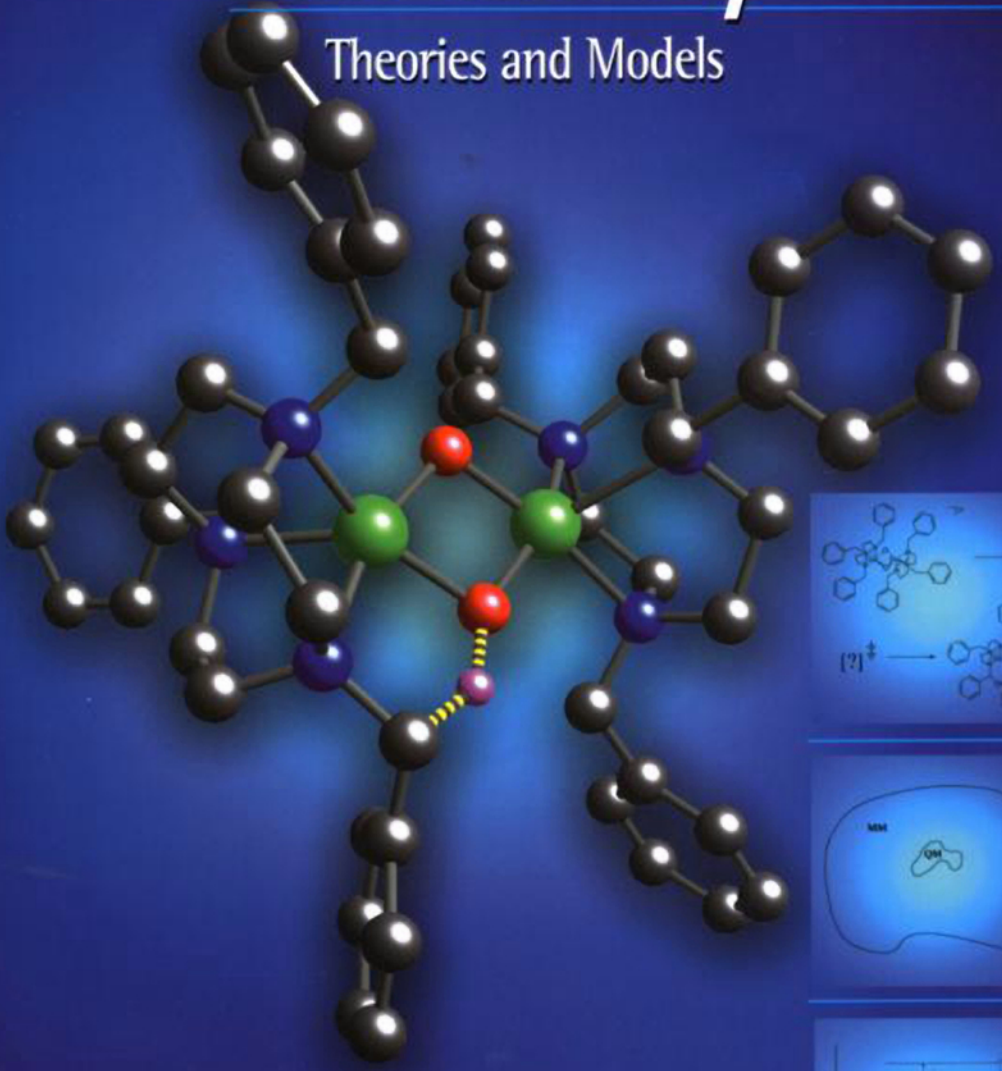
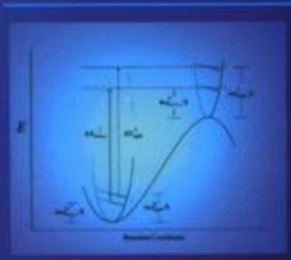
WILEY

Essentials of Computational Chemistry

Theories and Models





Christopher J. Cramer

Essentials of Computational Chemistry

Theories and Models

Christopher J. Cramer

University of Minnesota, USA

JOHN WILEY & SONS, LTD

Baffins Lane, Chichester,
West Sussex, PO19 1UD, England

National 01243 779777
International (+44) 1243 779777

e-mail (for orders and customer service enquiries): cs-books@wiley.co.uk

Visit our Home Page on http://www.wileyeurope.com
or
http://www.wiley.com

Other Wiley Editorial Offices

John Wiley & Sons, Inc., 605 Third Avenue,
New York, NY 10158-0012, USA

Wiley-VCH Verlag GmbH, Pappelallee 3,
D-69469 Weinheim, Germany

John Wiley, Australia, Ltd, 33 Park Road, Milton,
Queensland 4064, Australia

John Wiley & Sons (Asia) Pte Ltd, 2 Clementi Loop #02-01,
Jin Xing Distripark, Singapore 129809

John Wiley & Sons (Canada) Ltd, 22 Worcester Road,
Rexdale, Ontario, M9W 1L1, Canada

Library of Congress Cataloguing in Publication Data

Cramer, Christopher J., 1961-
Essentials of computational chemistry : theories and models / Christopher J. Cramer.
p.cm.
Includes bibliographical references and index.
ISBN 0-471-48551-9 ISBN 0-471-48552-7 (pbk.)
1. Chemistry, Physical and theoretical – Data processing. 2. Chemistry, Physical and theoretical – Mathematical models. I. Title

QD455.3.E4 C73 2002
541'.0285 – dc21 2001057380

British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library

ISBN 0 471-48551 9 (Hardback) 0-471 48552 7 (Paperback)

Typeset by Laserwords Private Limited, Chennai, India
Printed and bound in Great Britain by T J International, Padstow, Cornwall
This book is printed on acid-free paper responsibly manufactured from sustainable forestry,
in which at least two trees are planted for each one used for paper production.

For Katherine

Contents

Preface

Computational chemistry, alternatively sometimes called theoretical chemistry or molecular modeling (reflecting a certain factionalization amongst practitioners), is a field that can be said to be both old and young. It is old in the sense that its foundation was laid with the development of quantum mechanics in the early part of the twentieth century. It is young, however, insofar as arguably no technology in human history has developed at the pace that digital computers have over the last 35 years or so. The digital computer being the 'instrument' of the computational chemist, workers in the field have taken advantage of this progress to develop and apply new theoretical methodologies at a similarly astonishing pace.

The evidence of this progress and its impact on Chemistry in general can be assessed in various ways. Boyd and Lipkowitz, in their book series *Reviews in Computational Chemistry*, have periodically examined such quantifiable indicators as numbers of computational papers published, citations to computational chemistry software packages, and citation rankings of computational chemists. While such metrics need not necessarily be correlated with 'importance', the exponential growth rates they document are noteworthy. My own personal (and somewhat more whimsical) metric is the staggering increase in the percentage of exposition floor space occupied by computational chemistry software vendors at various chemistry meetings worldwide – *someone* must be buying those products!

Importantly, the need for at least a cursory understanding of theory/computation/modeling is by no means restricted to practitioners of the art. Because of the broad array of theoretical tools now available, it is a rare problem of interest that does not occupy the attention of both experimental *and* theoretical chemists. Indeed, the synergy between theory and experiment has vastly accelerated progress in any number of areas (as one example, it is hard to imagine a modern paper on the matrix isolation of a reactive intermediate and its identification by infrared spectroscopy not making a comparison of the experimental spectrum to one obtained from theory/calculation). To take advantage of readily accessible theoretical tools, and to understand the results reported by theoretical collaborators (or competitors), even the wettest of wet chemists can benefit from some familiarity with theoretical chemistry. My objective in this book is to provide a survey of computational chemistry – its underpinnings, its jargon, its strengths and weaknesses – that will be accessible to both the experimental and theoretical communities. The level of the presentation assumes exposure to quantum

and statistical mechanics; particular topics/examples span the range of inorganic, organic, and biological chemistry. As such, this text could be used in a course populated by senior undergraduates and/or beginning graduate students without regard to specialization.

The scope of theoretical methodologies presented in the text reflects my judgment of the degree to which these methodologies impact on a broad range of chemical problems, i.e., the degree to which a practicing chemist may expect to encounter them repeatedly in the literature and thus should understand their applicability (or lack thereof). In some instances, methodologies that do not find much modern use are discussed because they help to illustrate in an intuitive fashion how more contemporary models developed their current form. Indeed, one of my central goals in this book is to render less opaque the fundamental natures of the various theoretical models. By understanding the assumptions implicit in a theoretical model, and the concomitant limitations imposed by those assumptions, one can make informed judgments about the trustworthiness of theoretical results (and economically sound choices of models to apply, if one is about to embark on a computational project).

With no wish to be divisive, it must be acknowledged: there are some chemists who are not fond of advanced mathematics. Unfortunately, it is simply not possible to describe computational chemistry without resort to a fairly hefty number of equations, and, particularly for modern electronic-structure theories, some of those equations are fantastically daunting in the absence of a detailed knowledge of the field. That being said, I offer a promise to present no equation without an effort to provide an intuitive explanation for its form and the various terms within it. In those instances where I don't think such an explanation *can* be offered (of which there are, admittedly, a few), I will provide a qualitative discussion of the area and point to some useful references for those inclined to learn more.

In terms of layout, it might be preferable from a historic sense to start with quantum theories and then develop classical theories as an approximation to the more rigorous formulation. However, I think it is more pedagogically straightforward (and far easier on the student) to begin with classical models, which are in the widest use by experimentalists and tend to feel very intuitive to the modern chemist, and move from there to increasingly more complex theories. In that same vein, early emphasis will be on single-molecule (gas-phase) calculations followed by a discussion of extensions to include condensed-phase effects. While the book focuses primarily on the calculation of equilibrium properties, excited states and reaction dynamics are dealt with as advanced subjects in later chapters.

The quality of a theory is necessarily judged by its comparison to (accurate) physical measurements. Thus, careful attention is paid to offering comparisons between theory and experiment for a broad array of physical observables (the first chapter is devoted in part to enumerating these). In addition, there *is* some utility in the computation of things which cannot be observed (e.g., partial atomic charges), and these will also be discussed with respect to the performance of different levels of theory. However, the best way to develop a feeling for the scope and utility of various theories is to apply them, and instructors are encouraged to develop computational problem sets for their students. To assist in that regard, case studies appear at the end of most chapters illustrating the employ of one or more of the models most recently presented. The studies are drawn from the chemical literature;

depending on the level of instruction, reading and discussing the original papers as part of the class may well be worthwhile, since any synopsis necessarily does away with some of the original content.

Perversely, perhaps, I do not include in this book specific problems. Indeed, I provide almost no discussion of such nuts and bolts issues as, for example, how to enter a molecular geometry into a given program. The reason I eschew these undertakings is not that I think them unimportant, but that computational chemistry software is not particularly well standardized, and I would like neither to tie the book to a particular code or codes nor to recapitulate material found in users' manuals. Furthermore, the hardware and software available in different venues varies widely, so individual instructors are best equipped to handle technical issues themselves. With respect to illustrative problems for students, there *are* reasonably good archives of such exercises provided either by software vendors as part of their particular package or developed for computational chemistry courses around the world. Chemistry 8021 at the University of Minnesota, for example, has several years worth of problem sets (with answers) available at pollux.chem.umn.edu/8021. Given the pace of computational chemistry development and of modern publishing, such archives are expected to offer a more timely range of challenges in any case.

A brief summary of the mathematical notation adopted throughout this text is in order. Scalar quantities, whether constants or variables, are represented by italic characters. Vectors and matrices are represented by boldface characters (individual matrix *elements* are scalar, however, and thus are represented by italic characters that are indexed by subscript(s) identifying the particular element). Quantum mechanical operators are represented by italic characters if they have scalar expectation values and boldface characters if their expectation values are vectors or matrices (or if they are typically *constructed* as matrices for computational purposes). The only deliberate exception to the above rules is that quantities represented by Greek characters typically are made neither italic nor boldface, irrespective of their scalar or vector/matrix nature.

Finally, as with most textbooks, the total content encompassed herein is such that only the most masochistic of classes would attempt to go through this book cover to cover in the context of a typical, semester-long course. My intent in coverage is not to act as a firehose, but to offer a reasonable degree of flexibility to the instructor in terms of optional topics. Thus, for instance, Chapters 3 and 11–13 could readily be skipped in courses whose focus is primarily on the modeling of small- and medium-sized molecular systems. Similarly, courses with a focus on macromolecular modeling could easily choose to ignore the more advanced levels of quantum mechanical modeling. And, clearly, time constraints in a typical course are unlikely to allow the inclusion of more than one of the last two chapters. These practical points having been made, one can always hope that the eager student, riveted by the content, will take time to read the rest of the book him- or herself!

Christopher J. Cramer

1

What are Theory, Computation, and Modeling?

1.1 Definition of Terms

A clear definition of terms is critical to the success of all communication. Particularly in the area of computational chemistry, there is a need to be careful in the nomenclature used to describe predictive tools, since this often helps clarify what approximations have been made in the course of a modeling 'experiment'. For the purposes of this textbook, we will adopt a specific convention for what distinguishes theory, computation, and modeling.

In general, 'theory' is a word most scientists are entirely comfortable with. A theory is one or more rules that are postulated to govern the behavior of physical systems. Often, in science at least, such rules are quantitative in nature and expressed in the form of a mathematical equation. Thus, for example, one has the theory of Einstein that the energy of a particle, E, is equal to its relativistic mass, m, times the speed of light in a vacuum, c, squared,

$$E = mc^2 \tag{1.1}$$

The quantitative nature of scientific theories allows them to be tested by experiment. This testing is the means by which the applicable range of a theory is elucidated. Thus, for instance, many theories of classical mechanics prove applicable to macroscopic systems but break down for very small systems, where one must instead resort to quantum mechanics. The observation that a theory has limits in its applicability might, at first glance, seem a sufficient flaw to warrant discarding it. However, if a sufficiently large number of 'interesting' systems falls within the range of the theory, practical reasons tend to motivate its continued use. Of course, such a situation tends to inspire efforts to find a more *general* theory that is not subject to the limitations of the original. Thus, for example, classical mechanics can be viewed as a special case of the more general quantum mechanics in which the presence of macroscopic masses and velocities leads to a simplification of the governing equations (and concepts).

Such simplifications of general theories under special circumstances can be key to getting anything useful done! One would certainly *not* want to design the pendulum for a mechanical

clock using the fairly complicated mathematics of quantal theories, for instance, although the process would ultimately lead to the same result as that obtained from the simpler equations of the more restricted classical theories. Furthermore, at least at the start of the twenty-first century, a generalized 'theory of everything' does not yet exist. For instance, efforts to link theories of quantum electromagnetics and theories of gravity continue to be pursued.

Occasionally, a theory has proven so robust over time, even if only within a limited range of applicability, that it is called a 'law'. For instance, Coulomb's law specifies that the energy of interaction (in arbitrary units) between two point charges is given by

$$E = \frac{q_1 q_2}{\varepsilon r_{12}} \tag{1.2}$$

where q is a charge, ε is the dielectric constant of a homogeneous medium (possibly vacuum) in which the charges are embedded, and r_{12} is the distance between them. However, the term 'law' is best regarded as honorific – indeed, one might regard it as hubris to imply that experimentalists *can* discern the laws of the universe within a finite span of time.

Theory behind us, let us now move on to 'model'. The difference between a theory and a model tends to be rather subtle, and largely a matter of intent. Thus, the goal of a theory tends to be to achieve as great a generality as possible, irrespective of the practical consequences. Quantum theory, for instance, has breathtaking generality, but the practical consequence is that the equations that govern quantum theory are intractable for all but the most ideal of systems. A model, on the other hand, typically involves the deliberate introduction of simplifying approximations into a more general theory so as to extend its practical utility. Indeed, the approximations sometimes go to the extreme of rendering the model deliberately qualitative. Thus, one can regard the valence-shell-electron-pair repulsion (VSEPR; an acronym glossary is provided as Appendix A of this text) model familiar to most students of inorganic chemistry as a drastic simplification of quantum mechanics to permit discrete choices for preferred conformations of inorganic complexes. (While serious theoreticians may shudder at the empiricism that often governs such drastic simplifications, and mutter gloomily about lack of 'rigor', the value of a model is not in its intrinsic beauty, of course, but in its ability to solve practical problems.)

Another feature sometimes characteristic of a *quantitative* 'model' is that it incorporates certain constants that are derived wholly from experimental data, i.e., they are empirically determined. Again, the degree to which this distinguishes a model from a theory can be subtle. The speed of light and the charge of the electron are fundamental constants of the universe that appear either explicitly or implicitly in Eqs. (1.1) and (1.2), and we know these values only through experimental measurement. So, again, the issue tends to be intent. A model is often designed to apply specifically to a restricted volume of what we might call chemical space. For instance, we might imagine developing a model that would predict the free energy of activation for the hydrolysis of substituted β-lactams in water. Our motivation, obviously, would be the therapeutic utility of these species as antibiotics. Because we are limiting ourselves to consideration of only very specific kinds of bond-making and bond-breaking, we may be able to construct a model that takes advantage of a few experimentally known free energies of activation and correlates them with some other measured or predicted

quantity. For example, we might find from comparison with X-ray crystallography that there is a linear correlation between the aqueous free energy of activation, $\Delta G^{\ddagger}$, and the length of the lactam C–N bond in the crystal, r_{CN} (Figure 1.1). Our 'model' would then be

$$\Delta G^{\ddagger} = a r_{CN} + b \tag{1.3}$$

where a would be the slope (in units of energy per length) and b the intercept (in units of energy) for the empirically determined correlation.

Equation (1.3) represents a very simple model, and that simplicity derives, presumably, from the small volume of chemical space over which it appears to hold. As it is hard to imagine deriving Eq. (1.3) from the fundamental equations of quantum mechanics, it might be more descriptive to refer to it as a 'relationship' rather than a 'model'. That is, we make some attempt to distinguish between correlation and causality. For the moment, we will not parse the terms too closely.

An interesting question that arises with respect to Eq. (1.3) is whether it may be more broadly applicable. For instance, might the model be useful for predicting the free energies of activation for the hydrolysis of γ-lactams? What about amides in general? What about imides? In a statistical sense, these chemical questions are analogous to asking about the degree to which a correlation may be trusted for extrapolation vs. interpolation. One might say that we have derived a correlation involving two axes of multi-dimensional chemical space, activation free energy for β-lactam hydrolysis and β-lactam C–N bond length. Like any correlation, our model is expected to be most robust when used in an interpolative sense, i.e., when applied to newly measured β-lactam C–N bonds with lengths that fall within the range of the data used to derive the correlation. Increasingly less certain will be application of Eq. (1.3) to β-lactam bond lengths that are *outside* the range used to derive the correlation,

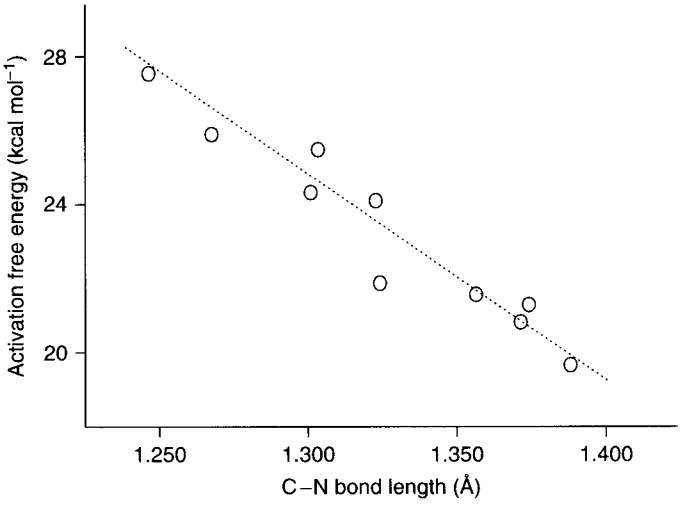


Figure 1.1 Correlation between activation free energy for aqueous hydrolysis of β-lactams and lactam C–N bond lengths as determined from X-ray crystallography (data entirely fictitious)

or assumption that other chemical axes, albeit qualitatively similar (like γ-lactam C–N bond lengths), will be coincident with the abscissa.

Thus, a key question in one's mind when evaluating any application of a theoretical model should be, 'How similar is the system being studied to systems that were employed in the development of the model?' The generality of a given model can only be established by comparison to experiment for a wider and wider variety of systems. This point will be emphasized repeatedly throughout this text.

Finally, there is the definition of 'computation'. While theories and models like those represented by Eqs. (1.1), (1.2), and (1.3), are not particularly taxing in terms of their mathematics, many others can only be efficiently put to use with the assistance of a digital computer. Indeed, there is a certain synergy between the development of chemical theories and the development of computational hardware, software, etc. If a theory cannot be tested, say because solution of the relevant equations lies outside the scope of practical possibility, then its utility cannot be determined. Similarly, advances in computational technology can permit existing theories to be applied to increasingly complex systems to better gauge the degree to which they are robust. These points are expanded upon in Section 1.4. Here we simply close with the concise statement that 'computation' is the use of digital technology to solve the mathematical equations defining a particular theory or model.

With all these definitions in hand, we may return to a point raised in the preface, namely, what is the difference between 'Theory', 'Molecular Modeling', and 'Computational Chemistry'? To the extent members of the community make distinctions, 'theorists' tend to have as their greatest goal the development of new theories and/or models that have improved performance or generality over existing ones. Researchers involved in 'molecular modeling' tend to focus on target systems having particular chemical relevance (e.g., for economic reasons) and to be willing to sacrifice a certain amount of theoretical rigor in favor of getting the right answer in an efficient manner. Finally, 'computational chemists' may devote themselves not to chemical aspects of the problem, *per se*, but to computer related aspects, e.g., writing improved algorithms for solving particularly difficult equations, or developing new ways to encode or visualize data, either as input to or output from a model. As with any classification scheme, there are no distinct boundaries recognized either by observers or by individual researchers, and certainly a given research undertaking may involve significant efforts undertaken within all three of the areas noted above. In the spirit of inclusiveness, we will treat the terms as essentially interchangeable.

1.2 Quantum Mechanics

The postulates and theorems of quantum mechanics form the rigorous foundation for the prediction of observable chemical properties from first principles. Expressed somewhat loosely, the fundamental postulates of quantum mechanics assert that microscopic systems are described by 'wave functions' that completely characterize all of the physical properties of the system. In particular, there are quantum mechanical 'operators' corresponding to each physical observable that, when applied to the wave function, allow one to predict the probability of finding the system to exhibit a particular value or range of values (scalar, vector,

etc.) for that observable. This text assumes prior exposure to quantum mechanics and some familiarity with operator and matrix formalisms and notation.

However, many successful chemical models exist that do not necessarily have obvious connections with quantum mechanics. Typically, these models were developed based on intuitive concepts, i.e., their forms were determined inductively. In principle, any successful model *must* ultimately find its basis in quantum mechanics, and indeed *a posteriori* derivations have illustrated this point in select instances, but often the form of a good model is more readily grasped when rationalized on the basis of intuitive chemical concepts rather than on the basis of quantum mechanics (the latter being desperately non-intuitive at first blush).

Thus, we shall leave quantum mechanics largely unreviewed in the next two chapters of this text, focusing instead on the intuitive basis for classical models falling under the heading of 'molecular mechanics'. Later in the text, we shall see how some of the fundamental approximations used in molecular mechanics can be justified in terms of well-defined approximations to more complete quantum mechanical theories.

1.3 Computable Quantities

What predictions can be made by the computational chemist? In principle, if one can measure it, one can predict it. In practice, some properties are more amenable to accurate computation than others. There is thus some utility in categorizing the various properties most typically studied by computational chemists.

1.3.1 Structure

Let us begin by focusing on isolated molecules, as they are the fundamental unit from which pure substances are constructed. The minimum information required to specify a molecule is its molecular formula, i.e., the atoms of which it is composed, and the manner in which those atoms are connected. Actually, the latter point should be put more generally. What is required is simply to know the relative positions of all of the atoms in space. Connectivity, or 'bonding', is itself a property that is open to determination. Indeed, the determination of the 'best' structure from a chemically reasonable (or unreasonable) guess is a very common undertaking of computational chemistry. In this case 'best' is defined as having the lowest possible energy given an overall connectivity roughly dictated by the starting positions of the atoms as chosen by the theoretician (the process of structure optimization is described in more detail in subsequent chapters).

This sounds relatively simple because we are talking about the modeling of an isolated, single molecule. In the laboratory, however, we are much more typically dealing with an equilibrium mixture of a very large number of molecules at some non-zero temperature. In that case, *measured* properties reflect thermal averaging, possibly over multiple discrete stereoisomers, tautomers, etc., that are structurally quite different from the idealized model system, and great care must be taken in making comparisons between theory and experiment in such instances.

1.3.2 Potential Energy Surfaces

The first step to making the theory more closely mimic the experiment is to consider not just one structure for a given chemical formula, but all possible structures. That is, we fully characterize the potential energy surface (PES) for a given chemical formula (this requires invocation of the Born–Oppenheimer approximation, as discussed in more detail in Chapters 4 and 15). The PES is a hypersurface defined by the potential energy of a collection of atoms over all possible atomic arrangements; the PES has $3N - 6$ coordinate dimensions, where N is the number of atoms ≥ 3. This dimensionality derives from the three-dimensional nature of Cartesian space. Thus each structure, which is a point on the PES, can be defined by a vector $\mathbf{X}$ where

$$\mathbf{X} \equiv (x_1, y_1, z_1, x_2, y_2, z_2, \ldots, x_N, y_N, z_N) \tag{1.4}$$

and x_i, y_i, and z_i are the Cartesian coordinates of atom i. However, this expression of $\mathbf{X}$ does not *uniquely* define the structure because it involves an arbitrary origin. We can reduce the dimensionality without affecting the structure by removing the three dimensions associated with translation of the structure in the x, y, and z directions (e.g., by insisting that the molecular center of mass be at the origin) and removing the three dimensions associated with rotation about the x, y, and z axes (e.g., by requiring that the principal moments of inertia align along those axes in increasing order).

A different way to appreciate this reduced dimensionality is to imagine constructing a structure vector atom by atom (Figure 1.2), in which case it is most convenient to imagine the dimensions of the PES being internal coordinates (i.e., bond lengths, valence angles, etc.). Thus, choice of the first atom involves no degrees of geometric freedom – the atom defines the origin. The position of the second atom is specified by its distance from the first. So, a two-atom system has a single degree of freedom, the bond length; this corresponds to $3N - 5$ degrees of freedom, as should be the case for a linear molecule. The third atom must be specified either by its distances to each of the preceding atoms, or by a distance to one and an angle between the two bonds thus far defined to a common atom. The three-atom system, if collinearity is not enforced, has 3 total degrees of freedom, as it should. Each additional atom requires three coordinates to describe its position. There are several ways to envision describing those coordinates. As in Figure 1.2, they can either be a bond length, a valence angle, and a dihedral angle, or they can be a bond length and two valence angles. Or, one can imagine that the first three atoms have been used to create a fixed Cartesian reference frame, with atom 1 defining the origin, atom 2 defining the direction of the positive x axis, and atom 3 defining the upper half of the xy plane. The choice in a given calculation is a matter of computational convenience. Note, however, that the *shapes* of particular surfaces necessarily depend on the choice of their coordinate systems, although they will map to one another in a one-to-one fashion.

Particularly interesting points on PESs include local minima, which correspond to optimal molecular structures, and saddle points (i.e., points characterized by having no slope in any direction, downward curvature for a single coordinate, and upward curvature for all of the other coordinates). Simple calculus dictates that saddle points are lowest energy barriers

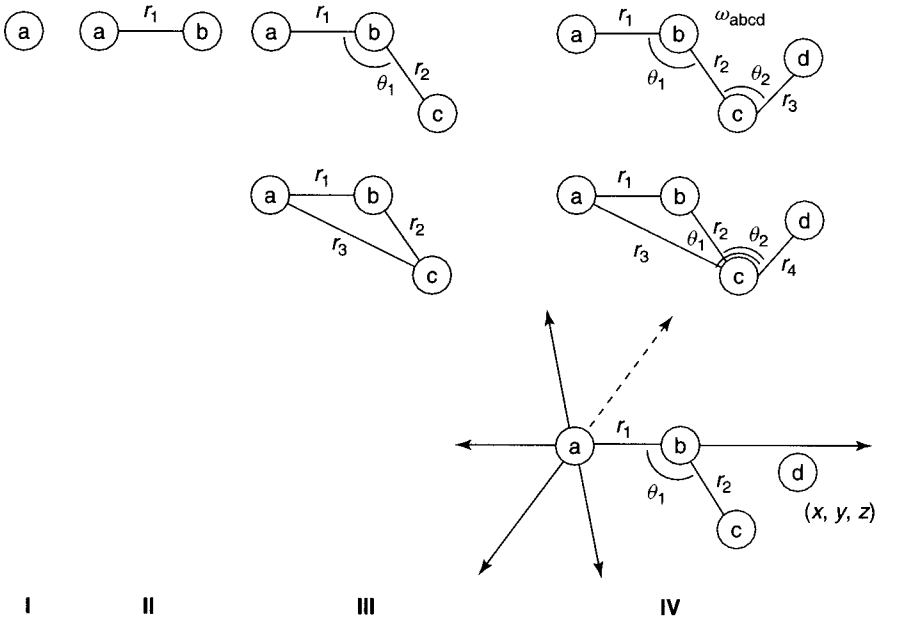


Figure 1.2 Different means for specifying molecular geometries. In frame **I**, there are no degrees of freedom as only the nature of atom 'a' has been specified. In frame **II**, there is a single degree of freedom, namely the bond length. In frame **III**, location of atom 'c' requires two additional degrees of freedom, either two bond lengths or a bond length and a valence angle. Frame **IV** illustrates various ways to specify the location of atom 'd'; note that in every case, three new degrees of freedom must be specified, either in internal or Cartesian coordinates

on paths connecting minima, and thus they can be related to the chemical concept of a transition state. So, a complete PES provides, for a given collection of atoms, complete information about all possible chemical structures and all isomerization pathways interconnecting them.

Unfortunately, complete PESs for polyatomic molecules are very hard to visualize, since they involve a large number of dimensions. Typically, we take slices through potential energy surfaces that involve only a single coordinate (e.g., a bond length) or perhaps two coordinates, and show the relevant reduced-dimensionality energy curves or surfaces (Figure 1.3). Note that some care must be taken to describe the nature of the slice with respect to the *other* coordinates. For instance, was the slice a hyperplane, implying that all of the non-visualized coordinates have fixed values, or was it a more general hypersurface? A typical example of the latter choice is one where the non-visualized coordinates take on values that minimize the potential energy given the value of the visualized coordinate(s). Thus, in the case of a single visualized dimension, the curve attempts to illustrate the minimum energy path associated with varying the visualized coordinate. [We must say 'attempts' here, because an actual continuous path connecting any two structures on a PES may involve any number of structures all of which have the same value for a single internal coordinate. When that

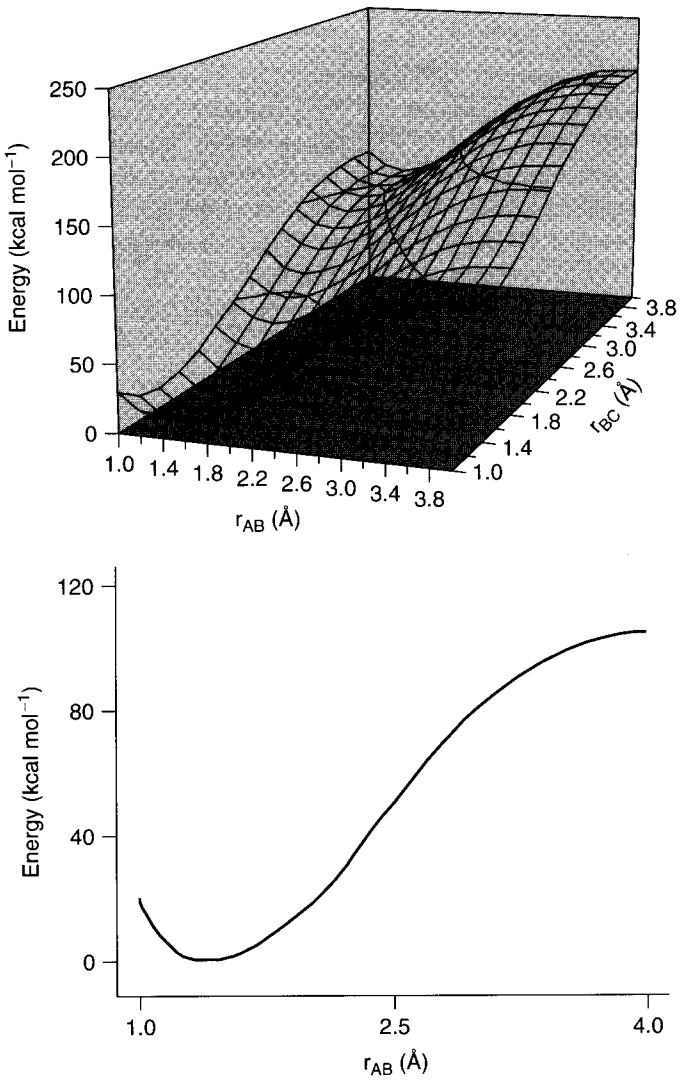


Figure 1.3 The full PES for the hypothetical molecule ABC requires four dimensions to display ($3N - 6 = 3$ coordinate degrees of freedom plus one dimension for energy). The three-dimensional plot (top) represents a hyperslice through the full PES showing the energy as a function of two coordinate dimensions, the AB and BC bond lengths, while taking a fixed value for the angle ABC (a typical choice might be the value characterizing the global minimum on the full PES). A further slice of this surface (bottom) now gives the energy as a function of a single dimension, the AB bond length, where the BC bond length is now also treated as frozen (again at the equilibrium value for the global minimum)

path is projected onto the dimension defined by that single coordinate (or any reduced number of dimensions including it) the resulting curve is a non-single-valued function of the dimension. When we arbitrarily choose to use the lowest energy point for each value of the varied coordinate, we may introduce discontinuities in the actual structures, even though the curve may appear to be smooth (Figure 1.4). Thus, the generation and interpretation of such 'partially relaxed' potential energy curves should involve a check of the individual structures to ensure that such a situation has not arisen.]

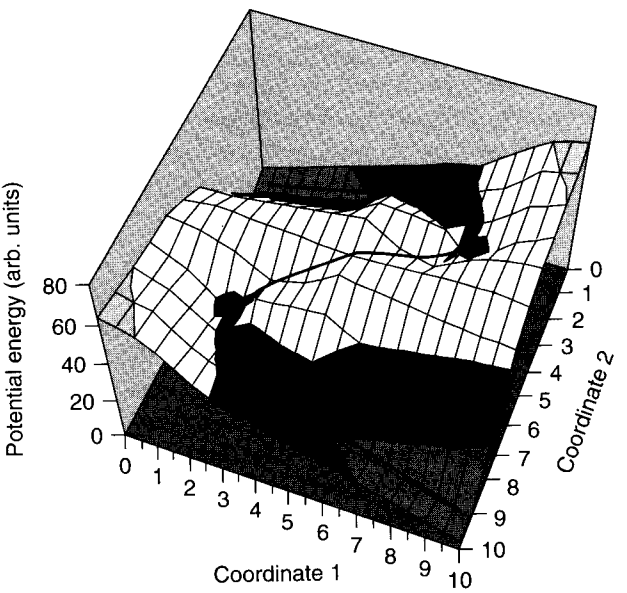


Figure 1.4 The bold line in (a) traces out a lowest-energy path connecting two minima of energy 0, located at coordinates (0,1) and (10,9), on a hypothetical three-dimensional PES – shaded regions correspond to contour levels spanning 20 energy units. Following the path starting from point (0,1) in the upper left, coordinate 1 initially smoothly increases to a value of about 7.5 while coordinate 2 undergoes little change. Then, however, because of the coupling between the two coordinates, coordinate 1 begins *decreasing* while coordinate 2 changes. The 'transition state structure' (saddle point) is reached at coordinates (5,5) and has energy 50. On this PES, the path downward is the symmetric reverse of the path up. If the full path is projected so as to remove coordinate 2, the two-dimensional potential energy diagram (b) is generated. The solid curve is what would result if we only considered lowest energy structures having a given value of coordinate 1. Of course, the solid curve is discontinuous in coordinate 2, since approaches to the 'barrier' in the solid curve from the left and right correspond to structures having values for coordinate 2 of about 1 and 9, respectively. The dashed curve represents the higher energy structures that appear on the smooth, continuous, three-dimensional path. If the lower potential energy diagram were to be generated by driving coordinate 1, and care were not taken to note the discontinuity in coordinate 2, the barrier for interconversion of the two minima would be underestimated by a factor of 2 in this hypothetical example. (For an actual example of this phenomenon, see Cramer *et al.* 1994.)

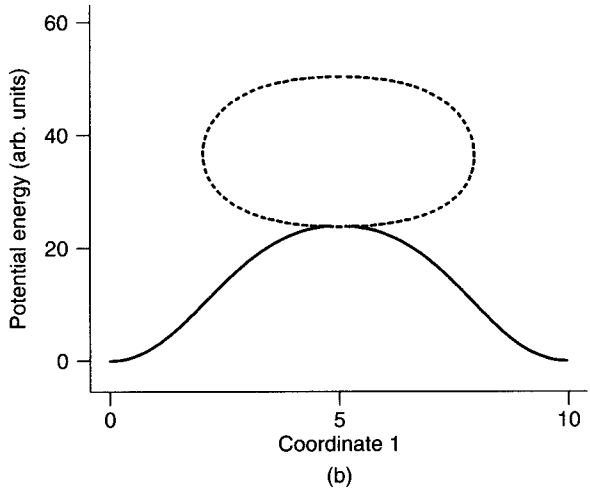


Figure 1.4 *(Continued)*

Finally, sometimes slices are chosen so that all structures in the slicing surface belong to a particular symmetry point group. The utility of symmetry will be illustrated in various situations throughout the text.

With the complete PES in hand (or, more typically, with the region of the PES that would be expected to be chemically accessible under the conditions of the experimental system being modeled), one can take advantage of standard precepts of statistical mechanics (see Chapter 10) to estimate equilibrium populations for situations involving multiple stable molecular structures and compute ensemble averages for physical observables.

1.3.3 Chemical Properties

One can arbitrarily divide the properties one might wish to estimate by computation into three classes. The first is 'single-molecule' properties, that is, properties that could in principle be measured from a single molecule, even though, in practice, use of a statistical ensemble may be required for practical reasons. Typical examples of such properties are spectral quantities. Thus, theory finds considerable modern application to predicting nuclear magnetic resonance (NMR) chemical shifts and coupling constants, electron paramagnetic resonance (EPR) hyperfine coupling constants, absorption maxima for rotational, vibrational, and electronic spectra (typically in the microwave, infrared, and ultraviolet/visible regions of the spectrum, respectively), and electron affinities and ionization potentials (see Chapter 9).

With respect to molecular energetics, one can, in principle, measure the total energy of a molecule (i.e., the energy required to separate it into its constituent nuclei and electrons all infinitely separated from one another and at rest). More typically, however, laboratory measurements focus on thermodynamic quantities such as enthalpy, free energy, etc., and

this is the second category into which predicted quantities fall. Theory is extensively used to estimate equilibrium constants, which are derived from free energy differences between minima on a PES, and rate constants, which, with certain assumptions (see Chapter 15), are derived from free energy differences between minima on a PES and connected transition-state structures. Thus, theory may be used to predict reaction thermochemistries, heats of formation and combustion, kinetic isotope effects, complexation energies (key to molecular recognition), acidity and basicity (e.g., pK_a values), 'stability', and hydrogen bond strengths, to name a few properties of special interest. With a sufficiently large collection of molecules being modeled, theory can also, in principle, compute bulk thermodynamic phenomena such as solvation effects, phase transitions, etc., although the complexity of the system may render such computations quite challenging.

Finally, there are computable 'properties' that do not correspond to physical observables. One may legitimately ask about the utility of such ontologically indefensible constructs! However, one should note that unmeasurable properties long predate computational chemistry – some examples include bond order, aromaticity, reaction concertedness, and isoelectronic, -steric, and -lobal behavior. These properties involve *conceptual* models that have proven sufficiently useful in furthering chemical understanding that they have overcome objections to their not being uniquely defined.

In cases where such models take measurable quantities as input (e.g., aromaticity models that consider heats of hydrogenation or bond-length alternation), clearly those measurable quantities are also computable. There are additional non-observables, however, that are unique to modeling, usually being tied to some aspect of the computational algorithm. A good example is atomic partial charge (see Chapter 9), which can be a very useful chemical concept for understanding molecular reactivity.

1.4 Cost and Efficiency

1.4.1 Intrinsic Value

Why has the practice of computational chemistry skyrocketed in the last few years? Try taking this short quiz: Chemical waste disposal and computational technology – which of these two keeps getting more and more expensive and which less and less? From an economic perspective, at least, theory is enormously attractive as a tool to reduce the costs of doing experiments.

Chemistry's impact on modern society is most readily perceived in the creation of materials, be they foods, textiles, circuit boards, fuels, drugs, packaging, etc. Thus, even the most ardent theoretician would be unlikely to suggest that theory could ever *supplant* experiment. Rather, most would opine that opportunities exist for *combining* theory with experiment so as to take advantage of synergies between them.

With that in mind, one can categorize efficient combinations of theory and experiment into three classes. In the first category, theory is applied *post facto* to a situation where some ambiguity exists in the interpretation of existing experimental results. For example, photolysis of a compound in an inert matrix may lead to a single product species as

analyzed by spectroscopy. However, the identity of this unique product may not be obvious given a number of plausible alternatives. A calculation of the energies and spectra for *all* of the postulated products provides an opportunity for comparison and may prove to be definitive.

In the second category, theory may be employed in a simultaneous fashion to optimize the design and progress of an experimental program. Continuing the above analogy, *a priori* calculation of spectra for plausible products may assist in choosing experimental parameters to permit the observation of minor components which might otherwise be missed in complicated mixture (e.g., theory may allow the experimental instrument to be tuned properly to observe a signal whose location would not otherwise be *a priori* predictable).

Finally, theory may be used to predict properties which might be especially difficult or dangerous (i.e., costly) to measure experimentally. In the difficult category are such data as rate constants for the reactions of trace, upper-atmospheric constituents that might play an important role in the ozone cycle. For sufficiently small systems, levels of quantum mechanical theory can now be brought to bear that have accuracies comparable to the best modern experimental techniques, and computationally derived rate constants may find use in complex kinetic models until such time as experimental data are available. As for dangerous experiments, theoretical pre-screening of a series of toxic or explosive compounds for desirable (or undesirable) properties may assist in prioritizing the order in which they are prepared, thereby increasing the probability that an acceptable product will be arrived at in a maximally efficient manner.

1.4.2 Hardware and Software

All of these points being made, even computational chemistry is not without cost. In general, the more sophisticated the computational model, the more expensive in terms of computational resources. The talent of the well-trained computational chemist is knowing how to maximize the accuracy of a prediction while minimizing the investment of such resources. A primary goal of this text is to render more clear the relationship between accuracy and cost for various levels of theory so that even relatively inexperienced users can make informed assessments of the likely utility (before the fact) or credibility (after the fact) of a given calculation.

To be more specific about computational resources, we may, without going into a great deal of engineering detail, identify three features of a modern digital computer that impact upon its utility as a platform for molecular modeling. The first feature is the speed with which it carries out mathematical operations. Various metrics are used when comparing the speed of 'chips', which are the fundamental processing units. One particularly useful one is the number of floating-point operations per second (FLOPS) that the chip can accomplish. That is, how many mathematical manipulations of decimally represented numbers can be carried out (the equivalent measure for integers is IPS). Various benchmark computer codes are available for comparing one chip to another, and one should always bear in mind that measured processor speeds are dependent on which code or set of codes was used. Different kinds of mathematical operations or different orderings of operations can have effects as large

as an order of magnitude on individual machine speeds because of the way the processors are designed and because of the way they interact with other features of the computational hardware.

The second feature affecting performance is memory. In order to carry out a floating-point operation, there must be floating-point numbers on which to operate. Numbers (or characters) to be processed are stored in a magnetic medium referred to as memory. In a practical sense, the size of the memory associated with a given processor sets the limit on the total amount of information to which it has 'instant' access. In modern multiprocessor machines, this definition has grown more fuzzy, as there tend to be multiple memory locations, and the speed with which a given processor can access a given memory location varies depending upon their physical locations with respect to one another. The somewhat unsurprising bottom line is that more memory and shorter access times tend to lead to improved computational performance.

The last feature is storage, typically referred to as disk since that has been the read/write storage medium of choice for the last several years. Storage is exactly like memory, in the sense that it holds number or character data, but it is accessible to the processing unit at a much slower rate than is memory. It makes up for this by being much cheaper and being, in principle, limitless and permanent. Calculations which need to read and/or write data to a disk necessarily proceed more slowly than do calculations that can take place entirely in memory. The difference is sufficiently large that there are situations where, rather than storing on disk data that will be needed later, it is better to throw them away (because memory limits require you to overwrite the locations in which they are stored), as subsequent recomputation of the needed data is faster than reading it back from disk storage. Such a protocol is usually called a 'direct' method (see Almlöf, Faegri, and Korsell 1982).

Processors, memory, and storage media are components of a computer referred to as 'hardware'. However, the efficiency of a given computational task depends also on the nature of the instructions informing the processor how to go about implementing that task. Those instructions are encoded in what is known as 'software'. In terms of computational chemistry, the most obvious piece of software is the individual program or suite of programs with which the chemist interfaces in order to carry out a computation. However, that is by no means the only software involved. Most computational chemistry software consists of a large set of instructions written in a 'high-level' programming language (e.g., FORTRAN, C++), and choices of the user dictate which sets of instructions are followed in which order. The collection of all such instructions is usually called a 'code' (listings of various computational chemistry codes can be found at websites such as http://cmm.info.nih.gov/modeling/software.html; in addition, the series *Reviews in Computational Chemistry* (Boyd, D. B. and Lipkowitz, K., Eds., VCH: New York) periodically publishes comprehensive listings of available software). But the language of the code cannot be interpreted directly by the processor. Instead, a series of other pieces of software (compilers, assemblers, etc.) translate the high-level language instructions into the step-by-step operations that are carried out by the processing unit. Understanding how to write code (in whatever language) that takes the best advantage of the total hardware/software environment on a particular computer is a key aspect to the creation of an efficient software package.

1.4.3 Algorithms

In a related sense, the manner in which mathematical equations are turned into computer instructions is also key to efficient software development. Operations like addition and subtraction do not allow for much in the way of innovation, needless to say, but operations like matrix diagonalization, numerical integration, etc., are sufficiently complicated that different algorithms leading to the same (correct) result can vary markedly in computational performance. A great deal of productive effort in the last decade has gone into the development of so-called 'linear-scaling' algorithms for various levels of theory. Such an algorithm is one that permits the cost of a computation to scale roughly linearly with the size of the system studied. At first, this may not sound terribly demanding, but a quick glance back at Coulomb's law [Eq. (1.2)] will help to set this in context. Coulomb's law states that the potential energy from the interaction of charged particles depends on the pairwise interaction of all such particles. Thus, one might expect any calculation of this quantity to scale as the *square* of the size of the system (there are $n(n-1)/2$ such interactions where n is the number of particles). However, for sufficiently large systems, sophisticated mathematical 'tricks' permit the scaling to be brought down to linear.

In this text, we will not be particularly concerned with algorithms – not because they are not important but because such concerns are more properly addressed in advanced textbooks aimed at future practitioners of the art. Our focus will be primarily on the conceptual aspects of particular computational models, and not necessarily on the most efficient means for implementing them.

We close this section with one more note on careful nomenclature. A 'code' renders a 'model' into a set of instructions that can be understood by a digital computer. Thus, if one applies a particular model, let us say the molecular mechanics model called MM3 (which will be described in the next chapter) to a particular problem, say the energy of chair cyclohexane, the results should be completely independent of which code one employs to carry out the calculation. If two pieces of software (let's call them MYPROG and YOURPROG) differ by more than the numerical noise that can arise because of different round-off conventions with different computer chips (or having set different tolerances for what constitutes a converged calculation) then one (or both!) of those pieces of software is *incorrect*. In colloquial terms, there is a 'bug' in the incorrect code(s).

Furthermore, it is never correct to refer to the results of a calculation as deriving from the code, e.g., to talk about one's 'MYPROG structure'. Rather, the results derive from the model, and the structure is an 'MM3 structure'. It is not simply incorrect to refer to the results of the calculation by the name of the code, it is confusing: MYPROG may well contain code for several *different* molecular mechanics models, not just MM3, so simply naming the program is insufficiently descriptive.

It is regrettable, but must be acknowledged, that certain models found in the chemical literature are themselves not terribly well defined. This tends to happen when features or parameters of a model are updated without any change in the name of the model as assigned by the original authors. When this happens, codes implementing older versions of the model will disagree with codes implementing newer versions even though each uses the same name for the model. Obviously, developers should scrupulously avoid ever allowing this situation

Table 1.1 Useful quantities in atomic and other units

Physical quantity (unit name)	Symbol	Value in a.u.	Value in SI units	Value(s) in other units
Angular momentum	$\hbar$	1	1.055×10^{-34} J s	2.521×10^{-35} cal s
Mass	m_e	1	9.109×10^{-31} kg	
Charge	e	1	1.602×10^{-19} C	1.519×10^{-14} statC
Vacuum permittivity	$4\pi\varepsilon_0$	1	1.113×10^{-10} C^2 J^{-1} m^{-1}	2.660×10^{-21} C^2 cal^{-1} $Å^{-1}$
Length (bohr)	a_0	1	5.292×10^{-11} m	0.529 Å 52.9 pm
Energy (hartree)	E_h	1	4.360×10^{-18} J	627.51 kcal mol^{-1} 2.626×10^3 kJ mol^{-1} 27.211 eV 2.195×10^5 cm^{-1}
Electric dipole moment	ea_0	1	8.478×10^{-30} C m	2.542 D
Electric polarizability	$e^2a_0^2E_h^{-1}$	1	1.649×10^{-41} C^2 m^2 J^{-1}	
Planck's constant	h	2π	6.626×10^{-34} J s	
Speed of light	c	1.370×10^2	2.998×10^8 m s^{-1}	
Bohr magneton	μ_B	0.5	9.274×10^{-24} J T^{-1}	
Nuclear magneton	μ_N	2.723×10^{-4}	5.051×10^{-27} J T^{-1}	

to arise. To be safe, scientific publishing that includes computational results should always state what code or codes were used, *to include version numbers*, in obtaining particular model results (clearly version control of computer codes is thus just as critical as it is for models).

1.5 Note on Units

In describing a computational model, a clear equation can be worth 1000 words. One way to render equations more clear is to work in atomic (or theorist's) units. In a.u., the charge on the proton, e, the mass of the electron, m_e, and $\hbar$ (i.e., Planck's constant divided by 2π) are all defined to have magnitude 1. When converting equations expressed in SI units (as opposed to Gaussian units), $4\pi\varepsilon_0$, where ε_0 is the permittivity of the vacuum, is also defined to have magnitude 1. As the magnitude of these quantities is unity, they are dropped from relevant equations, thereby simplifying the notation. Other atomic units having magnitudes of unity can be derived from these three by dimensional analysis. For instance, $\hbar^2/m_e e^2$ has units of distance and is defined as 1 a.u.; this atomic unit of distance is also called the 'bohr' and symbolized by a_0. Similarly, e^2/a_0 has units of energy, and defines 1 a.u. for this quantity, also called 1 hartree and symbolized by E_h. Table 1.1 provides notation and values for several useful quantities in a.u. and also equivalent values in other commonly used units. Greater precision and additional data are available at http://www.physics.nist.gov/PhysRefData/contents.html.

Bibliography and Suggested Additional Reading

Cramer, C. J., Famini, G. R., and Lowrey, A. 1993. 'Use of Quantum Chemical Properties as Analogs for Solvatochromic Parameters in Structure–Activity Relationships', *Acc. Chem. Res.*, **26**, 599.

Irikura, K. K., Frurip, D. J., Eds. 1998. *Computational Thermochemistry*, American Chemical Society Symposium Series, Vol. **677**, American Chemical Society: Washington, DC.

Jensen, F. 1999. *Introduction to Computational Chemistry*, Wiley: Chichester.

Levine, I. N. 2000. *Quantum Chemistry*, 5th Edn., Prentice Hall: New York.

Truhlar, D. G. 2000. 'Perspective on "Principles for a direct SCF approach to LCAO-MO *ab initio* calculations"' *Theor. Chem. Acc.*, **103**, 349.

References

Almlöf, J., Faegri, K., Jr., and Korsell, K. 1982. *J. Comput. Chem.*, **3**, 385.

Cramer, C. J., Denmark, S. E., Miller, P. C., Dorow, R. L., Swiss, K. A., and Wilson, S. R. 1994. *J. Am. Chem. Soc.*, **116**, 2437.

2

Molecular Mechanics

2.1 History and Fundamental Assumptions

Let us return to the concept of the PES as described in Chapter 1. To a computational chemist, the PES is a surface that can be generated point by point by use of some computational method which determines a molecular energy for each point's structure. However, the concept of the PES predates any serious efforts to "compute" such surfaces. The first PESs (or slices thereof) were constructed by molecular spectroscopists.

A heterodiatomic molecule represents the simplest case for study by vibrational spectroscopy, and it also represents the simplest PES, since there is only the single degree of freedom, the bond length. Vibrational spectroscopy measures the energy separations between different vibrational levels, which are quantized. Most chemistry students are familiar with the simplest kind of vibrational spectroscopy, where allowed transitions from the vibrational ground state ($\nu = 0$) to the first vibrationally excited state ($\nu = 1$) are monitored by absorption spectroscopy; the typical photon energy for the excitation falls in the infrared region of the optical spectrum. More sensitive experimental apparati are capable of observing other allowed absorptions (or emissions) between more highly excited vibrational states, and/or forbidden transitions between states differing by more than 1 vibrational quantum number. Isotopic substitution perturbs the vibrational energy levels by changing the reduced mass of the molecule, so the number of vibrational transitions that can be observed is arithmetically related to the number of different isotopomers that can be studied. Taking all of these data together, spectroscopists are able to construct an extensive ladder of vibrational energy levels to a very high degree of accuracy (tenths of a wavenumber in favorable cases), as illustrated in Figure 2.1.

The spacings between the various vibrational energy levels depend on the potential energy associated with bond stretching (see Section 9.3.2). The data from the spectroscopic experiments thus permit the derivation of that potential energy function in a straightforward way.

Let us consider for the moment the potential energy function in an abstract form. A useful potential energy function for a bond between atoms A and B should have an analytic form. Moreover, it should be continuously differentiable. Finally, assuming the dissociation energy for the bond to be positive, we will define the minimum of the function to have a potential energy of zero; we will call the bond length at the minimum r_{eq}. We can determine the value

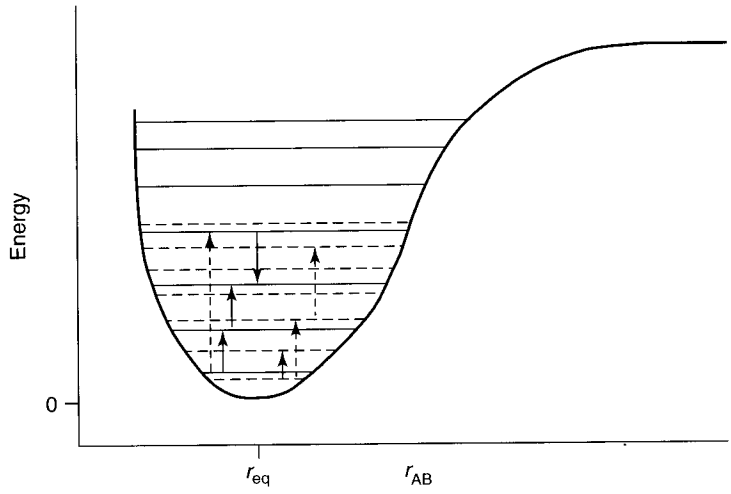


Figure 2.1 The first seven vibrational energy levels for a lighter (solid horizontal lines) and heavier (horizontal dashed lines) isotopomer of diatomic AB. Allowed vibrational transitions are indicated by solid vertical arrows, forbidden transitions are indicated by dashed vertical arrows

of the potential energy at an arbitrary point by taking a Taylor expansion about r_{eq}

$$U(r) = U(r_{eq}) + \left.\frac{dU}{dr}\right|_{r=r_{eq}} (r - r_{eq}) + \frac{1}{2!}\left.\frac{d^2U}{dr^2}\right|_{r=r_{eq}} (r - r_{eq})^2 + \frac{1}{3!}\left.\frac{d^3U}{dr^3}\right|_{r=r_{eq}} (r - r_{eq})^3 + \cdots \tag{2.1}$$

Note that the first two terms on the r.h.s. of Eq. (2.1) are zero, the first by arbitrary choice, the second by virtue of r_{eq} being the minimum. If we truncate after the first non-zero term, we have the simplest possible expression for the vibrational potential energy

$$U(r_{AB}) = \tfrac{1}{2}k_{AB}(r_{AB} - r_{AB,eq})^2 \tag{2.2}$$

where we have replaced the second derivative of U by the symbol k. Equation (2.2) is Hooke's law for a spring, where k is the 'force constant' for the spring; the same term is used for k in spectroscopy and molecular mechanics. Subscripts have been added to emphasize that force constants and equilibrium bond lengths may vary from one pair of atoms to another.

Indeed, one might expect that force constants and equilibrium lengths might vary substantially even when A and B remain constant, but the bond itself is embedded in different molecular frameworks (i.e., surroundings). However, as more and more spectroscopic data became available in the early 20th century, particularly in the area of organic chemistry, where hundreds or thousands of molecules having similar bonds (e.g., C–C single bonds)

could be characterized, it became empirically evident that the force constants and equilibrium bond lengths were largely the same from one molecule to the next. This phenomenon came to be called 'transferability'.

Concomitant with these developments in spectroscopy, thermochemists were finding that, to a reasonable approximation, molecular enthalpies could be determined as a sum of bond enthalpies. Thus, assuming transferability, if two different molecules were to be composed of identical bonds (i.e., they were to be isomers of one kind or another), the sum of the differences in the 'strains' of those bonds from one molecule to the other (which would arise from different bond lengths in the two molecules – the definition of strain in this instance is the positive deviation from the zero of energy) would allow one to predict the difference in enthalpies. Such prediction was a major goal of the emerging area of organic conformational analysis.

One might ask why any classical mechanical bond would deviate from its equilibrium bond length, insofar as that represents the zero of energy. The answer is that in polyatomic molecules, other energies of interaction must also be considered. For instance, repulsive van der Waals interactions between nearby groups may force some bonds connecting them to lengthen. The same argument can be applied to bond angles, which also have transferable force constants and optimal values (vide infra). Energetically unfavorable non-bonded, non-angle-bending interactions have come to be called 'steric effects' following the terminology suggested by Hill (1946), who proposed that a minimization of overall steric energy could be used to predict optimal structures. The first truly successful reduction of this general idea to practice was accomplished by Westheimer and Mayer (1946), who used potential energy functions to compute energy differences between twisted and planar substituted biphenyls and were able to rationalize racemization rates in these molecules.

The rest of this chapter examines the various components of the molecular energy and the force-field approaches taken for their computation. The discussion is, for the most part, general. At the end of the chapter, a comprehensive listing of reported/available force fields is provided with some description of their form and intended applicability.

2.2 Potential Energy Functional Forms

2.2.1 Bond Stretching

Before we go on to consider functional forms for all of the components of a molecule's total steric energy, let us consider the limitations of Eq. (2.2) for bond stretching. Like any truncated Taylor expansion, it works best in regions near its reference point, in this case r_{eq}. Thus, if we are interested primarily in molecular structures where no bond is terribly distorted from its optimal value, we may expect Eq. (2.2) to have reasonable utility. However, as the bond is stretched to longer and longer r, Eq. (2.2) predicts the energy to become infinitely positive, which is certainly not chemically realistic. The practical solution to such inaccuracy is to include additional terms in the Taylor expansion. Inclusion of the cubic term provides a potential energy function of the form

$$U(r_{AB}) = \tfrac{1}{2}[k_{AB} + k_{AB}^{(3)}(r_{AB} - r_{AB.eq})](r_{AB} - r_{AB.eq})^2 \tag{2.3}$$

where we have added the superscript '(3)' to the cubic force constant (also called the 'anharmonic' force constant) to emphasize that it is different from the quadratic one. The cubic force constant is negative, since its function is to reduce the overly high stretching energies predicted by Eq. (2.2). This leads to an unintended complication, however; Eq. (2.3) diverges to *negative* infinity with increasing bond length. Thus, the lowest possible energy for a molecule whose bond energies are described by functions having the form of Eq. (2.3) corresponds to all bonds being dissociated, and this can play havoc with automated minimization procedures.

Again, the simple, practical solution is to include the next term in the Taylor expansion, namely the quartic term, leading to an expression of the form

$$U(r_{AB}) = \tfrac{1}{2}[k_{AB} + k_{AB}^{(3)}(r_{AB} - r_{AB,eq}) + k_{AB}^{(4)}(r_{AB} - r_{AB,eq})^2](r_{AB} - r_{AB,eq})^2 \tag{2.4}$$

Such quartic functional forms are used in the general organic force field, MM3 (a large taxonomy of existing force fields appears at the end of the chapter). Many force fields that are designed to be used in reduced regions of chemical space (e.g., for specific biopolymers), however, use quadratic bond stretching potentials because of their greater computational simplicity.

The alert reader may wonder, at this point, why there has been no discussion of the Morse function

$$U(r_{AB}) = D_{AB}[1 - e^{-\alpha_{AB}(r_{AB} - r_{AB,eq})}]^2 \tag{2.5}$$

where D_{AB} is the dissociation energy of the bond and α_{AB} is a fitting constant. The hypothetical potential energy curve shown in Figure 2.1 can be reproduced over a much wider range of r by a Morse potential than by a quartic potential. Most force fields decline to use the Morse potential because it is computationally much less efficient to evaluate the exponential function than to evaluate a polynomial function (vide infra). Moreover, most force fields are designed to study the energetics of molecules whose various degrees of freedom are all reasonably close to their equilibrium values, say within 10 kcal/mol. Over such a range, the deviation between the Morse function and a quartic function is usually negligible.

Even in these instances, however, there is some utility to considering the Morse function. If we approximate the exponential in Eq. (2.5) as its infinite series expansion truncated at the cubic term, we have

$$\begin{aligned} U(r_{AB}) = D_{AB}\{1 - [1 - \alpha_{AB}(r_{AB} - r_{AB,eq}) + \tfrac{1}{2}\alpha_{AB}^2(r_{AB} - r_{AB,eq})^2 \\ - \tfrac{1}{6}\alpha_{AB}^3(r_{AB} - r_{AB,eq})^3]\}^2 \end{aligned} \tag{2.6}$$

Squaring the quantity in braces and keeping only terms through quartic gives

$$U(r_{AB}) = D_{AB}\left[\alpha_{AB}^2 - \alpha_{AB}^3(r_{AB} - r_{AB,eq}) + \frac{7}{12}\alpha_{AB}^4(r_{AB} - r_{AB,eq})^2\right](r_{AB} - r_{AB,eq})^2 \tag{2.7}$$

where comparison of Eqs. (2.4) and (2.7) makes clear the relationship between the various force constants and the parameters D and α of the Morse potential. In particular,

$$k_{AB} = 2\alpha_{AB}^2 D_{AB} \tag{2.8}$$

Typically, the simplest parameters to determine from experiment are k_{AB} and D_{AB}. With these two parameters available, α_{AB} can be determined from Eq. (2.8), and thus the cubic and quartic force constants can also be determined from Eqs. (2.4) and (2.7). Direct measurement of cubic and quartic force constants requires more spectral data than are available for many kinds of bonds, so this derivation facilitates parameterization. We will discuss parameterization in more detail later in the chapter, but turn now to consideration of other components of the total molecular energy.

2.2.2 Valence Angle Bending

Vibrational spectroscopy reveals that, for small displacements from equilibrium, energy variations associated with bond angle deformation are as well modeled by polynomial expansions as are variations associated with bond stretching. Thus, the typical force field function for angle strain energy is

$$U(\theta_{ABC}) = \tfrac{1}{2}[k_{ABC} + k_{ABC}^{(3)}(\theta_{ABC} - \theta_{ABC,eq}) + k_{ABC}^{(4)}(\theta_{ABC} - \theta_{ABC,eq})^2 + \cdots]$$
$$(\theta_{ABC} - \theta_{ABC,eq})^2 \tag{2.9}$$

where θ is the valence angle between bonds AB and BC (note that in a force field, a bond is *defined* to be a vector connecting two atoms, so there is no ambiguity about what is meant by an angle between two bonds), and the force constants are now subscripted ABC to emphasize that they are dependent on three atoms. Whether Eq. (2.9) is truncated at the quadratic term or whether more terms are included in the expansion depends entirely on the balance between computational simplicity and generality that any given force field chooses to strike. Thus, to note two specific examples, the general organic force field MM3 continues the expansion through to the sextic term for some ABC combinations, while the biomolecular force field of Cornell *et al.* (see Table 2.1, first row) limits itself to a quadratic expression in all instances. (Original references to all the force fields discussed in this chapter will be found in Table 2.1.)

While the above prescription for angle bending seems useful, certain issues do arise. First, note that no power expansion having the form of Eq. (2.9) will show the appropriate chemical behavior as the bond angle becomes linear, i.e., at $\theta = \pi$. Another flaw with Eq. (2.9) is that, particularly in inorganic systems, it is possible to have *multiple* equilibrium values; for instance, in the trigonal bipyramidal system PCl_5 there are stable Cl–P–Cl angles of $\pi/2$, $\pi/3$, and π for axial/equatorial, equatorial/equatorial, and axial/axial combinations of chlorine atoms, respectively. Finally, there is another kind of angle bending that is sometimes discussed in molecular systems, namely 'out-of-plane' bending. Prior to addressing these

various issues, it is instructive to consider the manner in which force fields typically handle potential energy variations associated with torsional motion.

2.2.3 Torsions

If we consider four atoms connected in sequence, ABCD, Figure 1.2 shows that a convenient means to describe the location of atom D is by means of a CD bond length, a BCD valence angle, and the torsional angle (or dihedral angle) associated with the ABCD linkage. As depicted in Figure 2.2, the torsional angle is defined as the angle between bonds AB and CD when they are projected into the plane bisecting the BC bond. The convention is to define the angle as positive if one must rotate the bond in front of the bisecting plane in a clockwise fashion to eclipse the bond behind the bisecting plane. By construction, the torsion angle is periodic. An obvious convention would be to use only the positive angle, in which case the torsion period would run from 0 to 2π radians (0 to 360°). However, the minimum energy for many torsions is for the antiperiplanar arrangement, i.e., $\omega = \pi$. Thus, the convention that $-\pi < \omega \leq \pi (-180^\circ \leq \omega \leq 180^\circ)$ also sees considerable use.

Since the torsion itself is periodic, so too must be the torsional potential energy. As such, it makes sense to model the potential energy function as an expansion of periodic functions, e.g., a Fourier series. In a general form, typical force fields use

$$U(\omega_{\mathrm{ABCD}}) = \tfrac{1}{2} \sum_{\{j\}_{\mathrm{ABCD}}} V_{j,\mathrm{ABCD}}[1 + (-1)^{j+1} \cos(j\omega_{\mathrm{ABCD}} + \psi_{j,\mathrm{ABCD}})] \tag{2.10}$$

where the values of the signed term amplitudes V_j and the set of periodicities $\{j\}$ included in the sum are specific to the torsional linkage ABCD (note that deleting a particular value of j from the evaluated set is equivalent to setting the term amplitude for that value of j

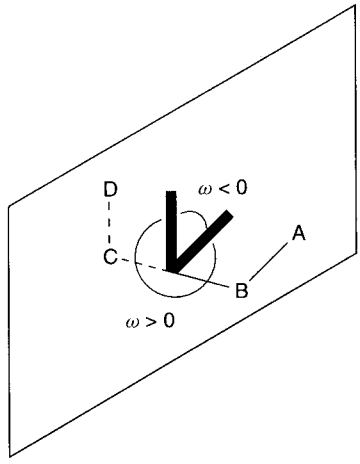


Figure 2.2 Definition and sign convention for dihedral angle ω. Note that the sign of ω is independent of whether one chooses to view the bisecting plane from the AB side or the CD side

equal to zero). Other features of Eq. (2.10) meriting note are the factor of 1/2 on the r.h.s., which is included so that the term amplitude V_j is equal to the maximum the particular term can contribute to U. The factor of $(-1)^{j+1}$ is included so that the function in brackets within the sum is zero for all j when $\omega = \pi$, if the phase angles ψ are all set to 0. This choice is motivated by the empirical observation that most (but not all) torsional energies are minimized for antiperiplanar geometries; the zero of energy for U in Eq. (2.10) thus occurs at $\omega = \pi$. Choice of phase angles ψ other than 0 permits a fine tuning of the torsional coordinate, which can be particularly useful for describing torsions in systems exhibiting large stereoelectronic effects, like the anomeric linkages in sugars (see, for instance, Woods 1996).

While the mathematical utility of Eq. (2.10) is clear, it is also well founded in a chemical sense, because the various terms can be associated with particular physical interactions when all phase angles ψ are taken equal to 0. Indeed, the magnitudes of the terms appearing in an individual fit can be informative in illuminating the degree to which those terms influence the overall rotational profile. We consider as an example the rotation about the C–O bond in fluoromethanol, the analysis of which was first described in detail by Wolfe *et al.* (1971) and Radom, Hehre and Pople (1971). Figure 2.3 shows the three-term Fourier decomposition of the complete torsional potential energy curve. Fluoromethanol is somewhat unusual insofar as the antiperiplanar structure is *not* the global minimum, although it is a local minimum. It is instructive to note the extent to which each Fourier term contributes to the overall torsional profile, and also to consider the physical factors implicit in each term.

One physical effect that would be expected to be onefold periodic in the case of fluoromethanol is the dipole–dipole interaction between the C–F bond and the O–H bond. Because of differences in electronegativity between C and F and O and H, the bond dipoles

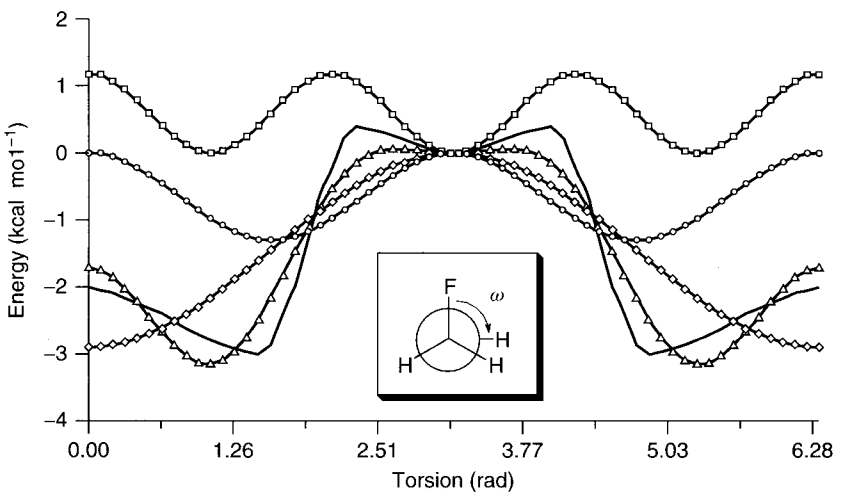


Figure 2.3 Fourier decomposition of the torsional energy for rotation about the C–O bond of fluoromethanol (bold black curve, energetics approximate). The Fourier sum (△) is composed of the onefold (◇), twofold (○), and threefold (□) periodic terms, respectively. In the Newman projection of the molecule, the oxygen atom lies behind the carbon atom at center

for these bonds point from C to F and from H to O, respectively. Thus, at $\omega = 0$, the dipoles are antiparallel (most energetically favorable) while at $\omega = \pi$ they are parallel (least energetically favorable). Thus, we would expect the V_1 term to be a minimum at $\omega = 0$, implying V_1 should be negative, and that is indeed the case. This term makes the largest contribution to the full rotational profile, having a magnitude roughly double either of the other two terms.

Twofold periodicity is associated with hyperconjugative effects. Hyperconjugation is the favorable interaction of a filled or partially filled orbital, typically a σ orbital, with a nearby empty orbital (hyperconjugation is discussed in more detail in Appendix D within the context of natural bond orbital (NBO) analysis). In the case of fluoromethanol, the filled orbital that is highest in energy is an oxygen lone pair orbital, and the empty orbital lowest in energy (and thus best able to interact in a resonance fashion with the oxygen lone pair) is the C–F σ^* antibonding orbital. Resonance between these orbitals, which is sometimes called negative hyperconjugation to distinguish it from resonance involving filled σ orbitals as donors, is favored by maximum overlap; this takes place for torsion angles of roughly $\pm\pi/2$. The contribution of this V_2 term to the overall torsional potential of fluoromethanol is roughly half that of the V_1 term, and of the expected sign.

The remaining V_3 term is associated with unfavorable bond–bond eclipsing interactions, which, for a torsion involving sp^3-hybridized carbon atoms, would be expected to show threefold periodicity. To be precise, true threefold periodicity would only be expected were each carbon atom to bear all identical substituents. Experiments suggest that fluorine and hydrogen have similar steric behavior, so we will ignore this point for the moment. As expected, the sign of the V_3 term is positive, and it has roughly equal weight to the hyperconjugative term.

[Note that, following the terminology introduced earlier, we refer to the unfavorable eclipsing of chemical bonds as a steric interaction. Since molecular mechanics in essence treats molecules as classical atomic balls (possibly charged balls, as discussed in more detail below) connected together by springs, this terminology is certainly acceptable. It should be borne in mind, however, that real atoms are most certainly not billiard balls bumping into one another with hard shells. Rather, the unfavorable steric interaction derives from exchange-repulsion between filled molecular orbitals as they come closer to one another, i.e., the effect is electronic in nature. Thus, the bromide that all energetic issues in chemistry can be analyzed as a combination of electronic and steric effects is perhaps overly complex... *all* energetic effects in chemistry, at least if we ignore nuclear chemistry, are *exclusively* electronic/electrical in nature.]

While this analysis of fluoromethanol is instructive, it must be pointed out that a number of critical issues have been either finessed or ignored. First, as can be seen in Figure 2.3, the actual rotational profile of fluoromethanol cannot be perfectly fit by restricting the Fourier decomposition to only three terms. This may sound like quibbling, since the 'perfect' fitting of an arbitrary periodic curve takes an infinite number of Fourier terms, but the poorness of the fit is actually rather severe from a chemical standpoint. This may be most readily appreciated by considering simply the four symmetry-unique stationary points – two minima and two rotational barriers. We are trying to fit their energies, but we also want their nature as stationary points to be correct, implying that we are trying to fit their first derivatives as

well (making the first derivative equal to zero defines them as stationary points). Thus, we are trying to fit eight constraints using only three variables (namely, the term amplitudes). By construction, we are actually guaranteed that 0 and π will have correct first derivatives, and that the energy value for π will be correct (since it is required to be the relative zero), but that still leaves five constraints on three variables. If we add non-zero phase angles ψ, we can do a better (but still not perfect) job.

Another major difficulty is that we have biased the system so that we can focus on a single dihedral interaction (FCOH) as being dominant, i.e., we ignored the HCOH interactions, and we picked a system where one end of the rotating bond had only a single substituent. To illustrate the complexities introduced by more substitution, consider the relatively simple case of *n*-butane (Figure 2.4). In this case, the three-term Fourier fit is in very good agreement with the full rotational profile, and certain aspects continue to make very good chemical sense. For instance, the twofold periodic term is essentially negligible, as would be expected since there are no particularly good donors or acceptors to interact in a hyperconjugative fashion. The onefold term, on the other hand, makes a very significant contribution, and this clearly cannot be assigned to some sort of dipole–dipole interaction, since the magnitude of a methylene–methyl bond dipole is very near zero. Rather, the magnitudes of the one- and threefold symmetric terms provide information about the relative steric strains associated with the two possible eclipsed structures, the lower energy of which has one H/H and two H/CH_3 eclipsing interactions, while the higher energy structure has two H/H and one CH_3/CH_3 interactions. While one might be tempted to try to derive some sort of linear combination rule for this still highly symmetric case, it should be clear that by the time one tries to analyze the torsion about a C–C bond bearing six different substituents, one's ability

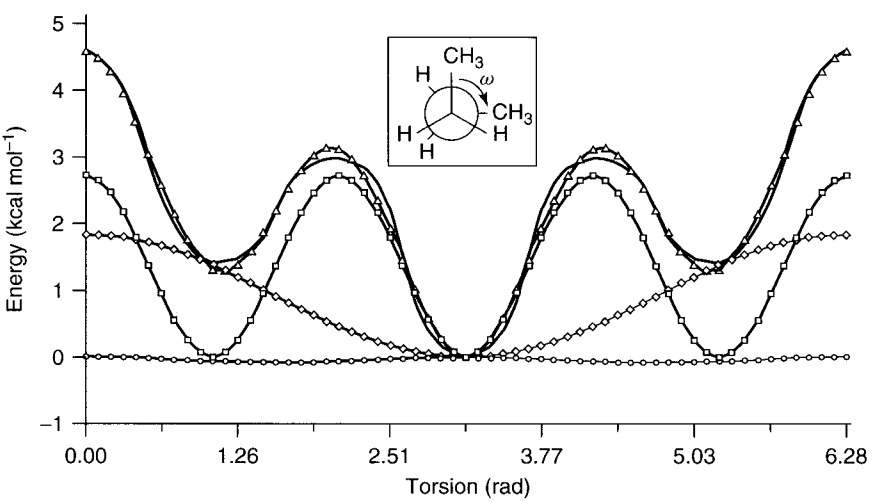


Figure 2.4 Fourier decomposition of the torsional energy for rotation about the C–C bond of *n*-butane (bold black curve, energetics approximate). The Fourier sum (△) has a close overlap, and is composed of the onefold (◇), twofold (○), and threefold (□) periodic terms, respectively

to provide a physically meaningful interpretation of the many different term amplitudes is quite limited.

Moreover, as discussed in more detail later, force field parameters are not statistically orthogonal, so optimized values can be skewed by coupling with other parameters. With all of these caveats in mind, however, there are still instances where valuable physical insights derive from a term-by-term analysis of the torsional coordinate.

Let us return now to a question raised above, namely, how to handle the valence angle bending term in a system where multiple equilibrium angles are present. Such a case is clearly analogous to the torsional energy, which also presents multiple minima. Thus, the inorganic SHAPES force field uses the following equations to compute angle bending energy

$$U(\theta_{ABC}) = \sum_{\{j\}_{ABC}} k_{j,ABC}^{Fourier}[1 + \cos(j\theta_{ABC} + \psi)] \tag{2.11}$$

$$k_{j,ABC}^{Fourier} = \frac{2k_{ABC}^{harmonic}}{j^2} \tag{2.12}$$

where ψ is a phase angle. Note that this functional form can also be used to ensure appropriate behavior in regions of bond angle inversion, i.e., where $\theta = \pi$. [As a digression, in metal coordination force fields an alternative formulation designed to handle multiple ligand–metal–ligand angles is simply to remove the angle term altogether. It is replaced by a non-bonded term specific to 1,3-interactions (a so-called 'Urey–Bradley' term) which tends to be repulsive. Thus, a given number of ligands attached to a central atom will tend to organize themselves so as to maximize the separation between any two. This 'points-on-a-sphere' (POS) approach is reminiscent of the VSEPR model of coordination chemistry.]

