Large Scale Quantum Mechanical Enzymology

Greg Lever

Magdalene College
University of Cambridge

A Dissertation Submitted for
The Degree of Doctor of Philosophy

March 2014

For my family

Contents

Preface

This dissertation describes work carried out between October 2010 and March 2014 in the Theory of Condensed Matter Group at the Cavendish Laboratory, Cambridge, under the supervision of Prof. Mike C. Payne FRS. This dissertation is the result of my own work and contains nothing which is the outcome of work done in collaboration with others, except where specifically indicated in the text. This dissertation has not been submitted in whole or in part for any degree or diploma at this or any other university. Parts of this dissertation have been published, or shall be submitted for publication, as follows:

Chapter 5
Greg Lever, Daniel J. Cole, Nicholas D. M. Hine, Peter D. Haynes and Mike C. Payne
J. Phys.: Conden. Matt. (2013), **25**, 152101

Chapter 6
Greg Lever, Daniel J. Cole, Kara E. Ranaghan, Richard Lonsdale, David J. Wales, Adrian J. Mulholland, Chris-Kriton Skylaris and Mike C. Payne
In preparation (2014)

Statement of Length

This dissertation does not exceed 60,000 words in length.

Greg Lever
Magdalene College, Cambridge
March 2014

Acknowledgments

"No duty is more urgent than that of returning thanks"

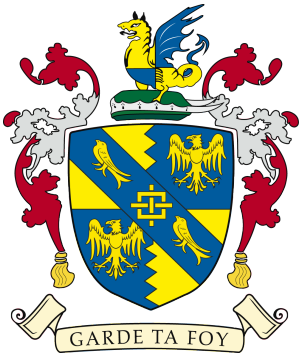
James Allen (1864 - 1912)

I hope to bring sufficient acknowledgment and thanks to those that have helped me along the way. First and foremost, I would like to thank my supervisor, **Mike Payne**, who, not only has proof-read this entire thesis, but has also ensured I have spent a happy and productive time in TCM. Mike has allowed me the freedom to pursue what I found interesting, while also being happy to provide advice, guidance and help wherever needed, along with always knowing the right person to introduce to me in order to further my collaborations. Mike also ensured my access to academic meetings and travel to meetings with collaborators was never hindered, harnessing the expertise of **Tracey Ingham**, **David Taylor**, **Helen Verrechia** and **Alan Clarke** who always did what they do best and made conference arrangements hassle-free and, crucially, always made sure there was plenty of coffee. I am also incredibly grateful to **David Wales**, **Peter Haynes**, **Adrian Mulholland**, **Nick Hine**, **Richard Lonsdale** and **Kara Ranaghan** for their valuable insight and input into my work. **Dave Bowler** and **Veronika Bràzdòva** were my academic mentors prior to postgraduate study. Were it not for their invaluable teaching and encouragement I may never have embarked upon this journey. Dave was the first person to teach me that whilst supervisors can provide overall guidance and ideas, post docs give a more detailed level of help. To this end I have had the fortune of working closely with a very talented researcher, **Danny Cole**, who has helped me a great deal over the years. **Simon Dubois** and **David O'Regan** have also helped me along the way, despite my incessant questions. Reflecting upon my time spent in TCM, it would be remiss of me to not mention **Michael Rutter** who has demonstrated, *inter alia*, a veritable masterclass in sarcasm, perhaps culminating in the adornment of two very large full-colour prints of Margaret Thatcher in my office upon the installation of new notice boards. On the more practical side of things, Michael has provided sterling support for everything computational in TCM, not to mention countless biscuits for afternoon tea, during which many words of wisdom were imparted on me that I shall remember for a long time. Of course I can not forget my office mates who have always made TCM fun and interesting who, in chronological order, are **Sam**, **Emma**, **Andrew**, **Daniel**, **Gen**, **Edgar** and **Aurelio**. I can only hope they did not tire too much of my presence in the office ! I must also mention the remainder of my TCM cohort with whom I met and soon developed great friendship with. **Rob**, **Matt**, **Pascal**, **Hannah**, and **Wave** always made sure there was a strong sense of community among us. I sincerely hope to remain friends with them as we all diverge away from TCM. I must also mention **Johann**, **Oli** and **Steve**, who, amongst everyone else for whom there is too little space to mention, managed to remind me there is life outside of the laboratory, and what an exciting one it always was ! Finally, my parents have given me tremendous support and endless encouragement during the entire process and gave willing help in reading my thesis, multiple times in the case of my father ! Their love and support for me knows no bounds. None of this would have been possible without them and the rest of my family, all to whom I dedicate this thesis.

Large Scale Quantum Mechanical Enzymology

Greg Lever

Magdalene College



Summary

There exists a concerted and continual effort to simulate systems of genuine biological interest to greater accuracy with methods of increasing transferability. More accurate descriptions of these systems at a truly atomistic and electronic level are irrevocably changing our understanding of biochemical processes. Broadly, classical techniques do not employ enough rigour, while conventional quantum mechanical approaches are too computationally expensive for systems of the requisite size. Linear-scaling density-functional theory (DFT) is an accurate method that can apply the predictive power of quantum mechanics to the system sizes required to study problems in enzymology. This dissertation presents methodological developments and protocols, including best practice, for accurate preparation and optimisation, combined with proof-of-principle calculations demonstrating reliable results for a range of small molecule and large biomolecular systems. Previous authors have shown that DFT calculations yield an unphysical, negligible energy gap between the highest occupied and lowest unoccupied molecular orbitals for proteins and large water clusters, a characteristic reproduced in this dissertation. However, whilst others use this phenomenon to question the applicability of Kohn-Sham DFT to large systems, it is shown within this dissertation that the vanishing gap is, in fact, an electrostatic artefact of the method used to prepare the system. Furthermore, practical solutions are demonstrated for ensuring a physical gap is maintained upon increasing system size. Harnessing these advances, the first application using linear-scaling DFT to optimise stationary points in the reaction pathway for the *Bacillus subtilis* chorismate mutase (CM) enzyme is made. Averaged energies of activation and reaction are presented for the rearrangement of chorismate to prephenate in CM and in water, for system sizes comprising up to 2000 atoms. Compared to the uncatalysed reaction, the calculated activation barrier is lowered by 10.5 kcal mol^{-1} in the presence of CM, in good agreement with experiment. In addition, a detailed analysis of the interactions between individual active-site residues and the bound substrate is performed, predicting the significance of individual enzyme sidechains in CM catalysis. These proof-of-principle applications of powerful large-scale DFT methods to enzyme catalysis will provide new insight into enzymatic principles from an atomistic and electronic perspective.

List of figures

List of tables

Chapter 1

Introduction

"The good news about computers is that they do what you tell them to do. The bad news is that they do what you tell them to do."

Ted Nelson

The very essence of science stands upon a foundation of observation. Experiments are devised, carried out, and interpreted, in an attempt to produce results concerning the nature of the world around us, adding to the existing body of knowledge. Theories are put in place in order to explain these results and aim to make further predictions to be tested through additional experiments, or perhaps to shine light on conflicting results. Whilst this is quite a simplistic view of the day-to-day undertakings of a scientist, the essential business of science remains true to reality. The theoretical and computational ideas harnessed in this dissertation have their roots in the seminal works of many distinguished scientists. Between 1918 and 1933, five Nobel prizes for Physics were awarded to the predominant developers of the theory of quantum mechanics (QM). These laureates were Max Planck, Niels Bohr, Louis de Broglie, Werner Heisenberg, Erwin Schrödinger and Paul Dirac, in chronological order. In addition, Albert Einstein's significant contributions cannot go unmentioned. These theoretical insights laid the foundations for the quantum chemical approach that won Walter Kohn and John Pople the prize for Chemistry in 1998. Considering earlier works, Johannes Diderik van der Waals and his eponymous interactions (awarded the prize for Physics in 1910), along with Charles-Augustin de Coulomb's influential contributions, inspired the development of models to describe intramolecular potentials based on approaches from classical physics. The classical models most widely used today have their origins in the work by the research groups of Frank Westheimer [1], Terrell Hill [2] and Sir Christopher Ingold [3], who in 1946 independently suggested how such approaches could be applied to molecules. It was Norman Allinger in 1965 who developed one of the first computer codes to optimise molecular structures in an empirical and classical framework [4]. Building upon Berni Alder and Thomas Wainwright's 1957 work on hard sphere molecular dynamics [5], not to mention Nicholas Metropolis' work on Markov chain Monte Carlo simulations predating it [6], Allinger's were the first real sets of methods that we would now recognise as molecular mechanics (MM). At the

same time, QM methods were being used in order to construct MM potentials. One of the first of such developments was Shneior Lifson and Arieh Warshel's consistent force field method [7] which Lifson used in collaboration with Michael Levitt to mininise the energy of a protein system [8]. Using a classical potential with terms constructed from underlying QM calculations is the basis of most of today's force fields. The problem was then how to accurately predict the structural coordinates of large systems. One of the greatest successes of attempting to meet this challenge stemmed from the initial work of Alder and Wainwright to develop a method of molecular dynamics (MD) to accurately predict the ionic configurations of a system. The use of classical potentials allows large systems and the associated time scales to be treated. However, the breaking and forming of chemical bonds can not be accurately treated in MM or MD. These processes can be described using conventional QM methods but, due to the computational costs, as will be discussed later in this dissertation, the system sizes accessible with these approaches do not reach the requirements for studying biomolecular systems. A particularly successful brand of QM-based approaches is density-functional theory (DFT). DFT calculations were initially employed mainly for the study of the electronic structure of simple solids, using a few atoms in a unit cell, with the use of periodic boundary conditions. However, following a huge effort to improve the accuracy and efficiency of the calculation techniques by Roberto Car and Michele Parrinello [9], the size of the target systems increased dramatically, but they were still not large enough to approach entire proteins. Density-functional approaches with a significantly reduced computational cost were first developed in 1991 by Weitao Yang [10] and have been developing ever since [11, 12]. However, large systems have many more degrees of freedom to explore and so require a greater computational effort in order to locate minimum energy structures. This requires the level of conformational sampling that is simply not feasible with QM-based calculations alone, so classical approaches are required. A significant step forward, that ties these two approaches together is that of combined quantum mechanics/molecular mechanics (QM/MM) [13]. It was the development of these "Multiscale Models for Complex Chemical Systems" that resulted in Martin Karplus, Michael Levitt and Arieh Warshel being awarded the 2013 Nobel prize for Chemistry. It is a strategy combining MM, QM/MM and full-QM that is presented in this dissertation. I therefore feel compelled to reproduce the oft-quoted line written by Sir Isaac Newton in a letter to Robert Hooke in 1676:

"If I have seen further it is by standing on the shoulders of giants"

and I feel this is an ever-present theme, not only in this dissertation but also in the natural sciences in general. I strongly believe this is something that practitioners in all fields should be aware of and generate a firm appreciation for.

