond vibrates with a repeat period of approximately 10 fs. The timescales of some typical motions together with appropriate time steps are shown in Table 7.1, which can be used to choose an appropriate time step.

The requirement that the time step is approximately one order of magnitude smaller than the shortest motion is clearly a severe restriction, particularly as these high-frequency motions are usually of relatively little interest and have a minimal effect on the overall behaviour of the system. One solution to this problem is to ‘freeze out’ such vibrations by constraining the appropriate bonds to their equilibrium values while still permitting the rest of the degrees of freedom to vary under the intramolecular and intermolecular forces present. This enables a longer time step to be used. We will consider such constraint dynamics methods in Section 7.5.

System	Types of motion present	Suggested time step (s)
Atoms	Translation	10^{-14}
Rigid molecules	Translation and rotation	5×10^{-15}
Flexible molecules, rigid bonds	Translation, rotation, torsion	2×10^{-15}
Flexible molecules, flexible bonds	Translation, rotation, torsion, vibration	10^{-15} or 5×10^{-16}

Table 7.1 The different types of motion present in various systems together with suggested time steps

7.3.5 Multiple Time Step Dynamics

Table 7.1 presents us with something of a dilemma. We would obviously desire to explore as much of the phase space as possible but this may be compromised by the need for a small time step. One possible approach is to use a multiple time step method. The underlying rationale is that certain interactions evolve more rapidly with time than other interactions. The twin-range method (Section 6.7.1) is a crude type of multiple time step approach, in that interactions involving atoms between the lower and upper cutoff distance remain constant and change only when the neighbour list is updated. However, this approach can lead to an accumulation of numerical errors in calculated properties. A more sophisticated approach is to approximate the forces due to these atoms using a Taylor series expansion [Streett *et al.* 1978]:

$$\mathbf{f}(t + \tau\delta t) = \mathbf{f}(t) + (\tau\delta t)\mathbf{d}\mathbf{f}(t)/\mathbf{dt} + \frac{1}{2}(\tau\delta t)^2\mathbf{d}^2\mathbf{f}(t)/\mathbf{dt}^2 + \dots \quad (7.30)$$

This series expansion is truncated at a specified order and is probably most easily implemented within a predictor–corrector type of algorithm, where the higher-order terms are already computed. This method has been applied to relatively simple systems such as molecular fluids [Streett *et al.* 1978] and alkane chain liquids [Swindoll and Haile 1984].

An alternative formulation of a multiple time step method is the ‘reversible reference system propagation algorithm’ (r-RESPA) method [Tuckerman *et al.* 1992]. In this method, the forces within a system are classified into a number of groups according to how rapidly the force varies over time. Each group then has its own time step while maintaining accuracy and numerical stability. The starting point for this algorithm is the Liouville equation, which defines how the state of the system, $\Gamma(t)$, evolves over time:

$$\Gamma(t) = e^{iLt}\Gamma \quad (t = 0) \quad (7.31)$$

The exponential $\exp(iLt)$ in Equation (7.31) involves the so-called *Liouville operator*, L , which in the case of a molecular system containing N atoms (and so $3N$ coordinates) can be expressed:

$$iL = \sum_{i=1}^{3N} \left[\frac{\partial x_i}{\partial t} \frac{\partial}{\partial x_i} + F_i(x) \frac{\partial}{\partial p_i} \right] \quad (7.32)$$

In the r-RESPA method this operator is decomposed into two or more parts, for example:

$$L = L_1 + L_2 + L_3 + L_4 \quad (7.33)$$

Each of these parts is then associated with specific terms in the force equation. For example, L_1 may correspond to the bond-stretching terms, L_2 to the angle-bending and torsional terms, L_3 to the short-range non-bonded interactions and L_4 to the long-range non-bonded interactions. Suppose the time step with which we evaluate the bond-stretching terms is δt_1 . Integers n_1 , n_2 and n_3 then define the time steps for the three other forces as follows.

$$\delta t_2 = n_1 \delta t_1; \quad \delta t_3 = n_1 n_2 \delta t_1; \quad \delta t_4 = n_1 n_2 n_3 \delta t_1 \quad (7.34)$$

The underlying theory of r-RESPA is somewhat involved, but the final result and consequent implementation is actually rather straightforward, being very closely related to the velocity Verlet integration scheme. For our four-way decomposition the algorithm would

be implemented as follows:

```

Calculate forces-1 (i.e.  $\mathbf{a}_1(t)$ )
Calculate forces-2 (i.e.  $\mathbf{a}_2(t)$ )
Calculate forces-3 (i.e.  $\mathbf{a}_3(t)$ )
Calculate forces-4 (i.e.  $\mathbf{a}_4(t)$ )
do step = 1,  $N_{\text{steps}}$ 
     $\mathbf{v} = \mathbf{v} + \frac{1}{2}n_1n_2n_3\delta t_1\mathbf{a}_4$ 
    do  $i_3 = 1, n_3$ 
         $\mathbf{v} = \mathbf{v} + \frac{1}{2}n_1n_2\delta t_1\mathbf{a}_3$ 
        do  $i_2 = 1, n_2$ 
             $\mathbf{v} = \mathbf{v} + \frac{1}{2}n_1\delta t_1\mathbf{a}_2$ 
            do  $i_1 = 1, n_1$ 
                 $\mathbf{v} = \mathbf{v} + \frac{1}{2}\delta t_1\mathbf{a}_1$ 
                 $\mathbf{r} = \mathbf{r} + \delta t_1\mathbf{v}$ 
                calculate forces-1 (i.e.  $\mathbf{a}_1$ )
                 $\mathbf{v} = \mathbf{v} + \frac{1}{2}\delta t_1\mathbf{a}_1$ 
            enddo
            calculate forces-2 (i.e.  $\mathbf{a}_2$ )
             $\mathbf{v} = \mathbf{v} + \frac{1}{2}n_1\delta t_1\mathbf{a}_2$ 
        enddo
        calculate forces-3 (i.e.  $\mathbf{a}_3$ )
         $\mathbf{v} = \mathbf{v} + \frac{1}{2}n_1n_2\delta t_1\mathbf{a}_3$ 
    enddo
    calculate forces-4 (i.e.  $\mathbf{a}_4$ )
     $\mathbf{v} = \mathbf{v} + \frac{1}{2}n_1n_2n_3\delta t_1\mathbf{a}_4$ 
enddo

```

In this scheme, \mathbf{v} and \mathbf{r} refer to one of the $3N$ velocities or positions, respectively. Note that the different types of force are calculated throughout the algorithm. It can be readily seen that the method reduces to the standard velocity Verlet method if n_1 , n_2 and n_3 are set equal to 1.

The r-RESPA method has been applied to a variety of systems, including simple model systems [Tuckerman *et al.* 1992] but also organic molecules [Watanabe and Karplus 1993], fullerene crystals [Procacci and Berne 1994] and also proteins [Humphreys *et al.* 1994, 1996]. In these studies the reduction in computational time compared with the standard velocity Verlet method varied between 4–5 and 20–40, depending upon the size of the system, without any noticeable loss in accuracy. Other developments of the r-RESPA algorithm include its coupling to the fast multipole method (see Section 6.8.3) [Zhou and Berne 1995].

7.4 Setting Up and Running a Molecular Dynamics Simulation

In this section we will examine some of the steps involved in performing a molecular dynamics simulation in the microcanonical ensemble. First, it is necessary to establish an

initial configuration of the system. As discussed in Section 6.4.2, the initial configuration may be obtained from experimental data, from a theoretical model or from a combination of the two. It is also necessary to assign initial velocities to the atoms. This can be done by randomly selecting from a Maxwell-Boltzmann distribution at the temperature of interest:

$$p(v_{ix}) = \left(\frac{m_i}{2\pi k_B T} \right)^{1/2} \exp \left[-\frac{1}{2} \frac{m_i v_{ix}^2}{k_B T} \right] \quad (7.35)$$

The Maxwell-Boltzmann equation provides the probability that an atom i of mass m_i has a velocity v_{ix} in the x direction at a temperature T . A Maxwell-Boltzmann distribution is a Gaussian distribution, which can be obtained using a random number generator. Most random number generators are designed to produce random numbers that are uniform in the range 0 to 1. However, it is relatively straightforward to convert such a random number generator to sample from a Gaussian distribution (or indeed from one of several other distributions [Rubinstein 1981]). The probability of generating a value from a Gaussian (normal) distribution with mean $\langle x \rangle$ and variance σ^2 ($\sigma^2 = \langle (x - \langle x \rangle)^2 \rangle$) is:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[-\frac{(x - \langle x \rangle)^2}{2\sigma^2} \right] \quad (7.36)$$

One option is to first generate two random numbers ξ_1 and ξ_2 between 0 and 1. The corresponding two numbers from the normal distribution are then calculated using

$$x_1 = \sqrt{-2 \ln \xi_1} \cos(2\pi\xi_2) \quad \text{and} \quad x_2 = \sqrt{-2 \ln \xi_1} \sin(\pi\xi_2) \quad (7.37)$$

An alternative approach is to generate twelve random numbers ξ_1, \dots, ξ_{12} and then calculate:

$$x = \sum_{i=1}^{12} \xi_i - 6 \quad (7.38)$$

These two methods generate random numbers in the normal distribution with zero mean and unit variance. A number (x) generated from this distribution can be related to its counterpart (x') from another Gaussian distribution with mean $\langle x' \rangle$ and variance σ using

$$x' = \langle x' \rangle + \sigma x \quad (7.39)$$

The initial velocities may also be chosen from a uniform distribution or from a simple Gaussian distribution. In either case the Maxwell-Boltzmann distribution of velocities is usually rapidly achieved.

The initial velocities are often adjusted so that the total momentum of the system is zero. Such a system then samples from the constant NVEP ensemble. To set the total linear momentum of the system to zero, the sum of the components of the atomic momenta along the x , y and z axes is calculated. This gives the total momentum of the system in each direction, which, when divided by the total mass, is subtracted from the atomic velocities to give an overall momentum of zero.

Having set up the system and assigned the initial velocities, the simulation proper can commence. At each step the force on each atom must be calculated by differentiating the potential function. The force on an atom may include contributions from the various

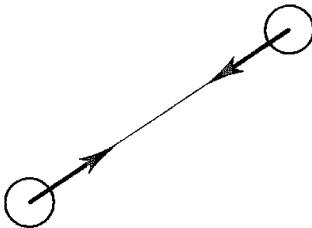


Fig. 7.7. The force between two particles acts along the line joining their centres of mass, in accordance with Newton's third law

terms in the force field such as bonds, angles, torsional terms and non-bonded interactions. The force is straightforward to calculate for two atoms interacting under the Lennard-Jones potential:

$$\mathbf{f}_{ij} = \frac{\mathbf{r}_{ij}}{|\mathbf{r}_{ij}|} \frac{24\epsilon}{\sigma} \left[2\left(\frac{\sigma}{r_{ij}}\right)^{13} - \left(\frac{\sigma}{r_{ij}}\right)^7 \right] \quad (7.40)$$

The force between the two atoms is equal in magnitude and opposite in direction and applies along the line connecting the two nuclear centres, in accordance with Newton's third law (Figure 7.7). It is necessary to calculate the force between each atom pair just once. This is most easily achieved by arranging to compute the force between an atom and those atoms with a higher index (i.e. for an atom i the forces are calculated with atoms $i + 1, i + 2, \dots, N$). Having calculated the force between an atom i and an atom with a higher index j , minus the force is added to the accumulating sum of the forces on j . The force calculation is most easily implemented using two loops, as outlined in the following pseudocode:

```

set elements on force array to zero
while atom1 = 1 to  $N - 1$ 
    while atom2 = atom1 + 1 to  $N$ 
        calculate force on atom1 due to interaction with atom2
        add the force to the array element, atom1
        subtract the force from the appropriate force array element atom2
    enddo
enddo

```

At the end of these two loops, the total force on each atom is known. A consequence of the fact that the force between the two atoms is equal and opposite is that the neighbour list for each atom need only contain those atoms with a higher number as the force on an atom due to interactions with lower numbered atoms will be calculated earlier in the loop. This organisation of the neighbour list contrasts with the structure used for Monte Carlo simulations, where all the neighbours of each atom (with both lower and higher indices) must be stored.

Analytical expressions for the forces due to other terms in the molecular mechanics potential function have been published for most of the functional forms encountered in common force

fields. These expressions can seem rather complicated because the intramolecular terms (e.g. bonds, angles, torsions) are calculated in terms of the internal coordinates, whereas molecular dynamics is typically performed using Cartesian coordinates. The chain rule must therefore be used to obtain the desired functional forms. However, the resulting expressions are relatively easy to implement in a computer program.

The first stage of a molecular dynamics simulation is the equilibration phase, the purpose of which is to bring the system to equilibrium from the starting configuration. During equilibration, various parameters are monitored together with the actual configurations. When these parameters achieve stable values then the production phase can commence. It is during the production phase that thermodynamic properties and other data are calculated. The parameters that are used to characterise whether equilibrium has been reached depend to some extent on the system being simulated but invariably include the kinetic, potential and total energies, the velocities, the temperature and the pressure. As we have indicated, the kinetic and potential energies would be expected to fluctuate in a simulation in the microcanonical ensemble but the total energy should remain constant. The components of the velocities should describe a Maxwell–Boltzmann distribution (in all three directions x , y and z) and the kinetic energy should be equally distributed among the three directions x , y and z . It is usually desired to perform a simulation at a specified temperature and so it is common practice to adjust the temperature of the system by scaling the velocities (see Section 7.7.1) during the equilibration phase. During the production phase the temperature is a variable of the system. Order parameters may be calculated to monitor changes in structure, which can supplement visual examination of the evolving trajectory.

When simulating an inhomogeneous system a more detailed equilibration procedure is usually desirable. A typical procedure suitable for a molecular dynamics simulation of a macromolecular solute, such as a protein in solution, would be as follows. First, the solvent alone together with any mobile counterions is subject to energy minimisation with the solute kept fixed in its initial conformation. The solvent and any counterions are then allowed to evolve using either a molecular dynamics (or indeed Monte Carlo) simulation, again keeping the structure of the solute molecule fixed. This solvent equilibration phase should be sufficiently extensive to allow the solvent to completely readjust to the potential field of the solute. For molecular dynamics this implies that the length of this solvent equilibration phase should be longer than the relaxation time of the solvent (the time taken for a molecule to lose any ‘memory’ of its original orientation, which for water is about 10 ps). Next, the entire system (solute and solvent) is minimised. Only then does the molecular dynamics simulation of the whole system commence.

At the start of the production phase all counters are set to zero and the system is permitted to evolve. In a microcanonical ensemble no velocity scaling is performed during the production phase and so the temperature becomes a calculated property of the system. Various properties are routinely calculated and stored during the production phase for subsequent analysis and processing. Careful monitoring of these properties during the simulation can show whether the simulation is ‘well behaved’ or not; it may be necessary to restart a simulation if problems are encountered. It is also usual to store the positions, energies

and velocities of configurations at regular intervals (e.g. every five to twenty time steps), from which other properties can be determined once the simulation has finished.

7.4.1 Calculating the Temperature

Many thermodynamic properties can be calculated from a molecular dynamics simulation. Most of these were dealt with in Section 6.2; here we just discuss the calculation of temperature. The instantaneous value of the temperature is related to the kinetic energy via the particles' momenta as follows:

$$\mathcal{K} = \sum_{i=1}^N \frac{|\mathbf{p}_i|^2}{2m_i} = \frac{k_B T}{2} (3N - N_c) \quad (7.41)$$

where N_c is the number of constraints and so $3N - N_c$ is the total number of degrees of freedom. For an isolated system (i.e. for a simulation of a system *in vacuo*) the total translational momentum of the system and the total angular momentum are conserved and can be made equal to zero by an appropriate choice of initial velocities. For a simulation performed using periodic boundary conditions, the total linear momentum is conserved but the total angular momentum is not. It is common practice to choose a set of initial velocities that ensures that the total linear momentum and the total angular momentum are zero. As the system evolves, the linear momentum should remain zero but the angular momentum will not. Molecular dynamics with periodic boundary conditions thus strictly samples from the constant *NVEP* ensemble where P is the total linear momentum. This differs trivially from the standard microcanonical ensemble but it should be remembered that the appropriate number of degrees of freedom must be subtracted from the total when calculating the kinetic energy per degree of freedom. Specifically, for a system *in vacuo* where the total linear and angular momenta have been set to zero, six degrees of freedom need to be subtracted. For a simulation using periodic boundary conditions three degrees of freedom need to be subtracted if the centre-of-mass motion of the system is removed. In constraint dynamics, discussed in the next section, rather more degrees of freedom may be fixed and N_c must be calculated accordingly.

7.5 Constraint Dynamics

The earliest molecular dynamics simulations using ‘realistic’ potentials were of atoms interacting under the Lennard-Jones potential. In such calculations the only forces on the atoms are those due to non-bonded interactions. It is rather more difficult to simulate molecules because the interaction between two non-spherical molecules depends upon their relative orientation as well as the distance between them. If the molecules are flexible then there will also be intramolecular interactions, which give rise to changes in conformation. Clearly, the simplest model is to treat the species present as rigid bodies with no intramolecular conformational freedom. In such cases the dynamics of each molecule can often be considered in terms of translations of its centre of mass and rotations about its centre of mass. The force on the molecule equals the vector sum of all the forces acting at the

centre of mass, and the rotational motion is determined by the torque about the centre of mass. To deal with these rotational motions is considerably more complicated than for the translational motions, but in favourable cases they can be programmed quite efficiently.

When the simulation involves conformationally flexible molecules then the motion is inevitably described in terms of the atomic Cartesian coordinates. The conformational behaviour of a flexible molecule is usually a complex superposition of different motions. The high frequency motions (e.g. bond vibrations) are usually of less interest than the lower frequency modes, which often correspond to major conformational changes. Unfortunately, the time step of a molecular dynamics simulation is dictated by the highest frequency motion present in the system. It would therefore be of considerable benefit to be able to increase the time step without prejudicing the accuracy of the simulation. Constraint dynamics enables individual internal coordinates or combinations of specified coordinates to be constrained, or 'fixed' during the simulation without affecting the other internal degrees of freedom.