A separate situation, also mentioned in the angle bending discussion, arises in the case of four-atom systems where a central atom is bonded to three otherwise unconnected atoms, e.g., formaldehyde. Such systems are good examples of the second case of step IV of Figure 1.2, i.e., systems where a fourth atom is more naturally defined by a bond length to the central atom and its two bond angles to the other two atoms. However, as Figure 2.5 makes clear, one *could* define the final atom's position using the first case of step IV of Figure 1.2, i.e., by assigning a length to the central atom, an angle to a third atom, *and then a dihedral angle to the fourth atom even though atoms three and four are not defined as connected.* Such an

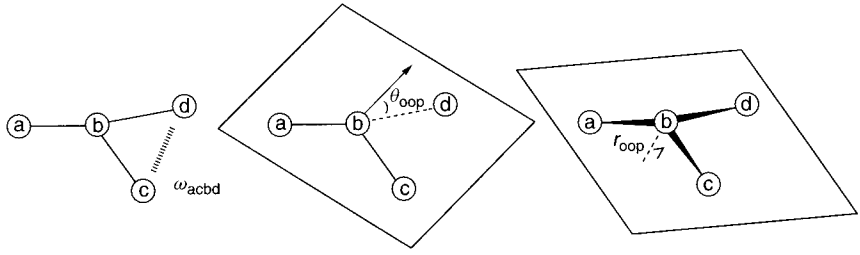


Figure 2.5 Alternative molecular coordinates that can be used to compute the energetics of distortions from planarity about a triply substituted central atom

assignment makes perfect sense from a geometric standpoint, even though it may seem odd from a chemical standpoint. Torsion angles defined in this manner are typically referred to as 'improper torsions'. In a system like formaldehyde, an improper torsion like OCHH would have a value of π radians (180°) in the planar, minimum energy structure. Increasing or decreasing this value would have the effect of moving the oxygen atom out of the plane defined by the remaining three atoms. Many force fields treat such improper torsions like any other torsion, i.e., they use Eq. (2.10). However, as Figure 2.5 indicates, the torsional description for this motion is only one of several equally reasonable coordinates that one might choose. One alternative is to quantify deviations from planarity by the angle $\theta_{\text{o.o.p.}}$ that one substituent makes with the plane defined by the other three (o.o.p. = 'out of plane'). Another is to quantify the elevation $r_{\text{o.o.p.}}$ of the central atom above/below the plane defined by the three atoms to which it is attached. Both of these latter modes have obvious connections to angle bending and bond stretching, respectively, and typically Eqs. (2.9) and (2.4), respectively, are used to model the energetics of their motion.

Let us return to the case of the butane rotational potential. As noted previously, the barriers in this potential are primarily associated with steric interactions between eclipsing atoms/groups. Anyone who has ever built a space-filling model of a sterically congested molecule is familiar with the phenomenon of steric congestion – some atomic balls in the space-filling model push against one another, creating strain (leading to the apocryphal 'drop test' metric of molecular stability: how great a height can the model be dropped from and remain intact?) Thus, in cases where dipole–dipole and hyperconjugative interactions are small about a rotating bond, one might question whether there is a need to parameterize a torsional function at all. Instead, one could represent atoms as balls, each having a characteristic radius, and develop a functional form quantifying the energetics of ball–ball interactions. Such a prescription provides an intuitive model for more distant 'non-bonded' interactions, which we now examine.

2.2.4 van der Waals Interactions

Consider the mutual approach of two noble gas atoms. At infinite separation, there is no interaction between them, and this defines the zero of potential energy. The isolated atoms are spherically symmetric, lacking any electric multipole moments. In a classical world (ignoring the chemically irrelevant gravitational interaction) there is no attractive force between them as they approach one another. When there are no dissipative forces, the relationship between force F in a given coordinate direction q and potential energy U is

$$F_q = -\frac{\partial U}{\partial q} \tag{2.13}$$

In this one-dimensional problem, saying that there is no force is equivalent to saying that the slope of the energy curve with respect to the 'bond length' coordinate is zero, so the potential energy remains zero as the two atoms approach one another. Associating non-zero size with our classical noble gas atoms, we might assign them hard-sphere radii r_{vdw}. In that case, when the bond length reaches twice the radius, the two cannot approach one another more

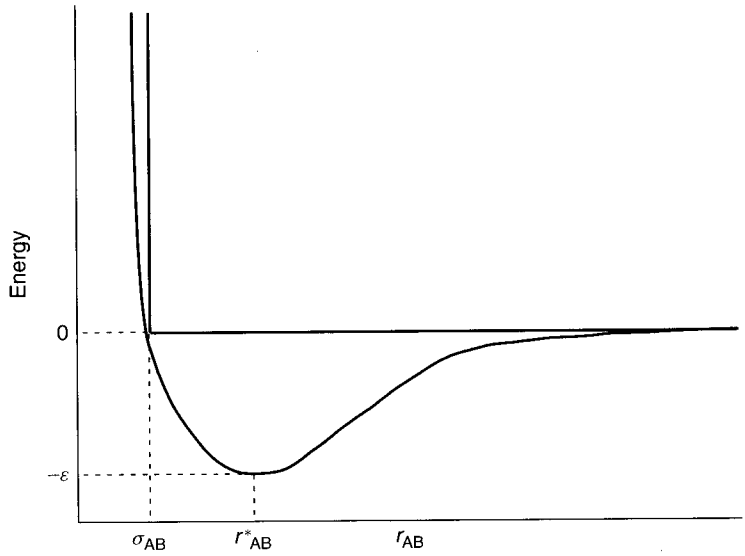


Figure 2.6 Non-attractive hard-sphere potential (straight lines) and Lennard–Jones potential (curve). Key points on the energy and bond length axes are labeled

closely, which is to say the potential energy discontinuously becomes infinite for $r < 2r_{vdw}$. This potential energy curve is illustrated in Figure 2.6.

One of the more profound manifestations of quantum mechanics is that this curve does *not* accurately describe reality. Instead, because the 'motions' of electrons are correlated (more properly, the electronic wave functions are correlated), the two atoms simultaneously develop electrical moments that are oriented so as to be mutually attractive. The force associated with this interaction is referred to variously as 'dispersion', the 'London' force, or the 'attractive van der Waals' force. In the absence of a permanent charge, the strongest such interaction is a dipole–dipole interaction, usually referred to as an 'induced dipole–induced dipole' interaction, since the moments in question are not permanent. Such an interaction has an inverse sixth power dependence on the distance between the two atoms. Thus, the potential energy becomes increasingly negative as the two noble gas atoms approach one another from infinity.

Dispersion is a fascinating phenomenon. It is sufficiently strong that even the dimer of He is found to have one bound vibrational state (Luo *et al.* 1993; with a vibrationally averaged bond length of 55 Å it is a remarkable member of the molecular bestiary). Even for molecules with fairly large *permanent* electric moments in the gas phase, dispersion is the dominant force favoring condensation to the liquid state at favorable temperatures and pressures (Reichardt 1990).

However, as the two atoms continue to approach one another, their surrounding electron densities ultimately begin to interpenetrate. In the absence of opportunities for bonding interactions, Pauli repulsion (or 'exchange repulsion') causes the energy of the system to rise rapidly with decreasing bond length. The sum of these two effects is depicted in Figure 2.6;

the contrasts with the classical hard-sphere model are that (i) an attractive region of the potential energy curve exists and (ii) the repulsive wall is not infinitely steep. [Note that at $r = 0$ the potential energy is that for an isolated atom having an atomic number equal to the sum of the atomic numbers for the two separated atoms; this can be of interest in certain formal and even certain practical situations, but we do no modeling of nuclear chemistry here.]

The simplest functional form that tends to be used in force fields to represent the combination of the dispersion and repulsion energies is

$$U(r_{AB}) = \frac{a_{AB}}{r_{AB}^{12}} - \frac{b_{AB}}{r_{AB}^{6}} \tag{2.14}$$

where a and b are constants specific to atoms A and B. Equation (2.14) defines a so-called 'Lennard–Jones' potential.

The inverse 12th power dependence of the repulsive term on interatomic separation has no theoretical justification – instead, this term offers a glimpse into the nuts and bolts of the algorithmic implementation of computational chemistry. Formally, one can more convincingly argue that the repulsive term in the non-bonded potential should have an exponential dependence on interatomic distance. However, the evaluation of the exponential function (and the log, square root, and trigonometric functions, *inter alia*) is roughly a factor of five times more costly in terms of central processing unit (cpu) time than the evaluation of the simple mathematical functions of addition, subtraction, or multiplication. Thus, the evaluation of r^{12} requires only that the theoretically justified r^6 term be multiplied by itself, which is a very cheap operation. Note moreover the happy coincidence that all terms in r involve *even* powers of r. The relationship between the internal coordinate r and Cartesian coordinates, which are typically used to specify atomic positions (see Section 2.4), is defined by

$$r_{AB} = \sqrt{(x_A - x_B)^2 + (y_A - y_B)^2 + (z_A - z_B)^2} \tag{2.15}$$

If only even powers of r are required, one avoids having to compute a square root. While quibbling over relative factors of five with respect to an operation that takes a tiny fraction of a second in absolute time may seem like overkill, one should keep in mind how many times the function in question may have to be evaluated in a given calculation. In a formal analysis, the number of non-bonded interactions that must be evaluated scales as N^2, where N is the number of atoms. In the process of optimizing a geometry, or of searching for many energy minima for a complex molecule, hundreds or thousands of energy evaluations may need to be performed for interim structures. Thus, seemingly small savings in time can be multiplied so that they are of practical importance in code development.

The form of the Lennard–Jones potential is more typically written as

$$U(r_{AB}) = 4\varepsilon_{AB}\left[\left(\frac{\sigma_{AB}}{r_{AB}}\right)^{12} - \left(\frac{\sigma_{AB}}{r_{AB}}\right)^{6}\right] \tag{2.16}$$

where the constants a and b of Eq. (2.14) are here replaced by the constants ε and σ. Inspection of Eq. (2.16) indicates that σ has units of length, and is the interatomic separation at which repulsive and attractive forces exactly balance, so that $U = 0$. If we differentiate Eq. (2.16) with respect to r_{AB}, we obtain

$$\frac{dU(r_{AB})}{dr_{AB}} = \frac{4\varepsilon_{AB}}{r_{AB}}\left[-12\left(\frac{\sigma_{AB}}{r_{AB}}\right)^{12} + 6\left(\frac{\sigma_{AB}}{r_{AB}}\right)^{6}\right] \tag{2.17}$$

Setting the derivative equal to zero in order to find the minimum in the Lennard–Jones potential gives, after rearrangement

$$r^*_{AB} = 2^{1/6}\sigma_{AB} \tag{2.18}$$

where r^* is the bond length at the minimum. If we use this value for the bond length in Eq. (2.16), we obtain $U = -\varepsilon_{AB}$, indicating that the parameter ε is the Lennard–Jones well depth (Figure 2.6).

The Lennard–Jones potential continues to be used in many force fields, particularly those targeted for use in large systems, e.g., biomolecular force fields. In more general force fields targeted at molecules of small to medium size, slightly more complicated functional forms, arguably having more physical justification, tend to be used (computational times for small molecules are so short that the efficiency of the Lennard–Jones potential is of little consequence). Such forms include the Morse potential [Eq. (2.5)] and the 'Hill' potential

$$U(r_{AB}) = \varepsilon_{AB}\left[\frac{6}{\beta_{AB} - 6}\exp\left(\beta_{AB}\frac{1 - r_{AB}}{r^*_{AB}}\right) - \frac{\beta_{AB}}{\beta_{AB} - 6}\left(\frac{r^*_{AB}}{r_{AB}}\right)^{6}\right] \tag{2.19}$$

where β is a new parameter and all other terms have the same meanings as in previous equations.

Irrespective of the functional form of the van der Waals interaction, some force fields reduce the energy computed for 1,4-related atoms (i.e., torsionally related) by a constant scale factor.

Our discussion of non-bonded interactions began with the example of two noble gas atoms having no permanent electrical moments. We now turn to a consideration of non-bonded interactions between atoms, bonds, or groups characterized by non-zero local electrical moments.

2.2.5 Electrostatic Interactions

Consider the case of two molecules A and B interacting at a reasonably large distance, each characterized by classical, non-polarizable, permanent electric moments. Classical electrostatics asserts the energy of interaction for the system to be

$$U_{AB} = \mathbf{M}^{(A)}\mathbf{V}^{(B)} \tag{2.20}$$

where $\mathbf{M}^{(A)}$ is an ordered vector of the multipole moments of A, e.g., charge (zeroth moment), x, y, and z components of the dipole moment, then the nine components of the quadrupole moment, etc., and $\mathbf{V}^{(B)}$ is a similarly ordered row vector of the electrical potentials deriving from the multipole moments of B. Both expansions are about single centers, e.g., the centers of mass of the molecules. At long distances, one can truncate the moment expansions at reasonably low order and obtain useful interaction energies.

Equation (2.20) can be used to model the behavior of a large collection of individual molecules efficiently because the electrostatic interaction energy is pairwise additive. That is, we may write

$$U = \sum_{A} \sum_{B>A} \mathbf{M}^{(A)} \mathbf{V}^{(B)} \tag{2.21}$$

However, Eq. (2.21) is not very convenient in the context of *intra*molecular electrostatic interactions. In a protein, for instance, how can one derive the electrostatic interactions between spatially adjacent amide groups (which have large local electrical moments)? In principle, one could attempt to define moment expansions for functional groups that recur with high frequency in molecules, but such an approach poses several difficulties. First, there is no good experimental way in which to measure (or even define) such local moments, making parameterization difficult at best. Furthermore, such an approach would be computationally quite intensive, as evaluation of the moment potentials is tedious. Finally, the convergence of Eq. (2.20) at short distances can be quite slow with respect to the point of truncation in the electrical moments.

Let us pause for a moment to consider the fundamental constructs we have used thus far to define a force field. We have introduced van der Waals balls we call atoms, and we have defined bonds, angles, and torsional linkages between them. What would be convenient would be to describe electrostatic interactions in some manner that is based on these available entities (this convenience derives in part from our desire to be able to optimize molecular geometries efficiently, as described in more detail below). The simplest approach is to assign to each van der Waals atom a partial charge, in which case the interaction energy between atoms A and B is simply

$$U_{AB} = \frac{q_A q_B}{\varepsilon_{AB} r_{AB}} \tag{2.22}$$

This assignment tends to follow one of three formalisms, depending on the intent of the modeling endeavor. In the simplest case, the charges are 'permanent', in the sense that all atoms of a given type are defined to carry that charge in all situations. Thus, the atomic charge is a fixed parameter.

Alternatively, the charge can be determined from a scheme that depends on the electronegativity of the atom in question, and also on the electronegativities of those atoms to which it is defined to be connected Thus, the atomic electronegativity becomes a parameter and some functional form is adopted in which it plays a role as a variable. In a force field with a reduced number of atomic 'types' (see below for more discussion of atomic types) this preserves flexibility in the recognition of different chemical environments. Such flexibility is critical for the charge because the electrostatic energy can be so large compared to other

components of the force field: Eq. (2.22) is written in a.u.; the conversion to energy units of kilocalories per mole and distance units of ångstorms involves multiplication of the r.h.s. by a factor of 332. Thus, even at 100 Å separation, the interaction energy between two unit charges in a vacuum would be more than 3 kcal/mol, which is of the same order of energy we expect for distortion of an individual stretching, bending, or torsional coordinate.

Finally, in cases where the force field is designed to study a particular molecule (i.e., generality is not an issue), the partial charges are often chosen to accurately reproduce some experimental or computed electrostatic observable of the molecule. Various schemes in common use are described in Chapter 9.

If, instead of the atom, we define charge polarization for the chemical bonds, the most convenient bond moment is the dipole moment. In this case, the interaction energy is defined between bonds AB and CD as

$$U_{AB/CD} = \frac{\mu_{AB}\mu_{CD}}{\varepsilon_{AB/CD} r_{AB/CD}^3}(\cos\chi_{AB/CD} - 3\cos\alpha_{AB}\alpha_{CD}) \tag{2.23}$$

where the bond moment vectors having magnitude μ are centered midway along the bonds and are collinear with them. The orientation vectors χ and α are defined in Figure 2.7.

Note that in Eqs. (2.22) and (2.23) the dielectric constant ε is subscripted. Although one might expect the best dielectric constant to be that for the permittivity of free space, such an assumption is not necessarily consistent with the approximations introduced by the use of atomic point charges. Instead, the dielectric constant must be viewed as a parameter of the model, and it is moreover a parameter that can take on multiple values. For use in Eq. (2.22),

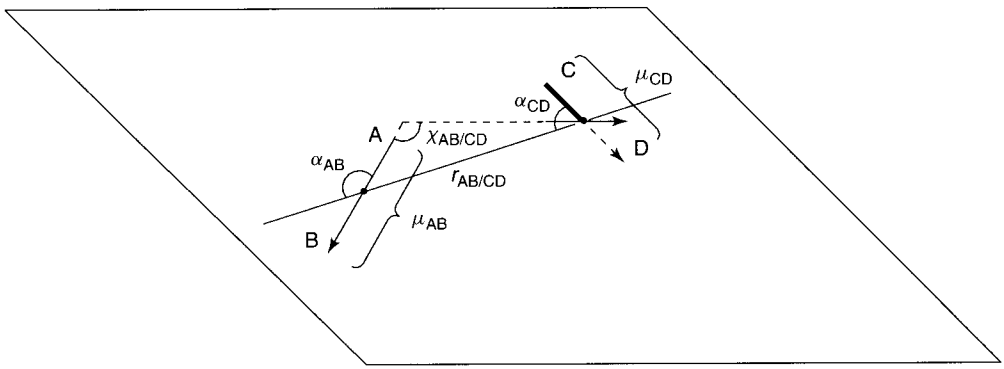


Figure 2.7 Prescription for evaluating the interaction energy between two dipoles. Each angle α is defined as the angle between the positive end of its respective dipole and the line passing through the two dipole centroids. The length of the line segment connecting the two centroids is r. To determine χ, the AB dipole and the centroid of the CD dipole are used to define a plane, and the CD dipole is projected into this plane. If the AB dipole and the projected CD dipole are parallel, χ is defined to be 0; if they are not parallel, they are extended as rays until they intersect. If the extension is from the same signed end of both dipoles, χ is the interior angle of the intersection (as illustrated), otherwise it is the exterior angle of the intersection

a plausible choice might be

$$\varepsilon_{AB} = \begin{cases} \infty & \text{if A and B are 1,2- or 1,3-related} \\ 3.0 & \text{if A and B are 1,4-related} \\ 1.5 & \text{otherwise} \end{cases} \tag{2.24}$$

which dictates that electrostatic interactions between bonded atoms or between atoms sharing a common bonded atom are not evaluated, interactions between torsionally related atoms *are* evaluated, but are reduced in magnitude by a factor of 2 relative to all other interactions, which are evaluated with a dielectric constant of 1.5. Dielectric constants can also be defined so as to have a continuous dependence on the distance between the atoms. Although one might expect the use of high dielectric constants to mimic to some extent the influence of a surrounding medium characterized by that dielectric (e.g., a solvent), this is rarely successful – more accurate approaches for including condensed-phase effects are discussed in Chapters 3, 11, and 12.

Bonds between heteroatoms and hydrogen atoms are amongst the most polar found in non-ionic systems. This polarity is largely responsible for the well-known phenomenon of hydrogen bonding, which is a favorable interaction (usually ranging from 3 to 10 kcal/mol) between a hydrogen and a heteroatom to which it is *not* formally bonded. Most force fields account for hydrogen bonding implicitly in the non-bonded terms, van der Waals and electrostatic. In some instances an additional non-bonded interaction term, in the form of a 10–12 potential, is added

$$U(r_{XH}) = \frac{a'_{XH}}{r_{XH}^{12}} - \frac{b'_{XH}}{r_{XH}^{10}} \tag{2.25}$$

where X is a heteroatom to which H is not bound. This term is analogous to a Lennard–Jones potential, but has a much more rapid decay of the attractive region with increasing bond length. Indeed, the potential well is so steep and narrow that one may regard this term as effectively forcing a hydrogen bond to deviate only very slightly from its equilibrium value.

Up to now, we have considered the interactions of *static* electric moments, but actual molecules have their electric moments *perturbed* under the influence of an electrical field (such as that deriving from the electrical moments of another molecule). That is to say, molecules are polarizable. To extend a force field to include polarizability is conceptually straightforward. Each atom is assigned a polarizability tensor. In the presence of the permanent electric field of the molecule (i.e., the field derived from the atomic charges or the bond–dipole moments), a dipole moment will be induced on each atom. Following this, however, the total electric field is the *sum* of the permanent electric field and that created by the induced dipoles, so the determination of the 'final' induced dipoles is an iterative process that must be carried out to convergence (which may be difficult to achieve). The total electrostatic energy can then be determined from the pairwise interaction of all moments and moment potentials (although the energy is determined in a pairwise fashion, note that many-body effects are incorporated by the iterative determination of the induced dipole moments). As a rough rule, computing the electrostatic interaction energy for a polarizable force field is about an order of magnitude more costly than it is for a static force field. Moreover, except for

the most accurate work in very large systems, the benefits derived from polarization appear to be small. Thus, with the possible exception of solvent molecules in condensed-phase models (see Section 12.4.1), most force fields tend to avoid including polarization.

2.2.6 Cross Terms

Bonds, angles, and torsions are not isolated molecular coordinates: they couple with one another. To appreciate this from a chemical point of view, consider BeH_2. In its preferred, linear geometry, one describes the Be hybridization as sp, i.e., each Be hybrid orbital used to bond with hydrogen has 50% 2s character and 50% 2p character. If we now decrease the bond angle, the p contribution increases until we stop at, say, a bond angle of $\pi/3$, which is the value corresponding to sp^2 hybridization. With more p character in the Be bonding hybrids, the bonds should grow longer. While this argument relies on rather basic molecular orbital theory, even from a *mechanical* standpoint, one would expect that as a bond angle is compressed, the bond lengths to the central atom will lengthen to decrease the non-bonded interactions between the terminal atoms in the sequence.

We can put this on a somewhat clearer mathematical footing by expanding the full molecular potential energy in a multi-dimensional Taylor expansion, which is a generalization of the one-dimensional case presented as Eq. (2.1). Thus

$$\begin{aligned} U(\mathbf{q}) = U(\mathbf{q}_{\text{eq}}) &+ \sum_{i=1}^{3N-6} (q_i - q_{i,\text{eq}}) \left.\frac{\partial U}{\partial q_i}\right|_{\mathbf{q}=\mathbf{q}_{\text{eq}}} \\ &+ \frac{1}{2!} \sum_{i=1}^{3N-6} \sum_{j=1}^{3N-6} (q_i - q_{i,\text{eq}})(q_j - q_{j,\text{eq}}) \left.\frac{\partial^2 U}{\partial q_i \partial q_j}\right|_{\mathbf{q}=\mathbf{q}_{\text{eq}}} \\ &+ \frac{1}{3!} \sum_{i=1}^{3N-6} \sum_{j=1}^{3N-6} \sum_{k=1}^{3N-6} (q_i - q_{i,\text{eq}})(q_j - q_{j,\text{eq}})(q_k - q_{k,\text{eq}}) \left.\frac{\partial^3 U}{\partial q_i \partial q_j \partial q_k}\right|_{\mathbf{q}=\mathbf{q}_{\text{eq}}} + \cdots \end{aligned} \tag{2.26}$$

where $\mathbf{q}$ is a molecular geometry vector of $3N - 6$ internal coordinates and the expansion is taken about an equilibrium structure. Again, the first two terms on the r.h.s. are zero by definition of U for $\mathbf{q}_{\text{eq}}$ and by virtue of all of the first derivatives being zero for an equilibrium structure. Up to this point, we have primarily discussed the 'diagonal' terms of the remaining summations, i.e., those terms for which all of the summation indices are equal to one another. However, if we imagine that index 1 of the double summation corresponds to a bond stretching coordinate, and index 2 to an angle bending coordinate, it is clear that our force field will be more 'complete' if we include energy terms like

$$U(r_{\text{AB}}, \theta_{\text{ABC}}) = \frac{1}{2} k_{\text{AB,ABC}} (r_{\text{AB}} - r_{\text{AB,eq}})(\theta_{\text{ABC}} - \theta_{\text{ABC,eq}}) \tag{2.27}$$

where $k_{\text{AB,ABC}}$ is the mixed partial derivative appearing in Eq. (2.26). Typically, the mixed partial derivative will be negligible for degrees of freedom that do not share common atoms.

In general force fields, stretch–stretch terms can be useful in modeling systems characterized by π conjugation. In amides, for instance, the coupling force constant between CO and CN stretching has been found to be roughly 15% as large as the respective diagonal bond-stretch force constants (Fogarasi and Balázs, 1985). Stretch–bend coupling terms tend to be most useful in highly strained systems, and for the computation of vibrational frequencies (see Chapter 9). Stretch–torsion coupling can be useful in systems where eclipsing interactions lead to high degrees of strain. The coupling has the form

$$U(r_{BC}, \omega_{ABCD}) = \tfrac{1}{2}k_{BC,ABCD}(r_{BC} - r_{BC,eq})[1 + \cos(j\omega + \psi)] \tag{2.28}$$

where j is the periodicity of the torsional term and ψ is a phase angle. Thus, if the term were designed to capture extra strain involving eclipsing interactions in a substituted ethane, the periodicity would require $j = 3$ and the phase angle would be 0. Note that the stretching bond, BC, is the *central* bond in the torsional linkage.

Other useful coupling terms include stretch–stretch coupling (typically between two adjacent bonds) and bend–bend coupling (typically between two angles sharing a common central atom). In force fields that aim for spectroscopic accuracy, i.e., the reproduction of vibrational spectra, still higher order coupling terms are often included. However, for purposes of general molecular modeling, they are typically not used.

2.2.7 Parameterization Strategies

At this stage, it is worth emphasizing the possibly obvious point that a force field is nothing but a (possibly very large) collection of functional forms and associated constants. With that collection in hand, the energy of a given molecule (whose atomic connectivity must in general be specified) can be evaluated by computing the energy associated with every defined type of interaction occurring in the molecule. Because there are typically a rather large number of such interactions, the process is facilitated by the use of a digital computer, but the mathematics is really extraordinarily simple and straightforward.

Thus, we have detailed how to construct a molecular PES as a sum of energies from chemically intuitive functional forms that depend on internal coordinates and on atomic (and possibly bond-specific) properties. However, we have not paid much attention to the individual parameters appearing in those functional forms (force constants, equilibrium coordinate values, phase angles, etc.) other than pointing out the relationship of many of them to certain spectroscopically measurable quantities. Let us now look more closely at the 'Art and Science' of the parameterization process.

In an abstract sense, parameterization can be a very well-defined process. The goal is to develop a model that reproduces experimental measurements to as high a degree as possible. Thus, step 1 of parameterization is to assemble the experimental data. For molecular mechanics, these data consist of structural data, energetic data, and, possibly, data on molecular electric moments. We will discuss the issues associated with each kind of datum further below, but for the moment let us proceed abstractly. We next need to define a 'penalty function', that is, a function that provides a measure of how much deviation there is between

our predicted values and our experimental values. Our goal will then be to select force-field parameters that minimize the penalty function. Choice of a penalty function is necessarily completely arbitrary. One example of such a function is

$$Z = \left[\sum_{i}^{\text{Observables}} \sum_{j}^{\text{Occurrences}} \frac{(\text{calc}_{i,j} - \text{expt}_{i,j})^2}{w_i^2} \right]^{1/2} \tag{2.29}$$

where observables might include bond lengths, bond angles, torsion angles, heats of formation, neutral molecular dipole moments, etc., and the weighting factors w carry units (so as to make Z dimensionless) and take into account not only possibly different numbers of data for different observables, but also the degree of tolerance the penalty function will have for the deviation of calculation from experiment for those observables. Thus, for instance, one might choose the weights so as to tolerate equally 0.01 Å deviations in bond lengths, 1° deviations in bond angles, 5° deviations in dihedral angles, 2 kcal/mol deviations in heats of formation, and 0.3 D deviations in dipole moment. Note that Z is evaluated using optimized geometries for all molecules; geometry optimization is discussed in Section 2.4. Minimization of Z is a typical problem in applied mathematics, and any number of statistical or quasi-statistical techniques can be used (see, for example, Schlick 1992). The minimization approach taken, however, is rarely able to remove the chemist and his or her intuition from the process.

To elaborate on this point, first consider the challenge for a force field designed to be general over the periodic table – or, for ease of discussion, over the first 100 elements. The number of unique bonds that can be formed from any two elements is 5050. If we were to operate under the assumption that bond-stretch force constants depend only on the atomic numbers of the bonded atoms (e.g., to make no distinction between so-called single, double, triple, etc. bonds), we would require 5050 force constants and 5050 equilibrium bond lengths to complete our force field. Similarly, we would require 100 partial atomic charges, and 5050 each values of σ and ε if we use Coulomb's law for electrostatics and a Lennard–Jones formalism for van der Waals interactions. If we carry out the same sort of analysis for bond angles, we need on the order of 10^6 parameters to complete the force field. Finally, in the case of torsions, somewhere on the order of 10^8 different terms are needed. If we include coupling terms, yet more constants are introduced.

Since one is unlikely to have access to 100 000 000+ relevant experimental data, minimization of Z is an underdetermined process, and in such a case there will be many different combinations of parameter values that give similar Z values. What combination is optimal? Chemical knowledge can facilitate the process of settling on a single set of parameters. For instance, a set of parameters that involved fluorine atoms being assigned a partial positive charge would seem chemically unreasonable. Similarly, a quick glance at many force constants and equilibrium coordinate values would rapidly eliminate cases with abnormally large or small values. Another approach that introduces the chemist is making the optimization process stepwise. One optimizes some parameters over a smaller data set, then holds those parameters frozen while optimizing others over a larger data set, and this process goes on until all parameters have been chosen. The process of choosing which parameters

to optimize in which order is as arbitrary as the choice of a penalty function, but may be justified with chemical reasoning.

Now, one might argue that no one would be foolish enough to attempt to design a force field that would be completely general over the first 100 elements. Perhaps if we were to restrict ourselves to organic molecules composed of {H, C, N, O, F, Si, P, Cl, Br, and I} – which certainly encompasses a large range of interesting molecules – then we could ameliorate the data sparsity problem. In principle, this is true, but in practice, the results are not very satisfactory. When large quantities of data are in hand, it becomes quite clear that atomic 'types' cannot be defined by atomic number alone. Thus, for instance, bonds involving two C atoms fall into at least four classes, each one characterized by its own particular stretching force constant and equilibrium distance (e.g., single, aromatic, double, and triple). A similar situation obtains for any pair of atoms when multiple bonding is an option. Different atomic hybridizations give rise to different angle bending equilibrium values. The same is true for torsional terms. If one wants to include metals, usually different oxidation states give rise to differences in structural and energetic properties (indeed, this segregation of compounds based on similar, discrete properties is what inorganic chemists sometimes use to *assign* oxidation state).

Thus, in order to improve accuracy, a given force field may have a very large number of atom types, even though it includes only a relatively modest number of nuclei. The primarily organic force fields MM3 and MMFF have 153 and 99 atom types, respectively. The two general biomolecular force fields (proteins, nucleic acids, carbohydrates) OPLS (optimized potentials for liquid simulations) and of Cornell *et al.* have 41 atoms types each. The completely general (i.e., most of the periodic table) universal force field (UFF) has 126 atom types. So, again, the chemist typically faces an underdetermined optimization of parameter values in finalizing the force field.

So, what steps can be taken to decrease the scope of the problem? One approach is to make certain parameters that depend on more than one atom themselves functions of single-atom-specific parameters. For instance, for use in Eq. (2.16), one usually defines

$$\sigma_{AB} = \sigma_A + \sigma_B \tag{2.30}$$

and

$$\varepsilon_{AB} = (\varepsilon_A \varepsilon_B)^{1/2} \tag{2.31}$$

thereby reducing in each case the need for $N(N+1)/2$ diatomic parameters to only N atomic parameters. [Indeed, truly general force fields, like DREIDING, UFF, and VALBOND attempt to reduce almost all parameters to being derivable from a fairly small set of atomic parameters. In practice, these force fields are not very robust, but as their limitations continue to be addressed, they have good long-range potential for broad, general utility.]

Another approach that is conceptually similar is to make certain constants depend on bond order or bond hybridization. Thus, for instance, in the VALBOND force field, angle bending energies at metal atoms are computed from orbital properties of the metal–ligand bonds; in the MM2 and MM3 force fields, stretching force constants, equilibrium bond lengths, and two-fold torsional terms depend on computed π bond orders between atoms.

Such additions to the force field somewhat strain the limits of a 'classical' model, since references to orbitals or computed bond orders necessarily introduce quantum mechanical aspects to the calculation. There is, of course, nothing wrong with moving the model in this direction – aesthetics and accuracy are orthogonal concepts – but such QM enhancements add to model complexity and increase the computational cost.

Yet another way to minimize the number of parameters required is to adopt a so-called 'united-atom' (UA) model. That is, instead of defining only atoms as the fundamental units of the force field, one also defines certain functional groups, usually hydrocarbon groups, e.g., methyl, methylene, aryl CH, etc. The group has its own single set of non-bonded and other parameters – effectively, this reduces the total number of atoms by one less than the total number incorporated into the united atom group.

Even with the various simplifications one may envision to reduce the number of parameters needed, a vast number remain for which experimental data may be too sparse to permit reliable parameterization (thus, for example, the MMFF94 force field has about 9000 defined parameters). How does one find the best parameter values? There are three typical responses to this problem.

The most common response nowadays is to supplement the experimental data with the highest quality *ab initio* data that can be had (either from molecular orbital or density functional calculations). A pleasant feature of using theoretical data is that one can compare regions on a PES that are far from equilibrium structures by direct computation rather than by trying to interpret vibrational spectra. Furthermore, one can attempt to make force-field energy derivatives correspond to those computed *ab initio*. The only limitation to this approach is the computational resources that are required to ensure that the *ab initio* data are sufficiently accurate.

The next most sensible response is to do nothing, and accept that there will be some molecules whose connectivity places them outside the range of chemical space to which the force field can be applied. While this can be very frustrating for the general user (typically the software package delivers a message to the effect that one or more parameters are lacking and then quits), if the situation merits, the necessary new parameters can be determined in relatively short order. Far more objectionable, when not well described, is the third response, which is to estimate missing parameter values and then carry on. The estimation process can be highly suspect, and unwary users can be returned nonsense results with no indication that some parameters were guessed at. If one suspects that a particular linkage or linkages in one's molecule may be outside the well-parameterized bounds of the force field, it is always wise to run a few test calculations on structures having small to moderate distortions of those linkages so as to evaluate the quality of the force constants employed.

It is worth noting that sometimes parameter estimation takes place 'on-the-fly'. That is, the program is designed to guess without human intervention parameters that were not explicitly coded. This is a somewhat pernicious aspect of so-called graphical user interfaces (GUIs): while they make the submission of a calculation blissfully simple – all one has to do is draw the structure – one is rather far removed from knowing what is taking place in the process of the calculation. Ideally, prominent warnings from the software should accompany any results derived from such calculations.

2.3 Force-field Energies and Thermodynamics

We have alluded above that one measure of the accuracy of a force field can be its ability to predict heats of formation. A careful inspection of all of the formulas presented thus far, however, should make it clear that we have not yet established any kind of connection between the force-field energy and any kind of thermodynamic quantity.

Let us review again the sense of Eqs. (2.4) and (2.9). In both instances, the minimum value for the energy is zero (assuming positive force constants and sensible behavior for odd power terms). An energy of zero is obtained when the bond length or angle adopts its equilibrium value. Thus, a 'strain-free' molecule is one in which every coordinate adopts its equilibrium value. Although we accepted a negative torsional term in our fluoromethanol example above, because it provided some chemical insight, by proper choice of phase angles in Eq. (2.10) we could also require this energy to have zero as a minimum (although not necessarily for the dihedral angle $\omega = \pi$). So, neglecting non-bonded terms for the moment, we see that the raw force-field energy can be called the 'strain energy', since it represents the positive deviation from a hypothetical strain-free system.

The key point that must be noted here is that strain energies for two different molecules *cannot be meaningfully compared unless the zero of energy is identical.* This is probably best illustrated with a chemical example. Consider a comparison of the molecules ethanol and dimethyl ether using the MM2(91) force field. Both have the chemical formula C_2H_6O. However, while ethanol is defined by the force field to be composed of two sp^3 carbon atoms, one sp^3 oxygen atom, five carbon-bound hydrogen atoms, and one alcohol hydrogen atom, dimethyl ether differs in that all six of its hydrogen atoms are of the carbon-bound type. Each strain energy will thus be computed relative to a different hypothetical reference system, and there is no *a priori* reason that the two hypothetical systems should be thermodynamically equivalent.

What is necessary to compute a heat of formation, then, is to define the heat of formation of each hypothetical, unstrained atom type. The molecular heat of formation can then be computed as the sum of the heats of formation of all of the atom types plus the strain energy. Assigning atom-type heats of formation can be accomplished using additivity methods originally developed for organic functional groups (Cohen and Benson 1993). The process is typically iterative in conjunction with parameter determination.

Since the assignment of the atomic heats of formation is really just an aspect of parameterization, it should be clear that the possibility of a negative force-field energy, which could derive from addition of net negative non-bonded interaction energies to small non-negative strain energies, is not a complication. Thus, a typical force-field energy calculation will report any or all of (i) a strain energy, which is the energetic consequence of the deviation of the internal molecular coordinates from their equilibrium values, (ii) a force-field energy, which is the sum of the strain energy and the non-bonded interaction energies, and (iii) a heat of formation, which is the sum of the force-field energy and the reference heats of formation for the constituent atom types (Figure 2.8).

For some atom types, thermodynamic data may be lacking to assign a reference heat of formation. When a molecule contains one or more of these atom types, the force field cannot

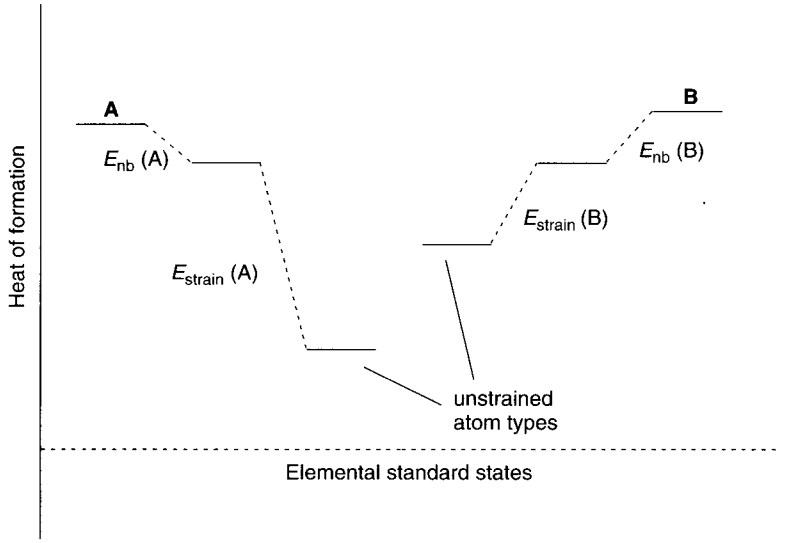


Figure 2.8 Molecules **A** and **B** are chemical isomers but are composed of different atomic types (atomic typomers?). Thus, the sums of the heats of formation of their respective unstrained atom types, which serve as their zeroes of force-field energy, are different. To each zero, strain energy and non-bonded energy (the sum of which are force-field energy) are added to determine heat of formation. In this example, note that **A** is predicted to have a lower heat of formation than **B** even though it has a substantially larger strain energy (and force-field energy); this difference is more than offset by the difference in the reference zeroes

compute a molecular heat of formation, *and energetic comparisons are necessarily limited to conformers, or other isomers that can be formed without any change in atom types.*

2.4 Geometry Optimization

One of the key motivations in early force-field design was the development of an energy functional that would permit facile optimization of molecular geometries. While the energy of an *arbitrary* structure can be interesting, real molecules vibrate thermally about their equilibrium structures, so finding minimum energy structures is key to describing equilibrium constants, comparing to experiment, etc. Thus, as emphasized above, one priority in force-field development is to adopt reasonably simple functional forms so as to facilitate geometry optimization. We now examine the optimization process in order to see how the functional forms enter into the problem.

2.4.1 Optimization Algorithms

Note that, in principle, geometry optimization could be a separate chapter of this text. In its essence, geometry optimization is a problem in applied mathematics. How does one find a minimum in an arbitrary function of many variables? [Indeed, we have already discussed that

problem once, in the context of parameter optimization. In the case of parameter optimization, however, it is not necessarily obvious how the penalty function being minimized *depends* on any given variable, and moreover the problem is highly underdetermined. In the case of geometry optimization, we are working with far fewer variables (the geometric degrees of freedom) and have, at least with a force field, analytic expressions for how the energy depends on the variables. The mathematical approach can thus be quite different.] As the problem is general, so, too, many of the details presented below will be general to any energy functional. However, certain special considerations associated with force-field calculations merit discussion, and so we will proceed first with an overview of geometry optimization, and then examine force-field specific aspects.

Because this text is designed primarily to illuminate the conceptual aspects of computational chemistry, and not to provide detailed descriptions of algorithms, we will examine only the most basic procedures. Much more detailed treatises of more sophisticated algorithms are available (see, for instance, Jensen 1999).

For pedagogical purposes, let us begin by considering a case where we do not know how our energy depends on the geometric coordinates of our molecule. To optimize the geometry, all we can do is keep trying different geometries until we are reasonably sure that we have found the one with the lowest possible energy (while this situation is atypical with force fields, there are still many sophisticated electronic structure methods for which it is indeed the only way to optimize the structure). How can one most efficiently survey different geometries?

It is easiest to proceed by considering a one-dimensional case, i.e., a diatomic with only the bond length as a geometric degree of freedom. One selects a bond length, and computes the energy. One then changes the bond length, let us say by shortening it 0.2 Å, and again computes the energy. If the energy goes down, we want to continue moving the bond length in that direction, and we should take another step (which need not necessarily be of the same length). If the energy goes up, on the other hand, we are moving in the wrong direction, and we should take a step in the opposite direction. Ultimately, the process will provide three adjacent points where the one in the center is lower in energy than the other two. Three non-collinear points uniquely define a parabola, and in this case the parabola must have a minimum (since the central point was lower in energy than the other two). We next calculate the energy for the bond length corresponding to the parabolic minimum (the degree to which the computed energy agrees with that from the parabolic equation will be an indication of how nearly harmonic the local bond stretching coordinate is). We again step left and right on the bond stretching coordinate, this time with smaller steps (perhaps an order of magnitude smaller) and repeat the parabolic fitting process. This procedure can be repeated until we are satisfied that our step size falls below some arbitrary threshold we have established as defining convergence of the geometry. Note that one can certainly envision variations on this theme. One could use more than three points in order to fit to higher order polynomial equations, step sizes could be adjusted based on knowledge of previous points, etc.

In the multi-dimensional case, the simplest generalization of this procedure is to carry out the process iteratively. Thus, for LiOH, for example, we might first find a parabolic minimum for the OH bond, then for the LiO bond, then for the LiOH bond angle (in each case holding

the other two degrees of freedom fixed), and then repeat the process to convergence. Of course, if there is strong coupling between the various degrees of freedom, this process will converge rather slowly.

What we really want to do at any given point in the multi-dimensional case is move not in the direction of a *single* coordinate, but rather in the direction of the greatest downward slope in the energy with respect to *all* coordinates. This direction is the opposite of the gradient vector, **g**, which is defined as

$$\mathbf{g}(\mathbf{q}) = \begin{bmatrix} \dfrac{\partial U}{\partial q_1} \\ \dfrac{\partial U}{\partial q_2} \\ \dfrac{\partial U}{\partial q_3} \\ \vdots \\ \dfrac{\partial U}{\partial q_n} \end{bmatrix} \tag{2.32}$$

where **q** is an n-dimensional coordinate vector ($n = 3N - 6$ where N is the number of atoms if we are working in internal coordinates, $n = 3N$ if we are working in cartesian coordinates, etc.) If we cannot compute the partial derivatives that make up **g** analytically, we can do so numerically. However, that numerical evaluation requires at least one additional energy calculation for each degree of freedom. Thus, we would increase (or decrease) every degree of freedom by some step size, compute the slope of the resulting line derived from the energies of our initial structure and the perturbed structure, and use this slope as an estimate for the partial derivative. Such a 'forward difference' estimation is typically not very accurate, and it would be better to take an additional point in the opposite direction for each degree of freedom, and then compute the 'central difference' slope from the corresponding parabola. It should be obvious that, as the number of degrees of freedom increases, it can be particularly valuable to have an energy function for which the first derivative is known *analytically*.

Let us examine this point a bit more closely for the force-field case. For this example, we will work in cartesian coordinates, in which case $\mathbf{q} = \mathbf{X}$ of Eq. (1.4). To compute, say, the partial derivative of the energy with respect to the x coordinate of atom A, we will need to evaluate the changes in energy for the various terms contributing to the full force-field energy as a function of moving atom A in the x direction. For simplicity, let us consider only the bond stretching terms. Clearly, only the energy of those bonds that have A at one terminus will be affected by A's movement. We may then use the chain rule to write

$$\frac{\partial U}{\partial x_A} = \sum_{\substack{i \text{ bonded} \\ \text{to A}}} \frac{\partial U}{\partial r_{Ai}} \frac{\partial r_{Ai}}{\partial x_A} \tag{2.33}$$

Differentiation of E with respect to r_{Ai} for Eq. (2.4) gives

$$\frac{\partial U}{\partial r_{Ai}} = \tfrac{1}{2}[2k_{Ai} + 3k_{Ai}^{(3)}(r_{Ai} - r_{Ai,eq}) + 4k_{Ai}^{(4)}(r_{Ai} - r_{Ai,eq})^2](r_{Ai} - r_{Ai,eq}) \tag{2.34}$$

The bond length r_{Ai} was defined in Eq. 2.15, and its partial derivative with respect to x_A is

$$\frac{\partial r_{Ai}}{\partial x_A} = \frac{(x_A - x_i)}{\sqrt{(x_A - x_i)^2 + (y_A - y_i)^2 + (z_A - z_i)^2}} \tag{2.35}$$

Thus, we may quickly assemble the bond stretching contributions to this particular component of the gradient. Contributions from the other terms in the force field can be somewhat more tedious to derive, but are nevertheless available analytically. This makes force fields highly efficient for the optimization of geometries of very large systems.

With $\mathbf{g}$ in hand, we can proceed in a fashion analogous to the one-dimensional case outlined above. We step along the direction defined by $-\mathbf{g}$ until we locate a minimum in the energy for this process; since we are taking points in a linear fashion, this movement is called a 'line search' (even though we may identify our minimum by fitting our points to a polynomial curve). Then, we recompute $\mathbf{g}$ at the located minimum and repeat the process. Our new search direction is necessarily orthogonal to our last one, since we minimized E in the last direction. This particular feature of a steepest descent curve can lead to *very* slow convergence in unfavorable cases.

A more robust method is the Newton–Raphson procedure. In Eq. (2.26), we expressed the full force-field energy as a multidimensional Taylor expansion in arbitrary coordinates. If we rewrite this expression in matrix notation, and truncate at second order, we have

$$U(\mathbf{q}^{(k+1)}) = U(\mathbf{q}^{(k)}) + (\mathbf{q}^{(k+1)} - \mathbf{q}^{(k)})\mathbf{g}^{(k)} + \tfrac{1}{2}(\mathbf{q}^{(k+1)} - \mathbf{q}^{(k)})^{\dagger}\mathbf{H}^{(k)}(\mathbf{q}^{(k+1)} - \mathbf{q}^{(k)}) \tag{2.36}$$

where the reference point is $\mathbf{q}^{(k)}$, $\mathbf{g}^{(k)}$ is the gradient vector for the reference point as defined by Eq. (2.32), and $\mathbf{H}^{(k)}$ is the 'Hessian' matrix for the reference point, whose elements are defined by

$$H_{ij}^{(k)} = \left.\frac{\partial^2 U}{\partial q_i \partial q_j}\right|_{\mathbf{q}=\mathbf{q}^{(k)}} \tag{2.37}$$

If we differentiate Eq. (2.36) term by term with respect to the ith coordinate of $\mathbf{q}^{(k+1)}$, noting that no term associated with point k has any dependence on a coordinate of point $k+1$ (and hence the relevant partial derivative will be 0), we obtain

$$\begin{aligned}\frac{\partial U(\mathbf{q}^{(k+1)})}{\partial q_i^{k+1}} = {} & \frac{\partial \mathbf{q}^{(k+1)}}{\partial q_i^{k+1}}\mathbf{g}^{(k)} + \frac{1}{2}\frac{\partial \mathbf{q}^{(k+1)^{\dagger}}}{\partial q_i^{k+1}}\mathbf{H}^{(k)}(\mathbf{q}^{(k+1)} - \mathbf{q}^{(k)}) \\ & + \frac{1}{2}(\mathbf{q}^{(k+1)} - \mathbf{q}^{(k)})^{\dagger}\mathbf{H}^{(k)}\frac{\partial \mathbf{q}^{(k+1)}}{\partial q_i^{k+1}}\end{aligned} \tag{2.38}$$

The l.h.s. of Eq. (2.38) is the ith element of the vector $\mathbf{g}^{(k+1)}$. On the r.h.s. of Eq. (2.38), since the partial derivative of $\mathbf{q}$ with respect to its ith coordinate is simply the unit vector in the ith coordinate direction, the various matrix multiplications simply produce the ith element of the multiplied vectors. Because mixed partial derivative values are independent of the order of differentiation, the Hessian matrix is Hermitian, and we may simplify Eq. (2.38) as

$$g_i^{(k+1)} = g_i^{(k)} + [\mathbf{H}^{(k)}(\mathbf{q}^{(k+1)} - \mathbf{q}^{(k)})]_i \tag{2.39}$$

where the notation $[]_i$ indicates the ith element of the product column matrix. The condition for a stationary point is that the l.h.s. of Eq. (2.39) be 0 for *all* coordinates, or

$$\mathbf{0} = \mathbf{g}^{(k)} + \mathbf{H}^{(k)}(\mathbf{q}^{(k+1)} - \mathbf{q}^{(k)}) \tag{2.40}$$

which may be rearranged to

$$\mathbf{q}^{(k+1)} = \mathbf{q}^{(k)} - (\mathbf{H}^{(k)})^{-1}\mathbf{g}^{(k)} \tag{2.41}$$

This equation provides a prescription for the location of stationary points. In principle, starting from an arbitrary structure having coordinates $\mathbf{q}^{(k)}$, one would compute its gradient vector $\mathbf{g}$ and its Hessian matrix $\mathbf{H}$, and then select a new geometry $\mathbf{q}^{(k+1)}$ according to Eq. (2.41). Equation (2.40) shows that the gradient vector for this new structure will be the $\mathbf{0}$ vector, so we will have a stationary point.

Recall, however, that our derivation involved a truncation of the full Taylor expansion at second order. Thus, Eq. (2.40) is only approximate, and $\mathbf{g}^{(k+1)}$ will not necessarily be $\mathbf{0}$. However, it will probably be smaller than $\mathbf{g}^{(k)}$, so we can repeat the whole process to pick a point $k + 2$. After a sufficient number of iterations, the gradient will hopefully become so small that structures $k + n$ and $k + n + 1$ differ by a chemically insignificant amount, and we declare our geometry to be converged.

There are a few points with respect to this procedure that merit discussion. First, there is the Hessian matrix. With n^2 elements, where n is the number of coordinates in the molecular geometry vector, it can grow somewhat expensive to construct this matrix at every step even for functions, like those used in most force fields, that have fairly simple analytical expressions for their second derivatives. Moreover, the matrix must be *inverted* at every step, and matrix inversion formally scales as n^3, where n is the dimensionality of the matrix. Thus, for purposes of efficiency (or in cases where analytic second derivatives are simply not available) approximate Hessian matrices are often used in the optimization process – after all, the truncation of the Taylor expansion renders the Newton–Raphson method *intrinsically* approximate. As an optimization progresses, second derivatives can be estimated reasonably well from finite differences in the analytic first derivatives over the last few steps. For the first step, however, this is not an option, and one typically either accepts the cost of computing an initial Hessian analytically for the level of theory in use, or one employs a Hessian obtained at a less expensive level of theory, when such levels are available (which is typically *not* the case for force fields). To speed up slowly convergent optimizations, it is often helpful to compute an analytic Hessian every few steps and replace

the approximate one in use up to that point. For *really* tricky cases (e.g., where the PES is fairly flat in many directions) one is occasionally forced to compute an analytic Hessian for *every* step.

Another key issue to note is that Eq. (2.41) provides a prescription to get to what is usually the *nearest* stationary point, but there is no guarantee that that point will be a *minimum*. The condition for a minimum is that all coordinate second derivatives (i.e., all diagonal elements of the Hessian matrix) be positive, but Eq. (2.41) places no constraints on the second derivatives. Thus, if one starts with a geometry that is very near a transition state (TS) structure, the Newton–Raphson procedure is likely to converge to that structure. This can be a pleasant feature, if one is looking for the TS in question, or an annoying one, if one is not. To verify the nature of a located stationary point, it is necessary to compute an accurate Hessian matrix and inspect its eigenvalues, as discussed in more detail in Chapter 9. With force fields, it is often cheaper and equally effective simply to 'kick' the structure, which is to say, by hand one moves one or a few atoms to reasonably distorted locations and then reoptimizes to verify that the original structure is again found as the lowest energy structure nearby.

Because of the importance of TS structures, a large number of more sophisticated methods exist to locate them. Many of these methods require that two minima be specified that the TS structure should 'connect', i.e., the TS structure intervenes in some reaction path that connects them. Within a given choice of coordinates, intermediate structures are evaluated and, hopefully, the relevant stationary point is located. Other methods allow the specification of a particular coordinate with respect to which the energy is to be maximized while minimizing it with respect to all other coordinates. When this coordinate is one of the normal modes of the molecule, this defines a TS structure. The bottom line for all TS structure location methods is that they work best when the chemist can provide a reasonably good initial guess for the structure, and they tend to be considerably more sensitive to the availability of a good Hessian matrix, since finding the TS essentially amounts to distinguishing between different local curvatures on the PES.

Most modern computational chemistry software packages provide some discussion of the relative merits of the various optimizers that they make available, at least on the level of providing practical advice (particularly where the user can set certain variables in the optimization algorithm with respect to step size between structures, tolerances, etc.), so we will not try to cover all possible tricks and tweaks here. We will simply note that it is usually a good idea to visualize the structures in an optimization as it progresses, as every algorithm can sometimes take a pathologically bad step, and it is usually better to restart the calculation with an improved guess than it is to wait and hope that the optimization ultimately returns to normalcy.

A final point to be made is that most optimizers are rather good at getting you to the *nearest* minimum, but an individual researcher may be interested in finding the *global* minimum (i.e., the minimum having the lowest energy of all minima). Again, this is a problem in applied mathematics for which no one solution is optimal (see, for instance, Leach 1991). Most methods involve a systematic or random sampling of alternative conformations, and this subject will be discussed further in the next chapter.

2.4.2 Optimization Aspects Specific to Force Fields

Because of their utility for very large systems, where their relative speed proves advantageous, force fields present several specific issues with respect to practical geometry optimization that merit discussion. Most of these issues revolve around the scaling behavior that the speed of a force-field calculation exhibits with respect to increasing system size. Although we raise the issues here in the context of geometry optimization, they are equally important in force-field simulations, which are discussed in more detail in the next chapter.

If we look at the scaling behavior of the various terms in a typical force field, we see that the internal coordinates have very favorably scaling – the number of internal coordinates is $3N - 6$, which is linear in N. The non-bonded terms, on the other hand, are computed from pairwise interactions, and therefore scale as N^2. However, this scaling assumes the evaluation of *all* pairwise terms. If we consider the Lennard–Jones potential, its long-range behavior decays proportional to r^{-6}. The total number of interactions should grow at most as r^2 (i.e., proportional to the surface area of a surrounding sphere), so the net energetic contribution should decay with an r^{-4} dependence. This quickly becomes negligible (particularly from a gradient standpoint) so force fields usually employ a 'cut-off' range for the evaluation of van der Waals energies – a typical choice is 10 Å. Thus, part of the calculation involves the periodic updating of a 'pair list', which is a list of all atoms for which the Lennard–Jones interaction needs to be calculated. The update usually occurs only once every several steps, since, of course, evaluation of interatomic distances *also* formally scales as N^2.

In practice, even though the use of a cut-off introduces only small disparities in the energy, the discontinuity of these disparities can cause problems for optimizers. A more stable approach is to use a 'switching function' which multiplies the van der Waals interaction and causes it (and possibly its first and second derivatives) to go smoothly to zero at some cut-off distance. This function must, of course, be equal to 1 at short distances.

The electrostatic interaction is more problematic. For point charges, the interaction energy decays as r^{-1}. As already noted, the number of interactions increases by up to r^2, so the total energy in an infinite system might be expected to diverge! Such formal divergence is avoided in most real cases, however, because in systems that are electrically neutral there are as many positive interactions as negative, and thus there are large cancellation effects. If we imagine a system composed entirely of neutral groups (e.g., functional groups of a single molecule or individual molecules of a condensed phase), the long-range interaction between groups is a dipole–dipole interaction, which decays as r^{-3}, and the total energy contribution should decay as r^{-1}. Again, the actual situation is more favorable because of positive and negative cancellation effects, but the much slower decay of the electrostatic interaction makes it significantly harder to deal with. Cut-off distances (again, ideally implemented with smooth switching functions) must be quite large to avoid structural artifacts (e.g., atoms having large partial charges of like sign anomalously segregating at interatomic distances just in excess of the cut-off).

In infinite periodic systems, an attractive alternative to the use of a cut-off distance is the Ewald sum technique, first described for chemical systems by York, Darden and Pedersen (1993). By using a reciprocal-space technique to evaluate long-range contributions, the total

electrostatic interaction can be calculated to a pre-selected level of accuracy (i.e., the Ewald sum limit is exact) with a scaling that, in the most favorable case (called 'Particle-mesh Ewald', or PME), is $N\log N$. Prior to the introduction of Ewald sums, the modeling of polyelectrolytes (e.g., DNA) was rarely successful because of the instabilities introduced by cut-offs in systems having such a high degree of localized charges (see, for instance, Beveridge and McConnell 2000).

In aperiodic systems, another important contribution has been the development of the so-called 'Fast Multipole Moment' (FMM) method (Greengard and Rokhlin 1987). In essence, this approach takes advantage of the significant cancellations in charge–charge interactions between widely separated regions in space, and the increasing degree to which those interactions can be approximated by highly truncated multipole–multipole interactions. In the most favorable case, FMM methods scale linearly with system size.

It should be remembered, of course, that scaling behavior is informative of the relative time one system takes compared to another of different size, and says *nothing* about the *absolute* time required for the calculation. Thus, FMM methods scale linearly, but the initial overhead can be quite large, so that it requires a very large system before it outperforms PME for the same level of accuracy. Nevertheless, the availability of the FMM method renders conceivable the molecular modeling of extraordinarily large systems, and refinements of the method are likely to be forthcoming.

An interesting question that arises with respect to force fields is the degree to which they can be used to study reactive processes, i.e., processes whereby one minimum-energy compound is converted into another with the intermediacy of some transition state. As noted at the beginning of this chapter, one of the first applications of force-field methodology was to study the racemization of substituted biphenyls. And, for such 'conformational reactions', there seems to be no reason to believe force fields would not be perfectly appropriate modeling tools. Unless the conformational change in question were to involve an enormous amount of strain in the TS structure, there is little reason to believe that any of the internal coordinates would be so significantly displaced from their equilibrium values that the force-field functional forms would no longer be accurate.

However, when it comes to reactions where bonds are being made and/or broken, it is clear that, at least for the vast majority of force fields that use polynomial expressions for the bond stretching energy, the 'normal' model is inapplicable. Nevertheless, substantial application of molecular mechanics to such TS structures has been reported, with essentially three different approaches having been adopted.

One approach, when sufficient data are available, is to define new atom types and associated parameters for those atoms involved in the bond-making/bond-breaking coordinate(s). This is rather tricky since, while there may be solid experimental data for activation energies, there are unlikely to be any TS structural data. Instead, one might choose to use structures computed from some QM level of theory for one or more members of the molecular data set. Then, if one assumes the reaction coordinate is highly transferable from one molecule to the next (i.e., this methodology is necessarily restricted to the study of a single reaction amongst a reasonably closely related set of compounds), one can define a force field where TS structures are treated as 'minima' – minima in quotes because the equilibrium distances

and force constants for the reactive coordinate(s) have values characteristic of the transition state.

This methodology has two chief drawbacks. A philosophical drawback is that movement along the reaction coordinate raises the force-field energy instead of lowering it, which is opposite to the real chemical system. A practical drawback is that it tends to be data limited – one may need to define a fairly large number of parameters using only a rather limited number of activation energies and perhaps some QM data. As noted in Section 2.2.7, this creates a tension between chemical intuition and statistical rigor. Two papers applying this technique to model the acid-catalyzed lactonization of organic hydroxy-acids illustrate the competing extremes to which such optimizations may be taken (Dorigo and Houk 1987; Menger 1990).

An alternative approach is one that is valence-bond like in its formulation. A possible TS structure is one whose molecular geometry is computed to have the same energy *irrespective of whether the atomic connectivity is that of the reactant or that of the product* (Jensen 1992). Consider the example in Figure 2.9 for a hypothetical hydride transfer from an alkoxide carbon to a carbonyl. When the C–H bond is stretched from the reactant structure, the energy of the reactant-bonded structure goes up, while the energy of the product-bonded structure goes down because that structure's C–H bond is coming closer to its equilibrium value (from which it is initially very highly displaced). The simplest way to view this process

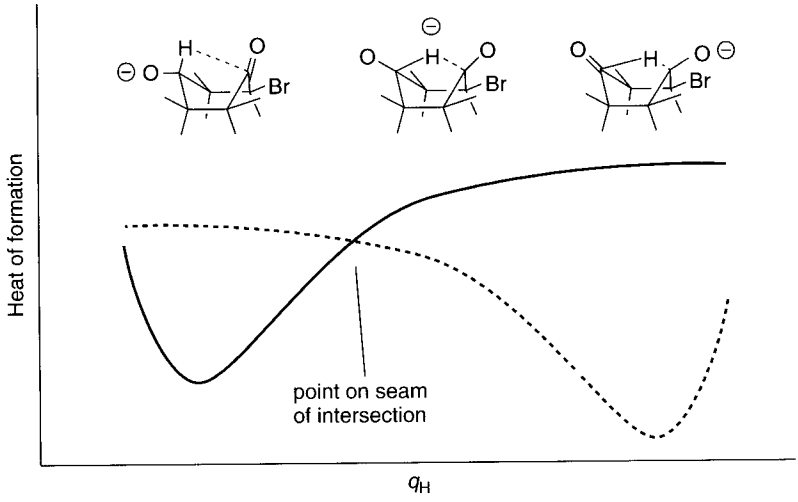


Figure 2.9 Slice through two intersecting enthalpy 'surfaces' along an arbitrary coordinate describing the location of a transferring H atom. The solid curve corresponds to bond stretching of the solid bond from carbon to the H atom being transferred. The dashed curve corresponds analogously to the dashed bond. At the point of intersection, the structure has the same energy irrespective of which bonding scheme is chosen. [For chemical clarity, the negative charge is shown shifting from one oxygen to the other, but for the method to be valid the two oxygen atom types could not change along either reaction coordinate. Note also that the bromine atom lifts the symmetry that would otherwise be present in this reaction.]

is to envision two PESs, one defined for the reactant and one for the product. These two surfaces will intersect along a 'seam', and this seam is where the energy is independent of which connectivity is employed. The TS structure is then defined as the *minimum* on the seam. This approach is only valid when the reactant and product energies are computed relative to a common zero (e.g., heats of formation are used; see Section 2.3), but one of its chief advantages is that it should properly reflect movement of the TS structure as a function of reaction thermicity. Because the seam of intersection involves structures having highly stretched bonds, care must be taken to use bond stretching functional forms that are accurate over larger ranges than are otherwise typical.

The third approach to finding TS structures involves either adopting bond making/breaking functional forms that are accurate at all distances (making evaluation of bond energies a rather unpleasant N^2 process), or mixing the force-field representation of the bulk of the molecule with a QM representation of the reacting region. Mixed QM/MM models are described in detail in Chapter 13.

2.5 Menagerie of Modern Force Fields

2.5.1 Available Force Fields

Table 2.1 contains an alphabetic listing of force fields which for the most part continue to be in use today. Nomenclature of force fields can be rather puzzling because developers rarely change the name of the force field as development progresses. This is not necessarily a major issue when new development extends a force field to functionality that had not previously been addressed, but can be singularly confusing if pre-existing parameters or functional forms are changed from one version to the next without an accompanying name change. Many developers have tried to solve this problem by adding to the force field name the last two digits of the year of the most recent change to the force field. Thus, one can have MM3(92) and MM3(96), which are characterized by, *inter alia*, different hydrogen bonding parameters. Similarly, one has consistent force field (CFF) and Merck molecular force field (MMFF) versions identified by trailing year numbers. Regrettably, the year appearing in a version number does not necessarily correspond to the year in which the modifications were published in the open literature. Moreover, even when the developers themselves exercise adequate care, there is a tendency for the user community to be rather sloppy in referring to the force field, so that the literature is replete with calculations inadequately described to ensure reproducibility.

Further confusing the situation, certain existing force fields have been used as starting points for development by new teams of researchers, and the name of the resulting product has not necessarily been well distinguished from the original (which may itself be in ongoing development by its original designers!). Thus, for instance, one has the MM2* and MM3* force fields that appear in the commercial program MACROMODEL and that are based on early versions of the unstarred force fields of the same name (the * indicates the use of point charges to evaluate the electrostatics instead of bond dipoles, the use of a non-directional 10–12 potential for hydrogen bonding in place of an MM3 Buckingham potential, and a

Table 2.1 Force fields

Name (if any)	Range	Comments	Refs	Σ(error)[a]
–	Biomolecules (2nd generation includes organics)	Sometimes referred to as AMBER force fields; new versions are first coded in software of that name. All-atom (AA) and united-atom (UA) versions exist.	Original: Weiner, S. J., Kollman, P. A., Nguyen, D. T. and Case, D. A. 1986. *J. Comput. Chem.*, **7**, 230. Latest generation: Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Jr., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W., and Kollman, P. A. 1995. *J. Am. Chem. Soc.* **117**, 5179. See also www.amber.ucsf.edu	
–	Organics and biomolecules	The program MACROMODEL contains many modified versions of other force fields, e.g., AMBER*, MM2*, MM3*, OPLSA*.	Mohamadi, F., Richards, N. J. G., Guida, W. C., Liskamp, R., Lipton, M., Caufield, C., Chang, G., Hendrickson, T., and Still, W. C. 1990. *J. Comput. Chem.* **11**, 440. Recent extension: Senderowitz, H. and Still, W. C. 1997. *J. Org. Chem.*, **62**, 1427. See also www.schrodinger.com	7 (AMBER*) 4 (MM2*) 5 (MM3*)
CHARMM	Biomolecules	Many versions of force field parameters exist, distinguished by ordinal number. All-atom and united-atom versions exist.	Original: Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S., and Karplus, M. 1983. *J. Comput. Chem.*, **4**, 187; Nilsson, L. and Karplus, M. 1986. *J. Comput. Chem.*, **7**, 591. Latest generation: MacKerell, A. D., Bashford, D., Bellott, M., Dunbrack, R. L., Evanseck, J. D., Field, M. J., Gao, J., Guo, H., Ha, S., Joseph-McCarthy, D., Kuchnir, L., Kuczera, K., Lau, T. F. K., Mattos, C., Michnick, S., Nago, T.,	

			Nguyen, D. T., Prodhom, B., Reiher, W. E., Roux, B., Schlenkrich, M., Smith, J. C., Stote, R., Straub, J., Watanabe, M., Wiórkievicz-Kuczera, J., Yin, D., and Karplus, M. 1998. *J. Phys. Chem. B*, **102**, 3586. See also yuri.harvard.edu	
CHARMm	Biomolecules and organics	Version of CHARMM somewhat extended and made available in Accelrys software products.	Momany, F. A. and Rone, R. 1992. *J. Comput. Chem.*, **13**, 888. See also www.accelrys.com	
Chem-X	Organics	Available in Chemical Design Ltd. software.	Davies, E. K. and Murrall, N. W. 1989. *J. Comput. Chem.*, **13**, 149.	12
CFF/CVFF	Organics and biomolecules	CVFF is the original; CFF versions are identified by trailing year digits. Bond stretching can be modeled with a Morse potential. Primarily available in Accelrys software.	CVFF: Lifson, S., Hagler, A. T., and Stockfisch, J. P. 1979. *J. Am. Chem. Soc.*, **101**, 5111, 5122, 5131.	13 (CVFF)
			CFF: Hwang, M.-J., Stockfisch, T. P., and Hagler, A. T. 1994. *J. Am. Chem. Soc.*, **116**, 2515; Maple, J. R., Hwang, M.-J., Stockfisch, T. P., Dinur, U., Waldman, M., Ewig, C. S., and Hagler, A. T. 1994. *J. Comput. Chem.*, **15**, 162; Maple, J. R., Hwang, M.-J., Jalkanen, K. J., Stockfisch, T. P., and Hagler, A. T. 1998. *J. Comput. Chem.*, **19**, 430; Ewig, C. S., Berry, R., Dinur, U., Hill, J.-R., Hwang, M.-J., Li, C., Maple, J., Peng, Z., Stockfisch, T. P., Thacher, T. S., Yan, L., Ni, X., and Hagler, A. T. 2001. *J. Comput. Chem.*, **22**, 1782. See also www.accelrys.com	7 (CFF91)

(*continued overleaf*)

Table 2.1 (*continued*)

Name (if any)	Range	Comments	Refs	Σ(error)[a]
COSMIC	Organics and biomolecules		Vinter, J. G., Davies, A., and Saunders, M. R. 1987. *J. Comput.-Aided Mol. Des.*, **1**, 31; Morley, S. D., Abraham, R. J., Hawarth, I. S., Jackson, D. E., Saunders, M. R., and Vinter, J. G. 1991. *J. Comput-Aided Mol. Des.*, **5**, 475.	
DREIDING	Main-group organics and inorganics	Bond stretching can be modeled with a Morse potential.	Mayo, S. L., Olafson, B. D., and Goddard, W. A., III, 1990. *J. Phys. Chem.* **94**, 8897.	10
ECEPP	Proteins	Computes only non-bonded interactions for fixed structures. Versions identified by /(ordinal number) after name.	Original: Némethy, G., Pottle, M. S., and Scheraga, H. A. 1983. *J. Phys. Chem.*, **87**, 1883. Latest generation: Némethy, G., Gibson, K. D., Palmer, K. A., Yoon, C. N., Paterlini, G., Zagari, A., Rumsey, S., and Scheraga, H. A. 1992. *J. Phys. Chem.*, **96**, 6472.	
ESFF	General	Bond stretching is modeled with a Morse potential.	Barlow, S., Rohl, A. L., Shi, S., Freeman, C. M., and O'Hare, D. 1996. *J. Am. Chem. Soc.*, **118**, 7578.	
GROMOS	Biomolecules	Coded primarily in the software having the same name.	Daura, X., Mark, A. E., and van Gunsteren, W. F. 1998. *J. Comput. Chem.*, **19**, 535.; Schuler, L. D., Daura, X., and van Gunsteren, W. F. 2001. *J. Comput. Chem.*, **22**, 1205. See also igc.ethz.ch/gromos	
MM2	Organics	Superseded by MM3 but still widely available in many modified forms.	Comprehensive: Burkert, U. and Allinger, N. L. 1982. *Molecular Mechanics*, ACS Monograph 177, American Chemical Society: Washington, DC.	5 (MM2(85), MM2(91), Chem-3D)

MM3	Organics and biomolecules	Widely available in many modified forms.	Original: Allinger, N. L., Yuh, Y. H., and Lii, J.-H. 1989. *J. Am. Chem. Soc.*, **111**, 8551. MM3(94): Allinger, N. L., Zhou, X., and Bergsma, J. 1994. *J. Mol. Struct. (Theochem)*, **312**, 69. Recent extension: Stewart, E. L., Nevins, N., Allinger, N. L., and Bowen, J. P. 1999. *J. Org. Chem.* **64**, 5350.	5 (MM3(92))
MM4	Hydrocarbons		Allinger, N. L., Chen, K. S., and Lii, J. H. 1996. *J. Comput. Chem.*, **17**, 642; Nevins, N., Chen, K. S., and Allinger, N. L. 1996. *J. Comput. Chem.*, **17**, 669; Nevins, N., Lii, J. H., and Allinger, N. L. 1996. *J. Comput. Chem.*, **17**, 695; Nevins, N. and Allinger, N. L. 1996. *J. Comput. Chem.*, **17**, 730; Allinger, N. L., Chen, K. S., Katzenellenbogen, J. A., Wilson, S. R., and Anstead, G. M. 1996. *J. Comput. Chem.*, **17**, 747.	
MMFF	Organics and biomolecules	Widely available in relatively stable form.	Halgren, T. A. 1996. *J. Comput. Chem.*, **17**, 490, 520, 553, 616; Halgren, T. A., and Nachbar, R. B. 1996. *J. Comput. Chem.*, **17**, 587. See also www.schrodinger.com	4 (MMFF93)
MMX	Organics, biomolecules, and inorganics	Based on MM2.	See www.serenasoft.com	5
MOMEC	Transition metal compounds		Comba, P., and Hambley, T. W. 1995. *Molecular Modeling of Inorganic Compounds*, VCH: New York.	

(continued overleaf)

Table 2.1 (*continued*)

Name (if any)	Range	Comments	Refs	Σ(error)[a]
OPLS	Biomolecules, some organics	Organic parameters are primarily for solvents. All-atom and united-atom versions exist.	Proteins: Jorgensen, W. L., and Tirado-Rives, J. 1988. *J. Am. Chem. Soc.*, **110**, 1657; Kaminski, G. A., Friesner, R. A., Tirado-Rives, J., Jorgensen, W. L. 2001. *J. Phys. Chem. B*, **105**, 6474. Nucleic acids: Pranata, J., Wierschke, S. G., and Jorgensen, W. L. 1991. *J. Phys. Chem. B*, **113**, 2810. Sugars: Damm, W., Frontera, A., Tirado-Rives, J., and Jorgensen, W. L. 1997. *J. Comput. Chem.*, **18**, 1955. Recent extensions: Rizzo, R. C., Jorgensen, W. L. 1999. *J. Am. Chem. Soc.*, **121**, 4827.	
PEF95SAC	Carbohydrates	Based on CFF form.	Fabricius, J., Engelsen, S. B., and Rasmussen, K. 1997. *J. Carbohydr. Chem.*, **16**, 751.	
SHAPES	Transition metal compounds		Allured, V. S., Kelly, C., and Landis, C. R. 1991. *J. Am. Chem. Soc.*, **113**, 1.	
SYBYL/Tripos	Organics and proteins	Available in Tripos and some other software.	Clark, M., Cramer, R. D., III, and van Opdenbosch, N. 1989. *J. Comput. Chem.*, **10**, 982. See also www.tripos.com and www.scivision.com	8–12
UFF	General	Bond stretching can be modeled with a Morse potential.	Rappé, A. K., Casewit, C. J., Colwell, K. S., Goddard, W. A., III, and Skiff, W. M. 1992. *J. Am. Chem. Soc.*, **114**, 10024, 10035, 10046.	21
VALBOND	Transition metal compounds	Atomic-orbital-dependent energy expressions.	Root, D. M., Landis, C. R., and Cleveland, T. 1993. *J. Am. Chem. Soc.*, **115**, 4201.	

[a] Kcal mol^{-1}. From Gundertofte *et al.* (1991, 1996); see text.

different formalism for handling conjugated systems). The commercial program Chem3D also has force fields based on MM2 and MM3, and makes no modification to the names of the originals.

As a final point of ambiguity, some force fields have not been given names, *per se*, but have come to be called by the names of the software packages in which they first became widely available. Thus, the force fields developed by the Kollman group (see Table 2.1) have tended to be referred to generically as AMBER force fields, because this software package is where they were originally coded. Kollman preferred that they be referred to by the names of the authors on the relevant paper describing their development, e.g., 'the force field of Cornell *et al.*' This is certainly more informative, since at this point the AMBER program includes within it *many* different force fields, so reference to the 'AMBER force field' conveys no information.

Because of the above ambiguities, and because it is scientifically unacceptable to publish data without an adequate description of how independent researchers might reproduce those data, many respected journals in the chemistry field now have requirements that papers reporting force-field calculations include as supplementary material a complete listing of all force field parameters (and functional forms, if they too cannot be adequately described otherwise) required to carry out the calculations described. This also facilitates the dissemination of information to those researchers wishing to develop their own codes for specific purposes.

Table 2.1 also includes a general description of the chemical space over which the force field has been designed to be effective; in cases where multiple subspaces are addressed, the order roughly reflects the priority given to these spaces during development. Force fields which have undergone many years worth of refinements tend to have generated a rather large number of publications, and the table does not try to be exhaustive, but effort is made to provide key references. The table also includes comments deemed to be particularly pertinent with respect to software implementing the force fields.

2.5.2 Validation

The vast majority of potential users of molecular mechanics have two primary, related questions: 'How do I pick the best force field for my problem?' and, 'How will I know whether I can trust the results?' The process of testing the utility of a force field for molecules other than those over which it was parameterized is known as 'validation'.

The answer to the first question is obvious, if not necessarily trivial: one should pick the force field that has previously been shown to be most effective for the most closely related problem one can find. That demonstration of effectiveness may have taken place within the process of parameterization (i.e., if one is interested in conformational properties of proteins, one is more likely to be successful with a force field specifically parameterized to model proteins than with one which has not been) or by post-development validation. Periodically in the literature, papers appear comparing a wide variety of force fields for some well-defined problem, and the results can be quite useful in guiding the choices of subsequent researchers (see also, Bays 1992). Gundertofte *et al.* (1991, 1996) studied the accuracy of

17 different force fields with respect to predicting 38 experimental conformational energy differences or rotational barriers in organic molecules. These data were grouped into eight separate categories (conjugated compounds, halocyclohexanes, haloalkanes, cyclohexanes, nitrogen-containing compounds, oxygen-containing compounds, hydrocarbons, and rotational barriers). A summary of these results appears for relevant force fields in Table 2.1, where the number cited represents the sum of the mean errors over all eight categories. In some cases a range is cited because different versions of the same force field and/or different software packages were compared. In general, the best performances are exhibited by the MM2 and MM3 force fields and those other force fields based upon them. In addition, MMFF93 had similar accuracy. Not surprisingly, the most general force fields do rather badly, with UFF faring quite poorly in every category other than hydrocarbons.

Broad comparisons have also appeared for small biomolecules. Barrows *et al.* (1998) compared 10 different force fields against well-converged quantum mechanical calculations for predicting the relative conformational energies of 11 different conformers of D-glucose. GROMOS, MM3(96), and the force field of Weiner *et al.* were found to have average errors of 1.5 to 2.1 kcal/mol in relative energy, CHARMM and MMFF had average errors of from 0.9 to 1.5, and AMBER*, Chem-X, OPLS, and an unpublished force field of Simmerling and Kollman had average errors from 0.6 to 0.8 kcal/mol, which compared quite well with vastly more expensive *ab initio* methods. Beachy *et al.* (1997) carried out a similar comparison for a large number of polypeptide conformations and found OPLS, MMFF, and the force field of Cornell *et al.* to be generally the most robust.

Of course, in looking for an optimal force field there is no guarantee that *any* system sufficiently similar to the one an individual researcher is interested in has *ever* been studied, in which case it is hard to make a confident assessment of force-field utility. In that instance, assuming some experimental data are available, it is best to do a survey of several force fields to gauge their reliability. When experimental data are *not* available, recourse to well-converged quantum mechanical calculations for a few examples is a possibility, assuming the computational cost is not prohibitive. QM values would then take the place of experimental data. Absent any of these alternatives, any force field calculations will simply carry with them a high degree of uncertainty and the results should be used with caution.

Inorganic chemists may be frustrated to have reached this point having received relatively little guidance on what force fields are best suited to *their* problems. Regrettably, the current state of the art does not provide any single force field that is both robust and accurate over a large range of inorganic molecules (particularly metal coordination compounds). As noted above, parameter transferability tends to be low, i.e., the number of atom types potentially requiring parameterization for a single metal atom, together with the associated very large number of geometric and non-bonded constants, tends to significantly exceed available data. Instead, individual problems tend to be best solved with highly tailored force fields, when they are available, or by combining QM and MM methods (see Chapter 13), or by accepting that the use of available highly generalized force fields increases the risk for significant errors and thus focusing primarily on structural perturbations over a related series of compounds rather than absolute structures or energetics is advised (see also Hay 1993; Norrby and Brandt 2001).

A last point that should be raised with regard to validation is that any comparison between theory and experiment must proceed in a consistent fashion. Consider molecular geometries. Chemists typically visualize molecules as having 'structure'. Thus, for example, single-crystal X-ray diffractometry can be used to determine a molecular structure, and at the end of a molecular mechanics minimization one has a molecular structure, but is it strictly valid to compare them?