1.1 Modelling and simulation: *In silico* techniques

A natural question might be asked as to where exactly computational simulations fit into the grand scheme of science. Simulations act to bridge pure theory, which is only able to give exact solutions to simpler, well-defined problems, and experiment, with all its inherent complexity. It has been remarked by multiple authors in the literature and in various academic meetings that developments in both simulation and experimental technologies have allowed us to approach the point whereby the system sizes addressed by both types of techniques coincide. Simulations can often fill a role of interpreting and interrogating experimental results, hopefully generating predictions or questions which help experiments. One fascinating example of this is the case of Hen Egg-White Lysozyme, where the work undertaken in the research group of Adrian Mulholland has resulted in the biochemistry text book explanation of the enzyme mechanism needing to be re-written because QM/MM calculations have revealed significantly different details about the reaction [14]. There are some cases where simulations can be used in order to investigate a new theory, but this is seldom done in the case of atomic calculations as many simulation methods have their roots in the well-founded theories of quantum mechanics and Newton's laws. At the very heart of atomistic simulations lies the calculation of the energy of a particular configuration of atoms and the associated forces. The bane of many computationalists' lives is ensuring the convergence of calculated properties with respect to simulation parameters. Convergence must be achieved in order to instil confidence in the results of the simulation. This aim is something that is continually strived for in this dissertation. Performing comparisons between well-known and relatively simple 'toy' problems before embarking on an investigation of a question of scientific interest is one way of attempting to ensure that simulations are adequately characterised and are capable of generating interesting and reliable data. Assuming that the outcome will be favourable without careful checking and treating simulation methods as a black box can be very dangerous and may lead to misleading results. Following on from careful testing of one's methods, the results of previous authors' simulations of your intended system of interest should be studied, where available. In doing so, this will demonstrate what previous methods have been applied to the system, give you a potential starting point for your own simulations, and highlight any associated problems with the system you intend to study. In general, the broad aims of the simulations should always be kept in mind. With a good set of aims to be adhered to, any new results or problems that will emerge throughout a study, can be reasonably managed.

1.2 Synergy between theory and experiment

An ever-present danger in having a tool as powerful as computational simulation is that it may be used as an end purely in itself. Whilst it may be possible to learn a lot about a

system from computational investigations alone, interaction with experiment and understanding the context within the larger scheme of the scientific process is key. Agreement between simulation and experiment is an important quantity to strive for in computational investigations. In the case of well-established experimental results, a new model can be tested to ensure it is performing correctly. Experimental results can also be explained and interpreted through the use of simulation if an existing mechanism is not currently decided upon. Computational methods can also push experiments forward through predicting what should be seen in otherwise unobserved scenarios. Just because a simulation can be run does not make it inherently interesting or experimentally relevant. This is a very difficult lesson to learn for computationalists. Collaboration with experimenters, or at least a firm understanding of the experimental process, should form a crucial part of any computational investigation. In a recent review covering sixty years of condensed matter physics [15], Phillipe Nozières stated:

> *"As a theorist I never forgot my short experimental stretch: I am always interested by what can be measured and how it can be measured. A dialogue between experiment and theory is a difficult venture, which requires a lot of patience on both sides to find a common language. When it succeeds it is incredibly rewarding."*

so clearly this difficult venture is one that will no doubt pay dividends if successful. There are many examples in the literature where a cursory nod is made in the direction of experiment only for the remainder of the paper to describe computational investigations with no relevance to the requirements of experiment. It is particularly easy to fall into this trap. For example, one could perform a calculation of a particular system at a temperature of 0 K but experimentally the observation may take place at 600 K ! Not only must one be able to translate between the language of the experimenter and the language of the theoretician, but one must also be able to ensure adequate exchange of ideas between researchers in separate disciplines. When starting to study systems of biological relevance from a physics perspective, one finds a convention that a physics education does not generally teach. This is the notion that biological systems have evolved, via natural selection, toward the powerful and complex functions necessary for life. This evolutionary process has left biological systems, at nearly every level, with an inherent heterogeneity. This presents a fundamental shift from what one encounters in physics, at the most basic level. For example, one is taught that an electron is exactly the same as all the other electrons that may surround it. This convention of the identical nature of particles is key to quantum mechanics and statistical physics. However, due to the inherent complexity encountered in biological systems, it is incredibly unlikely that any two 'identical' systems, be they protein molecules or two different cells in an organism, will ever be entirely identical. It is becoming increasingly apparent, in this ever interdisciplinary world, that the tools and expertise from one community can often be of use in investigating the problems of another community. One such example is when the importance of structural

heterogeneity in proteins was first revealed through experiments on myoglobin [16], using the cryogenic tools available to condensed matter physicists. Different processes of the myoglobin protein were separated using a wide range of temperatures, but this provoked widespread complaint that, as myoglobin operates at around room temperature, experiments performed at a temperature of 4 K were a waste of time. However, it was the knowledge that at room temperature the protein will be rapidly moving through many different configurations that led to the understanding that to probe the true molecular events involved in the reaction process, one has to decompose it into its components.

When it comes to comparison with experiment for biological systems, total energies are essentially meaningless. It is the differences between these values, or energy differences, that are more relevant, however, these still omit important entropic contributions. Therefore one should be aware of the experimental conditions and preparation of the system. Perhaps one of the most obvious quantities one might wish to compare with experiment would be the atomic geometry of the system. While reading off the three-dimensional coordinate and species of an atom from your simulation is easy, such structural data is often difficult to extract from experiment. In the case of biological systems such as proteins that need to be crystallised, there are structural domains that can be very difficult to resolve. Methods such as NMR imaging allow certain structural information to be extracted, such as interatomic distances and torsion angles. In the all important quest for experimental agreement, it would be all too easy to assert that if simulations generate the same result as experiment then the methods agree, and if not, they don't. However, investigations are rarely that simple, and it is frequently found that in investigating whether there is agreement between simulation and experiment that a further understanding of a system's underlying science is truly found. The observed differences between experiment and computation can arise from many factors, relating both to the methods used and the differences in what is being measured by each particular method. Returning to the ideas raised in the previous section regarding convergence, it can be tempting to halt convergence testing when sufficient agreement with experiment is achieved, but this is not a good code of practice to adhere to. The simulation is converged when it is converged, not just when it happens to match an experimental observation. Frequently, simulations are performed at zero temperature whereas the corresponding experiments are not, and this should, in general, give rise to a difference in the results. There are many other factors that can complicate comparisons between simulation and experiment, such as the level of impurities in an experimental sample, which may profoundly affect the material's properties, but which can be difficult to include in a simulation. Complicating factors in biological simulations include the concentration of substrate or any sort of buffers, which are solutions used in many biochemical techniques to maintain the pH of a solution in a fairly narrow range suitable for the particular process being investigated in experiment, which can be difficult to match in computational work. The calculated values of certain properties can often be affected by others and, therefore, in simulations, a key question

to consider when simulating water, for example, should be, is it better to simulate using experimental densities and temperatures or would it in fact be best to work relative to the calculated freezing point ? Whichever field it is applied to and whatever science underlies it, the testing of a simulation method is a vital part of any investigation and is a central theme in this dissertation.

1.3 Dissertation outline

The overview presented here is intended to act as an historical introduction to the relevant themes discussed within this dissertation. The following chapter serves as an introduction to – but far from a comprehensive outline of – the biological concepts relevant to the work presented in this dissertation. Chapter 3 outlines the computational methods that have been extensively developed by other authors, and those that have been implemented to study the systems discussed in later chapters. Chapter 4 presents a validation of the methods described in Chapter 3, by demonstrating their ability to accurately treat small molecules, and includes a discussion of their potential for application to larger biological systems. Chapter 5 investigates claims outlined by other authors that the methods described in Chapter 3 of this dissertation are unsuitable for the correct and reliable treatment of biomolecular systems. In so doing it aims to provide a methodology roadmap starting from experimentally resolved structures, moving through system preparation to eventual optimisation of these structures to ensure the resulting configurations make physical sense and demonstrate realistic properties whilst, at the same time, producing robust and reliable results. Using the methodological advances shown in Chapter 5, Chapter 6 investigates the rearrangement of chorismate to prephenate in water and in the presence of the *Bacilus subtilus* chorismate mutase (CM) enzyme. The aim of the chapter is to provide a demonstration of proof-of-principle calculations and to show that no single method, of those presented in Chapter 3, is necessarily the best choice for the task, but that a strategy combining these approaches can allow an accurate investigation of CM and biomolecular systems in general. Chapter 7 summarises the findings of this dissertation, provides a discussion of the implications of the investigations presented here, suggests ideas for additional development of the work and further areas into which the ideas presented in this dissertation can be explored.

Chapter 2

Proteins, enzymes and biological catalysis

"In biology, proteins are uniquely important... the most significant thing about proteins is that they can do almost anything. But their main function is to act as enzymes"

Francis Crick, *Society for Experimental Biology Symposium 1957*

In their simplest form, proteins essentially comprise unbranched polymer chains formed as a result of chemical bonding that takes place between the amino acid building blocks. This sequence of building blocks can be readily and rapidly determined via experimental means, building upon the pioneering work of Frederick Sanger who, in 1951, obtained the amino acid sequence of insulin [17], the first protein to have its sequence determined. However, these sequences give as much information about the biology of the system as a London telephone directory gives about the function and wonder of the city. The primary sequence formed by the amino acids then forms complex secondary structure through an intricate process of folding. The resulting structure of the protein in turn defines its function and in the case of enzymes, the particular type of reaction it catalyses. The objective of this chapter is to outline the key biological concepts used throughout this dissertation.