Before we consider in detail the use of constraint dynamics, it is helpful to establish the difference between *constraints* and *restraints*; we shall discuss the method of restrained molecular dynamics in a later chapter (see Section 9.10). A constraint is a requirement that the system is forced to satisfy. As we shall see, in constraint dynamics bonds or angles are forced to adopt specific values throughout a simulation. When a bond or angle is restrained then it is able to deviate from the desired value; the restraint only acts to 'encourage' a particular value. Restraints are most easily incorporated using additional terms in the force field which impose a penalty for deviations from the reference value. An additional difference is that restrained degrees of freedom still have an energy $k_B T/2$ associated with them, whereas constrained degrees of freedom do not.

The most commonly used method for applying constraints, particularly in molecular dynamics, is the SHAKE procedure of Ryckaert, Ciccotti and Berendsen [Ryckaert *et al.* 1977]. In constraint dynamics the equations of motion are solved while simultaneously satisfying the imposed constraints. Constrained systems have been much studied in classical mechanics; we shall illustrate the general principles using a simple system comprising a box sliding down a frictionless slope in two dimensions (Figure 7.8). The box is constrained to remain on the slope and so the box's x and y coordinates must always satisfy the equation of the slope (which we shall write as $y = mx + c$). If the slope were not present then the box

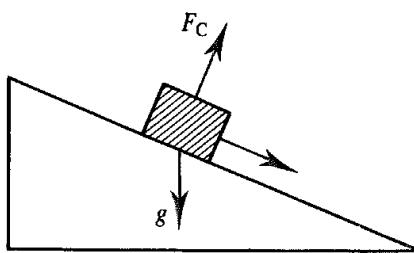


Fig. 7.8 A box sliding down a slope under the influence of gravity is subject to the constraint that it must remain on the slope. The constraint force F_c acts perpendicular to the direction of motion.

would fall vertically downwards. Constraints are often categorised as *holonomic* or *non-holonomic*. Holonomic constraints can be expressed in the form

$$f(q_1, q_2, q_3, \dots, t) = 0 \quad (7.42)$$

q_1, q_2 , etc., are the coordinates of the particles. Non-holonomic constraints cannot be expressed in this way. For example, the motion of a particle constrained to lie on the surface of a sphere is subject to a holonomic constraint, but if the particle is able to fall off the sphere under the influence of gravity then the constraint becomes non-holonomic. A holonomic constraint that keeps a particle on the surface of a sphere can be written:

$$r^2 - a^2 = 0 \quad (7.43)$$

r is the distance of the particle from the origin where the sphere of radius a is centred. The equivalent non-holonomic constraint is written as an inequality:

$$r^2 - a^2 \geq 0 \quad (7.44)$$

SHAKE uses holonomic constraints. In a constrained system the coordinates of the particles are not independent and the equations of motion in each of the coordinate directions are connected. A second difficulty is that the magnitude of the constraint forces is unknown. Thus in the case of the box on the slope, the gravitational force acting on the box is in the y direction whereas the motion is down the slope. The motion is thus not in the same direction as the gravitational force. As such, the total force on the box can be considered to arise from two sources one due to gravity and the other a constraint force that is perpendicular to the motion of the box (Figure 7.8). As there is no motion perpendicular to the surface of the slope, the constraint force does no work.

As we know, the motion of a system of N particles can be described in terms of $3N$ independent coordinates or degrees of freedom. If there are k holonomic constraints then the number of degrees of freedom is reduced to $3N - k$. It is possible, in principle at least, to find $3N - k$ independent coordinates (the *generalised coordinates*), which can then be used to solve the problem directly. For example, the motion of the box can be described using the single coordinate, q , along the direction of the slope. The component of the gravitational force that acts along the slope is $Mg \sin \theta$ and so the acceleration down the slope is $g \sin \theta$. The position at any time t can thus be obtained by integrating the following equation of motion:

$$\frac{d^2 q}{dt^2} = g \sin \theta \quad (7.45)$$

The solution to this equation is:

$$q(t) = q(0) + t\dot{q}(0) + \frac{t^2}{2} g \sin \theta \quad (7.46)$$

where $q(0)$ is the value of ξ at time $t = 0$ and $\dot{q}(0)$ is the initial velocity of the box along the slope. In this simple example it is quite easy to identify the single generalised coordinate that can be used to describe the motion in the constrained system. When there are many constraints, it can be difficult to determine the generalised coordinates. In any case, it is usually desirable to work with the atomic Cartesian coordinates. The motion of the box can be more generally described in terms of the Cartesian (x, y) coordinates of the box as follows

Newton's equations in the x and y directions are:

$$M \frac{d^2x}{dt^2} = F_{cx} \quad (7.47)$$

$$M \frac{d^2y}{dt^2} = -Mg + F_{cy} \quad (7.48)$$

where F_{cx} and F_{cy} are the components of the as yet unknown constraint force in the x and the y directions, respectively. We know that the constraint force acts perpendicular to the slope, and so the ratio of its x and y components must be:

$$\frac{F_{cx}}{F_{cy}} = -m \quad (7.49)$$

The constraint force can be introduced into Newton's equations as a Lagrange multiplier (see Section 1.10.5). To achieve consistency with the usual Lagrangian notation, we write F_{cy} as $-\lambda$ and so F_{cx} equals λm . Thus:

$$M \frac{d^2x}{dt^2} = \lambda m \quad (7.50)$$

$$M \frac{d^2y}{dt^2} = -Mg - \lambda \quad (7.51)$$

Equations (7.50) and (7.51) contain three unknowns (d^2x/dt^2 , d^2y/dt^2 and λ). A third equation that links x and y is the equation of the slope, which can be written in the following form:

$$\sigma = mx - y + c = 0 \quad (7.52)$$

This constraint equation is expressed in terms of x and y rather than their second derivatives. However, as $\sigma(x, y) = 0$ holds for all x, y , it follows that $d\sigma = 0$ and $d^2\sigma = 0$ also. Consequently, the constraint equation can be written:

$$m \frac{d^2x}{dt^2} - \frac{d^2y}{dt^2} = 0 \quad (7.53)$$

Solving the three equations gives:

$$\frac{d^2x}{dt^2} = -g \frac{m}{1 + m^2} \quad (7.54)$$

$$\frac{d^2y}{dt^2} = -g \frac{m^2}{1 + m^2} \quad (7.55)$$

The x and y coordinates at time t are thus given by:

$$x(t) = x(0) + t \frac{dx(0)}{dt} - g \frac{t^2}{2} \frac{m}{(1 + m^2)} \quad (7.56)$$

$$y(t) = y(0) + t \frac{dy(0)}{dt} - g \frac{t^2}{2} \frac{m^2}{(1 + m^2)} \quad (7.57)$$

In the general case, the equations of motion for a constrained system involve two types of force: the ‘normal’ forces arising from the intra- and intermolecular interactions, and the forces due to the constraints. We are particularly interested in the case where the constraint σ_k requires the bond between atoms i and j to remain fixed. The constraint influences the Cartesian coordinates of atoms i and j . The force due to this constraint can be written as follows:

$$F_{ckx} = \lambda_k \frac{\partial \sigma_k}{\partial x} \quad (7.58)$$

where λ_k is the Lagrange multiplier and x represents one of the Cartesian coordinates of the two atoms. Applying Equation (7.58) to the above example, we would write $F_{cx} = \lambda \partial \sigma / \partial x = \lambda m$ and $F_{cy} = \lambda \partial \sigma / \partial y = -\lambda$. If an atom is involved in a number of constraints (because it is involved in more than one constrained bond) then the total constraint force equals the sum of all such terms. The nature of the constraint for a bond between atoms i and j is:

$$\sigma_{ij} = (\mathbf{r}_i - \mathbf{r}_j)^2 - d_{ij}^2 = 0 \quad (7.59)$$

The constraint force lies along the bond at all times. For each constrained bond, there is an equal and opposite force on the two atoms that comprise the bond. The overall effect is that the constraint forces do no work. Suppose the constraint k corresponds to the bond length between atoms i and j . The constraint forces are obtained by differentiating the constraint with respect to the coordinates of atoms i and j and multiplying by an (as yet) undetermined multiplier:

$$\frac{\partial \sigma_k}{\partial \mathbf{r}_i} = 2(\mathbf{r}_i - \mathbf{r}_j) \quad \text{so} \quad F_{ci} = \lambda(\mathbf{r}_i - \mathbf{r}_j) \quad (7.60)$$

$$\frac{\partial \sigma_k}{\partial \mathbf{r}_j} = -2(\mathbf{r}_i - \mathbf{r}_j) \quad \text{and} \quad F_{cj} = -\lambda(\mathbf{r}_i - \mathbf{r}_j) \quad (7.61)$$

The factor of 2 that arises when we differentiate the square term has been incorporated into the Lagrange multiplier λ . The above expression for the forces can be incorporated into the Verlet algorithm as follows:

$$\mathbf{r}_i(t + \delta t) = 2\mathbf{r}_i(t) - \mathbf{r}_i(t - \delta t) + \frac{\delta t^2}{m_i} \mathbf{F}_i(t) + \sum_k \frac{\lambda_k \delta t^2}{m_i} \mathbf{r}_{ij}(t) \quad (7.62)$$

Recall that the positions that would be obtained from the Verlet algorithm without constraints are $\mathbf{r}'_i(t + \delta t) = 2\mathbf{r}_i(t) - \mathbf{r}_i(t - \delta t) + \delta t^2 \mathbf{F}_i(t)/m_i$. The summation in Equation (7.62) is over all constraints k that affect atom i . These constraints perturb the positions that would otherwise have been obtained from the integration algorithm, and so the above expression can be written:

$$\mathbf{r}_i(t + \delta t) = \mathbf{r}'_i(t + \delta t) + \sum_k \frac{\lambda_k \delta t^2}{m_i} \mathbf{r}_{ij}(t) \quad (7.63)$$

The next problem is to determine the multipliers λ_k that enable all the constraints to be satisfied simultaneously. This can be done algebraically in simple cases. Suppose we wish to fix the bond in a diatomic molecule. There is just one constraint in this case, and if the

atoms are labelled 1 and 2 we can write:

$$\mathbf{r}_1(t + \delta t) = \mathbf{r}'_1(t + \delta t) + \lambda_{12}(\delta t^2/m_1)(\mathbf{r}_1(t) - \mathbf{r}_2(t)) \quad (7.64)$$

$$\mathbf{r}_2(t + \delta t) = \mathbf{r}'_2(t + \delta t) - \lambda_{12}(\delta t^2/m_2)(\mathbf{r}_1(t) - \mathbf{r}_2(t)) \quad (7.65)$$

A third equation is derived from the requirement that the new positions keep the bond at the required distance:

$$|\mathbf{r}_1(t + \delta t) - \mathbf{r}_2(t + \delta t)|^2 = |\mathbf{r}_1(t) - \mathbf{r}_2(t)|^2 = d_{12}^2 \quad (7.66)$$

We now have three equations and three unknowns ($\mathbf{r}_1(t + \delta t)$, $\mathbf{r}_2(t + \delta t)$ and λ_{12}). Subtracting, and putting $\mathbf{r}_{12}(t) = (\mathbf{r}_1(t) - \mathbf{r}_2(t))$ and $\mathbf{r}'_{12}(t + \delta t) = \mathbf{r}'_1(t + \delta t) - \mathbf{r}'_2(t + \delta t)$ gives:

$$\mathbf{r}_1(t + \delta t) - \mathbf{r}_2(t + \delta t) = \mathbf{r}'_{12}(t + \delta t) + \lambda_{12}\delta t^2(1/m_1 + 1/m_2)\mathbf{r}_{12}(t) \quad (7.67)$$

Squaring both sides and imposing the constraint gives:

$$\mathbf{r}'_{12}(t + \delta t)^2 + 2\lambda_{12}\delta t^2(1/m_1 + 1/m_2)\mathbf{r}_{12}(t) + \lambda_{12}^2\delta t^4(1/m_1 + 1/m_2)^2\mathbf{r}_{12}(t)^2 = d_{12}^2 \quad (7.68)$$

Solving this quadratic equation for λ_{12} enables the new positions $\mathbf{r}_1(t + \delta t)$ and $\mathbf{r}_2(t + \delta t)$ to be determined.

In the case of a triatomic molecule with two bonds (between atoms 1,2 and 2,3), two constraint equations are obtained:

$$\mathbf{r}_{12}(t + \delta t) = \mathbf{r}'_{12}(t + \delta t) + \delta t^2(1/m_1 + 1/m_2)\lambda_{12}\mathbf{r}_{12}(t) - (\delta t^2/m_2)\lambda_{23}\mathbf{r}_{23}(t) \quad (7.69)$$

$$\mathbf{r}_{23}(t + \delta t) = \mathbf{r}'_{23}(t + \delta t) + \delta t^2(1/m_2 + 1/m_3)\lambda_{23}\mathbf{r}_{23}(t) - (\delta t^2/m_2)\lambda_{12}\mathbf{r}_{12}(t) \quad (7.70)$$

These expressions could be solved algebraically but even in this simple case the algebra becomes rather complicated. A solution can be obtained by ignoring the terms that are quadratic in λ as this produces equations which are linear in the Lagrange multipliers λ . When there are many constraints, the problem is equivalent to inverting a $k \times k$ matrix, even when the quadratic terms are ignored. The SHAKE method uses an alternative approach in which each constraint is considered in turn and solved. Satisfying one constraint may cause another constraint to be violated, and so it is necessary to iterate around the constraints until they are all satisfied to within some tolerance. The tolerance should be tight enough to ensure that the fluctuations in the simulation due to the SHAKE algorithm are much smaller than the fluctuations due to other sources, such as the use of cutoffs. Another important requirement is that the constrained degrees of freedom should be only weakly coupled to the remaining degrees of freedom, so that the motion of the molecule is not affected by the application of the constraints. The sampling of unconstrained degrees of freedom should also be unaffected. For example, if the bond lengths and angles are constrained in butane then the only degree of freedom remaining is the torsion angle. It is important that this torsion is able to explore its entire range of values in a way that is not biased because of the SHAKE procedure.

Our discussion so far has considered the use of SHAKE with the Verlet algorithm. Versions have also been derived for other integration schemes, such as the leap-frog algorithm, the predictor-corrector methods and the velocity Verlet algorithm. In the case of the velocity Verlet algorithm, the method has been named RATTLE [Anderson 1983].

When velocities appear in the integration algorithm these must be corrected as well as the positions.

Angle constraints can be easily accommodated in the SHAKE scheme by recognising that an angle constraint simply corresponds to an additional distance constraint. The angle in a triatomic molecule could thus be maintained at the desired value by requiring the distance between the two end atoms to adopt the appropriate value. This is how some small molecules (e.g. water) are maintained in a rigid geometry. For example, the simple point-charge (SPC) model of water uses three distance constraints. However, it is generally accepted that to constrain the bond angles in simulations of conformationally flexible molecules can have a deleterious effect on the efficiency with which the system explores configurational space. This is because many conformational transitions involve some opening or closing of angles as well as rotations about bonds. The most common use of SHAKE is for constraining bonds involving hydrogen atoms due to their much higher vibrational frequencies. This can enable the time step in a molecular dynamics simulation to be increased (e.g. from 1 fs to 2 fs).

The SHAKE method has been extended by Tobias and Brooks [Tobias and Brooks 1988] to enable constraints to be applied to arbitrary internal coordinates. This enables the torsion angle of a rotatable bond to be constrained to a particular value during a molecular dynamics simulation, which is particularly useful when used in conjunction with methods for calculating free energies (see Section 11.7)

7.6 Time-dependent Properties

Molecular dynamics generates configurations of the system that are connected in time and so an MD simulation can be used to calculate time-dependent properties. This is a major advantage of molecular dynamics over the Monte Carlo method. Time-dependent properties are often calculated as *time correlation coefficients*.

7.6.1 Correlation Functions

Suppose we have two sets of data values, x and y , and we wish to determine what correlation (if any) exists between them. For example, imagine that we are performing a simulation of a fluid in a capillary, and that we wish to determine the correlation between the absolute velocity of an atom and its distance from the wall of the tube. One way to do this would be to plot the sets of values as a graph. A correlation function (also known as a correlation coefficient) provides a numerical value that encapsulates the data and quantifies the strength of the correlation. A series of simulations with different capillary diameters could then be compared by examining the correlation coefficients. A variety of correlation functions can be defined, a commonly used one being:

$$C_{xy} = \frac{1}{M} \sum_{i=1}^M x_i y_i \equiv \langle x_i y_i \rangle \quad (7.71)$$

We have assumed that there are M values of x_i and y_i in the data sets. This correlation function can be normalised to a value between -1 and $+1$ by dividing by the root-mean-square values of x and y :

$$c_{xy} = \frac{\frac{1}{M} \sum_{i=1}^M x_i y_i}{\sqrt{\left(\frac{1}{M} \sum_{i=1}^M x_i^2\right) \left(\frac{1}{M} \sum_{i=1}^M y_i^2\right)}} = \frac{\langle x_i y_i \rangle}{\sqrt{\langle x_i^2 \rangle \langle y_i^2 \rangle}} \quad (7.72)$$

A value of 0 indicates no correlation and an absolute value of 1 indicates a high degree of correlation (which may be positive or negative). We will use a lowercase c to indicate a normalised correlation coefficient.