It is best to consider this question in a series of steps. First, recall that the goal of a MM minimization is to find a local minimum on the PES. That local minimum has a unique structure and each molecular coordinate has a precise value. What about the structure from experiment? Since most experimental techniques for assigning structure sample an ensemble of molecules (or one molecule many, many times), the experimental measurement is properly referred to as an expectation value, which is denoted by angle brackets about the measured variable. Real molecules vibrate, even at temperatures arbitrarily close to absolute zero, so measured structural parameters are actually expectation values over the molecular vibrations. Consider, for example, the length of the bond between atoms A and B in its ground vibrational state. For a quantum mechanical harmonic oscillator, $\langle r_{AB} \rangle = r_{AB,eq}$, but real bond stretching coordinates are anharmonic, and this inevitably leads to $\langle r_{AB} \rangle > r_{AB,eq}$ (see Section 9.3.2). In the case of He_2, mentioned above, the effect of vibrational averaging is rather extreme, leading to a difference between $\langle r_{AB} \rangle$ and $r_{AB,eq}$ of more than 50 Å! Obviously, one should not judge the quality of the calculated $r_{AB,eq}$ value based on comparison to the experimental $\langle r_{AB} \rangle$ value. Note that discrepancies between $\langle r_{AB} \rangle$ and $r_{AB,eq}$ will increase if the experimental sample includes molecules in excited vibrational states. To be rigorous in comparison, either the calculation should be extended to compute $\langle r_{AB} \rangle$ (by computation of the vibrational wave function(s) and appropriate averaging) or the experiment must be analyzed to determine $r_{AB,eq}$, e.g., as described in Figure 2.1.

Moreover, the above discussion assumes that the experimental technique measures exactly what the computational technique does, namely, the separation between the nuclear centroids defining a bond. X-ray crystallography, however, measures maxima in scattering amplitudes, and X-rays scatter not off nuclei but off electrons. Thus, if electron density maxima do not correspond to nuclear positions, there is no reason to expect agreement between theory and experiment (for heavy atoms this is not much of an issue, but for very light ones it can be). Furthermore, the conditions of the calculation typically correspond to an isolated molecule acting as an ideal gas (i.e., experiencing no intermolecular interactions), while a technique like X-ray crystallography obviously probes molecular structure in a condensed phase where crystal packing and dielectric effects may have significant impact on the determined structure.

The above example illustrates some of the caveats in comparing theory to experiment for a structural datum (see also Allinger, Zhou and Bergsma 1994). Care must also be taken in assessing energetic data. Force-field calculations typically compute potential energy, whereas equilibrium distributions of molecules are dictated by free energies (see Chapter 10). Thus, the force-field energies of two conformers should not necessarily be expected to reproduce an experimental equilibrium constant between them. The situation can become still more confused for transition states, since experimental data typically are either activation free energies or Arrhenius activation energies, neither of which corresponds directly with the

difference in potential energy between a reactant and a TS structure (see Chapter 15). Even in those cases where the force field makes possible the computation of heats of formation and the experimental data are available as enthalpies, it must be remembered that the effect of zero-point vibrational energy is accounted for in an entirely average way when atom-type reference heats of formation are parameterized, so some caution in comparison is warranted.

Finally, any experimental measurement carries with it some error, and obviously a comparison between theory and experiment should never be expected to do better than the experimental error. *The various points discussed in this last section are all equally applicable to comparisons between experiment and QM theories as well, and the careful practitioner would do well always to bear them in mind.*

2.6 Case Study: (2*R**,4*S**)-1-Hydroxy-2,4-dimethylhex-5-ene

Synopsis of Stahl *et al.* (1991), 'Conformational Analysis with Carbon-Carbon Coupling Constants. A Density Functional and Molecular Mechanics Study'.

Many natural products contain one or more sets of carbon backbones decorated with multiple stereogenic centers. A small such fragment that might be found in propiogenic natural products is illustrated in Figure 2.10. From a practical standpoint, the assignment of absolute configuration to each stereogenic center (*R* or *S*), or even of the relative configurations between centers, can be difficult in the absence of single-crystal X-ray data. When many possibilities exist, it is an unpleasant task to synthesize each one.

An alternative means to assign the stereochemistry is to use nuclear magnetic resonance (NMR). Coupling constant data from the NMR experiment can be particularly useful in assigning stereochemistry. However, if the fragments are highly flexible, the interpretation of the NMR data can be complicated when the interconversion of conformers is rapid on the NMR timescale. In that case, rather than observing separate, overlapping spectra for every conformer, only a population-averaged spectrum is obtained.

Deconvolution of such spectra can be accomplished in a computational fashion by (i) determining the energies of all conformers contributing to the equilibrium population, (ii) predicting the spectral constants associated with each conformer, and (iii) averaging over all spectral data weighted by the fractional contribution of each conformer to the equilibrium (the fractional contribution is determined by a Boltzmann average over the energies, see Eq. (10.49)). The authors adopted this approach for (2*R**,4*S**)-1-hydroxy-2,4-dimethylhex-5-ene, where the conformer energies were determined using the MM3 force field and the NMR coupling constants were predicted at the density functional level of theory. As density functional theory is the subject of Chapter 8 and the prediction of NMR data is not discussed until Section 9.4, we will focus here simply on the performance of MM3 for predicting conformer energies and weighting spectral data.

In order to find the relevant conformers, the authors employed a Monte Carlo/minimization strategy that is described in more detail in the next chapter – in practice, (2*R**,4*S**)-1-hydroxy-2,4-dimethylhex-5-ene is sufficiently small that one *could* survey every possible torsional isomer by brute force, but it would be very tedious. Table 2.2 shows, for the nine lowest energy conformers, their predicted energies, their contribution to the 300 K equilibrium population, their individual $^3J_{CC}$ coupling constants between atoms C(2)C(5), C(2)C(8), C(1)C(4), and C(4)C(7), and the mean absolute error in these coupling

constants compared to experiment (see Figure 2.10 for atom-numbering convention). In addition, the spectral data predicted from a population-weighted equilibrium average over the nine conformers making up 82% of the equilibrium population are shown.

The population-averaged data are those in best agreement with experiment. Conformer G shows similar agreement (the increased error is within the rounding limit for the table), but is predicted to be sufficiently high in energy that it is unlikely that MM3 could be sufficiently in error for it to be the only conformer at equilibrium. As a separate assessment of this point, the authors carry out *ab initio* calculations at a correlated level of electronic structure theory (MP2/TZ2P//HF/TZ2P; this notation and the relevant theories are discussed in Chapters 6 and 7, but exact details are not important here), and observe what they characterize as very good agreement between the force-field energies and the *ab initio* energies (the data are not provided).

In principle, then, when the relative configurations are not known for a flexible chain in some natural product backbone, the technique outlined above could be used to predict the expected NMR spectra for all possibilities, and presuming one prediction matched to experiment significantly more closely than any other, the assignment would be regarded

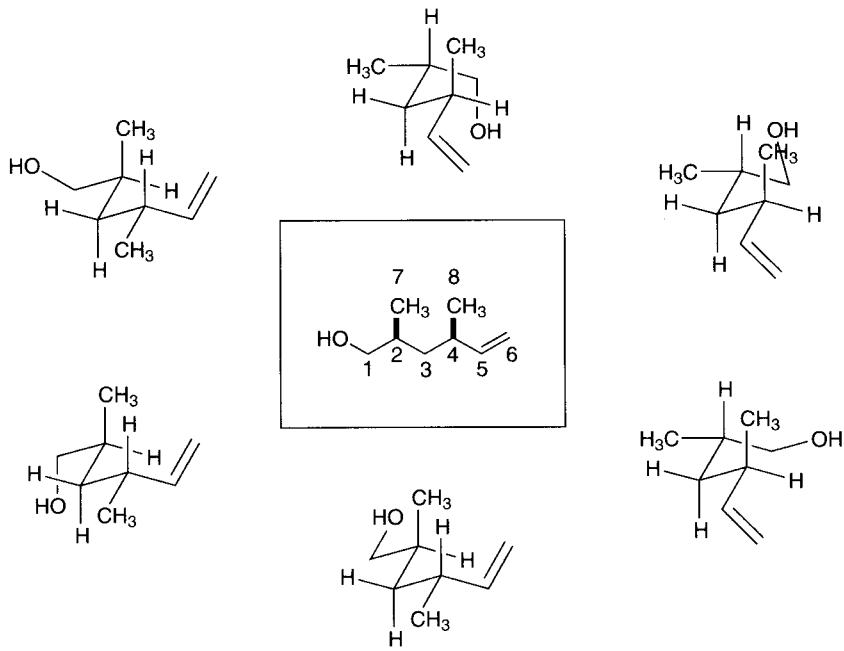


Figure 2.10 Some plausible conformations of (2*R**,4*S**)-1-hydroxy-2,4-dimethylhex-5-ene. How many different torsional isomers might one need to examine, and how would you go about generating them? [Note that the notation 2*R**,4*S** implies that the *relative* stereochemical configuration at the 2 and 4 centers is *R*,*S* – by convention, when the absolute configuration is not known the first center is always assigned to be *R**. However, the absolute conformations that are drawn here are *S*,*R* so as to preserve correspondence with the published illustrations of Stahl and coworkers. Since NMR in an achiral solvent does not distinguish between enantiomers, one can work with either absolute configuration in this instance.]

Table 2.2 Relative MM3 energies (kcal mol^{-1}), fractional equilibrium populations F (%), predicted NMR coupling constants (Hz), and mean unsigned error in predicted coupling constants for different conformers and the equilibrium average of (2*R**,4*S**)-1-hydroxy-2,4-dimethylhex 5-ene at 300 K.

			^3J				
Conformer	rel E	F	C(2)C(5)	C(2)C(8)	C(1)C(4)	C(4)C(7)	MUE
A	0.0	24	1.1	4.2	3.9	1.3	**0.6**
B	0.1	21	1.1	4.0	5.8	1.2	**1.0**
C	0.2	19	1.0	4.2	4.1	1.2	**0.7**
D	0.9	5	3.8	1.5	1.7	4.5	**2.2**
E	1.1	4	4.1	0.8	1.1	4.4	**2.5**
F	1.3	3	4.1	0.9	0.4	5.3	**2.9**
G	1.4	2	1.2	3.7	3.8	1.5	**0.3**
H	1.4	2	1.4	4.2	5.7	1.4	**0.9**
I	1.5	2	0.1	5.1	0.0	5.3	**2.5**
average		*82*	*1.4*	*3.7*	*4.1*	*1.8*	***0.3***
experiment			**1.4**	**3.3**	**3.8**	**2.2**	

as reasonably secure. At the least, it would suggest how to prioritize synthetic efforts that would be necessary to provide the ultimate proof.

In that regard, this paper might have been improved by including a prediction (and ideally an experimental measurement) for the NMR coupling data of (2*R**,4*R**)-1-hydroxy-2,4-dimethylhex-5-ene, i.e., the stereoisomer having the *R**,*R** relative configuration between the stereogenic centers instead of the *R**,*S** configuration. If each predicted spectrum matched its corresponding experimental spectrum significantly more closely than it matched the non-corresponding experimental spectrum, the utility of the methodology would be still more convincingly demonstrated. Even in the absence of this demonstration, however, the work of Stahl and his coworkers nicely illustrates how accurate force fields can be for 'typical' C,H,O-compounds, and also how different levels of theory can be combined to address different parts of a computational problem in the most efficient manner. In this case, inexpensive molecular mechanics is used to provide an accurate map of the wells on the conformational potential energy surface and the vastly more expensive DFT method is employed only thereafter to predict the NMR spectral data.

Bibliography and Suggested Additional Reading

Bowen, J. P. and Allinger, N. L. 1991. 'Molecular Mechanics: The Art and Science of Parameterization', in *Reviews in Computational Chemistry*, Vol. 2, Lipkowitz, K. B. and Boyd, D. B. Eds., VCH: New York, 81.

Cramer, C. J. 1994. 'Problems and Questions in the Molecular Modeling of Biomolecules', *Biochem. Ed.* **22**, 140.

Dinur, U. and Hagler, A. T. 1991. 'New Approaches to Empirical Force Fields', in *Reviews in Computational Chemistry*, Vol. 2, Lipkowitz, K. B. and Boyd, D. B., Eds., VCH; New York, 99.

Dykstra, C. E. 1993. 'Electrostatic Interaction Potentials in Molecular Force Fields', *Chem. Rev.* **93**, 2339.

Eksterowicz, J. E. and Houk, K. N. 1993. 'Transition-state Modeling with Empirical Force Fields', *Chem. Rev.* **93**, 2439.
Jensen, F. 1999. *Introduction to Computational Chemistry*, Wiley: Chichester.
Landis, C. R., Root, D. M., and Cleveland, T. 1995. 'Molecular Mechanics Force Fields for Modeling Inorganic and Organometallic Compounds' in *Reviews in Computational Chemistry*, Vol. 6, Lipkowitz, K. B. and Boyd, D. B. Eds., VCH: New York, 73.
Levine, I. N. 2000. *Quantum Chemistry*, 5th Edn., Prentice Hall: New York.
Norrby, P.-O. 2001. 'Recipe for an Organometallic Force Field', in *Computational Organometallic Chemistry*, Cundari, T. Ed., Marcel Dekker: New York 7.
Pettersson, I. and Liljefors, T. 1996. 'Molecular Mechanics Calculated Conformational Energies of Organic Molecules: A Comparison of Force Fields', in *Reviews in Computational Chemistry*, Vol. 9, Lipkowitz, K. B. and Boyd, D. B., Eds., VCH: New York, 167.
Schlegel, H. B. 1989. 'Some Practical Suggestions for Optimizing Geometries and Locating Transition States', in *New Theoretical Concepts for Understanding Organic Reactions*, Bertrán, J. and Csizmadia, I. G., Eds., Kluwer: Dordrecht, 33.

References

Allinger, N. L., Zhou, X., and Bergsma, J. 1994. *J. Mol. Struct. (Theochem.)*, **312**, 69.
Barrows, S. E., Storer, J. W., Cramer, C. J., French, A. D., and Truhlar, D. G. 1998. *J. Comput. Chem.*, **19**, 1111.
Bays, J. P. 1992. *J. Chem. Edu.*, **69**, 209.
Beachy, M. D., Chasman, D., Murphy, R. B., Halgren, T. A., and Friesner, R. A. 1997. *J. Am. Chem. Soc.*, **119**, 5908.
Beveridge, D. L. and McConnell, K. J. 2000. *Curr. Opin. Struct. Biol.*, **10**, 182.
Cohen, N. and Benson, S. W. 1993. *Chem. Rev.*, **93**, 2419.
Dorigo, A. E. and Houk, K. N. 1987. *J. Am. Chem. Soc.*, **109**, 3698.
Fogarasi, G. and Balázs, A. 1985. *J. Mol. Struct. (Thochem)*, **133**, 105.
Greengard, L. and Rokhlin, V. 1987. *J. Comput. Phys.*, **73**, 325.
Gundertofte, K., Liljefors, T., Norrby, P.-O., and Petterson, I. 1996. *J. Comput. Chem.*, **17**, 429.
Gundertofte, K., Palm, J., Petterson, I., and Stamvick, A. 1991. *J. Comput. Chem.*, **12**, 200.
Hay, B. P. 1993. *Coord. Chem. Rev.*, **126**, 177.
Hill, T. L. 1946. *J. Chem. Phys.*, **14**, 465.
Jensen, F. 1992. *J. Am. Chem. Soc.*, **114**, 1597.
Jensen, F. 1999. *Introduction to Computational Chemistry*, Wiley: Chichester, Chapter 14 and references therein.
Leach, R. 1991. *Rev. Comput. Chem.*, **2**, 1.
Menger, F. 1990. *J. Am. Chem. Soc.*, **112**, 8071.
Norrby, P.-O. and Brandt, P. 2001. *Coord. Chem. Rev.*, **212**, 79.
Radom, L., Hehre, W. J., and Pople, J. A. 1971. *J. Am. Chem. Soc.*, **93**, 289.
Reichardt, C. 1990. *Solvents and Solvent Effects in Organic Chemistry*, VCH: New York, 12.
Schlick, T. 1992. *Rev. Comput. Chem.*, **3**, 1.
Stahl, M., Schopfer, U., Frenking, G., and Hoffmann, R. W. 1991. *J. Am. Chem. Soc.*, **113**, 4792.
Westheimer, F. H. and Mayer, J. E. 1946. *J. Chem. Phys.*, **14**, 733.
Wolfe, S., Rauk, A., Tel, L. M., and Csizmadia, I. G. 1971. *J. Chem. Soc. B*, **136**.
Woods, R. J. 1996. *Rev. Comput. Chem.*, **9**, 129.
York, D. M., Darden, T. A., and Pederson, L. G. 1993. *J. Chem. Phys.*, **99**, 8345.

3

Simulations of Molecular Ensembles

3.1 Relationship Between MM Optima and Real Systems

As noted in the last chapter within the context of comparing theory to experiment, a minimum-energy structure, i.e., a local minimum on a PES, is sometimes afforded more importance than it deserves. Zero-point vibrational effects dictate that, even at 0 K, the molecule probabilistically samples a range of different structures. If the molecule is quite small and is characterized by fairly 'stiff' molecular coordinates, then its 'well' on the PES will be 'narrow' and 'deep' and the range of structures it samples will all be fairly close to the minimum-energy structure; in such an instance it is not unreasonable to adopt the simple approach of thinking about the 'structure' of the molecule as being the minimum energy geometry. However, consider the case of a large molecule characterized by many 'loose' molecular coordinates, say polyethyleneglycol, (PEG,$-(OCH_2CH_2)_n-$), which has 'soft' torsional modes: What is the structure of a PEG molecule having $n = 50$? Such a query is, in some sense, ill defined. Because the probability distribution of possible structures is not compactly localized, as is the case for stiff molecules, the very concept of structure as a time-independent property is called into question. Instead, we have to accept the flexibility of PEG as an intrinsic characteristic of the molecule, and any attempt to understand its other properties must account for its structureless nature. Note that polypeptides, polynucleotides, and polysaccharides all are *also* large molecules characterized by having many loose degrees of freedom. While nature has tended to select for particular examples of these molecules that are less flexible than PEG, nevertheless their utility as biomolecules sometimes derives from their ability to sample a wide range of structures under physiological conditions, and attempts to understand their chemical behavior must address this issue.

Just as zero-point vibration introduces probabilistic weightings to single-molecule structures, so too thermodynamics dictates that, given a large collection of molecules, probabilistic distributions of structures will be found about *different* local minima on the PES at non-zero absolute temperatures. The relative probability of clustering about any given minimum is a function of the temperature and some particular thermodynamic variable characterizing the system (e.g., Helmholtz free energy), that variable depending on what experimental conditions are being held constant (e.g., temperature and volume). Those variables being held constant define the 'ensemble'.

We will delay a more detailed discussion of ensemble thermodynamics until Chapter 10; indeed, in this chapter we will make use of ensembles designed to render the operative equations as transparent as possible without much discussion of extensions to other ensembles. The point to be re-emphasized here is that the vast majority of *experimental* techniques measure molecular properties as averages – either time averages or ensemble averages or, most typically, both. Thus, we seek computational techniques capable of accurately reproducing these aspects of molecular behavior. In this chapter, we will consider Monte Carlo (MC) and molecular dynamics (MD) techniques for the simulation of real systems. Prior to discussing the details of computational algorithms, however, we need to briefly review some basic concepts from statistical mechanics.

3.2 Phase Space and Trajectories

The state of a classical system can be completely described by specifying the positions and momenta of all particles. Space being three-dimensional, each particle has associated with it six coordinates – a system of N particles is thus characterized by $6N$ coordinates. The $6N$-dimensional space defined by these coordinates is called the 'phase space' of the system. At any instant in time, the system occupies one point in phase space

$$\mathbf{X}' = (x_1, y_1, z_1, p_{x,1}, p_{y,1}, p_{z,1}, x_2, y_2, z_2, p_{x,2}, p_{y,2}, p_{z,2}, \ldots) \tag{3.1}$$

For ease of notation, the position coordinates and momentum coordinates are defined as

$$\mathbf{q} = (x_1, y_1, z_1, x_2, y_2, z_2, \ldots) \tag{3.2}$$

$$\mathbf{p} = (p_{x,1}, p_{y,1}, p_{z,1}, p_{x,2}, p_{y,2}, p_{z,2}, \ldots) \tag{3.3}$$

allowing us to write a (reordered) phase space point as

$$\mathbf{X} = (\mathbf{q}, \mathbf{p}) \tag{3.4}$$

Over time, a dynamical system maps out a 'trajectory' in phase space. The trajectory is the curve formed by the phase points the system passes through. We will return to consider this dynamic behavior in Section 3.2.2.

3.2.1 Properties as Ensemble Averages

Because phase space encompasses every possible state of a system, the average value of a property A at equilibrium (i.e., its expectation value) for a system having a constant temperature, volume, and number of particles can be written as an integral over phase space

$$\langle A \rangle = \int\int A(\mathbf{q}, \mathbf{p}) P(\mathbf{q}, \mathbf{p}) d\mathbf{q} d\mathbf{p} \tag{3.5}$$

where P is the probability of being at a particular phase point. From statistical mechanics, we know that this probability depends on the energy associated with the phase point according to

$$P(\mathbf{q},\mathbf{p}) = Q^{-1}e^{-E(\mathbf{q},\mathbf{p})/k_{\mathrm{B}}T} \tag{3.6}$$

where E is the total energy (the sum of kinetic and potential energies depending on $\mathbf{p}$ and $\mathbf{q}$, respectively) k_{B} is Boltzmann's constant, T is the temperature, and Q is the system partition function

$$Q = \int\int e^{-E(\mathbf{q},\mathbf{p})/k_{\mathrm{B}}T} d\mathbf{q}d\mathbf{p} \tag{3.7}$$

which may be thought of as the normalization constant for P.

How might one go about evaluating Eq. (3.5)? In a complex system, the integrands of Eqs. (3.5) and (3.7) are unlikely to allow for analytic solutions, and one must perforce evaluate the integrals numerically. The numerical evaluation of an integral is, in the abstract, straightforward. One determines the value of the integrand at some finite number of points, fits those values to some function that *is* integrable, and integrates the latter function. With an increasing number of points, one should observe this process to converge to a particular value (assuming the original integral is finite) and one ceases to take more points after a certain tolerance threshold has been reached.

However, one must remember just how vast phase space is. Imagine that one has only a very modest goal: One will take only a single phase point from each 'hyper-octant' of phase space. That is, one wants all possible combinations of signs for all of the coordinates. Since each coordinate can take on two values (negative or positive), there are 2^{6N} such points. Thus, in a system having $N = 100$ particles (which is a very small system, after all) one would need to evaluate A and E at 4.15×10^{180} points! Such a process might be rather time consuming ...

The key to making this evaluation more tractable is to recognize that phase space is, for the most part, a wasteland. That is, there are enormous volumes characterized by energies that are far too high to be of any importance, e.g., regions where the positional coordinates of two different particles are such that they are substantially closer than van der Waals contact. From a mathematical standpoint, Eq. (3.6) shows that a high-energy phase point has a near-zero probability, and thus the integrand of Eq. (3.5) will also be near-zero (as long as property A does not go to infinity with increasing energy). As the integral of zero is zero, such a phase point contributes almost nothing to the property expectation value, and simply represents a waste of computational resources. So, what is needed in the evaluation of Eqs. (3.5) and (3.7) is some prescription for picking *important* (i.e., high-probability) points.

The MC method, described in Section 3.4, is a scheme designed to do exactly this in a pseudo-random fashion. Before we examine that method, however, we first consider a somewhat more intuitive way to sample 'useful' regions of phase space.

3.2.2 Properties as Time Averages of Trajectories

If we start a system at some 'reasonable' (i.e., low-energy) phase point, its energy-conserving evolution over time (i.e., its trajectory) seems *likely* to sample relevant regions of phase space.

Certainly, this is the picture most of us have in our heads when it comes to the behavior of a real system. In that case, a reasonable way to compute a property average simply involves computing the value of the property periodically at times t_i and assuming

$$\langle A \rangle = \frac{1}{M} \sum_{i}^{M} A(t_i) \tag{3.8}$$

where M is the number of times the property is sampled. In the limit of sampling continuously and following the trajectory indefinitely, this equation becomes

$$\langle A \rangle = \lim_{t \to \infty} \frac{1}{t} \int_{t_0}^{t_0+t} A(\tau) d\tau \tag{3.9}$$

The 'ergodic hypothesis' assumes Eq. (3.9) to be valid and independent of choice of t_0. It has been proven for a hard-sphere gas that Eqs. (3.5) and (3.9) are indeed equivalent (Ford 1973). No such proof is available for more realistic systems, but a large body of empirical evidence suggests that the ergodic hypothesis is valid in most molecular simulations.

This point being made, we have not yet provided a description of how to 'follow' a phase-space trajectory. This is the subject of molecular dynamics, upon which we now focus.

3.3 Molecular Dynamics

One interesting property of a phase point that has not yet been emphasized is that, since it is defined by the positions and momenta of all particles, it *determines* the location of the next phase point in the absence of outside forces acting upon the system. The word 'next' is used loosely, since the trajectory is a continuous curve of phase points (i.e., between any two points can be found another point) – a more rigorous statement is that the forward trajectory is completely determined by the initial phase point. Moreover, since time-independent Hamiltonians are necessarily invariant to time reversal, a single phase point completely determines a full trajectory. As a result, phase space trajectories cannot cross themselves (since there would then be two *different* points leading away (in both time directions) from a single point of intersection). To illuminate further some of the issues involved in following a trajectory, it is helpful to begin with an example.

3.3.1 Harmonic Oscillator Trajectories

Consider a one-dimensional classical harmonic oscillator (Figure 3.1). Phase space in this case has only two dimensions, position and momentum, and we will define the origin of this phase space to correspond to the ball of mass m being at rest (i.e., zero momentum) with the spring at its equilibrium length. This phase point represents a stationary state of the system. Now consider the dynamical behavior of the system starting from some point other than the origin. To be specific, we consider release of the ball at time t_0 from

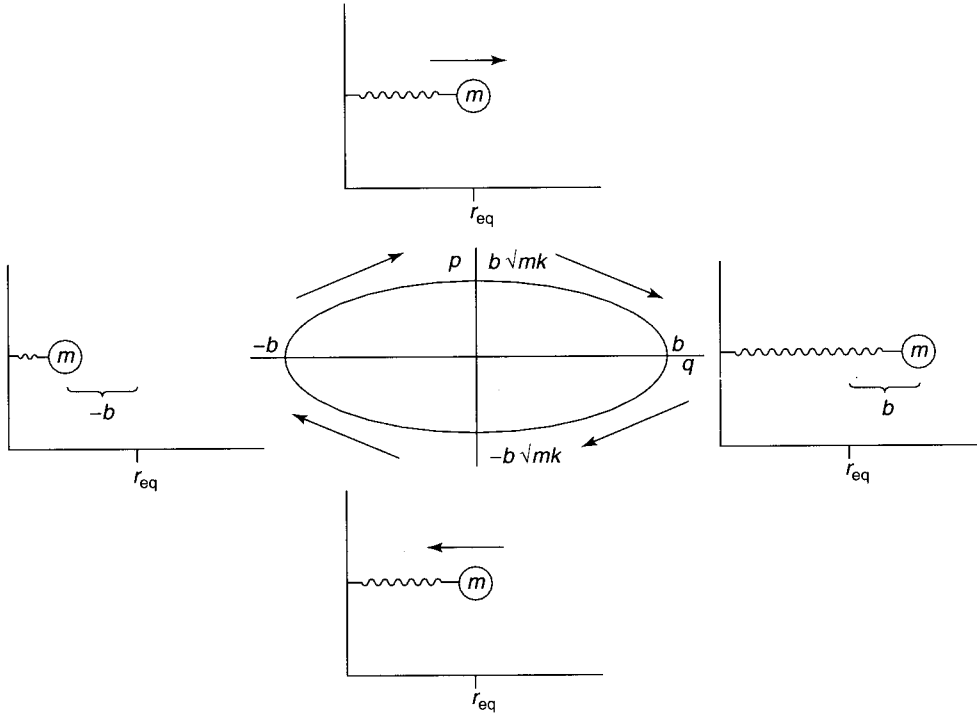


Figure 3.1 Phase-space trajectory (center) for a one-dimensional harmonic oscillator. As described in the text, at time zero the system is represented by the rightmost diagram ($q = b$, $p = 0$). The system evolves clockwise until it returns to the original point, with the period depending on the mass of the ball and the force constant of the spring

a position b length units displaced from equilibrium. The frictionless spring, characterized by force constant k, begins to contract, so that the position coordinate decreases. The momentum coordinate, which was 0 at t_0, also decreases (momentum is a vector quantity, and we here define negative momentum as movement towards the wall). As the spring passes through coordinate position 0 (the equilibrium length), the *magnitude* of the momentum reaches a maximum, and then decreases as the spring begins resisting further motion of the ball. Ultimately, the momentum drops to zero as the ball reaches position $-b$, and then grows increasingly positive as the ball moves back towards the coordinate origin. Again, after passing through the equilibrium length, the magnitude of the momentum begins to decrease, until the ball returns to the same point in phase space from which it began.

Let us consider the phase space trajectory traced out by this behavior beginning with the position vector. Over any arbitrary time interval, the relationship between two positions is

$$q(t_2) = q(t_1) + \int_{t_1}^{t_2} \frac{p(t)}{m} dt \tag{3.10}$$

where we have used the relationship between velocity and momentum

$$v = \frac{p}{m} \tag{3.11}$$

Similarly, the relationship between two momentum vectors is

$$p(t_2) = p(t_1) + m \int_{t_1}^{t_2} a(t)dt \tag{3.12}$$

where a is the acceleration. Equations (3.10) and (3.12) are Newton's equations of motion. Now, we have from Newton's Second Law

$$a = \frac{F}{m} \tag{3.13}$$

where F is the force. Moreover, from Eq. (2.13), we have a relationship between force and the position derivative of the potential energy. The simple form of the potential energy expression for a harmonic oscillator [Eq. (2.2)] permits analytic solutions for Eqs. (3.10) and (3.12). Applying the appropriate boundary conditions for the example in Figure 3.1 we have

$$q(t) = b\cos\left(\sqrt{\frac{k}{m}}t\right) \tag{3.14}$$

and

$$p(t) = -b\sqrt{mk}\sin\left(\sqrt{\frac{k}{m}}t\right) \tag{3.15}$$

These equations map out the oval phase space trajectory depicted in the figure.

Certain aspects of this phase space trajectory merit attention. We noted above that a phase space trajectory cannot cross itself. However, it *can* be periodic, which is to say it can trace out the same path again and again; the harmonic oscillator example is periodic. Note that the complete set of *all* harmonic oscillator trajectories, which would completely fill the corresponding two-dimensional phase space, is composed of concentric ovals (concentric circles if we were to choose the momentum metric to be $(mk)^{-1/2}$ times the position metric). Thus, as required, these (periodic) trajectories do not cross one another.

3.3.2 Non-analytical Systems

For systems more complicated than the harmonic oscillator, it is almost never possible to write down analytical expressions for the position and momentum components of the phase space trajectory as a function of time. However, if we *approximate* Eqs. (3.10) and (3.12) as

$$\mathbf{q}(t + \Delta t) = \mathbf{q}(t) + \frac{\mathbf{p}(t)}{m}\Delta t \tag{3.16}$$

and

$$\mathbf{p}(t + \Delta t) = \mathbf{p}(t) + m\mathbf{a}(t)\Delta t \tag{3.17}$$

(this approximation, Euler's, being exact in the limit of $\Delta t \to 0$) we are offered a prescription for *simulating* a phase space trajectory. [Note that we have switched from the scalar notation of the one-dimensional harmonic oscillator example to a more general vector notation. Note also that although the approximations in Eqs. (3.16) and (3.17) are introduced here from Eqs. (3.10) and (3.12) and the definition of the definite integral, one can also derive Eqs. (3.16) and (3.17) as Taylor expansions of $\mathbf{q}$ and $\mathbf{p}$ truncated at first order; this is discussed in more detail below.]

Thus, given a set of initial positions and momenta, and a means for computing the forces acting on each particle at any instant (and thereby deriving the acceleration), we have a formalism for 'simulating' the true phase-space trajectory. In general, initial positions are determined by what a chemist thinks is 'reasonable' – a common technique is to build the system of interest and then energy minimize it partially (since one is interested in dynamical properties, there is no point in looking for an absolute minimum) using molecular mechanics. As for initial momenta, these are usually assigned randomly to each particle subject to a temperature constraint. The relationship between temperature and momentum is

$$T(t) = \frac{1}{(3N - n)k_B} \sum_{i=1}^{N} \frac{|\mathbf{p}_i(t)|^2}{m_i} \tag{3.18}$$

where N is the total number of atoms, n is the number of constrained degrees of freedom (vide infra), and the momenta are relative to the reference frame defined by the motion of the center of mass of the system. A force field, as emphasized in the last chapter, is particularly well suited to computing the accelerations at each time step.

While the use of Eqs. (3.16) and (3.17) seems entirely straightforward, the finite time step introduces very real practical concerns. Figure 3.2 illustrates the variation of a single momentum coordinate of some arbitrary phase space trajectory, which is described by a smooth curve. When the acceleration is computed for a point on the true curve, it will be a vector tangent to the curve. If the curve is not a straight line, any mass-weighted step along the tangent (which is the process described by Eq. (3.17)) will necessarily result in a point *off* the true curve. There is no guarantee that computing the acceleration at this new point will lead to a step that ends in the vicinity of the true curve. Indeed, with each additional step, it is quite possible that we will move further and further away from the true trajectory, thereby ending up sampling non-useful regions of phase space. The problem is compounded for position coordinates, since the velocity vector being used is already only an estimate derived from Eq. (3.17), i.e., there is no guarantee that it will even be tangent to the true curve when a point on the true curve is taken. (The atomistic picture, for those finding the mathematical discussion opaque, is that if we move the atoms in a single direction over too long a time, we will begin to ram them into one another so that they are far closer than van der Waals contact. This will lead to huge repulsive forces, so that still larger atomic movements will occur over the next time step, until our system ultimately looks like a nuclear furnace, with

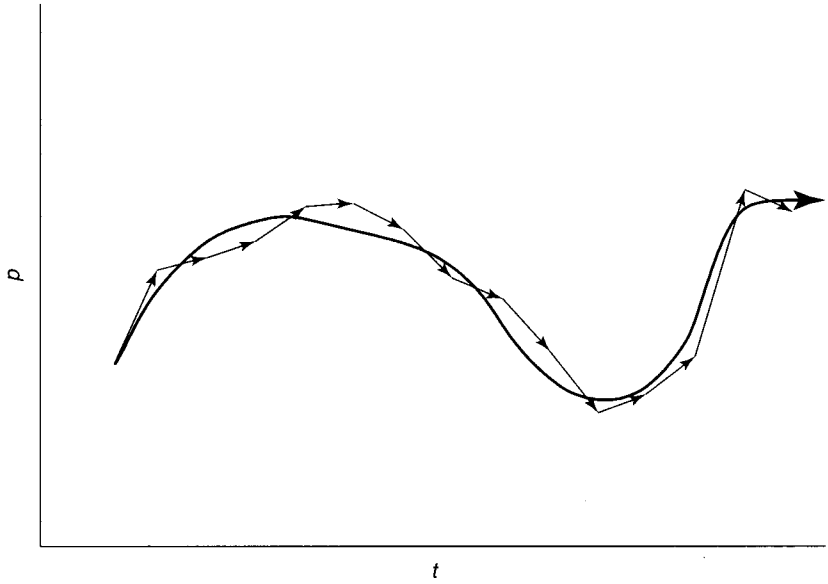


Figure 3.2 An actual phase-space trajectory (bold curve) and an approximate trajectory generated by repeated application of Eq. (3.17) (series of arrows representing individual time steps). Note that each propagation step has an identical Δt, but individual Δp values can be quite different. In the illustration, the approximate trajectory hews relatively closely to the actual one, but this will not be the case if too large a time step is used

atoms moving seemingly randomly. The very high energies of the various steps will preclude their contributing in a meaningful way to any property average.)

Of course, we know that in the limit of an infinitesimally small time step, we will recover Eqs. (3.10) and (3.12). But, since each time step requires a computation of all of the molecular forces (and, presumably, of the property we are interested in), which is computationally intensive, we do not want to take too *small* a time step, or we will not be able to propagate our trajectory for any chemically interesting length of time. What then is the optimal length for a time step that balances numerical stability with chemical utility? The general answer is that it should be at least one and preferably two orders of magnitude smaller than the fastest periodic motion within the system. To illustrate this, reconsider the 1-D harmonic oscillator example of Figure 3.1: if we estimate the first position of the mass after its release, given that the acceleration will be computed to be towards the wall, we will estimate the new position to be displaced in the negative direction. But, if we take too large a time step, i.e., we keep moving the mass towards the wall without ever accounting for the change in the acceleration of the spring with position, we might end up with the mass at a position more negative than $-b$. Indeed, we could end up with the mass behind the wall!

In a typical (classical) molecular system, the fastest motion is bond vibration which, for a heavy-atom–hydrogen bond has a period of about 10^{-14} s. Thus, for a system containing such bonds, an integration time step Δt should not much exceed 0.1 fs. This rather short time

step means that modern, large-scale MD simulations (e.g., on biopolymers in a surrounding solvent) are rarely run for more than some 10 ns of simulation time (i.e., 10^7 computations of energies, forces, etc.) That many interesting phenomena occur on the microsecond timescale or longer (e.g., protein folding) represents a severe limitation to the application of MD to these phenomena. Methods to efficiently integrate the equations of motion over longer times are the subject of substantial modern research (see, for instance, Olender and Elber 1996; Grubmuller and Tavan 1998; Feenstra, Hess and Berendsen 1999).

3.3.3 Practical Issues in Propagation

Using Euler's approximation and taking integration steps in the direction of the tangent is a particularly simple integration approach, and as such is not particularly stable. Considerably more sophisticated integration schemes have been developed for propagating trajectories. If we restrict ourselves to consideration of the position coordinate, most of these schemes derive from approximate Taylor expansions in $\mathbf{r}$, i.e., making use of

$$\mathbf{q}(t + \Delta t) = \mathbf{q}(t) + \mathbf{v}(t)\Delta t + \frac{1}{2!}\mathbf{a}(t)(\Delta t)^2 + \frac{1}{3!}\frac{d^3\mathbf{q}(\tau)}{dt^3}\bigg|_{\tau=t}(\Delta t)^3 + \cdots \tag{3.19}$$

where we have used the abbreviations $\mathbf{v}$ and $\mathbf{a}$ for the first (velocity) and second (acceleration) time derivatives of the position vector $\mathbf{q}$.

One such method, first used by Verlet (1967), considers the sum of the Taylor expansions corresponding to forward and reverse time steps Δt. In that sum, all odd-order derivatives disappear since the odd powers of Δt have opposite sign in the two Taylor expansions. Rearranging terms and truncating at second order (which is equivalent to truncating at third-order, since the third-order term has a coefficient of zero) yields

$$\mathbf{q}(t + \Delta t) = 2\mathbf{q}(t) - \mathbf{q}(t - \Delta t) + \mathbf{a}(t)(\Delta t)^2 \tag{3.20}$$

Thus, for any particle, each subsequent position is determined by the current position, the previous position, and the particle's acceleration (determined from the forces on the particle and Eq. (3.13)). For the very first step (for which no position $\mathbf{q}(t - \Delta t)$ is available) one might use Eqs. (3.16) and (3.17).

The Verlet scheme propagates the position vector with no reference to the particle velocities. Thus, it is particularly advantageous when the position coordinates of phase space are of more interest than the momentum coordinates, e.g., when one is interested in some property that is independent of momentum. However, often one wants to control the simulation temperature. This can be accomplished by scaling the particle velocities so that the temperature, as defined by Eq. (3.18), remains constant (or changes in some defined manner), as described in more detail in Section 3.6.3. To propagate the position and velocity vectors in a *coupled* fashion, a modification of Verlet's approach called the leapfrog algorithm has been proposed. In this case, Taylor expansions of the position vector truncated at second order

(not third) about $t + \Delta t/2$ are employed, in particular

$$\mathbf{q}\left(t + \frac{1}{2}\Delta t + \frac{1}{2}\Delta t\right) = \mathbf{q}\left(t + \frac{1}{2}\Delta t\right) + \mathbf{v}\left(t + \frac{1}{2}\Delta t\right)\frac{1}{2}\Delta t + \frac{1}{2!}\mathbf{a}\left(t + \frac{1}{2}\Delta t\right)\left(\frac{1}{2}\Delta t\right)^2 \tag{3.21}$$

and

$$\mathbf{q}\left(t + \frac{1}{2}\Delta t - \frac{1}{2}\Delta t\right) = \mathbf{q}\left(t + \frac{1}{2}\Delta t\right) - \mathbf{v}\left(t + \frac{1}{2}\Delta t\right)\frac{1}{2}\Delta t + \frac{1}{2!}\mathbf{a}\left(t + \frac{1}{2}\Delta t\right)\left(\frac{1}{2}\Delta t\right)^2 \tag{3.22}$$

When Eq. (3.22) is subtracted from Eq. (3.21) one obtains

$$\mathbf{q}(t + \Delta t) = \mathbf{q}(t) + \mathbf{v}\left(t + \frac{1}{2}\Delta t\right)\Delta t \tag{3.23}$$

Similar expansions for $\mathbf{v}$ give

$$\mathbf{v}\left(t + \frac{1}{2}\Delta t\right) = \mathbf{v}\left(t - \frac{1}{2}\Delta t\right) + \mathbf{a}(t)\Delta t \tag{3.24}$$

Note that in the leapfrog method, position depends on the velocities as computed one-half time step out of phase, thus, scaling of the velocities can be accomplished to control temperature. Note also that no force-field calculations actually take place for the fractional time steps. Forces (and thus accelerations) in Eq. (3.24) are computed at integral time steps, half-time-step-forward velocities are computed therefrom, and these are then used in Eq. (3.23) to update the particle positions. The drawbacks of the leapfrog algorithm include ignoring third-order terms in the Taylor expansions and the half-time-step displacements of the position and velocity vectors – both of these features can contribute to decreased stability in numerical integration of the trajectory.

Considerably more stable numerical integration schemes are known for arbitrary trajectories, e.g., Runge–Kutta (Press *et al.* 1986) and Gear predictor-corrector (Gear 1971) methods. In Runge–Kutta methods, the gradient of a function is evaluated at a number of different intermediate points, determined iteratively from the gradient at the current point, prior to taking a step to a new trajectory point on the path; the 'order' of the method refers to the number of such intermediate evaluations. In Gear predictor-corrector algorithms, higher order terms in the Taylor expansion are used to predict steps along the trajectory, and then the actual particle accelerations computed for those points are compared to those that were predicted by the Taylor expansion. The differences between the actual and predicted values are used to correct the position of the point on the trajectory. While Runge–Kutta and Gear predictor-corrector algorithms enjoy very high stability, they find only limited use in MD simulations because of the high computational cost associated with computing multiple first derivatives, or higher-order derivatives, for every step along the trajectory.

A different method of increasing the time step without decreasing the numerical stability is to remove from the system those degrees of freedom having the highest frequency (assuming,

of course, that any property being studied is independent of those degrees of freedom). Thus, if heavy-atom–hydrogen bonds are constrained to remain at a constant length, the next highest frequency motions will be heavy-atom–heavy-atom vibrations; these frequencies are typically a factor of 2–5 smaller in magnitude. While a factor of 2 is of only marginal utility, reducing the number of available degrees of freedom generally offers some savings in time and integration stability. So, when the system of interest is some solute immersed in a large bath of surrounding solvent molecules, it can be advantageous to freeze some or all of the degrees of freedom within the solvent molecules.

A commonly employed algorithm for eliminating these degrees of freedom is called SHAKE (Ryckaert, Ciccotti, and Berendsen 1977). In the context of the Verlet algorithm, the formalism for freezing bond lengths involves defining distance constraints d_{ij} between atoms i and j according to

$$|\mathbf{r}_{ij}|^2 - d_{ij}^2 = 0 \tag{3.25}$$

where $\mathbf{r}_{ij}$ is the instantaneous interatomic distance vector. The position constraints can be applied iteratively in the Verlet algorithm, for example, by first taking an *unconstrained* step according to Eq. (3.20). The constraints are then taken account of according to

$$\mathbf{r}_i(t + \Delta t) = \mathbf{r}_i^0(t + \Delta t) + \Delta\mathbf{r}_i(t) \tag{3.26}$$

where $\mathbf{r}_i^0(t + \Delta t)$ is the position after taking the unconstrained step, and $\Delta\mathbf{r}_i(t)$ is the displacement vector required to satisfy a set of coupled constraint equations. These equations are defined as

$$\Delta\mathbf{r}_i(t) = \frac{2(\Delta t)^2}{m_i} \sum_j \lambda_{ij}\mathbf{r}_{ij}(t) \tag{3.27}$$

where the Lagrange multipliers λ_{ij} are determined iteratively following substitution of Eqs. (3.25) and (3.26) into Eq. (3.20).

Finally, there are a number of entirely mundane (but still very worthwhile!) steps that can be taken to reduce the total computer time required for a MD simulation. As a single example, note that any force on a particle derived from a force-field non-bonded energy term is induced by some *other* particle (i.e., the potential is pairwise). Newton's Third Law tells us that

$$\mathbf{F}_{ij} = -\mathbf{F}_{ji} \tag{3.28}$$

so we can save roughly a factor of two in computing the non-bonded forces by only evaluating terms for $i < j$ and using Eq. (3.28) to establish the rest.

3.3.4 Stochastic Dynamics

When the point of a simulation is not to determine accurate thermodynamic information about an ensemble, but rather to watch the dynamical evolution of some particular system immersed in a larger system (e.g., a solute in a solvent), then significant computational savings can be

had by modeling the larger system stochastically. That is, the explicit nature of the larger system is ignored, and its influence is made manifest by a continuum that interacts with the smaller system, typically with that influence including a degree of randomness.

In Langevin dynamics, the equation of motion for each particle is

$$\mathbf{a}(t) = -\zeta \mathbf{p}(t) + \frac{1}{m}[\mathbf{F}_{\text{intra}}(t) + \mathbf{F}_{\text{continuum}}(t)] \tag{3.29}$$

where the continuum is characterized by a microscopic friction coefficient, ζ, and a force, $\mathbf{F}$, having one or more components (e.g., electrostatic and random collisional). Intramolecular forces are evaluated in the usual way from a force field. Propagation of position and momentum vectors proceeds in the usual fashion.

In Brownian dynamics, the momentum degrees of freedom are removed by arguing that for a system that does not change shape much over very long timescales (e.g., a molecule, even a fairly large one) the momentum of each particle can be approximated as zero relative to the rotating center of mass reference frame. Setting the l.h.s. of Eq. (3.29) to zero and integrating, we obtain the Brownian equation of motion

$$\mathbf{r}(t) = \mathbf{r}(t_0) + \frac{1}{\zeta}\int_{t_0}^{t} [\mathbf{F}_{\text{intra}}(\tau) + \mathbf{F}_{\text{continuum}}(\tau)]\mathrm{d}\tau \tag{3.30}$$

where we now propagate only the position vector.

Langevin and Brownian dynamics are very efficient because a potentially very large surrounding medium is represented by a simple continuum. Since the computational time required for an individual time step is thus reduced compared to a full deterministic MD simulation, much longer timescales can be accessed. This makes stochastic MD methods quite attractive for studying system properties with relaxation times longer than those that can be accessed with deterministic MD simulations. Of course, if those properties involve the surrounding medium in some explicit way (e.g., a radial distribution function involving solvent molecules, vide infra), then the stochastic MD approach is not an option.

3.4 Monte Carlo

3.4.1 Manipulation of Phase-space Integrals

If we consider the various MD methods presented above, the Langevin and Brownian dynamics schemes introduce an increasing degree of stochastic behavior. One may imagine carrying this stochastic approach to its logical extreme, in which event there are no equations of motion to integrate, but rather phase points for a system are selected entirely at random. As noted above, properties of the system can then be determined from Eq. (3.5), but the integration converges very slowly because most randomly chosen points will be in chemically meaningless regions of phase space.

One way to reduce the problem slightly is to recognize that for many properties A, the position and momentum dependences of A are separable. In that case, Eq. (3.5) can be written as

$$\langle A \rangle = \int A(\mathbf{q}) \left[\int P(\mathbf{p},\mathbf{q}) d\mathbf{p} \right] d\mathbf{q} + \int A(\mathbf{p}) \left[\int P(\mathbf{p},\mathbf{q}) d\mathbf{q} \right] d\mathbf{p} \tag{3.31}$$

Since the Hamiltonian is also separable, the integrals in brackets on the r.h.s. of Eq. (3.31) may be simplified and we write

$$\langle A \rangle = \int A(\mathbf{q}) P(\mathbf{q}) d\mathbf{q} + \int A(\mathbf{p}) P(\mathbf{p}) d\mathbf{p} \tag{3.32}$$

where $P(\mathbf{q})$ and $P(\mathbf{p})$ are probability functions analogous to Eq. (3.6) related only to the potential and kinetic energies, respectively. Thus, we reduce the problem of evaluating a $6N$-dimensional integral to the problem of evaluating two $3N$-dimensional integrals. Of course, if the property is independent of either the position or momentum variables, then there is only one $3N$-dimensional integral to evaluate.

Even with so large a simplification, however, the convergence of Eq. (3.32) for a realistically sized chemical system and a random selection of phase points is too slow to be useful. What is needed is a scheme to select important phase points in a biased fashion.

3.4.2 Metropolis Sampling

The most significant breakthrough in Monte Carlo modeling took place when Metropolis *et al.* (1953) described an approach where 'instead of choosing configurations randomly, then weighting them with $\exp(-E/k_B T)$, we choose configurations with a probability $\exp(-E/k_B T)$ and weight them evenly'.

For convenience, let us consider a property dependent only on position coordinates. Expressing the elegantly simple Metropolis idea mathematically, we have

$$\langle A \rangle = \frac{1}{X} \sum_{i=1}^{X} A(\mathbf{q}_i) \tag{3.33}$$

where X is the total number of points $\mathbf{q}$ sampled according to the Metropolis prescription. Note the remarkable similarity between Eq. (3.33) and Eq. (3.8). Equation (3.33) resembles an ensemble average from an MD trajectory where the order of the points, i.e., the temporal progression, has been lost. Not surprisingly, as time does not enter into the MC scheme, it is not possible to establish a time relationship between points.

The Metropolis prescription dictates that we choose points with a Boltzmann-weighted probability. The typical approach is to begin with some 'reasonable' configuration $\mathbf{q}_1$. The value of property A is computed as the first element of the sum in Eq. (3.33), and then $\mathbf{q}_1$ is randomly perturbed to give a new configuration $\mathbf{q}_2$. In the constant particle number, constant

volume, constant temperature ensemble (*NVT* ensemble), the probability p of 'accepting' point $\mathbf{q}_2$ is

$$p = \min\left[1, \frac{\exp(-E_2/k_B T)}{\exp(-E_1/k_B T)}\right] \tag{3.34}$$

Thus, if the energy of point $\mathbf{q}_2$ is not higher than that of point $\mathbf{q}_1$, the point is always accepted. If the energy of the second point *is* higher than the first, p is compared to a random number z between 0 and 1, and the move is accepted if $p \geq z$. Accepting the point means that the value of A is calculated for that point, that value is added to the sum in Eq. (3.33), and the entire process is repeated. If second point is *not* accepted, then the first point 'repeats', i.e., the value of A computed for the first point is added to the sum in Eq. (3.33) a second time and a new, random perturbation is attempted. Such a sequence of phase points, where each new point depends only on the immediately preceding point, is called a 'Markov chain'.

The art of running an MC calculation lies in defining the perturbation step(s). If the steps are very, very small, then the volume of phase space sampled will increase only slowly over time, and the cost will be high in terms of computational resources. If the steps are too large, then the rejection rate will grow so high that again computational resources will be wasted by an inefficient sampling of phase space. Neither of these situations is desirable.

In practice, MC simulations are primarily applied to collections of molecules (e.g., molecular liquids and solutions). The perturbing step involves the choice of a single molecule, which is randomly translated and rotated in a Cartesian reference frame. If the molecule is flexible, its internal geometry is also randomly perturbed, typically in internal coordinates. The ranges on these various perturbations are adjusted such that 20–50% of attempted moves are accepted. Several million individual points are accumulated, as described in more detail in Section 3.6.4.

Note that in the MC methodology, only the energy of the system is computed at any given point. In MD, by contrast, forces are the fundamental variables. Pangali, Rao, and Berne (1978) have described a sampling scheme where forces are used to choose the direction(s) for molecular perturbations. Such a force-biased MC procedure leads to higher acceptance rates and greater statistical precision, but at the cost of increased computational resources.

3.5 Ensemble and Dynamical Property Examples

The range of properties that can be determined from simulation is obviously limited only by the imagination of the modeler. In this section, we will briefly discuss a few typical properties in a general sense. We will focus on structural and time-correlation properties, deferring thermodynamic properties to Chapters 10 and 12.

As a very simple example, consider the dipole moment of water. In the gas phase, this dipole moment is 1.85 D (Demaison, Hütner, and Tiemann 1982). What about water in liquid water? A zeroth order approach to answering this problem would be to create a molecular mechanics force field defining the water molecule (a sizable number exist) that gives the correct dipole moment for the isolated, gas-phase molecule at its equilibrium

geometry, which moment is expressed as

$$\mu = \sum_{i=1}^{3} \frac{q_i}{\mathbf{r}_i} \tag{3.35}$$

where the sum runs over the one oxygen and two hydrogen atoms, q_i is the partial atomic charge assigned to atom i, and $\mathbf{r}_i$ is the position of atom i (since the water molecule has no net charge, the dipole moment is independent of the choice of origin for $\mathbf{r}$). In a liquid simulation (see Section 3.6.1 for more details on simulating condensed phases), the expectation value of the moment would be taken over *all* water molecules. Since the liquid is isotropic, we are not interested in the average vector, but rather the average magnitude of the vector, i.e.,

$$\langle|\mu|\rangle = \frac{1}{N} \sum_{n=1}^{N} \left| \sum_{i=1}^{3} \frac{q_{i,n}}{\mathbf{r}_{i,n}} \right| \tag{3.36}$$

where N is the number of water molecules in the liquid model. Then, to the extent that in liquid water the average geometry of a water molecule changes from its gas-phase equilibrium structure, the expectation value of the magnitude of the dipole moment will reflect this change. Note that Eq. (3.36) gives the ensemble average for a single snapshot of the system; that is, the 'ensemble' that is being averaged over is intrinsic to each phase point by virtue of their being multiple copies of the molecule of interest. By MD or MC methods, we would generate multiple snapshots, either as points along an MD trajectory or by MC perturbations, so that we would finally have

$$\langle|\mu|\rangle = \frac{1}{M \bullet N} \sum_{m=1}^{M} \sum_{n=1}^{N} \left| \sum_{i=1}^{3} \frac{q_{i,n,m}}{\mathbf{r}_{i,n,m}} \right| \tag{3.37}$$

where M is the total number of snapshots. [If we were considering the dipole moment of a solute molecule that was present in only one copy (i.e., a dilute solution), then the sum over N would disappear.]

Note that the expectation value compresses an enormous amount of information into a single value. A more complete picture of the moment would be a probability distribution, as depicted in Figure 3.3. In this analysis, the individual water dipole moment magnitudes (all $M \bullet N$ of them) are collected into bins spanning some range of dipole moments. The moments are then plotted either as a histogram of the bins or as a smooth curve reflecting the probability of being in an individual bin (i.e., equivalent to drawing the curve through the midpoint of the top of each histogram bar). The width of the bins is chosen so as to give maximum resolution to the lineshape of the curve without introducing statistical noise from underpopulation of individual bins.

Note that, although up to this point we have described the expectation value of A as though it were a scalar value, it is also possible that A is a function of some experimentally (and computationally) accessible variable, in which case we may legitimately ask about its expectation value at various points along the axis of its independent variable. A good

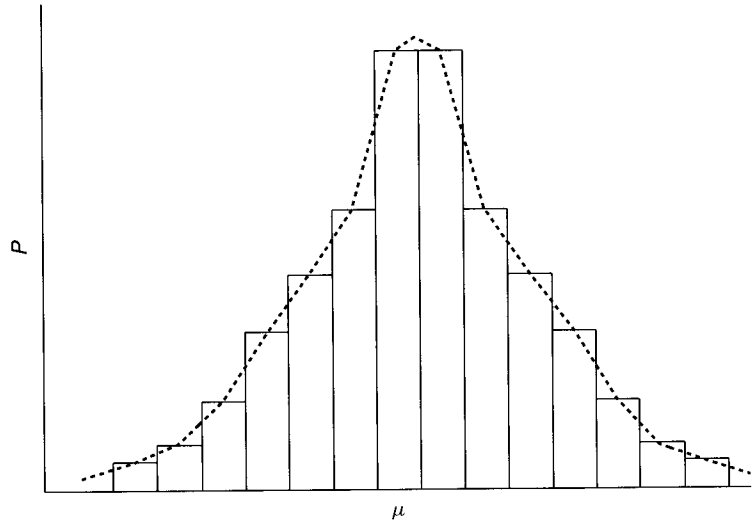


Figure 3.3 Hypothetical distribution of dipole moment magnitudes from a simulation of liquid water. The dashed curve is generated by connecting the tops of histogram bins whose height is dictated by the number of water molecules found to have dipole moments in the range spanned by the bin. Note that although the example is illustrated to be symmetric about a central value (which will thus necessarily be $\langle\mu\rangle$) this need not be the case

example of such a property is a radial distribution function (r.d.f.), which can be determined experimentally from X-ray or neutron diffraction measurements. The r.d.f. for two atoms A and B in a spherical volume element is defined by

$$\frac{1}{V}g_{AB}(r) = \frac{1}{N_A \bullet N_B}\left\langle \sum_{i=1}^{N_A}\sum_{j=1}^{N_B} \delta\left[r - r_{A_iB_j}\right]\right\rangle \tag{3.38}$$

where V is the volume, N is the total number of atoms of a given type within the volume element, δ is the Dirac delta function (the utility of which will become apparent momentarily), and r is radial distance. The double summation within the ensemble average effectively counts for each distance r the number of AB pairs separated by that distance. If we integrate over the full spherical volume, we obtain

$$\begin{aligned}\frac{1}{V}\int g_{AB}(r)d\mathbf{r} &= \frac{1}{N_A \bullet N_B}\left\langle \sum_{i=1}^{N_A}\sum_{j=1}^{N_B} \int \delta\left[r - r_{A_iB_j}\right] d\mathbf{r}\right\rangle \\ &= 1\end{aligned} \tag{3.39}$$

where we have made use of the property of the Dirac delta that its integral is unity. As there are $N_A \bullet N_B$ contributions of unity to the quantity inside the ensemble average, the r.h.s. of Eq. (3.39) is 1, and we see that the $1/V$ term is effectively a normalization constant on g.

We may thus interpret the l.h.s. of Eq. (3.39) as a probability function. That is, we may express the probability of finding two atoms of A and B within some range Δr of distance r from one another as

$$P\{\mathrm{A}, \mathrm{B}, r, \Delta r\} = \frac{4\pi r^2}{V} g_{\mathrm{AB}}(r)\Delta r \tag{3.40}$$

where, in the limit of small Δr, we have approximated the integral as $g_{\mathrm{AB}}(r)$ times the volume of the thin spherical shell $4\pi r^2 \Delta r$.

Note that its contribution to the probability function makes certain limiting behaviors on $g_{\mathrm{AB}}(r)$ intuitively obvious. For instance, the function should go to zero very rapidly when r becomes less than the sum of the van der Waals radii of A and B. In addition, at very large r, the function should be independent of r in homogeneous media, like fluids, i.e., there should be an equal probability for any interatomic separation because the two atoms no longer influence one another's positions. In that case, we could move g outside the integral on the l.h.s. of Eq. (3.39), and then the normalization makes it apparent that $g = 1$ under such conditions. Values other than 1 thus indicate some kind of structuring in a medium – values greater than 1 indicate preferred locations for surrounding atoms (e.g., a solvation shell) while values below 1 indicate underpopulated regions. A typical example of a liquid solution r.d.f. is shown in Figure 3.4. Note that with increasing order, e.g., on passing from a liquid to a solid phase, the peaks in g become increasingly narrow and the valleys increasingly wide and near zero, until in the limit of a motionless, perfect crystal, g would be a spectrum of Dirac δ functions positioned at the lattice spacings of the crystal.

It often happens that we consider one of our atoms A or B to be privileged, e.g., A might be a sodium ion and B the oxygen atom of a water and our interests might focus

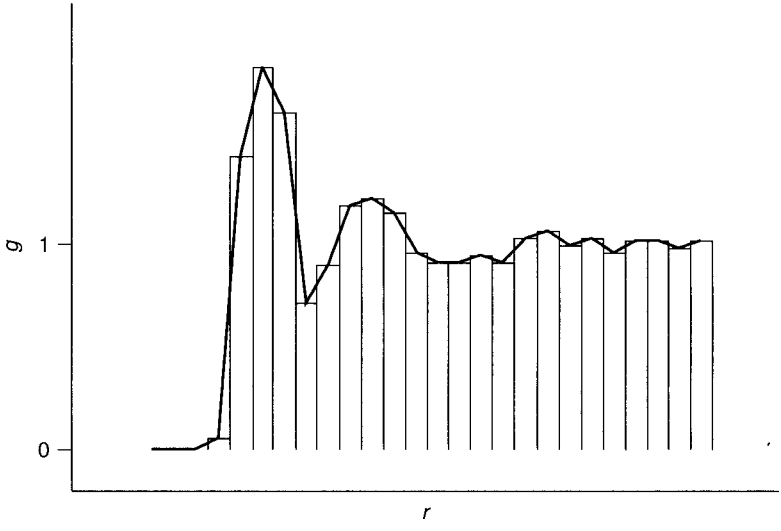


Figure 3.4 A radial distribution function showing preferred ($g > 1$) and disfavored ($g < 1$) interparticle distances. Random fluctuation about $g = 1$ is observed at large r

on describing the solvation structure of water about sodium ions in general. Then, we can define the total number of oxygen atoms n_B within some distance range about *any* sodium ion (atom A) as

$$n_B\{r, \Delta r\} = N_B P\{A, B, r, \Delta r\} \tag{3.41}$$

We may then use Eq. (3.40) to write

$$n_B\{r, \Delta r\} = 4\pi r^2 \rho_B g_{AB}(r)\Delta r \tag{3.42}$$

where ρ_B is the number density of B in the total spherical volume. Thus, if instead of $g_{AB}(r)$ we plot $4\pi r^2 \rho_B g_{AB}(r)$, then the area under the latter curve provides the number of molecules of B for arbitrary choices of r and Δr. Such an integration is typically performed for the distinct peaks in $g(r)$ so as to determine the number of molecules in the first, second, and possibly higher solvation shells or the number of nearest neighbors, next-nearest neighbors, etc., in a solid.

Determining $g(r)$ from a simulation involves a procedure quite similar to that described above for determining the continuous distribution of a scalar property. For each snapshot of an MD or MC simulation, all A–B distances are computed, and each occurrence is added to the appropriate bin of a histogram running from $r = 0$ to the maximum radius for the system (e.g., one half the narrowest box dimension under periodic boundary conditions, vide infra). Normalization now requires taking account not only of the total number of atoms A and B, but also the number of snapshots, i.e.,

$$g_{AB}(r) = \frac{V}{4\pi r^2 \Delta r M N_A N_B} \sum_{m=1}^{M} \sum_{i=1}^{N_A} \sum_{j=1}^{N_B} Q_m\left(r; r_{A_i B_j}\right) \tag{3.43}$$

where Δr is the width of a histogram bin, M is the total number of snapshots, and Q_m is the counting function

$$Q\left(r; r_{A_i B_j}\right) = \begin{cases} 1 & \text{if } r - \Delta r/2 \le r_{A_i B_j} < r + \Delta r/2 \\ 0 & \text{otherwise} \end{cases} \tag{3.44}$$

for snapshot m.

The final class of dynamical properties we will consider are those defined by time-dependent autocorrelation functions. Such a function is defined by

$$C(t) = \langle a(t_0)a(t_0 + t)\rangle_{t_0} \tag{3.45}$$

where the ensemble average runs over *time* snapshots, and hence can only be determined from MD, *not* MC. Implicit in Eq. (3.45) is the assumption that C does not depend on the value of t_0 (since the ensemble average is over different choices of this quantity), and this will only be true for a system at equilibrium. The autocorrelation function provides a measure of the degree to which the value of property a at one time influences the value at a later time. An autocorrelation function attains its maximum value for a time delay of zero (i.e.,

no time delay at all), and this quantity, $\langle a^2 \rangle$ (which *can* be determined from MC simulations since no time correlation is involved) may be regarded as a normalization constant.

Now let us consider the behavior of C for long time delays. In a system where property a is not periodic in time, like a typical chemical system subject to effectively random thermal fluctuations, two measurements separated by a sufficiently long delay time should be completely uncorrelated. If two properties x and y are uncorrelated, then $\langle xy \rangle$ is equal to $\langle x \rangle \langle y \rangle$, so at long times C decays to $\langle a \rangle^2$.

While notationally burdensome, the discussion above makes it somewhat more intuitive to consider a reduced autocorrelation function defined by

$$\hat{C}(t) = \frac{\langle [a(t_0) - \langle a \rangle][a(t_0 + t) - \langle a \rangle] \rangle_{t_0}}{\langle [a - \langle a \rangle]^2 \rangle} \tag{3.46}$$

which is normalized and, because the arguments in brackets fluctuate about their mean (and thus have individual expectation values of zero) decays to zero at long delay times. Example autocorrelation plots are provided in Figure 3.5. The curves can be fit to analytic expressions to determine characteristic decay times. For example, the characteristic decay time for an autocorrelation curve that can be fit to $\exp(-\zeta t)$ is ζ^{-1} time units.

Different properties have different characteristic decay times, and these decay times can be quite helpful in deciding how long to run a particular MD simulation. Since the point of a simulation is usually to obtain a statistically meaningful sample, one does not want to compute an average over a time shorter than several multiples of the characteristic decay time.

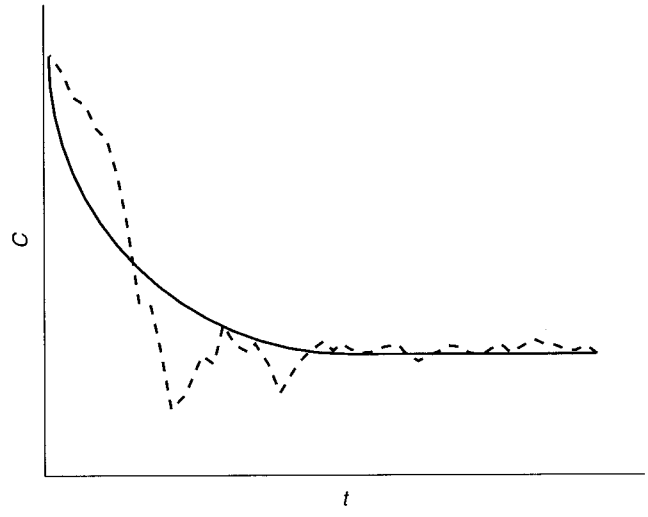


Figure 3.5 Two different autocorrelation functions. The solid curve is for a property that shows no significant statistical noise and appears to be well characterized by a single decay time. The dashed curve is quite noisy and, at least initially, shows a slower decay behavior. In the absence of a very long sample, decay times can depend on the total time sampled as well

As for the properties themselves, there are many chemically useful autocorrelation functions. For instance, particle position or velocity autocorrelation functions can be used to determine diffusion coefficients (Ernst, Hauge, and van Leeuwen 1971), stress autocorrelation functions can be used to determine shear viscosities (Haile 1992), and dipole autocorrelation functions are related to vibrational (infrared) spectra as their reverse Fourier transforms (Berens and Wilson 1981). There are also many useful correlation functions between two *different* variables (Zwanzig 1965). A more detailed discussion, however, is beyond the scope of this text.

3.6 Key Details in Formalism

The details of MC and MD methods laid out thus far can realistically be applied in a rigorous fashion only to systems that are too small to meaningfully represent actual chemical systems. In order to extend the technology in such a way as to make it useful for interpreting (or predicting) chemical phenomena, a few other approximations, or practical simplifications, are often employed. This is particularly true for the modeling of condensed phases, which are macroscopic in character.

3.6.1 Cutoffs and Boundary Conditions

As a spherical system increases in size, its volume grows as the cube of the radius while its surface grows as the square. Thus, in a truly macroscopic system, surface effects may play little role in the chemistry under study (there are, of course, exceptions to this). However, in a typical simulation, computational resources inevitably constrain the size of the system to be so small that surface effects may *dominate* the system properties. Put more succinctly, the modeling of a cluster may not tell one much about the behavior of a macroscopic system. This is particularly true when electrostatic interactions are important, since the energy associated with these interactions has an r^{-1} dependence.

One approach to avoid cluster artifacts is the use of 'periodic boundary conditions' (PBCs). Under PBCs, the system being modeled is assumed to be a unit cell in some ideal crystal (e.g., cubic or orthorhombic, see Theodorouo and Suter 1985). In practice, cut-off distances are usually employed in evaluating non-bonded interactions, so the simulation cell need be surrounded by only one set of nearest neighbors, as illustrated in Figure 3.6. If the trajectory of an individual atom (or a MC move of that atom) takes it outside the boundary of the simulation cell in any one or more cell coordinates, its image simultaneously enters the simulation cell from the point related to the exit location by lattice symmetry.

Thus, PBCs function to preserve mass, particle number, and, it can be shown, total energy in the simulation cell. In an MD simulation, PBCs also conserve linear momentum; since linear momentum is *not* conserved in real contained systems, where container walls disrupt the property, this is equivalent to reducing the number of degrees of freedom by 3. However, this effect on system properties is typically negligible for systems of over 100 atoms. Obviously, PBCs do *not* conserve angular momentum in the simulation cell of an MD simulation, but over time the movement of atoms in and out of each wall of the cell will be such that

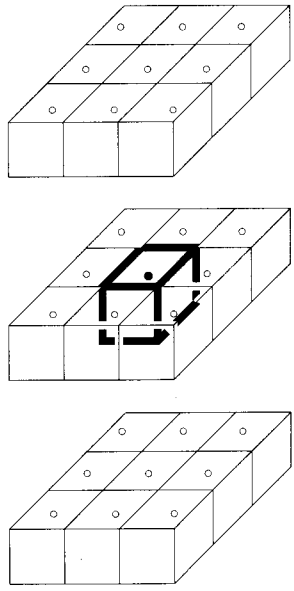


Figure 3.6 Exploded view of a cubic simulation cell surrounded by the 26 periodic images generated by PBCs. If the solid particle translates to a position that is outside the simulation cell, one of its periodic images, represented by open particles, will translate in

fluctuations will take place about a well-defined average. The key aspect of imposing PBCs is that no molecule within the simulation cell sees 'vacuum' within the range of its interaction cutoffs, and thus surface artifacts are avoided. Other artifacts associated with periodicity may be introduced, particularly with respect to correlation times in dynamics simulations (Berne and Harp 1970), but these can in principle be eliminated by moving to larger and larger simulation cells so that periodicity takes place over longer and longer length scales.

Of course, concerns about periodicity only relate to systems that are *not* periodic. The discussion above pertains primarily to the simulations of liquids, or solutes in liquid solutions, where PBCs are a useful approximation that helps to model solvation phenomena more realistically than would be the case for a small cluster. If the system truly is periodic, e.g., a zeolite crystal, then PBCs are integral to the model. Moreover, imposing PBCs can provide certain advantages in a simulation. For instance, Ewald summation, which accounts for electrostatic interactions to infinite length as discussed in Chapter 2, can only be carried out within the context of PBCs.

An obvious question with respect to PBCs is how large the simulation cell should be. The simple answer is that all cell dimensions must be at least as large as the largest cut-off length employed in the simulation. Otherwise, some interatomic interactions would be at least double counted (once within the cell, and once with an image outside of the cell). In practice, one would like to go well beyond this minimum requirement if the system being modeled is supposedly homogeneous and non-periodic. Thus, for instance, if one is modeling a large, dilute solute in a solvent (e.g., a biomolecule), a good choice for cell size might be

the dimensions of the molecule plus at least twice the largest cut-off distance. Thus, no two solute molecules interact with one another nor does any solvent molecule see two copies of the solute. (Note, however, that this does not change the fundamentally periodic nature of the system; it simply increases the number of molecules over which it is made manifest.)

As already noted in Chapter 2, for electrostatic interactions, Ewald sums are generally to be preferred over cut-offs because of the long-range nature of the interactions. For van der Waals type terms, cut-offs do not introduce significant artifacts provided they are reasonably large (typically 8–12 Å).

Because of the cost of computing interatomic distances, the evaluation of non-bonded terms in MD is often handled with the aid of a 'pairlist', which holds in memory all pairs of atoms within a given distance of one another. The pairlist is updated periodically, but less often than every MD step. Note that a particular virtue of MC compared to MD is that the only changes in the potential energy are those associated with a moved particle – all other interactions remain constant. This makes evaluation of the total energy a much simpler process in MC.

3.6.2 Polarization

As noted in Chapter 2, computation of charge–charge (or dipole–dipole) terms is a particularly efficient means to evaluate electrostatic interactions because it is pairwise additive. However, a more realistic picture of an actual physical system is one that takes into account the polarization of the system. Thus, different regions in a simulation (e.g., different functional groups, or different atoms) will be characterized by different local polarizabilities, and the local charge moments, by adjusting in an iterative fashion to their mutual interactions, introduce many-body effects into a simulation.

Simulations including polarizability, either only on solvent molecules or on all atoms, have begun to appear with greater frequency as computational resources have grown larger. In addition, significant efforts have gone into introducing polarizability into force fields in a general way by replacing fixed atomic charges with charges that fluctuate based on local environment (Winn, Ferenczy and Reynolds 1999; Banks *et al.* 1999), thereby preserving the simplicity of a pairwise interaction potential. However, it is not yet clear that the greater 'realism' afforded by a polarizable model greatly improves the accuracy of simulations. There are certain instances where polarizable force fields seem better suited to the modeling problem. For instance, Dang *et al.* (1991) have emphasized that the solvation of ions, because of their concentrated charge, is more realistically accounted for when surrounding solvent molecules are polarizable and Soetens *et al.* (1997) have emphasized its importance in the computation of ion–ion interaction potentials for the case of two guanidinium ions in water.

In general, however, the majority of properties do not yet seem to be more accurately predicted by polarizable models than by unpolarizable ones, provided adequate care is taken in the parameterization process. Of course, if one wishes to examine issues associated with polarization, it must necessarily be included in the model. In the area of solvents, for instance, Bernardo *et al.* (1994) and Zhu and Wong (1994) have carefully studied the properties of polarizable water models. In addition, Gao, Habibollazadeh, and Shao (1995) have developed

alcohol force fields reproducing the thermodynamic properties of these species as liquids with a high degree of accuracy, and have computed the polarization contribution to the total energy of the liquids to be 10–20%.

However, the typically high cost of including polarization is not attractive. Jorgensen has argued against the utility of including polarization in most instances, and has shown that bulk liquid properties can be equally well reproduced by fixed-charge force fields given proper care in the parameterization process (see, for instance, Mahoney and Jorgensen 2000). A particularly interesting example is provided by the simple amines ammonia, methylamine, dimethylamine, and trimethylamine. In the gas phase, the basicity of these species increases with increasing methylation in the expected fashion. In water, however, solvation effects compete with intrinsic basicity so that the four amines span a fairly narrow range of basicity, with methylamine being the most basic and trimethylamine and ammonia the least. Many models of solvation (see Chapters 11 and 12 for more details on solvation models) have been applied to this problem, and the failure of essentially all of them to correctly predict the basicity ordering led to the suggestion that in the case of explicit models, the failure derived from the use of non-polarizable force fields. Rizzo and Jorgensen (1999), however, parameterized non-polarizable classical models for the four amines that accurately reproduced their liquid properties and then showed that they further predicted the correct basicity ordering in aqueous simulations, thereby refuting the prior suggestion. [As a point of philosophy, the above example provides a nice illustration that a model's failure to accurately predict a particular quantity does *not* necessarily imply that a more expensive model needs to be developed – sometimes all that is required is a more careful parameterization of the existing model.] At least for the moment, then, it appears that errors associated with other aspects of simulation technology typically continue to be as large or larger than any errors introduced by use of non-polarizable force fields, so the use of such force fields in everyday simulations seems likely to continue for some time.

3.6.3 Control of System Variables

Our discussion of MD above was for the 'typical' MD ensemble, which holds particle number, system volume, and total energy constant – the *NVE* or 'microcanonical' ensemble. Often, however, there are other thermodynamic variables that one would prefer to hold constant, e.g., temperature. As temperature is related to the total kinetic energy of the system (if it is at equilibrium), as detailed in Eq. (3.18), one could in principle scale the velocities of each particle at each step to maintain a constant temperature. In practice, this is undesirable because the adjustment of the velocities, occasionally by fairly significant scaling factors, causes the trajectories to be no longer Newtonian. Properties computed over such trajectories are less likely to be reliable. An alternative method, known as Berendsen coupling (Berendsen *et al.* 1984), slows the scaling process by envisioning a connection between the system and a surrounding bath that is at a constant temperature T_0. Scaling of each particle velocity is accomplished by including a dissipative Langevin force in the equations of motion according to

$$\mathbf{a}_i(t) = \frac{\mathbf{F}_i(t)}{m_i} + \frac{\mathbf{p}_i(t)}{m_i \tau}\left[\frac{T_0}{T(t)} - 1\right] \tag{3.47}$$

where $T(t)$ is the instantaneous temperature, and τ has units of time and is used to control the strength of the coupling. The larger the value of τ the smaller the perturbing force and the more slowly the system is scaled to T_0 (i.e., τ is an effective relaxation time).

Note that, to start an MD simulation, one must necessarily generate an initial snapshot. It is essentially impossible for a chemist to simply 'draw' a large system that actually corresponds to a high-probability region of phase space. Thus, most MD simulations begin with a so-called 'equilibration' period, during which time the system is allowed to relax to a realistic configuration, after which point the 'production' portion of the simulation begins, and property averages are accumulated. A temperature coupling is often used during the equilibration period so that the temperature begins very low (near zero) and eventually ramps up to the desired system temperature for the production phase. This has the effect of damping particle movement early on in the equilibration (when there are presumably very large forces from a poor initial guess at the geometry).

In practice, equilibration protocols can be rather involved. Large portions of the system may be held frozen initially while subregions are relaxed. Ultimately, the entire system is relaxed (i.e., all the degrees of freedom that are being allowed to vary) and, once the equilibration temperature has reached the desired average value, one can begin to collect statistics.

With respect to other thermodynamic variables, many experimental systems are not held at constant volume, but instead at constant pressure. Assuming ideal gas statistical mechanics and pairwise additive forces, pressure P can be computed as

$$P(t) = \frac{1}{V(t)}\left[Nk_{\mathrm{B}}T(t) + \frac{1}{3}\sum_{i}^{N}\sum_{j>1}^{N} F_{ij}r_{ij}\right] \tag{3.48}$$

where V is the volume, N is the number of particles, F and r are the forces and distances between particles, respectively. To adjust the pressure in a simulation, what is typically modified is the volume. This is accomplished by scaling the location of the particles, i.e., changing the size of the unit cell in a system with PBCs. The scaling can be accomplished in a fashion exactly analogous with Eq. (3.47) (Andersen 1980).

An alternative coupling scheme for temperature and pressure, the Nosé–Hoover scheme, adds new, independent variables that control these quantities to the simulation (Nosé 1984; Hoover 1985). These variables are then propagated along with the position and momentum variables.

In MC methods, the 'natural' ensemble is the *NVT* ensemble. Carrying out MC simulations in other ensembles simply requires that the probabilities computed for steps to be accepted or rejected reflect dependence on factors other than the internal energy. Thus, if we wish to maintain constant pressure instead of constant volume, we can treat volume as a variable (again, by scaling the particle coordinates, which is equivalent to expanding or contracting the unit cell in a system described by PBCs). However, in the *NPT* ensemble, the deterministic thermodynamic variable is no longer the internal energy, but the enthalpy (i.e., $E + PV$) and, moreover, we must account for the effect of a change in system volume (three dimensions) on the total volume of phase space ($3N$ dimensions for position) since probability is related

to phase-space volume. Thus, in the *NPT* ensemble, the probability for accepting a new point 2 over an old point 1 becomes

$$p = \min\left\{1, \frac{V_2^N \exp[-(E_2 + PV_2)/k_B T]}{V_1^N \exp[-(E_1 + PV_1)/k_B T]}\right\} \tag{3.49}$$

(lower case 'p' is used here for probability to avoid confusion with upper case 'P' for pressure).

The choices of how often to scale the system volume, and by what range of factors, obviously influence acceptance ratios and are adjusted in much the same manner as geometric variables to maintain a good level of sampling efficiency. Other ensembles, or sampling schemes other than those using Cartesian coordinates, require analogous modifications to properly account for changes in phase space volume.