2.1 Amino acids

Hydrogen, carbon, nitrogen and oxygen constitute 96.5% of the mass of living cells [18]. This rises to 98% when taking sulphur and phosphorous into consideration. Therefore, it is clear that the chemistry of life is dominated by the lighter elements. It is these elements that also form the amino acids, or residues, which are the subunits of proteins and can be seen in Figure 2.1. The general chemical formula for an amino acid is $NH_2C^{\alpha}RHC'O_2H$. The four-fold coordinated central alpha-carbon atom (C^{α}) is sp^3 hybridised and is attached to a hydrogen atom, along with the amino group (NH_2) and carboxylic acid group ($C'O_2H$)

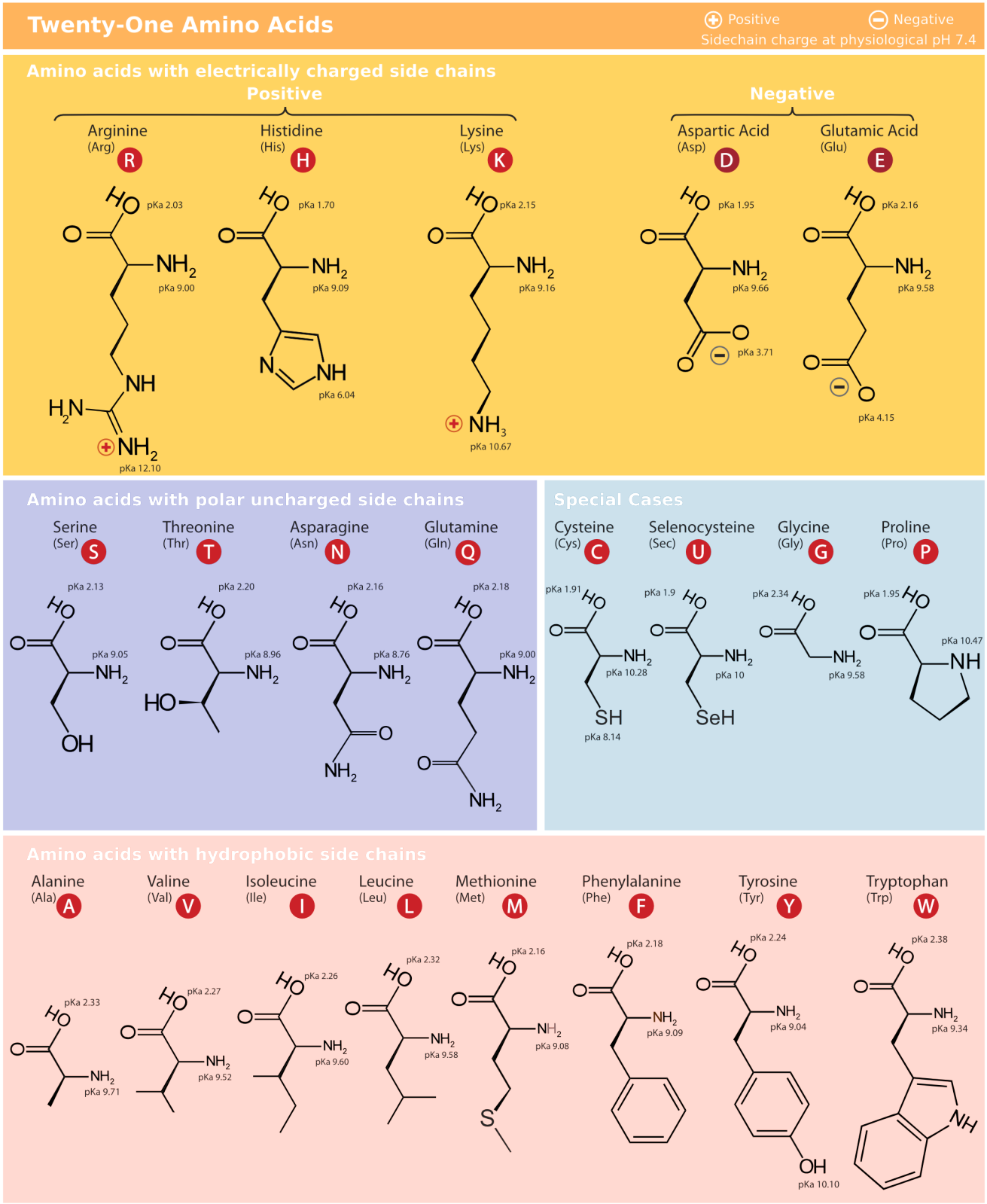


Figure 2.1: The twenty-one naturally occurring amino acids. Charged, polar uncharged and hydrophobic side chains are also highlighted. Figure adapted, and pKa data acquired, from Ref. [19].

and a side chain (R) by σ bonds. This bonding pattern is common to all the twenty-one amino acids. What distinguishes one amino acid from another is the side chain (R) attached to the alpha-carbon. The amino acids are usually divided into three classes, depending on the chemical nature of the side chain. Classes consist of amino acids with strictly hydrophobic side chains, those with charged residues and those with polar side chains. Most of the twenty-one naturally occurring amino acids were discovered in the 19^{th} century. How and why exactly this precise set of amino acids came to be chosen as the building blocks of life is one of the mysteries of evolution.

2.2 Protein structure

Proteins are formed in cells and are synthesised in ribosomes. Amino acids are joined into linked chains during this synthesis process when the carboxyl group of one amino acid condenses with the amino group of the next in order to eliminate water. This can be seen in Figure 2.2. This formation of peptide bonds is repeated as the chain elongates, generating the so-called 'backbone' from which the side chains project. The six atoms that surround each peptide bond are constrained in an arrangement close to planar, comprising the alpha carbon (C^{α}), carboxyl carbon (C') and amide nitrogen atoms [20]. The nitrogen, oxygen and subsequent alpha-carbon atoms are also close to coplanar. This is due to the adjacent nitrogen and carbon atoms in the N−H−C′=O unit being sp^2 hybridised. Their positions and resultant secondary structure can be defined in terms of the angles of rotation about the bonds connecting the three atoms. These angles of rotation are conventionally labeled as ψ, ϕ and ω, respectively. The peptide backbone dihedral, or torsion, angles are illustrated in Figure 2.3. The angle ϕ defines the rotation of the plane containing C_i^{α}, C_i', O_i and N_{i+1} around the $N_i{-}C_i^{\alpha}$ bond, controlling the $C'{-}C'$ distance. ψ defines the rotation of the plane containing C_i', O_i and N_{i+1} around the $C_i^{\alpha}{-}C_i$ bond and controls the N−N distance. ω defines rotation around the peptide bond $C'{-}N_{i+1}$ and controls the $C^{\alpha}{-}C^{\alpha}$ distance but in general is restricted to be close to 180° by the planar nature of the peptide bond, therefore ω describes any deviation from planarity. One consequence of the condensation process that leads to the formation of proteins is that the amino group of the first amino acid and the carboxyl group of the last amino acid remain intact. Thus a polypeptide is said to run from its amino (N−) terminus to its carboxy (C−) terminus. The sequence of amino acids from which a protein is built is termed its primary structure. One of the first important general principles to emerge from protein structure studies was the fact that amino acids in the interior of proteins have almost exclusively hydrophobic side chains. However, in order to form the compact and folded protein structure seen in nature, new interactions are required to compensate for the solvent interactions lost from the peptide background. Thus, there is a major barrier to creating such a hydrophobic core from a protein chain. In order to bring the side chains into the core, the main chain also needs to fold into the interior. Each peptide

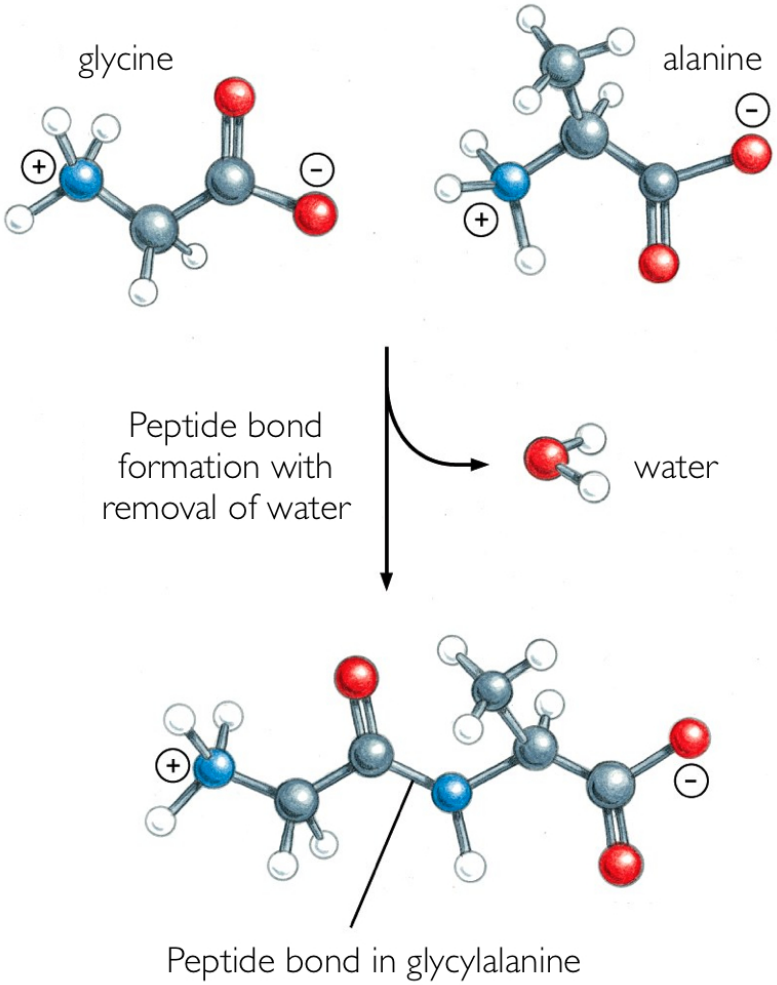


Figure 2.2: The condensation process that leads to peptide bond formation. Figure adapted from Ref. [18].

unit on the backbone has one hydrogen-bond donor (the N−H group) and one hydrogen-bond acceptor (the C′=O group) resulting in a very polar and hydrophilic backbone. In order to replace the favourable interactions the backbone would have with the solvent in an unfolded state, a more compact, folded structure is required. Proteins solve this problem by forming secondary structures where the backbone N−H and C′=O groups form intramolecular Hydrogen bonds with each other.