Sometimes the quantities x and y will fluctuate about non-zero mean values $\langle x \rangle$ and $\langle y \rangle$. Under such circumstances it is typical to consider just the fluctuating part and to define the correlation function as:

$$c_{xy} = \frac{\frac{1}{M} \sum_{i=1}^M (x_i - \langle x \rangle)(y_i - \langle y \rangle)}{\sqrt{\left(\frac{1}{M} \sum_{i=1}^M (x_i - \langle x \rangle)^2\right) \left(\frac{1}{M} \sum_{i=1}^M (y_i - \langle y \rangle)^2\right)}} = \frac{\langle (x_i - \langle x \rangle)(y_i - \langle y \rangle) \rangle}{\sqrt{\langle (x_i - \langle x \rangle)^2 \rangle \langle (y_i - \langle y \rangle)^2 \rangle}} \quad (7.73)$$

c_{xy} can also be written in the following useful way:

$$c_{xy} = \frac{\sum_{i=1}^M x_i y_i - \frac{1}{M} \left(\sum_{i=1}^M x_i \right) \left(\sum_{i=1}^M y_i \right)}{\sqrt{\left[\sum_{i=1}^M x_i^2 - \frac{1}{M} \left(\sum_{i=1}^M x_i \right)^2 \right] \left[\sum_{i=1}^M y_i^2 - \frac{1}{M} \left(\sum_{i=1}^M y_i \right)^2 \right]}} \quad (7.74)$$

Equation (7.74) does not require the mean values $\langle x \rangle$ and $\langle y \rangle$ to be determined before the correlation coefficient can be calculated and so values can be accumulated as the simulation proceeds.

A molecular dynamics simulation provides data values at specific times. This enables the value of some property at some instant to be correlated with the value of the same or another property at a later time t . The resulting values are known as *time correlation coefficients*. The correlation function is then written:

$$C_{xy}(t) = \langle x(t)y(0) \rangle \quad (7.75)$$

The following two results are useful:

$$\lim_{t \rightarrow 0} C_{xy}(0) = \langle xy \rangle \quad (7.76)$$

$$\lim_{t \rightarrow \infty} C_{xy}(t) = \langle x \rangle \langle y \rangle \quad (7.77)$$

If the quantities x and y are different, then the correlation function is sometimes referred to as a *cross-correlation function*. When x and y are the same then the function is usually called an *autocorrelation function*. An autocorrelation function indicates the extent to which the system retains a ‘memory’ of its previous values (or, conversely, how long it takes the system to ‘lose’ its memory). A simple example is the velocity autocorrelation coefficient whose value indicates how closely the velocity at a time t is correlated with the velocity at time 0. Some correlation functions can be averaged over all the particles in the system (as can the velocity autocorrelation function) whereas other functions are a property of the entire system (e.g. the dipole moment of the sample). The value of the velocity autocorrelation coefficient can be calculated by averaging over the N atoms in the simulation:

$$C_{vv}(t) = \frac{1}{N} \sum_{i=1}^N \mathbf{v}_i(t) \cdot \mathbf{v}_i(0) \quad (7.78)$$

To normalise the function, we divide by $\langle \mathbf{v}_i(0) \cdot \mathbf{v}_i(0) \rangle$:

$$c_{vv}(t) = \frac{1}{N} \sum_{i=1}^N \frac{\langle \mathbf{v}_i(t) \cdot \mathbf{v}_i(0) \rangle}{\langle \mathbf{v}_i(0) \cdot \mathbf{v}_i(0) \rangle} \quad (7.79)$$

In general, an autocorrelation function such as the velocity autocorrelation coefficient has an initial value of 1 and at long times has the value 0. The time taken to lose the correlation is often called the *correlation time*, or the *relaxation time*. If the duration of the simulation is significantly longer than the relaxation time (as it should be), then many sets of data can be extracted from the simulation to calculate the correlation function and to reduce the uncertainty in the calculation. If P steps of molecular dynamics are required for complete relaxation, and the simulation has been run for a total of Q steps, then $(Q - P)$ different sets of values could be used to calculate a value for the correlation function. The first set would run from step 1 to step N ; the second from step 2 to step $N + 1$, and so on (Figure 7.9). In fact, as we saw in Section 6.9 the high degree of correlation between successive time steps means that it is common to use time origins that are separated by several time steps, as shown in Figure 7.9. If we use M time origins (t_j) then the velocity autocorrelation function is given by:

$$C_{vv}(t) = \frac{1}{MN} \sum_{j=1}^M \sum_{i=1}^N \mathbf{v}_i(t_j) \cdot \mathbf{v}_i(t_j + t) \quad (7.80)$$

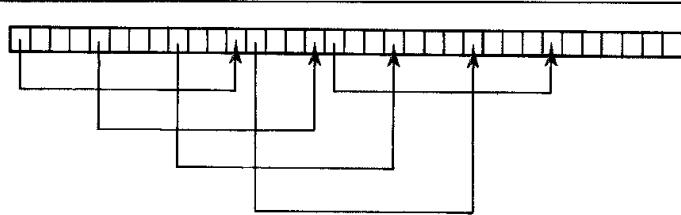


Fig. 7.9. The use of different time origins improves the accuracy with which time correlation functions can be calculated.

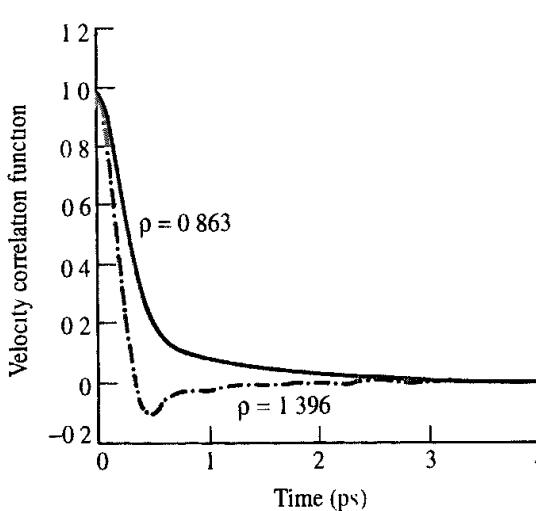


Fig. 7.10. Velocity autocorrelation functions for liquid argon at densities of 1.396 g cm^{-3} and 0.863 g cm^{-3}

Quantities with small relaxation times can thus be determined with greater statistical precision, as it will be possible to include a greater number of data sets from a given simulation. Moreover, no quantity with a relaxation time greater than the length of the simulation can be determined accurately.

The velocity autocorrelation functions obtained from molecular dynamics simulations of argon at two different densities are shown in Figure 7.10. The correlation coefficient has an initial value of 1 and is then quadratic at short times, a result that is predicted theoretically. The subsequent behaviour depends upon the density of the fluid. For a low-density fluid, the velocity autocorrelation coefficient gradually decreases to zero. At high densities $c_{vv}(t)$ crosses the axis and then becomes negative. A negative correlation coefficient simply means that the particle is now moving in the direction opposite to that at $t = 0$. This result has been interpreted in terms of a 'cage' structure of the liquid; the atom hits the side of the cage formed by its nearest neighbours and rebounds, reversing the direction of its motion. At both low density and high density the decay towards zero is significantly slower than the exponential decay predicted by kinetic theory. In fact, the decay varies as $t^{-3/2}$. This was one of the most interesting results obtained from the early molecular dynamics simulations and can be observed even with a hard-sphere model [Alder and Wainwright 1970]. The phenomenon is ascribed to the formation of a 'hydrodynamic vortex'. As the atom moves through the fluid it pushes other atoms out of the way. These atoms circle around and eventually give it a final 'push' so resulting in a less rapid decrease to zero (Figure 7.11).

The slow decay of the velocity autocorrelation function can present practical problems when deriving other properties, such as the transport coefficients, that require the correlation function to be integrated between $t = 0$ and $t = \infty$. The so-called 'long time-tail' of the autocorrelation function makes a significant contribution to the integral, but unfortunately the statistical uncertainty with which this part of the function can be calculated is greater as fewer segments of the appropriate length can be extracted from the simulation.

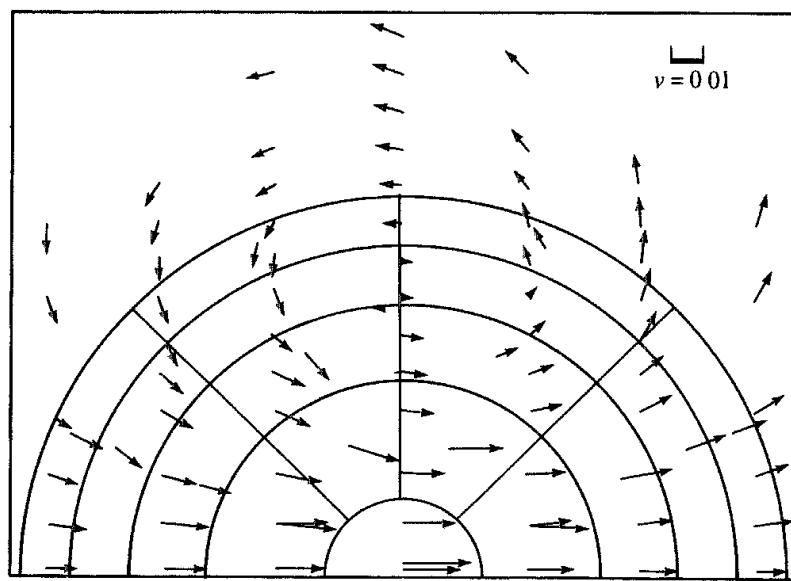


Fig 7.11 The slow decay of the velocity autocorrelation function towards zero can be explained in terms of the formation of a hydrodynamic vortex (Figure adapted from Alder B J and T E Wainwright 1970. Decay of the Velocity Autocorrelation Function Physical Review A 1:18-21)

The velocity autocorrelation function is an example of a single-particle correlation function, in which the average is calculated not only over time origins but also over all the atoms. Some properties are calculated for the entire system. One such property is the net dipole moment of the system, which is the vector sum of all the individual dipoles of the molecules in the system (clearly the dipole moment of the system can be non-zero only if each individual molecule has a dipole). The magnitude and orientation of the net dipole will change with time and is given by:

$$\mu_{\text{tot}}(t) = \sum_{i=1}^N \mu_i(t) \quad (7.81)$$

$\mu_i(t)$ is the dipole moment of molecule i at time t . The total dipolar correlation function is given by:

$$c_{\text{dipole}}(t) = \frac{\langle \mu_{\text{tot}}(t) \cdot \mu_{\text{tot}}(0) \rangle}{\langle \mu_{\text{tot}}(0) \cdot \mu_{\text{tot}}(0) \rangle} \quad (7.82)$$

The dipole correlation time of the system is a valuable quantity to calculate as it is related to the sample's absorption spectrum. Liquids usually absorb in the infrared region of the electromagnetic spectrum, a typical spectrum being shown in Figure 7.12. As can be seen, the spectrum is very broad with none of the sharp peaks characteristic of a well-resolved spectrum for a species in the gas phase. This is because the overall dipole of a liquid does not change at a constant rate but, rather, there is a distribution of frequencies. The intensity of absorption at any frequency depends upon the relative contribution of that frequency to the overall distribution. If, on average, the overall dipole changes very rapidly (i.e. the relaxation time is short) then the maximum in the absorption spectrum will occur at a

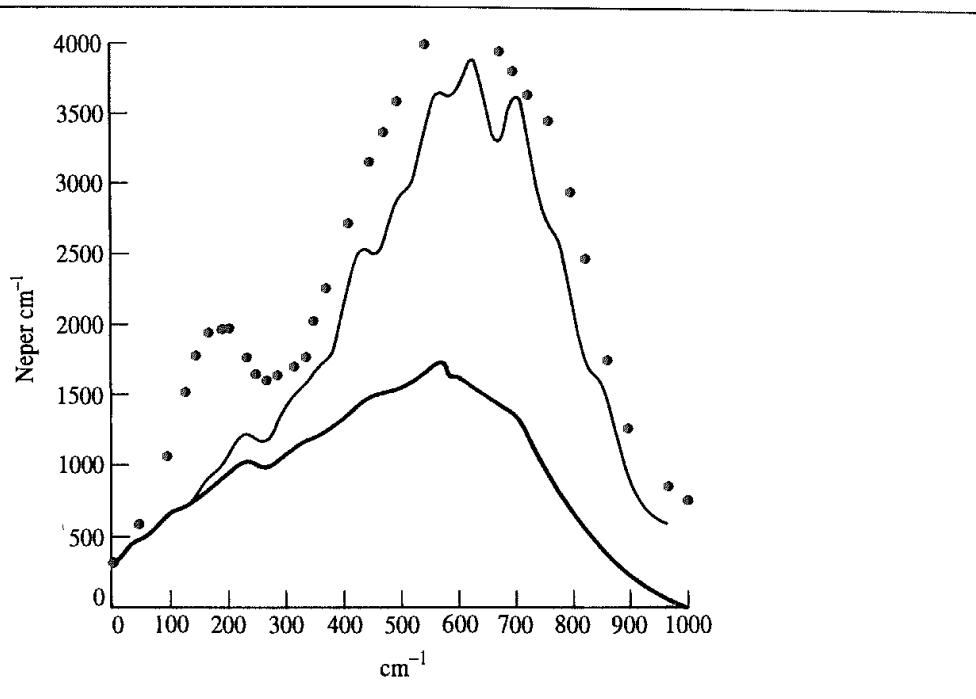


Fig 7.12. Experimental and calculated infrared spectra for liquid water. The black dots are the experimental values. The thick curve is the classical profile produced by the molecular dynamics simulation. The thin curve is obtained by applying quantum corrections. (Figure redrawn from Guillot B 1991. A Molecular Dynamics Study of the Infrared Spectrum of Water. Journal of Chemical Physics 95 1543–1551.)

higher frequency than if the relaxation time is long. To predict the spectrum from the correlation function it is therefore necessary to extract the relative contribution of each dipole fluctuation. This is done using Fourier analysis techniques, in which the correlation function is transformed from the time domain into the frequency domain (an introduction to Fourier analysis is provided in Section 1.10.8). The Fourier analysis picks out the intensity of dipole fluctuation at each frequency using the following relationship:

$$\hat{c}_{\text{dipole}}(\nu) = \int_{-\infty}^{\infty} c_{\text{dipole}}(t) \exp(-i2\pi\nu t) dt \quad (7.83)$$

Having calculated the Fourier transform it is then possible to plot the simulated spectrum and compare it to that obtained by experiment, as in Figure 7.12.

7.6.2 Orientational Correlation Functions

Other orientational correlation coefficients can be calculated in the same way as the correlation coefficients that we have discussed already. Thus, the reorientational correlation coefficient of a single rigid molecule indicates the degree to which the orientation of a molecule at a time t is related to its orientation at time 0. The angular velocity autocorrelation function is the rotational equivalent of the velocity correlation function:

$$c_{\omega\omega}(t) = \frac{\langle \omega_i(t) \cdot \omega_i(0) \rangle}{\langle \omega_i(0) \cdot \omega_i(0) \rangle} \quad (7.84)$$

In a liquid, the rotation of a molecule is influenced by neighbouring molecules and over time the correlation will decay to zero. The information embodied in the orientational correlation functions can be compared to a variety of spectroscopic experiments, including infrared, Raman and NMR spectra. For non-spherical molecules it can be useful to derive separate auto-correlation functions for the angular velocity along each of the principal axes of rotation. For example, for a spherical molecule such as CBr_4 neighbouring molecules have a relatively small influence on the loss in correlation in the angular velocity. By contrast, a linear molecule such as CS_2 experiences significant torques as it rotates. This has the effect of damping the rotational motion more severely than for the spherical case, and indeed the correlation function can change sign, indicating that the molecule is now rotating in the opposite direction. For some molecules such as water the presence of specific interactions between molecules (for example, due to hydrogen bonding) can give rise to very rapid damping and several minima in $c_{\omega\omega}(t)$.

7.6.3 Transport Properties

Transport refers to a phenomenon that gives rise to a flow of material from one region to another. For example, if a solution is prepared with a non-equilibrium solute distribution, then the solute diffuses until the concentration is equal throughout. If a thermal gradient is created, energy flows until the temperature is equalised. A momentum gradient gives rise to viscosity. The very existence of transport implies that the system is not in equilibrium. Techniques have been developed to perform non-equilibrium molecular dynamics simulations from which transport properties can be calculated, but we will not consider these here. We are thus faced with the problem of calculating non-equilibrium properties from equilibrium simulations. This may seem an impossible task but can be achieved by considering the microscopic local fluctuations that occur even in systems at equilibrium. We should be aware, however, that non-equilibrium molecular dynamics simulations can be a more efficient way to calculate transport properties and other quantities [Allen and Tildesley 1987].

To a first approximation the rate at which transport of the relevant quantity occurs (called the *flux*) is proportional to the gradient of the property with the constant of proportionality being the relevant transport property coefficient. For example, the flux of matter J_z (i.e. the amount passing through unit area in unit time) equals the diffusion coefficient (D) multiplied by the concentration gradient; this is Fick's first law of diffusion:

$$J_z = -D(d\mathcal{N}/dz) \quad (7.85)$$

\mathcal{N} is the number density (the number of atoms per unit volume). Equation (7.85) refers to diffusion in the z direction. The minus sign indicates that flux increases in the direction of negative concentration gradient. The time dependence of diffusive behaviour (which applies if a distribution is established at some time and is then allowed to evolve) is governed by Fick's second law, which gives the rate of change of concentration with time:

$$\frac{\partial \mathcal{N}(z, t)}{\partial t} = D \frac{\partial^2 \mathcal{N}(z, t)}{\partial z^2} \quad (7.86)$$

To solve Fick's second law equation it is necessary to impose two boundary conditions for the spatial dependence and one boundary condition for the temporal dependence (the

equation is second order in space and first order in time). For example, we might require that at time zero all N_0 particles have $z = 0$. The solution to the equation is then.

$$\mathcal{N}(z, t) = \frac{N_0}{A\sqrt{\pi Dt}} \exp\left[-\frac{z^2}{4Dt}\right] \quad (7.87)$$

where A is the cross-sectional area of the sample. Equation (7.87) is a Gaussian function which initially has a sharp peak at $z = 0$ but which gradually becomes more spread out as time progresses. When the material being simulated is a pure liquid then the coefficient D is often referred to as a *self-diffusion coefficient*. The diffusion coefficient is related to the mean square distance, $\langle |\mathbf{r}(t) - \mathbf{r}(0)|^2 \rangle$, which Einstein showed was equal to $2Dt$. In three dimensions, the mean square displacement is given by:

$$3D = \lim_{t \rightarrow \infty} \frac{\langle |\mathbf{r}(t) - \mathbf{r}(0)|^2 \rangle}{2t} \quad (7.88)$$

As indicated, the relationship strictly holds only in the limit as $t \rightarrow \infty$.