Just as with MD methods, MC simulations require an initial equilibration period so that property averages are not biased by very poor initial values. Typically various property values are monitored to assess whether they appear to have achieved a reasonable level of convergence prior to proceeding to production statistics. We now focus more closely on this issue.

3.6.4 Simulation Convergence

Convergence is defined as the acquisition of a sufficient number of phase points, through either MC or MD methods, to thoroughly sample phase space in a proper, Boltzmann-weighted fashion, i.e., the sampling is ergodic. While simple to define, convergence is *impossible* to prove, and this is either terribly worrisome or terribly liberating, depending on one's personal outlook.

To be more clear, we should separate the analysis of convergence into what might be termed 'statistical' and 'chemical' components. The former tends to be more tractable than the latter. Statistical convergence can be operatively defined as being *likely* to have been achieved when the average values for all properties of interest appear to remain roughly constant with increased sampling. In the literature, it is fairly standard to provide one or two plots of some particular properties as a function of time so that readers can agree that, to their eyes, the plots appear to have flattened out and settled on a particular value. For instance, in the simulation of macromolecules, the root-mean-square deviation (RMSD) of the simulation structure from an X-ray or NMR structure is often monitored. The RMSD for a particular snapshot is defined as

$$\mathrm{RMSD} = \sqrt{\frac{\sum_{i=1}^{N}(r_{i,\mathrm{sim}} - r_{i,\mathrm{expt}})^2}{N}} \tag{3.50}$$

where N is the number of atoms in the macromolecule, and the positions r are determined in a coordinate system having the center of mass at the origin and aligning the principle

moments of inertia along the cartesian axes (i.e., the simulated and experimental structures are best aligned prior to computing the RMSD). Monitoring the RMSD serves the dual purpose of providing a particular property whose convergence can be assessed and also of offering a quantitative measure of how 'close' the simulated structure is to the experimentally determined one. A typical RMSD plot is provided in Figure 3.7.

[Note that the information content in Figure 3.7 is often boiled down, when reported in the literature, to a single number, namely ⟨RMSD⟩. However, the magnitude of the fluctuation about the mean, which can be quantified by the standard deviation, is also an important quantity, and should be reported wherever possible. This is true for all expectation values derived from simulation. The standard deviation can be interpreted as a combination of the statistical noise (deriving from the limitations of the method) and the thermal noise (deriving from the 'correct' physical nature of the system). Considerably more refined methods of error analysis for average values from simulations have been promulgated (Smith and Wells 1984; Straatsma, Berendsen and Stam 1986; Kolafa 1986; Flyvberg and Petersen 1989).]

Another check of convergence in MD simulations, as alluded to above, is to ensure that the sampling length is longer than the autocorrelation decay time for a particular property by several multiples of that time. In practice, this analysis is performed with less regularity than is the simple monitoring of individual property values.

It must be borne in mind, however, that the typical simulation lengths that can be achieved with modern hardware and software are very, very rarely in excess of 1 μs. It is thus quite possible that the simulation, although it appears to be converged with respect to the analyses noted above, is trapped in a metastable state having a lifetime in excess of 1 μs, and as a

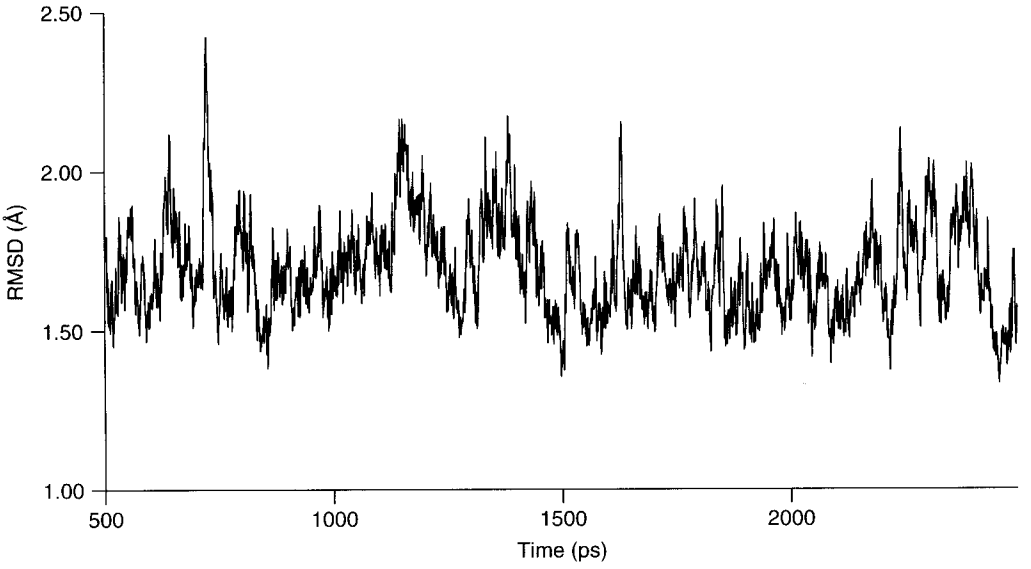


Figure 3.7 RMSD plot after 500 ps of equilibration for a solvated tRNA microhelix relative to its initial structure (Nagan *et al.* 1999)

result the statistics are not meaningful to the true system at equilibrium. The only way to address this problem is either to continue the simulation for a longer time or to run one or more additional simulations with different starting conditions or both. Entirely separate trajectories are more likely to provide data that are statistically uncorrelated with the original, but they are also more expensive since equilibration periods are required prior to collecting production mode data.

Problems associated with statistical convergence and/or metastability are vexing ones, but more daunting still can be the issue of chemical convergence. This is probably best illustrated with an example. Imagine that one would like to simulate the structure of a protein in water at pH 7 and that the protein contains nine histidine residues. At pH 7, the protein could, in principle, exist in many different protonation states (i.e., speciation) since the pK_a of histidine is quite near 7. Occam's razor and a certain amount of biochemical experience suggest that, in fact, only one or two states are likely to be populated under biological conditions, but how to choose which one(s) for simulation, since most force fields will not allow for protonation/deprotonation to take place? If the wrong state is chosen, it may be possible to acquire very good statistical convergence for the associated region of phase space, but that region is statistically unimportant compared to other regions which were *not* sampled.

3.6.5 The Multiple Minima Problem

A related problem, and one that is commonly encountered, has to do with molecules possessing multiple conformations. Consider *N*-methylacetamide, which can exist in *E* and *Z* forms. The latter stereoisomer is favored over the former by about 3 kcal/mol, but the barrier to interconversion is in excess of 18 kcal/mol. Thus, a simulation of *N*-methylacetamide starting with the statistically less relevant *E* structure is highly unlikely ever to sample the *Z* form, either using MD (since the high barrier implies an isomerization rate that will be considerably slower than the simulation time) or MC (since with small steps, the probability of going so far uphill would be very low, while with large steps it might be possible for the isomers to interconvert, but the rejection rate would be enormous making the simulation intractable). A related example with similar issues has to do with modeling phase transfer by MC methods, e.g., the movement of a solute between two immiscible liquids, or of a molecule from the gas phase to the liquid phase. In each case, the likelihood of moving a molecule in its entirety is low.

A number of computational techniques have been proposed to address these limitations. The simplest approach conceptually, which can be applied to systems where all possible conformations can be readily enumerated, is to carry out simulations for each one and then weight the respective property averages according to the free energies of the conformers (means for estimating these free energies are discussed in Chapter 12). This approach is, of course, cumbersome when the number of conformers grows large. This growth can occur with startling rapidity. For example, 8, 18, 41, 121, and 12 513 distinct minima have been identified for cyclononane, -decane, -undecane, -dodecane, and -heptadecane, respectively (Weinberg and Wolfe 1994). And cycloalkanes are relatively simple molecules compared, say, to a protein, where the holy grail of conformational analysis is prediction of a properly folded structure from only sequence information.

One approach to the identification of multiple minima is to periodically heat the system to a very high temperature. Since most force fields do not allow bond-breaking to occur, high temperature simply has the effect of making conformational interconversions more likely. After a certain amount of time, the system is cooled again to the temperature of interest, and statistics are collected. In practice, this technique is often used for isolated molecules in the gas phase in the hope of finding a global minimum energy structure, in which case it is referred to as 'simulated annealing'. In condensed phases, it is difficult to converge the statistical weights of the different accessed conformers. Within the context of MC simulations, other techniques to force the system to jump between minimum-energy wells in a properly energy-weighted fashion have been proposed (see, for instance, Guarnieri and Still 1994; Senderowitz and Still 1998).

An alternative to adjusting the temperature to help the system overcome high barriers is to artificially lower the barrier by adding a potential energy term that is counterbalancing. For instance, if the barrier is associated with a rotation, a so-called 'biasing potential' can be added such that the barrier is eliminated. The system can now sample freely between wells, but computed properties must be corrected for the proper free energy difference(s) in the absence of the biasing potential(s) (Straatsma and McCammon 1994). An interesting alternative suggested by Verkhivker, Elber, and Nowak (1992) is to have multiple conformers present *simultaneously* in a 'single' molecule. In the so-called 'locally enhanced sampling' method, the molecule of interest is represented as a sum of different conformers, each contributing fractionally to the total force field energy expression. Another conceptually similar method is to artificially lower barrier heights in certain regions of a phase space that has been artificially expanded by a single *extra* coordinate introduced for just that purpose – an idea analogous to the way catalysts lower barrier heights without affecting local minima (Stolovitzky and Berne 2000).

Just as with statistical convergence, there can be no guarantee that any of the techniques above will provide a thermodynamically accurate sampling of phase space, even though on the timescale of the simulation various property values may *appear* to be converged. As with most theoretical modeling, then, it is best to assess the likely utility of the predictions from a simulation by first comparing to experimentally well-known quantities. When these are accurately reproduced, other predictions can be used with greater confidence. As a corollary, the modeling of systems for which few experimental data are available against which to compare is perilous.

3.7 Case Study: Silica Sodalite

Synopsis of Nicholas *et al.* (1991) 'Molecular Modeling of Zeolite Structure. 2. Structure and Dynamics of Silica Sodalite and Silicate Force Field'.

Zeolites are mesoporous materials that are crystalline in nature. The simplest zeolites are made up of Al and/or Si and O atoms. Also known as molecular sieves, they find use as drying agents because they are very hygroscopic, but from an economic standpoint they are of greatest importance as size-selective catalysts in various reactions involving hydrocarbons and functionalized molecules of low molecular weight (for instance, they can be used to convert methanol to gasoline). The mechanisms by which zeolites operate

are difficult to identify positively because of the heterogeneous nature of the reactions in which they are involved (they are typically solids suspended in solution or reacting with gas-phase molecules), and the signal-to-noise problems associated with identifying reactive intermediates in a large background of stable reactants and products. As a first step toward possible modeling of reactions taking place inside the zeolite silica sodalite, Nicholas and co-workers reported the development of an appropriate force field for the system, and MD simulations aimed at its validation.

The basic structural unit of silica sodalite is presented in Figure 3.8. Because there are only two atomic types, the total number of functional forms and parameters required to define a force field is relatively small (18 parameters total). The authors restrict themselves to an overall functional form that sums stretching, bending, torsional, and non-bonded interactions, the latter having separate LJ and electrostatic terms. The details of the force field are described in a particularly lucid manner. The Si–O stretching potential is chosen to be quadratic, as is the O–Si–O bending potential. The flatter Si–O–Si bending potential is modeled with a fourth-order polynomial with parameters chosen to fit a bending potential computed from *ab initio* molecular orbital calculations (such calculations are the subject of Chapter 6). A Urey–Bradley Si–Si non-bonded harmonic stretching potential is added to couple the Si–O bond length to the Si–O–Si bond angle. Standard torsional potentials and LJ expressions are used, although, in the former case, a switching function is applied to allow the torsion energy to go to zero if one of the bond angles in the four-atom link becomes linear (which can happen at fairly low energy). With respect to electrostatic interactions, the authors note an extraordinarily large range of charges previously proposed for Si and O in this and related systems (spanning about 1.5 charge units). They choose a value for Si roughly midway through this range (which, by charge neutrality, determines the O charge as well), and examine the sensitivity of their model to the electrostatics by carrying out MD simulations with dielectric constants of 1, 2, and 5. The simulation cell is composed of 288 atoms (quite small, which makes the simulations computationally simple). PBCs and Ewald sums are used to account for the macroscopic nature of the real zeolite in simulations. Propagation of MD trajectories is accomplished using a leapfrog algorithm and

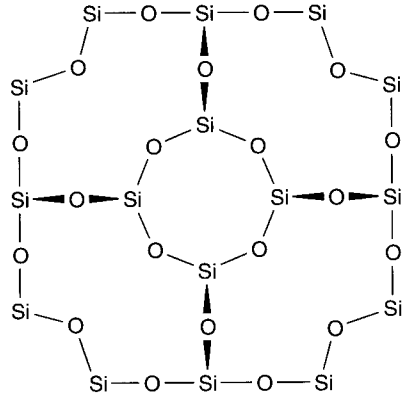


Figure 3.8 The repeating structural unit (with connections not shown) that makes up silica sodalite. What kinds of terms would be required in a force field designed to model such a system?

1.0 fs time steps following 20 ps or more of equilibration at 300 K. Each MD trajectory is 20 ps, which is very short by modern standards, but possibly justified by the limited dynamics available within the crystalline environment.

The quality of the parameter set is evaluated by comparing various details from the simulations to available experimental data. After testing a small range of equilibrium values for the Si–O bond, they settle on 1.61 Å, which gives optimized values for the unit cell Si–O bond length, and O–Si–O and Si–O–Si bond angles of 1.585 Å and 110.1° and 159.9°, respectively. These compare very favorably with experimental values of 1.587 Å and 110.3° and 159.7°, respectively. Furthermore, a Fourier transform of the total dipole correlation function (see Section 3.5) provides a model IR spectrum for comparison to experiment. Again, excellent agreement is obtained, with dominant computed bands appearing at 1106, 776, and 456 cm^{-1}, while experimental bands are observed at 1107, 787, 450 cm^{-1}. Simulations with different dielectric constants showed little difference from one another, suggesting that overall, perhaps because of the high symmetry of the system, sensitivity to partial atomic charge choice was low.

In addition, the authors explore the range of thermal motion of the oxygen atoms with respect to the silicon atoms they connect in the smallest ring of the zeolite cage (the eight-membered ring in the center of Figure 3.8). They determine that motion inward and outward and above and below the plane of the ring takes place with a fair degree of facility, while motion parallel to the Si–Si vector takes place over a much smaller range. This behavior is consistent with the thermal ellipsoids determined experimentally from crystal diffraction.

The authors finish by exploring the transferability of their force field parameters to a different zeolite, namely, silicalite. In this instance, a Fourier transform of the total dipole correlation function provides another model infrared (IR) spectrum for comparison to experiment, and again excellent agreement is obtained. Dominant computed bands appear at 1099, 806, 545, and 464 cm^{-1}, while experimental bands are observed at 1100, 800, 550, and 420 cm^{-1}. Some errors in band intensity are observed in the lower energy region of the spectrum.

As a first step in designing a general modeling strategy for zeolites, this paper is a very good example of how to develop, validate, and report force field parameters and results. The authors are pleasantly forthcoming about some of the assumptions employed in their analysis (for instance, all experimental data derive from crystals incorporating ethylene glycol as a solvent, while the simulations have the zeolite filled only with vacuum) and set an excellent standard for modeling papers of this type.

Bibliography and Suggested Additional Reading

Allen, M. P. and Tildesley, D. J. 1987. *Computer Simulation of Liquids*, Clarendon: Oxford.

Beveridge, D. L. and McConnell, K. J. 2000. 'Nucleic acids: theory and computer simulation, Y2K' *Curr. Opin. Struct. Biol.*, **10**, 182.

Brooks, C. L., III and Case, D. A. 1993. 'Simulations of Peptide Conformational Dynamics and Thermodynamics' *Chem. Rev.*, **93**, 2487.

Cheatham, T. E., III and Brooks, B. R. 1998. 'Recent Advances in Molecular Dynamics Simulation Towards the Realistic Representation of Biomolecules in Solution' *Theor. Chem. Acc.*, **99**, 279.

Frenkel, D. and Smit, B. 1996. *Understanding Molecular Simulation: From Algorithms to Applications*, Academic Press: San Diego.

Haile, J. 1992. *Molecular Dynamics Simulations*, Wiley: New York.
Jensen, F. 1999. *Introduction to Computational Chemistry*, Wiley: Chichester.
Jorgensen, W. L. 2000. 'Perspective on "Equation of State Calculations by Fast Computing Machines"' *Theor. Chem. Acc.*, **103**, 225.
Lybrand, T. P. 1990. 'Computer Simulation of Biomolecular Systems Using Molecular Dynamics and Free Energy Perturbation Methods' in *Reviews in Computational Chemistry*, Vol. 1, Lipkowitz, K. B. and Boyd, D. B., Eds., VCH: New York, 295.
McQuarrie, D. A. 1973. *Statistical Thermodynamics*, University Science Books: Mill Valley, CA.
Straatsma, T. P. 1996. 'Free Energy by Molecular Simulation' in *Reviews in Computational Chemistry*, Vol. 9, Lipkowitz, K. B. and Boyd, D. B., Eds., VCH: New York, 81.

References

Andersen, H. C. 1980. *J. Chem. Phys.*, **72**, 2384.
Banks, J. L., Kaminski, G. A., Zhou, R., Mainz, D. T., Berne, B. J., and Friesner, R. A. 1999. *J. Chem. Phys.*, **110**, 741.
Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A., and Haak, J. R. 1984. *J. Chem. Phys.*, **81**, 3684.
Berens, P. H., and Wilson, K. R. 1981. *J. Chem. Phys.*, **74**, 4872.
Bernardo, D. N., Ding, Y., Krogh-Jespersen, K., and Levy, R. M. 1994. *J. Phys. Chem.*, **98**, 4180.
Berne, B. J. and Harp, G. D. 1970. *Adv. Chem. Phys.*, **17**, **63**, 130.
Dang, L. X., Rice, J. E., Caldwell, J., and Kollman, P. A. 1991. *J. Am. Chem. Soc.*, **113**, 2481.
Demaison, J., Hütner, W., and Tiemann, E. 1982. In: *Molecular Constants, Landolt-Börstein, New Series, Group II*, Vol. 14a, Hellwege, K. -H. and Hellwege, A. M., Eds., Springer-Verlag: Berlin, 584.
Ernst, M. H., Hauge, E. H., and van Leeuwen, J. M. J. 1971. *Phys. Rev. A*, **4**, 2055.
Feenstra, K. A., Hess, B., and Berendsen, H. J. C. 1999. *J. Comput. Chem.*, **20**, 786.
Flyvberg, H. and Petersen, H. G. 1989. *J. Chem. Phys.*, **91**, 461.
Ford, J. 1973. *Adv. Chem. Phys.*, **24**, 155.
Gao, J., Habibollazadeh, D., and Shao, L. 1995. *J. Phys. Chem.*, **99**, 16460.
Gear, C. W. 1971. *Numerical Initial Value Problems in Ordinary Differential Equations*, Prentice-Hall: Englewood Cliffs, N.J.
Grubmuller, H. and Tavan, P. 1998. *J. Comput. Chem.*, **19**, 1534.
Guarnieri, F. and Still, W. C. 1994. *J. Comput. Chem.*, **15**, 1302.
Haile, J. 1992. *Molecular Dynamics Simulations*, Wiley: New York, 291.
Hoover, W. G. 1985. *Phys. Rev. A*, **31**, 1695.
Kolafa, J. 1986. *Mol. Phys.*, **59**, 1035.
Mahoney, W. and Jorgensen, W. L. 2000. *J. Chem. Phys.*, **112**, 8910.
Metropolis, N., Rosenbluth, A. E., Rosenbluth, M. N., Teller, A. H., and Teller, E. 1953. *J. Chem. Phys.*, **21**, 1087.
Nagan, M. C., Kerimo, S. S., Musier-Forsyth, K., and Cramer, C. J. 1999. *J. Am. Chem. Soc.*, **121**, 7310.
Nicholas, J. B., Hopfinger, A. J., Trouw, F. R., and Iton, L. E. 1991. *J. Am. Chem. Soc.*, **113**, 4792.
Nosé, S. 1984. *Mol. Phys.*, **52**, 255.
Olender, R. and Elber, R., 1996. *J. Chem. Phys.*, **105**, 9299.
Pangali, C., Rao, M., and Berne, B. J. 1978. *Chem. Phys. Lett.*, **55**, 413.
Press, W. H., Flannery, B. P., Teukolsky, S. A., and Vetterling, W. T. 1986. *Numerical Recipes*, Cambridge University Press: New York.
Rizzo, R. C. and Jorgensen, W. L. 1999. *J. Am. Chem. Soc.*, **121**, 4827.

Ryckaert, J. P., Ciccotti, G., and Berendsen, H. J. C. 1977. *J. Comput. Phys.*, **23**, 327.
Senderowitz, H. and Still, W. C. 1998. *J. Comput. Chem.*, **19**, 1736.
Smith, E. B. and Wells, B. H. 1984. *Mol. Phys.*, **53**, 701.
Soetens, J.-C., Millot, C., Chipot, C., Jansen, G., Ángyán, J. G., and Maigret, B. 1997. *J. Phys. Chem. B*, **101**, 10910.
Stolovitzky, G. and Berne, B. J. 2000. *Proc. Natl. Acad. Sci. (USA)*, **21**, 11164.
Straatsma, T. P. and McCammon, J. A. 1994. *J. Chem. Phys.*, **101**, 5032.
Straatsma, T. P., Berendsen, H. J. C., and Stam, A. J. 1986. *Mol. Phys.*, **57**, 89.
Theodorouo, D. N. and Suter, U. W. 1985. *J. Chem. Phys.*, **82**, 955.
Verkhivker, G., Elber, R., and Nowak, W. 1992. *J. Chem. Phys.*, **97**, 7838.
Verlet, L. 1967. *Phys. Rev.*, **159**, 98.
Weinberg, N. and Wolfe, S. 1994. *J. Am. Chem. Soc.*, **116**, 9860.
Winn, P. J., Ferenczy, G., and Reynolds, C. A. 1999. *J. Comput. Chem.*, **20**, 704.
Zhu, S.-B. and Wong, C. F. 1994. *J. Phys. Chem.*, **98**, 4695.
Zwanzig, R. 1965. *Ann. Rev. Phys. Chem.*, **16**, 67.

4

Foundations of Molecular Orbital Theory

4.1 Quantum Mechanics and the Wave Function

To this point, the models we have considered for representing microscopic systems have been designed based on classical, which is to say, macroscopic, analogs. We now turn our focus to contrasting models, whose foundations explicitly recognize the fundamental difference between systems of these two size extremes. Early practitioners of chemistry and physics had few, if any, suspicions that the rules governing microscopic and macroscopic systems should be different. Then, in 1900, Max Planck offered a radical proposal that blackbody radiation emitted by microscopic particles was limited to certain discrete values, i.e., it was 'quantized'. Such quantization was essential to reconciling large differences between predictions from classical models and experiment.

As the twentieth century progressed, it became increasingly clear that quantization was not only a characteristic of light, but also of the fundamental particles from which matter is constructed. Bound electrons in atoms, in particular, are clearly limited to discrete energies (levels) as indicated by their ultraviolet and visible line spectra. This phenomenon has no classical correspondence – in a classical system, obeying Newtonian mechanics, energy can vary continuously.

In order to describe microscopic systems, then, a different mechanics was required. One promising candidate was wave mechanics, since standing waves are also a quantized phenomenon. Interestingly, as first proposed by de Broglie, matter can indeed be shown to have wavelike properties. However, it also has particle-like properties, and to properly account for this dichotomy a new mechanics, quantum mechanics, was developed. This chapter provides an overview of the fundamental features of quantum mechanics, and describes in a formal way the fundamental equations that are used in the construction of computational models. In some sense, this chapter is historical. However, in order to appreciate the differences between modern computational models, and the range over which they may be expected to be applicable, it is important to understand the foundation on which all of them are built. Following this exposition, Chapter 5 overviews the approximations inherent

in so-called semiempirical QM models, Chapter 6 focuses on *ab initio* Hartree–Fock (HF) models, and Chapter 7 describes methods for accounting for electron correlation.

We begin with a brief recapitulation of some of the key features of quantum mechanics. The fundamental postulate of quantum mechanics is that a so-called wave function, Ψ, exists for any (chemical) system, and that appropriate operators (functions) which act upon Ψ return the observable properties of the system. In mathematical notation,

$$\vartheta\Psi = e\Psi \tag{4.1}$$

where ϑ is an operator and e is a scalar value for some property of the system. When Eq. (4.1) holds, Ψ is called an eigenfunction and e an eigenvalue, by analogy to matrix algebra were Ψ to be an N-element column vector, ϑ to be an $N \times N$ square matrix, and e to remain a scalar constant. Importantly, the product of the wave function Ψ with its complex conjugate (i.e., $|\Psi^*\Psi|$) has units of probability density. For ease of notation, and since we will be working almost exclusively with real, and not complex, wave functions, we will hereafter drop the complex conjugate symbol '*'. Thus, the probability that a chemical system will be found within some region of multi-dimensional space is equal to the integral of $|\Psi|^2$ over that region of space.

These postulates place certain constraints on what constitutes an acceptable wave function. For a bound particle, the normalized integral of $|\Psi|^2$ over all space must be unity (i.e., the probability of finding it somewhere is one) which requires that Ψ be quadratically integrable. In addition, Ψ must be continuous and single-valued.

From this very formal presentation, the nature of Ψ can hardly be called anything but mysterious. Indeed, perhaps the best description of Ψ at this point is that it is an oracle – when queried with questions by an operator, it returns answers. By the end of this chapter, it will be clear the precise way in which Ψ is expressed, and we should have a more intuitive notion of what Ψ represents. However, the view that Ψ is an oracle is by no means a bad one, and will be returned to again at various points.

4.2 The Hamiltonian Operator

4.2.1 General Features

The operator in Eq. (4.1) that returns the system energy, E, as an eigenvalue is called the Hamiltonian operator, H. Thus, we write

$$H\Psi = E\Psi \tag{4.2}$$

which is the Schrödinger equation. The typical form of the Hamiltonian operator with which we will be concerned takes into account five contributions to the total energy of a system (from now on we will say molecule, which certainly includes an atom as a possibility): the kinetic energies of the electrons and nuclei, the attraction of the electrons to the nuclei, and the interelectronic and internuclear repulsions. In more complicated situations, e.g., in

the presence of an external electric field, in the presence of an external magnetic field, in the event of significant spin–orbit coupling in heavy elements, taking account of relativistic effects, etc., other terms are required in the Hamiltonian. We will consider some of these at later points in the text, but we will not find them necessary for general purposes. Casting the Hamiltonian into mathematical notation, we have

$$H = -\sum_i \frac{\hbar^2}{2m_e}\nabla_i^2 - \sum_k \frac{\hbar^2}{2m_k}\nabla_k^2 - \sum_i \sum_k \frac{e^2 Z_k}{r_{ik}} + \sum_{i<j} \frac{e^2}{r_{ij}} + \sum_{k<l} \frac{e^2 Z_k Z_l}{r_{kl}} \tag{4.3}$$

where i and j run over electrons, k and l run over nuclei, $\hbar$ is Planck's constant divided by 2π, m_e is the mass of the electron, m_k is the mass of nucleus k, ∇^2 is the Laplacian operator, e is the charge on the electron, Z is an atomic number, and r_{ab} is the distance between particles a and b. Note that Ψ is thus a function of $3n$ coordinates where n is the total number of particles (nuclei and electrons), e.g., the x, y, and z Cartesian coordinates specific to each particle. If we work in Cartesian coordinates, the Laplacian has the form

$$\nabla_i^2 = \frac{\partial^2}{\partial x_i^2} + \frac{\partial^2}{\partial y_i^2} + \frac{\partial^2}{\partial z_i^2} \tag{4.4}$$

Note that the Hamiltonian operator in Eq. (4.3) is composed of kinetic energy and potential energy parts. The potential energy terms (the last three) appear exactly as they do in classical mechanics. The kinetic energy for a QM particle, however, is not expressed as $|\mathbf{p}|^2/2m$, but rather as the eigenvalue of the kinetic energy operator

$$T = -\frac{\hbar^2}{2m}\nabla^2 \tag{4.5}$$

Note also that, as described in Chapter 1, most of the constants appearing in Eq. (4.3) are equal to 1 when atomic units are chosen.

In general, Eq. (4.2) has *many* acceptable eigenfunctions Ψ for a given molecule, each characterized by a different associated eigenvalue E. That is, there is a complete set (perhaps infinite) of Ψ_i with eigenvalues E_i. For ease of future manipulation, we may assume without loss of generality that these wave functions are orthonormal, i.e., for a one particle system where the wave function depends on only three coordinates,

$$\int\int\int \Psi_i \Psi_j dx dy dz = \delta_{ij} \tag{4.6}$$

where δ_{ij} is the Krönecker delta (equal to one if $i = j$ and equal to zero otherwise). Orthonormal actually implies two qualities simultaneously: 'orthogonal' means that the integral in Eq. (4.6) is equal to zero if $i \neq j$ and 'normal' means that when $i = j$ the value of the integral is one. For ease of notation, we will henceforth replace all multiple integrals over Cartesian space with a single integral over a generalized $3n$-dimensional volume element $d\mathbf{r}$, rendering Eq. (4.6) as

$$\int \Psi_i \Psi_j d\mathbf{r} = \delta_{ij} \tag{4.7}$$

Now, consider the result of taking Eq. (4.2) for a specific Ψ_i, multiplying on the left by Ψ_j, and integrating. This process gives

$$\int \Psi_j H \Psi_i d\mathbf{r} = \int \Psi_j E_i \Psi_i d\mathbf{r} \tag{4.8}$$

Since the energy E is a scalar value, we may remove it outside the integral on the r.h.s. and use Eq. (4.7) to write

$$\int \Psi_j H \Psi_i d\mathbf{r} = E_i \delta_{ij} \tag{4.9}$$

This equation will prove useful later on, but it is worth noting at this point that it also offers a prescription for determining the molecular energy. With a wave function in hand, one simply constructs and solves the integral on the left (where i and j are identical and index the wave function of interest). Of course, we have not yet said much about the form of the wave function, so the nature of the integral in Eq. (4.8) is not obvious ... although one suspects it might be unpleasant to solve.

4.2.2 The Variational Principle

The power of quantum theory, as expressed in Eq. (4.1), is that if one has a molecular wave function in hand, one can calculate physical observables by application of the appropriate operator in a manner analogous to that shown for the Hamiltonian in Eq. (4.8). Regrettably, none of these equations offers us a prescription for *obtaining* the orthonormal set of molecular wave functions. Let us assume for the moment, however, that we can pick an arbitrary function, Φ, which is indeed an eigenfunction for the specific case of Eq. (4.2). Since we defined the set of orthonormal wave functions Ψ_i to be complete (and perhaps infinite), the function Φ must be some linear combination of the Ψ_i, i.e.,

$$\Phi = \sum_i c_i \Psi_i \tag{4.10}$$

where, of course, since we don't yet know the individual Ψ_i, we certainly don't know the coefficients c_i either! Note that the normality of Φ imposes a constraint on the coefficients, however, deriving from

$$\begin{aligned} \int \Phi^2 d\mathbf{r} = 1 &= \int \sum_i c_i \Psi_i \sum_j c_j \Psi_j d\mathbf{r} \\ &= \sum_{ij} c_i c_j \int \Psi_i \Psi_j d\mathbf{r} \\ &= \sum_{ij} c_i c_j \delta_{ij} \\ &= \sum_i c_i^2 \end{aligned} \tag{4.11}$$

Now, let us consider evaluating the energy associated with wave function Φ. Taking the approach of multiplying on the left and integrating as outlined above, we have

$$\begin{aligned}\int \Phi H \Phi d\mathbf{r} &= \int \left(\sum_i c_i \Psi_i\right) H \left(\sum_j c_j \Psi_j\right) d\mathbf{r} \\ &= \sum_{ij} c_i c_j \int \Psi_i H \Psi_j d\mathbf{r} \\ &= \sum_{ij} c_i c_j E_j \delta_{ij} \\ &= \sum_i c_i^2 E_i \end{aligned} \tag{4.12}$$

where we have used Eq. (4.9) to simplify the r.h.s. Thus, the energy associated with the generic wave function Φ is determinable from all of the coefficients c_i (that define how the orthonormal set of Ψ_i combine to form Φ) and their associated energies E_i. Regrettably, we still don't know the values for *any* of these quantities. However, let us take note of the following. In the set of all E_i there must be a lowest energy value (i.e., the set is bounded from below); let us call that energy, corresponding to the 'ground state', E_0. [Notice that this boundedness is a critical feature of quantum mechanics! In a classical system, one could imagine always finding a state lower in energy than another state by simply 'shrinking the orbits' of the electrons to increase nuclear–electronic attraction while keeping the kinetic energy constant.]

We may now combine the results from Eqs. (4.11) and (4.12) to write

$$\int \Phi H \Phi d\mathbf{r} - E_0 \int \Phi^2 d\mathbf{r} = \sum_i c_i^2 (E_i - E_0) \tag{4.13}$$

Assuming the coefficients to be real numbers, each term c_i^2 must be greater than or equal to zero. By definition of E_0, the quantity $(E_i - E_0)$ must also be greater than or equal to zero. Thus, we have

$$\int \Phi H \Phi d\mathbf{r} - E_0 \int \Phi^2 d\mathbf{r} \geq 0 \tag{4.14}$$

which we may rearrange to

$$\frac{\int \Phi H \Phi d\mathbf{r}}{\int \Phi^2 d\mathbf{r}} \geq E_0 \tag{4.15}$$

(note that when Φ is normalized, the denominator on the l.h.s. is 1, but it is helpful to have Eq. (4.15) in this more general form for future use).

Equation (4.15) has extremely powerful implications. If we are looking for the best wave function to define the ground state of a system, we can judge the quality of wave functions that we arbitrarily guess by their associated energies: *the lower the better.* This result is critical because it shows us that we do not have to construct our guess wave function Φ as a linear combination of (unknown) orthonormal wave functions Ψ_i, but we may construct it in any manner we wish. The quality of our guess will be determined by how low a value we calculate for the integral in Eq. (4.15). Moreover, since we would like to find the lowest possible energy within the constraints of how we go about constructing a wave function, we can use all of the tools that calculus makes available for locating extreme values.

4.2.3 The Born–Oppenheimer Approximation

Up to now, we have been discussing many-particle molecular systems entirely in the abstract. In fact, accurate wave functions for such systems are extremely difficult to express because of the correlated motions of particles. That is, the Hamiltonian in Eq. (4.3) contains pairwise attraction and repulsion terms, implying that no particle is moving independently of all of the others (the term 'correlation' is used to describe this interdependency). In order to simplify the problem somewhat, we may invoke the so-called Born–Oppenheimer approximation. This approximation is described with more rigor in Section 15.5, but at this point we present the conceptual aspects without delving deeply into the mathematical details.

Under typical physical conditions, the nuclei of molecular systems are moving much, much more slowly than the electrons (recall that protons and neutrons are about 1800 times more massive than electrons and note the appearance of mass in the denominator of the kinetic energy terms of the Hamiltonian in Eq. (4.3)). For practical purposes, electronic 'relaxation' with respect to nuclear motion is instantaneous. As such, it is convenient to decouple these two motions, and compute electronic energies for *fixed* nuclear positions. That is, the nuclear kinetic energy term is taken to be independent of the electrons, correlation in the attractive electron–nuclear potential energy term is eliminated, and the repulsive nuclear–nuclear potential energy term becomes a simply evaluated constant for a given geometry. Thus, the *electronic* Schrödinger equation is taken to be

$$(H_{\text{el}} + V_N)\Psi_{\text{el}}(\mathbf{q}_i; \mathbf{q}_k) = E_{\text{el}}\Psi_{\text{el}}(\mathbf{q}_i; \mathbf{q}_k) \tag{4.16}$$

where the subscript 'el' emphasizes the invocation of the Born–Oppenheimer approximation, H_{el} includes only the first, third, and fourth terms on the r.h.s. of Eq. (4.3), V_N is the nuclear–nuclear repulsion energy, and the electronic coordinates $\mathbf{q}_i$ are independent variables but the nuclear coordinates $\mathbf{q}_k$ are parameters (and thus appear following a semicolon rather than a comma in the variable list for Ψ). The eigenvalue of the electronic Schrödinger equation is called the 'electronic energy'. Note that the term V_N is a constant for a given set of fixed nuclear coordinates. Wave functions are invariant to the appearance of constant terms in the Hamiltonian, so in practice one almost always solves Eq. (4.16) without the inclusion of V_N, in which case the eigenvalue is sometimes called the 'pure electronic energy', and one then adds V_N to this eigenvalue to obtain E_{el}.

In general, the Born–Oppenheimer assumption is an extremely mild one, and it is entirely justified in most cases. It is worth emphasizing that this approximation has very profound consequences from a conceptual standpoint – so profound that they are rarely thought about but simply accepted as dogma. Without the Born–Oppenheimer approximation we would lack the concept of a potential energy surface: The PES is the surface defined by E_{el} over all possible nuclear coordinates. We would further lack the concepts of equilibrium and transition state geometries, since these are defined as critical points on the PES; instead we would be reduced to discussing high-probability regions of the nuclear wave functions. Of course, for some problems in chemistry, we *do* need to consider the quantum mechanical character of the nuclei, but the advantages afforded by the Born–Oppenheimer approximation should be manifest.

4.3 Construction of Trial Wave Functions

Equation (4.16) is simpler than Eq. (4.2) because electron–nuclear correlation has been removed. The remaining correlation, that between the individual electrons, is considerably more troubling. For the moment we will take the simplest possible approach and ignore it; we do this by considering systems with only a single electron. The electronic wave function has thus been reduced to depending only on the fixed nuclear coordinates and the three cartesian coordinates of the single electron. The eigenfunctions of Eq. (4.16) for a molecular system may now be properly called molecular orbitals (MOs; rather unusual ones in general, since they are for a molecule having only one electron, but MOs nonetheless). To distinguish a one-electron wave function from a many-electron wave function, we will designate the former as ψ_{el} and the latter as Ψ_{el}. We will hereafter drop the subscript 'el' where not required for clarity; unless otherwise specified, all wave functions are electronic wave functions.

The pure electronic energy eigenvalue associated with each molecular orbital is the energy of the electron in that orbital. Experimentally, one might determine this energy by measuring the ionization potential of the electron when it occupies the orbital (fairly easy for the hydrogen atom, considerably more difficult for polynuclear molecules). To measure E_{el}, which includes the nuclear repulsion energy, one would need to determine the 'atomization' energy, that is, the energy required to ionize the electron *and* to remove all of the nuclei to infinite separation. In practice, atomization energies are not measured, but instead we have compilations of such thermodynamic variables as heats of formation. The relationship between these computed and thermodynamic quantities is discussed in Chapter 10.

4.3.1 The LCAO Basis Set Approach

As noted earlier, we may imagine constructing wave functions in any fashion we deem reasonable, and we may judge the quality of our wave functions (in comparison to one another) by evaluation of the energy eigenvalues associated with each. The one with the lowest energy will be the most accurate and presumably the best one to use for computing other properties by the application of other operators. So, how might one go about choosing

mathematical functions with which to construct a trial wave function? This is a typical question in mathematics – how can an arbitrary function be represented by a combination of more convenient functions? The convenient functions are called a 'basis set'. Indeed, we have already encountered this formalism – Eq. (2.10) of Chapter 2 illustrates the use of a basis set of cosine functions to approximate torsional energy functions.

In our QM systems, we have temporarily restricted ourselves to systems of one electron. If, in addition, our system were to have only one nucleus, then we would not need to guess wave functions, but instead we could solve Eq. (4.16) *exactly*. The eigenfunctions that are determined in that instance are the familiar hydrogenic atomic orbitals, 1s, 2s, 2p, 3s, 3p, 3d, etc., whose properties and derivation are discussed in detail in standard texts on quantum mechanics. For the moment, we will not investigate the mathematical representation of these hydrogenic atomic orbitals in any detail, but we will simply posit that, as functions, they may be useful in the construction of more complicated *molecular* orbitals. In particular, just as in Eq. (4.10) we constructed a guess wave function as a linear combination of exact wave functions, so here we will construct a guess wave function ϕ as a linear combination of atomic wave functions φ, i.e.,

$$\phi = \sum_{i=1}^{N} a_i \varphi_i \tag{4.17}$$

where the set of N functions φ_i is called the 'basis set' and each has associated with it some coefficient a_i. This construction is known as the linear combination of atomic orbitals (LCAO) approach.

Note that Eq. (4.17) does not specify the locations of the basis functions. Our intuition suggests that they should be centered on the atoms of the molecule, but this is certainly not a requirement. If this comment seems odd, it is worth emphasizing at this point that we should not let our chemical intuition limit our mathematical flexibility. As chemists, we choose to use atomic orbitals (AOs) because we anticipate that they will be efficient functions for the representation of MOs. However, as mathematicians, we should immediately stop thinking about our choices as orbitals, and instead consider them only to be *functions*, so that we avoid being conceptually influenced about how and where to use them.

Recall that the wave function squared has units of probability density. In essence, the electronic wave function is a road map of where the electrons are more or less likely to be found. Thus, we want our basis functions to provide us with the flexibility to allow electrons to 'go' where their presence at higher density lowers the energy. For instance, to describe the bonding of a hydrogen atom to a carbon, it is clearly desirable to use a p function on hydrogen, oriented along the axis of the bond, to permit electron density to be localized in the bonding region more efficiently than is possible with only a spherically symmetric s function. Does this imply that the hydrogen atom is somehow sp-hybridized? Not necessarily – the p function is simply serving the purpose of increasing the flexibility with which the *molecular* orbital may be described. If we took away the hydrogen p function and instead placed an s function *in between* the C and H atoms, we could also build up electron density in the bonding region (see Figure 4.1). Thus, the *chemical* interpretation of the coefficients in Eq. (4.17) should only be undertaken with caution, as further described in Chapter 9.

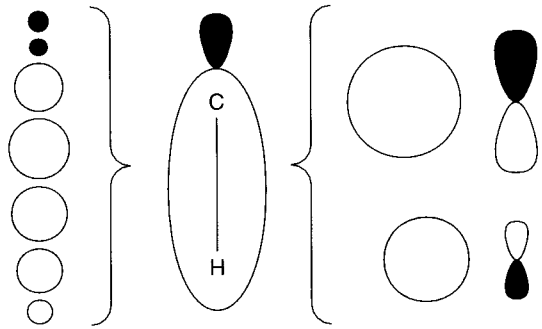


Figure 4.1 Two different basis sets for representing a C–H σ bonding orbital with the size of the basis functions roughly illustrating their weight in the hybrid MO. The set on the right is the more chemically intuitive since all basis functions are centered on the atoms. Note, however, that the use of a p function to polarize the hydrogen density goes beyond a purely minimalist approach. The set on the left is composed entirely of s functions distributed along the bond. Such a basis set may seem odd in concept, but is quite capable of accurately representing the electron density in space. Indeed, the basis set on the left would have certain computational advantages, chief among them the greater simplicity of working with s functions than with p functions

One should also note that the summation in Eq. (4.17) has an upper limit N; we cannot work with an infinite basis in any convenient way (at least not when the basis is AOs). However, the more atomic orbitals we allow into our basis, the closer our basis will come to 'spanning' the true molecular orbital space. Thus, the chemical idea that we would limit ourselves to, say, at most one 1s function on each hydrogen atom is needlessly confining from a mathematical standpoint. Indeed, there may be very many 'true' one-electron MOs that are very high in energy. Accurately describing these MOs may require some unusual basis functions, e.g., very diffuse functions to describe weakly bound electrons, like those found in Rydberg states. We will discuss these issues in much more detail in Section 6.2, but it is worth emphasizing here, at the beginning, that the distinction between orbitals and functions is a critical one in computational molecular orbital theory.

4.3.2 The Secular Equation

All that being said, let us now turn to evaluating the energy of our guess wave function. From Eqs. (4.15) and (4.17) we have

$$E = \frac{\int \left(\sum_i a_i \varphi_i\right) H \left(\sum_j a_j \varphi_j\right) d\mathbf{r}}{\int \left(\sum_i a_i \varphi_i\right) \left(\sum_j a_j \varphi_j\right) d\mathbf{r}}$$

$$= \frac{\sum_{ij} a_i a_j \int \varphi_i H \varphi_j d\mathbf{r}}{\sum_{ij} a_i a_j \int \varphi_i \varphi_j d\mathbf{r}}$$

$$= \frac{\sum_{ij} a_i a_j H_{ij}}{\sum_{ij} a_i a_j S_{ij}} \tag{4.18}$$

where we have introduced the shorthand notation H_{ij} and S_{ij} for the integrals in the numerator and denominator, respectively. These so-called 'matrix elements' are no longer as simple as they were in prior discussion, since the atomic orbital basis set, while likely to be efficient, is no longer orthonormal. These matrix elements have more common names, H_{ij} being called a 'resonance integral', and S_{ij} being called an 'overlap integral'. The latter has a very clear physical meaning, namely the extent to which any two basis functions overlap in a phase-matched fashion in space. The former integral is not so easily made intuitive, but it is worth pointing out that orbitals which give rise to large overlap integrals will similarly give rise to large resonance integrals. One resonance integral which *is* intuitive is H_{ii}, which corresponds to the energy of a single electron occupying basis function i, i.e., it is essentially equivalent to the ionization potential of the AO in the environment of the surrounding molecule.

Now, it is useful to keep in mind our objective. The variational principle instructs us that as we get closer and closer to the 'true' one-electron ground-state wave function, we will obtain lower and lower energies from our guess. Thus, once we have selected a basis set, we would like to choose the coefficients a_i so as to *minimize* the energy for all possible linear combinations of our basis functions. From calculus, we know that a necessary condition for a function (i.e., the energy) to be at its minimum is that its derivatives with respect to all of its free variables (i.e., the coefficients a_i) are zero. Notationally, that is

$$\frac{\partial E}{\partial a_k} = 0 \qquad \forall k \tag{4.19}$$

(where we make use of the mathematical abbreviation $\forall$ meaning 'for all'). Performing this fairly tedious partial differentiation on Eq. (4.18) for each of the N variables a_k gives rise to N equations which must be satisfied in order for Eq. (4.19) to hold true, namely

$$\sum_{i=1}^{N} a_i (H_{ki} - E S_{ki}) = 0 \qquad \forall k \tag{4.20}$$

This set of N equations (running over k) involves N unknowns (the individual a_i). From linear algebra, we know that a set of N equations in N unknowns has a non-trivial solution if and only if the determinant formed from the coefficients of the unknowns (in this case the 'coefficients' are the various quantities $H_{ki} - E S_{ki}$) is equal to zero. Notationally again,

that is

$$\begin{vmatrix} H_{11} - ES_{11} & H_{12} - ES_{12} & \cdots & H_{1N} - ES_{1N} \\ H_{21} - ES_{21} & H_{22} - ES_{22} & \cdots & H_{2N} - ES_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ H_{N1} - ES_{N1} & H_{N2} - ES_{N2} & \cdots & H_{NN} - ES_{NN} \end{vmatrix} = 0 \tag{4.21}$$

Equation (4.21) is called a secular equation. In general, there will be N roots E which permit the secular equation to be true. That is, there will be N energies E_j (some of which may be equal to one another, in which case we say the roots are 'degenerate') where each value of E_j will give rise to a different set of coefficients, a_{ij}, which can be found by solving the set of linear Eqs. (4.20) using E_j, and these coefficients will define an optimal wave function ϕ_j within the given basis set, i.e.,

$$\phi_j = \sum_{i=1}^{N} a_{ij}\varphi_i \tag{4.22}$$

In a one-electron system, the lowest energy molecular orbital would thus define the 'ground state' of the system, and the higher energy orbitals would be 'excited states'. Obviously, as these are different MOs, they have different basis function coefficients. Although we have not formally proven it, it is worth noting that the variational principle holds for the excited states as well: the calculated energy of a guess wave function for an excited state will be bounded from below by the true excited state energy (MacDonald 1933).

So, in a nutshell, to find the optimal one-electron wave functions for a molecular system, we:

1. Select a set of N basis functions.
2. For that set of basis functions, determine all N^2 values of both H_{ij} and S_{ij}.
3. Form the secular determinant, and determine the N roots E_j of the secular equation.
4. For each of the N values of E_j, solve the set of linear Eqs. (4.20) in order to determine the basis set coefficients a_{ij} for that MO.

All of the MOs determined by this process are mutually orthogonal. For degenerate MOs, some minor complications arise, but those are not discussed here.

4.4 Hückel Theory

4.4.1 Fundamental Principles

To further illuminate the LCAO variational process, we will carry out the steps outlined above for a specific example. To keep things simple (and conceptual), we consider a flavor of molecular orbital theory developed in the 1930s by Erich Hückel to explain some of the unique properties of unsaturated and aromatic hydrocarbons (Hückel 1931; for historical

insights, see also, Berson 1996; Frenking 2000). In order to accomplish steps 1–4 of the last section, Hückel theory adopts the following conventions:

(a) The basis set is formed entirely from parallel carbon 2p orbitals, one per atom. [Hückel theory was originally designed to treat only planar hydrocarbon π systems, and thus the 2p orbitals used are those that are associated with the π system.]

(b) The overlap matrix is defined by

$$S_{ij} = \delta_{ij} \tag{4.23}$$

Thus, the overlap of any carbon 2p orbital with itself is unity (i.e., the p functions are normalized), and that between any two p orbitals is zero.

(c) Matrix elements H_{ii} are set equal to the negative of the ionization potential of the methyl radical, i.e., the orbital energy of the singly occupied 2p orbital in the prototypical system defining sp^2 carbon hybridization. This choice is consistent with our earlier discussion of the relationship between this matrix element and an ionization potential. This energy value, which is defined so as to be negative, is rarely actually written as a numerical value, but is instead represented by the symbol α.

(d) Matrix elements H_{ij} between neighbors are also derived from experimental information. A 90° rotation about the π bond in ethylene removes all of the bonding interaction between the two carbon 2p orbitals. That is, the (positive) cost of the following process,

is $\Delta E = 2E_p - E_\pi$. The (negative) stabilization energy for the pi bond is distributed equally to the two p orbitals involved (i.e., divided in half) and this quantity, termed β, is used for H_{ij} between neighbors. (Note, based on our definitions so far, then, that $E_p = \alpha$ and $E_\pi = 2\alpha + 2\beta$.)

(e) Matrix elements H_{ij} between carbon 2p orbitals more distant than nearest neighbors are set equal to zero.

4.4.2 Application to the Allyl System

Let us now apply Hückel MO theory to the particular case of the allyl system, C_3H_3, as illustrated in Figure 4.2. Because we have three carbon atoms, our basis set is determined from convention (a) and will consist of 3 carbon 2p orbitals, one centered on each atom. We will arbitrarily number them 1, 2, 3, from left to right for bookkeeping purposes.

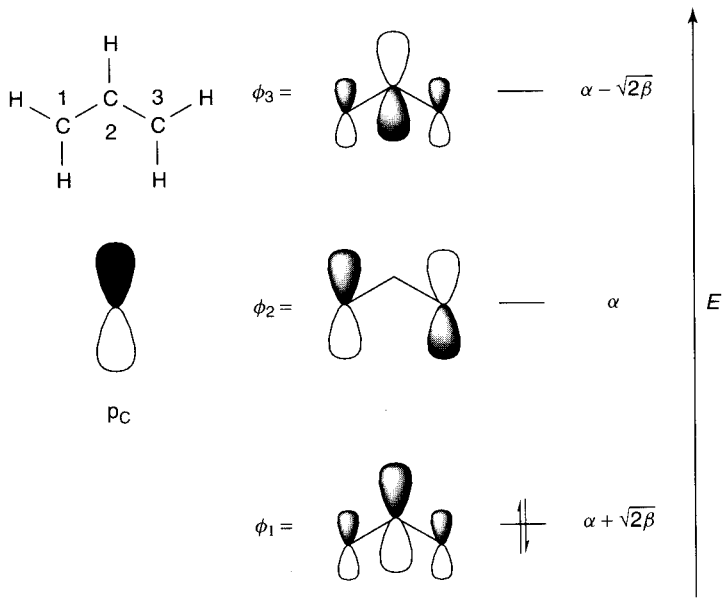


Figure 4.2 Hückel MOs for the allyl system. One p_C orbital per atom defines the basis set. Combinations of these 3 AOs create the 3 MOs shown. The electron occupation illustrated corresponds to the allyl cation. One additional electron in ϕ_2 would correspond to the allyl radical, and a second (spin-paired) electron in ϕ_2 would correspond to the allyl anion

The basis set size of three implies that we will need to solve a 3×3 secular equation. Hückel conventions (b)–(e) tell us the value of each element in the secular equation, so that Eq. (4.21) is rendered as

$$\begin{vmatrix} \alpha - E & \beta & 0 \\ \beta & \alpha - E & \beta \\ 0 & \beta & \alpha - E \end{vmatrix} = 0 \tag{4.24}$$

The use of the Krönecker delta to define the overlap matrix ensures that E appears only in the diagonal elements of the determinant. Since this is a 3×3 determinant, it may be expanded using Cramer's rule as

$$(\alpha - E)^3 + (\beta^2 \cdot 0) + (0 \cdot \beta^2) - [0 \cdot (\alpha - E) \cdot 0] - \beta^2(\alpha - E) - (\alpha - E)\beta^2 = 0 \tag{4.25}$$

which is a fairly simple cubic equation in E that has three solutions, namely

$$E = \alpha + \sqrt{2}\beta, \quad \alpha, \quad \alpha - \sqrt{2}\beta \tag{4.26}$$

Since α and β are negative by definition, the lowest energy solution is $\alpha + \sqrt{2}\beta$. To find the MO associated with this energy, we employ it in the set of linear Eqs. (4.20), together

with the various necessary H and S values to give

$$\begin{aligned}
a_1[\alpha - (\alpha + \sqrt{2}\beta) \cdot 1] + a_2[\beta - (\alpha + \sqrt{2}\beta) \cdot 0] + a_3[0 - (\alpha + \sqrt{2}\beta) \cdot 0] &= 0 \\
a_1[\beta - (\alpha + \sqrt{2}\beta) \cdot 0] + a_2[\alpha - (\alpha + \sqrt{2}\beta) \cdot 1] + a_3[\beta - (\alpha + \sqrt{2}\beta) \cdot 0] &= 0 \\
a_1[0 - (\alpha + \sqrt{2}\beta) \cdot 0] + a_2[\beta - (\alpha + \sqrt{2}\beta) \cdot 0] + a_3[\alpha - (\alpha + \sqrt{2}\beta) \cdot 1] &= 0
\end{aligned} \tag{4.27}$$

Some fairly trivial algebra reduces these equations to

$$\begin{aligned}
a_2 &= \sqrt{2}a_1 \\
a_3 &= a_1
\end{aligned} \tag{4.28}$$

While there are infinitely many values of a_1, a_2, and a_3 which satisfy Eq. (4.28), the requirement that the wave function be normalized provides a final constraint in the form of Eq. (4.11). The unique values satisfying both equations are

$$a_{11} = \frac{1}{2}, \quad a_{21} = \frac{\sqrt{2}}{2}, \quad a_{31} = \frac{1}{2} \tag{4.29}$$

where we have now emphasized that these coefficients are specific to the *lowest energy* molecular orbital by adding the second subscript '1'. Since we now know both the coefficients and the basis functions, we may construct the lowest energy molecular orbital, i.e.,

$$\varphi_1 = \frac{1}{2}p_1 + \frac{\sqrt{2}}{2}p_2 + \frac{1}{2}p_3 \tag{4.30}$$

which is illustrated in Figure 4.2.

By choosing the higher energy roots of Eq. (4.24), we may solve the sets of linear equations analogous to Eq. (4.27) in order to arrive at the coefficients required to construct ϕ_2 (from $E = \alpha$) and ϕ_3 (from $E = \alpha - \sqrt{2}\beta$). Although the algebra is left for the reader, the results are

$$\begin{aligned}
a_{12} &= \frac{\sqrt{2}}{2}, \quad a_{22} = 0, \quad a_{32} = -\frac{\sqrt{2}}{2} \\
a_{13} &= \frac{1}{2}, \quad a_{23} = -\frac{\sqrt{2}}{2}, \quad a_{33} = \frac{1}{2}
\end{aligned} \tag{4.31}$$

and these orbitals are also illustrated in Figure 4.2. The three orbitals we have derived are the bonding, non-bonding, and antibonding allyl molecular orbitals with which all organic chemists are familiar.

Importantly, Hückel theory affords us certain insights into the allyl system, one in particular being an analysis of the so-called 'resonance' energy arising from electronic delocalization in the π system. By delocalization we refer to the participation of more than two atoms in a

given MO. Consider for example the allyl cation, which has a total of two electrons in the π system. If we adopt a molecular aufbau principle of filling lowest energy MOs first and further make the assumption that each electron has the energy of the one-electron MO that it occupies (ϕ_1 in this case) then the total energy of the allyl cation π system is $2(\alpha + \sqrt{2}\beta)$. Consider the alternative 'fully localized' structure for the allyl system, in which there is a full (doubly-occupied) π bond between two of the carbons, and an empty, non-interacting p orbital on the remaining carbon atom. (This could be achieved by rotating the cationic methylene group 90° so that the p orbital becomes orthogonal to the remaining π bond, but that could no longer be described by simple Hückel theory since the system would be non-planar – the non-interaction we are considering here is purely a thought-experiment). The π energy of such a system would simply be that of a double bond, which by our definition of terms above is $2(\alpha + \beta)$. Thus, the Hückel resonance energy, which is equal to $H_\pi - H_{\text{localized}}$, is 0.83β (remember β is negative by definition, so resonance is a favorable phenomenon). Recalling the definition of β, the resonance energy in the allyl cation is predicted to be about 40% of the rotation barrier in ethylene.

We may perform the same analysis for the allyl radical and the allyl anion, respectively, by adding the energy of ϕ_2 to the cation with each successive addition of an electron, i.e., $H_\pi(\text{allyl radical}) = 2(\alpha + \sqrt{2}\beta) + \alpha$ and $H_\pi(\text{allyl anion}) = 2(\alpha + \sqrt{2}\beta) + 2\alpha$. In the hypothetical fully π-localized non-interacting system, each new electron would go into the non-interacting p orbital, also contributing each time a factor of α to the energy (by definition of α). Thus, the Hückel resonance energies of the allyl radical and the allyl anion are the same as for the allyl cation, namely, 0.83β.

Unfortunately, while it is clear that the allyl cation, radical, and anion all enjoy some degree of resonance stabilization, neither experiment, in the form of measured rotational barriers, nor higher levels of theory support the notion that in all three cases the magnitude is the same (see, for instance, Gobbi and Frenking 1994; Mo *et al.* 1996). So, what aspects of Hückel theory render it incapable of accurately distinguishing between these three allyl systems?

4.5 Many-electron Wave Functions

In our Hückel theory example, we derived molecular orbitals and molecular orbital energies using a *one-electron formalism*, and we then assumed that the energy of a many-electron system could be determined simply as the sum of the energies of the occupied one-electron orbitals (we used our chemical intuition to limit ourselves to two electrons per orbital). We further assumed that the orbitals themselves are invariant to the number of electrons in the π system. One might be tempted to say that Hückel theory thus ignores electron–electron repulsion. This is a bit unfair, however. By deriving our Hamiltonian matrix elements from experimental quantities (ionization potentials and rotational barriers) we have implicitly accounted for electron–electron repulsion in some sort of average way, but such an approach, known as an 'effective Hamiltonian' method, is necessarily rather crude. Thus, while Hückel theory continues to find use even today in qualitative studies of conjugated systems, it is

rarely sufficiently accurate for quantitative assessments. To improve our models, we need to take a more sophisticated accounting of many-electron effects.

4.5.1 Hartree-product Wave Functions

Let us examine the Schrödinger equation in the context of a one-electron Hamiltonian a little more carefully. When the only terms in the Hamiltonian are the one-electron kinetic energy and nuclear attraction terms, the operator is 'separable' and may be expressed as

$$H = \sum_{i=1}^{N} h_i \tag{4.32}$$

where N is the total number of electrons and h_i is the one-electron Hamiltonian defined by

$$h_i = -\frac{1}{2}\nabla_i^2 - \sum_{k=1}^{M} \frac{Z_k}{r_{ik}} \tag{4.33}$$

where M is the total number of nuclei (note that Eq. (4.33) is written in atomic units).

Eigenfunctions of the one-electron Hamiltonian defined by Eq. (4.33) must satisfy the corresponding one-electron Schrödinger equation

$$h_i \psi_i = \varepsilon_i \psi_i \tag{4.34}$$

Because the Hamiltonian operator defined by Eq. (4.32) is separable, its many-electron eigenfunctions can be constructed as products of one-electron eigenfunctions. That is

$$\Psi_{\mathrm{HP}} = \psi_1 \psi_2 \cdots \psi_N \tag{4.35}$$

A wave function of the form of Eq. (4.35) is called a 'Hartree-product' wave function.

The eigenvalue of Ψ is readily found from proving the validity of Eq. (4.35), viz.,

$$\begin{aligned}
H\Psi_{\mathrm{HP}} &= H\psi_1\psi_2\cdots\psi_N \\
&= \sum_{i=1}^{N} h_i \psi_1 \psi_2 \cdots \psi_N \\
&= (h_1\psi_1)\psi_2\cdots\psi_N + \psi_1(h_2\psi_2)\cdots\psi_N + \ldots + \psi_1\psi_2\cdots(h_N\psi_N) \\
&= (\varepsilon_1\psi_1)\psi_2\cdots\psi_N + \psi_1(\varepsilon_2\psi_2)\cdots\psi_N + \ldots + \psi_1\psi_2\cdots(\varepsilon_N\psi_N) \\
&= \sum_{i=1}^{N} \varepsilon_i \psi_1 \psi_2 \cdots \psi_N \\
&= \left(\sum_{i=1}^{N} \varepsilon_i\right) \Psi_{\mathrm{HP}}
\end{aligned} \tag{4.36}$$

where repeated application of Eq. (4.34) is used in proving that the energy eigenvalue of the many-electron wave function is simply the sum of the one-electron energy eigenvalues. Note that Eqs. (4.32)–(4.36) provide the mathematical rigor behind the Hückel theory example presented more informally above. Note that if every ψ is normalized then Ψ_{HP} is also normalized, since $|\Psi_{HP}|^2 = |\psi_1|^2 |\psi_2|^2 \cdots |\psi_N|^2$.

4.5.2 The Hartree Hamiltonian

As noted above, however, the Hamiltonian defined by Eqs. (4.32) and (4.33) does *not* include interelectronic repulsion, computation of which is vexing because it depends not on one electron, but instead on all possible (simultaneous) pairwise interactions. We may ask, however, how useful is the Hartree-product wave function in computing energies from the *correct* Hamiltonian? That is, we wish to find orbitals ψ that minimize $\langle \Psi_{HP} | H | \Psi_{HP} \rangle$. By applying variational calculus, one can show that each such orbital ψ_i is an eigenfunction of its own operator h_i defined by

$$h_i = -\frac{1}{2}\nabla_i^2 - \sum_{k=1}^{M} \frac{Z_k}{r_{ik}} + V_i\{j\} \tag{4.37}$$

where the final term represents an interaction potential with all of the other electrons occupying orbitals $\{j\}$ and may be computed as

$$V_i\{j\} = \sum_{j \neq i} \int \frac{\rho_j}{r_{ij}} d\mathbf{r} \tag{4.38}$$

where ρ_j is the charge (probability) density associated with electron j. The repulsive third term on the r.h.s. of Eq. (4.37) is thus exactly analogous to the attractive second term, except that nuclei are treated as point charges, while electrons, being treated as wave functions, have their charge spread out, so an integration over all space is necessary. Recall, however, that $\rho_j = |\psi_j|^2$. Since the point of undertaking the calculation is to determine the individual ψ, how can they be used in the one-electron Hamiltonians before they are known?

To finesse this problem, Hartree (1928) proposed an iterative 'self-consistent field' (SCF) method. In the first step of the SCF process, one *guesses* the wave functions ψ for all of the occupied MOs (AOs in Hartree's case, since he was working exclusively with atoms) and uses these to construct the necessary one-electron operators h. Solution of each differential Eq. (4.34) (in an atom, with its spherical symmetry, this is relatively straightforward, and Hartree was helped by his retired father who enjoyed the mathematical challenge afforded by such calculations) provides a *new* set of ψ, presumably different from the initial guess. So, the one-electron Hamiltonians are formed anew using these presumably more accurate ψ to determine each necessary ρ, and the process is repeated to obtain a still better set of ψ. At some point, the difference between a newly determined set and the immediately preceding set falls below some threshold criterion, and we refer to the final set of ψ as the 'converged' SCF orbitals. (An example of a threshold criterion might be that the total electronic energy change by no more than 10^{-6} a.u., and/or that the energy eigenvalue for each MO change by

no more than that amount – such criteria are, of course, entirely arbitrary, and it is typically only by checking computed properties for wave functions computed with varying degrees of imposed 'tightness' that one can determine an optimum balance between convergence and accuracy – the tighter the convergence, the more SCF cycles required, and the greater the cost in computational resources.)

Notice, from Eq. (4.36), that the sum of the individual operators h defined by Eq. (4.37) defines a separable Hamiltonian operator for which Ψ_{HP} is an eigenfunction. This separable Hamiltonian corresponds to a 'non-interacting' system of electrons (in the sense that each individual electron sees simply a constant potential with which it interacts – the nomenclature can be slightly confusing since the potential *does* derive in an average way from the other electrons, but the point is that their interaction is not accounted for instantaneously). The non-interacting Hamiltonian is *not* a good approximation to the true Hamiltonian, however, because each h includes the repulsion of its associated electron with all of the other electrons, i.e., h_i includes the repulsion between electron i and electron j, but so too does h_j. Thus, if we were to sum all of the one-electron eigenvalues for the operators h_i, which according to Eq. (4.36) would give us the eigenvalue for our non-interacting Hamiltonian, we would double-count the electron–electron repulsion. It is a straightforward matter to correct for this double-counting, however, and we may in principle compute $E = \langle \Psi_{HP} | H | \Psi_{HP} \rangle$ not directly but rather as

$$E = \sum_i \varepsilon_i - \frac{1}{2} \sum_{i \neq j} \int\int \frac{|\psi_i|^2 |\psi_j|^2}{r_{ij}} d\mathbf{r}_i d\mathbf{r}_j \tag{4.39}$$

where i and j run over all the electrons, ε_i is the energy of MO i from the solution of the one-electron Schrödinger equation using the one-electron Hamiltonian defined by Eq. (4.37), and we have replaced ρ with the square of the wave function to emphasize how it is determined (again, the double integration over all space derives from the wave function character of the electron – the double integral appearing on the r.h.s. of Eq. (4.39) is called a 'Coulomb integral' and is often abbreviated as J_{ij}). In spite of the significant difference between the non-interacting Hamiltonian and the correct Hamiltonian, operators of the former type have important utility, as we will see in Sections 7.4.2 and 8.3 within the contexts of perturbation theory and density functional theory, respectively.

At this point it is appropriate to think about our Hartree-product wave function in more detail. Let us say we have a system of eight electrons. How shall we go about placing them into MOs? In the Hückel example above, we placed them in the lowest energy MOs first, because we wanted ground electronic states, but we also limited ourselves to two electrons per orbital. Why? The answer to that question requires us to introduce something we have ignored up to this point, namely spin.

4.5.3 Electron Spin and Antisymmetry

All electrons are characterized by a spin quantum number. The electron spin function is an eigenfunction of the operator S_z and has only two eigenvalues, $\pm\hbar/2$; the spin eigenfunctions

are orthonormal and are typically denoted as α and β (not to be confused with the α and β of Hückel theory!) The spin quantum number is a natural consequence of the application of relativistic quantum mechanics to the electron (i.e., accounting for Einstein's theory of relativity in the equations of quantum mechanics), as first shown by Dirac. Another consequence of relativistic quantum mechanics is the so-called Pauli exclusion principle, which is usually stated as the assertion that no two electrons can be characterized by the same set of quantum numbers. Thus, in a given MO (which defines all electronic quantum numbers except spin) there are only two possible choices for the remaining quantum number, α or β, and thus only two electrons may be placed in any MO.

Knowing these aspects of quantum mechanics, if we were to construct a ground-state Hartree-product wave function for a system having two electrons of the same spin, say α, we would write

$$^3\Psi_{\mathrm{HP}} = \psi_a(1)\alpha(1)\psi_b(2)\alpha(2) \tag{4.40}$$

where the left superscript 3 indicates a triplet electronic state (two electrons spin parallel) and ψ_a and ψ_b are different from one another (since otherwise electrons 1 and 2 would have all identical quantum numbers) and orthonormal. However, the wave function defined by Eq. (4.40) is fundamentally flawed. The Pauli exclusion principle is an important mnemonic, but it actually derives from a feature of relativistic quantum field theory that has more general consequences, namely that electronic wave functions must *change sign* whenever the coordinates of two electrons are interchanged. Such a wave function is said to be 'antisymmetric'. For notational purposes, we can define the permutation operator P_{ij} as the operator that interchanges the coordinates of electrons i and j. Thus, we would write the Pauli principle for a system of N electrons as

$$\begin{aligned} P_{ij}\Psi[\mathbf{q}_1(1), \ldots, \mathbf{q}_i(i), \ldots, \mathbf{q}_j(j), \ldots, \mathbf{q}_N(N)] \\ = \Psi[\mathbf{q}_1(1), \ldots, \mathbf{q}_j(i), \ldots, \mathbf{q}_i(j), \ldots, \mathbf{q}_N(N)] \\ = -\Psi[\mathbf{q}_1(1), \ldots, \mathbf{q}_i(i), \ldots, \mathbf{q}_j(j), \ldots, \mathbf{q}_N(N)] \end{aligned} \tag{4.41}$$

where $\mathbf{q}$ now includes not only the three cartesian coordinates but also the spin function.

If we apply P_{12} to the Hartree-product wave function of Eq. (4.40),

$$\begin{aligned} P_{ij}[\psi_a(1)\alpha(1)\psi_b(2)\alpha(2)] &= \psi_b(1)\alpha(1)\psi_a(2)\alpha(2) \\ &\neq -\psi_a(1)\alpha(1)\psi_b(2)\alpha(2) \end{aligned} \tag{4.42}$$

we immediately see that it does *not* satisfy the Pauli principle. However, a slight modification to Ψ_{HP} can be made that causes it to satisfy the constraints of Eq. (4.41), namely

$$^3\Psi_{\mathrm{SD}} = \frac{1}{\sqrt{2}}[\psi_a(1)\alpha(1)\psi_b(2)\alpha(2) - \psi_a(2)\alpha(2)\psi_b(1)\alpha(1)] \tag{4.43}$$

(the reader is urged to verify that $^3\Psi_{SD}$ does indeed satisfy the Pauli principle; for the 'SD' subscript, see next section). Note that if we integrate $|^3\Psi_{SD}|^2$ over all space we have

$$\begin{aligned}\int |^3\Psi_{SD}|^2 \, d\mathbf{r}_1 d\omega_1 d\mathbf{r}_2 d\omega_2 = \frac{1}{2}\Bigg[& \int |\psi_a(1)|^2 \, |\alpha(1)|^2 \, |\psi_b(2)|^2 \, |\alpha(2)|^2 d\mathbf{r}_1 d\omega_1 d\mathbf{r}_2 d\omega_2 \\ & - 2 \int \psi_a(1)\psi_b(1) \, |\alpha(1)|^2 \, \psi_b(2)\psi_a(2) \, |\alpha(2)|^2 \, d\mathbf{r}_1 d\omega_1 d\mathbf{r}_2 d\omega_2 \\ & + \int |\psi_a(2)|^2 \, |\alpha(2)|^2 \, |\psi_b(1)|^2 \, |\alpha(1)|^2 d\mathbf{r}_1 d\omega_1 d\mathbf{r}_2 d\omega_2 \Bigg] \\ = \frac{1}{2}(1 - 0 + 1) \\ = 1 \end{aligned} \tag{4.44}$$

where ω is a spin integration variable, the simplification of the various integrals on the r.h.s. proceeds from the orthonormality of the MOs and spin functions, and we see that the prefactor of $2^{-1/2}$ in Eq. (4.43) is required for normalization.

4.5.4 Slater Determinants

A different mathematical notation can be used for Eq. (4.43)

$$^3\Psi_{SD} = \frac{1}{\sqrt{2}} \begin{vmatrix} \psi_a(1)\alpha(1) & \psi_b(1)\alpha(1) \\ \psi_a(2)\alpha(2) & \psi_b(2)\alpha(2) \end{vmatrix} \tag{4.45}$$

where the difference of MO products has been expressed as a determinant. Note that the permutation operator P applied to a determinant has the effect of interchanging two of the rows. It is a general property of a determinant that it changes sign when any two rows (or columns) are interchanged, and the utility of this feature for use in constructing antisymmetric wave functions was first exploited by Slater (1929). Thus, the 'SD' subscript used in Eqs. (4.43)–(4.45) stands for 'Slater determinant'. On a term-by-term basis, Slater-determinantal wave functions quickly become rather tedious to write down, but determinantal notation allows them to be expressed reasonably compactly as, in general,

$$\Psi_{SD} = \frac{1}{\sqrt{N!}} \begin{vmatrix} \chi_1(1) & \chi_2(1) & \cdots & \chi_N(1) \\ \chi_1(2) & \chi_2(2) & \cdots & \chi_N(2) \\ \vdots & \vdots & \ddots & \vdots \\ \chi_1(N) & \chi_2(N) & \cdots & \chi_N(N) \end{vmatrix} \tag{4.46}$$

where N is the total number of electrons and χ is a spin-orbital, i.e., a product of a spatial orbital and an electron spin eigenfunction. A still more compact notation that finds widespread use is

$$\Psi_{SD} = |\chi_1 \chi_2 \chi_3 \cdots \chi_N\rangle \tag{4.47}$$

where the prefactor $(N!)^{-1/2}$ is implicit. Furthermore, if two spin orbitals differ only in the spin eigenfunction (i.e., together they represent a doubly filled orbital) this is typically represented by writing the spatial wave function with a superscript 2 to indicate double occupation. Thus, if χ_1 and χ_2 represented α and β spins in spatial orbital ψ_1, one would write

$$\Psi_{SD} = |\psi_1^2 \chi_3 \cdots \chi_N\rangle \tag{4.48}$$

Slater determinants have a number of interesting properties. First, note that every electron appears in every spin orbital somewhere in the expansion. This is a manifestation of the indistinguishability of quantum particles (which is *violated* in the Hartree-product wave functions). A more subtle feature is so-called quantum mechanical exchange. Consider the energy of interelectronic repulsion for the wave function of Eq. (4.43). We evaluate this as

$$\begin{aligned}
&\int {}^3\Psi_{SD} \frac{1}{r_{12}} {}^3\Psi_{SD} d\mathbf{r}_1 d\omega_1 d\mathbf{r}_2 d\omega_2 \\
&= \frac{1}{2}\Bigg[\int |\psi_a(1)|^2 |\alpha(1)|^2 \frac{1}{r_{12}} |\psi_b(2)|^2 |\alpha(2)|^2 d\mathbf{r}_1 d\omega_1 d\mathbf{r}_2 d\omega_2 \\
&\quad - 2\int \psi_a(1)\psi_b(1) |\alpha(1)|^2 \frac{1}{r_{12}} \psi_b(2)\psi_a(2) |\alpha(2)|^2 d\mathbf{r}_1 d\omega_1 d\mathbf{r}_2 d\omega_2 \\
&\quad + \int |\psi_a(2)|^2 |\alpha(2)|^2 \frac{1}{r_{12}} |\psi_b(1)|^2 |\alpha(1)|^2 d\mathbf{r}_1 d\omega_1 d\mathbf{r}_2 d\omega_2\Bigg] \\
&= \frac{1}{2}\Bigg[\int |\psi_a(1)|^2 \frac{1}{r_{12}} |\psi_b(2)|^2 d\mathbf{r}_1 d\mathbf{r}_2 \\
&\quad - 2\int \psi_a(1)\psi_b(1) \frac{1}{r_{12}} \psi_b(2)\psi_a(2) d\mathbf{r}_1 d\mathbf{r}_2 \\
&\quad + \int |\psi_a(2)|^2 \frac{1}{r_{12}} |\psi_b(1)|^2 d\mathbf{r}_1 d\mathbf{r}_2\Bigg] \\
&= \frac{1}{2}\left(J_{ab} - 2\int \psi_a(1)\psi_b(1) \frac{1}{r_{12}} \psi_a(2)\psi_b(2) d\mathbf{r}_1 d\mathbf{r}_2 + J_{ab}\right) \\
&= J_{ab} - K_{ab}
\end{aligned} \tag{4.49}$$

Equation (4.49) indicates that for this wave function the classical Coulomb repulsion between the electron clouds in orbitals a and b is reduced by K_{ab}, where the definition of this integral may be inferred from comparing the third equality to the fourth. This fascinating consequence of the Pauli principle reflects the reduced probability of finding two electrons of the same spin close to one another – a so-called 'Fermi hole' is said to surround each electron.

Note that this property is a correlation effect *unique to electrons of the same spin*. If we consider the contrasting Slater determinantal wave function formed from different spins

$$\Psi_{SD} = \frac{1}{\sqrt{2}}[\psi_a(1)\alpha(1)\psi_b(2)\beta(2) - \psi_a(2)\alpha(2)\psi_b(1)\beta(1)] \tag{4.50}$$

and carry out the same evaluation of interelectronic repulsion we have

$$\begin{aligned}
&\int \Psi_{SD}\frac{1}{r_{12}}\Psi_{SD}d\mathbf{r}_1 d\omega_1 d\mathbf{r}_2 d\omega_2 \\
&= \frac{1}{2}\left[\int |\psi_a(1)|^2\,|\alpha(1)|^2\,\frac{1}{r_{12}}\,|\psi_b(2)|^2\,|\beta(2)|^2\,d\mathbf{r}_1 d\omega_1 d\mathbf{r}_2 d\omega_2\right. \\
&\quad - 2\int \psi_a(1)\psi_b(1)\alpha(1)\beta(1)\frac{1}{r_{12}}\psi_b(2)\psi_a(2)\alpha(2)\beta(2)d\mathbf{r}_1 d\omega_1 d\mathbf{r}_2 d\omega_2 \\
&\quad \left.+\int |\psi_a(2)|^2\,|\alpha(2)|^2\,\frac{1}{r_{12}}\,|\psi_b(1)|^2\,|\beta(1)|^2\,d\mathbf{r}_1 d\omega_1 d\mathbf{r}_2 d\omega_2\right] \\
&= \frac{1}{2}\left[\int |\psi_a(1)|^2\,\frac{1}{r_{12}}\,|\psi_b(2)|^2\,d\mathbf{r}_1 d\mathbf{r}_2\right. \\
&\quad - 2\cdot 0 \\
&\quad \left.+\int |\psi_a(2)|^2\,\frac{1}{r_{12}}\,|\psi_b(1)|^2\,d\mathbf{r}_1 d\mathbf{r}_2\right] \\
&= \frac{1}{2}(J_{ab} + J_{ab}) \\
&= J_{ab}
\end{aligned} \tag{4.51}$$

Note that the disappearance of the exchange correlation derives from the orthogonality of the α and β spin functions, which causes the second integral in the second equality to be zero when integrated over either spin coordinate.

4.5.5 The Hartree-Fock Self-consistent Field Method

Fock first proposed the extension of Hartree's SCF procedure to Slater determinantal wave functions. Just as with Hartree product orbitals, the HF MOs can be individually determined as eigenfunctions of a set of one-electron operators, but now the interaction of each electron with the static field of all of the other electrons (this being the basis of the SCF approximation) includes exchange effects on the Coulomb repulsion. Some years later, in a paper that was critical to the further development of practical computation, Roothaan described matrix algebraic equations that permitted HF calculations to be carried out using a basis set representation for the MOs (Roothaan 1951; for historical insights, see Zerner 2000). We will forego a formal derivation of all aspects of the HF equations, and simply present them in their typical form for closed-shell systems (i.e., all electrons spin-paired, two per occupied orbital) with wave functions represented as a single Slater determinant. This formalism is called 'restricted Hartree-Fock' (RHF); alternative fomalisms are discussed in Chapter 6.