A protein will fold into a stable configuration, or secondary structure, determined by its primary structure of amino acids. Although the secondary structure of proteins can be incredibly varied, there are two commonly recurring motifs. These α helices and β sheets, as illustrated in Figure 2.4, are recurring patterns in protein structures and are recognisably similar in virtually all natural proteins, despite varying in size and amino acid composition. These ideas were first put forward by William Astbury in 1933 when investigating keratin and collagen. Astbury proposed that unstretched protein molecules formed a helix, which he called the α-form, and stretching caused the helix to uncoil,

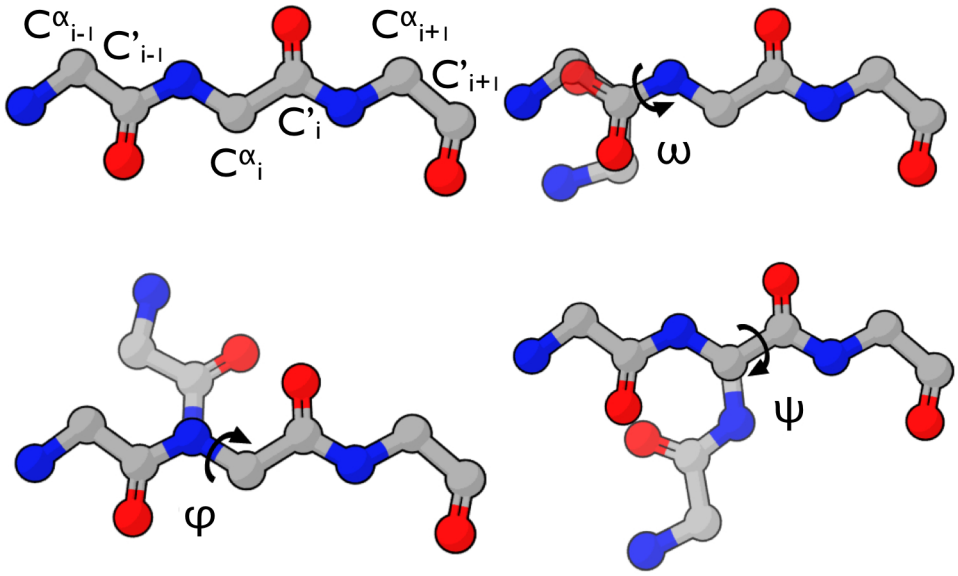


Figure 2.3: Peptide backbone dihedral angles.

forming an extended state which he called the β-form [21]. Whilst the details of the Astbury model were incorrect, they correspond to the modern ideas of secondary structure which were later refined by Pauling, Robert Corey and Herman Branson in 1951, where Astbury's original α and β notation was retained [22]. The comformation formed by an entire protein chain, including many secondary structure motifs, is termed its tertiary structure. In addition, if a protein is part of a complex of multiple polypeptide chains then the complete structure is termed the quaternary structure. This last concept remains outside the scope of this dissertation. Proteins are involved in many diverse functions ranging from maintaining the chemical potential across cell membranes to replicating DNA. However, most importantly and most relevant to this dissertation, proteins are actively engaged as enzymes in the catalysis of complex chemical reactions.

2.3 Enzyme catalysis

The cells in a living organism carry out a never-ending series of chemical reactions. This very often involves rearranging small organic molecules in a set of steps along some metabolic pathway. The molecules at the start of this process will usually be the result of photosynthesis in plants or the ingestion of food in mammals. The subsequent pathway will then modify the input molecules sufficiently to meet the requirements of the cells in the living system. Each cell performs many millions of these reactions every second. However, the vast majority of the reactions that take place would normally not happen at the mild temperatures and pressures found in the cell. The key to accelerating, or catalysing, these reactions comes in the form of enzymes. The primary function of

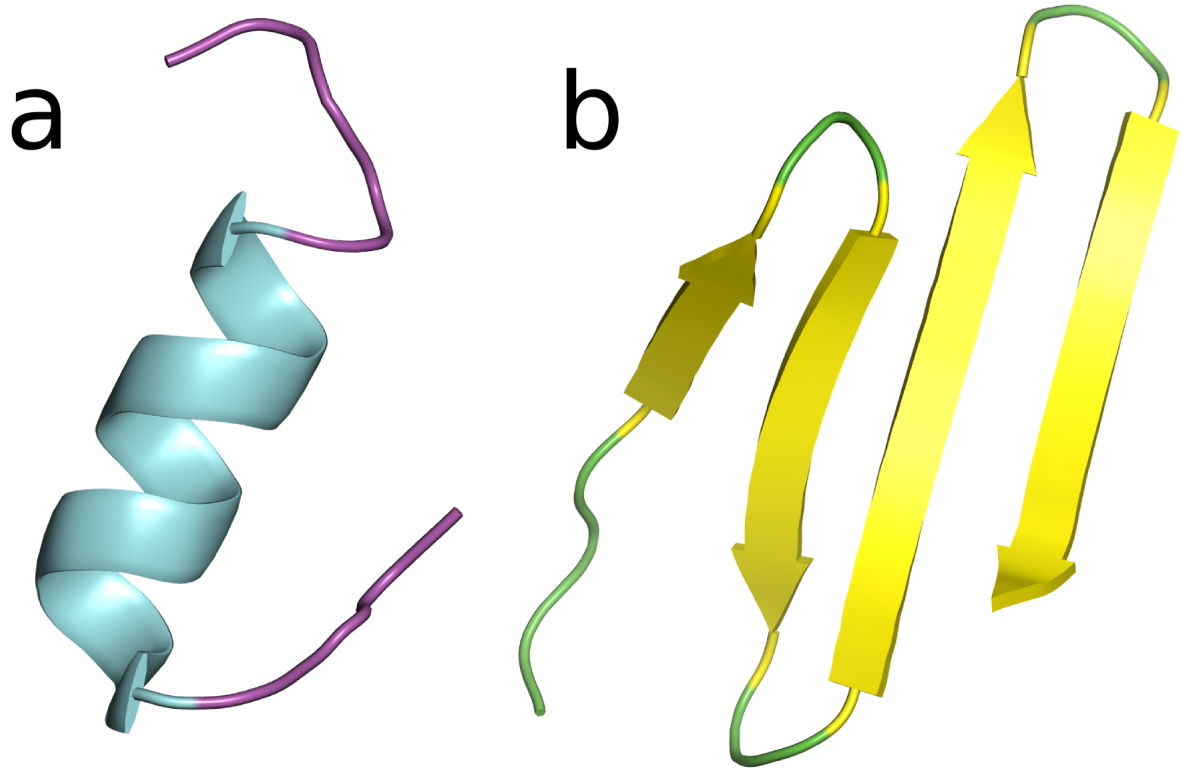


Figure 2.4: (a) α helix and (b) β sheet secondary structure motifs.

an enzyme is to accelerate the reaction rate of a particular chemical reaction relative to the equivalent uncatalysed reaction, or to make a reaction happen that would not occur spontaneously. Enzymes are known to catalyse around 4,000 biochemical reactions [23] with many reaction rates on the order of millions of times faster than the equivalent uncatalysed reactions. The initial ideas laid down by Emil Fischer and his 'lock-and-key' model [24] were used to explain the specificity found in enzymes through the fact that both the enzyme (the 'lock') and the substrate (the 'key') were thought to possess specific complementary geometric shapes that fit exactly into one another. However, while this is an excellent model for describing enzyme-substrate specificity, it does not adequately explain how enzymes manage to catalyse these chemical reactions. The work of Henry Eyring, Meredith Evans and Michael Polanyi [25, 26] in the 1930s, and Linus Pauling [27] in the 1940s, revolutionised the theory of enzyme catalysis by hypothesising transition state structures. It was Pauling's further proposal that the powerful catalytic action of enzymes could be explained by specific tight binding to the transition state species in Ref. [27] that initially led the ideas of transition state stabilisation by enzymes, and would lead the way to the modern transition state theory [28].