The Einstein relationship can thus be used to calculate the diffusion coefficient from an equilibrium simulation, by plotting the mean square displacement as a function of time and then attempting to obtain the limiting behaviour as $t \rightarrow \infty$ (Fick's law is inapplicable at short times). The quantity $|\mathbf{r}(t) - \mathbf{r}(0)|$ can be averaged over the particles in the system to reduce the statistical error. It is also usual practice to average over time origins where possible. When using this method for calculating the diffusion coefficient the mean-squared distances should not be limited by the edges of the periodic box. In other words, we require a set of positions that have not been translated back into the central simulation cell. This can be achieved either by storing a set of 'uncorrected' positions or indeed by not correcting any of the positions during the simulation and simply generating the appropriate minimum image positions as required for the calculation of the energy or forces.

Einstein relationships hold for other transport properties, e.g. the shear viscosity, the bulk viscosity and the thermal conductivity. For example, the shear viscosity η is given by:

$$\eta_{xy} = \frac{1}{V k_B T} \lim_{t \rightarrow \infty} \frac{\langle (\sum_{i=1}^N m \dot{x}_i(t) y_i(t) - \sum_{i=1}^N m \dot{y}_i(t) x_i(t))^2 \rangle}{2t} \quad (7.89)$$

The shear viscosity is a tensor quantity, with components η_{xy} , η_{xz} , η_{yx} , η_{yz} , η_{zy} , η_{zx} . It is a property of the whole sample rather than of individual atoms and so cannot be calculated with the same accuracy as the self-diffusion coefficient. For a homogeneous fluid the components of the shear viscosity should all be equal and so the statistical error can be reduced by averaging over the six components. An estimate of the precision of the calculation can then be determined by evaluating the standard deviation of these components from the average. Unfortunately, Equation (7.89) cannot be directly used in periodic systems, even if the positions have been unfolded, because the 'unfolded' distance between two particles may not correspond to the distance of the minimum image that is used to calculate the force. For this reason alternative approaches are required.

One alternative approach to the calculation of the diffusion and other transport coefficients is via an appropriate autocorrelation function. For example, the diffusion coefficient

depends upon the way in which the atomic position $\mathbf{r}(t)$ changes with time. At a time t the difference between $\mathbf{r}(t)$ and $\mathbf{r}(0)$ is given by:

$$|\mathbf{r}(t) - \mathbf{r}(0)| = \int_0^t \mathbf{v}(t') dt' \quad (7.90)$$

If both sides of Equation (7.90) are now squared then we obtain the following for the mean-square value:

$$\langle |\mathbf{r}(t) - \mathbf{r}(0)|^2 \rangle = \int_0^t dt' \int_0^t dt'' \langle \mathbf{v}(t') \cdot \mathbf{v}(t'') \rangle \quad (7.91)$$

The crucial feature to recognise is that the relevant correlation functions are unaffected by changing the origin, which means that the following holds:

$$\langle \mathbf{v}(t') \cdot \mathbf{v}(t'') \rangle = \langle \mathbf{v}(t'' - t') \cdot \mathbf{v}(0) \rangle \quad (7.92)$$

Integration of the double integral, Equation (7.91) leads to the *Green-Kubo* formula:

$$\frac{\langle |\mathbf{r}(t) - \mathbf{r}(0)|^2 \rangle}{2t} = \int_0^t \langle \mathbf{v}(\tau) \cdot \mathbf{v}(0) \rangle \left(1 - \frac{\tau}{t}\right) d\tau \quad (7.93)$$

In the limit:

$$\int_0^\infty \langle \mathbf{v}(\tau) \cdot \mathbf{v}(0) \rangle d\tau = \lim_{t \rightarrow \infty} \frac{\langle |\mathbf{r}(t) - \mathbf{r}(0)|^2 \rangle}{2t} = 3D \quad (7.94)$$

We can now see why long time-tails in the autocorrelation functions are so important. The area under the curve during the slow decay towards zero may be a significant part of the integral in the Green-Kubo formula. In practice, these integrals are determined numerically. The long time-tail may be dealt with by fitting a function to the curve and then attempting to integrate to infinity.

7.7 Molecular Dynamics at Constant Temperature and Pressure

Molecular dynamics is traditionally performed in the constant NVE (or $NVEP$) ensemble. Although thermodynamic results can be transformed between ensembles, this is strictly only possible in the limit of infinite system size ('the thermodynamic limit'). It may therefore be desired to perform the simulation in a different ensemble. The two most common alternative ensembles are the constant NVT and the constant NPT ensembles. In this section we will therefore consider how molecular dynamics simulations can be performed under conditions of constant temperature and/or constant pressure.

7.7.1 Constant Temperature Dynamics

There are several reasons why we might want to maintain or otherwise control the temperature during a molecular dynamics simulation. Even in a constant NVE simulation it is common practice to adjust the temperature to the desired value during the equilibration phase. A constant temperature simulation may be required if we wish to determine how

the behaviour of the system changes with temperature, such as the unfolding of a protein or glass formation. Simulated annealing algorithms require the temperature of the system to be reduced gradually while the system explores its degrees of freedom. Simulated annealing is used in searching conformational space and in the elucidation of macromolecular structure from NMR and X-ray data and is discussed in Section 9.9.2.

The temperature of the system is related to the time average of the kinetic energy, which for an unconstrained system is given by:

$$\langle \mathcal{K} \rangle_{NVT} = \frac{3}{2} N k_B T \quad (7.95)$$

An obvious way to alter the temperature of the system is thus to scale the velocities [Woodcock 1971]. If the temperature at time t is $T(t)$ and the velocities are multiplied by a factor λ , then the associated temperature change can be calculated as follows:

$$\Delta T = \frac{1}{2} \sum_{i=1}^N \frac{2}{3} \frac{m_i(\lambda v_i)^2}{N k_B} - \frac{1}{2} \sum_{i=1}^N \frac{2}{3} \frac{m_i v_i^2}{N k_B} \quad (7.96)$$

$$\Delta T = (\lambda^2 - 1)T(t) \quad (7.97)$$

$$\lambda = \sqrt{T_{\text{new}}/T(t)} \quad (7.98)$$

The simplest way to control the temperature is thus to multiply the velocities at each time step by the factor $\lambda = \sqrt{T_{\text{req}}/T_{\text{curr}}}$, where T_{curr} is the current temperature as calculated from the kinetic energy and T_{req} is the desired temperature.

An alternative way to maintain the temperature is to couple the system to an external heat bath that is fixed at the desired temperature [Berendsen *et al.* 1984]. The bath acts as a source of thermal energy, supplying or removing heat from the system as appropriate. The velocities are scaled at each step, such that the rate of change of temperature is proportional to the difference in temperature between the bath and the system:

$$\frac{dT(t)}{dt} = \frac{1}{\tau} (T_{\text{bath}} - T(t)) \quad (7.99)$$

τ is a coupling parameter whose magnitude determines how tightly the bath and the system are coupled together. This method gives an exponential decay of the system towards the desired temperature. The change in temperature between successive time steps is:

$$\Delta T = \frac{\delta t}{\tau} (T_{\text{bath}} - T(t)) \quad (7.100)$$

The scaling factor for the velocities is thus:

$$\lambda^2 = 1 + \frac{\delta t}{\tau} \left(\frac{T_{\text{bath}}}{T(t)} - 1 \right) \quad (7.101)$$

If τ is large, then the coupling will be weak. If τ is small, the coupling will be strong and when the coupling parameter equals the time step ($\tau = \delta t$) then the algorithm is equivalent to the simple velocity scaling method. A coupling constant of approximately 0.4 ps has been suggested as an appropriate value to use when the time step is 1 fs, giving $\delta t/\tau \approx 0.0025$. The advantage of this approach is that it does permit the system to fluctuate about the desired temperature.

These two relatively simple temperature scaling methods do not generate rigorous canonical averages. Velocity scaling artificially prolongs any temperature differences among the components of the system, which can lead to the phenomenon of 'hot solvent, cold solute', in which the 'temperature' of the solute is lower than that of the solvent, even though the overall temperature of the system is at the desired value. One 'solution' to this problem is to apply temperature coupling separately to the solute and to the solvent but the problem of unequal distribution of energy between the various components (and between the various modes of motion) may still remain. Two methods that do generate rigorous canonical ensembles if properly implemented are the *stochastic collisions* method and the *extended system* method.

In the stochastic collisions method a particle is randomly chosen at intervals and its velocity is reassigned by random selection from the Maxwell-Boltzmann distribution [Anderson 1980]. This is equivalent to the system being in contact with a heat bath that randomly emits 'thermal particles' which collide with the atoms in the system. Between each collision the system is simulated at constant energy and so the overall effect is equivalent to a series of microcanonical simulations, each performed at a slightly different energy. The distribution of the energies of these 'mini microcanonical' simulations will be a Gaussian function. The stochastic collisions method does not, of course, generate a smooth trajectory, which may be a drawback. By calculating the energy change due to a collision, Anderson showed that the mean rate (ν) at which each particle should suffer a stochastic collision is given by:

$$\nu = \frac{2a\kappa}{3k_B \mathcal{N}^{1/3} N^{2/3}} \quad (7.102)$$

a is a dimensionless constant, κ is the thermal conductivity and \mathcal{N} is the number density of the particles. If the thermal conductivity is not known then a suitable value of ν can be obtained from the intermolecular collision frequency ν_c :

$$\nu = \nu_c / N^{2/3} \quad (7.103)$$

If the collision rate is too low then the system does not sample from a canonical distribution of energies. If it is too high then the temperature control algorithm dominates and the system does not show the expected fluctuations in kinetic energy. The velocity of more than one particle can be changed in the stochastic collision method; in the limit the velocities of all the particles are changed simultaneously, though it is preferable to do this at quite long intervals. A distinction can thus be made between 'minor' collisions, in which only one (or a few) particles are affected, and 'major' (or 'massive') collisions, where the velocities of all particles are changed. It is also possible to use a combined approach, with minor collisions occurring relatively frequently and major collisions at longer intervals.

Extended system methods, originally introduced for performing constant temperature molecular dynamics by Nosé [Nosé 1984] and subsequently developed by Hoover [Hoover 1985], consider the thermal reservoir to be an integral part of the system. The reservoir is represented by an additional degree of freedom, labelled s . The reservoir has potential energy $(f+1)k_B T \ln s$, where f is the number of degrees of freedom in the physical system and T is the desired temperature. The reservoir also has kinetic energy $(Q/2)(ds/dt)^2$. Q is a parameter with the dimensions of energy \times (time) 2 and can be considered the

(fictitious) mass of the extra degree of freedom. The magnitude of Q determines the coupling between the reservoir and the real system and so influences the temperature fluctuations.

Each state of the extended system that is generated by the molecular dynamics simulation corresponds to a unique state of the real system. There is not, however, a direct correspondence between the velocities and the time in the real and the extended systems. The velocities of the atoms in the real system are given by:

$$\mathbf{v}_i = s \frac{d\mathbf{r}_i}{dt} \quad (7.104)$$

\mathbf{r}_i is the position of particle i in the simulation and \mathbf{v}_i is considered to be the real velocity of the particle. The time step $\delta t'$ is related to the time step in 'real time' δt by

$$\delta t = s\delta t' \quad (7.105)$$

The value of the additional degree of freedom s can change and so the time step in real time can fluctuate. Thus regular time intervals in the extended system correspond to a trajectory of the real system which is unevenly spaced in time.

The parameter Q controls the energy flow between the system and the reservoir. If Q is large then the energy flow is slow; in the limit of infinite Q , conventional molecular dynamics is regained and there is no energy exchange between the reservoir and the real system. However, if Q is too small then the energy oscillates, resulting in equilibration problems. Nosé has suggested that Q should be proportional to $f k_B T$; the constant of proportionality can then be obtained by performing a series of trial simulations for a test system and observing how well the system maintains the desired temperature.

7.7.2 Constant Pressure Dynamics

Just as one may wish to specify the temperature in a molecular dynamics simulation, so it may be desired to maintain the system at a constant pressure. This enables the behaviour of the system to be explored as a function of the pressure, enabling one to study phenomena such as the onset of pressure-induced phase transitions. Many experimental measurements are made under conditions of constant temperature and pressure, and so simulations in the isothermal-isobaric ensemble are most directly relevant to experimental data. Certain structural rearrangements may be achieved more easily in an isobaric simulation than in a simulation at constant volume. Constant pressure conditions may also be important when the number of particles in the system changes (as in some of the 'test particle' methods for calculating free energies and chemical potentials; see Section 8.9).

The pressure often fluctuates much more than quantities such as the total energy in a constant NVE molecular dynamics simulation. This is as expected because the pressure is related to the virial, which is obtained as the product of the positions and the derivative of the potential energy function. This product, $r_{ij} dV(r_{ij})/dr_{ij}$, changes more quickly with r than does the internal energy, hence the greater fluctuation in the pressure.

A macroscopic system maintains constant pressure by changing its volume. A simulation in the isothermal-isobaric ensemble also maintains constant pressure by changing the volume

of the simulation cell. The amount of volume fluctuation is related to the isothermal compressibility, κ .

$$\kappa = -\frac{1}{V} \left(\frac{\partial V}{\partial P} \right)_T \quad (7.106)$$

An easily compressible substance has a larger value of κ , so larger volume fluctuations occur at a given pressure than in a more incompressible substance. Conversely, in a constant volume simulation a less compressible substance shows larger fluctuations in the pressure. The isothermal compressibility is the pressure analogue of the heat capacity, which is related to the energy fluctuations.

A volume change in an isobaric simulation can be achieved by changing the volume in all directions, or in just one direction. It is instructive to consider the range of volume changes that one might expect to observe in a constant pressure simulation of a ‘typical’ system. The isothermal compressibility is related to the mean square volume displacement by:

$$\kappa = \frac{1}{k_B T} \frac{\langle V^2 \rangle - \langle V \rangle^2}{\langle V^2 \rangle} \quad (7.107)$$

The isothermal compressibility of an ideal gas is approximately 1 atm^{-1} . So for a simulation in a box of side 20 \AA (volume 8000 \AA^3) at 300 K , the root mean square change in the volume is approximately 18100 \AA^3 . This is larger than the initial size of the box! For a relatively incompressible substance such as water ($\kappa = 44.75 \times 10^{-6} \text{ atm}^{-1}$) the fluctuation is 121 \AA^3 , which corresponds to the box only changing by about 0.1 \AA in each direction. These values have clear implications for the appropriate size of the simulation system.

Many of the methods used for pressure control are analogous to those used for temperature control. Thus, the pressure can be maintained at a constant value by simply scaling the volume. An alternative is to couple the system to a ‘pressure bath’, analogous to a temperature bath [Berendsen *et al.* 1984]. The rate of change of pressure is given by:

$$\frac{dP(t)}{dt} = \frac{1}{\tau_p} (P_{\text{bath}} - P(t)) \quad (7.108)$$

τ_p is the coupling constant, P_{bath} is the pressure of the ‘bath’, and $P(t)$ is the actual pressure at time t . The volume of the simulation box is scaled by a factor λ , which is equivalent to scaling the atomic coordinates by a factor $\lambda^{1/3}$. Thus:

$$\lambda = 1 - \kappa \frac{\delta t}{\tau_p} (P - P_{\text{bath}}) \quad (7.109)$$

The new positions are given by:

$$\mathbf{r}'_i = \lambda^{1/3} \mathbf{r}_i \quad (7.110)$$

The constant κ can be combined with the relaxation constant τ_p as a single constant. This expression can be applied isotropically (i.e. such that the scaling factor is equal for all three directions) or anisotropically (where the scaling factor is calculated independently for each of the three axes). In general, it is best to use the anisotropic approach as this enables the box dimensions to change independently. Unfortunately, it has not been possible to determine from which ensemble this method samples.

In the extended pressure-coupling system methods, first introduced by Anderson [Anderson 1980], an extra degree of freedom, corresponding to the volume of the box, is added to the system. The kinetic energy associated with this degree of freedom (which can be considered to be equivalent to a piston acting on the system) is $\frac{1}{2}Q(dV/dt)^2$, where Q is the ‘mass’ of the piston. The piston also has potential energy PV , where P is the desired pressure and V is the volume of the system. A piston of small mass gives rise to rapid oscillations in the box, whereas a large mass has the opposite effect. An infinite mass returns normal molecular dynamics behaviour. The volume can vary during the simulation, with the average volume being determined by the balance between the internal pressure of the system and the desired external pressure. The extended-system temperature-scaling method of Nosé uses a scaled time; in the extended pressure method the coordinates of the extended system are related to the ‘real’ coordinates by:

$$\mathbf{r}'_i = V^{-1/3} \mathbf{r}_i \quad (7.111)$$

7.8 Incorporating Solvent Effects into Molecular Dynamics: Potentials of Mean Force and Stochastic Dynamics

In many simulations of solute–solvent systems the primary focus is the behaviour of the solute; the solvent is of relatively little interest, particularly in regions far from the solute molecule. The use of non-rectangular periodic boundary conditions, stochastic boundaries and ‘solvent shells’ can all help to reduce the number of solvent molecules required and enable a larger proportion of the computing time to be spent simulating the solute. In this section we consider a group of techniques that incorporate the effects of solvent without requiring any explicit specific solvent molecules to be present.

One approach to this problem is to use a *potential of mean force* (PMF), which describes how the free energy changes as a particular coordinate (such as the separation of two atoms or the torsion angle of a bond) is varied. The free energy change described by the potential of mean force includes the averaged effects of the solvent.

Potentials of mean force may be determined using a molecular dynamics or Monte Carlo simulation using the techniques of umbrella sampling or free energy perturbation, which will be discussed in Chapter 11. Here we illustrate the concept using an example. The energy difference between the *trans* and *gauche* conformations for an isolated molecule of 1,2-dichloroethane (i.e. in the gas phase) is approximately 1.14 kcal mol⁻¹ with a population containing 77% *trans* and 23% *gauche* conformers. In liquid 1,2-dichloroethane, however, the relative population of the *gauche* conformer is significantly increased relative to the *trans* conformer by comparison with the isolated molecule, with 44% *trans* and 56% *gauche*. These experimental results were reproduced by Jorgensen (see Figure 7.13) using Monte Carlo simulations [Jorgensen *et al.* 1981]. The potential of mean force would be designed to reproduce this new population and so enable a single 1,2-dichloroethane molecule to be simulated as if it were present in the liquid.