The one-electron Fock operator is defined for each electron i as

$$f_i = -\frac{1}{2}\nabla_i^2 - \sum_k^{\text{nuclei}} \frac{Z_k}{r_{ik}} + V_i^{\text{HF}}\{j\} \tag{4.52}$$

where the final term, the HF potential, is $2J_i - K_i$, and the J_i and K_i operators are defined so as to compute the J_{ij} and K_{ij} integrals previously defined above. To determine the MOs using the Roothaan approach, we follow a procedure analogous to that previously described for Hückel theory. First, given a set of N basis functions, we solve the secular equation

$$\begin{vmatrix} F_{11} - ES_{11} & F_{12} - ES_{12} & \cdots & F_{1N} - ES_{1N} \\ F_{21} - ES_{21} & F_{22} - ES_{22} & \cdots & F_{2N} - ES_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ F_{N1} - ES_{N1} & F_{N2} - ES_{N2} & \cdots & F_{NN} - ES_{NN} \end{vmatrix} = 0 \tag{4.53}$$

to find its various roots E_j. In this case, the values for the matrix elements F and S are computed explicitly.

Matrix elements S are the overlap matrix elements we have seen before. For a general matrix element $F_{\mu\nu}$ (we here adopt a convention that basis functions are indexed by lower-case Greek letters, while MOs are indexed by lower-case Roman letters) we compute

$$F_{\mu\nu} = \left\langle \mu \left| -\frac{1}{2}\nabla^2 \right| \nu \right\rangle - \sum_{k}^{\text{nuclei}} Z_k \left\langle \mu \left| \frac{1}{r_k} \right| \nu \right\rangle + \sum_{\lambda\sigma} P_{\lambda\sigma} \left[(\mu\nu|\lambda\sigma) - \frac{1}{2}(\mu\lambda|\nu\sigma) \right] \tag{4.54}$$

The notation $\langle \mu | g | \nu \rangle$ where g is some operator which takes basis function ϕ_ν as its argument, implies a so-called one-electron integral of the form

$$\langle \mu | g | \nu \rangle = \int \phi_\mu (g\phi_\nu) d\mathbf{r}. \tag{4.55}$$

Thus, for the first term in Eq. (4.54) g involves the Laplacian operator and for the second term g is the distance operator to a particular nucleus. The notation $(\mu\nu|\lambda\sigma)$ also implies a specific integration, in this case

$$(\mu\nu|\lambda\sigma) = \int\int \phi_\mu(1)\phi_\nu(1)\frac{1}{r_{12}}\phi_\lambda(2)\phi_\sigma(2) d\mathbf{r}(1) d\mathbf{r}(2) \tag{4.56}$$

where ϕ_μ and ϕ_ν represent the probability density of one electron and ϕ_λ and ϕ_σ the other. The exchange integrals $(\mu\lambda|\nu\sigma)$ are preceded by a factor of 1/2 because they are limited to electrons of the same spin while Coulomb interactions are present for any combination of spins.

The final sum in Eq. (4.54) weights the various so-called 'four-index integrals' by elements of the 'density matrix' $\mathbf{P}$. This matrix in some sense describes the degree to which individual basis functions contribute to the many-electron wave function, and thus how energetically important the Coulomb and exchange integrals should be (i.e., if a basis function fails to contribute in a significant way to any occupied MO, clearly integrals involving that basis

function should be of no energetic importance). The elements of **P** are computed as

$$P_{\lambda\sigma} = 2 \sum_{i}^{\text{occupied}} a_{\lambda i} a_{\sigma i} \tag{4.57}$$

where the coefficients $a_{\zeta i}$ specify the (normalized) contribution of basis function ζ to MO i and the factor of two appears because with RHF theory we are considering only singlet wave functions in which all orbitals are doubly occupied.

While the process of solving the HF secular determinant to find orbital energies and coefficients is quite analogous to that already described above for effective Hamiltonian methods, it is characterized by the same paradox present in the Hartree formalism. That is, we need to know the orbital coefficients to form the density matrix that is used in the Fock matrix elements, but the purpose of solving the secular equation is to determine those orbital coefficients. So, just as in the Hartree method, the HF method follows a SCF procedure, where first we guess the orbital coefficients (e.g., from an effective Hamiltonian method) and then we iterate to convergence. The full process is described schematically by the flow chart in Figure 4.3. The energy of the HF wavefunction can be computed in a fashion analogous to Eq. (4.39).

Hartree–Fock theory as constructed using the Roothaan approach is quite beautiful in the abstract. This is not to say, however, that it does not suffer from certain chemical and practical limitations. Its chief chemical limitation is the one-electron nature of the Fock operators. Other than exchange, all electron correlation is ignored. It is, of course, an interesting question to ask just how important such correlation is for various molecular properties, and we will examine that in some detail in following chapters.

Furthermore, from a practical standpoint, HF theory posed some very challenging technical problems to early computational chemists. One problem was choice of a basis set. The LCAO approach using hydrogenic orbitals remains attractive in principle; however, this basis set requires numerical solution of the four-index integrals appearing in the Fock matrix elements, and that is a very tedious process. Moreover, the *number* of four-index integrals is daunting. Since each index runs over the total number of basis functions, there are in principle N^4 total integrals to be evaluated, and this quartic scaling behavior with respect to basis-set size proves to be the bottleneck in HF theory applied to essentially any molecule.

Historically, two philosophies began to emerge at this stage with respect to how best to make further progress. The first philosophy might be summed up as follows: The HF equations are very powerful but still, after all, chemically flawed. Thus, other approximations that may be introduced to simplify their solution, and possibly at the same time improve their accuracy (by some sort of parameterization to reproduce key experimental quantities), are well justified. Many computational chemists continue to be guided by this philosophy today, and it underlies the motivation for so-called 'semiempirical' MO theories, which are discussed in detail in the next chapter.

The second philosophy essentially views HF theory as a stepping stone on the way to exact solution of the Schrödinger equation. HF theory provides a very well defined energy, one which can be converged in the limit of an infinite basis set, and the difference between that

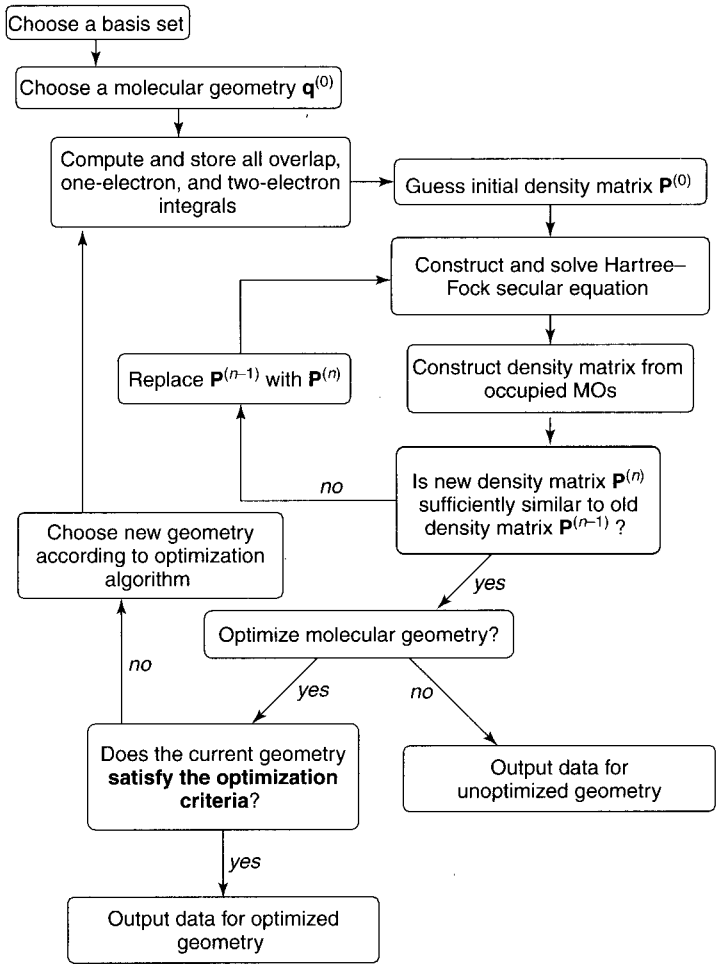


Figure 4.3 Flow chart of the HF SCF procedure. Note that data for an unoptimized geometry is referred to as deriving from a so-called 'single-point calculation'

converged energy and reality is the electron correlation energy (ignoring relativity, spin–orbit coupling, etc.). It was anticipated that developing the technology to achieve the HF limit *with no further approximations* would not only permit the evaluation of the chemical utility of the HF limit, but also probably facilitate moving on from that base camp to the Schrödinger equation summit. Such was the foundation for further research on '*ab initio*' HF theory, which forms the subject of Chapter 6.

Bibliography and Suggested Additional Reading

Frenking, G. 2000. "Perspective on 'Quantentheoretische Beiträge zum Benzolproblem. I. Die Elektronenkonfiguration des Benzols und verwandter Beziehungen" *Theor. Chem. Acc.*, **103**, 187.

Hehre, W. J., Radom, L., Schleyer, P. v. R., and Pople, J. A. 1986. Ab Initio *Molecular Orbital Theory*, Wiley: New York.
Levine, I. N. 2000. *Quantum Chemistry*, 5th Edn., Prentice Hall: New York.
Lowry, T. H. and Richardson, K. S. 1981. *Mechanism and Theory in Organic Chemistry*, 2nd Edn., Harper & Row: New York, 82–112.
Szabo, A. and Ostlund, N. S. 1982. *Modern Quantum Chemistry*, Macmillan: New York.

References

Berson, J. A. 1996. *Angew. Chem., Int. Ed. Engl.*, **35**, 2750.
Frenking. G. 2000. *Theor. Chem. Acc.*, **103**, 187.
Gobbi, A. and Frenking, G. 1994. *J. Am. Chem. Soc.*, **116**, 9275.
Hartree, D. R. 1928. *Proc. Cambridge Phil. Soc.*, **24**, 89, 111, 426.
Hückel, E. 1931. *Z. Phys.*, **70**, 204.
MacDonald, J. K. L. 1933. *Phys. Rev.*, **43**, 830.
Mo, Y. R., Lin, Z. Y., Wu, W., and Zhang, Q. N. 1996. *J. Phys. Chem.*, **100**, 6469.
Roothaan, C. C. J. 1951. *Rev. Mod. Phys.*, **23**, 69.
Slater, J. C., 1930. *Phys. Rev.*, **35**, 210.
Zerner, M. C. 2000. *Theor. Chem. Acc.*, **103**, 217.

5

Semiempirical Implementations of Molecular Orbital Theory

5.1 Semiempirical Philosophy

In the last chapter, the full formalism of Hartree–Fock theory was developed. While this theory is impressive as a physical and mathematical construct, it has several limitations in a practical sense. Particularly during the early days of computational chemistry, when computational power was minimal, carrying out HF calculations without any further approximations, even for small systems with small basis sets, was a challenging task.

In spite of the technical hurdles, however, many chemists recognized the potentially critical role that theory could play in furthering experimental progress on any number of fronts. And the interests of that population of chemists were by no means restricted to molecules composed of only a small handful of atoms. Accepting HF theory as a framework, several research groups turned their attention to implementations of the theory that would make it more tractable, and perhaps more accurate, for molecules of moderate size. These steps 'sideways', if you will, led to a certain bifurcation of effort in the area of molecular orbital theory (although certainly some research groups pursued topics in both directions) that persists to this day. Semiempirical calculations continue to appear in large numbers in the chemical literature; since there will always be researchers interested in molecules that exceed the size of those practically accessible by *ab initio* methods, semiempirical levels of MO theory are certain to continue to be developed and applied. This chapter describes the underlying approximations of semiempirical methods (organizing them roughly in chronological order of appearance) and provides detailed comparisons between methods now in common use for the prediction of various chemical properties. Section 5.7 describes recent developments in the area of improving/extending semiempirical models.

5.1.1 Chemically Virtuous Approximations

Let us consider how one might go about making formal Hartree–Fock theory less computationally intensive without necessarily sacrificing its accuracy. The most demanding step

of an HF calculation, in terms of computational resources, is the assembly of the two-electron (also called four-index) integrals, i.e., the J and K integrals appearing in the Fock matrix elements defined by Eq. (4.54). Not only is numerical solution of the integrals for an arbitrary basis set arduous, but there are so many of them (formally N^4 where N is the number of basis functions). One way to save time would be to estimate their value accurately in an *a priori* fashion, so that no numerical integration need be undertaken.

For which integrals is it easiest to make such an estimation? To answer that question, it is helpful to keep in mind the intuitive meaning of the integrals. Coulomb integrals measure the repulsion between two electrons in regions of space defined by the basis functions. It seems clear, then, that when the basis functions in the integral for one electron are very far from the basis functions for the other, the value of that integral will approach zero (the same holds true for the one-electron integrals describing nuclear attraction, i.e., if the basis functions for the electron are very far from the nucleus the attraction will go to zero, but these integrals are much less computationally demanding to solve). In a large molecule, then, one might be able to avoid the calculation of a very large number of integrals simply by assuming them to be zero, and one would still have a reasonable expectation of obtaining a Hartree–Fock energy close to that that would be obtained from a full calculation.

Such an approximation is what we might call a numerical approximation. That is, it introduces error to the extent that values employed are not exact, but the calculation can be converged to arbitrary accuracy by tightening the criteria for employing the approximation, e.g., in the case of setting certain two-electron integrals to zero, the threshold could be the average inter-basis-function distance, so that in the limit of choosing a distance of infinity, one recovers exact HF theory. Other approximations in semiempirical theory, however, are guided by a slightly different motivation, and these approximations might be well referred to as 'chemically virtuous approximations'. It is important to keep in mind that HF wave functions for systems having two or more electrons are *not* eigenfunctions of the corresponding non-relativistic Schrödinger equations. Because of the SCF approximation for how each electron interacts with all of the others, some electronic correlation is ignored, and the HF energy is necessarily higher than the exact energy.

How important is the correlation energy? Let us consider a very simple system: the helium atom. The energy of this two-electron system in the HF limit (i.e., converged with respect to basis-set size for the number of digits reported) is −2.861 68 a.u. (Clementi and Roetti 1974). The exact energy for the helium atom, on the other hand, is −2.903 72 a.u. (Pekeris 1959). The difference is 0.042 04 a.u., which is about 26 kcal mol^{-1}. Needless to say, as systems increase in size, greater numbers of electrons give rise to considerably larger correlation energies – hundreds or thousands of kcal mol^{-1} for moderately sized organic and inorganic molecules.

At first glance, this is a terrifying observation. At room temperature (298 K), it requires a change of 1.4 kcal mol^{-1} in a free energy of reaction to change an equilibrium constant by an order of magnitude. Similarly, a change of 1.4 kcal mol^{-1} in a rate-determining free energy of activation will change the rate of a chemical reaction by an order of magnitude. Thus, chemists typically would prefer theoretical accuracies to be no worse than 1.4 kcal mol^{-1},

so that room-temperature predictions can be trusted at least to within an order of magnitude (and obviously it would be nice to do much better). How then can we ever hope to use a theory that is intrinsically inaccurate by hundreds or thousands of kilocalories per mole to make chemically useful predictions? Michael J. S. Dewar, who made many contributions in the area of semiempirical MO theory, once offered the following analogy to using HF theory to make chemical predictions: It is like weighing the captain of a ship by first weighing the ship with the captain on board, then weighing the ship without her, and then taking the difference – the errors in the individual measurements are likely to utterly swamp the small difference that is the goal of the measuring.

In practice, as we shall see in Chapter 6, the situation with HF theory is not really as bad as our above analysis might suggest. Errors from neglecting correlation energy cancel to a remarkable extent in favorable instances, so that chemically useful interpretations of HF calculations *can* be valid. Nevertheless, the intrinsic inaccuracy of *ab initio* HF theory suggests that modifications of the theory introduced in order to simplify its formalism may actually *improve on* a rigorous adherence to the full mathematics, provided the new 'approximations' somehow introduce an accounting for correlation energy. Since this improves chemical accuracy, at least in intent, we may call it a chemically virtuous approximation. Most typically, such approximations involve the adoption of a parametric form for some aspect of the calculation where the parameters involved are chosen so as best to reproduce experimental data – hence the term 'semiempirical'.

5.1.2 Analytic Derivatives

If it is computationally demanding to carry out a single electronic structure calculation, how much more daunting to try to optimize a molecular geometry. As already discussed in detail in Section 2.4, chemists are usually interested not in arbitrary structures, but in stationary points on the potential energy surface. In order to find those points efficiently, many of the optimization algorithms described in Section 2.4 make use of derivatives of the energy with respect to nuclear motion – when those derivatives are available analytically, instead of numerically, rates of convergence are typically enhanced.

This is particularly true when the stationary point of interest is a transition-state structure. Unlike the case with molecular mechanics, the HF energy has no obvious bias for minimum-energy structures compared to TS structures – one of the most exciting aspects of MO theory, whether semiempirical or *ab initio*, is that it provides an energy functional from which reasonable TS structures may be identified. However, in the early second half of the twentieth century, it was not at all obvious how to compute analytic derivatives of the HF energy with respect to nuclear motion. Thus, another motivation for introducing semiempirical approximations into HF theory was to facilitate the computation of derivatives so that geometries could be more efficiently optimized. Besides the desire to attack TS geometries, there were also very practical motivations for geometry optimization. In the early days of semiempirical parameterization, experimental structural data were about as widely available as energetic data, and parameterization of semiempirical methods against both kinds of data would be expected to generate a more robust final model.

5.2 Extended Hückel Theory

Prior to considering semiempirical methods designed on the basis of HF theory, it is instructive to revisit one-electron effective Hamiltonian methods like the Hückel model described in Section 4.4. Such models tend to involve the most drastic approximations, but as a result their rationale is tied closely to experimental concepts and they tend to be intuitive. One such model that continues to see extensive use today is the so-called extended Hückel theory (EHT). Recall that the key step in finding the MOs for an effective Hamiltonian is the formation of the secular determinant for the secular equation

$$\begin{vmatrix} H_{11} - ES_{11} & H_{12} - ES_{12} & \cdots & H_{1N} - ES_{1N} \\ H_{21} - ES_{21} & H_{22} - ES_{22} & \cdots & H_{2N} - ES_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ H_{N1} - ES_{N1} & H_{N2} - ES_{N2} & \cdots & H_{NN} - ES_{NN} \end{vmatrix} = 0 \tag{5.1}$$

The dimension of the secular determinant for a given molecule depends on the choice of basis set. EHT adopts two critical conventions. First, all core electrons are ignored. It is assumed that core electrons are sufficiently invariant to differing chemical environments that changes in their orbitals as a function of environment are of no chemical consequence, energetic or otherwise. *All modern semiempirical methodologies make this approximation.* In EHT calculations, if an atom has occupied d orbitals, typically the highest occupied level of d orbitals is considered to contribute to the set of valence orbitals.

Each remaining valence orbital is represented by a so-called Slater-type orbital (STO). The mathematical form of a normalized STO used in EHT (in atom-centered polar coordinates) is

$$\varphi(r, \theta, \phi; \zeta, n, l, m) = \frac{(2\zeta)^{n+1/2}}{[(2n)!]^{1/2}} r^{n-1} e^{-\zeta r} Y_l^m(\theta, \phi) \tag{5.2}$$

where ζ is an exponent that can be chosen according to a simple set of rules developed by Slater that depend, *inter alia*, on the atomic number (Slater 1930), n is the principal quantum number for the valence orbital, and the spherical harmonic functions $Y_l^m(\theta, \phi)$, depending on the angular momentum quantum numbers l and m, are those familiar from solution of the Schrödinger equation for the hydrogen atom and can be found in any standard quantum mechanics text. Thus, the size of the secular determinant in Eq. (5.2) is dictated by the total number of valence orbitals in the molecule. For instance, the basis set for the MnO_4^- anion would include a total of 25 STO basis functions: one 2s and three 2p functions for each oxygen (for a subtotal of 16) and one 4s, three 4p, and five 3d functions for manganese.

STOs have a number of features that make them attractive. The orbital has the correct exponential decay with increasing r, the angular component is hydrogenic, and the 1s orbital has, as it should, a cusp at the nucleus (i.e., it is not smooth). More importantly, from a practical point of view, overlap integrals between two STOs as a function of interatomic distance are readily computed (Mulliken Rieke and Orloff 1949; Bishop 1966). Thus, in contrast to simple Hückel theory, overlap matrix elements in EHT are not assumed to be equal to the Krönecker delta, but are directly computed in every instance.

The only terms remaining to be defined in Eq. (5.1), then, are the resonance integrals H. For diagonal elements, the same convention is used in EHT as was used for simple Hückel theory. That is, the value for $H_{\mu\mu}$ is taken as the negative of the average ionization potential for an electron in the appropriate valence orbital. Thus, for instance, when μ is a hydrogen 1s function, $H_{\mu\mu} = -13.6$ eV. Of course in many-electron atoms, the valence-shell ionization potential (VSIP) for the ground-state atomic term may not necessarily be the best choice for the atom in a molecule, so this term is best regarded as an adjustable parameter, although one with a clear, physical basis. VSIPs have been tabulated for most of the atoms in the periodic table (Pilcher and Skinner 1962; Hinze and Jaffé 1962; Hoffmann 1963; Cusachs, Reynolds and Barnard 1966). Because atoms in molecular environments may develop fairly large partial charges depending on the nature of the atoms to which they are connected, schemes for adjusting the neutral atomic VSIP as a function of partial atomic charge have been proposed (Rein *et al.* 1966; Zerner and Gouterman 1966). Such an adjustment scheme characterizes so-called Fenske–Hall effective Hamiltonian calculations, which still find considerable use for inorganic and organometallic systems composed of atoms having widely different electronegativities (Hall and Fenske 1972).

The more difficult resonance integrals to approximate are the off-diagonal ones. Wolfsberg and Helmholz (1952) suggested the following convention

$$H_{\mu\nu} = \frac{1}{2} C_{\mu\nu}(H_{\mu\mu} + H_{\nu\nu}) S_{\mu\nu} \tag{5.3}$$

where C is an empirical constant and S is the overlap integral. Thus, the energy associated with the matrix element is proportional to the average of the VSIPs for the two orbitals μ and ν times the extent to which the two orbitals overlap in space (note that, by symmetry, the overlap between different STOs on the same atom is zero). Originally, the constant C was given a different value for matrix elements corresponding to σ- and π-type bonding interactions. In modern EHT calculations, it is typically taken as 1.75 for *all* matrix elements, although it can still be viewed as an adjustable parameter when such adjustment is warranted.

All of the above conventions together permit the complete construction of the secular determinant. Using standard linear algebra methods, the MO energies and wave functions can be found from solution of the secular equation. Because the matrix elements do not depend on the final MOs in any way (unlike HF theory), the process is not iterative, so it is very fast, even for very large molecules (however, the process *does* become iterative if VSIPs are adjusted as a function of partial atomic charge as described above, since the partial atomic charge depends on the occupied orbitals, as described in Chapter 9).

The very approximate nature of the resonance integrals in EHT makes it insufficiently accurate for the generation of PESs since the locations of stationary points are in general very poorly predicted. Use of EHT is thus best restricted to systems for which experimental geometries are available. For such cases, EHT tends to be used today to generate qualitatively correct MOs, in much the same fashion as it was used by Wolfsberg and Helmholz 50 years ago. Wolfsberg and Helmholz used their model to explain differences in the UV spectroscopies of MnO_4^-, CrO_4^{2-}, and ClO_4^- by showing how the different VSIPs of the central atom and differing bond lengths gave rise to different energy separations between

the relevant filled and empty orbitals in spectroscopic transitions. In the 21st century, such a molecular problem has become amenable to more accurate treatments, so the province of EHT is now primarily very large systems, like extended solids, where its speed makes it a practical option for understanding band structure (a 'band' is a set of MOs so densely spread over a range of energy that for practical purposes it may be regarded as a continuum; bands derive from combinations of molecular orbitals in a solid much as MOs derive from combinations of AOs in a molecule).

Thus, for example, EHT has been used by Genin and Hoffmann (1998) to characterize the band structure of a series of organic polymers with the intent of suggesting likely candidates for materials exhibiting organic ferromagnetism. Certain polymers formed from repeating heterocycle units having seven π electrons were identified as having narrow, half-filled valence bands, such bands being proposed as a necessary, albeit not sufficient, condition for ferromagnetism.

Note that one drawback of EHT is a failure to take into account electron spin. There is no mechanism for distinguishing between different multiplets, except that a chemist can, by hand, decide which orbitals are occupied, and thus enforce the Pauli exclusion principle. However, the energy computed for a triplet state is exactly the same as the energy for the corresponding 'open-shell' singlet (i.e., the state that results from spin-flip of one of the unpaired electrons in the triplet) – the electronic energy is the sum of the occupied orbital energies irrespective of spin – such an equality occurs experimentally only when the partially occupied orbitals fail to interact with each other either for symmetry reasons or because they are infinitely separated.

5.3 CNDO Formalism

Returning to the SCF formalism of HF theory, one can proceed in the spirit of an effective Hamiltonian method by developing a recipe for the replacement of matrix elements in the HF secular equation, Eq. (4.53). One of the first efforts along these lines was described by Pople and co-workers in 1965 (Pople, Santry, and Segal 1965; Pople and Segal 1965). The complete neglect of differential overlap (CNDO) method adopted the following conventions:

1. Just as in EHT, the basis set is formed from valence STOs, one STO per valence orbital. In the original CNDO implementation, only atoms having s and p valence orbitals were addressed.
2. In the secular determinant, overlap matrix elements are defined by

$$S_{\mu\nu} = \delta_{\mu\nu} \tag{5.4}$$

   where δ is the Krönecker delta.
3. All two-electron integrals are parameterized according to the following scheme. First, define

$$(\mu\nu|\lambda\sigma) = \delta_{\mu\nu}\delta_{\lambda\sigma}(\mu\mu|\lambda\lambda) \tag{5.5}$$

Thus, the only integrals that are non-zero have μ and ν as identical orbitals on the same atom, and λ and σ also as identical orbitals on the same atom, but the second atom might be different than the first (the decision to set to zero any integrals involving overlap of different basis functions gives rise to the model name).

4. For the surviving two-electron integrals,

$$(\mu\mu|\lambda\lambda) = \gamma_{AB} \tag{5.6}$$

where A and B are the atoms on which basis functions μ and λ reside, respectively. The term γ can either be computed explicitly from s-type STOs (note that since γ depends only on the atoms A and B, $(s_A s_A|s_B s_B) = (p_A p_A|s_B s_B) = (p_A p_A|p_B p_B)$, etc.) or it can be treated as a parameter. One popular parametric form involves using the so-called Pariser–Parr approximation for the one-center term (Pariser and Parr 1953).

$$\gamma_{AA} = IP_A - EA_A \tag{5.7}$$

where IP and EA are the atomic ionization potential and electron affinity, respectively. For the two-center term, the Mataga–Nishomoto formalism adopts

$$\gamma_{AB} = \frac{\gamma_{AA} + \gamma_{BB}}{2 + r_{AB}(\gamma_{AA} + \gamma_{BB})} \tag{5.8}$$

where r_{AB} is the interatomic distance (Mataga and Nishimoto 1957). Note the intuitive limits on γ in Eq. (5.8). At large distance, it goes to $1/r_{AB}$, as expected for widely separated charge clouds, while at short distances, it approaches the average of the two one-center parameters.

5. One-electron integrals for diagonal matrix elements are defined by

$$\left\langle \mu \left| -\frac{1}{2}\nabla^2 - \sum_k \frac{Z_k}{r_k} \right| \mu \right\rangle = -IP_\mu - \sum_k (Z_k - \delta_{Z_A Z_k})\gamma_{Ak} \tag{5.9}$$

where μ is centered on atom A. Equation (5.9) looks a bit opaque at first glance, but it is actually quite straightforward. Remember that the full Fock matrix element $F_{\mu\mu}$ is the *sum* of the one-electron integral Eq. (5.9) and a series of two-electron integrals. If the number of valence electrons on each atom is exactly equal to the valence nuclear charge (i.e., every atom has a partial atomic charge of zero) then the repulsive two-electron terms will exactly cancel the attractive nuclear terms appearing at the end of Eq. (5.9) and we will recapture the expected result, namely that the energy associated with the diagonal matrix element is the ionization potential of the orbital. The Krönecker delta affecting the nuclear charge for atom A itself simply avoids correcting for a non-existent two-electron repulsion of an electron in basis function μ with itself. (Removing the attraction to nuclei *other* than A from the r.h.s. of Eq. (5.9) defines a commonly tabulated semiempirical parameter that is typically denoted U_μ.)

6. The only terms remaining to be defined in the assembly of the HF secular determinant are the one-electron terms for off-diagonal matrix elements. These are defined as

$$\left\langle \mu \left| -\frac{1}{2}\nabla^2 - \sum_k \frac{Z_k}{r_k} \right| \nu \right\rangle = \frac{(\beta_A + \beta_B)S_{\mu\nu}}{2} \tag{5.10}$$

where μ and ν are centered on atoms A and B, respectively, the β values are semiempirical parameters, and $S_{\mu\nu}$ is the overlap matrix element computed using the STO basis set. Note that computation of overlap is carried out for every combination of basis functions, even though in the secular determinant itself **S** is defined by Eq. (5.4). There are, in effect, two different **S** matrices, one for each purpose. The β parameters are entirely analogous to the parameter of the same name we saw in Hückel theory – they provide a measure of the strength of through space interactions between atoms. As they are intended for completely general use, it is not necessarily obvious how to assign them a numerical value, unlike the situation that obtains in Hückel theory. Instead, β values for CNDO were originally adjusted to reproduce certain experimental quantities.

While the CNDO method may appear to be moderately complex, it represents a vast simplification of HF theory. Equation (5.5) reduces the number of two-electron integrals having non-zero values from formally N^4 to simply N^2. Furthermore, those N^2 integrals are computed by trivial algebraic formulae, not by explicit integration, and between any pair of atoms all of the integrals have the same value irrespective of the atomic orbitals involved. Similarly, evaluation of one-electron integrals is also entirely avoided, with numerical values for those portions of the relevant matrix elements coming from easily evaluated formulae. Historically, a number of minor modifications to the conventions outlined above were explored, and the different methods had names like CNDO/1, CNDO/2, CNDO/BW, etc.; as these methods are all essentially obsolete, we will not itemize their differences. One CNDO model that does continue to see some use today is the Pariser–Parr–Pople (PPP) model for conjugated π systems (Pariser and Parr 1953; Pople 1953). It is in essence the CNDO equivalent of Hückel theory (only π-type orbitals are included in the secular equation), and improves on the latter theory in the prediction of electronic state energies.

The computational simplifications inherent in the CNDO method are not without chemical cost, as might be expected. Like EHT, CNDO is quite incapable of accurately predicting good molecular structures. Furthermore, the simplification inherent in Eq. (5.6) has some fairly dire consequences; two examples are illustrated in Figure 5.1. Consider the singlet and triplet states of methylene. Clearly, repulsion between the two highest energy electrons in each state should be quite different: they are spin-paired in an sp^2 orbital for the singlet, and spin-parallel, one in the sp^2 orbital and one in a p orbital, for the triplet. However, in each case the interelectronic Coulomb integral, by Eq. (5.6), is simply γ_{CC}. And, just as there is no distinguishing between different types of atomic orbitals, there is also no distinguishing between the orientation of those orbitals. If we consider the rotational coordinate for hydrazine, it is clear that one factor influencing the energetics will be the repulsion of the two lone pairs, one

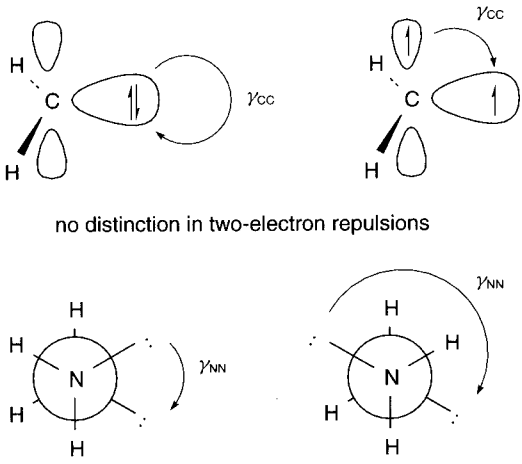


Figure 5.1 The CNDO formalism for estimating repulsive two-electron interactions fails to distinguish in one-center cases between different orbitals (top example for the case of methylene) and in two-center cases either between different orbitals or different orbital orientations (bottom example for the case of hydrazine)

on each nitrogen. However, we see from Eq. (5.8) that this repulsion, γ_{NN}, depends only on the distance separating the two nitrogen atoms, not on the orientation of the lone pair orbitals.

5.4 INDO Formalism

5.4.1 INDO and INDO/S

Of the two deficiencies specifically noted above for CNDO, the methylene problem is atomic in nature – it involves electronic interactions on a single center – while the hydrazine problem is molecular insofar as it involves two centers. Many ultraviolet/visible (UV/Vis) spectroscopic transitions in molecules are reasonably highly localized to a single center, e.g., transitions in mononuclear inorganic complexes. Pople, Beveridge, and Dobosh (1967) suggested modifications to the CNDO formalism to permit a more flexible handling of electron–electron interactions on the same center in order to model such spectroscopic transitions, and referred to this new formalism as 'intermediate neglect of differential overlap' (INDO). The key change is simply to use different values for the unique one-center two-electron integrals. When the atom is limited to a basis set of s and p orbitals, there are five such unique integrals

$$(ss|ss) = G_{ss}$$
$$(ss|pp) = G_{sp}$$
$$(pp|pp) = G_{pp}$$

$$(pp|p'p') = G_{pp'}$$
$$(sp|sp) = L_{sp} \tag{5.11}$$

The G and L values may be regarded as free parameters, but in practice they can be estimated from spectroscopic data. When the atomic valence orbitals include d and f functions, the number of unique integrals increases considerably, and the estimation of appropriate values from spectroscopy becomes considerably more complicated.

One effect of the greater flexibility inherent in the INDO scheme is that valence bond angles are predicted with much greater accuracy than is the case for CNDO. Nevertheless, overall molecular geometries predicted from INDO tend to be rather poor. However, if a good molecular geometry is available from some other source (ideally experiment) the INDO method has considerable potential for modeling the UV/Vis spectroscopy of the compound because of its better treatment of one-center electronic interactions.

Ridley and Zerner (1973) first described a careful parameterization of INDO specifically for spectroscopic problems, and designated that model INDO/S. Over the course of many years, Zerner and co-workers extended the model to most of the elements in the periodic table, including the lanthanides. Table 5.1 lists the energetic separations between various electronic states for three cases studied by INDO/S, ranging from the organic molecule pyridine, to the transition metal complex $Cr(H_2O)_6^{3+}$, to the metalloenzyme oxyhemocyanin which has a bimetallic Cu_2O_2 core ligated by enzyme histidine residues. Even just the ligated core of the latter system is daunting in size, but the simplifications intrinsic in semiempirical MO theory render it tractable. All of the geometries used were based on experiment, and excited state energies were computed using a CIS formalism (see Chapter 14 for details on CIS).

Table 5.1 Relative state energies (units of 1000 cm^{-1}) as computed by the INDO/S model

System (ground state)	State (transition)	INDO/S prediction	Experiment
Pyridine ($^1A_1$)[a]	$^1B_1$ ($n \rightarrow \pi^*$)	34.7	35.0
	$^1B_2$ ($\pi \rightarrow \pi^*$)	38.6	38.4
	$^1A_2$ ($n \rightarrow \pi^*$)	43.9	–
	$^1A_1$ ($\pi \rightarrow \pi^*$)	49.7	49.8
	$^1A_1$ ($\pi \rightarrow \pi^*$)	56.9	55.0
$Cr(H_2O)_6^{3+}$ ($^4A_{1g}$)[b]	$^4T_{2g}$ ($t \rightarrow e$)	12.4	12.4
	$^4T_{1g}$ ($t \rightarrow e$)	17.5	18.5
	$^2T_{1g}$ ($t \rightarrow t$)	13.2	13.1
	^2E_g ($t \rightarrow t$)	13.6	13.1
Oxyhemocyanin[c]	d→d	15.0	14.3–15.0
	$\pi \rightarrow$SOMO	17.8	17.5–18.1
	$\pi \rightarrow$SOMO	18.3	17.5–18.1
	$\pi^* \rightarrow \pi^*$	25.3	23.5–23.6
	$\pi \rightarrow$SOMO	36.3	29.4–30.4

[a]Ridley, J. E. and Zerner, M. C. 1973. *Theor. Chim. Acta*, **32**, 111.
[b]Anderson, W. P., Edwards, W. D., and Zerner, M. C. 1986. *Inorg. Chem.*, **25**, 2728.
[c]Estiú, G. L. and Zerner, M. C. 1999. *J. Am. Chem. Soc.* **121**, 1893

The INDO/S model is very successful for $d \rightarrow d$ transitions within transition metal complexes (typical accuracies are within 2000 cm^{-1}), potentially less robust for spectroscopic transitions that are not well localized to a single center (e.g., metal-to-ligand or ligand-to-metal excitations or excitations in extended π systems), and does not do very well in predicting Rydberg excitations since very diffuse orbitals are not a part of the basis set. The INDO/S model also exhibits good accuracy for the prediction of ionization potentials and oscillator strengths for weak electronic transitions (oscillator strengths for strong transitions tend to be overestimated).

It is perhaps appropriate at this point to make a distinction between a semiempirical 'method' and a semiempirical 'model'. The method describes the overall formalism for constructing the relevant secular equation. Within that formalism, however, different choices for free parameters may be made depending on the problem to which the method is being applied. Thus, INDO is a method, but the INDO/S model is a particular parameterization of the method designed for spectroscopic applications. One might say that the model is the set of all of the parameters required by the method. While we make this distinction here in a purely terminological sense, it has profound philosophical origins that should not be taken for granted as they led to the development of the first truly general semiempirical model, as described next.

5.4.2 MINDO/3 and SINDO1

Michael J. S. Dewar was trained as an organic chemist, but very early on he saw the potential for using MO theory, and in particular quantitative formulations of MO theory, to rationalize structure and reactivity in organic chemistry. Prior to Dewar's work, semiempirical models tended to be problem-specific. That is, while there were general methods like CNDO and INDO, most specific parameterizations (i.e., models) were carried out only within the context of narrowly defined chemical problems. Of course, if a particular chemical problem is important, there is certainly nothing wrong with developing a model specific to it, but the generality of the overall methods remained, for the most part, unexplored. Dewar established, as a goal, the development of a parameter set that would be as robust as possible across the widest possible spectrum (at the time, that spectrum was chosen to be primarily organic chemistry, together with a few inorganic compounds comprised of second- and third-row elements).

Dewar also recognized that a truly thorough test of a general model would require the efforts of more than one research group. To that end, he adopted a philosophy that not only should the best parameterization be widely promulgated but so too should computer code implementing it. Dewar's code included geometry optimization routines, which made it particularly attractive to non-developers interested in using the code for chemical purposes.

The first general parameterization to be reported by Dewar and co-workers was a third-generation modified INDO model (MINDO/3; Bingham, Dewar, and Lo, 1975). Some of the specific modifications to the INDO framework included the use of different ζ exponents in s and p type STOs on the same atom, the definition of pair parameters β_{AB} between two atoms A and B that were *not* averages of atomic parameters (actually, four such parameters

exist per pair, corresponding to $s_A s_B$, $s_A p_B$, $p_A s_B$, and $p_A p_B$ orbital interactions), adoption of a slightly different form for γ_{AB} than that of Eq. (5.8), and some empirical modifications to the nuclear repulsion energy.

Moreover, rather than following any particular set of rules to generate parameter values (e.g., Slater's rules for orbital exponents), every parameter was treated as a free variable subject to 'chemical common-sense' restraints. That is, parameter values were allowed to vary freely so long as they did not clearly become physically unrealistic; thus, for instance, a parameter set where an atomic U_s value was smaller in magnitude than the corresponding U_p value would not be acceptable, since ionization of a valence s electron on an atom cannot be more favorable than ionization of a valence p electron. To optimize parameter values, Dewar initially took a set of 138 small molecules containing C, H, O, and N, and constructed a penalty function depending on bond distances, valence angles, torsional angles, dipole moments, ionization potentials, and heats of formation (most molecules had experimental data available only for a subset of the penalty function components).

The use of the last experimental observable, the heat of formation, merits a brief digression. Molecular heats of formation are the most widely available thermochemical quantities against which one might imagine carrying out a parameterization (this was even more the case in the early 1970s), but these are enthalpic quantities, not potential energies. The expectation value from a MO calculation is the potential energy to separate all of the electrons and nuclei to infinite separation. How can these two be compared? Dewar's approach was to compute or estimate the MINDO/3 SCF energy for the elements in their standard states (on a per atom basis for those elements whose standard states are not monatomic) and to record this as an atomic parameter. To compute the heat of formation for a molecule, the individual standard-state atomic SCF energies were subtracted from the molecular SCF energy, and the difference was summed with the experimental heats of formation for all the constituent atoms (this is in effect rather similar to the atomic-type-equivalents scheme discussed for force-field calculations in Chapter 2, except that in the semiempirical approach, there is only a single type for each atom). Most computer codes implementing semiempirical models continue to print out the energy in this fashion, as a so-called heat of formation.

Note, however, that zero-point vibrational energy and thermal contributions to the experimental enthalpies have been ignored (or, perhaps more accurately, treated in some sort of average fashion by the parameterization process). At the time, computing those contributions was far from trivial. Now, however, it is quite straightforward to do so (see Chapter 10 for details), leading to some ambiguity in how energies from semiempirical calculations should be reported. As a general rule, it would probably be better to consider the semiempirical SCF energies to have the status of potential energies, and explicitly to account for thermochemical contributions when necessary, but the literature is full of examples where this issue is confused. Of course, if the goal of the calculation is truly to estimate the heat of formation of a particular molecule, then the semiempirical SCF energy should be used uncorrected, since that enthalpic quantity was one of the targets for which parameters were optimized.

The performance of the MINDO/3 model was impressive overall. The mean absolute error in predicted heats of formation was 11 kcal/mol (all molecules), the corresponding

error for ionization potentials was 0.7 eV (46 molecules), for heavy-atom bond lengths 0.022 Å (81 molecules), and for dipole moments 0.45 D (31 molecules). While mean errors of this size exceed what would be tolerated today, they were unprecedentedly small in 1975. Dewar's subsequent work on other semiempirical models (see below) rendered MINDO/3 effectively obsolete, but its historical importance remains unchanged.

A modified INDO model that is *not* entirely obsolete is the symmetric orthogonalized INDO (SINDO1) model of Jug and co-workers, first described in 1980 (Nanda and Jug 1980). The various conventions employed by SINDO1 represent slightly different modifications to INDO theory than those adopted in the MINDO/3 model, but the more fundamental difference is the inclusion of d functions for atoms of the second row in the periodic table. Inclusion of such functions in the atomic valence basis set proves critical for handling hypervalent molecules containing these atoms, and thus SINDO1 performs considerably better for phosphorus-containing compounds, for instance, than do other semiempirical models that lack d functions (Jug and Schulz Chem. 1988).

5.5 Basic NDDO Formalism

The INDO model extends the CNDO model by adding flexibility to the description of the one-center two-electron integrals. In INDO, however, there continues to be only a single two-center two-electron integral, which takes on the value γ_{AB} irrespective of which orbitals on atoms A and B are considered. As already noted, this can play havoc with the accurate representation of lone pair interactions.

The neglect of diatomic differential overlap (NDDO) method relaxes the constraints on two-center two-electron integrals in a fashion analogous to that for one-center integrals in the INDO method. Thus, all integrals $(\mu\nu|\lambda\sigma)$ are retained provided μ and ν are on the same atomic center and λ and σ are on the same atomic center, but not necessarily the center hosting μ and ν. How many different integrals are permitted? The order of μ and ν does not affect the value of the integral, so we need only worry about combinations, not permutations, in which case there are 10 unique combinations of s, p_x, p_y, and p_z. With 10 unique combinations on each atom, there are 100 possible combinations of combinations for the integrals. If we include d functions, the number of unique integrals increases to 2025.

Although these numbers seem large, this is still a considerable improvement over evaluating *every* possible integral, as would be undertaken in *ab initio* HF theory. Most modern semiempirical models are NDDO models. After examining the differences in their formulation, we will examine their performance characteristics in some detail in Section 5.6.

5.5.1 MNDO

Dewar and Thiel (1977) reported a modified neglect of differential overlap (MNDO) method based on the NDDO formalism for the elements C, H, O, and N. With the conventions specified by NDDO for which integrals to keep, which to discard, and how to model one-electron integrals, it is possible to write the NDDO Fock matrix elements individually for

inspection. For the most complex element, a diagonal element, we have

$$F_{\mu\mu} = U_\mu - \sum_{B\neq A} Z_B(\mu\mu|s_B s_B) + \sum_{\nu\in A} P_{\nu\nu}\left[(\mu\mu|\nu\nu) - \frac{1}{2}(\mu\nu|\mu\nu)\right] + \sum_B \sum_{\lambda\in B}\sum_{\sigma\in B} P_{\lambda\sigma}(\mu\mu|\lambda\sigma) \tag{5.12}$$

where μ is located on atom A. The first term on the r.h.s. is the atomic orbital ionization potential, the second term the attraction to the other nuclei where each nuclear term is proportional to the repulsion with the valence s electron on that nucleus, the third term reflects the Coulomb and exchange interactions with the other electrons on atom A, and the final term reflects Coulomb repulsion with electrons on other atoms B.

An off-diagonal Fock matrix element for two basis functions μ and ν on the same atom A is written as

$$F_{\mu\nu} = -\sum_{B\neq A} Z_B(\mu\nu|s_B s_B) + P_{\mu\nu}\left[\frac{3}{2}(\mu\nu|\mu\nu) - \frac{1}{2}(\mu\mu|\nu\nu)\right] + \sum_B \sum_{\lambda\in B}\sum_{\sigma\in B} P_{\lambda\sigma}(\mu\nu|\lambda\sigma) \tag{5.13}$$

where each term on the r.h.s. has its analogy in Eq. (5.12). When μ is on atom A and ν on atom B, this matrix element is written instead as

$$F_{\mu\nu} = \frac{1}{2}(\beta_\mu + \beta_\nu)S_{\mu\nu} - \frac{1}{2}\sum_{\lambda\in A}\sum_{\sigma\in B} P_{\lambda\sigma}(\mu\lambda|\nu\sigma) \tag{5.14}$$

where the first term on the r.h.s. is the resonance integral that encompasses the one-electron kinetic energy and nuclear attraction terms; it is an average of atomic resonance integrals 'β' times the overlap of the orbitals involved. The second term on the r.h.s. captures favorable exchange interactions. Note that the MNDO model did *not* follow Dewar's MINDO/3 approach of having β parameters specific to pairs of atoms. While the latter approach allowed for some improved accuracy, it made it quite difficult to add new elements, since to be complete all possible pairwise β combinations with already existing elements would require parameterization.

The only point not addressed in Eqs. (5.12) to (5.14) is how to go about evaluating all of the necessary two-electron integrals. Unlike one-center two-electron integrals, it is not easy to analyze spectroscopic data to determine universal values, particularly given the large number of integrals not taken to be zero. The approach taken by Dewar and co-workers was to evaluate these integrals by replacing the continuous charge clouds with classical multipoles. Thus, an ss product was replaced with a point charge, an sp product was replaced with a classical dipole (represented by two point charges slightly displaced from the nucleus along the p orbital axis), and a pp product was replaced with a classical quadrupole (again represented by point charges). The magnitudes of the moments, being one-center in nature, are related to the parameterized integrals in Eq. (5.11). By adopting such a form for the integrals, their evaluation is made quite simple, and so too is evaluation of their analytic derivatives with respect to nuclear motion.

To complete the energy evaluation by the MNDO method, the nuclear repulsion energy is added to the SCF energy. The MNDO nuclear repulsion energy is computed as

$$V_N = \sum_{k<l}^{\text{nuclei}} Z_k Z_l (s_k s_k | s_l s_l) \left(1 + \frac{1}{\tau} e^{-\alpha_{Z_k} r_{kl}} + e^{-\alpha_{Z_l} r_{kl}}\right) \tag{5.15}$$

where Z is the valence atomic number, α is a parameter having a specific value for each atom type, r is the interatomic distance, and τ is equal to 1 unless the two nuclei k and l are an O/H or N/H pair, in which case it is rXH. Thus, internuclear repulsion is proportional to the repulsion between s electrons on the same centers, and the repulsion is empirically increased slightly at short bond lengths to make up for imbalances in the electronic part of the calculation.

As with MINDO/3, Dewar and Thiel optimized the parameters of the MNDO model against a large test set of molecular properties. Within the assumption of a valence orbital set comprised only of s and p orbitals, MNDO parameters are now available for H, He, Li, Be, B, C, N, O, F, Al, Si, P, S, Cl, Zn, Ge, Br, Sn, I, Hg, and Pb. The MNDO model is typically not used as often as the NDDO models discussed next, but MNDO calculations still appear in the literature. MNDO forms the foundation for MNDO/d, which is discussed in Section 5.7.2. In addition, a modified MNDO model explicitly adding electron correlation effects (MNDOC) by second-order perturbation theory (see Section 7.4) was described by Thiel in 1981 (Thiel 1981; Schweig and Thiel 1981). By explicitly accounting for electron correlation in the theory, the parameters do not have to absorb the effects of its absence from HF theory in some sort of average way. Thus, in principle, MNDOC should be more robust in application to problems with widely varying degrees of electron correlation. In practice, the model has not yet been compared to other NDDO models to the degree necessary to evaluate whether the formalism lives up to that potential.

5.5.2 AM1

Although a detailed discussion of the performance of MNDO is deferred until Section 5.6, one critical flaw in the method is that is does very poorly in the prediction of hydrogen bonding geometries and energies. Recognizing this to be a major drawback, particularly with respect to modeling systems of biological interest, Dewar and co-workers modified the functional form of their NDDO model; since the primary error was one involving bond lengths, the key modification was to the nuclear repulsion term. In Austin Model 1 (AM1; Dewar, at the time, was a faculty member at the University of Texas, Austin), originally described in 1985 for the elements C, H, O, and N (Dewar *et al.* 1985), the nuclear repulsion energy between any two nuclei A and B is computed as

$$V_N(A, B) = Z_A Z_B (s_A s_A | s_B s_B) + \frac{Z_A Z_B}{r_{AB}} \sum_{i=1}^{4} [a_{A,i} e^{-b_{A,i}(r_{AB} - c_{A,i})^2} + a_{B,i} e^{-b_{B,i}(r_{AB} - c_{B,i})^2}] \tag{5.16}$$

where the variables are for the most part those in Eq. (5.15) and in addition every atom has up to 4 each parameters a, b, and c describing Gaussian functions centered at various distances

c that modify the potential of mean force between the two atoms. Simultaneous optimization of the original MNDO parameters with the Gaussian parameters led to markedly improved performance, although the Gaussian form of Eq. (5.16) is sufficiently force-field-like in nature that one may quibble about this method being entirely quantum mechanical in nature.

Since the report for the initial four elements, AM1 parameterizations for B, F, Al, Si, P, S, Cl, Zn, Ge, Br, I, and Hg have been reported. Because AM1 calculations are so fast (for a quantum mechanical model), and because the model is reasonably robust over a large range of chemical functionality, AM1 is included in many molecular modeling packages, and results of AM1 calculations continue to be reported in the chemical literature for a wide variety of applications.

5.5.3 PM3

One of the authors on the original AM1 paper and a major code developer in that effort, James J. P. Stewart, subsequently left Dewar's labs to work as an independent researcher. Stewart felt that the development of AM1 had been potentially non-optimal, from a statistical point of view, because (i) the optimization of parameters had been accomplished in a stepwise fashion (thereby potentially accumulating errors), (ii) the search of parameter space had been less exhaustive than might be desired (in part because of limited computational resources at the time), and (iii) human intervention based on the perceived 'reasonableness' of parameters had occurred in many instances. Stewart had a somewhat more mathematical philosophy, and felt that a sophisticated search of parameter space using complex optimization algorithms might be more successful in producing a best possible parameter set within the Dewar-specific NDDO framework.

To that end, Stewart set out to optimize *simultaneously* parameters for H, C, N, O, F, Al, Si, P, S, Cl, Br, and I. He adopted an NDDO functional form identical to that of AM1, except that he limited himself to two Gaussian functions per atom instead of the four in Eq. (5.16). Because his optimization algorithms permitted an efficient search of parameter space, he was able to employ a significantly larger data set in evaluating his penalty function than had been true for previous efforts. He reported his results in 1989; as he considered his parameter set to be the third of its ilk (the first two being MNDO and AM1), he named it Parameterized Model 3 (PM3; Stewart 1989).

There is a possibility that the PM3 parameter set may actually be the global minimum in parameter space for the Dewar-NDDO functional form. However, it must be kept in mind that even if it *is* the global minimum, it is a minimum for a particular penalty function, which is itself influenced by the choice of molecules in the data set, and the human weighting of the errors in the various observables included therein (see Section 2.2.7). Thus, PM3 will not necessarily outperform MNDO or AM1 for any particular problem or set of problems, although it is likely to be optimal for systems closely resembling molecules found in the training set. As noted in the next section, some features of the PM3 parameter set can lead to very unphysical behaviors that were not assessed by the penalty function, and thus were not avoided. Nevertheless, it is a very robust NDDO model, and continues to be used at least as widely as AM1.

In addition to the twelve elements noted above, PM3 parameters for Li, Be, Mg, Zn, Ga, Ge, As, Se, Cd, In, Sn, Sb, Te, Hg, Tl, Pb, Bi, Po, and At have been reported. The PM3 methodology is available in essentially all molecular modeling packages that carry out semiempirical calculations.

5.6 General Performance Overview of Basic NDDO Models

Many comparisons of the most widely used semiempirical models have been reported. They range from narrowly focused anecdotal discussions for specific molecules to detailed tests over large sets of molecules for performance in the calculation of various properties. We will discuss here a subset of these comparisons that have the broadest impact – those looking for a more thorough overview are referred to the bibliography at the end of the chapter.

5.6.1 Energetics

The primary energetic observable against which NDDO models were parameterized was heat of formation. Table 5.2 compares the mean unsigned errors for MNDO, AM1, and PM3 for various classes of molecules (the column labeled MNDO/d is discussed in Section 5.7.2). The greater accuracies of AM1 and PM3 compared to MNDO are manifest in every case. PM3 appears to offer a slight advantage over AM1 for estimating the heats of formation of molecules composed of lighter elements (C, H, O, N, F), and a clear advantage for

Table 5.2 Mean unsigned errors (kcal mol^{-1}) in predicted heats of formation from basic NDDO models

Elements (number)	Subset (number)	MNDO	AM1	PM3	MNDO/d
Lighter (181)		7.35	5.80	4.71	
	CH (58)	5.81	4.89	3.79	
	CHN (32)	6.24	4.65	5.02	
	CHNO (48)	7.12	6.79	4.04	
	CHNOF (43)	10.50	6.76	6.45	
	Radicals (14)	9.3	8.0	7.4	
Heavier (488)		29.2	15.3	10.0	4.9
	Al (29)	22.1	10.4	16.4	4.9
	Si (84)	12.0	8.5	6.0	6.3
	P (43)	38.7	14.5	17.1	7.6
	S (99)	48.4	10.3	7.5	5.6
	Cl (85)	39.4	29.1	10.4	3.9
	Br (51)	16.2	15.2	8.1	3.4
	I (42)	25.4	21.7	13.4	4.0
	Hg (37)	13.7	9.0	7.7	2.2
	Normal (421)	11.0	8.0	8.4	4.8
	Hypervalent (67)	143.2	61.3	19.9	5.4
Cations (34)		9.55	7.62	9.46	
Anions (13)		11.36	7.11	8.81	

heavier elements. However, in the latter case, the difference is essentially entirely within the subset of hypervalent molecules included in the test set, e.g., PBr_5, IF_7, etc. Over the 'normal' subset of molecules containing heavy atoms, the performance of AM1 and PM3 is essentially equivalent. Analysis of the errors in predicted heats of formation suggests that they are essentially random, i.e., they reflect the 'noise' introduced into the Schrödinger equation by the NDDO approximations and cannot be corrected for in a systematic fashion without changing the theory. This random noise can be problematic when the goal is to determine the relative energy differences between two or more isomers (conformational or otherwise), since one cannot be as confident that errors will cancel as is the case for more complete quantum mechanical methods.

Errors for charged and open-shell species tend to be somewhat higher than the corresponding errors for closed-shell neutrals. This may be at least in part due to the greater difficulty in measuring accurate experimental data for some of these species, but some problems with the theory are equally likely. For instance, the more loosely held electrons of an anion are constrained to occupy the same STO basis functions as those used for uncharged species, so anions are generally predicted to be anomalously high in energy. Radicals are systematically predicted to be too stable (the mean signed error over the radical test set in Table 5.2 is only very slightly smaller than the mean unsigned error) meaning that bond dissociation energies are usually predicted to be too low. Note that for the prediction of radicals all NDDO methods were originally parameterized with a so-called 'half-electron RHF method', where the formalism of the closed-shell HF equations is used even though the molecule is open-shell (Dewar, Hashmall, and Venier 1968). Thus, while use of so-called 'unrestricted Hartree–Fock (UHF)' technology (see Section 6.3.3) is technically permitted for radicals in semiempirical theory, it tends to lead to unrealistically low energies and is thus less generally useful for thermochemical prediction (Pachkovski and Thiel 1996).

Another energetic quantity of some interest is the ionization potential (IP). Recall that in HF theory, the eigenvalue associated with each MO is the energy of an electron in that MO. Thus, a good estimate of the negative of the IP is the energy of the highest occupied MO – this simple approximation is one result from a more general statement known as Koopmans' theorem (Koopmans, 1933). Employing this approximation, all of the semiempirical methods do reasonably well in predicting IPs for organic molecules. On a test set of 207 molecules containing H, C, N, O, F, Al, S, P, Cl, Br, and I, the average error in predicted IP for MNDO, AM1, and PM3 is 0.7, 0.6, and 0.5 eV, respectively. For purely inorganic compounds, PM3 shows essentially unchanged performance, while MNDO and AM1 have errors increased by a few tenths of an electron volt.

With respect to the energetics associated with conformational changes and reactions, a few general comments can be made. MNDO has some well-known shortcomings; steric crowding tends to be too strongly disfavored and small ring compounds are predicted to be too stable. The former problem leads to unrealistically high heats of formation for sterically congested molecules (e.g., neopentane) and similarly too high heats of activation for reactions characterized by crowded TS structures. For the most part, these problems are corrected in AM1 and PM3 through use of Eq. (5.16) to modify the non-bonded interactions. Nevertheless, activation enthalpies are still more likely to be too high than too low for the semiempirical

methods because electron correlation energy tends to be more important in TS structures than in minima (see also Table 8.3), and since correlation energy is introduced in only an average way by parameterization of the semiempirical HF equations, it cannot distinguish well between the two kinds of structures.

For intermolecular interactions that are weak in nature, e.g., those arising from London forces (dispersion) or hydrogen bonding, semiempirical methods are in general unreliable. Dispersion is an electron correlation phenomenon, so it is not surprising that HF-based semiempirical models fail to make accurate predictions. As for hydrogen bonding, one of the primary motivations for moving from MNDO to AM1 was to correct for the very weak hydrogen bond interactions predicted by the former. Much of the focus in the parameterization efforts of AM1 and PM3 was on reproducing the enthalpy of interaction of the water dimer, and both methods do better in matching the experimental value of 3.6 kcal mol^{-1} than does MNDO. However, detailed analyses of hydrogen bonding in many different systems have indicated that in most instances the interaction energies are systematically too small by up to 50 percent and that the basic NDDO methods are generally not well suited to the characterization of hydrogen bonded systems (Dannenberg 1997). Bernal-Uruchurtu *et al.* (2000) have suggested that the form of Eq. 5.16 is inadequate for describing hydrogen bonding; by use of an alternative parameterized interaction function, they were able to modify PM3 so that the PES for the water dimer was significantly improved.

Energetic barriers to rotation about bonds having partial double bond character tend to be significantly too low at semiempirical levels. In amides, for instance, the rotation barrier about the C–N bond is underestimated by about 15 kcal/mol. In several computer programs implementing NDDO methods, an *ad hoc* molecular mechanics torsional potential can be added to amide bond linkages to correct for this error. Smaller errors, albeit still large as a fraction of total barrier height, are observed about C–C single bonds in conjugated chains.

With respect to conformational analysis, the NDDO models are not quantitatively very accurate. Hehre has reported calculations for eight different sets of conformer pairs having an average energy difference between pairs of 2.3 kcal mol^{-1}. Predictions from MNDO, AM1, and PM3 gave mean unsigned errors of 1.4, 1.3, and 1.8 kcal mol^{-1}, respectively, although in four of the eight cases AM1 was within 0.5 kcal mol^{-1}. In addition, AM1 and PM3 have been compared for the 11 D-glucopyranose conformers discussed in Chapter 2 in the context of analyzing force field performance; AM1 and PM3 had mean unsigned errors of 1.4 and 0.8 kcal mol^{-1}, respectively, making them less accurate than the better force fields. The PM3 number is misleadingly good in this instance; although the method does reasonably well for the 11 conformers studied, the PM3 PES also includes highly unusual minimum energy structures not predicted by any other method (vide infra).

5.6.2 Geometries

Correct molecular structures are dependent on the proper location of wells in the PES, so they are intimately related to the energetics of conformational analysis. For organic molecules, most gross structural details are modeled with a reasonable degree of accuracy. Dewar, Jie, and Yu (1993) evaluated AM1 and PM3 for 344 bond lengths and 146 valence angles

in primarily organic molecules composed of H, C, N, O, F, Cl, Br, and I; the average unsigned errors were 0.027 and 0.022 Å, respectively, for the bond lengths, and 2.3 and 2.8°, respectively, for the angles. In the parameterization of PM3, Stewart (1991) performed a similar analysis for a larger set of molecules, some of them including Al, Si, P, and S. For 460 bond lengths, the mean unsigned errors were 0.054, 0.050, and 0.036 Å for MNDO, AM1, and PM3, respectively. For 196 valence angles, the errors were 4.3, 3.3, and 3.9°, respectively. In the case of MNDO, bond angles at the central O and S atoms in ethers and sulfides, respectively, were found to be up to 9° too large, presumably owing to the overestimation of steric repulsion between the substituting groups.

Comparing all of the sets of comparisons, it is evident that the geometries for the molecules containing second-row elements are considerably more difficult to predict accurately than are those for simpler organics. Furthermore, MNDO and AM1 are less successful when extended to these species than is PM3.

Stewart also carried out an analysis for dihedral angles, and found errors of 21.6, 12.5, and 14.9°, respectively, for MNDO, AM1, and PM3. However, only 16 data points were available and the accurate measurement of dihedral angles is challenging. Nevertheless, there appear to be systematic errors in dihedral angles for small- to medium-sized ring systems, where predicted geometries tend to be too 'flat', again probably because of overestimated steric repulsion between non-bonded ring positions (Ferguson *et al.* 1992). Four-membered rings are typically predicted to be planar instead of puckered.

A few additional geometric pathologies have been discovered over the years for the various semiempirical methods. While many are for species that might be described as exotic, others have considerably more potential to be troublesome.

Heteroatom–heteroatom linkages are often problematic. MNDO and AM1 both predict peroxide O–O bonds to be 0.018 Å too short. In hydrazines, the N–N bond rotamer placing the two nitrogen lone pairs antiperiplanar to one another is usually overstabilized relative to the *gauche* rotamer. Thus, even though experimentally hydrazine has been determined to have a C_2 *gauche* structure, all of the methods predict the global minimum to be the trans C_{2v} structure (PM3 does not find the C_2 structure to be a stationary point at all). In nitroxyl compounds, N–N bonds are predicted to be too short by up to 0.7 Å. AM1 has similar problems with P–P bond lengths. In silyl halides, Si–X bonds are predicted to be too short by tenths of an ångström by PM3.

PM3 shows additional problems that are disturbing. Nitrogen atoms formally possessing a lone pair tend to be significantly biased towards pyramidal geometries. In addition, there is an anomalous, deep well in the non-bonded H–H interaction expressed by Eq. (5.16) at a distance of about 1.4 Å (Csonka 1993), which can lead to such odd situations as hydroxyl groups preferring to interact with one another by H–H contacts instead of typical hydrogen bonding contacts in D-glucopyranose conformers (Barrows *et al.* (1995)).

As already noted above, the energetics of normal hydrogen bonding is not handled well by any semiempirical method; geometries are similarly problematic. PM3 predicts the expected near-linear single hydrogen bond for most systems, but typically it is too short by as much as 0.2 Å. AM1 predicts heavy-atom–heavy-atom bond distances in hydrogen bonds that are about right, but strongly favors bifurcated hydrogen bonds in those systems where that

is possible (e.g., in the water dimer, the water molecule acting as a hydrogen bond donor interacts with the other water molecule through *both* its protons equally). MNDO hydrogen bonds are much, much too long, since the interaction energies at this level are predicted to be far too small.

5.6.3 Charge Distributions

One of the most useful features of a QM model is its ability to provide information about the molecular charge distribution. It is a general rule of thumb that even very low quality QM methods tend to give reasonable charge distributions. For neutral molecules, the dominant moment in the overall charge distribution is the usually dipole moment (unless symmetry renders the dipole moment zero). For a 125-molecule test set including H, C, N, O, F, Al, Si, P, S, Cl, Br, and I functionality, Stewart found mean unsigned errors in dipole moments of 0.45, 0.35, and 0.38 D, respectively, for MNDO, AM1, and PM3 (Stewart 1989). PM3 seems to be somewhat more robust for compounds incorporating phosphorus.

An alternative measure of the charge distribution involves a partitioning into partial atomic charges. While such partitioning is always arbitrary (see Chapter 9) simple methods tend to give reasonably intuitive results when small basis sets are used, as is the case for the NDDO models. While MNDO and AM1 present no particular issues for such analysis, PM3 tends to predict nitrogen atoms to be too weakly electronegative. Thus, in the ammonium cation, PM3 predicts the charge on nitrogen to be +1.0 while the charge on each hydrogen is predicted to be 0.0 (Storer *et al.* (1995)).

Finally, some attention has been paid to the quality of the complete electrostatic potential about the molecule at the NDDO level. This topic is discussed in Chapter 9, as are additional details associated with the performance of semiempirical models in comparison to other levels of electronic structure theory for a variety of more specialized properties.

5.7 Ongoing Developments in Semiempirical MO Theory

Semiempirical theory is still in widespread use today not because it competes effectively with more sophisticated theories in terms of accuracy, but because it competes effectively in terms of computational resources. Indeed, if one has either an enormously large molecule, or an enormously large number of small molecules to be compared at a consistent level (the next section describes a particular example of this case), semiempirical theory is the only practical option. Of course, with each improvement in technology, the size horizon of the more sophisticated levels expands, but there seems little danger that chemists will not always be able to imagine still larger systems meriting quantum chemical study. Therefore, considerable interest remains in improving semiempirical models in a variety of directions. We close this chapter with a brief overview of some of the most promising of these.

5.7.1 Use of Semiempirical Properties in SAR

This area is a development in the *usage* of NDDO models that emphasizes their utility for large-scale problems. Structure–activity relationships (SARs) are widely used in the pharmaceutical industry to understand how the various features of biologically active molecules

contribute to their activity. SARs typically take the form of equations, often linear equations, that quantify activity as a function of variables associated with the molecules. The molecular variables could include, for instance, molecular weight, dipole moment, hydrophobic surface area, octanol–water partition coefficient, vapor pressure, various descriptors associated with molecular geometry, etc. Once a SAR is developed, it can be used to prioritize further research efforts by focusing first on molecules predicted by the SAR to have highest activity.

Thus, if a drug company has a database of several hundred thousand molecules that it has synthesized over the years, and it has measured molecular properties for those compounds, once it identifies a SAR for some particular bio-target, it can quickly run its database through the SAR to identify other molecules that should be examined. However, this process is not very useful for identifying *new* molecules that might be better than any presently existing ones. It can be quite expensive to synthesize new molecules randomly, so how can that process be similarly prioritized?

One particularly efficient alternative is to develop SARs not with experimental molecular properties, but with predicted ones. Thus, if the drug company database is augmented with predicted values, and a SAR on predicted values proves useful based on data for compounds already assayed, potential new compounds can be examined in a purely computational fashion to evaluate whether they should be priority targets for synthesis. In 1998, Beck *et al.* (1998) optimized the geometries of a database of 53 000 compounds with AM1 in 14 hours on a 128-processor Origin 2000 computer. Such speed is presently possible only for semiempirical levels of theory. Once the geometries and wave functions are in hand, it is straightforward (and typically much faster) to compute a very wide variety of molecular properties in order to survey possible SARs. Note that for the SAR to be useful, the absolute values of the computed properties do not necessarily need to be accurate – only their variation relative to their activity is important.

5.7.2 d Orbitals in NDDO Models

To extend NDDO methods to elements having occupied valence d orbitals that participate in bonding, it is patently obvious that such orbitals need to be included in the formalism. However, to accurately model even non-metals from the third row and lower, particularly in hypervalent situations, d orbitals are tremendously helpful to the extent they increase the flexibility with which the wave function may be described. As already mentioned above, the d orbitals present in the SINDO1 and INDO/S models make them extremely useful for spectroscopy. However, other approximations inherent in the INDO formalism make these models poor choices for geometry optimization, for instance. As a result, much effort over the last decade has gone into extending the NDDO formalism to include d orbitals.

Thiel and Voityuk (1992, 1996) described the first NDDO model with d orbitals included, called MNDO/d. For H, He, and the first-row atoms, the original MNDO parameters are kept unchanged. For second-row and heavier elements, d orbitals are included as a part of the basis set. Examination of Eqs. (5.12) to (5.14) indicates what is required parametrically to add d orbitals. In particular, one needs U_d and β_d parameters for the one-electron integrals, additional one-center two-electron integrals analogous to those in Eq. (5.11) (there are

formally 12 such integrals), and a prescription for handling two-center two-electron integrals including d functions. In MNDO/d, the U, β, and G_{dd} terms are treated as adjustable parameters, the remaining one-center two-electron integrals are analytic functions of G_{dd} and the integrals in Eq. (5.11), and the Dewar convention whereby two-center two-electron integrals are evaluated using classical multipole expansions is retained, except that multipolar representations beyond quadrupole (e.g., a dd cloud would be a hexadecapole) are ignored, since testing indicates they typically contribute negligibly to the total electronic energy. Parameterization of the various new terms proceeds in the same fashion as for prior NDDO models, with a penalty function focused on molecular thermochemical and structural data. The performance of the model for heavy elements is summarized in Table 5.2 (for light elements MNDO/d is identical to MNDO). MNDO/d represents an enormous improvement over AM1 and PM3 in its ability to handle hypervalent molecules, and in most cases the error over the various test sets is reduced by half or more when MNDO/d is used.

It appears, then, that MNDO/d has high utility for thermochemical applications. In addition to the elements specified in Table 5.2, MNDO/d parameters have been determined for Na, Mg, Zn, Zr, and Cd. However, since the model is based on MNDO and indeed identical to MNDO for light elements, it still performs rather poorly with respect to intermolecular interactions, and with respect to hydrogen bonding in particular.

The approach of Thiel and Voityuk has also been adopted by Hehre and co-workers, who have applied it in extending the PM3 Hamiltonian to include d orbitals. This model, which to date has not been fully described in the literature and is only available as part of a commercial software package (SPARTAN), is called PM3(tm), where the 'tm' emphasizes a focus on transition metals. The parameterization philosophy has been different from prior efforts insofar as only geometrical data (primarily from X-ray crystallography) have been included in the penalty function. This choice was motivated at least in part by the general scarcity of thermochemical data for molecules containing transition metals. Thus, the model may be regarded as an efficient way to generate reasonable molecular geometries whose energies may then be evaluated using more complete levels of theory. For example, a study by Goh and Marynick (2001) found that the geometries of metallofullerenes predicted at the PM3(tm) level compared well with those predicted from much more expensive density functional calculations.

A semiempirical model including d orbitals has also been reported by Dewar and coworkers (Dewar, Jie, and Yu 1993; Holder, Dennington, and Jie 1994), although the full details of its functional form still await publication. Semi-*ab initio* model 1 (SAM1, or SAM1D if d orbitals are included), however, is not quite so straightforward an extension of the NDDO formalism, but represents a rather different approach to constructing the Fock matrix. In SAM1, the valence-orbital basis set is made up not of Slater-type orbitals but instead of Gaussian-type orbitals; in particular the STO-3G basis set is used (see Section 6.2.2). Using this basis set, one- and two-electron integrals not explicitly set to zero in the NDDO formalism are analytically calculated in an *ab initio* fashion (see Section 6.1), but the resulting values are then treated as input to parameterized scaling functions depending on, *inter alia*, interatomic distance. Parameters exist for H, Li, C, N, O, F, Si, P, S, Cl, Fe, Cu, Br, and I. For molecules made up of light elements, SAM1 performs better than AM1

and very slightly better than PM3. The same is true for non-hypervalent molecules made up of heavier elements, while very large improvements are observed for molecules containing hypervalent heavy atoms – across 404 compounds containing Si, P, S, Cl, Br, and/or I as heavy elements, the mean unsigned errors in heats of formation for AM1, PM3, SAM1, and MNDO/d are 16.2, 9.5, 9.3, and 5.1 kcal/mol (Thiel and Voityuk 1996).

5.7.3 SRP Models

SRP, a term first coined by Rossi and Truhlar (1995), stands for 'specific reaction (or range) parameters'. An SRP model is one where the standard parameters of a semiempirical model are adjusted so as to foster better performance on a particular problem or class of problems. In a sense, the SRP concept represents completion of a full circle in the philosophy of semiempirical modeling. It tacitly recognizes the generally robust character of some underlying semiempirical model, and proceeds from there to optimize that model for a particular system of interest. In application, then, SRP models are similar to the very first semiempirical models, which also tended to be developed on an *ad hoc*, problem-specific basis. The difference, however, is that the early models typically were developed essentially from scratch, while SRP models may be viewed as perturbations of more general models.

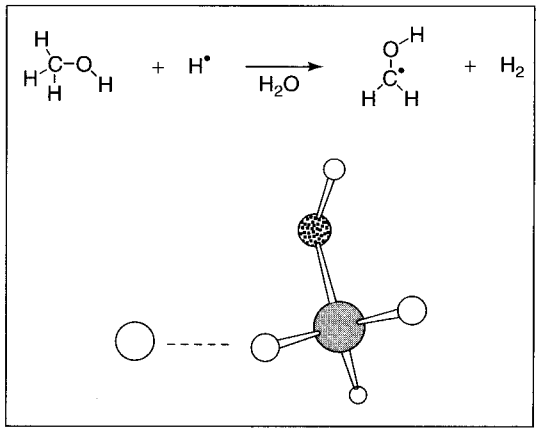
The concept is best illustrated with an example. Chuang *et al.* (1999) used an AM1-SRP model to study the hydrogen-atom-mediated destruction of organic alcohols in water. As illustrated in Figure 5.2, the AM1 model itself makes a very poor prediction for the dissociation energy of the C–H bond in methanol, and hence for the reaction exothermicity. By minor adjustment of a few of the AM1 parameters, however, the SRP model gives good agreement with experiment. The resulting SRP model in this case was used as a very efficient QM method for generation of a PES from which tunneling contributions to the reaction rate constant could be estimated (see Section 15.3). The very large number of QM calculations required to generate the PES made use of an SRP model preferable to more complete levels of electronic structure theory like those discussed in Chapter 7.

5.7.4 Linear Scaling

As already touched upon in Section 2.4.2, the development of methods that scale linearly with respect to system size opens the door to the modeling of very large systems with maximal computational efficiency. Because the NDDO approximation is already rather efficient when it comes to forming the Fock matrix (because so many integrals are assumed to be zero, etc.), it serves as an excellent basis on which to build a linear-scaling QM model. Such models have been reported; the details associated with achieving linear scaling are sufficiently technical that interested readers are referred to the original literature (van der Vaart *et al.* 2000; Khandogin, Hu, and York 2000; see also Stewart 1996).

It is worth a pause, however, to consider how such models should best be used. Part of the motivation for developing linear scaling models has been to permit QM calculations to be carried out on biomolecules, e.g., proteins or polynucleic acids. However, one may legitimately ask whether there is any point in such a calculation, beyond demonstrating that

Parameter	AM1	AM1-SRP
C		
U_s	−52.03	−49.85
U_p	−39.61	−40.34
β_s	−15.72	−16.91
β_p	−7.72	−9.19
O		
U_s	−97.83	−99.18
U_p	−78.26	−80.76
β_s	−29.27	−29.00
β_p	−29.27	−29.25



Source	ΔE_{rxn}	D_e(C–H)	D_e(H–H)
AM1	−28.0	81.4	109.4
AM1-SRP	−4.9	104.4	109.4
Expt.	−5.1	104.4	109.5

Figure 5.2 AM1 and AM1-SRP parameters (eV) optimized to reproduce the C–H bond dissociation energy of methanol, the H–H bond dissociation energy of hydrogen, and the experimental energy for the illustrated hydrogen-atom transfer (kcal mol^{-1}). Note that in all cases but one, the magnitude of the parameter change on going from AM1 to AM1-SRP is less than 10 percent

it can be done. Because of the relatively poor fashion with which semiempirical models handle non-bonded interactions, there is every reason to expect that such models would be disastrously bad at predicting biomolecular geometries – or at the very least inferior to the far more efficient force fields developed and optimized for this exact purpose.

Instead, the virtue of the semiempirical models when applied to such molecules tends to be that they permit the charge distribution to be predicted more accurately given a particular structure. To the extent that biomolecules often employ charge–charge interactions to enhance reactivity and or specificity in the reaction and recognition of smaller molecules, such predictions can be quite useful. Since the QM calculation intrinsically permits polarization of the overall electronic structure, it is capable of showing greater sensitivity to group–group interactions as they modify the charge distribution than is the case for the typical fixed-atomic-charge, non-polarizable force field.

Of course, one may also be interested in the modeling of a bond-making/bond-breaking reaction that takes place within a very large molecular framework, in which case the availability of appropriate force-field models is extremely limited and one must perforce resort to some QM approach in practice. Recognition of the complementary strengths and weaknesses of QM and MM models has led to extensive efforts to combine them in ways that allow maximum advantage to be taken of the good points of both; such QM/MM hybrid models are the subject of Chapter 13.

5.8 Case Study: Asymmetric Alkylation of Benzaldehyde

Synopsis of Goldfuss and Houk (1998) 'Origin of Enantioselectivities in Chiral β-Amino Alcohol Catalyzed Asymmetric Additions of Organozinc Reagents to Benzaldehyde: PM3 Transition State Modeling'.

A major goal of organic synthesis is the preparation of chiral molecules in an optically pure fashion, i.e., as single enantiomers. Any such process must involve discrimination between enantiomerically related transition states based on a chiral environment, and a popular method for establishing such an environment is to employ a so-called chiral auxiliary as part of one or more of the involved reagents. Since the auxiliary must itself be pure in order to be maximally effective, and since optically pure molecules can be expensive, even when derived from natural products, it is especially desirable to design processes where the chiral auxiliary forms part of a catalyst rather than part of a stoichiometric reagent. An example of such a process is the addition of organozinc reagents to aldehydes. In the presence of β-amino alcohols, one equivalent of dialkylzinc reacts with the alcohol to liberate ethane and form an amino-coordinated zinc alkoxide. This alkoxide catalyzes the addition of a second equivalent of dialkylzinc to aldehydes by forming supermolecular complexes like those illustrated in Figure 5.3.

When the β-amino alcohol ligand is chiral and optically pure, there are four potentially low-energy TS structures that may lead to products. Several chiral ligands have been shown

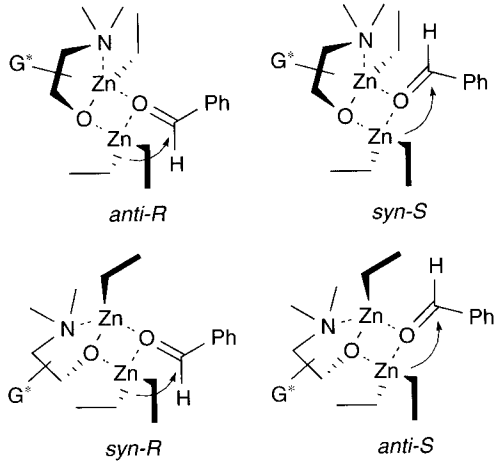


Figure 5.3 Four alternative TS structures for catalyzed addition of diethylzinc to benzaldehyde. The descriptors refer to the side of the four-membered ring on which the aldehyde carbon is found relative to the alkoxide carbon – same (*syn*) or opposite (*anti*) – and the absolute configuration of the new stereogenic center formed following ethyl transfer, *R* or *S*. In the absence of chirality in the β-amino alcohol ligand, indicated by the G* group(s), the TS structures at opposite corners would be enantiomeric with one another, and no preference for *R* over *S* product would be observed. At least four other TS structures can be readily imagined while maintaining the atomic connectivities of those shown here. What are they and why might they be intuitively discounted? Are there still other TS structures one might imagine? How does one decide when all relevant TS structures have been considered?

to give high enantioselectivities in the alkyl addition, indicating that either a single one of the four TS structures is significantly lower in energy, or, if not, the two associated with one enantiomer are significantly lower than either of the two for the other enantiomer.