It is now widely accepted that enzymes function to stabilise the transition states lying between the reactants and products in the chemical reactions they are catalysing. This

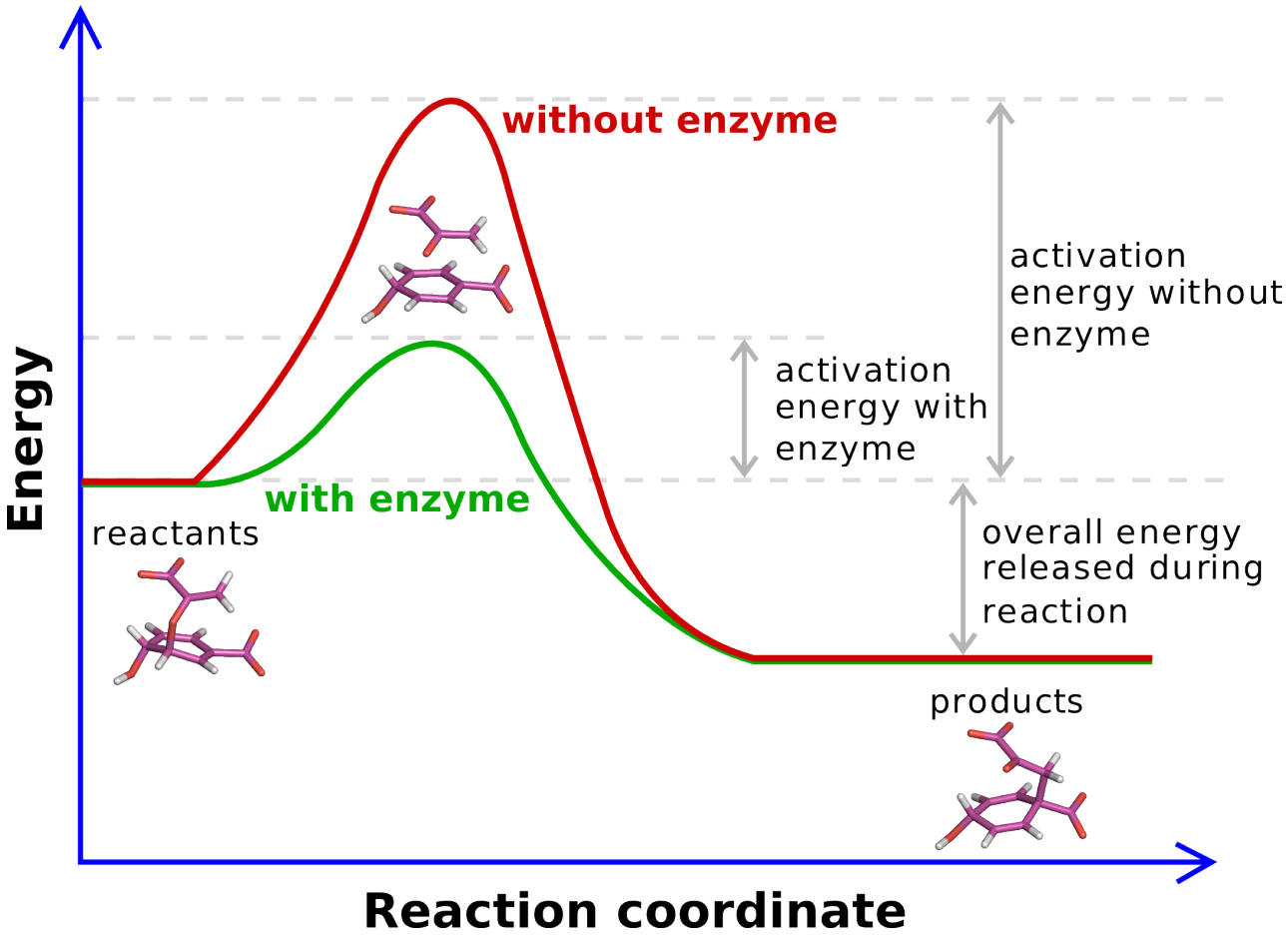


Figure 2.5: Schematic diagram illustrating how the activation barrier is reduced, compared to the equivalent reaction in the absence of the enzyme, through stabilising the transition state in the presence of the enzyme.

stabilisation dramatically reduces the activation energy required for the reaction to take place, therefore greatly accelerating the rate of reaction. An illustration comparing the activation barriers for a reaction in the presence of an enzyme and the equivalent reaction in the absence of the enzyme can be seen in Figure 2.5. Drawing from the conclusions that an enzyme binds strongly to its particular transition state, the enzyme could also be expected to bind strongly to any synthesised molecule which closely resembles the ionic structure of such a transition state. Whereas reactant and products often participate in several enzyme reactions, the transition state tends to be characteristic of one particular enzyme. Therefore any inhibitor, or transition state analogue, would need to be specific for that particular enzyme. The identification of transition state analogues, for a range of targets [29–32], further supports the transition state stabilisation hypothesis for enzymatic catalysis. How exactly this transition state stabilisation arises is still a topic of debate amongst enzymologists, and the study of the precise mechanisms involved in, and origins of, enzyme catalysis is an active area of research.

Many authors propose that the stabilisation arises mainly due to the favourable

Coulombic interactions between the enzyme and the substrate. Therefore, it is crucial to treat the electrostatics of the, often polar, enzyme active site accurately. Enzymes can alter the electronic structure of their constituent substrates via protonation, proton abstraction, electron transfer, geometric distortion, hydrophobic partitioning and interaction with Lewis acids and bases. This is usually achieved through short-range forces from noncovalent bonds such as van der Waals interactions, electrostatic interactions and hydrogen bonds. A hydrogen bond is the attractive interaction between polar molecules where hydrogen is bound to a highly electronegative atom, such as nitrogen or oxygen, forming an attractive interaction with another atom, such as OH $---$ N. The hydrogen bond is directional and so is at its strongest when the three atoms involved are aligned. Electrostatic interactions occur between partially charged groups on polar molecules, such as the charged amino acids. At very short distances, any two atoms will show a weak van der Waals interaction, due to their fluctuating electron densities. These three types of weak bonds have less than 1/20 the strength of a standard covalent bond [18]. However, despite a single example of any of these bonds being relatively weak compared to a covalent bond, many of them can form together to create a strong bonding arrangement that stabilises a particular three-dimensional structure. These bonds involve atoms not only in the polypeptide backbone but also the amino acid side chains. The stability of each folded shape is significantly dependent upon the combined strength of large numbers of these noncovalent bonds.

Whilst performing as an enzyme may seem like just another function in the long list of jobs that proteins carry out in the cell, the collosal, unmitigated catalytic power of enzymes is extraordinary [33]. The incredible efficiency demonstrated by the OMPase enzyme, taking a reaction that would otherwise have a half life of 78 million years in solution, to complete in just 18 milliseconds [34] is simply breathtaking. The role of enzymes as biological catalysts is clearly critical for life as just under half of all gene products are annotated as having enzymatic function [35]. In addition, enzymes are often the targets of pharmaceutical development with a significant fraction of approved clinical drugs modifying the behaviour of enzymes implicated in human disease along with disease-causing pathogens [36]. Nearly half of all marketed small molecule therapeutics are designed as enzyme inhibitors [37]. It is argued that ligand design can benefit greatly from improved knowledge of enzyme mechanisms and key active-site interactions [38]. In addition, increasing importance is being given to the prediction of enzyme-mediated adverse reactions [39] and drug metabolism [40]. An understanding of the electronic, atomic and molecular origins of how enzymes achieve their catalytic rate enhancements is a long-standing problem in biochemistry and, increasingly, within computational biology. In many cases, experimental observation alone is not able to establish the mechanisms of enzyme-catalysed reactions and the origins of catalysis, due to a lack of detailed microscopic information regarding the transition state of the reaction in the enzyme.

Transition states are central to many of the fundamental questions that surround

chemical reactivity; the stabilisation of such states is a highly important process for the efficiency of catalysis within enzymes. Transition state complexes can often prove very difficult to observe directly in experiment due to their extremely short lifetimes, typically picoseconds. However, it should be noted that the development of femtosecond transition state spectroscopic techniques is currently an active area of experimental research [41–43]. Computational modelling could, potentially, complement experiment in this task as it has the ability to probe and analyse enzyme transition state configurations directly. It is becoming increasingly apparent that molecular simulation has a vital role to play in elucidating the complex processes involved in these outstanding natural catalysts. From the perspective of practical applications, modelling techniques that can shed new light on enzyme-catalysed reactions can then help to contribute toward the design of new drugs or the development of novel industrial catalysts via biomimetic approaches. The concept of using atomistic simulations to model enzyme-catalysed reactions, starting from the first pioneering works of Arieh Warshel [13] and Steve Scheiner [44], has risen to prominence in recent years and is now at the point where the field of computational enzymology has securely laid foundations [45, 46]. However, there still remains little consensus about the ideal methodology to perform calculations on an enzyme of choice. Whilst it is outside the scope of this dissertation to discuss the matter in detail, a more comprehensive discussion of methods currently used, along with their associated advantages and disadvantages, can be found in an elegant recent review by Richard Lonsdale [47]. Elucidation of the origins of enzyme catalysis involves understanding the origin of the difference between the uncatalysed activation barrier and the activation barrier in the protein, along with the associated enzyme mechanisms. The primary focus is on the factor that governs the reduction of the activation barrier of the chemical step. This is, of course, a question of energetics. One of the main objectives of this dissertation is to provide energies of activation and reaction for the rearrangement of chorismate to prephenate, both in the presence of the *Bacillus subtilis* chorismate mutase (CM) enzyme and also the uncatalysed equivalent reaction in water.

2.4 Summary

This chapter has outlined the biological concepts relevant to the investigations presented in this dissertation. Starting from the fundamental building blocks of nature, the amino acids, the discussion moved on to show how these component parts form larger polypeptide chains. The types of secondary structure motifs that these long chains fold into was then outlined. One of the essential dogmas of biology is that structure informs function. Once the polypeptide chains discussed fold into their correct structures they can then perform a variety of functions. One such function is to catalyse reactions that would otherwise take too long to be of biological relevance, demonstrating how enzymes are critical for life. The next chapter will discuss the computational methods used in this work to study

proteins and the tests required to ensure that these methods accurately describe the physical properties we know are crucial for correctly describing the biochemistry of these systems. Chapter 4 will investigate the properties of small molecules that can adopt torsion angles, which correspond to those outlined in this chapter, to match those of amino acids that form extended α-helices and β-sheets discussed here. If these properties can be reproduced using the ONETEP and `OPTIM` codes, discussed in the following chapter, then the resulting simulations performed on larger systems in Chapters 5 and 6 can be trusted. Transition state stabilisation, a key process in reducing the activation barrier, occurs, partially, as a result of efficient overlap of electron orbitals between the residues in the enzyme active site. The following chapter will outline accurate and efficient methods for optimising the ionic and electronic structure of enzyme systems and discuss how the bonding interactions between active-site residues and a substrate can be probed. These interactions in the active site of the CM enzyme can be analysed in detail, and this will be investigated in Chapter 6.