A simulation performed using a potential of mean force enables the modulating effects of the solvent to be taken into account. The solvent also influences the dynamic behaviour of the

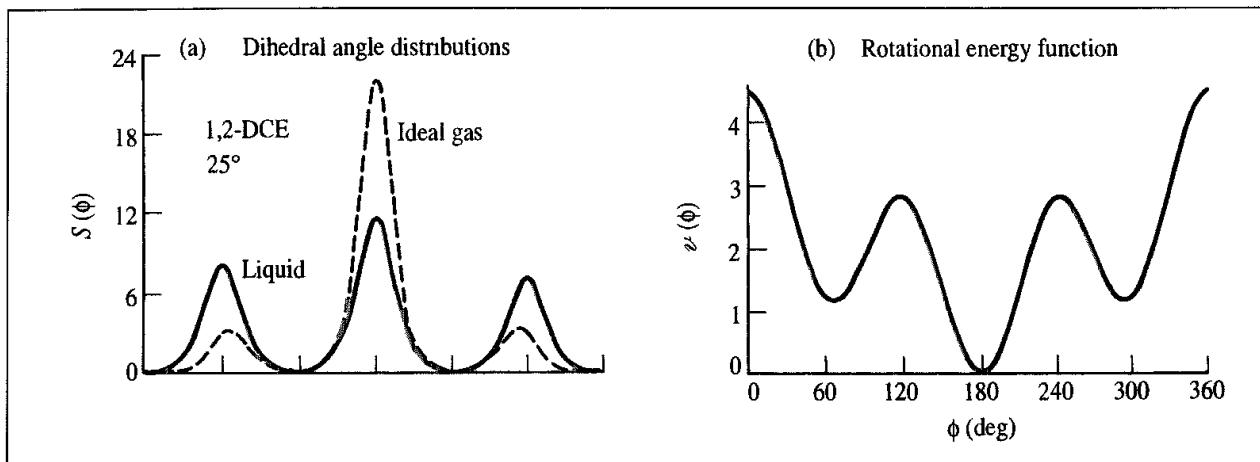


Fig 7.13: Population distribution for 1,2-dichloroethane in the gas and liquid phases (Figure redrawn from Jorgensen WL, R C Binning Jr and B Bigot 1981. Structures and Properties of Organic Liquids. n-Butane and 1,2-Dichloroethane and Their Conformational Equilibria. Journal of the American Chemical Society **103** 4393–4399)

solute via random collisions, and by imposing a frictional drag on the motion of the solute through the solvent. The Langevin equation of motion is the starting point for the *stochastic dynamics* models, which also incorporate these two effects. In stochastic dynamics the force on a particle is considered to arise from three sources. The first component is due to interactions between the particle and other particles. This force (\mathbf{F}_i) depends upon the position of the particle relative to the other particles and is modelled using a potential of mean force. The second force arises from the motion of the particle through the solvent and is equivalent to the frictional drag on the particle due to the solvent. This frictional force is proportional to the speed of the particle with the constant of proportionality being the friction coefficient:

$$\mathbf{F}_{\text{frictional}} = -\xi \mathbf{v} \quad (7.112)$$

where \mathbf{v} is the velocity and ξ is the friction coefficient. The friction coefficient is related to the collision frequency (γ) by $\gamma = \xi/m$ (m is the mass of the particle). γ^{-1} can be considered as the time taken for the particle to lose memory of its initial velocity (the velocity relaxation time). For a spherical particle the friction coefficient is related to the diffusion constant D by:

$$\xi = k_B T / D \quad (7.113)$$

If the radius of the spherical particle is a then the frictional force is given by Stokes' law:

$$\mathbf{F}_{\text{frictional}} = 6\pi a \eta \mathbf{v} \quad (7.114)$$

where η is the viscosity of the fluid.

The third contribution to the force on the particle is due to random fluctuations caused by interactions with solvent molecules. We will write this force as $\mathbf{R}(t)$. The Langevin equation of motion for a particle i can therefore be written*:

$$m_i \frac{d^2 \mathbf{x}_i(t)}{dt^2} = \mathbf{F}_i\{\mathbf{x}_i(t)\} - \gamma_i \frac{d\mathbf{x}_i(t)}{dt} m_i + \mathbf{R}_i(t) \quad (7.115)$$

* γ_i in Equation (7.115) is often referred to as the friction coefficient in the literature.

A number of simulation methods based on Equation (7.115) have been described. These differ in the assumptions that are made about the nature of frictional and random forces. A common simplifying assumption is that the collision frequency γ is independent of time and position. The random force $R(t)$ is often assumed to be uncorrelated with the particle velocities, positions and the forces acting on them, and to obey a Gaussian distribution with zero mean. The force F_i is assumed to be constant over the time step of the integration.

Three different situations can be considered, depending upon the relative magnitudes of the integration time step and the velocity relaxation time. The first category corresponds to timescales that are short relative to the velocity relaxation time ($\gamma\delta t \ll 1$). Under such circumstances the solvent does not activate or deactivate the particle to any significant extent. In the limit of zero γ (when there are no effects due to solvent) then the Langevin Equation (7.115) reduces to that obtained from Newton's laws of motion. At the other extreme the velocity relaxation time is much smaller than the time step. This corresponds to the diffusive regime, where the motion is rapidly damped by the solvent. The third situation is intermediate between these two extremes. Various methods have been proposed for integrating the Langevin equation of motion in these three regions.

In the region where $\gamma\delta t \ll 1$ the following is a simple integration algorithm [van Gunsteren *et al.* 1981]:

$$x_{i+1} = x_i + v_i \delta t + \frac{1}{2}(\delta t)^2 \{-\gamma v_i + m^{-1}(F_i + R_i)\} \quad (7.116)$$

$$v_{i+1} = v_i + (\delta t) \{-\gamma v_i + m^{-1}(F_i + R_i)\} \quad (7.117)$$

The average random force over the time step is taken from a Gaussian with a variance $2mk_B T \gamma(\delta t)^{-1}$. x_i is one of the $3N$ coordinates at time step i ; F_i and R_i are the relevant components of the frictional and random forces at that time; v_i is the velocity component.

An alternative expression is based on the following finite difference approximations [Brunger *et al.* 1984]:

$$d^2x/dt^2 \approx (x_{i+1} - 2x_i + x_{i-1})/\delta t^2 \quad (7.118)$$

$$dx/dt \approx (x_{i+1} - x_{i-1})/2\delta t \quad (7.119)$$

This leads to the following expressions for the coordinates x_{i+1} :

$$x_{i+1} = x_i + (x_i - x_{i-1}) \frac{1 - \frac{1}{2}\gamma\delta t}{1 + \frac{1}{2}\gamma\delta t} + \left(\frac{\delta t^2}{m} \right) \frac{F_i + R_i}{1 + \frac{1}{2}\gamma\delta t} \quad (7.120)$$

In the region where $\gamma\delta t \gg 1$ then if the interparticle force is assumed to be constant over the integration time step the following result is obtained [van Gunsteren *et al.* 1981]:

$$x_{i+1} = x_i + F_i(m\gamma)^{-1}\delta t + X_i(\delta t) \quad (7.121)$$

where X_i is a Gaussian distribution with zero mean and a variance of $2k_B T(m\gamma)^{-1} = 2D\delta t$. An extension of this treatment is to permit force F_i to vary linearly over the time step, giving:

$$x_{i+1} = x_i + \frac{\delta t}{m\gamma} (F_i + \frac{1}{2}\dot{F}_i\delta t) + X_i \quad (7.122)$$

\dot{F}_i is the derivative of the force at the time step i and is obtained numerically:

$$\dot{F}_i = (F_i - F_{i-1})/\delta t \quad (7.123)$$

In the intermediate region, where there are no restrictions on $\gamma\delta t$, then integration of the equations of motion gives the following rather complicated result [van Gunsteren and Berendsen 1982]:

$$\begin{aligned} x_{i+1} &= x_i + v_i \gamma^{-1} (1 - \exp(-\gamma\delta t)) + F_i (m\gamma)^{-1} [\delta t - \gamma^{-1} (1 - \exp(-\gamma\delta t))] \\ &\quad + (mg)^{-1} \int_{t_i}^{t_{i+1}} [1 - \exp(-\gamma(t_{i+1} - t'))] R(t') dt' \end{aligned} \quad (7.124)$$

$$\begin{aligned} v_{i+1} &= v_i \exp(-\gamma\delta t) + F_i (m\gamma)^{-1} (1 - \exp(-\gamma\delta t)) \\ &\quad + (m)^{-1} \int_{t_i}^{t_{i+1}} \exp(-\gamma(t_{i+1} - t')) R(t') dt' \end{aligned} \quad (7.125)$$

The important feature of these two equations is that the new positions and the new velocities both depend upon an integral over the random force, $R(t)$ (the final terms in Equations (7.124) and (7.125)). As both of these integrals depend upon $R_i(t)$ they are correlated. Specifically, they obey a *bivariate Gaussian* distribution. Such a distribution provides the probability that a particle located at x_i at time t with velocity v_i and experiencing a force F_i will be at x_{i+1} at time $t + \delta t$ with velocity v_{i+1} . In practice, this means that the distribution for the second variable depends upon the value selected for the first variable. It can be difficult to properly sample from such distributions, but van Gunsteren and Berendsen showed that the equations can be reformulated in terms of sampling from two independent Gaussian functions.

More complex stochastic dynamics treatments are possible; our treatment has only provided a rather simple treatment of solvent effects. For example, we have assumed that the frictional force at a given instant is proportional only to its velocity at the same time. A more realistic model assumes that the frictional forces are correlated; they have a ‘memory’ of previous values. The friction coefficient can also be made to depend on the coordinates of the other particles.

7.8.1 Practical Aspects of Stochastic Dynamics Simulations

A stochastic dynamics simulation requires a value to be assigned to the collision frequency friction coefficient γ . For simple particles such as spheres this can be related to the diffusion constant in the fluid. For the simulation of a rigid molecule it may be possible to derive γ via the diffusion coefficient from a standard molecular dynamics situation. In the more general case we require the friction coefficient of each atom. For simple molecules such as butane the friction coefficient can be considered to be the same for all atoms. The optimal value for γ can be determined by trial and error, performing a stochastic dynamics simulation for different values of γ and comparing the results with those from experiment (where available) or from standard molecular dynamics simulations. For large molecules the atomic friction coefficient is considered to depend upon the degree to which each atom is in contact with the solvent and is usually taken to be proportional to the accessible surface area of the atom (as defined in Section 1.5).

One of the main advantages of the stochastic dynamics methods is that dramatic time savings can be achieved, which enables much longer stimulations to be performed. For example, Widmalm and Pastor performed 1 ns molecular dynamics and stochastic dynamics simulations of an ethylene glycol molecule in aqueous solution of the solute and 259 water molecules [Widmalm and Pastor 1992]. The molecular dynamics simulation required 300 hours whereas the stochastic dynamics simulation of the solute alone required just 24 minutes. The dramatic reduction in time for the stochastic dynamics calculation is due not only to the very much smaller number of molecules present but also to the fact that longer time steps can often be used in stochastic dynamics simulations.

Stochastic dynamics has been widely used to study the behaviour of long-chain molecules and polymers. The advantages of stochastic dynamics are especially important for polymers [Helfand 1984], where many interesting phenomena occur over relatively long time periods, so putting them beyond the scope of conventional molecular dynamics. However, one must take care with the Langevin method when simulating systems in which specific solute-solvent interactions are present. For example, Yun-Yu, Lu and van Gunsteren used both stochastic dynamics and molecular dynamics to study the immunosuppressant drug cyclosporin (Figure 7.14) in two solvents: carbon tetrachloride and water [Yun-Yu *et al.* 1988]. The time-averaged structures obtained from each method were compared to determine the similarity between the average structure obtained for each simulation. Fluctuations in torsion angles were also compared. The analysis showed that the structures obtained from the molecular dynamics and stochastic dynamics simulations of cyclosporin in carbon tetrachloride were very similar, but that the results were very different for the Langevin and molecular dynamics simulations performed in water. This was due to an excessive degree of internal hydrogen bonding in the stochastic dynamics simulation; the equivalent

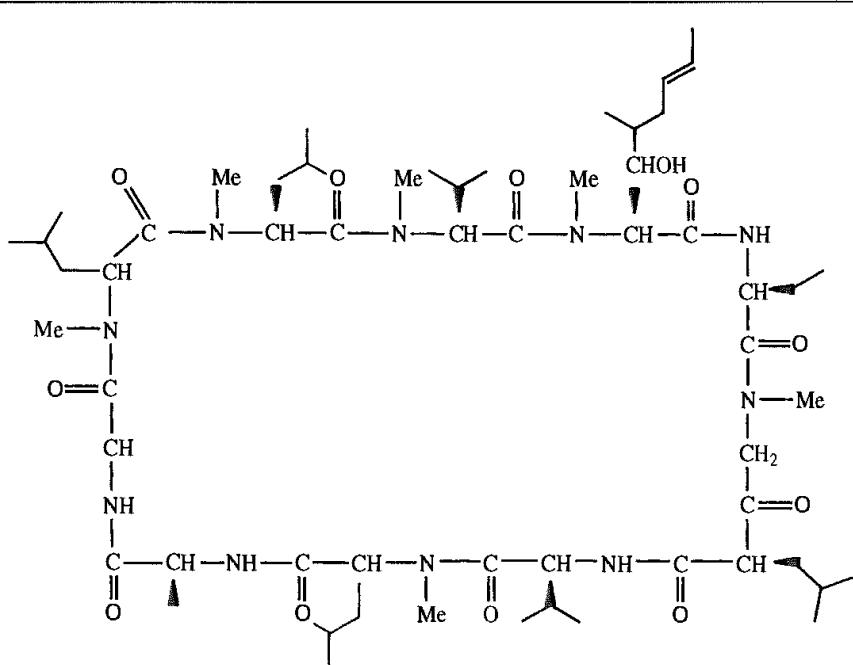


Fig 7.14: Cyclosporin

molecular dynamics simulation contained much more hydrogen bonding between cyclosporin and the solvent.

7.9 Conformational Changes from Molecular Dynamics Simulations

Molecular dynamics can provide information about the conformational properties of molecular systems and the way in which the conformation changes with time. Molecular graphics programs can facilitate the analysis of such simulations by displaying the structural parameters of interest in a manner that enables the time dimension to be taken into account. Perhaps the most direct way to demonstrate the conformational behaviour of the system is as a movie, where coordinate sets saved at regular intervals are displayed in sequence. For publication purposes, time-dependent data can be displayed graphically, with one of the axes corresponding to the time, such as the plots of energy or autocorrelation function versus time (Figures 7.3 and 7.10). The representation of bond rotations is difficult using x/y plots due to the 2π periodicity of a torsion angle. Lavery and Sklenar have developed a method to represent torsion data as a polar plot [Lavery and Sklenar 1988], where the distance from the origin corresponds to the time (Figure 7.15). Such 'dials' are very useful for detecting the presence of correlated conformational changes.

When viewing a movie of a molecular dynamics simulation of a complex molecule one is often struck by the chaotic nature of the motion. This should be expected; the motion of complex molecules *is* chaotic, but there are often underlying low-frequency motions which correspond to more significant and more interesting conformational changes. Fourier analysis techniques can be used to filter out the unwanted high-frequency motions, enabling the important low-frequency changes to be observed unhindered. Here we describe the filtering method of Dauber-Osguthorpe and Osguthorpe [Dauber-Osguthorpe and Osguthorpe 1990, 1993].

A Fourier transform enables one to convert the variation of some quantity as a function of time into a function of frequency, and vice versa. Thus, if we represent the quantity that varies in time as $x(t)$, then Fourier analysis enables us to also represent that quantity as a function $X(\nu)$, where ν is the frequency ($-\infty < \nu < \infty$). Fourier analysis is usually introduced by considering functions that vary in a periodic manner with time which can be written as a superposition of sine and cosine functions (a Fourier series; see Section 1.10.8). If the period of the function $x(t)$ is τ then the cosine and sine terms in the Fourier series are functions of frequencies $2\pi n/\tau$, where n can take integer values 1, 2, 3,

A Fourier series is rarely relevant to the interpretation of a molecular dynamics simulation as the movement of the atoms is not periodic but chaotic. The Fourier transform enables a non-periodic function to be converted into the equivalent frequency function (and vice versa). The Fourier transform can be developed from the Fourier series simply by considering the effect of increasing the period of a periodic function to infinity. The frequency function obtained from a Fourier transform is a continuous function rather than one written as a series of discrete frequencies. Further details are provided in Section 1.10.8.