To better determine the specific steric and/or electronic influences giving rise to high observed enantioselectivities, Goldfuss and Houk studied the energies of the four TS structures in Figure 5.3 for different chiral β-amino alcohols at the PM3 level of theory.

One possible concern in such an approach is the quality of the Zn parameters in PM3, since experimental data for zinc compounds are considerably more sparse than for more quotidian organic compounds. Thus, as a first step, Goldfuss and Houk considered the small complex formed from formaldehyde, di*methyl*zinc, and unsubstituted β-aminoethanol. They compared the geometries of the two TS structures predicted at the PM3 level to those previously obtained by another group at the *ab initio* HF/3-21G level (note that since the amino alcohol is not chiral, there are two TS structures, not four); they observed that agreement was reasonable for the gross shapes of the TS structures, although there were fairly substantial differences in various bond lengths – up to 0.2 Å in Zn–O bonds and the forming C–C bond. They also compared the relative energies for the two TS structures at the PM3 level to those previously reported from small, correlated *ab initio* calculations. Agreement was at best fair, with PM3 giving an energy difference between the two structures of 6.8 kcal mol^{-1}, compared to the prior result of 2.9 kcal mol^{-1}.

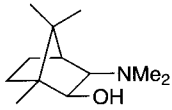
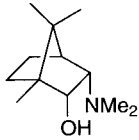
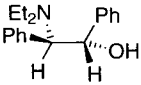
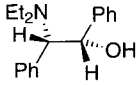
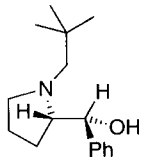
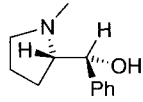
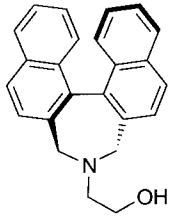
Comparison between PM3 and the previously reported levels of theory is interesting from a methodological perspective. However, to the extent that there are significant disagreements between the methods, PM3 is as likely to be the most accurate as any, given the rather low levels of *ab initio* theory employed (*ab initio* theory is discussed in detail in the next two chapters). Insofar as the size of the chemical problem makes it impractical to seek converged solutions of the Schrödinger equation, Goldfuss and Houk turned instead to a comparison of PM3 to available experimental data. In particular, they computed product ratios based on the assumption that these would reflect a 273 K Boltzmann distribution of corresponding TS structures (this follows from transition state theory, discussed in Section 15.3, for a reaction under kinetic control). For the TS energies, they employed the relative PM3 electronic energies plus zero-point vibrational energies obtained from frequency calculations (see Section 10.2). Then, for a variety of different chiral β-amino alcohols, they compared predicted enantiomeric excess, defined as

$$\%\text{ee} = |\%R - \%S| \tag{5.17}$$

to experimental values obtained under a variety of different conditions. This comparison, summarized in Table 5.3, shows remarkably good agreement between PM3 and experiment.

It is worth a brief digression to note that, from a theoretical standpoint, it is rather easy to make predictions in cases where a single product is observed. When experimentalists report a single product, they typically mean that to within the detection limits of their analysis, they observe only a single compound – unless special efforts are undertaken, this might imply no better than 20:1 excess of the observed product over any other possibilities. At 298 K, this implies that the TS structure of lowest energy lies at least 2 kcal mol^{-1} below any competing TS structures. Of course, it might be 20 or 200 kcal mol^{-1} below competing TS structures – when experiment reports only a single product, there is no way to quantify this. Thus, even if theory is badly in error, as long as the correct TS structure is predicted to be lowest by more than 2 kcal mol^{-1}, there will be 'perfect' agreement with

Table 5.3 Comparison of predicted and experimental enantiomeric excesses for diethylzinc addition to benzaldehyde in the presence of various β-amino alcohols

β-Amino alcohol	Configuration	%ee	
		PM3	Experiment
	*S*	100	99
	*R*	99	95
	*S*	100	94
	*S*	98	81
	*R*	97	100
	*R*	82	72
	*S*	33	49

experiment. If, however, two or more products are reported with a quantitative ratio, the quality of the theoretical results can be much more accurately judged. At 298 K, every error of 1.4 kcal mol^{-1} in predicted relative energies between two TS structures will change the ratios of predicted products by an order of magnitude. Thus, in the case of two competing

TS structures leading to different enantiomers, a %ee of 0 would results from equal TS energies, a %ee of 82 from relative energies of 1.4 kcal mol^{-1}, and a %ee of 98 from relative energies of 2.8 kcal mol^{-1}. Given this analysis, the near quantitative agreement between PM3 and those experimental cases showing %ee values below 90 reflects startlingly good accuracy for a semiempirical level of theory.

Armed with such solid agreement between theory and experiment, Goldfuss and Houk go on to analyze the geometries of the various TS structures to identify exactly which interactions lead to unfavorably high energies and can be used to enhance chiral discrimination. They infer in particular that the optimal situation requires that the alkoxide carbon atom be substituted by two groups of significantly different size, e.g., a hydrogen atom and a bulky alkyl or aryl group. This work thus provides a nice example of how preliminary experimental work can be used to validate an economical theoretical model that can then be used to suggest future directions for further experimental optimization. However, it must not be forgotten that the success of the model must derive in part from favorable cancellation of errors – the theoretical model, after all, fails to account for solvent, thermal contributions to free energies, and various other possibly important experimental conditions. As such, application of the model in a predictive mode should be kept within reasonable limits, e.g., results for new β-amino alcohol structures would be expected to be more secure than results obtained for systems designed to use a substituted 1,2-diaminoethane ligand in place of the β-amino alcohol.

Bibliography and Suggested Additional Reading

Clark, T. 2000. 'Quo Vadis Semiempirical MO-theory?' *J. Mol. Struct. (Theochem)*, **530**, 1.

Dewar, M. J. S. 1975. *The PMO Theory of Organic Chemistry*, Plenum: New York.

Hall, M. B. 2000. 'Perspective on "The Spectra and Electronic Structure of the Tetrahedral Ions MnO_4^-, CrO_4^{2-}, and ClO_4^-"' *Theor. Chem. Acc.*, **103**, 221.

Hehre, W. J. 1995. *Practical Strategies for Electronic Structure Calculations*, Wavefunction: Irvine, CA.

Jensen, F. 1999. *Introduction to Computational Chemistry*, Wiley: Chichester.

Levine, I. N. 2000. *Quantum Chemistry*, 5th Edn., Prentice Hall: New York.

Pople, J. A. and Beveridge, D. A. 1970. *Approximate Molecular Orbital Theory*, McGraw-Hill: New York.

Stewart, J. J. P. 1990. 'Semiempirical Molecular Orbital Methods' in *Reviews in Computational Chemistry*, Vol. 1, Lipkowitz, K. B. and Boyd, D. B., Eds., VCH: New York, 45.

Thiel, W. 1998. 'Thermochemistry from Semiempirical Molecular Orbital Theory' in *Computational Thermochemistry, ACS Symposium Series*, Vol. 677, Irikura, K. K. and Frurip, D. J., Eds., American Chemical Society: Washington, DC, 142.

Whangbo, M.-H. 2000. "Perspective on 'An extended Hückel theory. I. Hydrocarbons"' *Theor. Chem. Acc.*, **103**, 252.

Zerner, M. 1991. 'Semiempirical Molecular Orbital Methods' in *Reviews in Computational Chemistry*, Vol. 2, Lipkowitz, K. B. and Boyd, D. B., Eds., VCH: New York, 313.

References

Barrows, S. E., Dulles, F. J., Cramer, C. J., French, A. D., and Truhlar, D. G. 1995. *Carbohydr. Res.*, **276**, 219.

Bernal-Uruchurtu, M. I., Martins-Costa, M. T. C., Millot, C., and Ruiz-Lopez, M. F. 2000. *J. Comput. Chem.*, **21**, 572.
Bingham, R. C., Dewar, M. J. S., and Lo, D. H. 1975. *J. Am. Chem. Soc.*, **97**, 1285, 1307.
Bishop, D. M. 1966. *J. Chem. Phys.*, **45**, 1880 and references therein.
Chuang, Y.-Y., Radhakrishnan, M. L., Fast, P. L., Cramer, C. J., and Truhlar, D. G. 1999. *J. Phys. Chem.*, **103**, 4893.
Clementi, E. and Roetti, C. 1974. *At. Data Nucl. Data Tables.*, **14**, 177.
Csonka, G. I. 1993. *J. Comput. Chem.*, **14**, 895.
Cusachs, L. C., Reynolds, J. W., and Barnard, D. 1966. *J. Chem. Phys.*, **44**, 835.
Dannenberg, J. A. 1997. *J. Mol. Struct. (Theochem)*, **401**, 287.
Dewar, M. J. S., Hashmall, J. A., and Venier, C. G. 1968. *J. Am. Chem. Soc.*, **90**, 1953.
Dewar, M. J. S., Jie, C., and Yu, J. 1993. *Tetrahedron*, **49**, 5003.
Dewar, M. J. S., Zoebisch, E. G., Healy, E. F., and Stewart, J. J. P. 1985. *J. Am. Chem. Soc.*, **107**, 3902.
Ferguson, D. M., Gould, W. A., Glauser, W. A., Schroeder, S., and Kollman, P. A. 1992. *J. Comput. Chem.*, **13**, 525.
Genin, H. and Hoffmann, R. 1998. *Macromolecules*, **31**, 444.
Goh, S. K. and Marynick, D. S. 2001. *J. Comput. Chem.*, **22**, 1881.
Goldfuss, B. and Houk, K. N. 1998. *J. Org. Chem.*, **63**, 8998.
Hall, M. B. and Fenske, R. F. 1972. *Inorg. Chem.*, **11**, 768.
Hinze, J. and Jaffé, H. H. 1962. *J. Am. Chem. Soc.*, **84**, 540.
Hoffmann, R. 1963. *J. Chem. Phys.*, **39**, 1397.
Holder, A., Dennington, R. D., and Jie, C. 1994. *Tetrahedron*, **50**, 627.
Jug, K. and Schulz, J. 1988. *J. Comput. Chem.*, **9**, 40.
Khandogin, L, Hu, A. G., and York, D. M. 2000. *J. Comput. Chem.*, **21**, 1562.
Koopmans, T. 1933. *Physica (Utrecht)*, **1**, 104.
Mataga, N. and Nishimoto, K. 1957. *Z. Phys. Chem.*, **13**, 140.
Mulliken, R. S., Rieke, C. A., and Orloff, H. 1949. *J. Chem. Phys.*, **17**, 1248.
Nanda, D. N. and Jug, K. 1980. *Theor. Chim. Acta*, **57**, 95.
Pachkovski, S. and Thiel, W. 1996. *J. Am. Chem. Soc.*, **118**, 7164.
Pariser, R. and Parr, R. G. 1953. *J. Chem. Phys.*, **21**, 466, 767.
Pekeris, C. L. 1959. *Phys. Rev.*, **115**, 1216.
Pilcher, G. and Skinner, H. A. 1962. *Inorg. Nucl. Chem.*, **24**, 937.
Pople, J. A. 1953. *Trans. Faraday Soc.*, **49**, 1375.
Pople, J. A. and Segal, G. A. 1965. *J. Chem. Phys.*, **43**, S136.
Pople, J. A., Beveridge, D. L., and Dobosh, P. A. 1967. *J. Chem. Phys.*, **47**, 2026.
Pople, J. A., Santry, D. P., and Segal, G. A. 1965. *J. Chem. Phys.*, **43**, S129.
Rein, R., Fukuda, N., Win, H., Clarke, G. A., and Harris, F. E. 1966. *J. Chem. Phys.*, **45**, 4743.
Ridley, J. E. and Zerner, M. C. 1973. *Theor. Chim. Acta*, **32**, 111.
Rossi, I. and Truhlar, D. G. 1995. *Chem. Phys. Lett.*, **233**, 231.
Schweig, A. and Thiel, W. 1981. *J. Am. Chem. Soc.*, **103**, 1425.
Slater, J. C. 1930. *Phys. Rev.*, **36**, 57.
Stewart, J. J. P. 1989. *J. Comput. Chem.*, **10**, 209, 221.
Stewart, J. J. P. 1991. *J. Comput. Chem.*, **12**, 320.
Stewart, J. J. P. 1996. *Int. J. Quantum Chem.*, **58**, 133.
Storer, J. W., Giesen, D. J., Cramer, C. J., and Truhlar, D. G. 1995. *J. Comput-Aided Mol. Des.*, **9**, 87.
Thiel, W. 1981. *J. Am. Chem. Soc.*, **103**, 1413, 1421.

Thiel, W. and Voityuk, A. A. 1992, *Theor. Chim. Acta,* **81**, 391.
Thiel, W. and Voityuk, A. A. 1996, *Theor. Chim. Acta,* **93**, 315.
Thiel, W. and Voityuk, A. A. 1996. *J. Phys. Chem.*, **100**, 616.
van der Vaart, A., Gogonea, V., Dixon, S. L., and Merz, K. M. 2000. *J. Comput. Chem.*, **21**, 1494.
Wolfsberg, M. and Helmholz, L. J. 1952. *J. Chem. Phys.*, **20**, 837.
Zerner, M. and Gouterman, M. 1966. *Theor. Chim. Acta*, **4**, 44.

6

Ab Initio Implementations of Hartree–Fock Molecular Orbital Theory

6.1 *Ab Initio* Philosophy

The fundamental assumption of HF theory, that each electron sees all of the others as an average field, allows for tremendous progress to be made in carrying out practical MO calculations. However, neglect of electron correlation can have profound chemical consequences when it comes to determining accurate wave functions and properties derived therefrom. As noted in the preceding chapter, the development of semiempirical theories was motivated in part by the hope that judicious parameterization efforts could compensate for this feature of HF theory. While such compensation has no rigorous foundation, to the extent it permits one to make accurate chemical predictions, it may have great practical utility.

Early developers of so-called '*ab initio*' (Latin for 'from the beginning') HF theory, however, tended to be less focused on making short-term predictions, and more focused on long-term development of a *rigorous* methodology that would be worth the wait (a dynamic tension between the need to make predictions now and the need to make better predictions tomorrow is likely to characterize computational chemistry well into the future). Of course, the ultimate rigor is the Schrödinger equation, but that equation is insoluble in a practical sense for all but the most simple of systems. Thus, HF theory, in spite of its fairly significant fundamental assumption, was adopted as useful in the *ab initio* philosophy because it provides a very well defined stepping stone on the way to more sophisticated theories (i.e., theories that come closer to accurate solution of the Schrödinger equation). To that extent, an enormous amount of effort has been expended on developing mathematical and computational techniques to reach the HF limit, which is to say to solve the HF equations with the equivalent of an infinite basis set, *with no additional approximations*. If the HF limit is achieved, then the energy error associated with the HF approximation for a given system, the so-called electron correlation energy E_{corr}, can be determined as

$$E_{\text{corr}} = E - E_{\text{HF}} \tag{6.1}$$

where E is the 'true' energy and E_{HF} is the system energy in the HF limit. Chapter 7 is devoted to the discussion of techniques for estimating E_{corr}.

Along the way it became clear that, perhaps surprisingly, HF energies could be chemically useful. Typically their utility was manifest for situations where the error associated with ignoring the correlation energy could be made unimportant by virtue of comparing two or more systems for which the errors could be made to cancel. The technique of using isodesmic equations, discussed in Section 10.6, represents one example of how such comparisons can successfully be made.

In addition, the availability of HF wave functions made possible the testing of how useful such wave functions might be for the prediction of properties *other* than the energy. Simply because the HF wave function may be arbitrarily far from being an eigenfunction of the Hamiltonian operator does not *a priori* preclude it from being reasonably close to an eigenfunction for some other quantum mechanical operator.

This chapter begins with a discussion of basis sets, the mathematical functions used to construct the HF wave function. Key technical details associated with open-shell vs. closed-shell systems are also addressed. A performance overview and case study are provided in conclusion.

6.2 Basis Sets

The basis set is the set of mathematical functions from which the wave function is constructed. As detailed in Chapter 4, each MO in HF theory is expressed as a linear combination of basis functions, the coefficients for which are determined from the iterative solution of the HF SCF equations (as flow-charted in Figure 4.3). The full HF wave function is expressed as a Slater determinant formed from the individual occupied MOs. In the abstract, the HF limit is achieved by use of an infinite basis set, which necessarily permits an optimal description of the electron probability density. In practice, however, one cannot make use of an infinite basis set. Thus, much work has gone into identifying mathematical functions that allow wave functions to approach the HF limit arbitrarily closely in as efficient a manner as possible.

Efficiency in this case involves three considerations. As noted in Chapter 4, in the absence of additional simplifying approximations like those present in semiempirical theory, the number of two-electron integrals increases as N^4 where N is the number of basis functions. So, keeping the total number of basis functions to a minimum is computationally attractive. In addition, however, it can be useful to choose basis set functional forms that permit the various integrals appearing in the HF equations to be evaluated in a computationally efficient fashion. Thus, a larger basis set can still represent a computational improvement over a smaller basis set if evaluation of the greater number of integrals for the former can be carried out faster than for the latter. Finally, the basis functions must be chosen to have a form that is useful in a chemical sense. That is, the functions should have large amplitude in regions of space where the electron probability density (the wave function) is also large, and small amplitudes where the probability density is small. The simultaneous optimization of these three considerations is at the heart of basis set development.

6.2.1 Functional Forms

Slater-type orbitals were introduced in Section 5.2 (Eq. (5.2)) as the basis functions used in extended Hückel theory. As noted in that discussion, STOs have a number of attractive features primarily associated with the degree to which they closely resemble hydrogenic atomic orbitals. In *ab initio* HF theory, however, they suffer from a fairly significant limitation. There is no analytical solution available for the general four-index integral (Eq. (4.56)) when the basis functions are STOs. The requirement that such integrals be solved by numerical methods severely limits their utility in molecular systems of any significant size.

Boys (1950) proposed an alternative to the use of STOs. All that is required for there to be an analytical solution of the general four-index integral formed from such functions is that the radial decay of the STOs be changed from e^{-r} to e^{-r^2}. That is, the AO-like functions are chosen to have the form of a Gaussian function. The general functional form of a normalized Gaussian-type orbital (GTO) in atom-centered Cartesian coordinates is

$$\phi(x, y, z; \alpha, i, j, k) = \left(\frac{2\alpha}{\pi}\right)^{3/4} \left[\frac{(8\alpha)^{i+j+k} i!j!k!}{(2i)!(2j)!(2k)!}\right]^{1/2} x^i y^j z^k e^{-\alpha(x^2+y^2+z^2)} \tag{6.2}$$

where α is an exponent controlling the width of the GTO, and i, j, and k are non-negative integers that dictate the nature of the orbital in a Cartesian sense.

In particular, when all three of these indices are zero, the GTO has spherical symmetry, and is called an s-type GTO. When exactly one of the indices is one, the function has axial symmetry about a single Cartesian axis and is called a p-type GTO. There are three possible choices for which index is one, corresponding to the p_x, p_y, and p_z orbitals.

When the sum of the indices is equal to two, the orbital is called a d-type GTO. Note that there are six possible combinations of index values (i, j, k) that can sum to two. In Eq. (6.2), this leads to possible Cartesian prefactors of x^2, y^2, z^2, xy, xz, and yz. These six functions are called the Cartesian d functions. In the solution of the Schrödinger equation for the hydrogen atom, only five functions of d-type are required to span all possible values of the z component of the orbital angular momentum for $l = 2$. These five functions are usually referred to as xy, xz, yz, $x^2 - y^2$, and $3z^2 - r^2$. Note that the first three of these canonical d functions are common with the Cartesian d functions, while the latter two can be derived as linear combinations of the Cartesian d functions. A remaining linear combination that can be formed from the Cartesian d functions is $x^2 + y^2 + z^2$, which, insofar as it has spherical symmetry, is actually an s-type GTO. Different Gaussian basis sets adopt different conventions with respect to their d functions: some use all six Cartesian d functions, others prefer to reduce the total basis set size and use the five linear combinations. [Note that if the extra function is kept, the linear combination having s-like symmetry still has the same exponent α governing its decay as the rest of the d set. As d orbitals are more diffuse than s orbitals having the same principal quantum number (which is to say the magnitude of α for the nd GTOs will be smaller than that for the α of the ns GTOs), the extra s orbital does not really contribute at the same principal quantum level, as discussed in more detail below.]

As one increases the indexing, the disparity between the number of Cartesian functions and the number of canonical functions increases. Thus, with f-type GTOs (indices summing

to 3) there are 10 Cartesian functions and 7 canonical functions, with g-type 15 and 10, etc. GTOs can be taken arbitrarily high in angular momentum.

6.2.2 Contracted Gaussian Functions

Although they are convenient from a computational standpoint, GTOs have specific features that diminish their utility as basis functions. One issue of key concern is the shape of the radial portion of the orbital. For s type functions, GTOs are smooth and differentiable at the nucleus ($r = 0$), but real hydrogenic AOs have a cusp (Figure 6.1). In addition, all hydrogenic AOs have a radial decay that is exponential in r while the decay of GTOs is exponential in r^2; this results in too rapid a reduction in amplitude with distance for the GTOs.

In order to combine the best feature of GTOs (computational efficiency) with that of STOs (proper radial shape), most of the first basis sets developed with GTOs used them as building blocks to approximate STOs. That is, the basis functions φ used for SCF calculations were not individual GTOs, but instead a linear combination of GTOs fit to reproduce as accurately as possible a STO, i.e.,

$$\varphi(x, y, z; \{\alpha\}, i, j, k) = \sum_{a=1}^{M} c_a \phi(x, y, z; \alpha_a, i, j, k) \tag{6.3}$$

where M is the number of Gaussians used in the linear combination, and the coefficients c are chosen to optimize the shape of the basis function sum and ensure normalization. When a basis function is defined as a linear combination of Gaussians, it is referred to as a 'contracted' basis function, and the individual Gaussians from which it is formed are called 'primitive' Gaussians. Thus, in a basis set of contracted GTOs, each basis function is defined by the contraction coefficients c and exponents α of each of its primitives. The 'degree of

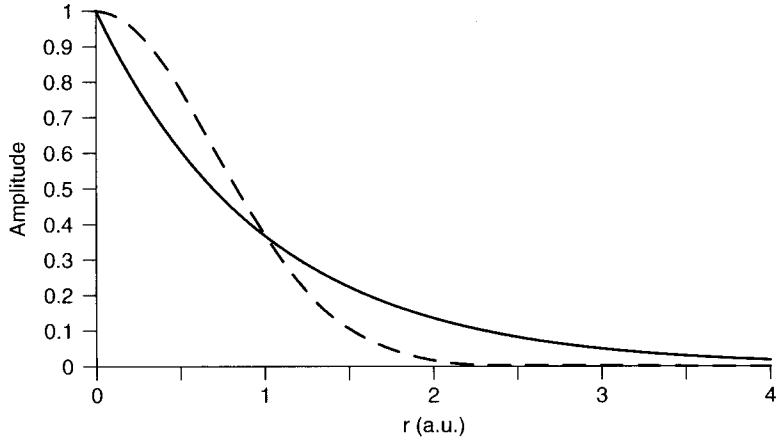


Figure 6.1 Behavior of e^x where $x = r$ (solid line, STO) and $x = r^2$ (dashed line, GTO)

contraction' refers to the total number of primitives used to make all of the contracted functions, as described in more detail below. Contracted GTOs when used as basis functions continue to permit analytical evaluation of all of the four-index integrals.

Hehre, Stewart, and Pople (1969) were the first to systematically determine optimal contraction coefficients and exponents for mimicking STOs with contracted GTOs for a large number of atoms in the periodic table. They constructed a series of different basis sets for different choices of M in Eq. (6.3). In particular, they considered $M = 2$ to 6, and they called these different basis sets STO-MG, for 'Slater-Type Orbital approximated by M Gaussians'. Obviously, the more primitives that are employed, the more accurately a contracted function can be made to match a given STO. However, note that a four-index two-electron integral becomes increasingly complicated to evaluate as each individual basis function is made up of increasingly many primitive functions, according to

$$
\begin{aligned}
(\mu\nu|\lambda\sigma) &= \iint \varphi_\mu(1)\varphi_\nu(1)\frac{1}{r_{12}}\varphi_\lambda(2)\varphi_\sigma(2)dr_1dr_2 \\
&= \iint \sum_{a_\mu=1}^{M_\mu} c_{a_\mu}\phi_{a_\mu}(1) \sum_{a_\nu=1}^{M_\nu} c_{a_\nu}\phi_{a_\nu}(1)\frac{1}{r_{12}} \sum_{a_\lambda=1}^{M_\lambda} c_{a_\lambda}\phi_{a_\lambda}(2) \sum_{a_\sigma=1}^{M_\sigma} c_{a_\sigma}\phi_{a_\sigma}(2)dr_1dr_2 \\
&= \sum_{a_\mu=1}^{M_\mu}\sum_{a_\nu=1}^{M_\nu}\sum_{a_\lambda=1}^{M_\lambda}\sum_{a_\sigma=1}^{M_\sigma} c_{a_\mu}c_{a_\nu}c_{a_\lambda}c_{a_\sigma}\int \phi_{a_\mu}(1)\phi_{a_\nu}(1)\frac{1}{r_{12}}\phi_{a_\lambda}(2)\phi_{a_\sigma}(2)dr_1dr_2
\end{aligned}
\tag{6.4}
$$

It was discovered that the optimum combination of speed and accuracy (when comparing to calculations using STOs) was achieved for $M = 3$. Figure 6.2 compares a 1s function using the STO-3G formalism to the corresponding STO and shows also the 3 primitives from which the contracted basis function is constructed. STO-3G basis functions have been defined for most of the atoms in the periodic table.

Gaussian functions have another feature that would be undesirable if they were to be used individually to represent atomic orbitals: they fail to exhibit radial nodal behavior. Thus, no choice of variables permits Eq. (6.3) to mimic a 2s orbital, which is negative near the origin and positive beyond a certain radial distance. Use of a contraction scheme, however, alleviates this problem; contraction coefficients c in Eq. (6.4) can be chosen to have either negative or positive sign, and thus fitting to functions having radial nodal behavior poses no special challenges.

While the acronym STO-3G is designed to be informative about the contraction scheme, it is appropriate to mention an older and more general notation that appears in much of the earlier literature, although it has mostly fallen out of use today. In that notation, the STO-3G H basis set would be denoted (3s)/[1s]. The material in parentheses indicates the number and type of primitive functions employed, and the material in brackets indicates the number and type of contracted functions. If first-row atoms are specified too, the notation for STO-3G would be (6s3p/3s)/[2s1p/1s]. Thus, for instance, lithium would require 3 each (since it is STO-3G) of 1s primitives, 2s primitives, and 2p primitives, so the total primitives are 6s3p, and the contraction schemes creates a single 1s, 2s, and 2p set, so the contracted functions are

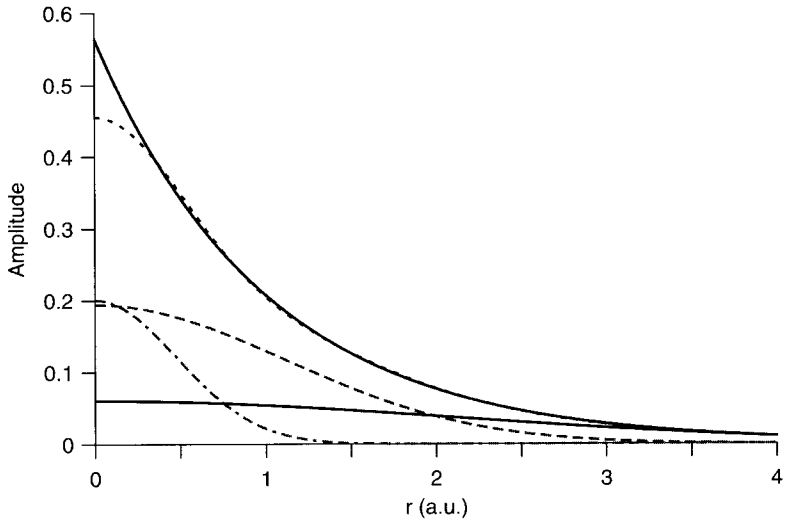


Figure 6.2 The radial behavior of various basis functions in atom-centered coordinates. The bold solid line at top is the STO ($\zeta = 1$) for the hydrogen 1s function; for the one-electron H system, it is also the exact solution of the Schrödinger equation. Nearest it is the contracted STO-3G 1s function (------) optimized to match the STO. It is the sum of a set of one each tight (-·-·-·-·-), medium (– – –), and loose (———) Gaussian functions shown below. The respective Gaussian primitive exponents α are 2.227660, 0.405771, and 0.109818, and the associated contraction coefficients c are 0.154329, 0.535328, and 0.444635. Note that from 0.5 to 4.0 a.u., the STO-3G orbital matches the correct orbital closely. However, near the origin there is a notable difference and, were the plot to extend to very large r, it would be apparent that the decay of the STO-3G orbital is more rapid than the correct orbital, in analogy to Figure 6.1

2s1p. These are separated from the hydrogenic details by a slash in each instance. Extensions to higher rows follow by analogy. Variations on this nomenclature scheme exist, but we will not examine them here.

As a final comment on the STO-*M*G series of basis sets, note that for higher rows than H and He, there is some efficiency to be gained by choosing the exponents used for the primitive Gaussians in the s and p contractions to be the same (then the radial parts of all four-index integrals are identical irrespective of whether they are (ss|ss), (ss|sp), (ss|pp), (sp|sp), etc.). Of course, the shape of s- and p-type functions are different, so the contraction coefficients are *not* identical. When common exponents are chosen in this fashion, the basis functions are sometimes called sp basis functions. Table 6.1 lists the exponents and contraction coefficients for the 2s and 2p functions of oxygen. Note the negative sign of the coefficient for the tightest function in the 2s expansion, thereby providing the proper radial nodal characteristics.

6.2.3 Single-ζ, Multiple-ζ, and Split-Valence

The STO-3G basis set is what is known as a 'single-ζ' basis set, or, more commonly, a 'minimal' basis set. This nomenclature implies that there is one and only one basis function

Table 6.1 STO-3G 2sp basis set for oxygen

α_{2sp}	c_{2s}	c_{2p}
5.0331527	−0.099967	0.155916
1.1695944	0.399513	0.607684
0.3803892	0.700115	0.391957

defined for each type of orbital core through valence. Thus for H and He, there is only a 1s function. For Li to Ne, there are five functions, 1s, 2s, $2p_x$, $2p_y$, and $2p_z$. For Na to Ar, 3s, $3p_x$, $3p_y$, and $3p_z$ are added to the second-row set, making a total of nine functions, etc. This number is the absolute minimum required, and it is certainly nowhere near the infinite basis set limit. Other minimal basis sets include the MINI sets of Huzinaga and co-workers, which are named MINI-1, MINI-2, etc., and vary in the number of primitives used for different kinds of functions.

One way to increase the flexibility of a basis set is to 'decontract' it. That is, we might imagine taking the STO-3G basis set, and instead of constructing each basis function as a sum of three Gaussians, we could construct *two* basis functions for each AO, the first being a contraction of the first two primitive Gaussians, while the second would simply be the normalized third primitive. This prescription would not double the size of our basis set, since we would have all the same individual integrals to evaluate as previously, but the size of our secular equation *would* be increased. A basis set with two functions for each AO is called a 'double-ζ' basis. Of course, we could decontract further, and treat each primitive as a full-fledged basis function, in which case we would have a 'triple-ζ' basis, and we could then decide to add more functions indefinitely creating higher and higher multiple-ζ basis sets. Modern examples of such basis sets are the cc-pCVDZ, cc-pCVTZ, etc. sets of Dunning and co-workers, where the acronym stands for 'correlation-consistent polarized Core and Valence (Double/Triple/etc.) Zeta' (Woon and Dunning 1995); correlation consistency and polarization are described in more detail below.

The advantage of such a scheme is, naturally, that these increasingly large basis sets must come closer and closer to the HF limit. Let us step back for a moment, however, and consider the *chemical* consequences of providing extra basis functions for a given AO. Recall that a final MO from an HF calculation is a linear combination of all of the basis functions. Indeed, if we were to examine the 1s core orbital resulting from an HF calculation on atomic oxygen using the fully uncontracted set of STO-3G Gaussian primitives as a basis (i.e., a triple-ζ basis), we might well find it to be a linear combination of the 1s functions very similar to that *defining* the STO-3G *contracted* oxygen 1s function. And, if we were to look at the MOs resulting from an equivalent calculation on, say, formaldehyde ($H_2C{=}O$), we would probably find another orbital which we would assign as the oxygen 1s orbital having very similar AO coefficients. Indeed, we would find this same orbital little changed in almost any molecule incorporating oxygen we might choose to examine. The reason for this is that core orbitals are only weakly affected by chemical bonding.

Valence orbitals, on the other hand, can vary widely as a function of chemical bonding. Atoms bonded to significantly more electronegative elements take on partial positive charge

from loss of valence electrons, and thus their remaining density is distributed more compactly. The reverse is true when the bonding is to a more electropositive element. From a chemical standpoint, then, there is more to be gained by having flexibility in the valence basis functions than in the core, and recognition of this phenomenon led to the development of so-called 'split-valence' or 'valence-multiple-ζ' basis sets. In such basis sets, core orbitals continue to be represented by a single (contracted) basis function, while valence orbitals are split into arbitrarily many functions.

Amongst the most widely used split-valence basis sets are those of Pople *et al.* These basis sets include 3-21G, 6-21G, 4-31G, 6-31G, and 6-311G. The nomenclature is a guide to the contraction scheme. The first number indicates the number of primitives used in the contracted core functions. The numbers after the hyphen indicate the numbers of primitives used in the valence functions – if there are two such numbers, it is a valence-double-ζ basis, if there are three, valence-triple-ζ. This notation is somewhat more informative than the older style noted in the previous section. Thus, for a calculation on water, for instance, the 6-311G basis would be represented (11s5p/5s)/[4s3p/3s]. The latter notation does not specify how many primitives are devoted to which contracted basis functions, while 6-311G makes this point clear. Like the STO-*M*G basis sets, the split-valence sets use sp basis functions having common exponents.

An interesting question arises for split-valence and multiple-ζ basis sets: how should one go about choosing exponents and coefficients for the contracted functions? As the basis is no longer minimal, there is no particular virtue in fitting to STOs (which were originally used because they were thought to represent the optimal single-function approximation to an AO). Pople and co-workers, like most other researchers in the field, relied on the variational principle. That is, some test set of atoms and/or molecules was established, and exponents and coefficients were optimized so as to give the minimum energy over the test set. In the end, just as the name of a force field refers to its functional form and a list of all its parameters, so too the name of a basis set refers to its contraction scheme and a list of all of its exponents and coefficients for each atom.

One feature of the Pople basis sets is that they use a so-called 'segmented' contraction. This implies that the primitives used for one basis function are not used for another of the same angular momentum (e.g., no common primitives between the 2s and 3s basis functions for phosphorus). Such a contraction scheme is typical of older basis sets. Other segmented split-valence basis sets include the MIDI and MAXI basis sets of Huzinaga and co-workers, which are named MIDI-1, MIDI-2, etc., MAXI-1, MAXI-2, etc. and vary in the number of primitives used for different kinds of functions.

The Pople basis sets have seen sufficient use in the literature that certain trends have clearly emerged. While a more complete discussion of the utility of HF theory and its basis-set dependence appears at the end of this chapter, we note here that, in general, the 4-31G basis set is inferior to the less expensive 3-21G, so there is little point in ever using it. The 6-21G basis set is obsolete.

An alternative method to carrying out a segmented contraction is to use a so-called 'general' contraction (Raffenetti 1973). In a general contraction, there is a single set of primitives that are used in *all* contracted basis functions, but they appear with different coefficients

in each. The general contraction scheme has some technical advantages over the segmented one. One advantage in terms of efficiency is that integrals involving the same primitives, i.e., those occurring in the final line of Eq. (6.4), need in principle be calculated only once, and the value can be stored for later reuse as needed. Examples of split-valence basis sets using general contractions are the cc-pVDZ, cc-pVTZ, etc. sets of Dunning and co-workers, where the acronym stands for 'correlation-consistent polarized Valence (Double/Triple/etc.) Zeta' (Dunning 1989; Woon and Dunning 1993). The 'correlation-consistent' part of the name implies that the exponents and contraction coefficients were variationally optimized not only for HF calculations, but also for calculations including electron correlation, methods for which are described in Chapter 7. The subject of polarization is what we turn to next.

6.2.4 Polarization Functions

The distinction between atomic orbitals and basis functions in molecular calculations has been emphasized several times now. An illustrative example of why the two should not necessarily be thought of as equivalent is offered by ammonia, NH_3. The inversion barrier for interconversion between equivalent pyramidal minima in ammonia has been measured to be 5.8 kcal mol^{-1}. However, a HF calculation with the equivalent of an infinite, atom-centered basis set of s and p functions predicts the planar geometry of ammonia to be a minimum-energy structure!

The problem with the calculation is that s and p functions centered on the atoms do not provide sufficient mathematical flexibility to adequately describe the wave function for the pyramidal geometry. This is true even though the atoms nitrogen and hydrogen can individually be reasonably well described entirely by s and p functions. The *molecular* orbitals, which are eigenfunctions of a Schrödinger equation involving multiple nuclei at various positions in space, require more mathematical flexibility than do the atoms.

Because of the utility of AO-like GTOs, this flexibility is almost always added in the form of basis functions corresponding to one quantum number of higher angular momentum than the valence orbitals. Thus, for a first-row atom, the most useful polarization functions are d GTOs, and for hydrogen, p GTOs. Figure 6.3 illustrates how a d function on oxygen can polarize a p function to improve the description of the O–H bonds in the water molecule. The use of p functions to polarize hydrogen s functions has already been mentioned in Section 4.3.1. [An alternative way to introduce polarization is to allow basis functions not to be centered on atoms. Such floating Gaussian orbitals (FLOGOs) are illustrated on the left-hand side of Figure 4.1. While the use of FLOGOs reduces the need to work with integrals involving high-angular-momentum functions, the process of geometry optimization is rendered considerably more complicated, so they are rarely employed in modern calculations.] Adding d functions to the nitrogen basis set causes HF theory to predict correctly a pyramidal minimum for ammonia, although some error in prediction of the inversion barrier still exists even at the HF limit because of the failure to account for electron correlation.

A variety of other molecular properties prove to be sensitive to the presence of polarization functions. While a more complete discussion occurs in Section 6.4, we note here that d functions on second-row atoms are absolutely required in order to make reasonable

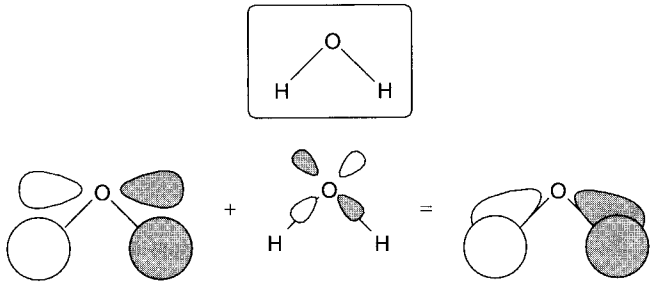


Figure 6.3 The MO formed by interaction between the antisymmetric combination of H 1s orbitals and the oxygen p_x orbital (see also Figure 6.7). Bonding interactions are enhanced by mixing a small amount of Od_{xz} character into the MO

predictions for the geometries of molecules including such atoms in formally hypervalent bonding situations, e.g., phosphates, sulfoxides, siliconates, etc.

A variety of empirical rules exist for choosing the exponent(s) for a set of polarization functions. If only a single set is desired, one possible choice is to make the maximum in the radial density function, $\langle r^2 \rangle$, equal to that for the existing valence set (e.g., the 3d functions that best 'overlap' the 2p functions for a first-row atom – note that the radial density is used instead of the actual overlap integral because the latter, by symmetry, must be zero).

Because of the expense associated with adding polarization functions – the total number of functions begins to grow rather quickly with their inclusion – early calculations typically made use of only a single set. Pople and co-workers introduced a simple nomenclature scheme to indicate the presence of these functions, the '*' (pronounced 'star'). Thus, 6-31G* implies a set of d functions added to polarize the p functions in 6-31G. A second star implies p functions on H and He, e.g., 6-311G** (Krishnan, Frisch, and Pople 1980).

Subsequent work has shown that there is a rough correspondence between the value of adding polarization functions and the value of decontracting the valence basis function(s). In particular, there is a rough equality between each decontraction step and adding one new set of polarization functions, including a new set of higher angular momentum. Put more succinctly, 'balanced' double-ζ basis sets should include d functions on heavy atoms and p functions on H, triple-ζ basis sets should include 1 set of f and 2 sets of d functions on heavy atoms, and 1 set of d and 2 sets of p functions on H, etc. This is the polarization prescription adopted by the cc-pV*n*Z basis sets of Dunning and co-workers already mentioned above, where n ranges over D (double), T (triple), Q (quadruple), five, and six (Wilson, van Mourik, and Dunning 1996). Thus, for cc-pV6Z, for example, each heavy atom has one i function, two h functions, three g functions, four f functions, five d functions, and six valence s and p functions, in addition to core functions (using the canonical numbers of these functions, we have 140 basis functions for a single second-row atom, so this basis set presently finds use only for the smallest of systems). Note that while it would be an unpleasant exercise to try to draw an i function, it is straightforwardly defined by taking the sum of i, j, and k equal to 6 in Eq. (6.2).

Recognizing the tendency to use more than one set of polarization functions in modern calculations, the standard nomenclature for the Pople basis sets now typically includes an

explicit enumeration of those functions instead of the star nomenclature. Thus, 6-31G(d) is to be preferred over 6-31G* because the former obviously generalizes to allow names like 6-31G(3d2fg,2pd), which implies heavy atoms polarized by three sets of d functions, two sets of f functions, and a set of g functions, and hydrogen atoms by two sets of p functions and one of d (note that since this latter basis set is only valence double-ζ, it is somewhat unbalanced by having so many polarization functions).

A partially polarized basis set, MIDI! (where the '!' is pronounced 'bang'; in some electronic structure programs, the abbreviation MIDIX is employed to avoid complications associated with interpretation of the exclamation point), has been introduced by Cramer and Truhlar and co-workers, who adopted a different philosophy in its development (Easton *et al.* 1996; Li, Cramer, and Truhlar 1998). Rather than optimizing the basis set with respect to molecular energies, they sought to design an economical basis set for geometry optimizations and partial charge calculations on medium-sized molecules, including neutrals, cations, and anions, with special emphasis on functional groups that are important for biomolecules. The MIDI! basis set has d functions on all atoms heavier than H for which it is defined with the exception of carbon (i.e., on heteroatoms). Although much smaller than the 6-31G(d) basis set, for instance, in direct comparisons it yields more accurate geometries and charges as judged by comparison to much higher level calculations.

Finally, an important nomenclature point is that most basis sets are *defined* to use the five spherical d functions, but an important exception is 6-31G* (or 6-31G(d)), which is defined to use the six Cartesian d functions. Some electronic structure programs are not flexible about permitting the user to choose how many d functions are used, so it is important to check that a consistent scheme has been employed when comparing to existing literature data. To avoid ambiguity, it is helpful to modify the basis set name when the number of d functions employed is not the same as that assumed as the default, e.g., MIDI!(6d) to denote use of the six Cartesian d functions instead of the normal spherical five with the MIDI! basis. Another nomenclature issue of importance involves the basis set '3-21G*'. While this notation pervades the literature, it is ambiguous and should be avoided. Pople and co-workers suggested taking from 6-31G* the polarization functions for second-row atoms and beyond and using them directly (i.e., without any reoptimization of exponents) with the smaller basis 3-21G. The motivation for this was to address the serious geometry problems that arise for hypervalent third-row atoms without d functions whilst maintaining a very cheap description of second-row atoms. To distinguish this situation from the normal '*', they named this basis set 3-21G$^{(*)}$, and that is the notation that should always be used to emphasize that no d functions are present on first-row atoms.

6.2.5 Diffuse Functions

The highest energy MOs of anions, highly excited electronic states, and loose supermolecular complexes, tend to be much more spatially diffuse than garden-variety MOs. When a basis set does not have the flexibility necessary to allow a weakly bound electron to localize far from the remaining density, significant errors in energies and other molecular properties can occur. To address this limitation, standard basis sets are often 'augmented' with diffuse basis functions when their use is warranted.

In the Pople family of basis sets, the presence of diffuse functions is indicated by a '+' in the basis set name. Thus, 6-31+G(d) indicates that heavy atoms have been augmented with an additional one s and one set of p functions having small exponents. A second plus indicates the presence of diffuse s functions on H, e.g., 6-311++G(3df,2pd). For the Pople basis sets, the exponents for the diffuse functions were variationally optimized on the anionic one-heavy-atom hydrides, e.g., BH_2^-, and are the same for 3-21G, 6-31G, and 6-311G. In the general case, a rough rule of thumb is that diffuse functions should have an exponent about a factor of four smaller than the smallest valence exponent.

In the Dunning family of cc-pV*n*Z basis sets, diffuse functions on all atoms are indicated by prefixing with 'aug'. Moreover, one set of diffuse functions is added for *each* angular momentum already present. Thus, aug-cc-pVTZ has diffuse f, d, p, and s functions on heavy atoms and diffuse d, p, and s functions on H and He.

Particularly for the calculation of acidities and electron affinities, diffuse functions are absolutely required. For instance, the acidity of HF (not Hartree-Fock in this case, but hydrogen fluoride) increases by 44 kcal/mol when the 6-31+G(d) basis set is used instead of unaugmented 6-31G(d).

6.2.6 The HF Limit

Solution of the HF equations with an infinite basis set is defined as the HF limit. Actually carrying out such a calculation is almost never a practical possibility. However, it is sometimes the case that one may extrapolate to the HF limit with a fair degree of confidence.

Of the basis sets discussed thus far, the cc-pV*n*Z and cc-pCV*n*Z examples were designed expressly for this purpose. As they increase in size in a consistent fashion with each increment of n, one can imagine plotting some particular computed property as a function of n^{-1} and extrapolating a curve fit through those points back to the intercept; the intercept corresponds to $n = \infty$, i.e., the infinite basis limit (Figure 6.4).

Note that certain issues do arise in how one should carry out this extrapolation. If the property is sensitive to geometry, should the geometry be optimized at each level, or should a single geometry be chosen, thereby permitting the extrapolating equation to account for basis-set effects only? Are there any fundamental principles dictating what form the extrapolating equation should take, or can any arbitrary curve fitting approach be applied? In general, the answers to these questions are case-dependent, and the chemist cannot be completely removed from the calculation.

Note that the cost of the extrapolation procedure outlined above becomes increasingly large as points for $n = 4$, 5, and 6 are added. For systems having more than five or six atoms, these calculations can be staggeringly demanding in terms of computational resources.

A somewhat more common approach is one that does not try explicitly to extrapolate to the HF limit but uses similar concepts to try to correct for some basis-set incompleteness. The assumption is made that the effects of 'orthogonal' increases in basis set size can be considered to be additive (a substantial amount of work suggests that this assumption is typically not too bad, at least for molecular energies), and thus the individual effects can be summed together to estimate the full-basis-set result. This is best illustrated by example. Consider

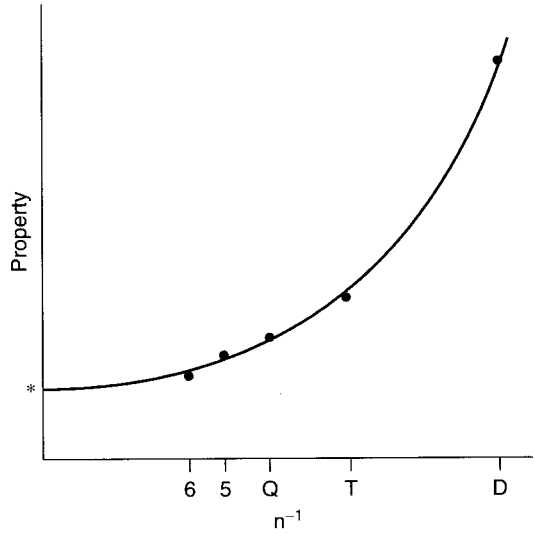


Figure 6.4 Use of an extrapolation procedure to estimate the expectation value for some property at the HF limit. The abscissa is marked off as n^{-1} in cc-pVnZ notation (see page 162). Note the sensitivity of the limiting value, which is to say the ordinate intercept, that might be expected based on the use of different curve-fitting procedures

HF calculations carried out for the chemical warfare agent VX ($C_{11}H_{26}NO_2PS$, Figure 6.5) with the following basis sets: 6-31G, 6-31++G, 6-31G(d,p), 6-311G, and 6-311++G(d,p). With these basis sets, the total number of basis functions for VX are 204, 378, 294, 294, and 542, respectively.

The additivity assumption can be expressed as

$$\begin{aligned} E[\text{HF/6-311++G(d,p)}] \approx\ & E[\text{HF/6-31G}] \\ & + \{E[\text{HF/6-31G(d,p)}] - E[\text{HF/6-31G}]\} \\ & + \{E[\text{HF/6-311G}] - E[\text{HF/6-31G}]\} \\ & + \{E[\text{HF/6-31++G}] - E[\text{HF/6-31G}]\} \end{aligned} \tag{6.5}$$

where the notation 'x/y' implies 'level of theory x using basis set y'. Each successive line on the r.h.s. of Eq. (6.5) reflects the incremental contribution from a particular basis set improvement – first polarization functions, then valence decontraction, then diffuse functions. As already noted, the calculation on the l.h.s. requires 542 basis functions. Although there are four *different* calculations on the r.h.s., if we recall that the amount of time for an HF calculation scales formally as the fourth power of the number of basis functions, the amount of time to carry out those four calculations, expressed as a fraction of the amount of time to carry out the full calculation, is $(204^4 + 378^4 + 294^4 + 294^4)/542^4 = 0.43$. That is, evaluation of the r.h.s. of Eq. (6.5) takes less than half the time of evaluation of the l.h.s.

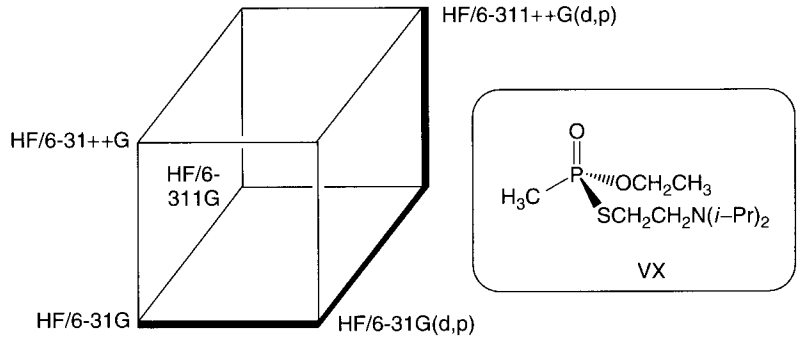


Figure 6.5 The chemical warfare agent VX and a conceptual illustration of the additivity concept embodied in Eq. (6.5). Each boldface line in the additivity cube represents one line on the r.h.s. of the equation

While the above schemes are interesting from a technical standpoint, it must be recalled that chemically there are potentially large errors associated with the HF approximation, so the HF limit is of more interest from a formal standpoint than from a chemical one. Thus, we will defer additional discussion of extrapolation and additivity concepts until Section 7.7, where it is re-examined in the additional context of accounting for electron correlation effects.

6.2.7 Effective Core Potentials

The periodic table is rich and complex, and very heavy elements pose rather distinct challenges to MO theory. First, there is the purely technical hurdle that such elements have large numbers of electrons, and there is thus a concomitant requirement to use a large number of basis functions to describe them. Of course, these extra electrons are mostly core electrons, and thus a minimal representation will probably be adequate. Nevertheless, if one wants to model a small cluster of uranium atoms, for instance, the basis set size quickly becomes intractable. Not surprisingly, more electrons means more energy associated with electron correlation, too.

It was Hellmann (1935) who first proposed a rather radical solution to this problem – replace the electrons with analytical functions that would reasonably accurately, and much more efficiently, represent the combined nuclear–electronic core to the remaining electrons. Such functions are referred to as effective core potentials (ECPs). In a sense, we have already seen ECPs in a very crude form in semiempirical MO theory, where, since only valence electrons are treated, the ECP is a nuclear point charge reduced in magnitude by the number of core electrons.

In *ab initio* theory, ECPs are considerably more complex. They properly represent not only Coulomb repulsion effects, but also adherence to the Pauli principle (i.e., outlying atomic orbitals must be orthogonal to core orbitals having the same angular momentum). This being said, we will not dwell on the technical aspects of their construction. Interested readers are referred to the bibliography at the end of the chapter.

Note that were ECPs to do nothing more than reduce the scope of the electronic structure problem for heavy elements, they would still have great value. However, they have another virtue as well. The core electrons in very heavy elements reach velocities sufficiently near the speed of light that they manifest relativistic effects. A non-relativistic Hamiltonian operator is incapable of accounting for such effects, which can be significant for many chemical properties. A full discussion of modeling relativistic effects, while a fascinating topic, is well beyond the scope of this book. We note here simply that, to the extent an ECP represents the behavior of an atomic core, relativistic effects can be folded in, and thereby removed from the problem of finding suitable wave functions for the remaining electrons.

A key issue in the construction of ECPs is just how many electrons to include in the core. So-called 'large-core' ECPs include everything but the outermost (valence) shell, while 'small-core' ECPs scale back to the next lower shell. Because polarization of the sub-valence shell can be chemically important in heavier metals, it is usually worth the extra cost to explicitly include that shell in the calculations. Thus, the most robust ECPs for the elements Sc–Zn, Y–Cd, and La–Hg, employ [Ne], [Ar], and [Kr] cores, respectively. There is less consensus on the small-core vs. large-core question for the non-metals.

Among the most widely used pseudopotentials are those of Hay and Wadt (sometimes also called the Los Alamos National Laboratory (or LANL) ECPs; Hay and Wadt 1985) and those of Stevens *et al.* (1992). The Hay–Wadt ECPs are non-relativistic for the first row of transition metals while most others are not; as relativistic effects are usually quite small for this region of the periodic table, the distinction is not particularly important.

6.2.8 Sources

Most electronic structure programs come with a library of built-in basis sets, to include many if not all of those mentioned above. A tremendously useful electronic resource is the Environmental Molecular Sciences Laboratory Gaussian Basis Set Order Form, a website that permits the download of a very large number of different basis sets formatted for a variety of different software packages. Moreover, the site has reference information that typically includes values for test calculations as published by the original authors. Since different software packages may have different conventions for how to deal with certain aspects of the basis set (e.g., five spherical vs. six Cartesian d functions), it is always a good idea to carry out such test calculations to ensure that the basis set is being used in a manner consistent with its definition and, hopefully, with previously reported calculations in the literature.

So, how to choose the 'best' basis set for the problem at hand? Obviously a fair rule of thumb is that bigger is better, keeping in mind issues of balance between valence decontraction and presence of polarization functions. As noted above, diffuse functions are warranted in certain specific situations, but in the absence of those situations, there tends to be no strong reason to include them.

Additionally, access to particular software packages may play some role in motivating the choice of basis set. Some packages are equipped to take advantage of efficiencies possible for such features as combined s and p exponents, or general contractions, while others are not, and there may thus be significant timing issues differentiating basis sets.

Finally, and perhaps most important for the vast majority of chemical problems where saturation of the basis set is not a practical possibility, the choice should consider the degree to which other results from that particular basis set at that particular level of theory are available for comparison. For instance, to the extent that there are an enormous number of HF/6-31G(d) results published, and thus a reasonably firm understanding of the specific successes and failures of the model, this can assist in the interpretation of new results – Pople has referred to the collection of all data from a given theoretical prescription as comprising a 'model chemistry' and emphasized the utility of analyzing theoretical performance (and future model development efforts) within such a framework.

6.3 Key Technical and Practical Points of Hartree–Fock Theory

A deep understanding of the underlying theory is, alas, of only limited value in successfully carrying out a HF calculation with any given software package. This section is not designed to supplant program users' manuals, the utility of reading which cannot be overemphasized, but discusses aspects of practical HF calculations that are often glossed over in formal presentations of the theory.

6.3.1 SCF Convergence

As noted in Chapter 4, there is never any guarantee that the SCF process will actually converge to a stable solution. A fairly common problem is so-called 'SCF oscillation'. This occurs when a particular density matrix, call it $\mathbf{P}^{(a)}$, is used to construct a Fock matrix $\mathbf{F}^{(a)}$ (and thus the secular determinant), diagonalization of which permits the construction of an updated density matrix $\mathbf{P}^{(b)}$; this is a general description of any step in the SCF cycle. In the oscillatory case, however, the diagonalization of the Fock matrix created using $\mathbf{P}^{(b)}$ (i.e., $\mathbf{F}^{(b)}$) gives a density matrix indistinguishable from $\mathbf{P}^{(a)}$. Thus, the SCF simply bounces back and forth from $\mathbf{P}^{(a)}$ to $\mathbf{P}^{(b)}$ and never converges. This behavior can be recognized easily by looking at the SCF energy for each step, which itself bounces back and forth between the two discrete values associated with the two different unconverged wave functions defined by $\mathbf{P}^{(a)}$ and $\mathbf{P}^{(b)}$.

In more pathological cases, the SCF behaves even more badly, with large changes occurring in the density matrix at every step. Again, observation of the energies associated with each step is diagnostic for this problem; they are observed to vary widely and seemingly randomly. Such behavior is not uncommon for the first three or four steps of a typical SCF, but usually beyond this point there is a 'zeroing-in' process that leads to convergence.

In the abstract sense, converging the SCF equations is a problem in applied mathematics, and many algorithms have been developed for this process. While the technical details are not presented here, the process is quite analogous to the process of finding a minimum on a PES as described in Chapter 2. In the SCF problem, instead of a space of molecular coordinates we operate in a space of orbital coefficients (so-called 'Fock space'), and there are certain constraints beyond the purely energetic ones, but many of the search strategies are analogous. Similarly analogous is the degree to which they tend to balance speed and

stability. Usually the default optimizer in a given program is the fastest one available, while other methods (e.g., quadratically convergent methods) typically take more steps to converge but are less likely to suffer from oscillation or other problems. Thus, one option for dealing with a system where convergence proves difficult is simply to run through all the different convergence schemes offered by the electronic structure package and hope that one proves sufficiently robust.

In general, however, it is more efficient to solve the problem using chemistry rather than mathematics. If the SCF equations are failing to converge, the problem lies in the initial guess (this is, of course, something of a truism, for if you were to guess the proper eigenfunction, obviously there would be no problem with convergence). Most programs use as their default option a semiempirical method to generate a guess wave function, e.g., EHT or INDO. The resulting wave function (remember that a wave function is simply the list of coefficients describing how the basis functions are put together to form the occupied MOs) is then used to construct a guess for the HF calculation by mapping coefficients from the basis set of the semiempirical method to the basis set for the HF calculation.

When the HF basis set is minimal, this is fairly simple (there is a one-to-one correspondence in basis functions) but when it is larger, some algorithmic choices are made about how to carry out the mapping (e.g., always map to the tightest function or map based on overlap between the semiempirical STO and the large-basis contracted GTO). Thus, it is *usually* easier to converge a small-basis-set HF calculation than a larger one. This suggests a method for bootstrapping one's way to the convergence of a large-basis-set calculation: First, obtain a wave function from a minimal basis set (e.g., STO-3G), then use that as an initial guess for a calculation with a small split-valence basis set (e.g., 3-21G), and repeat this process with increasingly larger basis sets until the target is reached. Because of the exponential scaling, the early calculations typically represent a negligible time investment, especially if they are saving steps in a slowly converging SCF for the full-sized basis set by providing a more accurate initial guess.

The above process has another possible utility that is associated with the molecular geometry. Often when an SCF is difficult to converge, the problem is that the molecular structure is very bad. If that is the case, there can be a very small separation between the highest occupied MO (HOMO) and the lowest unoccupied MO (LUMO). Such small separations wreak havoc on the SCF process, because it is possible that occupation of *either* orbital could lead to HF eigenfunctions of similar energy. In that case, the characters of the two orbitals are very sensitive to all the remaining occupied orbitals, which generate the static potential felt by the highest energy electrons, and their coefficients can undergo large changes that fail to converge (an issue of non-dynamical electron correlation, see Section 7.1). Optimizing the geometry at a low level of theory, where the wave function *can* be coaxed to converge, is typically an efficient way to overcome this problem. Some care must be exercised, however, in systems where the lowest levels of theory may not be reliable for molecular geometries. As a general rule, however, visualization of the structure, and some thoughtful analysis of it by comparison to whatever analogs or prior calculations may be available, is nearly always worth the effort.

Very complete basis sets, or those with many diffuse functions, pose some of the worst problems for SCF convergence because of near-linear dependencies amongst the basis

functions. That is, some basis functions may be fairly well described as linear combinations of other basis functions. This is most readily appreciated by considering two very diffuse s orbitals on adjacent atoms; if they have maxima in their radial density at 40 Å but the two atoms are only 1.5 Å apart, the two basis functions are really almost indistinguishable from one another throughout most of space. If a basis set has a *true* linear dependence, then it is necessarily impossible to assure orthogonality of all of the MOs (a division by zero occurs at a particular point of the SCF process), so very near-linear dependence can lead to numerical instabilities. Thus, it is again important to have a good guess. In a case like this, sometimes it is useful not only to carry out bootstrap calculations in terms of basis sets, but in terms of electrons. Thus, if one is interested in an anion, for instance, one can first try to converge a large-basis-set wave function for the neutral (or the cation), to get a good estimate of the more compact MOs, and then import that wave function as a guess for the anionic system, trying thereby to reduce the impact of possible numerical instabilities.

6.3.2 Symmetry

The presence of symmetry in a molecule can be used to great advantage in electronic structure calculations, although some care is required to avoid possible pitfalls that are simultaneously introduced (Appendix B provides a brief overview of nomenclature (e.g., the term "irrep", which is used below) and key principles of group theory as they apply to MO calculations and symmetry). The advantages of symmetry are primarily associated with computational efficiency.

The most obvious advantage is one step removed from the electronic structure problem, namely geometry optimization. The presence of symmetry elements removes some of the $3N - 6$ degrees of molecular freedom (where N is the number of atoms) that would otherwise be present for an asymmetric molecule. This reduction in the dimensionality of the PES can make the search for stationary points more efficient. Consider benzene (C_6H_6), for example. With 12 atoms, the PES formally has 30 dimensions, and a complete representation would be graphically challenging. However, if we restrict ourselves to structures of D_{6h} symmetry, then there are only two degrees of freedom – one's first choice for defining those degrees of freedom might be the C–C and C–H bond lengths; an equally valid choice, which may be more useful as input to a software program that will ensure preservation of symmetry, is the O–C and O–H radial distances, where O is the point at the center of the benzene ring. Finding a minimum on a two-dimensional PES is obviously quite a bit simpler than on a 30-dimensional one (Figure 6.6).

Symmetry is also tremendously useful in several aspects of solving the SCF equations. A key feature is the degree to which it simplifies evaluation of the four-index integrals. In particular, if the totally symmetric representation is not included in the product of the irreducible representations of basis functions μ, ν, λ, and σ, then $(\mu\nu|\lambda\sigma) = 0$. The analogous rule holds for the one-electron integrals. In general, of course, the atomic basis functions do not belong to *any* irreducible representation, since they themselves do not transform with all the symmetry elements of the molecule. Thus, as a first step to taking advantage of symmetry, linear combinations of the various basis functions must be formed that *do*

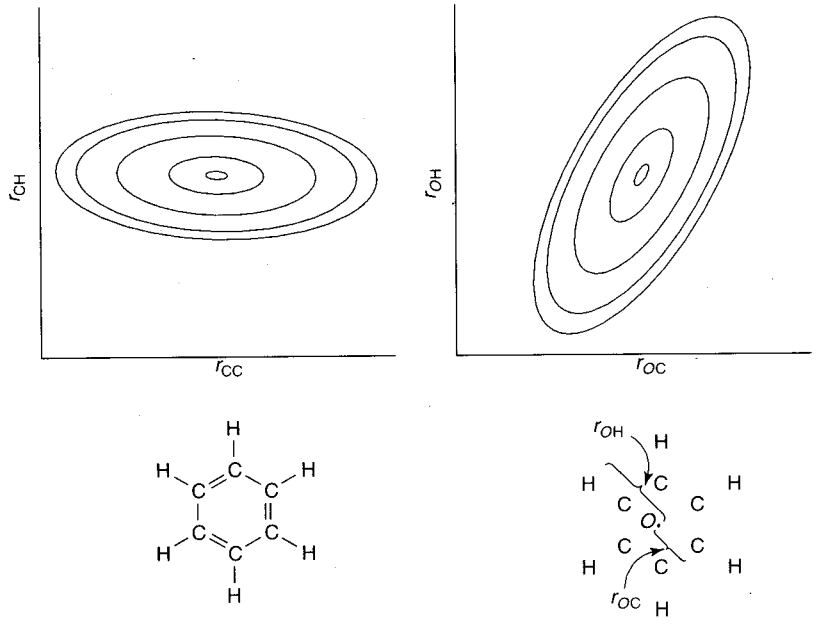


Figure 6.6 Illustration of two two-dimensional PESs for benzene in D_{6h} symmetry. The surfaces differ in choice of coordinates, which may affect optimizer efficiency, ease of input, etc., but will have no effect on the equilibrium structure. Contour lines reflect constant energy intervals of arbitrary magnitude. No attempt is made to illustrate the full 30-dimensional PES, on which it would be considerably more taxing to search for a minimum-energy structure

belong to irreps of the molecular point group. This process is illustrated in Figure 6.7 for a HF/STO-3G calculation on water, which belongs to the C_{2v} point group. The seven atomic basis functions can be linearly transformed to four, two, and one functions belonging to the a_1, b_2, and b_1 irreps, respectively (no combination of basis functions belongs to the a_2 irrep with STO-3G; were d functions to be present on oxygen, the d_{xy} function would be a_2).

In the C_{2v} point group, the totally symmetric representation (a_1) is contained only in the product of any irrep with itself. Thus, any matrix element F_{ab} where transformed basis functions a and b belong to *different* irreps is zero, and the Fock matrix expressed in the transformed basis is block diagonal (see Figure 6.7). Recall that the process of solving the secular equation is equivalent to diagonalization of the Fock matrix. Diagonalization of a block diagonal matrix can be accomplished by separate diagonalization of each block. Noting that diagonalization scales as N^3, where N is the dimensionality of the matrix, the total time to diagonalize our symmetrized Fock matrix for water compared to the unsymmetrized alternative is $(4^3 + 2^3 + 1^3)/7^3 = 0.21$, i.e., a saving of almost 80 percent.

Strangely enough, with all of the advantages of symmetry, one often finds in the literature statements by authors that they deliberately did *not* employ symmetry, as though such a protocol has some virtue associated with it. What motivates this choice? For some, it reflects a reluctance to work on a reduced-dimensionality PES because minima on that PES may *not* be

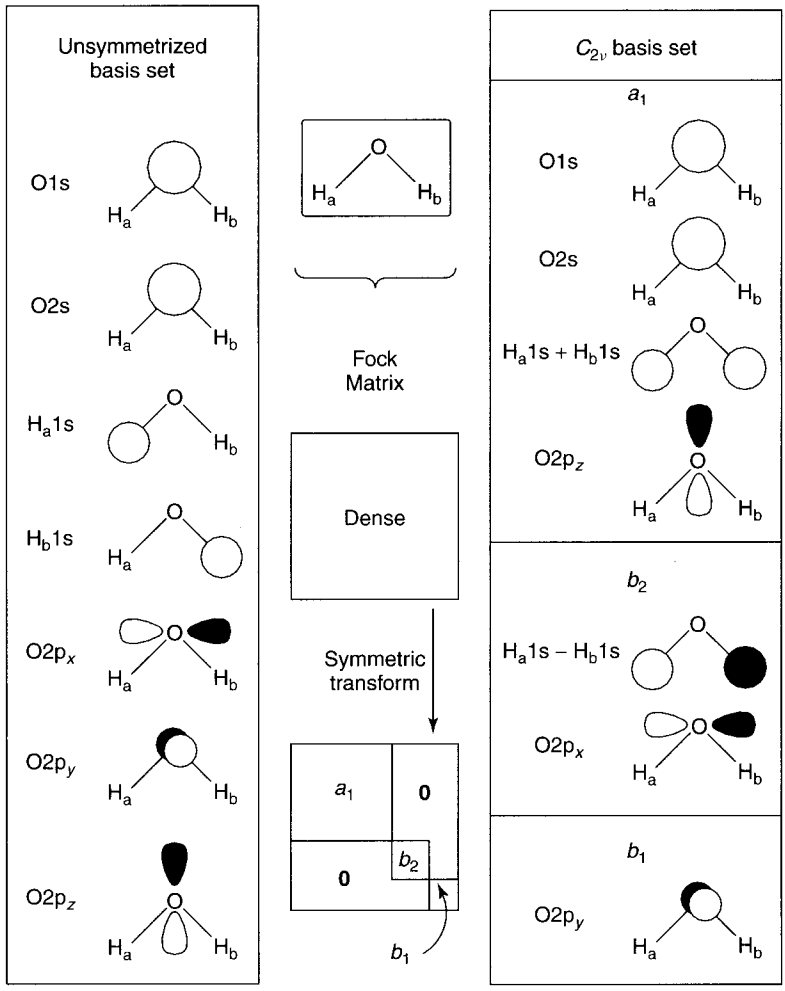


Figure 6.7 Transformation of the minimal water AO basis set to one appropriate for the C_{2v} point group. The effect on the form of the Fock matrix is also illustrated

minima on the full PES. This is best illustrated with an example. Consider the chloride/methyl chloride system with D_{3h} symmetry imposed upon it (Figure 6.8). The system under these constraints has only two degrees of freedom, the C–H bond length and the C–Cl bond length; the overall structure, however, is that associated with the exchange of one chloride ion for another in a bimolecular nucleophilic substitution (i.e., an S_N2 reaction). Minimizing the energy of the system subject to the D_{3h} constraint will give the best possible energy for this arrangement, and hence the TS structure for the reaction. The reason that it is a TS structure and not a true minimum is that the degree(s) of freedom that would change in order to reach a minimum energy structure, i.e., to generate different C–Cl bond lengths, are not included in the reduced-dimensionality PES. Had no symmetry been imposed, however, the

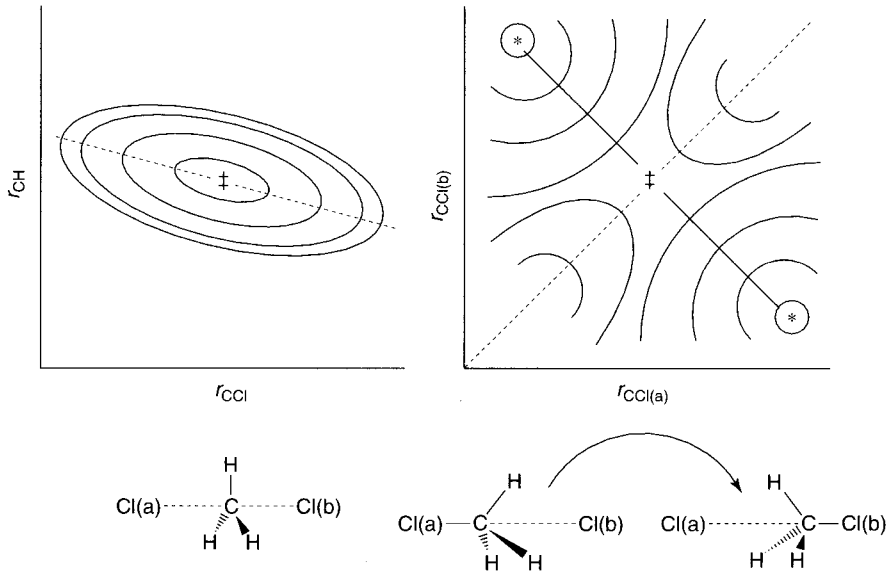


Figure 6.8 Reduced-dimensionality PESs for the chloride/methyl chloride system. On the left is the two-dimensional surface associated with a D_{3h} symmetry constraint. On this surface, the point marked ‡ is a minimum. The simultaneous shortening or lengthening of the C–Cl bonds (simultaneous to preserve D_{3h} symmetry) while allowing the C–H bond lengths to relax is indicated by the dashed line on this surface. The same process is indicated by the dashed line on the surface to the right, whose coordinates are the individual C–Cl bond lengths, and point ‡ again represents the minimum on this line. However, movement *off* the dashed line can lower the energy further. Movement along the *solid* line, which involves lengthening one C–Cl bond whilst shortening the other, corresponds to the reaction path for nucleophilic substitution from one equilibrium structure to another (points marked *), and illustrates that the minimum-energy structure under the D_{3h} constraint is actually a TS structure on the full PES

system would eventually have moved in this direction (given a competent optimizer) unless a transition-state search had been specified.

This issue is really of little importance for modern purposes, however. The best way to evaluate the nature of a stationary point, irrespective of whether it was located using symmetry or not, is to carry out a calculation of the full-dimensional Hessian matrix (see Sections 2.4.1 and 9.3.2). Such a calculation is definitive. In the event that a symmetric stationary point is found *not* to have the character desired, it is often of interest in any case (if it is a TS structure) because inspection of the mode(s) having negative force constants permits an efficient start at optimizing to the desired point using lower symmetry. Since higher symmetry calculations tend to be quite efficient in any case, there is little to be lost by imposing symmetry at the start, and deciding along the way whether some or all symmetry constraints must be relaxed. Note, however, that symmetry constraints must arise from *molecular* symmetry, not an erroneous idea of local symmetry. Thus, for instance, the three C–H bonds of a methyl group should not be constrained to have the same length unless they are truly symmetrically related by a molecular C_3 axis.

Another potential pitfall with symmetry constraints involves the nature of the wave function. Consider the nitroxyl radical H_2NO, which has C_s symmetry (Figure 6.9). The unpaired electron can either reside in an MO dominated by an oxygen p orbital that is of a″ symmetry, or in an MO having π^*_{NO} character that is of a′ symmetry. These two electronic states are fundamentally different. The symmetry of a doublet electronic state is simply the symmetry of the half-filled orbital if all other orbitals are doubly occupied, so we would refer to the two possible electronic states here as ^{2}A″ and ^{2}A′, respectively. When symmetry is imposed, we will have a block diagonal Fock matrix and the unpaired electron will appear in either the a′ block or the a″ block, depending on the initial guess. Once placed there, most SCF convergence procedures will not provide any means for the electronic state symmetry to change, i.e., if the initial guess is a ^{2}A′ wave function, then the calculation will proceed for that state, and if the initial guess is a ^{2}A″ wave function, then it will instead be that state that is optimized. The two states both exist, but one is the ground state and the other an excited state, and one must take care to ensure that one is not working with the undesired state.

Typically, one can assess the nature of the state (ground vs. excited) after convergence of the wave function. Continuing with our example, let us say that we have optimized the ^{2}A′ state. We can then take that wave function, alter it so that the occupation number of the

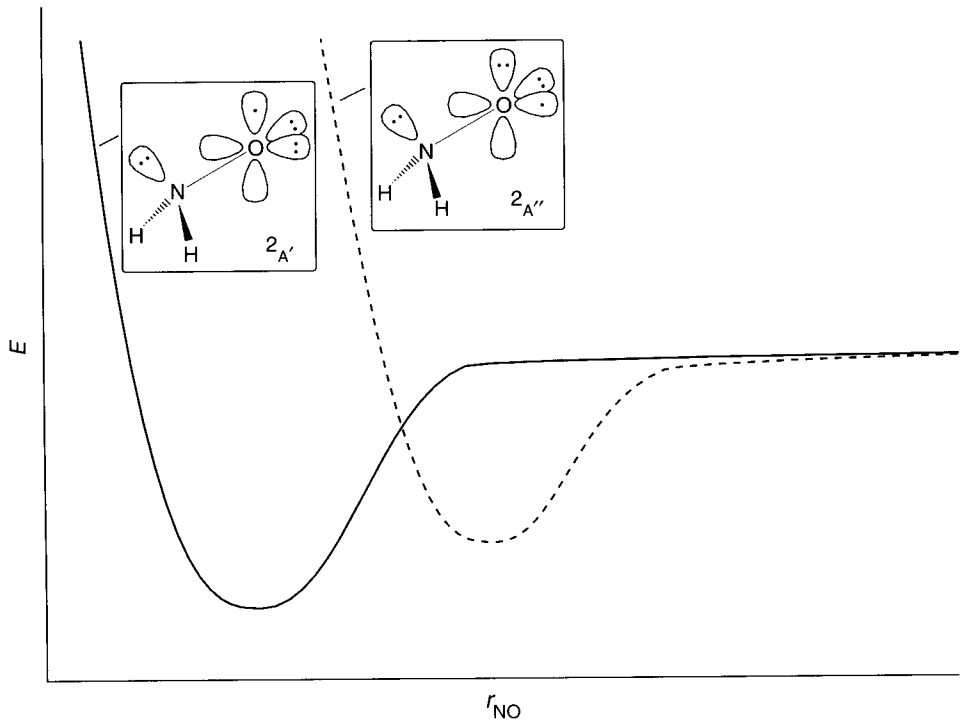


Figure 6.9 Curve crossing of the two lowest energy doublets for C_s $H_2NO^{\bullet}$. The question of which one is the ground state depends on the NO bond length. One can thus be misled in looking for the lowest energy structure if one fails to optimize the geometry for each

highest occupied a′ orbital is zero instead of one, and the occupation of the lowest formerly unoccupied a″ orbital is one instead of zero (i.e., construct a ^{2}A″ wave function using the MOs of the ^{2}A′ wave function) and carry out an SCF calculation using this construction as the initial guess. If the energy drops relative to the first wave function, then the first was an excited state. Many electronic structure programs offer the option to do this in a systematic fashion, i.e., to consider *every* possible switch of an electron from one orbital to another (such a calculation is really a CIS calculation, see Section 14.2.2). Note that there can be some challenging subtleties in working with systems where many states are close to one another in energy. For instance, it could occur in the nitroxyl example above that the two electronic state PESs cross one another in such a fashion that each of the two states is the ground state *at its respective optimized geometry*. Such a situation can only be determined by a fairly careful analysis of the PESs for both states. Note that failing to impose symmetry on the system does not in any way alleviate this problem. Instead, it obscures it, since no symmetry labels can be applied to the orbitals and thus, in the absence of visualization of the half-filled orbital, there is no simple means to differentiate between the two states.

Note that the problem just discussed above is rarely encountered for closed-shell singlets. That is because any excitation from an orbital of one symmetry type to one of a different symmetry type must be a *double* excitation if the closed-shell character of the wave function is to be preserved. Typically the difference in energy between these two possible configurations is so large that no reasonable means for guessing the initial wave function generates the higher energy possibility. This is one of the advantages of closed-shell states compared to open-shell ones. Certain other aspects of dealing with open-shell systems also merit attention.

6.3.3 Open-shell Systems

The presentation of the HF equations in Chapter 4 assumed a closed-shell singlet for simplicity, but what if there are one or more singly occupied orbitals? Let us proceed with an example to guide the discussion, in this case, the methyl radical, which is planar in its equilibrium structure (Figure 6.10). The most intuitive description of the wave function for this system (ignoring symmetry for ease of discussion) would be

$$^2\Psi = \left|\mathrm{C1s}^2\sigma^2_{\mathrm{CH_a}}\sigma^2_{\mathrm{CH_b}}\sigma^2_{\mathrm{CH_c}}\mathrm{C2p}^1_z\right\rangle \tag{6.6}$$

Thus, there is a doubly occupied carbon 1s core, three C–H bonding orbitals, and the unpaired electron in a carbon 2p orbital. Given this configuration, it might seem natural to envision an extension of HF theory where all of the orbitals continue to be evaluated using essentially the restricted formalism (RHF) for closed-shell systems, but the density matrix elements for the singly occupied orbital(s) are not multiplied by the factor of two appearing in Eq. (4.57). In essence, this describes so-called restricted open-shell HF theory (ROHF). In its completely general form, certain complications arise for systems whose descriptions require more than a single determinant (i.e., unlike Eq. (6.6)), so we will not extend this qualitative description of the nature of the theory to specific equations (such details are

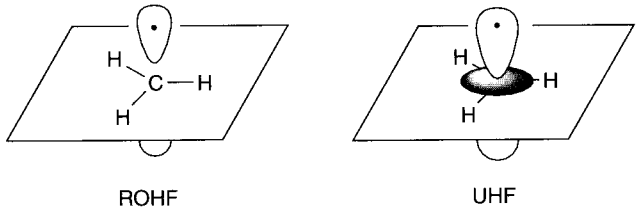


Figure 6.10 In the absence of spin polarization, which corresponds to the ROHF picture, there is zero spin density in the plane containing the atoms of the methyl radical. Accounting for spin polarization, which corresponds to the UHF picture, results in a build-up of negative spin density (represented as a shaded region) in the same plane

available in Veillard (1975)). It suffices to note that most electronic structure packages offer ROHF as an option for open-shell calculations.