Chapter 3

Computational techniques

"Let us, as nature directs, begin with first principles"

Aristotle, *Poetics* I

The major clash between the mechanics of the classical and that of the quantum is the fundamental property of uncertainty. Einstein's retort that *"God does not play dice"* would be forever a hallmark of the stubbornness of staunch believers in a classical deterministic view of everything in the universe from the galaxies in the furthest region imaginable to the individual protons and electrons that make up everything around us. The unreasonable nature of the fear of uncertainty and the probabilistic nature of matter on the small scale can be adequately reflected through an excerpt from the writings of John Locke [48]:

> *"If we will disbelieve everything, because we can not certainly know all things; we shall do muchwhat as wisely as he, who would not use his legs, but sit still and perish because he had no wings to fly"*

In essence, quantum mechanics brings a vast arsenal of machinery at the disposal of physicists, chemists and now, seemingly, even biologists [49], allowing scientists of any discipline to benefit from the elegant principles contained within the theory. However, as outlined in the introduction to this dissertation, it is not quantum mechanical approaches alone that are required to accurately describe biomolecular systems. In order to reduce computational expense and to greatly expand the sample size from which results can be extracted, classical approaches must also be used. This chapter should be regarded as an overview of the many key theoretical ideas and methods implemented in this dissertation and not a place for detailed discussion.

3.1 Many-body quantum mechanics

Many problems related to the electronic structure of matter – not including relativistic effects, magnetic fields and quantum electrodynamics, can be adequately and accurately

described by the equation due to Erwin Schrödinger [50]. The equation was published in 1926 and was soon applied to multi-electronic atoms and to polyatomic systems such as molecules [51] and solids [52]. The aim of these works was to find a description of matter at the atomic scale, i.e. in terms of atomic nuclei and electrons. In general terms, one can imagine a piece of matter as a collection of interacting atoms, This ensemble of particles may be in the gas phase (molecules, clusters) or in a condensed phase (bulk solids, surfaces, wires). It could be in a solid, liquid or amorphous phase, either homogeneous or heterogeneous (molecules in solution, interfaces, adsorbates on surfaces). However, at this scale, one can unambiguously describe all these systems as a set of atomic nuclei and electrons interacting via coulombic, electrostatic forces. Formally, one can write the Hamiltonian of such a system in the following general form:

$$\begin{aligned}\hat{\mathcal{H}} = &-\frac{\hbar^2}{2m_e}\sum_{i=1}^{N}\nabla_i^2 - e^2\sum_{I=1}^{P}\sum_{i=1}^{N}\frac{Z_I}{|\mathbf{r}_i-\mathbf{R}_I|} + \frac{e^2}{2}\sum_{i\neq j}^{N}\sum_{j\neq i}^{N}\frac{1}{|\mathbf{r}_i-\mathbf{r}_j|} - \frac{\hbar^2}{2}\sum_{I=1}^{P}\frac{\nabla_I^2}{M_I} \\ &+ \frac{e^2}{2}\sum_{I=1}^{P}\sum_{J\neq I}^{P}\frac{Z_I Z_J}{|\mathbf{R}_I-\mathbf{R}_J|} \\ &= \hat{T}_{\mathrm{e}} + \hat{V}_{\mathrm{ne}} + \hat{V}_{\mathrm{ee}} + \hat{T}_{\mathrm{nn}} + \hat{V}_{\mathrm{nn}} \end{aligned} \tag{3.1}$$

with electrons, with charge e and mass m_e denoted by lower case subscripts and nuclei with charge Z_I and mass M_I denoted by upper case subscripts. $\mathbf{R} = \{\mathbf{R}_I, I = 1, \ldots, P\}$ is a set of P nuclear coordinates. $\mathbf{r} = \{\mathbf{r}_i, i = 1, \ldots, N\}$ is a set of N electronic coordinates. $\hat{T}_{\mathrm{e}}$ is the electron kinetic energy, $\hat{V}_{\mathrm{ne}}$ is the electron-ion interaction, $\hat{V}_{\mathrm{ee}}$ is the electron-electron interaction, $\hat{T}_{\mathrm{nn}}$ is the nuclear kinetic energy and $\hat{V}_{\mathrm{nn}}$ is the ion-ion interaction. One can then use this Hamiltonian to solve the time-independent Schrödinger equation:

$$\hat{\mathcal{H}}\Psi_n(\mathbf{R},\mathbf{r}) = \epsilon_n\Psi_n(\mathbf{R},\mathbf{r}) \tag{3.2}$$

where ϵ_n are the energy eigenvalues and $\Psi_n(\mathbf{R},\mathbf{r})$ are the corresponding eigenstates, or wave functions, which must be antisymmetric with respect to exchange of electronic coordinates in $\mathbf{r}$, since the electrons are fermions and the total electronic wave function should change sign whenever the coordinates of any two electrons are exchanged, and symmetric or antisymmetric with respect to exchange of nuclear variables in $\mathbf{R}$. Different nuclear species are distinguishable, but nuclei of the same species also obey a specific statistics according to the nuclear spin. They are fermions for half-integer nuclear spin (e.g. H, ^{3}He) and bosons for integer spin (e.g. D, ^{4}He, H_2). At the atomic energy scales which are the focus of this dissertation, the nuclei are extremely well-described as massive point charges and their internal structure is safely neglected. The wave functions are single-valued, square-integrable functions of the system parameters and provides a complete description of the system. Linear Hermitian operators act on a wave function and correspond to the physical observables, those dynamical variables which can be measured, e.g. position, momentum and energy.

In practice, the problem posed in equation (3.2) is almost impossible to treat within a full quantum mechanical framework. There are only a few cases, such as hydrogenoid atoms or the H_2^+ molecule, where a complete analytic solution available. Exact numerical solutions are also limited to a few cases, mostly atoms and very small molecules. There are several features that contribute to this difficulty, but the most important is that this is a multi-component, many-body system, and the two-body nature of the Coulomb interaction makes equation (3.2) not separable.

Confining the problem to the case of an atom with Z electrons, and focusing on the electronic wave function, in order to respect the antisymmetry of the wave function against electron exchange, such a wave function can in principle be written as an antisymmetrised product of one-electron wave functions (a so-called Slater determinant). This assumes, however, some kind of separability of the Schrödinger equation, implying that the probability of finding an electron at some point in space is essentially independent of where the other electrons are located. The repulsive electron-electron interaction is quite at odds with this picture, because an electron located at point $\mathbf{r}$ in space precludes other electrons from approaching this location. Hence, the probability of finding an electron at $\mathbf{r}$ depends on the location of the other $Z-1$ electrons. This phenomenon is known as correlation, and it implies that the exact many-body wave function should contain factors depending on two electronic coordinates. Therefore, the image in terms of one-electron wave functions can be somewhat crude in many cases. This means that the full Schrödinger equation cannot be easily decoupled into a set of equations, so that, in general, we have to deal with $3(P+N)$ coupled degrees of freedom. The usual choice is to resort to a few reasonable and well-controlled approximations, which encompass a wide variety of problems of interest. This can be achieved through two major approximations: the adiabatic separation of the nuclear and electronic degrees of freedom, and the classical treatment of atomic nuclei.

The time scale associated with the motion of nuclei is usually much slower than that associated with electrons. The most unfavorable case of a single proton already corresponds to a mass ratio of 1:1836, i.e. less than 1%. Within a classical picture one could say that, under typical conditions, the velocity of the electron is much larger than that of the heavy particle (the proton). In 1927, Max Born and his student, Julius Robert Oppenheimer, proposed a scheme for separating the motion of nuclei from that of the electrons [53]. They showed that no mixing of different electronic stationary states happened due to the interaction with the nuclei. Therefore, under appropriate conditions, the electrons do not undergo transitions between stationary states. This is called the adiabatic approximation. The electrons can then be thought of as instantaneously following the motion of the nuclei, while remaining always in the same stationary state of the electronic Hamiltonian. As the nuclei follow their dynamics, the electrons instantaneously adjust their wave function according to the nuclear wave function. This approximation ignores the possibility of having non-radiative transitions between different electronic eigenstates. Transitions can only arise through the coupling with an external electromagnetic field,

but this issue will not be addressed in this dissertation.

These ideas can be cast in a more mathematical framework by proposing a solution to equation (3.2) in the form of:

$$\Psi(\mathbf{R},\mathbf{r},t)=\sum_{n}\Theta_{n}(\mathbf{R},t)\Phi_{n}(\mathbf{R},\mathbf{r}) \tag{3.3}$$

where $\Theta_n(\mathbf{R},t)$ are wave functions describing the evolution of the nuclear sub-system in each one of the adiabatic electronic eigenstates $\Phi_n(\mathbf{R},\mathbf{r})$. These satisfy the time-independent Schrödinger equation:

$$\hat{h}_{\mathrm{e}}\Phi_{n}(\mathbf{R},\mathbf{r})=E_{n}(\mathbf{R})\Phi_{n}(\mathbf{R},\mathbf{r}) \tag{3.4}$$

where the electronic Hamiltonian is defined as:

$$\hat{h}_{\mathrm{e}}=\hat{T}+\hat{U}_{\mathrm{ee}}+\hat{V}_{\mathrm{ne}}=\hat{\mathcal{H}}-\hat{T}_{\mathrm{n}}-\hat{V}_{\mathrm{nn}} \tag{3.5}$$

In this partial differential equation on the $\mathbf{r}$ variables, the $3P$ nuclear coordinates $\mathbf{R}$ enter as parameters. This expansion, which is always mathematically possible, is called the expansion in the adiabatic basis, because $\Phi_n(\mathbf{R},\mathbf{r})$ are solutions of the time-independent electronic Schrödinger equation, corresponding to a particular nuclear configuration. Equation (3.4) has to be solved for all nuclear configurations $\mathbf{R}$ where the nuclear wave function is non-vanishing. By replacing this ansatz into the full Schrödinger equation one obtains a set (infinite, in principle) of coupled partial differential equations containing off-diagonal terms. The off-diagonal terms will mix (excite) the different electronic eigenstates along the temporal evolution. These are precisely the non-radiative transitions alluded to previously. If this is the case, then the dynamics is said to be non-adiabatic. However, if the off-diagonal terms can be neglected, then an expression like (3.3) is valid because the nuclear dynamics has no means to cause electronic transitions, and the electrons remain always in the same (n) adiabatic state (ground or excited). In this case, the dynamics is said to be adiabatic. The necessary condition for neglecting the non-adiabatic couplings is that:

$$\frac{m}{M}\left|\frac{\hbar\Omega_{v}}{E_{q}(\mathbf{R})-E_{n}(\mathbf{R})}\right|\ll 1 \tag{3.6}$$

where Ω_v is the maximum frequency of rotation of the electronic wave function due to the nuclear motion, and the energies in the denominator correspond to the electronic adiabatic eigenstates (the energy gap if $q=1$ and $n=0$). The ratio of electronic to nuclear mass m/M is always smaller than 5×10^{-4}, thus justifying the adiabatic approximation unless a very small gap occurs, as for open-shell, conical intersections or Jahn-Teller systems. The case of lighter particles such as muons would be different. Typical electronic excitations are of the order of 1 eV, while typical nuclear excitations (phonons) are of the order of 0.01 eV. This indicates that there is a clear separation of energy (and consequently time) scales. There are situations in which this approximation is not adequate, but they are rather exceptional cases and shall not be addressed in this dissertation.