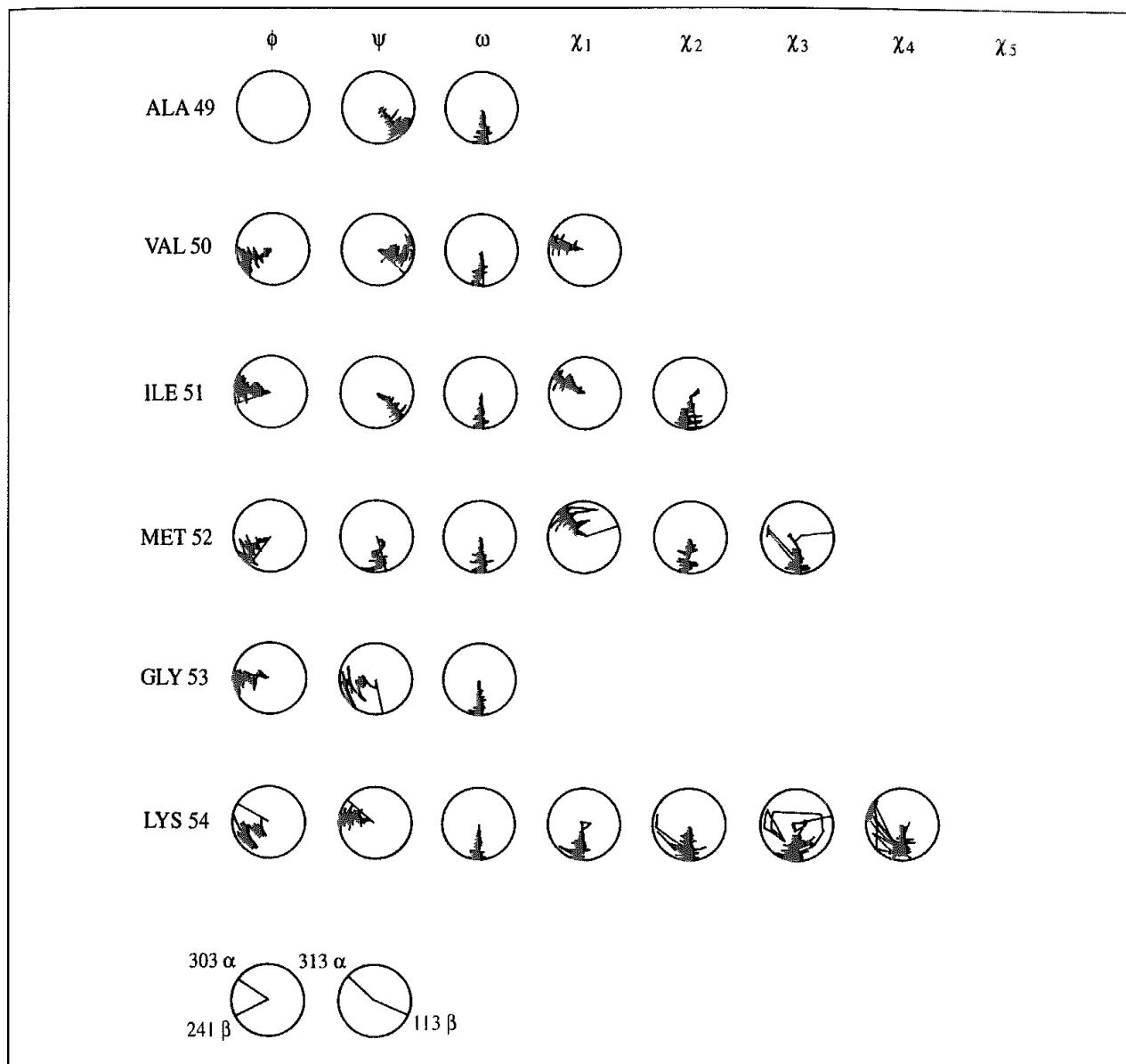


Fig 7.15 The variation in torsion angles can be effectively represented as a series of 'dials', where the time corresponds to the distance from the centre of the dial. Data from a molecular dynamics simulation of an intermolecular complex between the enzyme dihydrofolate reductase and a triazine inhibitor [Leach and Klein 1995]

At each step of the Fourier analysis of a molecular dynamics simulation the variation with time of one of the Cartesian coordinates of one of the atoms in the system is converted into the corresponding frequency function. Fast Fourier techniques are usually employed for this step. The frequency spectrum can then be filtered to remove high frequencies. This is achieved simply by setting the coefficients of the unwanted frequencies in the frequency function to zero. The resulting spectrum is then converted back to the time domain to give a new set of coordinate values at each of the time steps in the trajectory. This new coordinate set includes only the selected frequencies. This process can be repeated for the three coordinates of each atom to give a filtered trajectory for the entire system. It is also possible to select just a single frequency (i.e. a single normal mode) from the frequency spectrum and view this in isolation.

7.10 Molecular Dynamics Simulations of Chain Amphiphiles

The molecular dynamics technique is widely used for simulating large molecular systems, some of which have many degrees of conformational freedom. In this section we will examine the application of molecular dynamics to chain amphiphiles, a class of molecules of interest to both the 'biological' and 'materials science' communities. These molecules have a polar head group attached to one or more hydrocarbon chains. Some examples are shown in Figure 7.16. The head group has a high affinity for water, whereas the hydrocarbon tail prefers to exist in a hydrophobic environment. The molecules therefore exist in both phases at a water/oil interface. A characteristic feature of these molecules is their ability

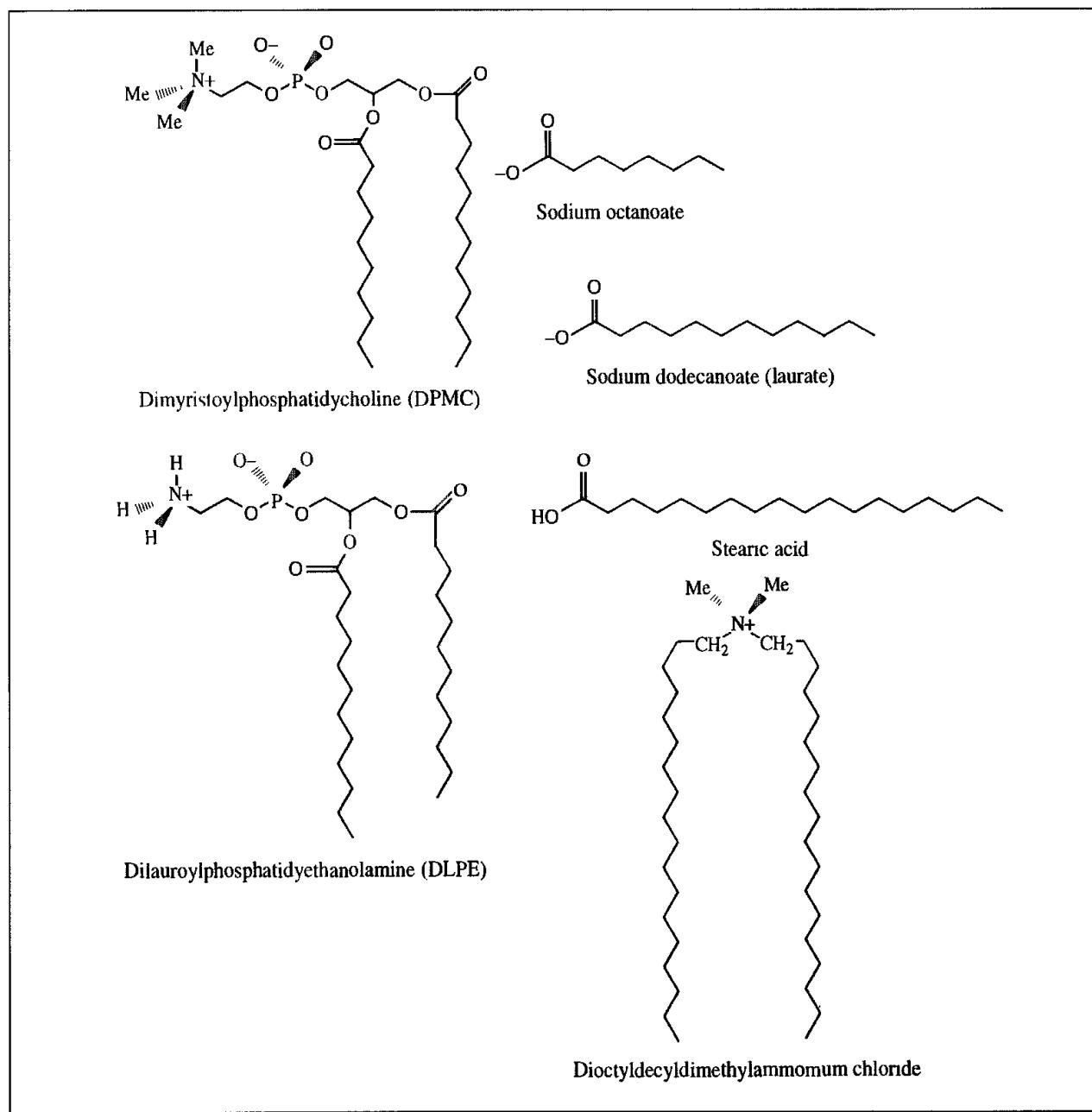


Fig. 7.16. Some typical amphiphiles

to form extended layer structures. Monolayers, bilayers and multiple layers are all possible. A monolayer at the water/air interface is known as a Langmuir film; when this is transferred to a solid substrate it is known as a Langmuir-Blodgett film. Langmuir-Blodgett films with many layers can be constructed in the laboratory but most simulation studies of these systems have been restricted to monolayers or bilayers. The ability to control the thickness of a Langmuir-Blodgett film and their high degree of order means that they are intensively investigated as insulators in semiconductors, filtration devices and as anti-reflective coatings. Amphiphiles are important in biology as cell membranes are formed from lipid bilayers. At a high enough concentration some amphiphiles can form micelles, which are globular structures that have the head groups all pointing into solution and the tails inside (Figure 7.17).

Amphiphiles often have a complex phase behaviour with several liquid crystalline phases. These liquid crystalline phases are often characterised by long-range order in one direction together with the formation of a layer structure. The molecules may nevertheless be able to move laterally within the layer and perpendicular to the surface of the layer. Structural information can be obtained using spectroscopic techniques including X-ray and neutron diffraction and NMR. The quadrupolar splitting in the deuterium NMR spectrum can be

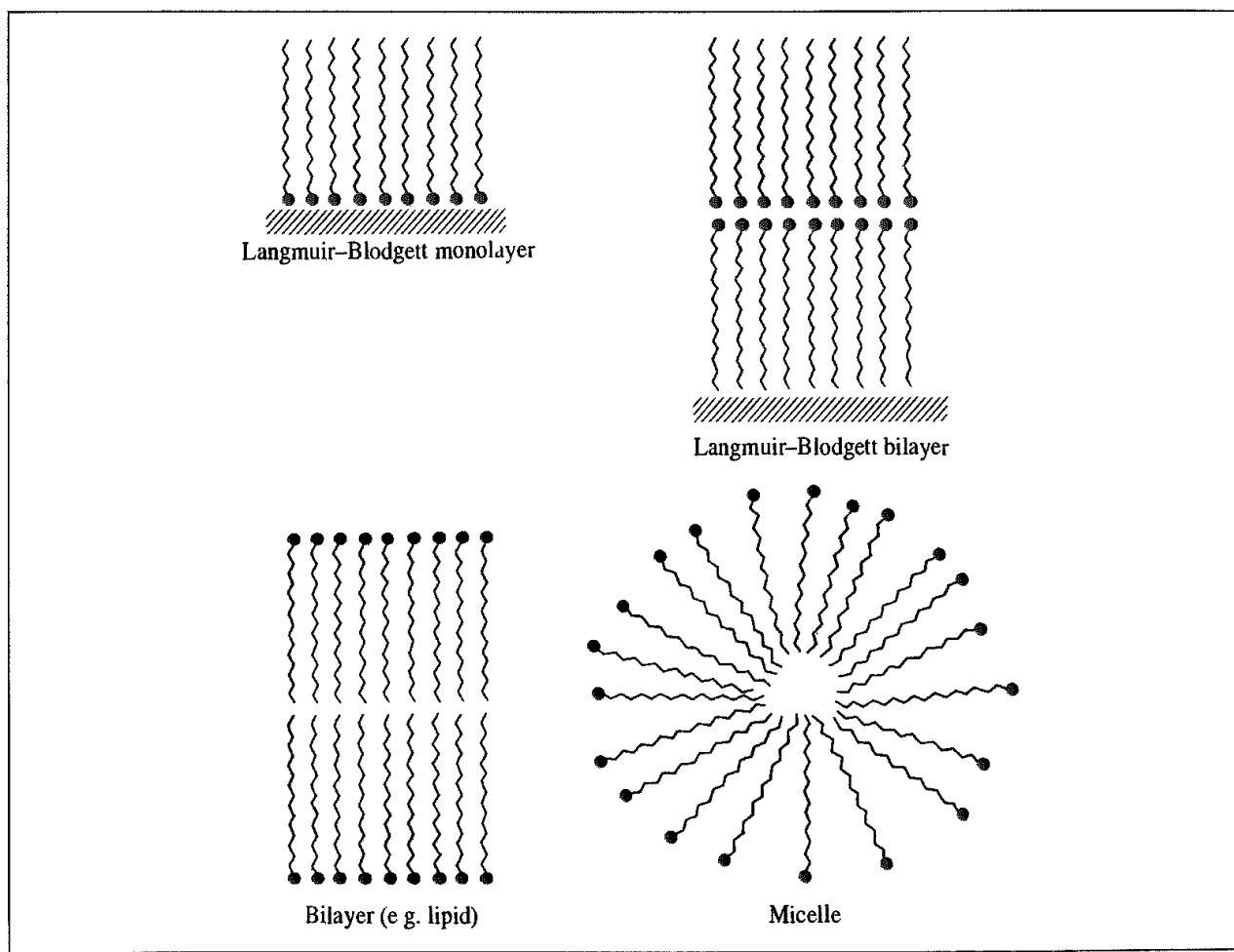


Fig 7.17 Some of the various phases that amphiphiles may form.

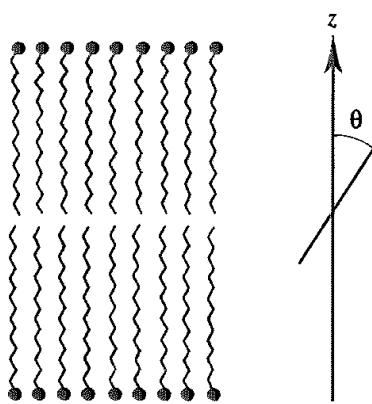


Fig 7.18 Definition of the order parameter

used to determine an *order parameter* for the carbon atoms on the hydrocarbon tail. The order parameter is defined as.

$$S = 0.5 \langle 3 \cos \theta_i \cos \theta_j - \delta_{ij} \rangle \quad (7.126)$$

θ_i is the angle between the i th molecular axis and the *director*, which is the average of the molecular axes over the sample. For a bilayer in the $L\alpha$ phase (the one present in cell membranes) the director is the same as the bilayer normal and is conventionally taken to be the z axis; see Figure 7.18. δ_{ij} is the Kronecker delta function ($\delta_{ij} = 1$ if $i = j$; $\delta_{ij} = 0$ if $i \neq j$). The expression for S is averaged over time and over molecules. The deuterium NMR experiment provides the order parameter S_{CD} , which indicates the average orientation of the C–D bond vector with respect to the bilayer normal. The experimental order parameters S_{CD} can range from 1.0 (indicating full order along the bilayer normal) to –0.5 (full order perpendicular to the bilayer normal) [Seelig and Seelig 1974]. A value of zero is considered to indicate full isotropic motion of the group. Experimental values are determined using molecules with deuterium-substituted methylene groups at positions along the hydrocarbon chain. Many simulations of amphiphiles are performed using united atom models for the hydrocarbon chains and it is therefore necessary to be able to relate the experimental order parameters to values that can be calculated from a simulation. This is done as follows [Essex *et al.* 1994]. Molecular axes are defined for each CH_2 unit in the chain as shown in Figure 7.19. These molecular axes are defined for the n th CH_2 unit as follows:

z : vector from C_{n-1} to C_{n+1}

y : vector perpendicular to z and in the plane through C_{n-1} , C_n and C_{n+1}

x : perpendicular to y and z

Using these definitions, components of the molecular order parameter tensor can be determined (for example, S_{zz} is determined by measuring the angle between the molecular z axis and the bilayer normal). The experimental order parameter can be related to the molecular order parameter using the equation:

$$S_{CD} = 2S_{xx}/3 + S_{yy}/3 \quad (7.127)$$

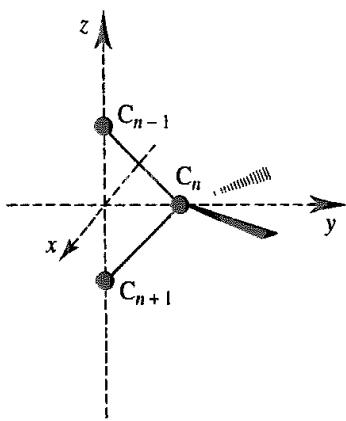


Fig 7.19. Calculation of the order parameter for united atom simulations.

With all-atom simulations the locations of the hydrogen atoms are known and so the order parameters can be calculated directly. Another structural property of interest is the ratio of *trans* conformations to *gauche* conformations for the CH₂–CH₂ bonds in the hydrocarbon tail. The *trans* : *gauche* ratio can be estimated using a variety of experimental techniques such as Raman, infrared and NMR spectroscopy.

7.10.1 Simulation of Lipids

There has been considerable interest in the simulation of lipid bilayers due to their biological importance. Early calculations on amphiphilic assemblies were limited by the computing power available, and so relatively simple models were employed. One of the most important of these is the mean field approach of Marcelja [Marcelja 1973, 1974], in which the interaction of a single hydrocarbon chain with its neighbours is represented by two additional contributions to the energy function. The energy of a chain in the mean field is given by:

$$\mathcal{V}_{\text{tot}} = \mathcal{V}_{\text{int}} + \mathcal{V}_{\text{disp}} + \mathcal{V}_{\text{rep}} \quad (7.128)$$

where \mathcal{V}_{int} is the internal energy of a chain, which can be calculated using standard force field methods. $\mathcal{V}_{\text{disp}}$ simulates the van der Waals interactions with the neighbouring molecules. It is often modelled using a Maier-Saupe potential:

$$\mathcal{V}_{\text{disp}} = -\Phi \sum_{i=1}^{\text{carbons}} \frac{1}{2} (3 \cos^2 \theta_i - 1) \quad (7.129)$$

The summation runs over all carbon atoms in the chain. θ_i is the angle between the bilayer normal and the molecular axis, as discussed above. Φ is the field strength; this may be parametrised to reproduce appropriate experimental data such as the deuterium NMR order parameters or it may be obtained by a self-consistent protocol, as described below. In his work on lipid bilayers Marcelja used a slightly different expression for $\mathcal{V}_{\text{disp}}$ which

involved the fraction of *trans* bonds in the system:

$$\mathcal{V}_{\text{disp}} = -\Phi \frac{n_{\text{trans}}}{n} \sum_{i=1}^{\text{carbons}} \frac{1}{2} (3 \cos^2 \theta_i - 1) \quad (7.130)$$

This additional factor was introduced to ensure the proper behaviour over both liquid crystalline and solid phases. In simulations of the liquid crystalline phase alone this term may be omitted for computational efficiency.