Besides being intuitively satisfying, ROHF theory produces wave functions that are eigenfunctions of the operator S^2 (just as the true wave function must be), having eigenvalues $S(S+1)$ where S is the magnitude of the vector sum of the spin magnetic moments for all of the unpaired electrons. However, ROHF theory fails to account for spin polarization in the doubly occupied orbitals. To appreciate this point, let us return to the methyl radical. Note that because the unpaired spin in this molecule is in the carbon $2p_z$ orbital, the plane containing the atoms, which is the nodal plane for the $2p_z$ orbital, must have zero spin density. This being the case, electron spin resonance experiments should detect zero hyperfine coupling between the magnetic moments of the hydrogen atoms (or of a ^{13}C nucleus) and the unpaired electron. However, even after correcting for the effects of molecular vibrations, it is clear that there *is* a coupling between the two.

Spin density is found in the molecular plane because of spin polarization, which is an effect arising from exchange correlation. The Fermi hole that surrounds the unpaired electron allows other electrons of the same spin to localize above and below the molecular plane slightly more than can electrons of opposite spin. Thus, if the unpaired electron is α, we would expect there to be a slight excess of β density in the molecular plane; as a result, the ^1H hyperfine splitting should be negative (see Section 9.1.3), and this is indeed the situation observed experimentally. An ROHF wave function, because it requires the spatial distribution of both spins in the doubly occupied orbitals to be identical, cannot represent this physically realistic situation.

To permit the α and β spins to occupy different regions of space, it is necessary to treat them individually in the construction of the molecular orbitals. Following this formalism, we would rewrite our methyl radical wave function Eq. (6.6) as

$$^2\Psi = \left|C1s^{\alpha}C1s'^{\beta}\sigma^{\alpha}_{CH_a}\sigma'^{\beta}_{CH_a}\sigma^{\alpha}_{CH_b}\sigma'^{\beta}_{CH_b}\sigma^{\alpha}_{CH_c}\sigma'^{\beta}_{CH_c}C2p^{\alpha}_{z}\right\rangle \tag{6.7}$$

where the prime notation on each β orbital emphasizes that while it may be spatially similar to the analogous α orbital, it need not be identical. The individual orbitals are found by carrying out separate HF calculations for each spin, with the spin-specific Fock operator

now defined as

$$F_{\mu\nu}^{\xi} = \left\langle \mu \left| -\frac{1}{2}\nabla^2 \right| \nu \right\rangle - \sum_{k}^{\text{nuclei}} Z_k \left\langle \mu \left| \frac{1}{r_k} \right| \nu \right\rangle + \sum_{\lambda\sigma} \left[\left(P_{\lambda\sigma}^{\alpha} + P_{\lambda\sigma}^{\beta} \right) (\mu\nu|\lambda\sigma) - P_{\lambda\sigma}^{\xi} (\mu\lambda|\nu\sigma) \right] \tag{6.8}$$

where ξ is either α or β, and the two spin-density matrices are defined as

$$P_{\lambda\sigma}^{\xi} = \sum_{i}^{\xi-\text{occupied}} a_{\lambda i}^{\xi} a_{\sigma i}^{\xi} \tag{6.9}$$

where the coefficients a are the usual ones expressing the MOs in the AO basis, but there are separate sets for the α and β orbitals. Notice, then, that in Eq. (6.8), the Coulomb repulsion (the first set of integrals in the double sum) is calculated with both spins, but exchange (the second set of integrals) is calculated only with identical spins. Because the SCF is being carried out separately for each spin, the two density matrices can differ, which is to say the MOs can be different, and this permits spin polarization. Equations (6.8) and (6.9) define unrestricted Hartree-Fock (UHF) theory.

While UHF wave functions have the desirable feature of including spin polarization, they are *not*, in general, eigenfunctions of S^2. By allowing the spatial parts of the different spin orbitals to differ, the final UHF wave function incorporates some degree of 'contamination' from higher spin states – specifically, states whose high-spin components would derive from flipping the spin of one or more electrons. Thus, doublets are contaminated by quartets, sextets, octets, etc., while triplets are contaminated by pentets, heptets, nonets, etc. The degree of spin contamination can be assessed by inspection of $\langle S^2 \rangle$, which should be 0.0 for a singlet, 0.75 for a doublet, 2.00 for a triplet, 3.75 for a quartet, etc. Values that vary from these proper eigenvalues by more than 5 percent or so should inspire great caution in working with the wave function, since other expectation values will also be skewed by differences between the property for the desired state and those for the contaminating states (see Section 9.1.4 and Appendix C for details on the calculation of $\langle S^2 \rangle$).

Various techniques have been developed to reduce or eliminate the contribution of contaminating states to the UHF wave function or expectation values derived from it. Some of these are described in Appendix C, which contains a more detailed description of spin algebra in general. In general, however, it should be noted that none of these approaches are convenient for geometry optimization, which makes characterization of an open-shell PES quite difficult when spin contamination effects are large. Thus, open-shell systems nearly always require more care than closed-shell singlets, because both the ROHF and the UHF formalisms are subject to intrinsically unphysical behavior. Depending on the nature of the system and the properties being calculated, such behavior may or may not be manifest.

Finally, note that some open-shell systems cannot be described by a single determinant. The classical example is an open-shell singlet, i.e., a system having electrons of α and β spin

in different spatial orbitals a and b. The wave function for such a system that is properly antisymmetric and preserves the indistinguishability of particles is

$$^1\Psi = \tfrac{1}{2}[a(1)b(2) + a(2)b(1)][\alpha(1)\beta(2) - \alpha(2)\beta(1)] \tag{6.10}$$

which *cannot* be expressed as a single determinant. Because RHF and UHF are defined to use single-determinantal wave functions, they are formally unable to address this wave function (cf. Appendix C). In its most general form, ROHF *is* defined for multideterminantal systems, but the more typical approach is to use multiconfiguration self-consistent field theory, as described in Section 7.2.

6.3.4 Efficiency of Implementation and Use

We have emphasized up to this point the formal N^4 scaling of HF theory. However, in practice, the situation is never so severe, and indeed linear scaling HF implementations have begun to appear. Of course, one should remember that scaling behavior is different from speed. Thus, for a system of a given size, a HF calculation using algorithms that scale linearly may take significantly longer than conventional algorithms – it is simply true that at *some* point the linear scaling algorithm will become more efficient given a system of large enough size. In any case, because there are several features present in electronic structure programs that allow some control over the efficiency of the calculation, we discuss here the most common ones, recapitulating a few that have already been mentioned above.

First, there is the issue of how to go about computing the four-index integrals, which are responsible for the formal N^4 scaling. One might imagine that the most straightforward approach is to compute every single one and, as it is computed, write it to storage – then, as the Fock matrix is assembled element by element, call back the computed values whenever they are required (most of the integrals are required several times). In practice, however, this approach is only useful when the time required to write to and read from storage is very, very fast. Otherwise, modern processors can actually recompute the integral from scratch faster than modern hardware can recover the previously computed value from, say, disk storage. The process of computing each integral as it is needed rather than trying to store them all is called 'direct SCF'. Only when the storage of all of the integrals can be accomplished in memory itself (i.e., not on an external storage device) is the access time sufficiently fast that the 'traditional' method is to be preferred over direct SCF.

As the size of the system increases, it becomes possible to take advantage of other features of the electronic structure that further improve the efficiency of direct SCF. For instance, it is possible to estimate upper bounds for four-index integrals reasonably efficiently, and if the upper bound is so small that the integral can make no significant contribution, there is no point evaluating it more accurately than assigning it to be zero. Such small integrals are legion in large systems, since if each of the four basis functions is distantly separated from all of the others simple overlap arguments make it clear that the integral cannot be very large.

With very, very large systems, fast-multipole methods analogous to those described in Section 2.4.2 can be used to reduce the scaling of Coulomb integral evaluation to linear

(see, for instance, Strain, Scuseria, and Frisch 1996; Challacombe and Schwegler 1997), and linear methods to evaluate the exchange integrals have also been promulgated (Ochensfeld, White, and Head-Gordon 1998). At this point, the bottleneck in HF calculations becomes diagonalization of the Fock matrix (a step having formal N^3 scaling), and early efforts to reduce the scaling of this step have also appeared (Millam and Scuseria 1997).

As already described above, efficiency in converging the SCF for systems with large basis sets can be enhanced by using as an initial guess the converged wave function from a different calculation, one using either a smaller basis set or a less negative charge. This same philosophy can be applied to geometry optimization, which can be quite time-consuming for very large calculations. It is often very helpful to optimize the geometry first at a more efficient level of theory. This is true not just because the geometry optimized with the lower level is probably a good place to start for the higher level, but also because typically one can compute the force constants at the lower level and use them as an initial guess for the higher level Hessian matrix that will be much better than the typical guess generated by the optimizing algorithm. As described in Section 2.4.1, the availability of a good Hessian matrix can make an enormous amount of difference in how quickly a geometry optimization can be induced to converge.

Also as already noted above, taking advantage of molecular symmetry can provide very large savings in time. However, structures optimized under the constraints of symmetry should always be checked by computation of force constants to verify their nature as stationary points on the full PES. Additionally, it is typically worthwhile to verify that open-shell wave functions obtained for symmetric molecules are stable with respect to orbital changes that would generate other electronic states.

Finally, the use of ECP basis sets for heavy elements improves efficiency by reducing the scale of the electronic structure problem. In addition, relativistic effects can be accounted for by construction of the pseudopotential.

6.4 General Performance Overview of *Ab Initio* HF Theory

6.4.1 Energetics

Because HF theory ignores correlation, and because in its *ab initio* (as opposed to semiempirical) formulation, no attempt is made to correct for this deficiency, HF theory cannot realistically be used to compute heats of formation. Indeed, Feller and Peterson (1998) examined the atomization energies of 66 small molecules at the HF level using the aug-cc-pVnZ basis sets with $n = D$, T, and Q, and obtained mean unsigned errors of 85, 66, and 62 kcal mol^{-1}, respectively. Thus, even as one approaches the HF limit, the intrinsic error in an absolute molecular energy calculation can be very large. In general, the energy associated with *any* process involving a change in the total number of paired electrons is very poorly predicted at the HF level because of the failure to account for electron correlation (cf. use of isodesmic reactions as described in Section 10.6).

Even if the number of paired electrons remains constant but the nature of the bonds is substantially changed, the HF level can show rather large errors. For instance, the atmospheric reaction converting CO and $HO^{\bullet}$ to $H^{\bullet}$ and CO_2 is known to be exoergic with an energy

change of about −23 kcal mol^{-1}. The HF level of theory using the STO-3G, 3-21G, 6-31G(d,p), and near-infinite quality basis sets predicts energy changes of 34.1, 3.1, −5.8, and −7.6 kcal mol^{-1}, respectively, which is quite far from accurate.

Note that isomerization is a process that can change bonding substantially as well. Hehre *et al.* have compared experimental data for 35 isomerization reactions to predictions from the HF/STO-3G, HF/3-21G, and HF/6-31G(d)//HF/3-21G levels (the latter notation, $w/x//y/z$, implies level of theory w with basis set x applied using a geometry optimized at level of theory y using basis set z, i.e., a single point calculation). The isomerizations were quite diverse, including for example acetone to methyl vinyl ether, acetaldehyde to oxetane, formamide to nitrosomethane, and ethanethiol to dimethyl sulfide. The energy differences spanned from 0.2 to 62.6 kcal mol^{-1}, and the dispersion in the data (i.e., the mean absolute error that would be generated by simply guessing the average energy difference over all reactions) was 12.7 kcal mol^{-1}. The mean unsigned errors for the above noted levels were 12.3, 4.8, and 3.2 kcal mol^{-1}, respectively. Thus, the minimal basis set does very badly, HF/3-21G is perhaps qualitatively useful, and the final method has some utility that might well be improved had geometry optimization taken place at the same level as the energy evaluation. Note however that the maximum errors were 51.2, 22.6, and 11.3 kcal mol^{-1}, respectively, which emphasizes that average performances are no guarantee of good behavior in any one system. For comparison, on a subset of 30 of the same isomerizations, MNDO, AM1, and PM3 had mean unsigned errors of 9.1, 7.4, and 5.8 kcal mol^{-1}, and maximal errors of 42, 24, and 23 kcal mol^{-1}, respectively.

The situation continues to improve when the changes in bonding are reduced to those associated with conformational changes. St.-Amant, Cornell, and Kollman (1995) examined 35 different conformational energy differences in a variety of primarily organic molecules, where the average difference in energy between conformers was 1.6 kcal mol^{-1}; at the HF/6-31+G(d,p)//HF/6-31G(d) level, the RMS error in predicted differences was 0.6 kcal mol^{-1}. A similar study on a set of eight organic molecules with an average conformational energy difference of 2.3 kcal mol^{-1} has been reported by Hehre; in that instance, mean unsigned errors of 1.0 and 0.7 kcal mol^{-1} were observed at the HF/3-21G(*) and HF/6-31G(d) levels, respectively (Hehre 1995).

Returning to the 11 glucose conformers already discussed in Chapters 2 and 5 in the context of molecular mechanics and semiempirical models, the performance of several levels of HF theory for predicting the relative conformer energies are listed in Table 6.2. The mean unsigned error associated with assuming all conformers to have the average energy is 1.2 kcal mol^{-1}, so the best HF models do very well by comparison. Note, however, that the small basis sets STO-3G and 3-21G do rather badly. Analysis suggests that these basis sets, which lack polarization functions on heteroatoms, significantly overestimate the energy of hydrogen bonds. Since the glucose conformers are characterized by differing numbers of intramolecular hydrogen bonds, this effect significantly increases the error for these small basis sets. Note also the interesting feature that the polarized double-ζ basis sets 6-31G(d) and cc-pVDZ provide better predictive accuracy than the more complete cc-pVTZ and cc-pVQZ sets. Such a situation is by no means unusual – it is often the case that basis set incompleteness and failure to account for electron correlation introduce errors of opposite

Table 6.2 Mean unsigned errors (kcal mol^{-1}) in 11 predicted glucose conformational energies for various basis sets at the HF level in order of basis set size

Basis set	Mean unsigned error
STO-3G	1.1
3-21G	2.0
6-31G(d)	0.2
cc-pVDZ	0.1
cc-pVTZ	0.6
cc-pVQZ	0.8

sign. If those errors are also of similar *magnitude*, then fortuitously good results can be obtained. A great deal of experience suggests that, very broadly speaking, polarized double-ζ basis sets are the ones most likely to enjoy such a favorable cancellation of errors at the HF level when it occurs. However, it is very risky to *rely* on this phenomenon for any particular calculation *in the absence of prior evidence that it is operative in one or more closely related systems.*

Among the simplest of conformational changes is that associated with rotation about a single bond. Given that this process involves very small changes in bonding, electron correlation effects on the rotation barrier are expected to be small, and indeed, even HF theory with very small basis sets for the most part performs adequately in the prediction of such barriers. For eight rotations about $H_mX\text{-}YH_n$ single bonds, where X,Y = {B, C, N, O, Si, S, P}, Hehre *et al.* found mean unsigned errors of 0.6, 0.6, and 0.3 kcal mol^{-1} at the HF/STO-3G, HF/3-21G(*) and HF/6-31G(d) levels, respectively. If $H_3B{-}NH_3$ is removed from the set, the error drops to 0.5, 0.2, and 0.2 kcal mol^{-1}, respectively. The dispersion in the data set was 0.6 kcal mol^{-1}.

Although HF theory fares poorly in computing most reaction energies, because of the substantial electron correlation effects associated with making/breaking bonds, it is reasonably robust for predicting protonation/deprotonation energies. Since the proton carries with it no electrons, one may think of these reactions as being considerably less sensitive to differential electron correlation in reactants and products. Provided basis sets of polarized valence-double-ζ quality or better are used, absolute proton affinities of neutral molecules are typically computed to an accuracy of better than 5 percent. Errors increase, however, if the cations are non-classical (e.g., bridging protons are present) since such structures tend to be found as minima only after accounting for electron correlation effects. Deprotonation energies of neutral compounds are computed with similar absolute accuracy ($\pm$8 kcal/mol or so) so long as diffuse functions are included in the basis set to balance the description of the anion. If smaller basis sets are used, very large errors are observed.

Another fairly conservative 'reaction' is the removal or attachment of a single electron from/to a molecule. As already discussed in Chapter 5, Koopmans' theorem equates the energy of the HOMO with the negative of the IP. This approximation ignores the effect of electronic relaxation in the ionized product, i.e., the degree to which the remaining electrons redistribute themselves following the detachment of one from the HOMO. If we were to

calculate the IP as the difference in HF energies for the closed-shell neutral and the open-shell product, we would obtain the so-called ΔSCF IP

$$\mathrm{IP}_{\Delta\mathrm{SCF}} = E_{\mathrm{HF}}(\mathrm{A}^{+\bullet}) - E_{\mathrm{HF}}(\mathrm{A}) \tag{6.11}$$

where orbital relaxation *is* included. Including relaxation results in a smaller predicted IP, since relaxation lowers the energy of the cation radical relative to the neutral (the HOMO energy used in Koopmans' theorem derives from orbitals already fully relaxed for the neutral). Note, however, that the neutral species has one more electron than the radical cation, and thus there will be larger electron correlation effects. By ignoring these effects through the use of HF theory, we destabilize the neutral more than the radical cation, and too small an IP is expected in any case. Thus, Koopmans' theorem benefits from a cancellation of errors: the orbital relaxation and the electron correlation effects offset one another. In practice, the cancellation can be remarkably good; Koopmans' theorem IPs are often within 0.3 eV or so of experiment provided basis sets of polarized valence-double-ζ quality or better are used in the HF calculation. However, this favorable cancellation begins to break down if IPs are computed for orbitals other then the HOMO. As more tightly held electrons are ionized, particularly core electrons, the relaxation effects are much larger than the correlation effects, and Koopmans' approximation should not be used.

Koopmans' theorem can be formally applied to electron affinities (EAs) as well, i.e., the EA can be taken to be the orbital energy of the lowest unoccupied (virtual) orbital. Here, however, relaxation effects and correlation effects both favor the radical anion, so rather than canceling, the errors are additive, and Koopmans' theorem estimates will almost always underestimate the EA. It is thus generally a better idea to compute EAs from a ΔSCF approach whenever possible.

A key point meriting discussion is the use of HF theory to model systems where two or more molecules are in contact, held together by non-bonded interactions. Such interactions in actual physical systems include electrostatic interactions between permanent and induced charge distributions, dispersion, and hydrogen bonding (the latter includes both of the prior two in addition to some possible degree of covalent interaction). It is important to note that HF theory is formally incapable of modeling dispersion, because this phenomenon is entirely a consequence of electron correlation, for which HF theory fails to account. Nevertheless, bimolecular interaction energies are often reasonably well predicted by HF theory, particularly with basis sets like 6-31G(d) and others of similar size. As might be expected based on preceding discussion, this again reflects a cancellation of errors.

Clearly, failure to account for dispersion would be expected to strongly reduce intermolecular interactions, so the remaining errors must be in the direction of overbinding. In this instance, there are two chief contributors to overbinding. The first is that, as noted in Section 6.4.3, HF charge distributions tend to be overpolarized, which gives rise to electrostatic interactions that are somewhat too large. The second effect is more technical in nature, and is referred to as 'basis set superposition error' (BSSE). If we consider a bimolecular interaction, the HF interaction energy can be trivially defined as

$$\Delta E_{\mathrm{bind}} = E_{\mathrm{HF}}^{\mathrm{a}\cup\mathrm{b}}(\mathrm{A}\bullet\mathrm{B}) - E_{\mathrm{HF}}^{\mathrm{a}}(\mathrm{A}) - E_{\mathrm{HF}}^{\mathrm{b}}(\mathrm{B}) \tag{6.12}$$

where a and b are the basis functions associated with molecules A and B, respectively. Note that if a and b are not both infinite basis sets, then there are more basis functions employed in the calculation of the complex than in either of the monomers. The greater flexibility of the basis set for the complex can provide an artifactual lowering of the energy when one of the monomers 'borrows' basis functions of the other to improve its own wave function.

One method proposed to correct for this phenomenon is the so-called counterpoise (CP) correction. Although some variations exist, one popular approach defines the CP corrected interaction energy as

$$\begin{aligned}\Delta E_{\text{bind}}^{\text{CP}} = {} & E_{\text{HF}}^{\text{a}\cup\text{b}}(\text{A}\bullet\text{B})_{\text{A}\bullet\text{B}} - E_{\text{HF}}^{\text{a}\cup\text{b}}(\text{A})_{\text{A}\bullet\text{B}} - E_{\text{HF}}^{\text{a}\cup\text{b}}(\text{B})_{\text{A}\bullet\text{B}} \\ & + [E_{\text{HF}}^{\text{a}}(\text{A})_{\text{A}\bullet\text{B}} - \text{E}_{\text{HF}}^{\text{a}}(\text{A})_{\text{A}}] + [E_{\text{HF}}^{\text{b}}(\text{B})_{\text{A}\bullet\text{B}} - E_{\text{HF}}^{\text{b}}(\text{B})_{\text{B}}]\end{aligned} \tag{6.13}$$

where the subscripts appearing after the molecular species describe the geometry employed. Thus, in the first line on the r.h.s., the energy of bringing the two monomers together, each monomer already having the geometry it has in the complex, is computed using a consistent basis set. Thus, in the monomer calculations, basis functions for the missing partner are included in the calculation, even though the nuclei on which those functions are centered are not actually there – such basis functions are sometimes called ghost functions. Since the ghost functions slightly lower the energies of the monomers, the overall binding energy is less than would be the case if they were not to be used. The second line on the r.h.s. of Eq. (6.13) then accounts for the energy required to distort each monomer from its preferred equilibrium structure to the structure found in the complex. Since it is not obvious where to put the ghost functions when the monomer adopts its equilibrium geometry, the geometry-distortion energies are computed using only the nuclei-centered monomer basis sets.

However, it must be noted that the borrowing of basis functions is only partly a mathematical artifact. To the extent that some charge transfer and charge polarization take place as part of forming the bimolecular complex, some of the borrowing simply reflects chemical reality. Thus, CP correction always overestimates the BSSE, and there is no clear way to correct for this overestimation. Indeed, Masamura (2001) has found from analysis of ion-hydrate clusters that interaction energies computed with basis sets of augmented-polarized-double-ζ quality or better were in closer agreement with complete basis-set results before CP correction than after. As a result, there tend to be two schools of thought on how best to deal with BSSE. Some researchers prefer to spend the time that would be required for CP correction instead on the evaluation of Eq. (6.12) with a more saturated basis set. Since, in the limit of an infinite basis, Eqs. (6.12) and (6.13) are equivalent, a demonstration of convergence of Eq. (6.12) with respect to basis-set size is a reasonable indication of accuracy, at least at the HF level.

6.4.2 Geometries

Optimization of the molecular geometry at the HF level appears at first sight to be a daunting task because of the difficulty of obtaining analytic derivatives (see Section 2.4.1). To take the first derivative of Eq. (4.54) with respect to the motion of an atom, we can exhaustively apply the chain rule term by term. Thus, we must determine derivatives of basis functions

and operators with respect to a particular coordinate, and this is not so hard, but we also need to know the derivatives of the density matrix elements with respect to atomic motion, and these derivatives are not obvious at all. However, Pulay (1969) discovered an elegant connection between these very complicated derivatives and the much simpler derivatives of the overlap matrix (which depend only on analytically known basis function derivatives). This breakthrough led to rapid developments in computing higher-order derivatives and optimization algorithms, and as a result, HF geometries are now quite efficiently available.

For minimum-energy structures, HF geometries are usually very good when using basis sets of relatively modest size. For basis sets of polarized valence-double-ζ quality, errors in bond lengths between heavy atoms average about 0.03 Å, and between heavy atoms and H about 0.015 Å. Bond angles are predicted to an average accuracy of about 1.5°, and dihedral angles are also generally well predicted, although available experimental data in the gas phase are scarce. Even with the 3-21G(*) basis set, this accuracy is not much degraded.

To the extent that HF theory is in error, it tends to overemphasize occupation of bonding orbitals (see Chapter 7). Thus, errors tend to be in the direction of predicting bonds to be too short, and this effect becomes more pronounced as one proceeds to saturated basis sets; Feller and Peterson (1998) observed predicted geometries at the HF level to *degrade* in quality with increasing basis-set size in the series aug-cc-pV*n*Z using n = D, T, Q. A good example is the case of the monocyclic singlet diradical 1,3-didehydrobenzene (Figure 6.11). RHF theory erroneously predicts this molecule to be bicyclic with a formal single bond between the radical positions.

There are some additional pathological cases that must be borne in mind in evaluating the quality of predicted HF geometries for minima. As already noted, polarization functions are absolutely required for geometric accuracy in systems characterized by hypervalent bonding; failure to include polarization functions on heteroatoms with single lone pairs can also cause them to be insufficiently pyramidalized. Furthermore, in systems crowding many pairs of non-bonding electrons into small regions of space (e.g., the four oxygen lone pairs in a

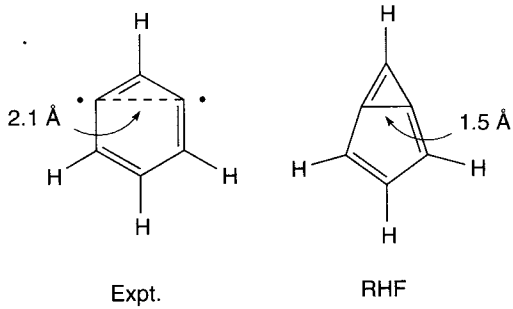


Figure 6.11 Structures of 1,3-didehydrobenzene (*m*-benzyne) from experiment and RHF calculations. Because of its tendency to overemphasize bonding interactions, RHF optimization results in a bicyclic structure. While the RHF error in bond length is very large, it should be noted that the 'bond-stretching' coordinate *is* known to be very flat (for very detailed analyses on the sensitivity of this system to different theoretical levels, see Kraka *et al.* 2001 and Winkler and Sander 2001)

peroxide) electron correlation effects on geometries, ignored by HF theory, can begin to be large, so some caution is warranted here as well. Finally, dative bonds (i.e., those where both electrons in the bonding pair formally come from only one of the atoms) are often poorly described at the HF level. For instance, at the HF/6-31G(d) level, the B–C and B–N distances in the complexes $H_3B{\bullet}CO$ and H_3BNH_3 are predicted to be too long by about 0.1 Å.

Geometries of TS structures are not readily available from experiment, but a fairly substantial body of theoretical work permits comparisons to be made with very high-level calculations. In the case of TS structures, the failure of HF theory to account for electron correlation can be more problematic, since correlation effects in partial bonds can be large. For example, the difference in C–Cl bond lengths predicted at the HF/6-31G(d) and higher levels of theory is more than 0.2 Å for the anti-S_N2' reaction of chloride anion with allyl chloride (Figure 6.12). Although this single example provides an indication of how large differences can be, Wiest, Montiel, and Houk (1997) have analyzed TS structures for many different organic reactions, particularly electrocyclic reactions, and have inferred that in such instances HF/6-31G(d) TS structures are generally of good quality. Nevertheless, the variation in possible bonding situations in TS structures is such that comparison of HF structures with those obtained at better levels of theory is almost always worthwhile in order to ensure quality.

As for the energies of non-bonded complexes, the failure of HF theory to account for dispersion tends to make such complexes too loose in structure, i.e., intermolecular distances are unrealistically large. Hydrogen bonded structures, on the other hand, are often quite good because errors in overestimating electrostatic interactions cancel the failure to account for dispersion. The HF structures show the expected preference for linear bond angles at hydrogen, when such are possible, and further exhibit reasonable distances between donor and acceptor heavy atoms in most instances.

6.4.3 Charge Distributions

HF dipole moments tend to be fairly insensitive to increases in basis-set size beyond valence-double-ζ. With such basis sets, there is a systematic error in dipole moment

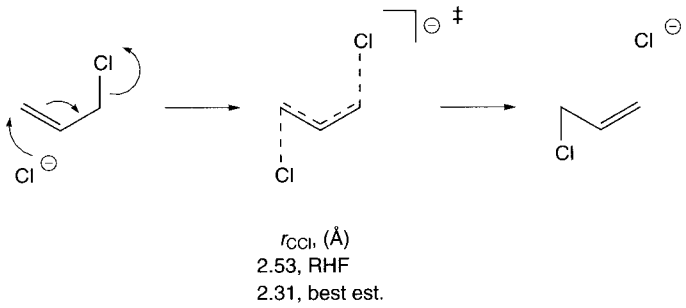


Figure 6.12 The anti-S_N2' reaction of chloride with allyl chloride. While RHF theory does well with the reactant/product geometries, it significantly overestimates the C–Cl bond lengths in the C_2 symmetric TS structure based on calculations at more reliable levels of theory

estimation – typically the magnitude of the dipole is overestimated by 10–25 percent, i.e., molecules are predicted to be too polar. Individual exceptions to this rule exist, of course. In an absolute sense, Scheiner, Baher, and Andzelm (1997) explored the performance of HF/6-31G(d,p) for 108 molecules and obtained a mean unsigned error of 0.23 D.

Results are erratic with smaller basis sets, in part due to lower quality wave functions and in part due to poorer geometries, which affect the dipole moment. An exception is the economical MIDI! basis set for which, as noted above, heteroatom d exponents were specifically optimized so that high-quality geometries and charge distributions (instead of minimal energies) are obtained from the HF wave function. Electrostatic potentials computed with MIDI! also give good agreement with correlated levels of electronic structure theory.

A more complete discussion of charge distributions is deferred until Chapter 9. The performance of HF theory for other molecular properties is also presented in more detail there.

6.5 Case Study: Polymerization of 4-Substituted Aromatic Enynes

Synopsis of Ochiai, Tomita, and Endo (2001) 'Investigation on Radical Polymerization Behavior of 4-Substituted Aromatic Enynes. Experimental, ESR, and Computational Studies'.

One strategy for making highly functionalized polymers is first to carry out polymerization of a system bearing functionalizable appendages, and then after polymerization to react those appendages to introduce new functionality into the polymer. Such an approach can be advantageous in instances where the monomer that would in principle lead directly to the functionalized polymer fails itself to be useful as a polymerization substrate.

Ochiai and co-workers developed an experimental protocol for the radical polymerization of one such reactive monomer, 4-phenylbut-1-en-3-yne. As illustrated in Figure 6.13, this polymerization creates a polyethylene chain functionalized with phenylethynyl substituents.

A factor that affects the kinetics of the polymerization, and, more critically, the utility of the monomer in copolymerizations with other monomers, e.g., methyl methacrylate, is the stability of the radical formed from addition of the growing polymer chain to the vinyl terminus. In order to gauge the stabilizing effect of the phenylethynyl group, and the sensitivity of the stabilization to substitution at the *para* position of the aromatic ring, Ochiai and co-workers carried out calculations at the UHF/3-21G level to evaluate (i) the spin density in the 1-phenylprop-1-yn-3-yl radical and (ii) the reaction energy for the process

$$RCH_3 + CH_3^{\bullet} \rightarrow RCH_2^{\bullet} + CH_4 \qquad (6.14)$$

where R was varied over a number of different functional groups. This so-called isodesmic reaction (see Section 10.4.3) essentially computes the C–H bond energy for the substituted system *relative to* the C–H bond energy in methane, thereby reducing absolute errors that would be associated with a small HF calculation for an absolute bond energy.

The spin density calculation, which analyzes the difference between each atom's Mulliken population (see Section 9.1.3.2) of α and β electrons, indicated the unpaired electron to be highly delocalized, with populations on the *ortho* and *para* carbons of the phenyl

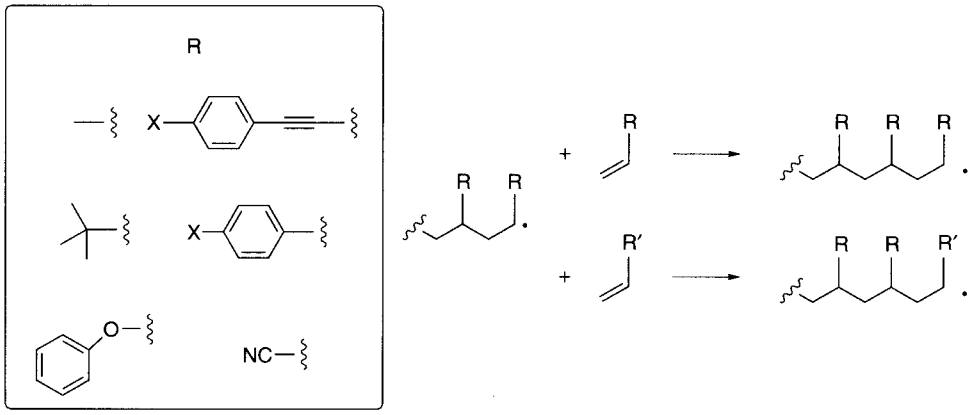


Figure 6.13 Radical polymerization of a growing polymer chain in the presence of two distinct monomers (i.e., copolymerization conditions) can at every step incorporate one monomer or the other. How might one *quantitatively* go about estimating the intrinsic preference for one monomer over the other? What other molecular properties expected to correlate with this discrimination might be subject to computation?

ring nearly equal to that found on the formal radical position (these large positive spin densities were balanced by large negative spin densities on the intervening carbon atoms, which is a typically observed situation). HF theory tends to overpolarize spin, so the magnitude of the spin polarization is probably not trustworthy, but the large degree of delocalization is probably qualitatively reasonable. The prediction that the ring *para* position carries substantial spin was found to be consistent with copolymerization reactivity studies that showed substantial sensitivity to the presence of the *para* substituents MeO, Me, Cl, and CF_3.

One measure of the resonance component of radical stability in polymerizations is the so-called Q value of the monomer, which quantifies the resonance stabilization of the radical (Stevens 1990). However, the experimentally determined value of Q can be influenced by other factors unrelated to resonance. To evaluate the extent to which their measured Q values were consistent with resonance stabilization of the monomer radical, the authors compared isodesmic energies from Eq. (6.14) to measured Q values for R = Me, *t*Bu, PhO, CN, Ph, vinyl, and phenylethynyl. The largest stabilization energy was computed for the R = phenylethynyl case, about 101 kJ mol^{-1}, although at the HF/3-21G level the expected linear correlation between $\log Q$ and stabilization energy was only fair ($R^2 = 0.86$; a better correlation for the non-phenylethynyl substituents had been obtained previously at a higher level of theory).

The authors also considered the relative influence of *para* substitution in the phenylethynyl compared to simply phenyl (i.e., compared to the analogous styrenes). They found that over the four substituents noted above, the stabilization energy from Eq. (6.14) varied by 5.2 kJ mol^{-1} for phenylethynyl and 7.0 kJ mol^{-1} for phenyl. Thus, insertion of the acetylene unit between the radical center and the aromatic ring is predicted to decrease the influence of the aryl substituent by only about 25 percent.

This study employs HF theory to answer only very qualitative questions, which is appropriate given the typically rather poor accuracy of the model in the absence of

accounting for electron correlation. Future use of HF/3-21G to predict Q values for monomers not yet experimentally characterized might be worthwhile, but quantitative differences between monomers should not be taken particularly seriously except to the extent they may be categorized as large, medium, or small.

Bibliography and Suggested Additional Reading

Almlöf, J. 1994. 'Notes on Hartree-Fock Theory and Related Topics' *Lecture Notes in Quantum Chemistry II*, Roos, B. O., Ed., Springer-Verlag: Berlin, 1.

Barrows, S. E., Storer, J. W., Cramer, C. J., French, A. D., and Truhlar, D. G. 1998. 'Factors Controlling the Relative Stability of Anomers and Hydroxymethyl Conformers of Glucopyranose' *J. Comput. Chem.*, **19**, 1111.

Carsky, P. and Urban, M. 1980. *Ab Initio Calculations*, Springer-Verlag: Berlin.

Cramer, C. J. 1991. 'The Fluorophosphoranyl Series: Computational Insights into Relative Stabilities and Localization of Spin' *J. Am. Chem. Soc.*, **113**, 2439.

Cundari, T., Benson, M. T., Lutz, M. L., and Sommerer, S. O. 1996. 'Effective Core Potential Approaches to the Chemistry of the Heavier Elements' in *Reviews in Computational Chemistry*, Vol. 8, Lipkowitz, K. B. and Boyd, D. B., Eds., VCH: New York, 145.

Feller, D. and Davidson, E. R. 1990. 'Basis Sets for *Ab Initio* Molecular Orbital Calculations and Intermolecular Interactions' in *Reviews in Computational Chemistry*, Vol. 1, Lipkowitz, K. B. and Boyd, D. B. Eds., VCH: New York, 1.

Frenking, G., Antes, I., Böhme, M., Dapprich, S., Ehlers, A. W., Jonas, V., Neuhaus, A., Otto, M., Stegmann, R., Veldkamp, A., and Vyboishchikov, S. F. 1996. 'Pseudopotential Calculations of Transition Metal Compounds: Scope and Limitations' in *Reviews in Computational Chemistry*, Vol. 8, Lipkowitz, K. B. and Boyd, D. B. Eds., VCH: New York, 63.

Hehre, W. J., Radom, L., Schleyer, P. v. R., and Pople, J. A. 1986. Ab Initio *Molecular Orbital Theory*, Wiley: New York.

Huzinaga, S., Ed. 1984. *Gaussian Basis Sets for Molecular Calculations*, Elsevier: Amsterdam.

Jensen, F. 1999. *Introduction to Computational Chemistry*, Wiley: Chichester.

Levine, I. N. 2000. *Quantum Chemistry*, 5th Edn., Prentice Hall: New York.

Petersson, G. A. 1998. 'Complete Basis-Set Thermochemistry and Kinetics' in *Computational Thermochemistry*, ACS Symposium Series, Volume 677, Irikura, K. K. and Frurip, D. J., Eds., American Chemical Society: Washington, DC, 237.

Schlegel, H. B. 2000. 'Perspective on "*Ab Initio* Calculation of Force Constants and Equilibrium Geometries in Polyatomic Molecules. I. Theory"' *Theor. Chem. Acc.*, **103**, 294.

Szabo, A. and Ostlund, N. S. 1982. *Modern Quantum Chemistry*, Macmillan: New York.

References

Challacombe, M. and Schwegler, E. 1997. *J. Chem. Phys.*, **106**, 5526.

Dunning, T. H. 1989. *J. Chem. Phys.*, **90**, 1007.

Easton, R. E., Giesen, D. J., Welch, A., Cramer, C. J., and Truhlar, D. G. 1996. *Theor. Chim. Acta*, **93**, 281.

Feller, D. and Peterson, K. A. 1998. *J. Chem. Phys.*, **108**, 154.

Hay, P. J. and Wadt, W. R. 1985. *J. Chem. Phys.*, **82**, 270.

Hehre, W. J. 1995. *Practical Strategies for Electronic Structure Calculations*, Wavefunction: Irvine, CA, 175.

Hehre, W. J., Stewart, R. F., and Pople, J. A. 1969. *J. Chem. Phys.*, **51**, 2657.

Hellmann, H. 1935. *J. Chem. Phys.*, **3**, 61.
Kraka, E., Anglada, J., Hjerpe, A., Filatov, M., and Cremer, D. 2001. *Chem. Phys. Lett.*, **348**, 115.
Krishnan, R., Frisch, M. J., and Pople, J. A. 1980. *J. Chem. Phys.*, **72**, 4244.
Li, J., Cramer, C. J., and Truhlar, D. G. 1998. *Theor. Chem. Acc.*, **99**, 192.
Masamura, M. 2001. *Theor. Chem. Acc.*, **106**, 301.
Millam, J. M. and Scuseria, G. E. 1997. *J. Chem. Phys.*, **106**, 5569.
Ochensfeld, C., White, C. A., and Head-Gordon, M. 1998. *J. Chem. Phys.*, **109**, 1663.
Ochiai, B., Tomita, I., and Endo, T. 2001. *Macromolecules*, **34**, 1634.
Pulay, P. 1969. *Mol. Phys.*, **17**, 197.
Raffenetti, R. C. 1973. *J. Chem. Phys.*, **58**, 4452.
St.-Amant, A., Cornell, W. D., and Kollman, P. A. 1995. *J. Comput. Chem.*, **16**, 1483.
Scheiner, A. C., Baker, J., and Andzelm, J. W. 1997. *J. Comput. Chem.*, **18**, 775.
Stevens, M. P. 1990. *Polymer Chemistry*, 2nd Edn., Oxford University Press: New York, 225.
Stevens, W. J., Krauss, M., Basch, H., and Jasien, P. G. 1992. *Can. J. Chem.*, **70**, 612.
Strain, M. C., Scuseria, G. E., and Frisch, M. J. 1996. *Science*, **271**, 51.
Veillard, A. 1975. In: *Computational Techniques in Quantum Chemistry and Molecular Physics*, NATO ASI Series C, Vol. 15, Diercksen, G. H. F., Sutcliffe, B. T., and Veillard, A., Eds., Reidel: Dordrecht, 201.
Wiest, O., Montiel, D. C., and Houk, K. N. 1997. *J. Phys. Chem. A*, **101**, 8378.
Wilson, A. K., van Mourik, T., and Dunning, T. H. 1996. *J. Mol. Struct.*, **388**, 339.
Winkler, M. and Sander, W. 2001. *J. Phys. Chem. A*, **105**, 10422.
Woon, D. and Dunning, T. H. 1993. *J. Chem. Phys.*, **98**, 1358.
Woon, D. and Dunning, T. H. 1995. *J. Chem. Phys.*, **103**, 4572.

7

Including Electron Correlation in Molecular Orbital Theory

7.1 Dynamical vs. Non-dynamical Electron Correlation

Hartree–Fock theory makes the fundamental approximation that each electron moves in the static electric field created by all of the other electrons, and then proceeds to optimize orbitals for all of the electrons in a self-consistent fashion subject to a variational constraint. The resulting wave function, when operated upon by the Hamiltonian, delivers as its expectation value the lowest possible energy for a single-determinantal wave function formed from the chosen basis set.

It is important to note that there is a key distinction between the Hamiltonian operator and the Fock operator. The former operator returns the electronic energy for the *many-electron* system; the latter is really not a single operator, but the set of all of the interdependent *one-electron* operators that are used to find the one-electron MOs from which the HF wave function is constructed as a Slater determinant.

So, the question arises of how we might modify the HF wave function to obtain a lower electronic energy when we operate on that modified wave function with the Hamiltonian. By the variational principle, such a construction would be a more accurate wave function. We cannot do better than the HF wave function with a single determinant, so one obvious choice is to construct a wave function as a linear combination of multiple determinants, i.e.,

$$\Psi = c_0\Psi_{\mathrm{HF}} + c_1\Psi_1 + c_2\Psi_2 + \cdots \tag{7.1}$$

where the coefficients c reflect the weight of each determinant in the expansion and also ensure normalization. For the moment, we will ignore the nature of the determinants, other than the first one, which is the HF determinant. A general expansion does not *have* to include the HF determinant, but since the HF wave function seems to be a reasonable one for many purposes, it is useful to think of it as a leading term in any more complete wave function.

For the majority of the chemical species we have discussed thus far, the chief error in the HF approximation derives from ignoring the correlated motion of each electron with every other. This kind of electron correlation is called 'dynamical correlation' because it refers

to the dynamical character of the electron–electron interactions. Empirically, it is observed that for most systems the HF wave function dominates in the linear combination expressed by Eq. (7.1) (i.e., c_0 is much larger than any other coefficient); even though the correlation energy may be large, it tends to be made up from a sum of individually small contributions from other determinants.

However, in some instances, one or more of these other determinants may have coefficients of similar magnitude to that for the HF wave function. It is easiest to illustrate this by consideration of a specific example. Consider the closed-shell singlet wave function for trimethylenemethane (TMM, Figure 7.1). TMM is a so-called non-Kekulé molecule – in D_{3h} symmetry, it has two degenerate frontier orbitals for which only two electrons are available. Following a molecular analog of Hund's rule, the molecule has a triplet ground state (i.e., the lowest energy state has one spin-aligned electron in each degenerate orbital), but here we are concerned with the closed-shell *singlet*.

If we carry out a restricted HF calculation, one or other of the degenerate frontier pair will be chosen to be occupied, the calculation will optimize the shapes of all of the occupied orbitals, and we will end up with a best possible single-Slater-determinantal wave function formed from those MOs. But it should be fairly obvious that an equally good wave function

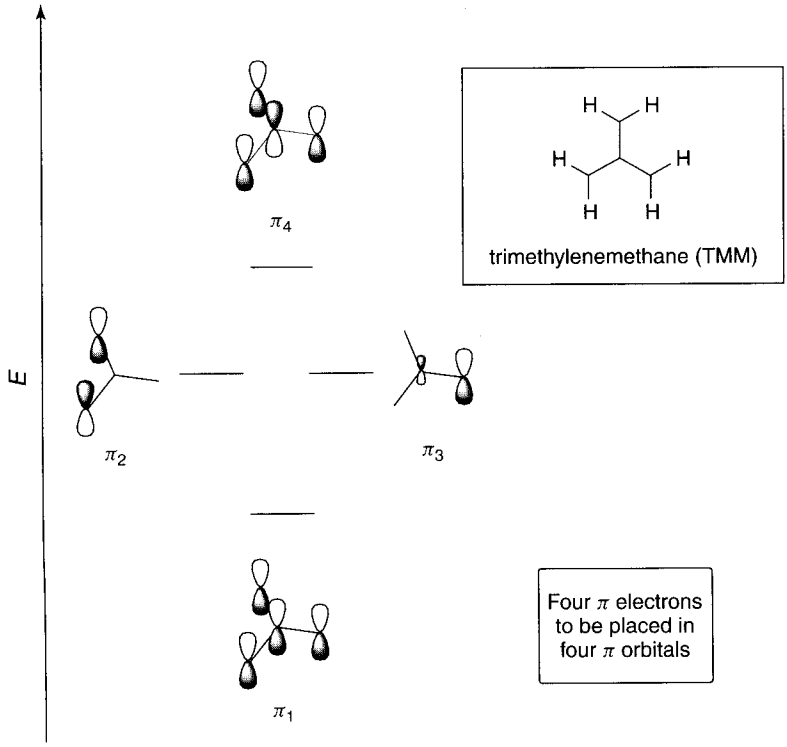


Figure 7.1 The π orbital system of TMM. Orbitals π_2 and π_3 are degenerate when TMM adopts D_{3h} symmetry

might have been formed if the original guess had chosen to populate the *other* of the two degenerate frontier orbitals. Thus, we might expect each of these two different RHF determinants to contribute with roughly equal weight to an expansion of the kind represented by Eq. (7.1). This kind of electron correlation, where different determinants have similar weights because of near (or exact) degeneracy in frontier orbitals, is called 'non-dynamical correlation' to distinguish it from dynamical correlation. This emphasizes that the error here is not so much that the HF approximation ignores the correlated motion of the electrons, but rather that the HF process is constructed in a fashion that is intrinsically single-determinantal, which is insufficiently flexible for some systems.

This chapter begins with a discussion of how to include non-dynamical and dynamical electron correlation into the wave function using a variety of methods. Because the mathematics associated with correlation techniques can be extraordinarily opaque, the discussion is deliberately restricted for the most part to a qualitative level; an exception is Section 7.4.1, where many details of perturbation theory are laid out – those wishing to dispense with those details can skip this subsection without missing too much. Practical issues associated with the employ of particular techniques are discussed subsequently. At the end of the chapter, some of the most modern recipes for accurately and efficiently estimating the exact correlation energy are described, and a particular case study is provided.

7.2 Multiconfiguration Self-Consistent Field Theory

7.2.1 Conceptual Basis

Continuing with our TMM example, let us say that we have carried out an RHF calculation where the frontier orbital that was chosen to be occupied was π_2. The determinant resulting after optimization will be

$$\Psi_{\mathrm{RHF}} = |\cdots \pi_1^2 \pi_2^2 \pi_3^0\rangle \tag{7.2}$$

and orbital π_3 will be empty (i.e., a virtual orbital). We emphasize this by including it in the Slater determinant with an occupation number of zero, although this notation is not standard. We might generate the alternative determinant by keeping the same MOs but simply switching the occupation numbers, i.e.,

$$\Psi_{\pi_2 \to \pi_3} = |\cdots \pi_1^2 \pi_2^0 \pi_3^2\rangle \tag{7.3}$$

An alternative, however, would be to require the RHF calculation to populate π_3 in the initial guess, in which case we would determine

$$\Psi'_{\mathrm{RHF}} = |\cdots \pi'^2_1 \pi'^0_2 \pi'^2_3\rangle \tag{7.4}$$

where the prime on the wave function and orbitals emphasizes that, since different orbitals were occupied during the SCF process, the shapes of *all* orbitals will be different comparing one RHF wave function to the other.

If we were to compare the energies of the wave functions from Eqs. (7.2), (7.3), and (7.4), we would find the energies of the first and third to be considerably lower than that of the second. Since the real system has degenerate frontier orbitals (neglecting Jahn–Teller distortion), it seems reasonable that the energies of wave functions Eq. (7.2) and Eq. (7.4) are similar, but why is the energy of Eq. (7.3) higher? The problem lies in the nature of the SCF process. Only occupied orbitals contribute to the electronic energy – virtual orbitals do not. As such, there is no driving force to optimize the shapes of virtual orbitals; all that is required is that they be orthogonal to the occupied MOs. Thus, the quality of the shape of orbital π_3 depends on whether it is determined as an occupied or a virtual orbital.

From the nature of the system, however, we would really like π_2 and π_3 to be treated *equivalently* during the orbital optimization process. That is, we would like to find the best orbital shapes for these MOs so as to minimize the energy of the *two*-configuration wave function

$$\Psi_{\text{MCSCF}} = a_1|\cdots\pi_1^2\pi_2^2\rangle + a_2|\cdots\pi_1^2\pi_3^2\rangle \tag{7.5}$$

where a_1 and a_2 account for normalization and relative weighting (and we expect them to be equal for D_{3h} TMM). Such a wave function is a so-called 'multiconfiguration self-consistent-field' (MCSCF) one, because the orbitals are optimized for a *combination* of configurations (the particular case where the expansion includes only two configurations is sometimes abbreviated TCSCF).

As a technical point, a 'configuration' or 'configuration state function' (CSF) refers to the molecular spin state and the occupation numbers of the orbitals. For closed-shell singlets, CSFs can always be represented as single determinants, so the terms can be used somewhat loosely. In many open-shell systems, however, proper CSFs can only be represented by a combination of two or more determinants (see Eq. (6.10), for example). MCSCF theory is designed to handle *both* multiple configurations and the possible multi-determinantal character of individual configurations. In that sense, MCSCF is a generalization of ROHF theory, which *can* handle multiple determinants but is *not* capable of handling multiple CSFs.

In general, then, an MCSCF calculation involves a specification of what MOs may be occupied in the CSFs appearing in the expansion of Eq. (7.1). Given that specification, the formalism finds a variational optimum for the shape of each MO (as a linear combination of basis functions) *and* for the weight of each CSF in the MCSCF wave function.

Because a particular 'active' orbital may be occupied by zero, one, or two electrons in any given determinant, these MCSCF orbitals do *not* have unique eigenvalues associated with them, i.e., one cannot discuss the energy of the orbital. Instead, one can describe the 'occupation number' of each such orbital i as

$$(\text{occ. no.})_{i,\text{MCSCF}} = \sum_{n}^{\text{CSFs}} (\text{occ. no.})_{i,n} a_n^2 \tag{7.6}$$

where the sum runs over all CSFs and the occupation number of the orbital in each CSF is multiplied by the percentage contribution of that CSF to the total wave function. Because of the orthogonality of the CSFs, for a normalized MCSCF wave function the sum of the

squares of all CSF coefficients is unity and the percent contribution of any CSF to the wave function is simply its expansion coefficient squared.

MCSCF calculations in practice require *much* more technical expertise than do single-configuration HF analogs. One particularly difficult problem is that spurious minima in coefficient space can often be found, instead of the variational minimum. Thus, convergence criteria are met for the self-consistent field, but the wave function is not really optimized. It usually requires a careful inspection of the orbital shapes and, where available, some data on relative energetics between related species or along a reaction coordinate to ascertain if this has happened.

A different issue requiring careful attention is how to go about selecting the orbitals that should be allowed to be partially occupied, and how to specify the 'flexibility' of the CSF expansion. We turn to this issue next.

7.2.2 Active Space Specification

Selection of orbitals to include in an MCSCF requires first and foremost a consideration of the chemistry being examined. For instance, in the TMM example above, a two-configuration wave function is probably not a very good choice in this system. When the orbitals being considered belong to a π system, it is typically a good idea to include *all* of them, because as a rule they are all fairly close to one another in energy. Thus, a more complete active space for TMM would consider all four π orbitals and the possible ways to distribute the four π electrons within them. MCSCF active space choices are often abbreviated as '(m,n)' where m is the number of electrons and n is the number of orbitals, so this would be a (4,4) calculation.

Sometimes reaction coordinates are studied that involve substantial changes in bonding. In such an instance, it is critical that a consistent choice of orbitals be made. For instance, consider the electrocyclization of 1,3-butadiene to cyclobutene (Figure 7.2). The frontier orbitals of butadiene are those associated with the π system, so, as just discussed, a (4,4) approach seems logical. However, the electrocyclization reaction transforms the two π bonds into one different π bond and one new σ bond. Thus, a consistent (4,4) choice in cyclobutene would involve the π and π^* orbitals and the σ and σ^* orbitals of the new single bond.

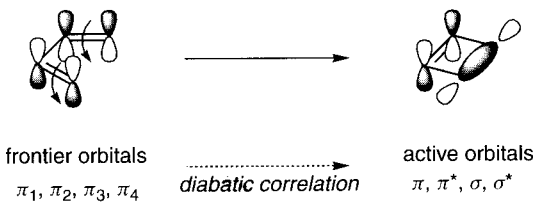


Figure 7.2 The frontier orbitals of s-*cis*-1,3-butadiene are the four π orbitals (π_2 is the specific example shown). If these orbitals are followed in a diabatic sense along the electrocyclization reaction coordinate, they correlate with the indicated orbitals of cyclobutadiene

While these orbitals would be easy to identify in butadiene and cyclobutene, it might be considerably more difficult to choose the corresponding orbitals in a TS structure, where symmetry is lower and mixing of σ and π character might complicate identification.

The next question to consider is how generally to allow the distribution of the electrons in the active space. Returning to TMM, it is clear that we want the CSFs already listed in Eq. (7.5), but in a (4,4) calculation, we might also want to be still more flexible, e.g., considering as the most important four perhaps

$$\begin{aligned}\Psi_{\mathrm{MCSCF}} = a_1|\cdots\pi_1^2\pi_2^2\pi_3^0\pi_4^0\rangle + a_2|\cdots\pi_1^2\pi_2^0\pi_3^2\pi_4^0\rangle \\ + a_3|\cdots\pi_1^2\pi_2^0\pi_3^0\pi_4^2\rangle + a_4|\cdots\pi_1^0\pi_2^2\pi_3^2\pi_4^0\rangle\end{aligned} \tag{7.7}$$

where we have again included orbitals with occupation numbers of zero in the notation for clarity. If we ignore symmetry for the moment, we could also take account of possibly important CSFs having partially occupied orbitals, e.g.,

$$\begin{aligned}\Psi_{\mathrm{MCSCF}} = a_1|\cdots\pi_1^2\pi_2^2\pi_3^0\pi_4^0\rangle + a_2|\cdots\pi_1^2\pi_2^0\pi_3^2\pi_4^0\rangle + a_3|\cdots\pi_1^2\pi_2^0\pi_3^0\pi_4^2\rangle \\ + a_4|\cdots\pi_1^0\pi_2^2\pi_3^2\pi_4^0\rangle + a_5\left(|\cdots\pi_1^2\pi_2^1\overline{\pi}_3^1\pi_4^0\rangle + |\cdots\pi_1^2\overline{\pi}_2^1\pi_3^1\pi_4^0\rangle\right)\end{aligned} \tag{7.8}$$

where the electron in a singly occupied orbital has α spin unless the orbital has a bar over it, in which case it has β spin. Note again that the open-shell singlet appearing after coefficient a_5 requires two determinants to specify.

If we were to try to decide, based on any more or less rational approach, which CSFs to include in some particular expansion along the lines of Eq. (7.8), this would constitute a general MCSCF calculation. However, an alternative to picking and choosing amongst CSFs is simply to include *all* possible configurations in the expansion. In general, the number N of singlet CSFs that can be formed from the distribution of m electrons in n orbitals is determined as

$$N = \frac{n!\,(n+1)!}{\left(\frac{m}{2}\right)!\left(\frac{m}{2}+1\right)!\left(n-\frac{m}{2}\right)!\left(n-\frac{m}{2}+1\right)!} \tag{7.9}$$

In the case of $m = n = 4$, $N = 20$ (it is a mildly diverting exercise to try to generate all 20 by hand). Permitting all possible arrangements of electrons to enter into the MCSCF expansion is typically referred to as having chosen a 'complete active space', and such calculations are said to be of the CASSCF, or just CAS, variety.

Notice that the factorial functions appearing in Eq. (7.9) quickly have daunting consequences. What if we were interested in carrying out a CASSCF calculation on methanol (CH_3OH) including all of the valence electrons in the active space? Such a calculation would be a (14,12) CAS (14 valence electrons, 5 pairs of σ and σ^* orbitals corresponding to the single bonds, and 2 oxygen lone pair orbitals). Neglecting symmetry, the total number of CSFs, from Eq. (7.9), would be 169 884. Recalling that the nature of the MCSCF process is to simultaneously optimize the MO coefficients *and* all of the CSF coefficients, one might imagine that such a calculation would be rather taxing. Indeed, CASSCF calculations on

systems having more than 1 000 000 CSFs are extraordinarily demanding of resources and are rarely undertaken.

Various schemes exist to try to reduce the number of CSFs in the expansion in a rational way. Symmetry can reduce the scope of the problem enormously. In the TMM problem, many of the CSFs having partially occupied orbitals correspond to an electronic state symmetry other than that of the totally symmetric irreducible representation, and thus make no contribution to the closed-shell singlet wave function (if symmetry is not used before the fact, the calculation itself will determine the coefficients of non-contributing CSFs to be zero, but no advantage in efficiency will have been gained). Since this application of group theory involves no approximations, it is one of the best ways to speed up a CAS calculation.

An alternative approach is taken with a formalism known as generalized valence bond (GVB). In a typical CASSCF calculation, one first carries out an HF calculation, and then expresses the CSFs in those orbitals for use in the MCSCF process. To improve convergence, one often undertakes a localization of the canonical HF virtual orbitals (which are otherwise rather diffuse, particularly with large basis sets), so that they are more chemically realistic (see Appendix D for more information on orbital localization schemes). Such a transformation of the orbitals is rigorously permitted and has no effect on wave function expectation values. In the case of GVB, in contrast to CASSCF, not only are the virtual orbitals localized but so too are the occupied orbitals. Thus, to the maximum extent possible, the transformed orbitals look like the canonical bonds, lone pairs, and anti-bonds of valence bond theory, i.e., like Lewis structures. In GVB, only excitations from certain occupied to certain unoccupied orbitals are allowed. For instance, in the so-called perfect-pairing (PP) scheme, the pair of electrons in any bonding orbital is allowed to excite only as a pair, and only into the corresponding antibonding orbital (assuming it is empty). The motivation, then, is to try to capture in an efficient and chemically localized way the most important contributions to non-dynamical correlation. Some mathematical difficulties arise in the GVB scheme because the localized orbitals are not necessarily orthogonal, but the method can be quite fast because of the reduced number of configurations, and one hopes that the retained configurations are the most chemically important ones.

Another means to reduce the scale of the problem is to shrink the size of the CAS calculation, but to allow a limited number of excitations from/to orbitals outside of the CAS space. This secondary space is called a 'restricted active space' (RAS), and usually the excitation level is limited to one or two electrons. Thus, while all possible configurations of electrons in the CAS space are permitted, only a limited number of RAS configurations is possible. Remaining occupied and virtual orbitals, if any, are restricted to occupation numbers of exactly two and zero, respectively.

There is one other step sometimes taken to make the CAS/RAS calculation more efficient, and that is to freeze the shapes of the core orbitals to those determined at the HF level. Thus, there may be four different types of orbitals in a particular MCSCF calculation: frozen orbitals, inactive orbitals, RAS orbitals, and CAS orbitals. Figure 7.3 illustrates the situation in detail. Again, symmetry is the theoretician's friend in keeping the size of the system manageable in favorable cases.

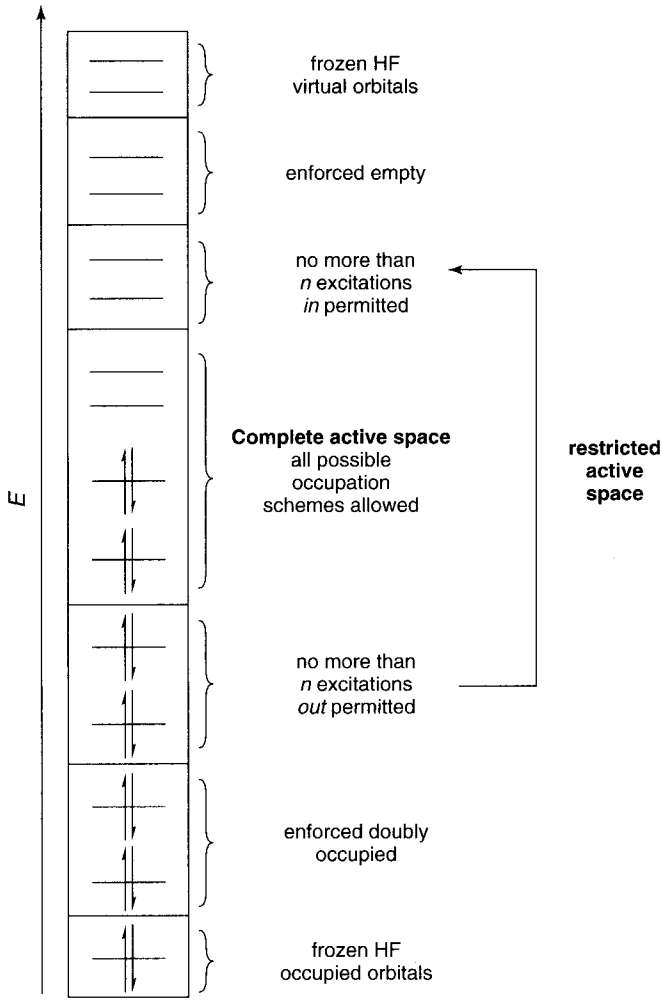


Figure 7.3 Possible assignment of different orbitals in a completely general MCSCF formalism. Frozen orbitals are not permitted to relax from their HF shapes, in addition to having their occupation numbers of zero (virtual) or two (occupied) enforced

While all of the above details are useful for making calculations more efficient, they still are not necessarily very helpful in evaluating just which orbitals should be included in any given space. Typically, a certain amount of trial and error is required in the selection of an active space. After selection of a given active space and convergence of the MCSCF wave function, one should inspect the occupation numbers of the active orbitals. A reasonable rule of thumb is that any orbital having an occupation number greater than 1.98 or less than 0.02 is not important enough to include in the CAS space, and should be removed to avoid instability. In addition, of course, it may be wise to add some orbitals not previously considered to see if *their* occupation numbers justify inclusion in the active space. And,

clearly, if one is considering a reaction coordinate or a series of isomers, the active space must be balanced, so any orbital contributing significantly in one calculation should probably be used in all calculations. While methods to include dynamical correlation after an MCSCF calculation *can* help to make up for a less than optimal choice of active space, it is best not to rely on this phenomenon.

7.2.3 Full Configuration Interaction

Having discussed ways to reduce the scope of the MCSCF problem, it is appropriate to consider the other limiting case. What if we carry out a CASSCF calculation for *all* electrons including *all* orbitals in the complete active space? Such a calculation is called 'full configuration interaction' or 'full CI'. Within the choice of basis set, it is the best possible calculation that can be done, because it considers the contribution of every possible CSF. Thus, a full CI with an infinite basis set is an 'exact' solution of the (non-relativistic, Born–Oppenheimer, time-independent) Schrödinger equation.

Note that no reoptimization of HF orbitals is required, since the set of all possible CSFs is 'complete'. However, that is not much help in a computational efficiency sense, since the number of CSFs in a full CI can be staggeringly large. The trouble is not the number of electrons, which is a constant, but the number of basis functions. Returning to our methanol example above, if we were to use the 6-31G(d) basis set, the total number of basis functions would be 38. Using Eq. (7.9) to determine the number of CSFs in our (14,38) full CI we find that we must optimize 2.4×10^{13} expansion coefficients (!), and this is really a rather small basis set for chemical purposes.

Thus, full CI calculations with large basis sets are usually carried out for only the smallest of molecules (it is partly as a result of such calculations that the relative contributions to basis-set quality of polarization functions vs. decontraction of valence functions, as discussed in Chapter 6, were discovered). In larger systems, the practical restriction to smaller basis sets makes full CI calculations less chemically interesting, but such calculations remain useful to the extent that, as an optimal limit, they permit an evaluation of the quality of other methodologies for including electron correlation using the same basis set. We turn now to a consideration of such other methods.

7.3 Configuration Interaction

7.3.1 Single-determinant Reference

If we consider all possible excited configurations that can be generated from the HF determinant, we have a full CI, but such a calculation is typically too demanding to accomplish. However, just as we reduced the scope of CAS calculations by using RAS spaces, what if we were to reduce the CI problem by allowing only a limited number of excitations? How many should we include? To proceed in evaluating this question, it is helpful to rewrite Eq. (7.1) using a more descriptive notation, i.e.,

$$\Psi = a_0\Psi_{\mathrm{HF}} + \sum_{i}^{\text{occ.}}\sum_{r}^{\text{vir.}} a_i^r \Psi_i^r + \sum_{i<j}^{\text{occ.}}\sum_{r<s}^{\text{vir.}} a_{ij}^{rs}\Psi_{ij}^{rs} + \cdots \tag{7.10}$$

where i and j are occupied MOs in the HF 'reference' wave function, r and s are virtual MOs in Ψ_{HF}, and the additional CSFs appearing in the summations are generated by exciting an electron from the occupied orbital(s) indicated by subscripts into the virtual orbital(s) indicated by superscripts. Thus, the first summation on the r.h.s. of Eq. (7.10) includes all possible single electronic excitations, the second includes all possible double excitations, etc.

If we assume that we do *not* have any problem with non-dynamical correlation, we may assume that there is little need to reoptimize the MOs even if we do not plan to carry out the expansion in Eq. (7.10) to its full CI limit. In that case, the problem is reduced to determining the expansion coefficients for each excited CSF that *is* included. The energies E of N different CI wave functions (i.e., corresponding to different variationally determined sets of coefficients) can be determined from the N roots of the CI secular equation

$$\begin{vmatrix} H_{11} - E & H_{12} & \dots & H_{1N} \\ H_{21} & H_{22} - E & \dots & H_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ H_{N1} & N_{N2} & \dots & H_{NN} - E \end{vmatrix} = 0 \tag{7.11}$$

where

$$H_{mn} = \langle \Psi_m | H | \Psi_n \rangle \tag{7.12}$$

H is the Hamiltonian operator and the numbering of the CSFs is arbitrary, but for convenience we will take $\Psi_1 = \Psi_{HF}$ and then all singly excited determinants, all doubly excited, etc. Solving the secular equation is equivalent to diagonalizing **H**, and permits determination of the CI coefficients associated with each energy. While this is presented without derivation, the formalism is entirely analogous to that used to develop Eq. (4.21).

To solve Eq. (7.11), we need to know how to evaluate matrix elements of the type defined by Eq. (7.12). To simplify matters, we may note that the Hamiltonian operator is composed only of one- and two-electron operators. Thus, if two CSFs differ in their occupied orbitals by 3 or more orbitals, every possible integral over electronic coordinates hiding in the r.h.s. of Eq. (7.12) will include a simple overlap between at least one pair of different, and hence orthogonal, HF orbitals, and the matrix element will necessarily be zero. For the remaining cases of CSFs differing by two, one, and zero orbitals, the so-called Condon–Slater rules, which can be found in most quantum chemistry textbooks, detail how to evaluate Eq. (7.12) in terms of integrals over the one- and two-electron operators in the Hamiltonian and the HF MOs.

A somewhat special case is the matrix element between the HF determinant and a singly excited CSF. The Condon–Slater rules applied to this situation dictate that

$$\begin{aligned} H_{1n} &= \langle \Psi_{HF} | H | \Psi_i^r \rangle \\ &= \langle \phi_r | F | \phi_i \rangle \end{aligned} \tag{7.13}$$

where F is the Fock operator and i and r are the occupied and virtual HF orbitals in the single excitation. Since these orbitals are eigenfunctions of the Fock operator, we have

$$\begin{aligned} \langle \phi_r | F | \phi_i \rangle &= \varepsilon_i \langle \phi_r | \phi_i \rangle \\ &= \varepsilon_i \delta_{ir} \end{aligned} \tag{7.14}$$

where ε_i is the MO eigenvalue. Thus, all matrix elements between the HF determinants and singly excited determinants are zero, since to be singly excited, r must not be equal to i. This result is known as Brillouin's theorem (Brillouin 1934).

It is *not* the case that arbitrary matrix elements between *other* determinants differing by only one occupied orbital are equal to zero. Nevertheless, the Condon–Slater rules and Brillouin's theorem ensure that the CI matrix in a broad sense is reasonably sparse, as illustrated in Figure 7.4. With that in mind, let us return to the question of which excitations to include in a 'non-full' CI. What if we only keep single excitations? In that case, we see from Figure 7.4 that the CI matrix will be block diagonal. One 'block' will be the HF energy, H_{11}, and the other will be the singles/singles region. Since a block diagonal matrix can be fully diagonalized block by block, and since the HF result is already a block by itself,

	Ψ_{HF}	Ψ_i^a	Ψ_{ij}^{ab}	Ψ_{ijk}^{abc}
Ψ_{HF}	E_{HF}	**0**	dense	**0**
Ψ_i^a	**0**	dense	sparse	very sparse
Ψ_{ij}^{ab}	dense	sparse	sparse	extremely sparse
Ψ_{ijk}^{abc}	**0**	very sparse	extremely sparse	extremely sparse

Figure 7.4 Structure of the CI matrix as blocked by classes of determinants. The HF block is the (1,1) position, the matrix elements between the HF and singly excited determinants are zero by Brillouin's theorem, and between the HF and triply excited determinants are zero by the Condon–Slater rules. In a system of reasonable size, remaining regions of the matrix become increasingly sparse, but the number of determinants in each block grows to be extremely large. Thus, the (1,1) eigenvalue is most affected by the doubles, then by the singles, then by the triples, etc

it is apparent that the lowest energy root, i.e., the ground-state HF root, is unaffected by inclusion of single excitations. Indeed, one way to think about the HF process is that it is an optimization of orbitals subject to the constraint that single excitations do not contribute to the wave function. Thus, the so-called CI singles (CIS) method finds no use for ground states, although it can be useful for excited states, as described in Section 14.2.2.

So, we might next consider including only double excitations (CID). It is worthwhile to do a very simple example, such as molecular hydrogen in a minimal basis set. In that case, there are only 2 HF orbitals, the σ and the σ^* orbitals associated with the H–H bond, in which case there is only one doubly excited state, corresponding to $|\sigma^{*2}>$. The CID state energies are found from solving

$$\begin{vmatrix} H_{11} - E & H_{12} \\ H_{21} & H_{22} - E \end{vmatrix} = 0 \tag{7.15}$$

This quadratic equation is simple to solve, and gives root energies

$$E = \frac{1}{2}\left[H_{11} + H_{22} \pm \sqrt{(H_{22} - H_{11})^2 + 4H_{12}}\right] \tag{7.16}$$

The Condon–Slater rules dictate that H_{12} is an electron-repulsion integral, and it thus has a positive sign (it is actually the exchange integral K_{12}). So, examining Eq. (7.16), we see that to the average of the two pure-state energies (ground and doubly excited) we should either add or subtract a value slightly larger than half the difference between the two state energies. Thus, when we subtract, our energy will be below the HF energy, and the difference will be the correlation energy. In the case of H_2 with the STO-3G basis set at a bond distance of 1.4 a.u., $E_{corr} = -0.02056$ a.u., or about 13 kcal/mol.

In bigger systems, this process can be carried out analogously. However, the size of the CI matrix can quickly become very, very large, in which case diagonalization is computationally taxing. More efficient methods than diagonalization exist for finding only one or a few eigenvalues of large matrices. These methods are typically iterative, and most modern electronic structure programs use them in preference to full matrix diagonalization.

What about triple excitations? While there are no non-zero matrix elements between the ground state and triply excited states, the triples do mix with the doubles, and can through them influence the lowest energy eigenvalue. So, there is some motivation for including them. On the other hand, there are a *lot* of triples, making their inclusion difficult in a practical sense. As a result, triples, and higher-level excitations, are usually not accounted for in truncated CI treatments.

Let us return, however, to singly excited determinants. While, like triples, they fail to interact with the ground state (although in this case because of Brillouin's theorem), they too mix with doubles and thus can have *some* influence on the lowest eigenvalue. In this instance, there are sufficiently few singles compared to doubles that it does not make the problem significantly more difficult to include them, and this level of theory is known as CISD.

The scaling for CISD with respect to system size is, in the large basis limit, on the order of N^6. Such scaling behavior is considerably worse than HF, and thus poses a more stringent

limit on the sizes of systems that can be practically addressed. Just as with MCSCF, symmetry can be used to significantly reduce the computational effort by facilitating the evaluation of matrix elements. Similarly, some orbitals can be frozen in the generation of excited states. A popular choice is to leave the core orbitals frozen in CISD.

One of the most appealing features of CISD is that it is variational. Thus, the CISD energy represents an upper bound on the exact energy. However, it has a particularly unattractive feature as well, and that is that it is not 'size consistent'. This property is best explained by example: consider the H_2 molecule case addressed above. We may construct the CID wave function as

$$\Psi_{\mathrm{CID}} = (1 - c)^2 \Psi_{\mathrm{HF}} + c^2 \Psi_{1\bar{1}}^{2\bar{2}} \tag{7.17}$$

where the coefficient c is determined from the diagonalization process. Now, consider the CID wave function for two molecules of H_2 separated by, say, 50 Å. For all practical purposes, there is no chemical interaction between them, so we could take the overall wave function simply to be a properly antisymmetrized product of Eq. (7.17) with itself. This expression would include a term, preceded by the coefficient c^4, corresponding to simultaneous double excitation within each molecule. However, that is a quadruply excited configuration. As such, if we carried out a CID calculation on the two molecules as a single system, it would not be permitted. Thus, twice the CID energy of one molecule of H_2 will be lower than the CID energy for two molecules of H_2 at large separations, which is a vexing result.

Various approaches to overcoming the size extensivity problem have been proposed. Owing to its simplicity, one of the more popular methods is that of Langhoff and Davidson (1974), which estimates the energy associated with the missing quadruple excitations as

$$E_{\mathrm{Q}} = (1 - a_0)^2 (E_{\mathrm{CISD}} - E_{\mathrm{HF}}) \tag{7.18}$$

where a_0 is the coefficient of the HF determinant in the normalized truncated CISD wave function (which itself is Eq. (7.10) without the ellipsis). This is typically abbreviated as CISD(Q). In modern work, there has been a tendency to avoid single-reference CI calculations in favor of other, size-extensive methods for including electron correlation (vide infra).

7.3.2 Multireference

The formalism for multireference configuration interaction (MRCI) is quite similar to that for single-reference CI, except that instead of the HF wave function serving as reference, an MCSCF wave function is used. While it is computationally considerably more difficult to construct the initial MCSCF wave function than a HF wave function, the significant improvement of the virtual orbitals in the former case can make the CI itself more rapidly convergent. Nevertheless, the number of matrix elements requiring evaluation in MRCI calculations is enormous, and they are usually undertaken only for small systems. Typically, MRCI is a useful method to study a large section of a PES, where significant changes in

bonding (and thus correlation energy) are taking place so a sophisticated method is needed to accurately predict dynamical and non-dynamical correlation energies.

As with single-reference CI, most MRCI calculations truncate the CI expansion to include only singles and doubles (MRCISD). An analog of Eq. (7.18) has been proposed to make up for the non-size-extensivity this engenders (Bruna, Peyerimhoff, and Buenker, 1980). MRCISD calculations with large basis sets can be better than similarly expensive full CI calculations with smaller basis sets, illustrating that most of the correlation energy can be captured by including only limited excitations, at least in those systems small enough to permit thorough evaluation.

7.4 Perturbation Theory

7.4.1 General Principles

Often in pseudoeigenvalue equations, the nature of a particular operator makes it difficult to work with. However, it is sometimes worthwhile to create a more tractable operator by removing some particularly unpleasant portion of the original one. Using exact eigenfunctions and eigenvalues of the simplified operator, it is possible to estimate the eigenfunctions and eigenvalues of the more complete operator. Rayleigh–Schrödinger perturbation theory provides a prescription for accomplishing this.

In the general case, we have some operator $\mathbf{A}$ that we can write as

$$\mathbf{A} = \mathbf{A}^{(0)} + \lambda \mathbf{V} \tag{7.19}$$

where $\mathbf{A}^{(0)}$ is an operator for which we can find eigenfunctions, $\mathbf{V}$ is a perturbing operator, and λ is a dimensionless parameter that, as it varies from 0 to 1, maps $\mathbf{A}^{(0)}$ into $\mathbf{A}$. If we expand our ground-state eigenfunctions and eigenvalues as Taylor series in λ, we have

$$\Psi_0 = \Psi_0^{(0)} + \lambda \frac{\partial \Psi_0^{(0)}}{\partial \lambda}\bigg|_{\lambda=0} + \frac{1}{2!}\lambda^2 \frac{\partial^2 \Psi_0^{(0)}}{\partial \lambda^2}\bigg|_{\lambda=0} + \frac{1}{3!}\lambda^3 \frac{\partial^3 \Psi_0^{(0)}}{\partial \lambda^3}\bigg|_{\lambda=0} + \cdots \tag{7.20}$$

and

$$a_0 = a_0^{(0)} + \lambda \frac{\partial a_0^{(0)}}{\partial \lambda}\bigg|_{\lambda=0} + \frac{1}{2!}\lambda^2 \frac{\partial^2 a_0^{(0)}}{\partial \lambda^2}\bigg|_{\lambda=0} + \frac{1}{3!}\lambda^3 \frac{\partial^3 a_0^{(0)}}{\partial \lambda^3}\bigg|_{\lambda=0} + \cdots \tag{7.21}$$

where $a_0^{(0)}$ is the eigenvalue for $\Psi_0^{(0)}$, which is the appropriate normalized ground-state eigenfunction for $\mathbf{A}^{(0)}$. For ease of notation, Eqs. (7.20) and (7.21) are usually written as

$$\Psi_0 = \Psi_0^{(0)} + \lambda \Psi_0^{(1)} + \lambda^2 \Psi_0^{(2)} + \lambda^3 \Psi_0^{(3)} + \cdots \tag{7.22}$$

and

$$a_0 = a_0^{(0)} + \lambda a_0^{(1)} + \lambda^2 a_0^{(2)} + \lambda^3 a_0^{(3)} + \cdots \tag{7.23}$$

where the terms having superscripts (n) are referred to as 'nth-order corrections' to the zeroth order term and are defined by comparison to Eqs. (7.20) and (7.21).