3.2 Density-functional theory

The full many-body wave function is not a single Slater determinant, otherwise Hartree-Fock theory would be exact. A different line of thought to solving the electronic Hamiltonian drove Llewellyn H. Thomas [54] and Enrico Fermi [55] to propose that the full electronic density was the fundamental variable of the many-body problem. From this idea they derived a differential equation for the density without resorting to one-electron orbitals. The Thomas-Fermi approach was developed in the hopes that the energy can in fact be written exclusively in terms of the electronic density. The original Thomas-Fermi approximation was actually too crude, mainly because the approximation used for the kinetic energy of the electrons was unable to sustain bound states. This idea, however, was intuitive at the time, but a proof that this was the case had to wait more than thirty years. In 1964, Pierre Hohenberg and Walter Kohn formulated and proved a theorem that put on solid mathematical grounds the former ideas [56]. The theorem is divided into two parts.

The first part of the Hohenberg-Kohn theorem (HK1) states that the external potential is uniquely determined by the electronic density, besides some trivial additive constant. In order to prove HK1, one should first assume the opposite to be true, i.e. that the external potential is not uniquely determined by the density. In this case one should be able to find two potentials, ν and ν', such that their ground state density n is the same. Let Φ and $E_0 = \langle\Phi|\hat{H}|\Phi\rangle$ be the ground state wave function and ground state energy of the Hamiltonian n $\hat{H} = \hat{T} + \hat{V}_{\text{ext}} + \hat{U}_{\text{ee}}$. One should also let Φ' and $E_0' = \langle\Phi'|\hat{H}'|\Phi'\rangle$ be the ground state wave function and ground state energy of the Hamiltonian $\hat{H}' = \hat{T} + \hat{V}'_{\text{ext}} + \hat{U}_{\text{ee}}$. According to the Rayleigh-Ritz variational theorem, HK1 then asserts:

$$\begin{aligned} E_0 < \langle\Phi'|\hat{H}|\Phi'\rangle &= \langle\Phi'|\hat{H}'|\Phi'\rangle + \langle\Phi'|\hat{H} - \hat{H}'|\Phi'\rangle \\ &= E_0' + \int n(\mathbf{r}) \left(\nu_{\text{ext}}(\mathbf{r}) - \nu'_{\text{ext}}(\mathbf{r})\right) \mathrm{d}\mathbf{r} \end{aligned} \tag{3.7}$$

where HK1 uses the fact that different Hamiltonians necessarily correspond to different ground states $\Phi \neq \Phi'$. Since the potential is a multiplicative operator, one can exchange the roles of Φ and Φ' (and $\hat{H}$ and $\hat{H}'$) to obtain:

$$\begin{aligned} E_0' < \langle\Phi|\hat{H}'|\Phi\rangle &= \langle\Phi|\hat{H}|\Phi\rangle + \langle\Phi'|\hat{H}' - \hat{H}|\Phi'\rangle \\ &= E_0 - \int n(\mathbf{r}) \left(\nu_{\text{ext}}(\mathbf{r}) - \nu'_{\text{ext}}(\mathbf{r})\right) \mathrm{d}\mathbf{r} \end{aligned} \tag{3.8}$$

where upon adding the inequalities (3.7) and (3.8) gives $E_0 + E_0' < E_0' + E_0$, which is absurd. Therefore, by proof *ab adsurdum*, there can not be $\nu_{\text{ext}}(\mathbf{r}) \neq \nu'_{\text{ext}}(\mathbf{r})$ that correspond to the same electronic density for the ground state, unless they differ by some trivial additive constant. There is a corollary to this proof in that since $n(\mathbf{r})$ uniquely determines $\nu_{\text{ext}}(\mathbf{r})$, it also determines the ground state wave function Φ, which should be obtained by obtaining the full many-body Schrödinger equation.

The second part of the Hohenberg-Kohn theorem (HK2) begins with $\tilde{n}(\mathbf{r})$ which is a non-negative density, normalised to N. One can then define the variational energy E_ν:

$$E_\nu[\tilde{n}] = F[\tilde{n}] + \int \tilde{n}(\mathbf{r})\nu_{\text{ext}}(\mathbf{r})\mathrm{d}\mathbf{r} \tag{3.9}$$

where $F[\tilde{n}]$ is defined as:

$$F[\tilde{n}] = \langle\Phi[\tilde{n}]|\hat{T} + \hat{U}_{\text{ee}}|\Phi[\tilde{n}]\rangle \tag{3.10}$$

where $\Phi[\tilde{n}]$ is the ground state of a potential which has $\tilde{n}$ as its ground state density, such that $E_0 = E_\nu[n]$ verifies $E_0 < E_\nu[\tilde{n}]$ for any $\tilde{n} \neq n$, and is thus the ground state energy. HK2 then considers:

$$\begin{aligned}\langle\Phi[\tilde{n}]|\hat{H}|\Phi[\tilde{n}]\rangle &= F[\tilde{n}] + \int \tilde{n}\nu_{\text{ext}}(\mathbf{r})\mathrm{d}\mathbf{r} \\ &= E_\nu[\tilde{n}] \geq E_\nu[n] = E_0 = \langle\Phi[n]|\hat{H}|\Phi[n]\rangle\end{aligned} \tag{3.11}$$

where the inequality effectively follows from the Rayleigh-Ritz variational principle for the wave function, but instead applied to the electronic density. Therefore, the variational principle states that:

$$\delta\left\{E_\nu[n] - \mu\left(\int n(\mathbf{r})\mathrm{d}\mathbf{r} - N\right)\right\} = 0 \tag{3.12}$$

which leads to:

$$\mu = \frac{\delta E_\nu[n]}{\delta n} = \nu_{\text{ext}}(\mathbf{r}) + \frac{\delta F[n]}{\delta n} \tag{3.13}$$

The knowledge of $F[n]$ implies the knowledge of the solution of the full many-body Schrödinger equation. $F[n]$ is the so-called *universal* functional, which does not depend explicitly on the external potential, it depends only on the electronic density. In the Hohenberg-Kohn formulation, $F[\tilde{n}] = \langle\Phi[\tilde{n}]|\hat{T} + \hat{U}_{\text{ee}}|\Phi[\tilde{n}]\rangle$ where Φ is the ground state many-body wave function. HK1 and HK2 form the mathematical basis of DFT.

In the Hohenberg-Kohn theorem the electronic density determines the external potential. However, it is also required that the density corresponds to some ground state antisymmetric wave function. While this is a necessary condition for the *true* density n, it may not be the case for other trial densities $\tilde{n}$. In fact, unacceptable densities can easily be obtained in a variational search strategy if this is not done carefully. With this observation in mind, in 1982 Mel Levy reformulated DFT in such a way that the antisymmetric origin of the density is guaranteed [57]. Levy used the constrained search method, which was then widely applied by several authors in similar contexts. The main idea is to redefine the universal functional $F[n]$ given by Expression (3.10) in the following way:

$$F[n] = \min_{\Phi\to n}\left\{\langle\Phi|\hat{T} + \hat{U}_{\text{ee}}|\Phi\rangle\right\} \tag{3.14}$$

where n is any non-negative density such that $\int n(\mathbf{r})\mathrm{d}\mathbf{r} = N$ and $\int |\nabla n^{\frac{1}{2}}(\mathbf{r})|^2\mathrm{d}\mathbf{r} < \infty$, with the additional constraint that the density should arise from an antisymmetric wave function. The search is thus constrained to the subspace of all the antisymmetric Φ that

give rise to the *same* density n, thus eliminating the conceptual difficulty of possible unphysical densities.

Using DFT one can determine the electronic ground state density and energy exactly, provided that $F[n]$ is known. In fact, since the density determines the potential uniquely, by solving the full many-body Schrödinger equation, one can determine uniquely the many-body wave functions, ground *and* excited states. In 1965, Kohn and Sham devised a practical scheme for determining the ground state [58]. The main problem at this stage is with the kinetic energy:

$$T = \langle \Phi | \hat{T} | \Phi \rangle = -\frac{\hbar^2}{2m} \sum_{i=1}^{N} \langle \Phi | \nabla_i^2 | \Phi \rangle = -\frac{\hbar^2}{2m_e} \int \left[\nabla_{\mathbf{r}}^2 \rho_1(\mathbf{r}, \mathbf{r}') \right] \tag{3.15}$$

because its explicit expression in terms of the electronic density is not known. According to equation (3.15) the exact calculation of the kinetic energy term requires the knowledge of the Laplacian of the one-body density matrix, which is not related to the density in an obvious manner. The main problem with the approach is that the kinetic operator is inherently non-local. The approach suggested by Kohn and Sham starts from the observation that a system of non-interacting electrons is exactly described by an antisymmetric wave function of the Slater determinant type, made of one-electron orbitals. As in Hartree-Fock theory, for such a wave function the kinetic energy can be easily obtained in terms of one-electron orbitals. In this case the ground state density matrix $\rho_1(\mathbf{r}, \mathbf{r}')$ is given by:

$$\rho_1(\mathbf{r}, \mathbf{r}') = \sum_{i=1}^{\infty} f_i \langle \phi_i | \nabla^2 | \phi_i \rangle \tag{3.16}$$