The repulsive contribution, \mathcal{V}_{rep} , is due to lateral pressure on each chain. In Marcelja's original treatment, this was set equal to the product of the lateral pressure, γ , and the cross-sectional area of the chain. The cross-sectional area was approximated by:

$$A = A_0 l_0 / l \quad (7.131)$$

where l_0 and A_0 are the length and cross-sectional area, respectively, of the hydrocarbon chain in a fully extended conformation. l is the length of the chain in the current conformation, projected onto the bilayer normal. If the bilayer normal is along the z axis then l is taken to be the z coordinate of the last carbon atom in the hydrocarbon chain. In other mean field models [Pastor *et al.* 1988] the product $\gamma A_0 / l_0$ is replaced with a single adjustable parameter Γ and so \mathcal{V}_{rep} is given by:

$$\mathcal{V}_{\text{rep}} = \sum_{\text{chains}} \frac{\Gamma}{(z_n - z_0)} \quad (7.132)$$

where z_n is the z coordinate of the last carbon in the chain and z_0 is the coordinate of the surface of the monolayer or bilayer. This force acts to keep the last carbon away from the surface; the closer it gets the larger the force pulling it away.

In his calculations, Marcelja generated all possible conformations of the hydrocarbon chain, restricting each carbon-carbon bond to the *trans* and *gauche* conformations. The energy of each conformation was evaluated. From the ensemble of conformations a partition function can be computed:

$$Z = \sum_{\text{all conformations}} \exp[-\mathcal{V}_{\text{tot}}/k_B T] \quad (7.133)$$

The molecular field is related to the partition function:

$$\Phi = \sum_{\text{all conformations}} \left\{ \frac{\frac{n_{\text{trans}}}{n} \sum_{i=1}^{\text{carbons}} \frac{1}{2} (3 \cos^2 \theta_i - 1) \exp[-\mathcal{V}_{\text{tot}}/k_B T]}{Z} \right\} \quad (7.134)$$

The molecular field is thus related to the partition function and so it is possible to generate a self-consistent value of the molecular field, Φ . Thermodynamic properties can then be calculated from the partition function. For example, Marcelja calculated the pressure as a function of the area per polar head group for surface monolayers at a variety of temperatures. His results showed good qualitative agreement with experimental results for such systems.

The mean field approach can be incorporated into a molecular dynamics simulation. It is particularly useful when used in conjunction with Langevin dynamics, as very long simulations can be performed. For example, Pearce and Harvey were able to perform simulations of three unsaturated phospholipids for 100 ns (i.e. 0.1 μ s) in single-molecule Langevin dynamics calculations [Pearce and Harvey 1993]. An extension of this strategy is to use a central 'core' containing one or more molecules that are simulated using molecular dynamics. This core is surrounded by a shell of molecules that are simulated using Langevin dynamics with the mean field. In this way one attempts to simulate a more 'realistic' system without incurring the computational penalty of a full molecular dynamics simulation of the entire system [De Loof *et al.* 1991].

The first molecular dynamics simulations of a lipid bilayer which used an explicit representation of all the molecules was performed by van der Ploeg and Berendsen in 1982 [van der Ploeg and Berendsen 1982]. Their simulation contained 32 decanoate molecules arranged in two layers of sixteen molecules each. Periodic boundary conditions were employed and a united atom force potential was used to model the interactions. The head groups were restrained using a harmonic potential of the form:

$$\nu(z) = \frac{k_h}{2}(z - \langle z \rangle)^2 \quad (7.135)$$

By writing the restraint in terms of the average z coordinates of the head groups ($\langle z \rangle$) van der Ploeg and Berendsen ensured that the bilayer was able to change its thickness to reach its equilibrium value. This restraining potential was designed to reproduce the interactions between the head groups and the water layer, neither of which was explicitly included in the calculation. A key feature of the simulation was the long equilibration time required. By explicitly representing all the molecules in the system it was possible to determine the collective motion of the system as a whole. One distinct feature was a slowly fluctuating collective tilt of the molecules away from the normal to the bilayer surface (Figure 7.20). The degree to which the molecules were aligned with each other was also correlated with the tilt angle. When the average tilt angle reached a maximum the chains were much more likely to be well aligned, but when the average tilt angle was close to zero (i.e. such that the average orientation of the chains was almost normal to the bilayer surface) much less order was observed. In their original simulations this collective tilt phenomenon was observed to extend over the entire simulation cell, suggesting that the cell dimensions were too small and that the use of periodic boundary conditions was enhancing the long-range correlations. Simulations using a larger system subsequently showed that this collective tilt could be observed for subsets of the molecules.

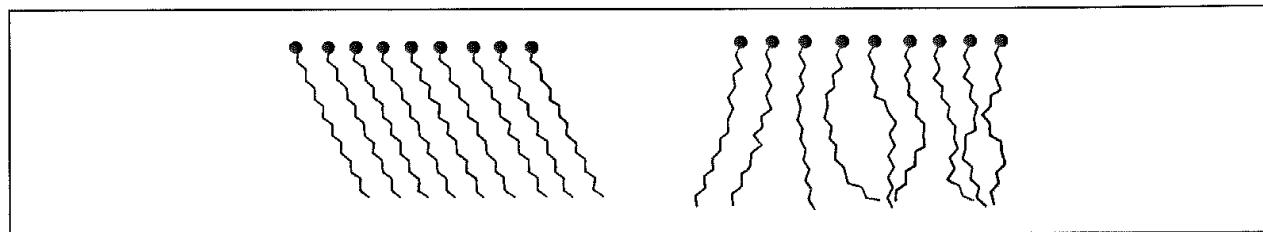


Fig. 7.20 Variation in alignment of chains in lipid simulation with tilt angle [van der Ploeg and Berendsen 1982]

Faster computers have enabled more realistic simulations of lipid bilayers to be performed, of larger systems, with more accurate models and for longer times [Stouch 1993; Tobias *et al.* 1997]. The trend is very much towards simulations that use full representations of all species present (i.e. ‘all-atom’ models with explicit solvent and counterions). The charged and highly polar nature of lipid head groups means that a proper representation of the long-range electrostatic forces can be critical, using a method such as the Ewald summation. Equilibration of such systems often requires hundreds of picoseconds and certain phenomena are only observed on a nanosecond timescale. In addition, molecules such as cholesterol and proteins may be included within the membrane. These simulations have revealed many hitherto unknown features of the behaviour of such systems. For example, considerable conformational mobility of the hydrocarbon chains is often observed in the liquid crystalline phases. This is illustrated in Figure 7.21 (colour plate section), which shows a snapshot of a lipid bilayer after a molecular dynamics simulation of several hundred picoseconds. The considerable degree of disorder in the hydrocarbon chains near the middle of the bilayer is clear from this figure and is very different to the idealised, ‘textbook’ pictures in which the chains are perfectly aligned in completely extended conformations. The distribution of *gauche* conformations tends to be higher towards the end of the chain, though in some systems a *gauche* link is required near the head group to enable the chain to lie perpendicular to the interface. ‘Kinks’ are often observed in the chains; these are arrangements of three successive bonds with *gauche*(+)-*trans*-*gauche*(-) torsion angle, which enable the chain to remain perpendicular to the surface.

7.10.2 Simulations of Langmuir–Blodgett films

The simulations of Langmuir–Blodgett systems can be difficult due to the need to correctly model the solid support. To illustrate the procedure we will describe the calculations of Kim, Moller, Tildesley and Quirke [Kim *et al.* 1994a] who simulated stearic acid ($\text{CH}_3(\text{CH}_2)_{16}\text{COOH}$) adsorbed onto graphite. The surface was modelled using a Lennard-Jones 9–3 potential that depends upon the height of the atom (α) above the surface (z_α):

$$\nu_{\alpha s}(z_\alpha) = \frac{2\pi\rho}{3}\varepsilon_{ss} \left[\frac{2}{15} \left(\frac{\sigma_{\alpha s}}{z_\alpha} \right)^9 - \left(\frac{\sigma_{\alpha s}}{z_\alpha} \right)^3 \right] \quad (7.136)$$

where ρ is the density of the solid and ε_{xx} and δ_{ss} are its Lennard-Jones parameters. An image-change method was also applied to the acid head group with the interaction between a charge and its image being:

$$\nu_{ic}(z) = \frac{1}{2} \frac{(\varepsilon - \varepsilon')}{(\varepsilon + \varepsilon')} \left[\frac{q_\alpha^2}{8\pi\varepsilon_0(z - z_{ip})} \right] \quad (7.137)$$

where ε' is the relative permittivity of the solid (taken to be 4.0) and ε is the permittivity above the surface ($\varepsilon = 1.0$). The image plane is located at $z_{ip} = \sigma_{ss}/2$. Each charge interacts with its own image and with the images of other charges, but there are no interactions between the image charges themselves. The hydrocarbon chain of the stearic acid was modelled using an all-atom model, with explicit hydrogen atoms.

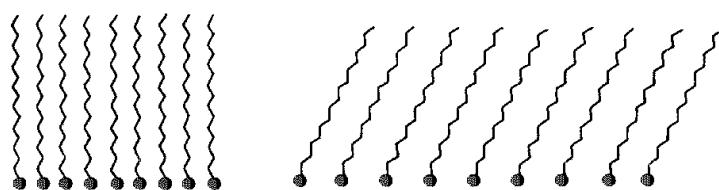


Fig 7.22 Simulations of a Langmuir-Blodgett film [Kim et al 1994a] as the area per head group increases the chains tilt away from the normal

A molecular dynamics simulation of 64 molecules with periodic boundary conditions confirmed the presence of a transition in which the collective tilt of the chains changed from being upright (i.e. perpendicular to the surface) to having an angle of around 20° (Figure 7.22). This transition was induced by increasing the area per head group. The proportion of molecules in the all-*trans* conformation decreased significantly as the head group area was increased (97.7% of molecules were fully extended for a head group area of 20.6 \AA^2 but only 66.9% for an area of 21.2 \AA^2). The bond linking the chain to the acid head growth showed a considerable degree of rotational disorder.

Bilayers of stearic acid were also simulated on a hydrophobic surface [Kim et al. 1994b]. In the bilayer the molecules are arranged head to head, with the hydrocarbon tail on the surface. In this arrangement hydrogen bonds form between the head groups (Figure 7.23). The bilayer also showed the tilt angle transition that was observed for the monolayer, though the degree of tilt was considerably less for the bilayer, suggesting that hydrogen bonding between the head groups was important in controlling the orientation of the molecules.

An extension of these calculations to cationic dialkylamide salts required an even more complex model [Adolf et al. 1995]. These molecules have the general formula $(\text{CH}_3)_2\text{N}^+[(\text{CH}_2)_{n-1}\text{CH}_3][(\text{CH}_2)_{m-1}\text{CH}_3]\text{Cl}^-$ and the isomer with $m = n = 18$ is one of the main active ingredients in commercial fabric softeners. The presence of two long alkyl

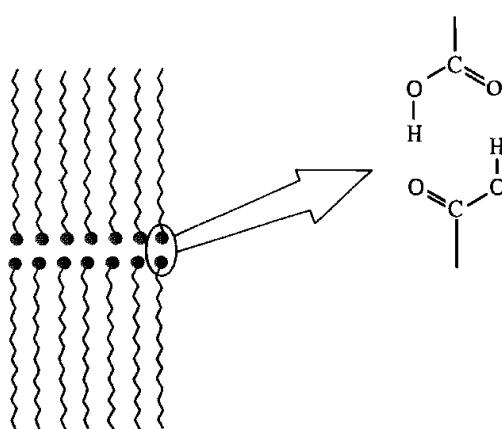


Fig 7.23 In simulations of stearic acid on a hydrophobic surface hydrogen bonding between the head groups is important in controlling the orientation of the molecules [Kim et al 1994b]

chains and an ionic head group means that these molecules are also structurally similar to phospholipids. A modified Ewald method was used to calculate electrostatic interactions in the two dimensions parallel to the surface, and the anisotropic potential model of Toxvaerd (see Section 4.15) was employed to retain the computational savings of a united-atom model. This system also showed a variation in the tilt with head group area, though the results at the highest head group densities were not as ‘solid-like’ as was suggested by the experimental data. Nevertheless, there were some areas where the model could be improved, including the need to incorporate water molecules and use a more appropriate representation of the chloride anion.

7.10.3 Mesoscale Modelling: Dissipative Particle Dynamics

The molecular dynamics methods that we have discussed in this chapter, and the examples that have been used to illustrate them, fall into the category of ‘atomistic simulations’, in that all of the actual atoms (or at least the non-hydrogen atoms) in the core system are represented explicitly. Atomistic simulations can provide very detailed information about the behaviour of the system, but as we have discussed this typically limits a simulation to the nanosecond timescale. Many processes of interest occur over a longer timescale. In the case of processes which occur on a ‘macroscopic’ timescale (i.e. of the order of seconds) then rather simple models may often be applicable. Between these two extremes are phenomena that occur on an intermediate scale (of the order of microseconds). This is the realm of the *mesoscale*. Dissipative particle dynamics (DPD) is particularly useful in this region, examples include complex fluids such as surfactants and polymer melts.

These three general regions (atomistic, mesoscopic and macroscopic) are not only characterised by different timescales but also varying length scales. Indeed, there is a general inverse relationship between the time and the length. In the case of the dissipative particle dynamics method the fast motion of the atoms is integrated out, leaving as the fundamental ‘unit’ a set of beads that interact with other beads via an appropriate potential [Koelman and Hoogerbrugge 1993]. Each bead represents a small ‘droplet’ of the fluid. The total force on each bead is due to a combination of direct interactions with other beads together with random and dissipative forces. The trajectory of the system is calculated by integrating Newton’s laws of motion in the usual way, from which properties can be derived.

The underlying model in dissipative particle dynamics is usually developed in such a way that the mass, length and timescales are all unity. This is similar to the use of reduced units for the Lennard-Jones potential (Section 4.10.5). A particular advantage of such an approach is that a single simulation may often be able to explain the behaviour of many different systems. With a mass of 1 the force acting on a particle is equal to its acceleration. In DPD there are three forces on each bead [Groot and Warren 1997]:

$$\mathbf{f}_i = \sum_{j=1; j \neq i}^N (\mathbf{F}_{ij}^C + \mathbf{F}_{ij}^D + \mathbf{F}_{ij}^R) \quad (7.138)$$

The summation is over all other particles j which are within a certain cutoff radius r_c of i . This cutoff radius becomes the unit of length in the subsequent treatment (i.e. $r_c = 1$). The

first of these forces, the conservative force, \mathbf{F}_{ij}^C , is a soft repulsion that acts along the line joining i to j .

$$\mathbf{F}_{ij}^C = \begin{cases} a_{ij}(1 - r_{ij})\hat{\mathbf{r}}_{ij} & r_{ij} < 1 \\ 0 & r_{ij} > 1 \end{cases} \quad (7.139)$$

r_{ij} is the distance between beads i and j and $\hat{\mathbf{r}}_{ij}$ is the corresponding unit vector. The second force is a dissipative (or drag) force, which is given by:

$$\mathbf{F}_{ij}^D = \begin{cases} -\gamma w^D(r_{ij})(\hat{\mathbf{r}}_{ij} \cdot \mathbf{v}_{ij})\hat{\mathbf{r}}_{ij} & r_{ij} < 1 \\ 0 & r_{ij} > 1 \end{cases} \quad (7.140)$$

This dissipative force is proportional to the relative velocity of the two beads and acts so as to reduce their relative momentum. \mathbf{v}_{ij} is the difference between the two velocities ($\mathbf{v}_{ij} = \mathbf{v}_i - \mathbf{v}_j$) and $w^D(r_{ij})$ is a weight function that depends upon the distance r_{ij} and disappears for inter-bead distances greater than unity (i.e. r_c).

The third and final force acting between any pair of beads is a random force:

$$\mathbf{F}_{ij}^R = \begin{cases} \sigma w^R(r_{ij})\theta_{ij}\hat{\mathbf{r}}_{ij} & r_{ij} < 1 \\ 0 & r_{ij} > 1 \end{cases} \quad (7.141)$$

$w^R(r_{ij})$ is a distance-dependent weight function similar to that for the dissipative force. θ_{ij} is a function which ensures that the random force between each pair of particles averages to zero over time and is independent of the force between every other pair of particles. The random force can be more usefully expressed in terms of the timestep in the integration scheme:

$$\mathbf{F}_{ij}^R = \frac{\sigma w^R(r_{ij})\zeta_{ij}\hat{\mathbf{r}}_{ij}}{\sqrt{\delta t}} \quad (7.142)$$

ζ_{ij} is a random number with zero mean and unit variance, chosen independently for each pair of particles and at each time step in the integration.

Both the dissipative force and the random force act along the line joining the pair of beads and also conserve linear and angular momentum. The model thus has two unknown functions $w^D(r_{ij})$ and $w^R(r_{ij})$ and two unknown constants γ and σ . In fact, only one of the two weight functions can be chosen arbitrarily as they are related [Espanol and Warren 1995]. Moreover, the temperature of the system relates the two constants:

$$w^D(r) = [w^R(r)]^2 \quad (7.143)$$

$$\sigma^2 = 2\gamma k_B T \quad (7.144)$$

The usual choice for the weight functions is to make the random force the same as the conservative force:

$$w^D(r) = [w^R(r)]^2 = \begin{cases} (1 - r)^2 & r < 1 \\ 0 & r > 1 \end{cases} \quad (7.145)$$

The equations of motion are integrated using a modified velocity Verlet algorithm. The modification is required because the force depends upon the velocity, the extra step involves

a prediction followed by a correction. If, in addition to the use of units of mass and length, we assume that $k_B T$ is equal to 1, then the unit of time is:

$$\tau = r_c \sqrt{m/k_B T} \quad (7.146)$$

It remains to assign values to the noise amplitude, σ , the time step for the integration, δt , and the repulsion parameter, a_{ij} . The effects of the first two of these upon the stability of the simulation are also related to the integration method. Groot and Warren determined that when the noise amplitude was larger than 8 the integration scheme became unstable and that a value of 3 gave good results over a range of temperatures [Groot and Warren 1997]. The integration time step with the modified Verlet algorithm should have a value between 0.04 and 0.06; any larger and the temperature would artificially increase by an unacceptable amount. The repulsion parameter is the key determinant of the interactions between the beads. This can be achieved by relating the DPD model to bulk properties. For example, to model the compressibility of water at room temperature the repulsion parameter is related to the density, ρ , by:

$$a_{ii}\rho = 75k_B T \quad (7.147)$$

These interaction parameters between particles of the same type can be used to derive the values of a_{ij} between unlike beads. For polymers which involve beads of different types the repulsion between unlike beads is made larger than between like beads.