Thus, we may write

$$(\mathbf{A}^{(0)} + \lambda\mathbf{V})|\Psi_0\rangle = a|\Psi_0\rangle \tag{7.24}$$

as

$$\begin{aligned}(\mathbf{A}^{(0)} + \lambda\mathbf{V})|\Psi_0^{(0)} + \lambda\Psi_0^{(1)} + \lambda^2\Psi_0^{(2)} + \lambda^3\Psi_0^{(3)} + \cdots\rangle = \\ (a_0^{(0)} + \lambda a_0^{(1)} + \lambda^2 a_0^{(2)} + \lambda^3 a_0^{(3)} + \cdots)|\Psi_0^{(0)} + \lambda\Psi_0^{(1)} + \lambda^2\Psi_0^{(2)} + \lambda^3\Psi_0^{(3)} + \cdots\rangle\end{aligned} \tag{7.25}$$

Since Eq. (7.25) is valid for any choice of λ between 0 and 1, we can expand the left and right sides and consider only equalities involving like powers of λ. Powers 0 through 3 require

$$\mathbf{A}^{(0)}|\Psi_0^{(0)}\rangle = a_0^{(0)}|\Psi_0^{(0)}\rangle \tag{7.26}$$

$$\mathbf{A}^{(0)}|\Psi_0^{(1)}\rangle + \mathbf{V}|\Psi_0^{(0)}\rangle = a_0^{(0)}|\Psi_0^{(1)}\rangle + a_0^{(1)}|\Psi_0^{(0)}\rangle \tag{7.27}$$

$$\mathbf{A}^{(0)}|\Psi_0^{(2)}\rangle + \mathbf{V}|\Psi_0^{(1)}\rangle = a_0^{(0)}|\Psi_0^{(2)}\rangle + a_0^{(1)}|\Psi_0^{(1)}\rangle + a_0^{(2)}|\Psi_0^{(0)}\rangle \tag{7.28}$$

$$\mathbf{A}^{(0)}|\Psi_0^{(3)}\rangle + \mathbf{V}|\Psi_0^{(2)}\rangle = a_0^{(0)}|\Psi_0^{(3)}\rangle + a_0^{(1)}|\Psi_0^{(2)}\rangle + a_0^{(2)}|\Psi_0^{(1)}\rangle + a_0^{(3)}|\Psi_0^{(0)}\rangle \tag{7.29}$$

where further generalization should be obvious. Our goal, of course, is to determine the various nth-order corrections. Equation (7.26) is the zeroth-order solution from which we are hoping to build, while Eq. (7.27) involves the two unknown first-order corrections to the wave function and eigenvalue.

To proceed, we first impose intermediate normalization of Ψ; that is

$$\langle\Psi_0|\Psi_0^{(0)}\rangle = 1 \tag{7.30}$$

By use of Eq. (7.22) and normalization of $\Psi_0^{(0)}$, it must then be true that

$$\langle\Psi_0^{(n)}|\Psi_0^{(0)}\rangle = \delta_{n0} \tag{7.31}$$

Now, we multiply on the left by $\Psi_0^{(0)}$ and integrate to solve Eqs. (7.27)–(7.29). In the case of Eq. (7.27), we have

$$\langle\Psi_0^{(0)}|\mathbf{A}^{(0)}|\Psi_0^{(1)}\rangle + \langle\Psi_0^{(0)}|\mathbf{V}|\Psi_0^{(0)}\rangle = a_0^{(0)}\langle\Psi_0^{(0)}|\Psi_0^{(1)}\rangle + a_0^{(1)}\langle\Psi_0^{(0)}|\Psi_0^{(0)}\rangle \tag{7.32}$$

Using

$$\langle\Psi_0^{(0)}|\mathbf{A}^{(0)}|\Psi_0^{(1)}\rangle = \langle\Psi_0^{(1)}|\mathbf{A}^{(0)}|\Psi_0^{(0)}\rangle^* \tag{7.33}$$

and Eqs. (7.26), (7.30), and (7.31), we can simplify Eq. (7.32) to

$$\langle\Psi_0^{(0)}|\mathbf{V}|\Psi_0^{(0)}\rangle = a_0^{(1)} \tag{7.34}$$

which is the well-known result that the first-order correction to the eigenvalue is the expectation value of the perturbation operator over the unperturbed wave function.

As for $\Psi_0^{(1)}$ like *any* function of the electronic coordinates, it can be expressed as a linear combination of the *complete* set of eigenfunctions of $\mathbf{A}^{(0)}$, i.e.,

$$\Psi_0^{(1)} = \sum_{i>0} c_i \Psi_i^{(0)} \tag{7.35}$$

To determine the coefficients c_i in Eq. (7.35), we can multiple Eq. (7.27) on the left by $\Psi_j^{(0)}$ and integrate to obtain

$$\langle \Psi_j^{(0)} | \mathbf{A}^{(0)} | \Psi_0^{(1)} \rangle + \langle \Psi_j^{(0)} | \mathbf{V} | \Psi_0^{(0)} \rangle = a_0^{(0)} \langle \Psi_j^{(0)} | \Psi_0^{(1)} \rangle + a_0^{(1)} \langle \Psi_j^{(0)} | \Psi_0^{(0)} \rangle \tag{7.36}$$

Using Eq. (7.35), we expand this to

$$\begin{aligned} \left\langle \Psi_j^{(0)} \left| \mathbf{A}^{(0)} \right| \sum_{i>0} c_i \Psi_i^{(0)} \right\rangle + \langle \Psi_j^{(0)} | \mathbf{V} | \Psi_0^{(0)} \rangle = \\ a_0^{(0)} \left\langle \Psi_j^{(0)} \left| \sum_{i>0} c_i \Psi_i^{(0)} \right. \right\rangle + a_0^{(1)} \langle \Psi_j^{(0)} | \Psi_0^{(0)} \rangle \end{aligned} \tag{7.37}$$

which, from the orthonormality of the eigenfunctions, simplifies to

$$c_j a_j^{(0)} + \langle \Psi_j^{(0)} | \mathbf{V} | \Psi_0^{(0)} \rangle = c_j a_0^{(0)} \tag{7.38}$$

or

$$c_j = \frac{\langle \Psi_j^{(0)} | \mathbf{V} | \Psi_0^{(0)} \rangle}{a_0^{(0)} - a_j^{(0)}} \tag{7.39}$$

With the first-order eigenvalue and wave function corrections in hand, we can carry out analogous operations to determine the second-order corrections, then the third-order, etc. The algebra is tedious, and we simply note the results for the eigenvalue corrections, namely

$$a_0^{(2)} = \sum_{j>0} \frac{|\langle \Psi_j^{(0)} | \mathbf{V} | \Psi_0^{(0)} \rangle|^2}{a_0^{(0)} - a_j^{(0)}} \tag{7.40}$$

and

$$a_0^{(3)} = \sum_{j>0,k>0} \frac{\langle \Psi_0^{(0)} | \mathbf{V} | \Psi_j^{(0)} \rangle [\langle \Psi_j^{(0)} | \mathbf{V} | \Psi_k^{(0)} \rangle - \delta_{jk} \langle \Psi_0^{(0)} | \mathbf{V} | \Psi_0^{(0)} \rangle] \langle \Psi_k^{(0)} | \mathbf{V} | \Psi_0^{(0)} \rangle}{(a_0^{(0)} - a_j^{(0)})(a_0^{(0)} - a_k^{(0)})} \tag{7.41}$$

Let us now examine the application of perturbation theory to the particular case of the Hamiltonian operator and the energy.

7.4.2 Single-reference

We now consider the use of perturbation theory for the case where the complete operator **A** is the Hamiltonian, **H**. Møller and Plesset (1934) proposed choices for $\mathbf{A}^{(0)}$ and **V** with this goal in mind, and the application of their prescription is now typically referred to by the acronym MPn where n is the order at which the perturbation theory is truncated, e.g., MP2, MP3, etc. Some workers in the field prefer the acronym MBPTn, to emphasize the more general nature of many-body perturbation theory (Bartlett 1981).

The MP approach takes $\mathbf{H}^{(0)}$ to be the sum of the one-electron Fock operators, i.e., the non-interacting Hamiltonian (see Section 4.5.2)

$$\mathbf{H}^{(0)} = \sum_{i=1}^{n} f_i \tag{7.42}$$

where n is the number of basis functions and f_i is defined in the usual way according to Eq. (4.52). In addition, $\Psi^{(0)}$ is taken to be the HF wave function, which is a Slater determinant formed from the occupied orbitals. By analogy to Eq. (4.36), it is straightforward to show that the eigenvalue of $\mathbf{H}^{(0)}$ when applied to the HF wave function is the sum of the occupied orbital energies, i.e.,

$$\mathbf{H}^{(0)}\Psi^{(0)} = \sum_{i}^{\text{occ.}} \varepsilon_i \Psi^{(0)} \tag{7.43}$$

where the orbital energies are the usual eigenvalues of the specific one-electron Fock operators. The sum on the r.h.s. thus defines the eigenvalue $a^{(0)}$.

Recall that this is *not* the way the electronic energy is usually calculated in an HF calculation – it is the expectation value for the *correct* Hamiltonian and the HF wave function that determines that energy. The 'error' in Eq. (7.43) is that each orbital energy includes the repulsion of the occupying electron(s) with all of the other electrons. Thus, each electron–electron repulsion is counted twice (once in each orbital corresponding to each pair of electrons). So, the correction term **V** that will return us to the correct Hamiltonian and allow us to use perturbation theory to improve the HF wave function and eigenvalues must be the difference between counting electron repulsion once and counting it twice. Thus,

$$\mathbf{V} = \sum_{i}^{\text{occ.}} \sum_{j>i}^{\text{occ.}} \frac{1}{r_{ij}} - \sum_{i}^{\text{occ.}} \sum_{j}^{\text{occ.}} \left(J_{ij} - \frac{1}{2} K_{ij} \right) \tag{7.44}$$

where the first term on the r.h.s. is the proper way to compute electron repulsion (and is exactly as it appears in the Hamiltonian of Eq. (4.3) and the second term is how it is computed from summing over the Fock operators for the occupied orbitals where J and K are the Coulomb and exchange operators defined in Section 4.5.5. Note that, since we are summing over occupied orbitals, we must be working in the MO basis set, not the AO one.

So, let us now consider the first-order correction $a^{(1)}$ to the zeroth-order eigenvalue defined by Eq. (7.43). In principle, from Eq. (7.34), we operate on the HF wave function $\Psi^{(0)}$

with **V** defined in Eq. (7.44), multiply on the left by $\Psi^{(0)}$, and integrate. By inspection, cognoscenti should not have much trouble seeing that the result will be the negative of the electron–electron repulsion energy. However, if that is not obvious, there is no need to carry through the integrations in any case. That is because we can write

$$\begin{aligned} a^{(0)} + a^{(1)} &= \langle \Psi^{(0)} | \mathbf{H}^{(0)} | \Psi^{(0)} \rangle + \langle \Psi^{(0)} | \mathbf{V} | \Psi^{(0)} \rangle \\ &= \langle \Psi^{(0)} | \mathbf{H}^{(0)} + \mathbf{V} | \Psi^{(0)} \rangle \\ &= \langle \Psi^{(0)} | \mathbf{H} | \Psi^{(0)} \rangle \\ &= E_{\mathrm{HF}} \end{aligned} \tag{7.45}$$

i.e., *the Hartree-Fock energy is the energy correct through first-order in Møller-Plesset perturbation theory*. Thus, the second term on the r.h.s. of the first line of Eq. (7.45) must indeed be the negative of the overcounted electron–electron repulsion already noted to be implicit in $a^{(0)}$.

As MP1 does not advance us beyond the HF level in determining the energy, we must consider the second-order correction to obtain an estimate of correlation energy. Thus, we must evaluate Eq. (7.40) using the set of all possible excited-state eigenfunctions and eigenvalues of the operator $\mathbf{H}^{(0)}$ defined in Eq. (7.42). Happily enough, that is a straightforward process, since within a finite basis approximation, the set of all possible excited eigenfunctions is simply all possible ways to distribute the electrons in the HF orbitals, i.e., all possible excited CSFs appearing in Eq. (7.10).

Let us consider the numerator of Eq. (7.40). Noting that **V** is $\mathbf{H} - \mathbf{H}^{(0)}$, we may write

$$\begin{aligned} \sum_{j>0} \langle \Psi_j^{(0)} | \mathbf{V} | \Psi_0^{(0)} \rangle &= \sum_{j>0} \langle \Psi_j^{(0)} | \mathbf{H} - \mathbf{H}^{(0)} | \Psi_0^{(0)} \rangle \\ &= \sum_{j>0} [\langle \Psi_j^{(0)} | \mathbf{H} | \Psi_0^{(0)} \rangle - \langle \Psi_j^{(0)} | \mathbf{H}^{(0)} | \Psi_0^{(0)} \rangle] \\ &= \sum_{j>0} \left[\langle \Psi_j^{(0)} | \mathbf{H} | \Psi_0^{(0)} \rangle - \sum_{i}^{\text{occ.}} \varepsilon_i \langle \Psi_j^{(0)} | \Psi_0^{(0)} \rangle \right] \\ &= \sum_{j>0} \langle \Psi_j^{(0)} | \mathbf{H} | \Psi_0^{(0)} \rangle \end{aligned} \tag{7.46}$$

where the simplification of the r.h.s. on proceeding from line 3 to line 4 derives from the orthogonality of the ground- and excited-state Slater determinants. As for the remaining integrals, from the Condon–Slater rules, we know that we need only consider integrals involving doubly and singly excited determinants. However, from Brillouin's theorem, we also know that the integrals involving the singly excited determinants will all be zero. The Condon–Slater rules applied to the remaining integrals involving doubly excited determinants

dictate that

$$\sum_{j>0} \langle \Psi_j^{(0)} | \mathbf{V} | \Psi_0^{(0)} \rangle = \sum_{i}^{\text{occ.}} \sum_{j>i}^{\text{occ.}} \sum_{a}^{\text{vir.}} \sum_{b>a}^{\text{vir.}} [(ij|ab) - (ia|jb)] \tag{7.47}$$

where the two-electron integrals are those defined by Eq. (4.56).

As for the denominator of Eq. (7.40), from inspection of Eq. (7.43), $a^{(0)}$ for each doubly excited determinant will differ from that for the ground state only by including in the sum the energies of the virtual orbitals into which excitation has occurred and excluding the energies of the two orbitals from which excitation has taken place. Thus, the full expression for the second-order energy correction is

$$a^{(2)} = \sum_{i}^{\text{occ.}} \sum_{j>i}^{\text{occ.}} \sum_{a}^{\text{vir.}} \sum_{b>a}^{\text{vir.}} \frac{[(ij|ab) - (ia|jb)]^2}{\varepsilon_i + \varepsilon_j - \varepsilon_a - \varepsilon_b} \tag{7.48}$$

The sum of $a^{(0)}$, $a^{(1)}$, and $a^{(2)}$ defines the MP2 energy.

MP2 calculations can be done reasonably rapidly because Eq. (7.48) can be efficiently evaluated. The scaling behavior of the MP2 method is roughly N^5, where N is the number of basis functions. Analytic gradients and second derivatives are available for this level of theory, so it can conveniently be used to explore PESs. MP2, and indeed all orders of MPn theory, are size-consistent, which is a particularly desirable feature. Finally, Saebø and Pulay have described a scheme whereby the occupied orbitals are localized and excitations out of these orbitals are not permitted if the accepting (virtual) orbitals are too far away (the distance being a user-defined variable; Pulay 1983; Saebø and Pulay 1987). This localized MP2 (LMP2) technique significantly decreases the total number of integrals requiring evaluation in large systems, and can also be implemented in a fashion that leads to linear scaling with system size. These features have the potential to increase computational efficiency substantially.

However, it should be noted that the Møller–Plesset formalism is potentially rather dangerous in design. Perturbation theory works best when the perturbation is small (because the Taylor expansions in Eqs. (7.20) and (7.21) are then expected to be quickly convergent). But, in the case of MP theory, the perturbation is the full electron–electron repulsion energy, which is a rather large contributor to the total energy. So, there is no reason to expect that an MP2 calculation will give a value for the correlation energy that is particularly good. In addition, the MPn methodology is *not* variational. Thus, it is possible that the MP2 estimate for the correlation energy will be too large instead of too small (however, this rarely happens in practice because basis set limitations always introduce error in the direction of underestimating the correlation energy).

Naturally, if one wants to improve convergence, one can proceed to higher orders in perturbation theory (note, however, that even at infinite order, there is no guarantee of convergence when a finite basis set has been used). At third order, it is still true that only matrix elements involving doubly excited determinants need be evaluated, so MP3 is not too much more expensive than MP2. A fair body of empirical evidence, however, suggests

that MP3 calculations tend to offer rather little improvement over MP2. Analytic gradients are not available for third and higher orders of perturbation theory.

At the MP4 level, integrals involving triply and quadruply excited determinants appear. The evaluation of the terms involving triples is the most costly, and scales as N^7. If one simply chooses to ignore the triples, the method scales more favorably and this choice is typically abbreviated MP4SDQ. In a small to moderately sized molecule, the cost of accounting for the triples is roughly equal to that for the rest of the calculation, i.e., triples double the time. In closed-shell singlets with large frontier orbital separations, the contributions from the triples tend to be rather small, so ignoring them may be worthwhile in terms of efficiency. However, when the frontier orbital separation drops, the contribution of the triples can become very large, and major errors in interpretation can derive from ignoring their effects. In such a situation, the triples in essence help to correct for the error involved in using a single-reference wave function.

Empirically, MP4 calculations can be quite good, typically accounting for more than 95% of the correlation energy with a good basis set. However, although ideally the MPn results for any given property would show convergent behavior as a function of n, the more typical observation is oscillatory, and it can be difficult to extrapolate accurately from only four points (MP1 = HF, MP2, MP3, MP4). As a rough rule of thumb, to the extent that the results of an MP2 calculation differ from HF, say for the energy difference between two isomers, the difference tends to be overestimated. MP3 usually pushes the result back in the HF direction, by a variable amount. MP4 increases the difference again, but in favorable cases by only a small margin, so that some degree of convergence may be relied upon. Additional performance details are discussed in Section 7.6.

7.4.3 Multireference

The generalization of MPn theory to the multireference case involves the obvious choice of using an MCSCF wave function for $\Psi^{(0)}$ instead of a single-determinant RHF or UHF one. However, it is much less obvious what should be chosen for $\mathbf{H}^{(0)}$, as the MCSCF MOs do not diagonalize any particular set of one-electron operators. Several different choices have been made by different authors, and each defines a unique 'flavor' of multireference perturbation theory (see, for instance, Andersson 1995; Davidson 1995; Finley and Freed 1995). One of the more popular choices is the so-called CASPT2N method of Roos and co-workers (Andersson, Malmqvist, and Roos 1992). Often this method is simply called CASPT2 – while this ignores the fact that different methods having other acronym endings besides N have been defined by these same authors (e.g., CASTP2D and CASPT2g1), the other methods are sufficiently inferior to CASPT2N that they are typically used only by specialists and confusion is minimized.

Most multireference methods described to date have been limited to second order in perturbation theory. Analytic gradients are not yet available. While some third order results have begun to appear, much like the single-reference case, they do not seem to offer much improvement over second order.

An appealing feature of multireference perturbation theory is that it can correct for some deficiencies associated with an incomplete active space. For instance, the relative energies

for various electronic states of TMM (Figure 7.1) were found to vary widely depending on whether a (2,2), (4,4), or (10,10) active space was used; however, the relative energies from corresponding CASPT2 calculations agreed well with one another. Thus, while the motivation for multireference perturbation theory is to address dynamical correlation after a separate treatment of non-dynamical correlation, it seems capable of handling a certain amount of the latter as well.

7.5 Coupled-cluster Theory

One of the more mathematically elegant techniques for estimating the electron correlation energy is coupled-cluster (CC) theory (Cizek 1966). We will avoid most of the formal details here, and instead focus on intuitive connections to CI and MPn theory (readers interested in a more mathematical development may examine Crawford and Schaefer 1996).

The central tenet of CC theory is that the full-CI wave function (i.e., the 'exact' one within the basis set approximation) can be described as

$$\Psi = e^{\mathbf{T}}\Psi_{\mathrm{HF}} \tag{7.49}$$

The cluster operator $\mathbf{T}$ is defined as

$$\mathbf{T} = \mathbf{T}_1 + \mathbf{T}_2 + \mathbf{T}_3 + \cdots + \mathbf{T}_n \tag{7.50}$$

where n is the total number of electrons and the various $\mathbf{T}_i$ operators generate all possible determinants having i excitations from the reference. For example,

$$\mathbf{T}_2 = \sum_{i<j}^{\text{occ.}} \sum_{a<b}^{\text{vir.}} t_{ij}^{ab} \Psi_{ij}^{ab} \tag{7.51}$$

where the amplitudes t are determined by the constraint that Eq. (7.49) be satisfied. The expansion of $\mathbf{T}$ ends at n because no more than n excitations are possible.

Of course, operating on the HF wave function with $\mathbf{T}$ is, in essence, full CI (more accurately, in full CI one applies $1 + \mathbf{T}$), so one may legitimately ask what advantage is afforded by the use of the exponential of $\mathbf{T}$ in Eq. (7.49). The answer lies in the consequences associated with truncation of $\mathbf{T}$. For instance, let us say that we only want to consider the double excitation operator, i.e., we make the approximation $\mathbf{T} = \mathbf{T}_2$. In that case, Taylor expansion of the exponential function in Eq. (7.49) gives

$$\begin{aligned} \Psi_{\mathrm{CCD}} &= e^{\mathbf{T}}\Psi_{\mathrm{HF}} \\ &= \left(1 + \mathbf{T}_2 + \frac{\mathbf{T}_2^2}{2!} + \frac{\mathbf{T}_2^3}{3!} + \cdots\right)\Psi_{\mathrm{HF}} \end{aligned} \tag{7.52}$$

where CCD implies coupled cluster with only the double-excitation operator. Note that the first two terms in parentheses, $1 + \mathbf{T}_2$, define the CID method described in Section 7.3.1.

The remaining terms, however, involve products of excitation operators. Each application of $\mathbf{T}_2$ generates double excitations, so the product of two applications (the square of $\mathbf{T}_2$) generates quadruple excitations. Similarly, the cube of $\mathbf{T}_2$ generates hextuple substitutions, etc. It is exactly the *failure* to include these excitations that makes CI non-size-consistent! So, using the exponential of $\mathbf{T}$ in Eq. (7.49) ensures size consistency. Moreover, through careful analysis of perturbation theory, one can show that CCD is equivalent to including all of the terms involving products of double substitutions out to infinite order, i.e., MP∞D using the notation developed earlier in the context of MP4.

The computational problem, then, is determination of the cluster amplitudes t for all of the operators included in the particular approximation. In the standard implementation, this task follows the usual procedure of left-multiplying the Schrödinger equation by trial wave functions expressed as determinants of the HF orbitals. This generates a set of coupled, non-linear equations in the amplitudes which must be solved, usually by some iterative technique. With the amplitudes in hand, the coupled-cluster energy is computed as

$$\langle \Psi_{\mathrm{HF}} | \mathbf{H} | e^{\mathbf{T}} \Psi_{\mathrm{HF}} \rangle = E_{\mathrm{CC}} \tag{7.53}$$

In practice, the cost of including single excitations (i.e., $\mathbf{T}_1$) in addition to doubles is worth the increase in accuracy, and this defines the CCSD model. The scaling behavior of CCSD is on the order of N^6. Inclusion of connected triple excitations (i.e., those arising with their own unique amplitudes from $\mathbf{T}_3$, not the 'disconnected' triples arising as products of $\mathbf{T}_1$ and $\mathbf{T}_2$) defines CCSDT, but this is very computationally costly (scaling as N^8), making it intractable for all but the smallest of molecules. Various approaches to estimating the effects of the connected triples using perturbation theory have been proposed (each with its own acronym...) Of these, the most robust, and thus most commonly used, is that in the so-called CCSD(T) method, which also includes a singles/triples coupling term (Raghavachari *et al.* 1989). Indeed, the CCSD(T) level has come to be the gold-standard level of theory for single-reference calculations. Note that coupled-cluster theory is not, however, variational.

The CCSD(T) level is reasonably forgiving even in instances where the single-determinant assumption is questionable. Some discussion of this point with examples is provided in the next section. Here, however, we note that one *measure* of the multireference character that is often reported is the so-called T_1 diagnostic of Lee and Taylor (1989), defined as

$$T_1 = \sqrt{\frac{\sum_{i}^{\text{occ.}} \sum_{a}^{\text{vir.}} (t_i^a)^2}{n}} \tag{7.54}$$

where n is the number of electrons and the singles amplitudes are defined analogously to those for the doubles appearing in Eq. (7.51). A value above 0.02 has been suggested as warranting some caution in the interpretation of single-reference CCSD results.

The presence of large singles amplitudes can also be problematic for the CCSD(T) method, because the perturbation theory estimate for the triples can become unstable. One possibility to eliminate that instability involves changing the orbitals used to express the reference

wave function from the canonical HF orbitals to so-called Brueckner orbitals. The Brueckner orbitals are found as linear combinations of the HF orbitals subject to the constraint that all of the singles amplitudes in the CCSD cluster operator be zero (a process that requires iteration). This approach is sometimes called Brueckner doubles (BD). The energetic effect of connected triples can again be estimated using a perturbative approach, which defines the BD(T) method.

A method that is closely related to coupled cluster theory is quadratic configuration interaction including singles and doubles (QCISD). Originally developed by Pople and co-workers as a way to correct for size-consistency errors in CISD (Pople, Head-Gordon, and Raghavachari 1987), it was later shown to be almost equivalent to CCSD in its construction (He and Cremer 1991). The QCISD(T) method includes the same perturbative correction for contributions from unlinked triples as that used in CCSD(T). Typically, CCSD and QCISD give results closely agreeing with one another and the same holds true for their (T) analogs (although in certain challenging systems the more complete coupled-cluster methods have been found to be more robust). Given their usually close correspondence in quality, any motivation to use one over the other tends to derive from the better features that may be associated with it in any given electronic structure code (e.g., inclusion of analytic gradients, a particularly efficient implementation, etc.) Note, however, that while coupled-cluster methods are in principle well defined for the inclusion of excitations up to any level–and in certain benchmark, small-molecule cases, full inclusion of triples, quadruples, etc., can be undertaken–the development of QCISD and QCISD(T) did not proceed from truncation of a general operator, but rather from augmentation of CISD to correct for size inconsistency. Thus, there do not exist any 'higher' levels of QCISD, although such levels could be defined to include additional excitations by analogy to CCSD.

7.6 Practical Issues in Application

The goal of most calculations is to obtain as high a level of accuracy as possible within the constraints of the available computational resources. As including electron correlation in a calculation can be critical to enhancing accuracy, but can also be excruciatingly expensive in large systems, it is important to appreciate the strengths and weaknesses of different correlation techniques with respect to various system characteristics. This section provides some discussion of factors affecting all correlation treatments, and compares and contrasts certain specific issues associated with individual treatments.

7.6.1 Basis Set Convergence

As noted in Chapter 6, basis-set flexibility is key to accurately describing the molecular wave function. When methods for including electron correlation are included, this only becomes more true. One can appreciate this in an intuitive fashion from thinking of the correlated wave function as a linear combination of determinants, as expressed in Eq. (7.1). Since the excited determinants necessarily include occupation of orbitals that are virtual in the HF determinant, and since the HF determinant in some sense 'uses up' the best

Table 7.1 Basis set convergence for HF and full CI energies of CO and O, respectively

Saturated basis functions	E_{HF}(CO) (a.u.)	E_{CI}(O) (a.u.)
s, p	−112.717	−74.935
s, p, d	−112.785	−75.032
s, p, d, f	−112.790	−75.053
s, p, d, f, g		−75.061
Infinite limit	−112.791	−75.069

combinations of basis functions for the occupied orbitals (from the requirement that the excited states be orthogonal to the ground state), the excited states are more dependent on basis-set completeness (this generalizes to the MCSCF case as well, although the discussion in this section is primarily focused on single-reference theories).

This differential sensitivity is illustrated in Table 7.1, which compares the convergence of the HF energy for CO with the convergence of the full-CI energy for just the O atom. In this case, the convergence is with respect to adding higher angular momentum basis functions into a set that is saturated with functions of lower angular momentum. Note that even though the O atom has only half as many basis functions as CO (accepting that the practical 'infinite' basis-set limit is still actually finite), the energetic gain derived from adding functions of higher angular momentum in the d to g range is typically 2 to 4 times larger in the former system than the latter. Note also that, although the HF energy of CO is effectively converged by the time a saturated basis including f functions is used, the CI energy is still more than 10 kcal mol^{-1} from being converged.

The greater dependence on basis-set quality of correlated calculations compared to those of the HF variety has prompted many developers of basis sets to optimize contractions via some scheme that includes evaluating results from the former. For instance, the 'correlation consistent' prefix of the cc-pVnZ basis sets discussed in Chapter 6 highlights this feature.

It has already been mentioned that one way to improve efficiency is to freeze core electrons in correlation treatments. One might think that it represents a more rigorous calculation if one does not freeze them, but unless the basis set includes extra core functions, there is some imbalance in the treatment of core–core vs. core–valence correlation. Put differently, correlating the core electrons requires that basis functions be provided that can be used for this purpose; split-valence basis sets with minimal cores are ill suited in this regard. Instead, basis sets of true multiple-ζ quality should be used. Recent examples of such basis sets include the correlation-consistent polarized core and valence multiple-ζ (cc-pCVnZ) basis sets of Woon and Dunning (1995), where extra core functions are added in increments, and including higher angular momenta, proportional to those employed for the valence space.

The correlation energy is sometimes separated into so-called radial and angular components. Again, an intuitive view of the correlated calculation is that, by considering the contribution of excited determinants having occupation of HF virtual orbitals, one is helping the electrons to get out of one another's way more effectively than they can within the SCF approximation. The radial component derives from decreasing the contraction for functions of a particular angular momentum, i.e., providing tighter and looser functions around each

atom. Alternatively, the space around each atom within a given distance range can be made more accessible by adding functions of increasingly higher angular momentum, i.e., polarization functions, and this is the second contributor. In general, the importance of angular correlation increases at the expense of radial correlation as the atomic number increases, but this is mostly an effect associated with the core electrons. As a rule of thumb, the same balance noted for HF theory is true for correlated calculations: each level of decreased contraction in a given shell is worth about as much as adding a set of polarization functions of the next higher angular momentum.

For the very small systems in Table 7.1, it is possible to approach the exact solution of the Schrödinger equation, but, as a rule, convergence of the correlation energy is depressingly slow. Mathematically, this derives from the poor ability of products of one-electron basis functions, which is what Slater determinants are, to describe the cusps in two-electron densities that characterize electronic structure. For the MP2 level of theory, Schwartz (1962) showed that, in the limit of large l, the error in the correlation energy between electrons of opposite spin goes as $(l + 1/2)^{-3}$ where l is the highest angular momentum that is saturated in the basis set. Thus, if we apply this formula to going from an (s,p) saturated basis set to an (s,p,d) basis set, our error drops by only 64%, i.e., we recover a little less than two-thirds of the missing correlation energy. Going from (s,p,d) to (s,p,d,f), the improvement drops to 53%, or, compared to the (s,p) starting point, about five-sixths of the original error. Since the correlation energy can be enormous, and since actually saturating the basis set in these functions of higher angular momentum can be expensive, such convergence behavior is not especially good.

The so-called R12 methods of Klopper and Kutzelnigg (1987) provide an interesting alternative to Slater-determinant-based methods with respect to analysis of convergence behavior. In this methodology, the wave function is not simply a product of one-electron orbitals but includes additionally all interelectronic distances r_{ij} (such wave functions were first pioneered by Hylleraas (1929) in the early days of quantum mechanics in an effort to treat the 2-electron helium atom as accurately as the one-electron hydrogen atom). With the interelectronic cusp explicitly included, the convergence behavior of the MP2 correlation energy improves to $(l + 1/2)^{-7}$. Now the error recovery on going from an (s,p) to an (s,p,d,f) basis set is in principle 98%. In practice, the R12 methods require very large basis sets for technical reasons, and so such calculations continue to be limited to relatively small systems, but they are still quite useful for benchmarking purposes and may see an expanded role with future developments.

Note that the scaling behavior of methods more highly correlated than MP2 is expected in general to be less favorable than MP2. This derives from the greater sensitivity to basis set exhibited by determinants involving excitations beyond double, since still more virtual orbitals must be occupied.

7.6.2 Sensitivity to Reference Wave Function

For single-reference correlated methods, there are several issues associated with the HF reference that can significantly affect the interpretation of the correlated calculation. First, there is the degree to which the wave function can indeed be reasonably well described by a

single configuration, i.e., the extent to which the HF determinant dominates in the expansion of Eq. (7.1). When a non-trivial degree of multireference character exists, perturbation theory is particularly sensitive to this feature, and can give untrustworthy results. To appreciate this, recall the TMM example with which this chapter began (Figure 7.1). Let us take the single-configuration HF wave function represented by Eq. (7.2), and consider the MP2 energy contribution from the double excitation taking both electrons from occupied orbital π_2 to virtual orbital π_3. From Eq. 7.48, we see that the denominator associated with this term is the difference in orbital energies. However, since these orbitals are formally degenerate, the denominator is zero and the perturbation theory expression for the energy associated with this term is infinite! Note that this is not a case of the actual electronic state being degenerate, but is entirely an artifact of using a single-reference wave function.

As a general rule, whenever the frontier orbital separation becomes small, the magnitude of the MPn energy terms will become large because of their inverse dependence on orbital energy separation, and perturbation theory will be very slowly convergent in such instances. Such decreased separations also increase the degree of multireference character in the wave function, so the tight coupling between these two phenomena is rationalized. Figure 7.5 provides an example of this phenomenon for the case of the energy of carbonyl oxide, which has a moderate degree of multireference character, relative to its isomer dioxirane. The comparatively small size of these systems permits the extension of perturbation theory through fifth order with the 6-31G(d) basis set, but even the difference between the MP4 and MP5 results remains a fairly large 1.3 kcal mol^{-1}.

CCSD is similarly sensitive to multireference character, although it is less obvious that this should be so based on the formalism presented above. However, inclusion of triples in the CCSD wave function is usually very effective in correcting for a single-reference treatment of a weakly to moderately multireference problem. Of course, the most common way to include the triples is by perturbation theory, i.e., CCSD(T), and as noted above, this level

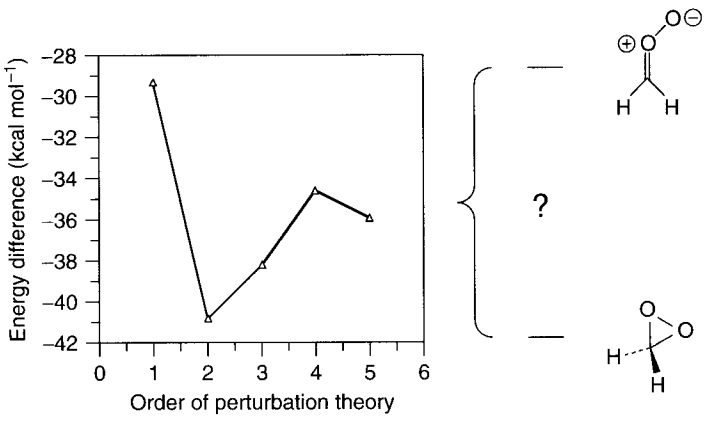


Figure 7.5 Slowly oscillatory behavior of MPn/6-31G(d)//HF/6-31G(d) theory for the energy separation between carbonyl oxide and dioxirane. Accurate extrapolation from this perturbation series is an unlikely prospect

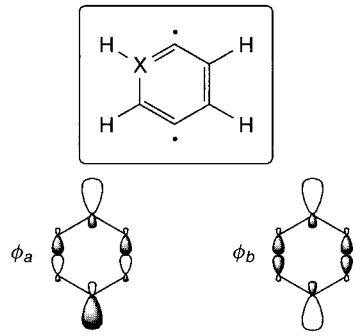


Figure 7.6 Frontier orbitals of a *para* aryne diradical. In the isoelectronic cases of X = C and X = NH^+, the energy of orbital ϕ_a is very slightly below that of ϕ_b, leading to a high degree of multiconfigurational character in the singlets. When X = C, the two orbitals belong to the b_{1u} and a_g irreps of the D_{2h} point group, respectively, and thus a single excitation from the former to the latter cannot contribute to the singlet ground state that has overall A_g symmetry; only a double excitation contributes. When X = NH^+, however, both orbitals are of a′ type symmetry within the C_s point group, so that a single excitation can (and does) make a large contribution to the singlet ground state that has overall A′ symmetry. The latter situation can contribute to instability in estimating the energetic effects of triples substitutions in '(T)' methods based on a single-determinantal reference

too can be unstable if singles amplitudes are large. In such an instance, BD(T) calculations, which eliminate the singles amplitudes, can be efficacious.

To illustrate some of these points in greater detail, consider the aryne diradicals *p*-benzyne and its isoelectronic but charged congener, *N*-protonated 2,5-pyridyne (Figure 7.6). In these systems, the frontier orbitals of interest are the bonding and antibonding combinations of the 'singly occupied' σ orbitals left after abstraction of two hydrogen atoms from the aromatic ring. Because the orbitals are *para*-related to one another, the interaction between them is weak, and thus the frontier orbital energy separation is small. As such, the closed-shell singlet and triplet states lie relatively near one another; in the case of *p*-benzyne, negative ion photoelectron spectroscopy has established the singlet–triplet (S-T) splitting to be -3.8 ± 0.5 kcal mol^{-1} (Wenthold *et al.* 1998). This corresponds to an *energy* splitting of -4.2 ± 0.5 kcal mol^{-1} (i.e., differences in zero-point vibrational energy have been removed). High-level calculations suggest, not surprisingly, that the S-T splitting in the *N*-protonated pyridyne system should be very nearly the same (Debbert and Cramer 2000). Table 7.2 illustrates the results from a variety of different levels of electronic structure theory applied to computing the S-T splitting using the cc-pVDZ basis set.

Notice, first, how spectacularly wrong the HF results are, this being indicative of significant multireference character for the singlets, which should really be described as about 60:40 mixtures of the two determinants corresponding to double occupation of the antibonding and bonding combinations of the σ orbitals. In the *p*-benzyne case, the MP2 calculation correctly predicts the singlet to be the preferred state, but drastically overshoots in doing so. As is typical, MP3 oscillates back to the HF prediction (triplet ground state) but with a smaller margin of error, and then MP4 corrects back again in the proper direction, with a somewhat

Table 7.2 Singlet–triplet splittings (kcal mol^{-1}) for *p*-benzyne and *N*-protonated 2,5-pyridyne[a]

Level of Theory	*p*-Benzyne	*N*-Protonated 2,5-Pyridyne
HF	87.8	87.4
MP2	−25.3	−4.1
MP2 (cc-pVTZ)	−27.8	11.4
MP3	22.8	42.0
MP4SDQ	14.3	10.1
MP4	−20.9	−21.0
CCSD	16.8	17.3
CCSD (cc-pVTZ)	18.4	18.9
CCSD(T)	−4.5	−29.4
QCISD	16.2	4.7
QCISD(T)	−4.1	−5.9
BD	17.0	17.0
BD(T)	−4.1	−5.1
CAS(8,8)	−2.7	−2.4
CASPT2(8,8)	−5.1	−5.1
Experiment or best estimate	**−4.2**	**−5.0**

[a]Basis set cc-pVDZ unless otherwise indicated; geometries from BPW91/cc-pVDZ density functional calculations (see Chapter 8).

smaller overestimation than was observed for MP2. Clearly, however, one could not with confidence extrapolate to an infinite-order perturbation theory result from these four points. The situation is much the same for the 2,5-pyridynium ion, except that the MP2 result is very close to experiment. Such fortuitous agreement is obviously entirely coincidental, as the perturbation series is wildly oscillating.

Note also the importance of triple excitations in correcting for the multideterminantal character. The change in the S-T splitting from inclusion of the triples at the MP4 level is as large as or larger than the change in going from MP3 to MP4SDQ. A similar effect is seen with the QCISD and CCSD formalisms – both incorrectly predict triplet ground states, but inclusion of triples via the (T) formalism gives for the most part rather good results. A significant exception is the CCSD(T) result for the 2,5-pyridynium ion, where the triples correction drastically overcorrects. This is an example of instability arising from large singles amplitudes in the CCSD expansion. In *p*-benzyne, the symmetry of the bonding and antibonding combinations of the σ orbitals is different, so a single excitation from one orbital to the other cannot contribute to the closed-shell wave function. In the less symmetric 2,5-pyridynium ion, however, these orbitals are the same symmetry, so such excitations are allowed and are major contributors to the wave function (Figure 7.6). In such instances, BD(T) calculations are to be preferred over CCSD(T), and indeed, the BD(T) level of theory performs very nicely for this problem (in this particular case the QCISD(T) level also seems to be more robust than CCSD(T) with respect to sensitivity to singles, but this is the reverse of the situation that normally obtains).

As for basis-set convergence, triple-ζ calculations at the MP2 and CCSD levels are provided for comparison to the double-ζ results. For this particular property, the results

for *p*-benzyne are not terribly sensitive to improvements in the flexibility of the basis set. In the pyridynium ion case, the CCSD results are also not very sensitive, but a large effect is seen at the MP2 level. This has more to do with the instability of the perturbation expansion than any intrinsic difference between the isoelectronic arynes.

Note that when multiconfigurational character is *explicitly* accounted for, by an MCSCF calculation using a complete active space including the relevant σ orbitals and electrons as well as the six π orbitals and electrons, the results even without accounting for dynamical electron correlation are fairly good. Including dynamical correlation at the CASPT2 level improves them to the point where they are quite good.

Insofar as CASPT2 uses a multiconfigurational reference, one might expect it to be less prone than MP2 to instability. This is entirely true, *so long as the MCSCF reference is adequate*. If the MCSCF has converged to a spurious solution, perturbation theory is often successful in identifying this because a very large contribution from one or more excitations will be observed. Alternatively, if the MCSCF failed to include one or more critical orbitals, again, large contributions will be obtained for corresponding excitations. From a formal standpoint, it is better to include those orbitals in the active space than it is to rely on CASPT2 to correct for both dynamical and non-dynamical behavior, even though in some instances it seems the latter approach can give good results.

A separate issue that can contribute to instability in correlated calculations is spin contamination. As noted in Chapter 6, spin contamination refers to the inclusion in the wave function of contributions from states of higher spin that mix in when unrestricted methods permit α and β spin orbitals to localize in different regions of space. As a rough rule, the sensitivity of different methods to spin contamination is about what it is to multiconfigurational character: MP*n* methods are to be avoided and inclusion of triples in CCSD or QCISD (or BD) is important. So-called projected MP*n* methods attempt to correct for spin contamination after the fact by projecting out states of higher spin from the correlated wave function (see Appendix B), and these methods tend to be helpful in cases where spin contamination is relatively small, say no more than 10%. Unfortunately, analytic gradients are not available for spin-projected methods, so they must be applied to geometries the optimization of which may have taken place at a considerably less reliable level.

An issue related to spin contamination is so-called Hartree–Fock instability. Various wave functions can exhibit different kinds of instabilities, often, but not always, associated with trying to describe a multiconfigurational system with a single-determinant approach. Thus, for instance, RHF solutions may be unstable with respect to breaking the identical character of the α and β orbitals – a so-called RHF $\rightarrow$ UHF instability (the UHF singlet wave function is usually highly contaminated with triplet character after reoptimization). UHF wave functions for symmetric systems can also be unstable, in this case with respect to spatial symmetry breaking of the individual orbitals. The MOs, if allowed to relax, fail to fall into the irreps of the molecular point group, adopting instead lower symmetry shapes even if the molecular framework is held fixed so as to continue to belong to the higher symmetry point group. Such instability tends to be associated with systems having delocalized spin – the allyl radical is a classical example. All of these cases prove very problematic for perturbation theory, but are handled with somewhat greater success by other correlated methods. In certain very highly

symmetric systems, the wave function can also be unstable to using complex instead of real MOs, but this situation is rare. Most modern electronic-structure programs allow one to check the stability of the HF wave function with respect to these various phenomena, and such steps are warranted in cases having narrow frontier orbital separations and/or delocalized spin. Resort to MCSCF wave functions can be required in particularly problematic systems.

7.6.3 Price/Performance Summary

For a typical equilibrium structure, the HF level of theory predicts bond lengths that are usually a little too short. It is simple to rationalize this using Eq. (7.1). To the extent that correlated methods include excited configurations in the wave function expansion, and to the extent that the orbitals into which excitations occur typically have some antibonding character, this tends to increase bond lengths in order to lower the energy. As a rule, the MP2 level is an excellent choice for geometry optimizations of *minima* that include correlation energy, and significant improvements can be obtained at fairly reasonable cost. Scheiner *et al.* (1997) examined a large number of bond lengths in 108 molecules containing from two to eight atoms and found that, with the 6-31G(d,p) basis set, the average error in bond length at the MP2 level was 0.015 Å, which may be compared to an error at the HF level of 0.021 Å. An improvement of roughly the same order was obtained by Feller and Peterson (1998) in a separate investigation of 184 small molecules using the aug-cc-pV*n*Z basis sets. Bond angles are already sufficiently accurate at the HF level that little improvement is observed at the MP2 level.

While analytic derivatives are available for several more highly correlated levels of theory, geometric improvements beyond the MP2 level tend to be so small for equilibrium structures that they are not worth the cost. This is *not* necessarily the case for TS structures, where the accurate description of a partial bond may well require correlation beyond the MP2 level. As a rough rule, if one observes a *large* change in some geometric property on going from the HF to the MP2 level, it is probably worthwhile to investigate the predictions from still higher levels of theory, since clearly the perturbations are large, and there is good reason to believe MP2 does not provide convergence in the property of interest.

With respect to energetics, MP2 must again be considered a very efficient level of theory for energy differences between minima. In many instances, one finds that the error in such differences is reduced by 25–50% on going from the HF level to the MP2 level. For instance, Hehre reports a sample of 45 isomerizations where errors in isomer energies were reduced from 2.9 to 1.9 kcal mol^{-1} on going from HF/6-31G(d) to MP2/6-31G(d). For the 11 glucose conformers discussed in Chapters 5 and 6, the average error in conformational energy is reduced from 0.6 to 0.4 kcal mol^{-1} on going from HF/cc-pVTZ//MP2/cc-pVDZ to MP2/cc-pVTZ//MP2/cc-pVDZ (Barrows *et al.* 1998). Note, though, that for the same glucose conformers, the error *increases* from 0.1 to 1.0 kcal mol^{-1} on going from HF/cc-pVDZ to MP2/cc-pVDZ, illustrating the degree to which errors in basis set and correlation approximations can sometimes offset one another.

However, the generally robust nature of MP2 in the above examples simply reflects the degree to which most minima are already fairly well described by HF wave functions. When

this is not the case, e.g., in TS structures, there are no hard and fast rules that can be cited with respect to the expected quality of any level of theory. Instead, one is thrown back on the twin responsibilities of demonstrating either (a) agreement with known experimental data of one kind or another in the same or related systems or (b) convergence with respect to treatment of electron correlation. A *rough* quality ordering that is often observed is

$$\begin{aligned} &\text{HF} < \text{MP2} \sim \text{MP3} \sim \text{CCD} < \text{CISD} \\ &< \text{MP4SDQ} \sim \text{QCISD} \sim \text{CCSD} < \text{MP4} \\ &< \text{QCISD(T)} \sim \text{CCSD(T)} \sim \text{BD(T)} \end{aligned} \tag{7.55}$$

Table 7.3 provides a more quantitative feel for the performance summary embodied in Eq. (7.55) using data provided by Bartlett (1995) for the absolute errors in various levels of theory compared to full CI for HB, H_2O, and HF using a polarized double-ζ basis set. In this case, calculations were carried out both at the equilibrium geometries, and also at geometries where the X–H bonds were stretched by 50% and 100%; correlation should become more important in describing these higher energy species. The ordering of the levels in the table is approximately that listed in Eq. (7.55), although CCD seems to do fortuitously well. As expected, the lower levels of correlation treatment degrade markedly compared to the higher levels when the bonds are stretched. A few levels not generally available in most electronic structure packages are included in the table for completeness, including MP5, MP6, and levels having full inclusion of triples, represented by 'T' instead of '(T)'. The heroically expensive CCSDTQ, which takes full account of all triple and quadruple excitations, is also included, and shows the extraordinarily high accuracy one might expect for so complete a treatment, albeit one that can only be applied to the smallest of molecules.

Table 7.3 Average errors in correlation energies (kcal mol^{-1}) compared to full CI for various methods applied to HB, H_2O, and HF at both equilibrium and bond-stretched geometries

Level of theory	Equilibrium geometry	Equilibrium and stretched geometries
MP2	10.4	17.4
MP3	5.0	14.4
CISD	5.8	13.8
CCD	2.4	8.0
MP4SDQ	2.7	7.1
CCSD	1.9	4.5
QCISD	1.7	4.0
MP4	1.3	3.7
MP5	0.8	3.2
MP6	0.3	0.9
CCSD(T)	0.3	0.6
QCISD(T)	0.3	0.5
CCSDT	0.2	0.5
CCSDTQ	0.01	0.02

Table 7.4 Formal scaling behavior, as a function of basis functions N, of various electronic structure methods

Scaling behavior	Method(s)
N^4	HF
N^5	MP2
N^6	MP3, CISD, MP4SDQ, CCSD, QCISD
N^7	MP4, CCSD(T), QCISD(T)
N^8	MP5, CISDT, CCSDT
N^9	MP6
N^{10}	MP7, CISDTQ, CCSDTQ

To further judge what level may be appropriate for a given problem, it is critical that cost be taken into account. The scaling behavior of the various levels in Eq. (7.55) varies widely, as indicated in Table 7.4. Given the price/performance ratios implied by comparing Eq. (7.55) with Table 7.4, there is, for instance, usually little point in doing an MP3 or CISD calculation when superior MP4SDQ or CCSD calculations may typically be accomplished at roughly similar cost (note that scaling similarity is *not* the same as overall time similarity, since the times for the benchmark 'one-basis-function' calculations may differ, but for small to moderately sized molecules, the overall times do not tend to be terribly dissimilar). It should also be recalled that in the large molecule limit, all scaling behaviors tend to reduce because prescreening techniques can avoid the calculation of many negligible integrals.

7.7 Parameterized Methods

Having ascended to the heights of theoretical rigor, it is perhaps time for a brief respite and a timely recapitulation of, of all things, the philosophy underlying the development of *semiempirical* methods: wouldn't it be nice to get the right answer for any problem in general? Although methods like full CI and CCSDTQ, when used in conjunction with large and flexible basis sets, are breathtakingly accurate as solutions of the Schrödinger equation, the bottom line is that they simply cannot be applied to more than the smallest fraction of chemically interesting systems because of their computational expense. And, with scaling behaviors on the order of N^{10}, this situation is unlikely to change anytime soon. As a result, particularly within the last decade, practitioners of *ab initio* MO theory have returned to the idea of introducing parameters to improve predictive accuracy, albeit with a considerably lighter touch than that associated with a full-blown semiempirical method. This section describes a variety of different approaches to improving the results from calculations including electron correlation.

7.7.1 Scaling Correlation Energies

The premise behind correlation scaling is particularly simple. Because of basis-set limitations and approximations in the correlation treatment, one is very rarely able to compute the full

correlation energy. However, with a given choice of basis set and level of theory, the fraction that *is* calculated is often quite consistent over a fairly large range of structure. Thus, we might define an improved electronic energy as

$$E_{\text{SAC-e.c.m.}} = E_{\text{HF}} + \frac{E_{\text{e.c.m.}} - E_{\text{HF}}}{A} \tag{7.56}$$

where 'e.c.m.' is a particular electron correlation method, A is an empirical scale factor typically less than one, and thus all of the correlation energy, computed as the difference between E_{ecm} and E_{HF}, is scaled by the constant factor of A^{-1}. SAC emphasizes this 'scaling all correlation' energy assumption.

As first proposed by Gordon and Truhlar (1986), typically one would go about selecting A by comparison to known experimental data in a system of interest and/or systems related to it. For example, if the subject of interest is the PES for the reaction of the hydroxyl radical with ethyl chloride, and if the overall energies of reaction are known for the abstraction of the α and β hydrogen atoms (to make water and the corresponding alkyl radicals), then A would be selected for a given electron correlation method (say, MP2) in order to make $E_{\text{SAC-MP2}}$ agree with experiment as closely as possible for those particular data points. This same value of A would then be used for any point on the PES. Of course, the more experimental details that can be included in the choice of A, the better the parameterization (and the better able one is to judge the utility of Eq. (7.56) by examination of the errors in a one-parameter fit).

Note that one particularly attractive feature of Eq. (7.56) is that if the particular electron correlation method has available analytic derivatives, so too must $E_{\text{SAC-e.c.m.}}$, since derivatives for the latter will be simply determined as appropriately scaled sums of the e.c.m. and HF derivatives. Geometry optimization, and indeed the entire calculation, can essentially be carried out for exactly the cost of the e.c.m.

While one might imagine that values of A might best be determined individually within any given system, Siegbahn and co-workers have examined a large number of primarily small inorganic systems and suggested that, for the modified coupled-pair functional (MCPF) treatment of correlation (which is analogous in spirit to coupled cluster) with a polarized double-ζ basis set, a value of 0.80 has broad applicability, and they name this choice PCI-80 (Siegbahn, Blomberg, and Svensson 1994; Blomberg and Siegbahn 1998). A summary of the utility of this level of theory for inorganic systems including comparison to density functional theory (DFT) can be found in Table 8.2.

Gordon and Truhlar (1986) have emphasized that variations on the theme of Eq. (7.56) can be useful in different circumstances. For instance, one might imagine carrying out multireference calculations and assuming two different scale factors, one applying to the non-dynamical correlation energy associated with some increase in active space size, and the other with the dynamical correlation energy associated with a CASPT2 calculation. Alternatively, one could have different scaling factors for the terms associated with different levels of electronic excitation, e.g., scaling the doubles differently than the triples. Choices along these lines should be guided by Occam's parameter-razor: in the absence of significant improvements, fewer is better.

7.7.2 Extrapolation

The most attractive feature of the SAC methods is their simplicity. A potential contributor to their possible failure, however, is that the factor A, by being based on experiment, hides within it corrections for both basis-set incompleteness and truncation in the correlation operator. It is not obvious over any particular range of chemical space that either one will be constant, in which case it seems particularly unlikely that either *both* will be constant simultaneously, or that their changes will exactly offset one another. There is thus some virtue in attempting to correct for the two approximations separately. As has already been noted in Chapter 6, estimates of the HF limit can be derived by carrying out calculations with increasingly larger basis sets and then assuming some asymptotic behavior as a function of basis-set size (see Figure 6.4). The same can be done with correlated methods, and many modern basis sets were developed specifically with this goal in mind.

Such a procedure may not seem to be properly classified as a 'parameterized' method, since no individual calculation incorporates a parameter, optimized or otherwise. However, in this instance it is the selection of the functional form for asymptotic behavior that may be considered to be parametric. As noted in Section 7.6.1, for certain levels of theory, like MP2, rigorous convergence behaviors have been derived, but it must be stressed that those behaviors are valid in the limit of a complete basis set, and the ability to fit points obtained with a smaller basis set to the limiting curve is by no means assured (see, for instance, Petersson and Frisch 2000).

In principle, then, in systems where computational costs are not prohibitively expensive, one might try to employ extrapolation so that the energies appearing in Eq. (7.56) represented complete-basis-set (CBS) energies, in which case A corrects only for approximations in the correlation treatment.

7.7.3 Multilevel Methods

In Section 6.2.6, we considered approaches to the HF limit derived under the assumption that various various aspects of basis-set incompleteness (radial, angular, etc.) could be accounted for in some additive fashion (see Eq. (6.5)). In essence, multilevel methods carry this approach one step further, and assume a similar behavior for the correlation energy. For instance, the QM energies for glucose conformers that have served as a benchmark for comparison with lower levels of theory in preceding chapters were computed at a composite (C) level as

$$\begin{aligned} E(\mathrm{C}) = & E(\mathrm{MP2/cc\text{-}pVTZ//MP2/cc\text{-}pVDZ}) \\ & + \{E(\mathrm{CCSD/6\text{-}31G(d)//MP2/6\text{-}31G(d)}) \\ & \quad - E(\mathrm{MP2/6\text{-}31G(d)})\} \\ & + \{E(\mathrm{HF/cc\text{-}p^{T}VQZ//MP2/cc\text{-}pVDZ}) \\ & \quad - E(\mathrm{HF/cc\text{-}pVTZ//MP2/cc\text{-}pVDZ})\} \end{aligned} \tag{7.57}$$

Thus, triple-ζ MP2 energies at double-ζ MP2 geometries are augmented with a correction for doubles contributions beyond second order (line 2 on the r.h.s. of Eq. (7.57)) and a correction for basis set size increase beyond triple-ζ (line 3 on the r.h.s. of Eq. (7.57) where the 'T' superscript in the first basis set implies that polarization functions from cc-pVTZ were used in conjunction with valence functions from cc-pVQZ).

While such *ad hoc* multilevel methods have been employed for rather a long time, only in the late 1980s were efforts undertaken to systematize the approach so as to define a model chemistry having broad applicability. The first such effort was the so-called G1 theory of Pople and co-workers, which was followed very rapidly by an improved modification called G2 theory, so that the former may be considered to be obsolete (Curtiss *et al.* 1991). The steps involved in a G2 calculation are detailed in Table 7.5. In this instance, the goal of the calculation is accurate *thermochemistry*, so some of the steps are devoted to computing thermal contributions to the enthalpy, as opposed to the electronic energy, as described in more detail in Chapter 10. Although the philosophy of the remaining steps is essentially the same as that predicating Eq. (7.57), there is considerably more attention to detail in specific aspects of the basis-set problem and the accounting for electron correlation. There is also a completely empirical correction procedure (step 8) to, *inter alia*, account for core-valence

Table 7.5 Steps in G2 and G3 theory for molecules[a,b]

Step	G2	G3
(1)	HF/6-31G(d) geometry optimization	HF/6-31G(d) geometry optimization
(2)	ZPVE from HF/6-31G(d) frequencies	ZPVE from HF/6-31G(d) frequencies
(3)	MP2(full)/6-31G(d) geometry optimization (all subsequent calculations use this geometry)	MP2(full)/6-31G(d) geometry optimization (all subsequent calculations use this geometry)
(4)	E[MP4/6-311+G(d,p)] $-E$[MP4/6-311G(d,p)]	E[MP4/6-31+G(d)] $-E$[MP4/6-31G(d)]
(5)	E[MP4/6-311G(2df,p)] $-E$[MP4/6-311G(d,p)]	E[MP4/6-31G(2df,p)] $-E$[MP4/6-31G(d)]
(6)	E[QCISD(T)/6-311G(d)] $-E$[MP4/6-311G(d)]	E[QCISD(T)/6-31G(d)] $-E$[MP4/6-31G(d)]
(7)	E[MP2/6-311+G(3df,2p)] $-E$[MP2/6-311G(2df,p)] $-E$[MP2/6-311+G(d,p)] $+E$[MP2/6-311G(d,p)]	E[MP2(full)/G3large[c]] $-E$[MP2/6-31G(2df,p)] $-E$[MP2/6-31+G(d)] $+E$[MP2/6-31G(d)]
(8)	$-0.00481 \times$ (number of valence electron pairs) $-0.00019 \times$ (number of unpaired valence electrons)	$-0.006386 \times$ (number of valence electron pairs) $-0.002977 \times$ (number of unpaired valence electrons)
$E_0 =$	$0.8929 \times (2) + E$[MP4/6-311G(d,p)] + (4) + (5) + (6) + (7) + (8)	$0.8929 \times (2) + E$[MP4/6-31G(d)] + (4) + (5) + (6) + (7) + (8)

[a]For atoms, G3 energies are defined to include a spin-orbit correction taken either from experiment or other high-level calculations. In addition, different coefficients are used in step (8).

[b]In the G2 method, the 6-311G basis set and its derivatives are not defined for second-row atoms; instead, a basis set optimized by McLean and Chandler (1980) is used.

[c]Available at http://chemistry.anl.gov/compmat/g3theory.htm. Defined to use canonical 5 d and 7 f functions.

correlation and to improve the performance of the model over systems having different numbers of unpaired spins. Note that 'full' following a correlation acronym implies that core electrons were included in the correlation treatment, as opposed to the more typical choice of freezing them to excitation. Over a test set of 148 enthalpies of formation, the average error of G2 theory is 1.6 kcal mol^{-1}.

Over time, many different groups have suggested minor modifications of G2 theory (each spawning a new acronym). Some trade accuracy for computational efficiency in order to permit application to larger systems; the most popular of these has been G2(MP2), which avoids the costly MP4 calculations in G2 at the expense of increasing the error over the test set to 1.8 kcal mol^{-1} (Curtiss, Raghavachari, and Pople 1993). Others emphasize alternative methods for obtaining molecular geometries, or attempt to correct for other deficiencies in G2 applied to specific classes of molecules (G2 does poorly on perfluorinated species, for instance).

A modification of G2 by Pople and co-workers was deemed sufficiently comprehensive that it is known simply as G3, and its steps are also outlined in Table 7.5. G3 is more accurate than G2, with an error for the 148-molecule heat-of-formation test set of 0.9 kcal mol^{-1}. It is also more efficient, typically being about twice as fast.

It should be noted that G2 and G3 potentially fail to be size extensive because of the correction term in step 8. If one is studying a homolytic dissociation into two components, at what point along the reaction coordinate are the formerly paired electrons considered to be unpaired? There will be a discontinuity in the energy at that point. In addition, G3 theory uses a different correction for atoms than for molecules, and this too fails to be size extensive.

Alternative multilevel methods that have some similarities to G2, G3, and their variants, are the CBS methods of Petersson and co-workers (see Bibliography at end of chapter). A key difference between the G*n* models and the CBS models is that, rather than assuming basis-set incompleteness effects to be completely accounted for by additive corrections, results for different levels of theory are extrapolated to the complete-basis-set limit in defining a composite energy. Four well-defined CBS models exist, CBS-4, CBS-q, CBS-Q, and CBS-APNO, these being in order of increasing accuracy and, naturally, cost. Over the same 148-molecule test set as used above to evaluate G2 and G3, the average absolute errors of CBS-4, CBS-q, and CBS-Q are 2.7, 2.3, and 1.2 kcal mol^{-1}, respectively. CBS-APNO reduces the error in CBS-Q by a factor of 2 (to only 0.5 kcal mol^{-1} on a somewhat smaller 125-molecule test set), but requires a very expensive QCISD(T)/6-311+G(2df,p) calculation. A particular feature of most of the CBS methods is that they include an (empirical) correction for spin contamination in open-shell species, for which unrestricted treatments potentially sensitive to such contamination are used. In terms of speed, CBS-Q is roughly the speed of G3.

A somewhat more obviously empirical variation on the multilevel approach is the multi-coefficient method of Truhlar and co-workers. Although many different variations of this approach have now been described, it is simplest to illustrate the concept for the so-called multi-coefficient G3 (MCG3) model (Fast, Sánchez, and Truhlar 1999). In essence, the model assumes a G3-like energy expression, but each term has associated with it a coefficient that

is not restricted to be unity, as is the case for G3. Specifically

$$E_{\mathrm{MCG3}} = \sum_{i=1}^{9} c_i(i) + E_{\mathrm{SO}} + E_{\mathrm{CC}} \tag{7.58}$$

where (i) represents a component of the G3 energy (actually, there are some rather slight variations involved with basis sets and frozen-core approximations that increase efficiency), E_{SO} and E_{CC} are empirically estimated spin-orbit and core-correlation energies, and the coefficients c_i are optimized over the usual G3 thermochemistry test set. One additional important difference in the use of G3 energy components is that the G3 empirical correction, which leads to non-size-extensivity, is *not* included. Thus, MCG3 is size extensive. The performance of MCG3 is very slightly better than G3 itself, but this accuracy is achieved at roughly half the cost in terms of computational resources for molecules having 2 or 3 very heavy atoms (the *larger* ones in the G3 test set...)

The real power in the multi-coefficient models, however, derives from the potential for the coefficients to make up for more severe approximations in the quantities used for (i) in Eq. (7.58). At present, Truhlar and co-workers have codified some 20 different multicoefficient models, some of which they term 'minimal', meaning that relatively few terms enter into analogs of Eq. (7.58), and in particular the optimized coefficients absorb the spin-orbit and core-correlation terms, so they are not separately estimated. Different models can thus be chosen for an individual problem based on error tolerance, resource constraints, need to optimize TS geometries at levels beyond MP2, etc. Moreover, for some of the minimal models, analytic derivatives are available on a term-by-term basis, meaning that analytic derivatives for the composite energy can be computed simply as the sum over terms.

A somewhat more chemically based empirical correction scheme is the bond-additivity correction (BAC) methodology. In the BAC-MP4 approach, for instance, the energy of a molecule is computed as

$$\begin{aligned} E(\mathrm{BAC\text{-}MP4}) =& E[\mathrm{MP4/6\text{-}31G(d,p)//HF/6\text{-}31G(d,p)}] \\ &+ \sum_{\mathrm{A,B}} \Delta E_{\mathrm{A-B}} + E_{\mathrm{SC}} + E_{\mathrm{MR}} \end{aligned} \tag{7.59}$$

where E_{SC} and E_{MR} correct for spin contamination (if any) and multireference character (if any) and the summation runs over all atom pairs and each 'bond' correction is a function of bond length (the correction goes to zero at infinite bond length) and a set of parameters, one parameter for each atom and two parameters for each possible pair of atoms. The parameters themselves are determined by fitting to experimental bond dissociation energies, heats of formation (corrected for zero-point vibrational energies and thermal contributions), or other useful thermochemical data. The central assumption of this model, then, is that the error can be decomposed in an additive fashion over the bonds.

In a study of 110 C1 and C2 molecules composed of C, H, O, and F, the average BAC-MP4 unsigned error in predicted heat of formation was 2.1 kcal mol^{-1} (Zachariah *et al.* 1996). As the MP4 calculation uses a relatively modest basis set size, the BAC procedure is

quite fast by comparison to some of the multilevel methods described above. On the other hand, as with any method relying on pairwise parameterization, extension to a large number of atoms requires a great deal of parameterization data, and this is a potential limitation of the BAC method when applied to systems containing atoms not already parameterized.

Because they include empirically derived parameters, multilevel models nearly always outperform single-level calculations at an equivalently expensive level of theory. That being said, one should avoid a slavish devotion to any particular multilevel model simply because it has been graced with an acronym defining it. For any given chemical problem, it is quite possible that an individual investigator can construct a specific multilevel model with relatively little effort that will outperform any of the already defined ones. The issue is simply whether sufficient data exist for the particular system of interest in order to make such a focused model possible. When the data do not, then that is the best time to rely on those previously defined models that have been demonstrated to be reasonably robust over relevant swaths of chemical space.

As for the utility of single-level models, it should be recalled that the goal of most multilevel models is *absolute* energy prediction, while many chemical studies are undertaken in order to better understand *relative* energy differences. Cancellation of errors makes the latter studies more tractable at less complete levels of theory, and single-level models can still be useful in both qualitative and quantitative senses. In addition, there is no wave function defined for the typical parametric model; there is only an energy functional that potentially depends on several different wave functions. Should one wish to know the expectation value for some property *other* than the energy, one will either have to devise a separate multilevel expression, or adopt a single-level formalism for which a wave function is indeed defined.

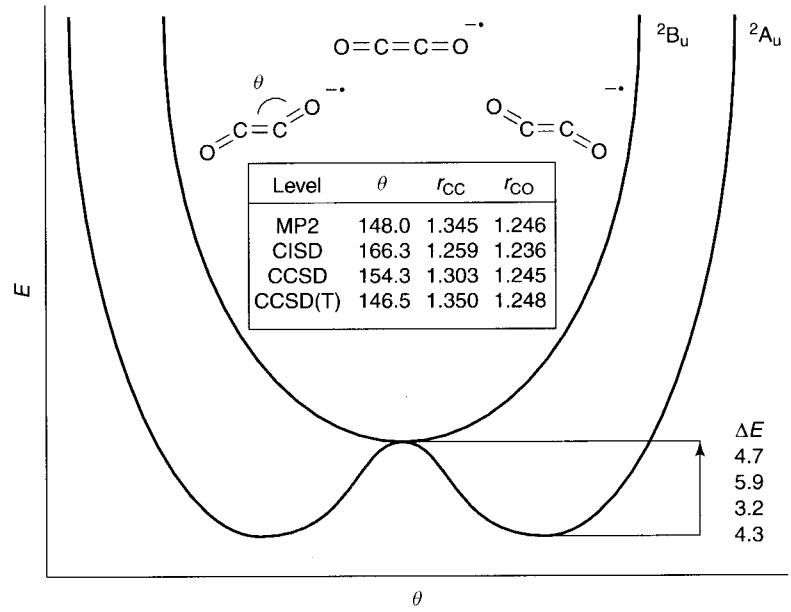
Note that most of the energetic performance data summarized above may also be found in tabular form, compared to density functional models, in Table 8.1

7.8 Case Study: Ethylenedione Radical Anion

Synopsis of Thomas, J. R. *et al.* (1995) 'The Ethylenedione Anion: Elucidation of the Intricate Potential Energy Hypersurface'.

The ground state of ethylenedione, the dimer of carbon monoxide, has been reliably predicted to be a triplet that is bound with respect to dissociation by virtue of its high spin state (two singlet carbon monoxide molecules are lower in energy, but the triplet cannot dissociate into two closed-shell singlets). As such, it has proven an interesting target for synthesis, albeit without success. One possible avenue for its synthesis is to detach electrons from negative ion precursors. This prompted Thomas and co-workers to characterize the radical anion of ethylenedione at a variety of correlated levels of electronic structure theory.

At the UHF level the linear form, which formally has a $^2\Pi_u$ electronic state (see Appendix B for details on group theoretical notation), is predicted to be the minimum energy structure. However, at almost all correlated levels the molecule bends to lift the degeneracy of a pair of a_u and b_u orbitals, leading to a so-called Renner–Teller potential energy surface, as illustrated in Figure 7.7. The lower energy state is ^2A_u and geometric details are provided in the figure for four different correlated levels, all using a large TZP+ basis set.



Level	θ	r_{CC}	r_{CO}
MP2	148.0	1.345	1.246
CISD	166.3	1.259	1.236
CCSD	154.3	1.303	1.245
CCSD(T)	146.5	1.350	1.248

Figure 7.7 Renner–Teller PES for ethylenedione radical anion. Geometrical data for the ^2A_u equilibrium structure are provided for various levels of theory using an augmented polarized triple-ζ basis set (TZP+). Barriers to linearity (ΔE, kcal mol^{-1}) are from CCSD(T) calculations using, from top to bottom, DZP, DZP+, TZP, and TZP+ basis sets. If the initial guess is for the ^2B_u state instead of the ^2A_u state, what will happen?

The details of the molecular structure are difficult to nail down because of the shallow nature of the PES in the vicinity of the linear form. Thus, even with a fairly complete basis set, there are large disagreements between CISD, CCSD, and CCSD(T), although there is a remarkably good (coincidental) agreement between MP2 and CCSD(T). The situation is still more dissatisfying insofar as further increases in basis-set size, in this case adding additional sets of polarization functions, result in bond length changes of up to 0.03 Å and bond angle changes of up to 14° at the MP2, CISD, and CCSD levels. The cost of the CCSD(T) computations is such that use of these larger basis sets is not practical, and thus it is not clear what the effect will be at this formally most complete level of theory.

To further clarify the situation, the authors examined two other quantities dependent on the shape of the PES in the vicinity of the linear form. First, they computed the barrier to double-inversion through the linear form. The data are listed in Figure 7.7., and show some basis set dependence. Note that the CCSD(T)/TZP+ result is approximated to within 0.1 kcal mol^{-1} by summing the CCSD(T)/TZP barrier with the difference between the CCSD(T)/DZP+ and CCSD(T)/DZP barriers. That is, the effect of diffuse functions evaluated with a double-ζ basis set can be treated as additive to the non-augmented triple-ζ results, along the lines described in Section 7.7.3.

The authors made a more exacting comparison for vibrational frequencies, where experimental data were available for the matrix isolated radical anion. Focusing on one fundamental and one combination band, the CCSD(T)/TZP+ predictions of 1527 and 1955 cm^{-1}

compared reasonably well to the experimental values of 1518 and 2042. Again, the flat nature of the PES in the vicinity of the linear form makes things difficult for theory, since this introduces potentially large anharmonicity that is not accounted for in the usual harmonic approximation employed to compute vibrational frequencies (see Section 9.3.2). Isotope shifts in the frequencies, however, showed very close agreement between theory and experiment, all data agreeing to within 5% for seven different isotopomers.

The authors did examine whether significant non-dynamical correlation effects complicated the system, but MCSCF calculations with large active spaces failed to identify any configurations other than the dominant one that entered with coefficients in excess of 0.09, suggesting that the use of single-reference methods was well justified. Part of the challenge for this particular system simply derives from its negative charge, which imposes a greater demand on basis-set saturation. In any case, this example illustrates how deceptively difficult it can be to converge solution of the Schrödinger equation even for seemingly simple chemical systems – a mere four heavy atoms.

Bibliography and Suggested Additional Reading

Bartlett, R. J. 2000. 'Perspective on "On the Correlation Problem in Atomic and Molecular Systems. Calculations of Wavefunction Components in Ursell-type Expansion Using Quantum-field Theoretical Methods"' *Theor. Chem. Acc.*, **103**, 273.

Cramer, C. J. 1998. 'Bergman, Aza-Bergman, and Protonated Aza-Bergman Cyclizations and Intermediate 2,5-Arynes: Chemistry and Challenges to Computation' *J. Am. Chem. Soc.*, **120**, 6261.

Cramer, C. J. and Smith, B. A. 1996. 'Trimethylenemethane. Comparison of Multiconfiguration Self-consistent Field and Density Functional Methods for a Non-Kekulé Hydrocarbon' *J. Phys. Chem.* **100**, 9664.

Curtiss, L. A., Raghavachari, K., Redfern, P. C., Rassolov, V., and Pople, J. A. 1998. 'Gaussian-3 (G3) Theory for Molecules Containing First and Second-row Atoms', *J. Chem. Phys.* **109**, 7764.

Feller, D. and Davidson, E. R. 1990. 'Basis Sets for Ab Initio Molecular Orbital Calculations and Intermolecular Interactions' in *Reviews in Computational Chemistry*, Vol. 1, Lipkowitz, K. B. Boyd, D. B., Eds., VCH: New York, 1.

Hehre, W. J. 1995. *Practical Strategies for Electronic Structure Calculations*, Wavefunction: Irvine, CA.

Hehre, W. J., Radom, L., Schleyer, P. v. R., and Pople, J. A. 1986. Ab Initio *Molecular Orbital Theory*, Wiley: New York.

Jensen, F. 1999. *Introduction to Computational Chemistry*, Wiley: Chichester.

Levine, I. N. 2000. *Quantum Chemistry*, 5th Edn., Prentice Hall: New York.

Martin, J. M. L. 1998. 'Calibration of Atomization Energies of Small Polyatomics' in *Computational Thermochemistry,* ACS Symposium Series, Vol. 677, Irikura, K. K. and Frurip, D. J. Eds., American Chemical Society, Washington, DC, 212.

Petersson, G. A. 1998. 'Complete Basis-set Thermochemistry and Kinetics' in *Computational Thermochemistry,* ACS Symposium Series, Vol. 677, Irikura, K. K. and Frurip, D. J., Eds., American Chemical Society, Washington, DC, 237.

Petersson, G. A., Malick, D. K., Wilson, W. G., Ochterski, J. W., Montgomery, J. A., Jr., and Frisch, M. J. 1998. 'Calibration and Comparison of the Gaussian-2, Complete Basis Set, and Density Functional Methods for Computational Thermochemistry' *J. Chem. Phys.*, **109**, 10570.

Szabo, A. and Ostlund, N. S. 1982. *Modern Quantum Chemistry*, Macmillan: New York.

Tratz, C. M., Fast, P. L., and Truhlar, D. G. 1999. 'Improved Coefficients for the Scaling All Correlations and Multi-coefficient Correlation Methods' *Phys. Chem. Comm.*, 14.

Werner, H.-J. 2000. 'Perspective on "Theory of Self-consistent Electron Pairs. An Iterative Method for Correlated Many-electron Wavefunctions"' *Theor. Chem. Acc.*, **103**, 322.

Zachariah, M. R. and Melius, C. F. 1998. 'Bond-additivity Correction of *Ab Initio* Computations for Accurate Prediction of Thermochemistry' in *Computational Thermochemistry*, ACS Symposium Series, Vol. 677, Irikura, K. K. and Frurip, D. J., Eds., American Chemical Society, Washington, DC, 162.

References

Andersson, K. 1995. *Theor. Chim. Acta,* **91**, 31.

Andersson, K., Malmqvist, P. -Å., and Roos, B. O. 1992. *J. Chem. Phys.*, **96**, 1218.

Barrows, S. E., Storer, J. W., Cramer, C. J., French, A. D., and Truhlar, D. G. 1998. *J. Comput. Chem.*, **19**, 1111.

Bartlett, R. J. 1981. *Ann. Rev. Phys. Chem.*, **32**, 359.

Bartlett, R. J. 1995. In: *Modern Electronic Structure Theory*, Yarkony, D. R., Ed., World Scientific: New York, Part 2, Chapter 6.

Blomberg, M. R. A. and Siegbahn, P. E. M. 1998. In: *Computational Chemistry*, ACS Symposium Series, Vol. 677, Irikura, K. K. and Frurip, D. J., Eds., American Chemical Society: Washington, DC, 197.

Brillouin, L. 1934. *Actualities Sci. Ind.*, **71**, 159.

Bruna, P. J., Peyerimhoff, S. D., and Buenker, R. J. 1980. *Chem. Phys. Lett.*, **72**, 278.

Cizek, J. 1966. *J. Chem. Phys.*, **45**, 4256.

Crawford, T. D. and Schaefer, H. F., III, 1996. In: *Reviews in Computational Chemistry*, Vol. 14, Lipkowitz, K. B. and Boyd, D. B., Eds., Wiley-VCH: New York, 33 and references therein.

Curtiss, L. A., Raghavachari, K., and Pople, J. A. 1993. *J. Chem. Phys.*, **98**, 1293.

Curtiss, L. A., Raghavachari, K., Trucks, G. W., and Pople, J. A. 1991. *J. Chem. Phys.*, **94**, 7221.

Davidson, E. R. 1995. *Chem. Phys. Lett.*, **241**, 432.

Debbert, S. L. and Cramer, C. J. 2000. *Int. J. Mass Spectrom.*, **201**, 1.

Fast, P. L., Sánchez, P. L., and Truhlar, D. G. 1999. *Chem. Phys. Lett.*, **306**, 407.

Feller, D. and Peterson, K. A. 1998. *J. Chem. Phys.*, **108**, 154.

Finley, J. P. and Freed, K. F. 1995. *J. Chem. Phys.*, **102**, 1306.

Gordon, M. S. and Truhlar, D. G. 1986. *J. Am. Chem. Soc.*, **108**, 5412.

He, Z. and Cremer, D. 1991. *Int. J. Quantum Chem., Quantum Chem. Symp.*, **25**, 43.

Hylleraas, E. A. 1929. *Z. Phys.*, **54**, 347.

Klopper, W. and Kutzelnigg, W. 1987. *Chem. Phys. Lett.*, **134**, 17.

Langhoff, S. R. and Davidson, E. R. 1974. *Int. J. Quantum Chem.*, **8**, 61.

Lee, T. J. and Taylor, P. R. 1989. *Int. J. Quantum Chem.*, **S23**, 199.

McLean, A. D. and Chandler, G. S. 1980. *J. Chem. Phys.*,**72**, 5639.

Møller, C. and Plesset, M. S. 1934. *Phys. Rev.*, **46**, 359.

Petersson, G. A. and Frisch, M. J. 2000. *J. Phys. Chem. A,* **104**, 2183.

Pople, J. A., Head-Gordon, M., and Raghavachari, K. 1987. *J. Chem. Phys.*, **87**, 5968.

Pulay, P. 1983. *Chem. Phys. Lett.*, **100**, 151.

Raghavachari, K., Trucks, G. W., Pople, J. A., and Head-Gordon, M. 1989. *Chem. Phys. Lett.*, **157**, 479.

Saebø, S. and Pulay, P. 1987. *J. Chem. Phys.*, **86**, 914.

Scheiner, A. C., Baker, J., and Andzelm, J. W. 1997. *J. Comput. Chem.*, **18**, 775.

Schwartz, C. 1962. *Phys. Rev.*, **126**, 1015.
Siegbahn, P. E. M., Blomberg, M. R., and Svensson, M. 1994. *Chem. Phys. Lett.*, **223**, 35.
Thomas, J. R., DeLeeuw, B. J., O'Leary, P., Schaefer, H. F., III, Duke, B. J., and O'Leary, B. 1995. *J. Chem. Phys.*, **102**, 652.
Wenthold, P. G., Squires, R. R., and Lineberger, W. C. 1998. *J. Am. Chem. Soc.*, **120**, 5279.
Woon, D. E. and Dunning, T. H., Jr., 1995 *J. Chem. Phys.*, **103**, 4572.
Zachariah, M. R., Westmoreland, P. R., Burgess, D. R., Jr., Tsang, W., and Melius, C. F. 1996. *J. Phys. Chem.*, **100**, 8737.