A good example of the use of DPD is the study by Groot and Madden of the microphase separation of diblock copolymer melts [Groot and Madden 1998; Groot *et al.* 1999]. Block copolymers are surfactants which are present in many consumer products such as foods (e.g. ice cream and margarine), detergents and personal care products (e.g. shampoo). The properties of these materials are strongly dependent upon their bulk organisation (or *morphology*), which in turn depends upon the relative sizes of the head and the tail groups and how they interact. The diblock copolymers of interest can be represented by the general formula $A_m B_n$ where A and B represent amalgamations of the smaller building blocks from which the polymer is constructed. Of particular interest was the way in which the behaviour of the system varied as the ratio of A to B was changed, for a fixed polymer length. In this particular case the length was fixed at 10 beads and the entire simulation contained a total of 40 000 particles. A variety of systems were investigated, such as A_2B_8 , A_3B_7 and A_5B_5 . The beads in each polymer chain were kept together by adding an extra term to the force (Equation (7.139)) of the form $C r_{ij}$ if i is connected to j .

Due to the greater degree of repulsion between unlike beads the final configuration of the system contains domains which are rich in either the A or the B type of bead. Some regions are rich in A; others are rich in B. The organisation of the A-rich and B-rich domains can be visualised by plotting a three-dimensional contour that connects regions where the density is intermediate between purely A and purely B.

For the 1:1 polymer (A_5B_5) a lamellar phase was obtained, in which the A- and B-rich domains form parallel planes of alternating A and B beads (Figure 7.24(a), colour plate section). For other configurations, however, different structures were observed. The A_3B_7 system evolved to a hexagonal phase (Figure 7.24(b)) and the A_2B_8 structure produces a set of peanut-shaped micelles (Figure 7.24(c))

Appendix 7.1 Energy Conservation in Molecular Dynamics

The total energy is the sum of the kinetic $\mathcal{K}(t)$ and potential energies $\mathcal{V}(t)$:

$$E(t) = \mathcal{K}(t) + \mathcal{V}(t) \quad (7.148)$$

We want to derive an expression for the rate of change of the energy with time, dE/dt . First, we differentiate the kinetic energy term with respect to time:

$$\frac{d\mathcal{K}}{dt} = \sum_{i=1}^N \frac{d}{dt} \left(\frac{1}{2} m_i v_i^2 \right) = \sum_{i=1}^N m_i v_i \frac{dv_i}{dt} \quad (7.149)$$

As $m_i dv_i/dt$ is equal to the force on the atom i , the result can be written:

$$\frac{d\mathcal{K}}{dt} = \sum_{i=1}^N v_i f_i \quad (7.150)$$

f_i is the force on atom i .

The potential energy is written as a series of pairwise interaction terms:

$$\mathcal{V}(t) = \sum_{i=1}^N \sum_{j=i+1}^N \nu(r_{ij}(t)) \quad (7.151)$$

The derivative of the potential energy with respect to time can be written:

$$\frac{d\mathcal{V}}{dt} = \sum_{i=1}^N \sum_{j=i+1}^N \frac{\partial \nu}{\partial r_{ij}} \frac{dr_{ij}}{dt} \quad (7.152)$$

$\partial \nu / \partial r_{ij}$ equals 1 for each pairwise combination i and j . Each term $\nu(r_{ij})$ is a function of the positions of atom i and j (\mathbf{r}_i and \mathbf{r}_j) and we can then write:

$$\frac{d\nu(r_{ij})}{dt} = \frac{d\nu(r_{ij})}{d\mathbf{r}_i} \frac{d\mathbf{r}_i}{dt} + \frac{d\nu(r_{ij})}{d\mathbf{r}_j} \frac{d\mathbf{r}_j}{dt} \quad (7.153)$$

For a given atom i , there will be a total of $N - 1$ terms of the form $\nu(r_{ij})$ in the expression for the potential energy due to the interactions between i and all other atoms j . Hence we can write $d\mathcal{V}/dt$ as follows:

$$\frac{d\mathcal{V}}{dt} = \sum_{i=1}^N \sum_{j=1, j \neq i}^N \frac{\partial \nu(r_{ij})}{\partial \mathbf{r}_i} \frac{d\mathbf{r}_i}{dt} = \sum_{i=1}^N \frac{d\mathbf{r}_i}{dt} \sum_{j=1, j \neq i}^N \frac{\partial \nu(r_{ij})}{\partial \mathbf{r}_i} \quad (7.154)$$

The force on atom i due to its interaction with atom j equals minus the gradient with respect to \mathbf{r}_i , or $-d\nu(r_{ij})/d\mathbf{r}_i$. Thus the total force on the atom is equal to

$$- \sum_{j=1; j \neq i}^N \frac{\partial \nu(r_{ij})}{\partial \mathbf{r}_i} \quad (7.155)$$

and so we have:

$$\frac{d\mathcal{V}}{dt} = - \sum_{i=1}^N \frac{d\mathbf{r}_i}{dt} f_i = - \sum_{i=1}^N v_i f_i \quad (7.156)$$

Thus $(d\mathcal{V}/dt) + (d\mathcal{K}/dt) = dE/dt = 0$, which implies that the energy is constant. In practice, the total energy fluctuates about a constant value.

Further Reading

- Allen M P and D J Tildesley 1987 *Computer Simulation of Liquids*. Oxford, Oxford University Press.
- Berendsen H C and W F van Gunsteren 1984. Molecular Dynamics Simulations Techniques and Approaches. In Barnes A J, W J Orville-Thomas and J Yarwood (Editors) *Molecular Liquids, Dynamics and Interactions*. NATO ASI Series C135, New York, Reidel, pp. 475–600
- Berendsen H C and W F van Gunsteren 1986. Practical Algorithms for Dynamic Simulations. Molecular Dynamics Simulation of Statistical Mechanical Systems. *Proceedings of the Enrico Fermi Summer School Varenna Soc Italiana di Fisica* Bologna, pp 43–65
- Brooks C L III, M Karplus and B M Pettitt 1988 Proteins. A Theoretical Perspective of Dynamics, Structure and Thermodynamics. *Advances in Chemical Physics* Volume LXXI. New York, John Wiley & Sons
- Goldstein H 1980. *Classical Mechanics* (2nd Edition). Reading, MA, Addison-Wesley.
- Haile J M 1992. *Molecular Dynamics Simulation Elementary Methods* New York, John Wiley & Sons
- McCammon J A and S C Harvey 1987. *Dynamics of Proteins and Nucleic Acids*. Cambridge, Cambridge University Press
- van Gunsteren W F 1994 Molecular Dynamics and Stochastic Dynamics Simulations: A Primer. In van Gunsteren W F, P K Weiner and A J Wilkinson (Editors) *Computer Simulations of Biomolecular Systems* Volume 2 Leiden, ESCOM
- van Gunsteren W F and H J C Berendsen 1990 Computer Simulation of Molecular Dynamics: Methodology, Applications and Perspectives in Chemistry *Angewandte Chemie International Edition in English* **29**:992–1023

References

- Adolf D B, D J Tildesley, M R S Pinches, J B Kingdon, T Madden and A Clark 1995. Molecular Dynamics Simulations of Dioctadecyldimethylammonium Chloride Monolayers. *Langmuir* **11**:237–246.
- Alder B J and T E Wainwright 1957 Phase Transition for a Hard-sphere System. *Journal of Chemical Physics* **27**:1208–1209
- Alder B J and T E Wainwright 1970. Decay of the Velocity Autocorrelation Function *Physical Review A* **1**:18–21.
- Allen M P and D J Tildesley 1987 *Computer Simulation of Liquids*. Oxford, Oxford University Press
- Anderson H C 1980. Molecular Dynamics Simulations at Constant Pressure and/or Temperature. *Journal of Chemical Physics* **72**:2384–2393.
- Anderson H C 1983. Rattle A ‘Velocity’ Version of the Shake Algorithm for Molecular Dynamics Calculations. *Journal of Computational Physics* **54**:24–34.
- Beeman D 1976 Some Multistep Methods for Use in Molecular Dynamics Calculations. *Journal of Computational Physics* **20** 130–139.

- Berendsen H J C, J P M Postma, W F van Gunsteren, A Di Nola and J R Haak 1984 Molecular Dynamics with Coupling to an External Bath. *Journal of Chemical Physics* **81**:3684-3690
- Brunger A, C B Brooks and M Karplus 1984 Stochastic Boundary Conditions for Molecular Dynamics Simulations of ST2 Water. *Chemical Physics Letters* **105**:495-500.
- Dauber-Osguthorpe P and D J Osguthorpe 1990 Analysis of Intramolecular Motions by Filtering Molecular Dynamics Trajectories. *Journal of the American Chemical Society* **112**:7921-7935
- Dauber-Osguthorpe P and D J Osguthorpe 1993 Partitioning the Motion in Molecular Dynamics Simulations into Characteristic Modes of Motion. *Journal of Computational Chemistry* **14** 1259-1271.
- De Loof H, S C Harvey, J P Segrest and R W Pastor 1991. Mean Field Stochastic Boundary Molecular Dynamics Simulation of a Phospholipid in a Membrane. *Biochemistry* **30** 2099-2113
- Espanol P and P B Warren 1995 Statistical Mechanics of Dissipative Particle Dynamics. *Europhysics Letters* **30**,191-196.
- Essex J W, M M Hann and W G Richards 1994. Molecular Dynamics of a Hydrated Phospholipid Bilayer. *Philosophical Transactions of the Royal Society of London* **B344**,239-260
- Fincham D and Heyes D M 1982 Integration Algorithms in Molecular Dynamics. *CCP5 Quarterly* **6**.4-10
- Gear C W 1971 *Numerical Initial Value Problems in Ordinary Differential Equations* Englewood Cliffs, NJ, Prentice Hall
- Groot R D and T J Madden 1998. Dynamic Simulation of Diblock Copolymer Microphase Separation. *Journal of Chemical Physics* **108**:8713-8724
- Groot R D, T J Madden and D J Tildesley 1999 On the Role of Hydrodynamic Interactions in Block Copolymer Microphase Separation. *Journal of Chemical Physics* **110** 9739-9749.
- Groot R D and P B Warren 1997. Dissipative Particle Dynamics: Bridging the Gap Between Atomistic and Mesoscopic Simulation. *Journal of Chemical Physics* **107** 4423-4435.
- Helfand E 1984 Dynamics of Conformational Transitions in Polymers. *Science* **226**:647-650.
- Hockney R W 1970. The Potential Calculation and Some Applications. *Methods in Computational Physics* **9** 136-211
- Hoover W G 1985. Canonical Dynamics. Equilibrium Phase-space Distributions. *Physical Review A* **31**:1695-1697.
- Humphreys D D, R A Friesner and B J Berne 1994. A Multiple Time-step Molecular Dynamics Algorithm for Macromolecules. *Journal of Physical Chemistry* **98**:6885-6892.
- Humphreys D D, R A Friesner and B J Berne 1995. Simulated Annealing of a Protein in a Continuum Solvent by Multiple Time-step Molecular Dynamics. *Journal of Physical Chemistry* **99**:10674-10685.
- Humphreys, D D, R A Friesner and B J Berne 1996 A Multiple Time-step Molecular Dynamics Algorithm for Macromolecules. *Journal of Physical Chemistry* **98**:6885-6892.
- Jorgensen W L, R C Binning Jr and B Bigot 1981 Structures and Properties of Organic Liquids: *n*-Butane and 1,2-Dichloroethane and Their Conformational Equilibria. *Journal of the American Chemical Society* **103**,4393-4399.
- Kim K S, M A Moller, D J Tildesley and N Quirke 1994a. Molecular Dynamics Simulations of Langmuir-Blodgett Monolayers with Explicit Head-group Interactions. *Molecular Simulation* **13**:77-99.
- Kim K S, D J Tildesley and N Quirke 1994b. Molecular Dynamics of Langmuir-Blodgett Films: II. Bilayers. *Molecular Simulation* **13**,101-114
- Koelman J M V A and P J Hoogerbrugge 1993 Dynamic Simulations of Hard-sphere Suspensions Under Steady Shear. *Europhysics Letters* **21** 363-368.
- Lavery R and H Sklenar 1988 The Definition of Generalized Helicoidal Parameters and of Axis Curvature for Irregular Nucleic Acids. *Journal of Biomolecular Structure and Dynamics* **6**,63-91
- Leach A R and T E Klein 1995 A Molecular Dynamics Study of the Inhibitors of Dihydrofolate Reductase by a Phenyl Triazine. *Journal of Computational Chemistry* **16**:1378-1393

- Marcelja S 1973. Molecular Model for Phase Transition in Biological Membranes *Nature* **241**:451–453
- Marcelja S 1974. Chain Ordering in Liquid Crystals. II. Structure of Bilayer Membranes. *Biochimica et Biophysica Acta* **367**:165–176
- Nosé S 1984. A Molecular Dynamics Method for Simulations in the Canonical Ensemble. *Molecular Physics* **53**:255–268.
- Pastor R W, R M Venable and M Karplus 1988. Brownian Dynamics Simulation of a Lipid Chain in a Membrane Bilayer. *Journal of Chemical Physics* **89**:1112–1127.
- Pearce L L and S C Harvey 1993. Langevin Dynamics Studies of Unsaturated Phospholipids in a Membrane Environment. *Biophysical Journal* **65**:1084–1092
- Procacci P and B Berne 1994. Computer Simulation of Solid C₆₀ Using Multiple Time-step Algorithms. *Journal of Chemical Physics* **101**:2421–2431.
- Rahman A 1964. Correlations in the Motion of Atoms in Liquid Argon. *Physical Review A* **136**:405–411.
- Rahman A and F H Stillinger 1971. Molecular Dynamics Study of Liquid Water. *Journal of Chemical Physics* **55**:3336–3359
- Robinson A J, W G Richards, P J Thomas and M M Hann 1994. Head Group and Chain Behaviour in Biological Membranes—A Molecular Dynamics Simulation. *Biophysical Journal* **67**:2345–2354.
- Rubinstein R Y 1981. *Simulation and Monte Carlo Methods*. New York, John Wiley & Sons.
- Ryckaert J P, G Cicotti and H J C Berendsen 1977. Numerical Integration of the Cartesian Equations of Motion of a System with Constraints. Molecular Dynamics of n-Alkanes. *Journal of Computational Physics* **23**:327–341
- Seelig A and J Seelig 1974. The Dynamics Structure of Fatty Acyl Chains in a Phospholipid Bilayer Measured by Deuterium Magnetic Resonance. *Biochemistry* **13**:4839–4845
- Stouch T R 1993. Lipid Membrane Structure and Dynamics Studied by All-atom Molecular Dynamics Simulations of Hydrated Phospholipid Bilayers. *Molecular Simulation* **10**:335–362.
- Streett W B, D Tildesley and G Saville 1978. Multiple Time-step Methods in Molecular Dynamics. *Molecular Physics* **35**:639–648.
- Swindoll R D and J M Haile 1984. A Multiple Time-step Method for Molecular Dynamics Simulations of Fluids of Chain Molecules. *Journal of Computational Physics* **53**:289–298
- Swope W C, H C Anderson, P H Berens and K R Wilson 1982. A Computer Simulation Method for the Calculation of Equilibrium Constants for the Formation of Physical Clusters of Molecules: Application to Small Water Clusters. *Journal of Chemical Physics* **76**:637–649
- Tobias D J and C L Brooks III 1988. Molecular Dynamics with Internal Coordinate Constraints. *Journal of Chemical Physics* **89**:5115–5126
- Tobias D J, K Tu and M L Klein 1997. Atomic-scale Molecular Dynamics Simulations of Lipid Membranes. *Current Opinion in Colloid and Interface Science* **2**:15–26
- Tuckerman M, B J Berne and G J Martyna 1992. Reversible Multiple Time Scale Molecular Dynamics. *Journal of Chemical Physics* **97**:1990–2001
- van der Ploeg P and H J C Berendsen 1982. Molecular Dynamics Simulation of a Bilayer Membrane. *Journal of Chemical Physics* **76**:3271–3276
- van Gunsteren W F and H J C Berendsen 1982. Algorithms for Brownian Dynamics. *Molecular Physics* **45**:637–547.
- van Gunsteren W F, H J C Berendsen and J A C Rullmann 1981. Stochastic Dynamics for Molecules with Constraints. Brownian Dynamics of n-Alkanes. *Molecular Physics* **44**:69–95.
- Verlet L 1967. Computer 'Experiments' on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. *Physical Review* **159**:98–103
- Watanabe M and M Karplus 1993. Dynamics of Molecules with Internal Degrees of Freedom by Multiple Time-step Methods. *Journal of Chemical Physics* **99**:8063–8074
- Widmalm G and R W Pastor 1992. Comparison of Langevin and Molecular Dynamics Simulations. *Journal of the Chemical Society Faraday Transactions* **88**:1747–1754

- Woodcock L V 1971 Isothermal Molecular Dynamics Calculations for Liquid Salts *Chemical Physics Letters* **10**:257–261.
- Yun-Yu S, W Lu and W F van Gunsteren 1988 On the Approximation of Solvent Effects on the Conformation and Dynamics of Cyclosporin A by Stochastic Dynamics Simulation Techniques *Molecular Simulation* **1** 369–383
- Zhou R and B J Berne 1995 A New Molecular Dynamics Method Combining the Reference System Propagator Algorithm with a Fast Multipole Method for Simulating Proteins and Other Complex Systems. *Journal of Chemical Physics* **103** 9444–9459.