Kohn and Sham's idea was that, if one can find a system of non-interacting electrons that produces the same electronic density of the interacting system, then the kinetic energy of the non-interacting system can be calculated exactly via equation (3.16). However, this is not the exact kinetic energy of the interacting system. The missing fraction is due to the fact that the true many-body wave function is not a Slater determinant. There is then a correlation contribution to the kinetic energy that is not taken into account, which must be included in the correlation energy term. Kohn and Sham's approach assumes that the equivalent non-interacting system, i.e. a system of non-interacting electrons whose ground state density coincides with that of the interacting system, does exist. This system will be called the non-interacting reference system of density $n(\mathbf{r})$, and is described by the Hamiltonian:

$$\hat{\mathcal{H}}_{\mathrm{R}} = \sum_{i=1}^{N} \left[-\frac{\hbar^2}{2m_e} \nabla_i^2 + \nu_{\mathrm{R}}(\mathbf{r}_i) \right] \tag{3.17}$$

where N is the number of electrons. Here, the reference potential $\nu_{\mathrm{R}}(\mathbf{r})$ is such that he ground state density of $\hat{\mathcal{H}}_{\mathrm{R}}$ equals $n(\mathbf{r})$ If that is the case, Hohenberg and Kohn's theorem ensures that the ground state energy equals the energy of the interacting system. This Hamiltonian has no electron-electron interactions. Therefore, its eigenstates can be

expressed in the form of Slater determinants. For this discussion the assumptions are that the occupation numbers are 2 for $i \leq N_\mathrm{s}$ and 0 for $i > N_\mathrm{s}$, with $N_\mathrm{s} = N/2$ the number of doubly occupied orbitals. For simplicity, a possible spin dependence is ignored. This would arise, for example, in magnetic or open shell systems. Within these assumptions, the density reads:

$$n(\mathbf{r}) = 2\sum_{i=1}^{N_\mathrm{s}} |\phi_i(\mathbf{r})|^2 \tag{3.18}$$

while the kinetic term is:

$$T_\mathrm{R}[n] = -\frac{\hbar^2}{m_e}\sum_{i=1}^{N_\mathrm{s}} \langle\phi_i|\nabla^2|\phi_i\rangle \tag{3.19}$$

The single-particle orbitals $\phi_i(\mathbf{r})$ are the N_s lowest-energy eigenfunctions of the one-electron Hamiltonian:

$$\hat{H}_\mathrm{KS} = -\frac{\hbar^2}{2m_e}\nabla^2 + \nu_\mathrm{R}(\mathbf{r}) \tag{3.20}$$

which are obtained by solving the one-electron Schrödinger equation:

$$\hat{H}_\mathrm{KS}\phi_i(\mathbf{r}) = \epsilon_i\phi_i(\mathbf{r}) \tag{3.21}$$

The universal density functional can be re-written to include $T_\mathrm{R}[n]$ from equation (3.19):

$$F[n] = T_\mathrm{R}[n] + \frac{1}{2}\int\int \frac{n(\mathbf{r})n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\mathrm{d}\mathbf{r}\mathrm{d}\mathbf{r}' + \tilde{E}_\mathrm{XC}[n] \tag{3.22}$$

which defines a modified exchange and correlation energy $\tilde{E}_\mathrm{XC}$, which accounts also for the kinetic correlation ignored in $T_\mathrm{R}[n]$. By substituting this expression for F into the total energy functional, $E_\nu[n] = F[n] + \int n(\mathbf{r})\nu_\mathrm{ext}(\mathbf{r})\mathrm{d}\mathbf{r}$, then the Kohn-Sham energy functional is obtained:

$$E_\mathrm{KS}[n] = T_\mathrm{R}[n] + \int n(\mathbf{r})\nu_\mathrm{ext}(\mathbf{r})\mathrm{d}\mathbf{r} + \frac{1}{2}\int\int \frac{n(\mathbf{r})n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\mathrm{d}\mathbf{r}\mathrm{d}\mathbf{r}' + \tilde{E}_\mathrm{XC} \tag{3.23}$$

In this way the energy functional is expressed in terms of the N_s orbitals that minimise the non-interacting electronic kinetic energy under the fixed density constraint. The one-electron orbitals are usually called the Kohn-Sham orbitals. The Kohn-Sham orbitals satisfy the one-electron Kohn-Sham equations (3.21), but so far there is no expression for the reference potential ν_R. All that is know is that ν_R is a potential that ensures that the density of the non-interacting reference system is the same as the true density of the interacting system. It should then be possible to determine it by minimising the KS functional (3.23) with respect to the density, under the constraint that this density integrates to N particles. The variational principle is the now applied to the Kohn-Sham functional:

$$\frac{\delta}{\delta n(\mathbf{r})}\left(E_\mathrm{KS}[n] - \mu\int n(\mathbf{r}\mathrm{d}\mathbf{r})\right) = 0 \tag{3.24}$$

obtaining the following equation for the minimising ground state density:

$$\frac{\delta T_\mathrm{R}[n]}{\delta n(\mathbf{r})} + \nu_\mathrm{ext}(\mathbf{r}) + \int \frac{n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\mathrm{d}\mathbf{r}' + \frac{\tilde{E}_\mathrm{XC}[n]}{\delta n(\mathbf{r})} = \mu \tag{3.25}$$

where the functional derivative $\delta T_{\mathrm{R}}[n]/\delta n(\mathbf{r})$ can be readily obtained by considering the non-interacting Hamiltonian $\hat{\mathcal{H}}_{\mathrm{R}}$ of equation (3.17). Since the particles in the reference system only interact with the reference potential, and not between themselves, this Hamiltonian corresponds to the energy functional:

$$E_{\nu_{\mathrm{R}}}[\tilde{n}] = T_{\mathrm{R}}[\tilde{n}] + \int \tilde{n}(\mathbf{r})\nu_{\mathrm{R}}(\mathbf{r})\mathrm{d}\mathbf{r} \tag{3.26}$$

whose ground state energy is the same as that of the interacting system because they share the same electronic density. Therefore, in general $E_{\nu_{\mathrm{R}}}[\tilde{n}] \geq E_0$ and the equality is verified only for the ground state density n. This means that the functional derivative of $E_{\nu_{\mathrm{R}}}[\tilde{n}]$ must vanish for the ground state density. Applying the variational principle to $E_{\nu_{\mathrm{R}}}[\tilde{n}]$, one obtains:

$$\frac{\delta T_{\mathrm{R}}[n]}{\delta n(\mathbf{r})} + \nu_{\mathrm{R}}(\mathbf{r}) = \mu_{\mathrm{R}} \tag{3.27}$$

where μ_{R} is the chemical potential of the non-interacting system, which should coincide with that of the interacting system μ. Otherwise, if the interacting and the equivalent non-interacting reference system were put into contact, there would be charge flow from one to the other. By comparing equations (3.25) and (3.27) and setting $\mu_{\mathrm{R}} = \mu$, one obtains the following expression for the reference potential:

$$\nu_{\mathrm{R}}(\mathbf{r}) = \nu_{\mathrm{ext}}(\mathbf{r}) + \int \frac{n(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|}\mathrm{d}\mathbf{r}' + \frac{\delta \tilde{E}_{\mathrm{XC}}[n]}{\delta n(\mathbf{r})} \tag{3.28}$$

The reference potential depends on the solutions of the one-electron Schrödinger equation (the Kohn-Sham orbitals) through the electronic density and so the equation must be solved self-consistently, making sure that the density used to construct the reference potential coincides (within some tolerance) with that obtained from the solutions of the equation via (3.18).

3.2.1 Exchange and correlation

The strategy to attack the many-body electronic problem presented in the previous section consisted of dividing the total energy of an electronic system into a number of different contributions. The classical electron-electron interaction, or Hartree term, and the interaction of the electrons with external fields, in particular that of the atomic nuclei, are known as explicit functionals of the electronic density. It can be seen that their only dependence on the electronic variables is through the electronic density. In that sense it is said that they are functionals of the density. The non-interacting kinetic energy and the exchange energy are known as functionals of the non-interacting orbitals, which are in turn (unknown) functionals of the density. The correlation energy is a big unknown. The exchange energy, although well known as a function of the single-particle orbitals, involves the calculation of computationally expensive integrals. In addition, up to date there is no approximation available where the correlation energy is treated at a comparable level of

accuracy. Therefore, if exchange is treated exactly as a functional of the orbitals, it will not be able to compensate for any errors introduced when approximating the correlation term.

Electrons will repel one another according to Coulomb's law with a repulsion energy of $\frac{1}{|\mathbf{r}-\mathbf{r}'|}$. Therefore electrons will move in order to avoid one another. In other words, their motion will be correlated. For a given basis set, to be discussed in the next section, the correlation energy is equal to the difference between the exact energy and the energy calculated using the Hartree-Fock approach. To illustrate the importance of this correlation energy, for the helium atom, the difference in energy between treating the interactions between the two electrons in an average, as opposed to an instantaneous, manner is on the order of 1 eV, or 23 kcal mol^{-1}, and in general this can be thought of as the error, per electron pair, in the Hartree-Fock approach [59]. In addition, it was in 1925 when Wolfgang Pauli formulated the quantum mechanical principle, that now bears his name, stating that no two identical fermions may simultaneously occupy the same quantum state [60]. The result of this is that the wave function of indistinguishable fermions must be antisymmetric, or change its sign, upon the exchange of two identical fermions. The resultant exchange interaction alters the expectation value of the energy, upon the overlap of wave functions of two or more electrons, as it increases the expectation value of the separation between the particles. The effects of this exchange interaction were discovered by Werner Heisenberg [61] and Paul Dirac [62] in 1926; it has no classical analogue.

It seems sensible to treat both the exchange and correlation terms to a similar level of approximation. The idea now is to look for consistent approximations to exchange and correlation where both terms are treated in a similar manner. One of the natural starting points is the homogeneous electron gas, which is a simplified model for metallic systems. This is the simplest system of correlated electrons, and as such has been studied in great detail. Using the homogeneous electron gas as a reference may not seem a particularly good idea for molecular systems, as their electronic densities are far from uniform. Perhaps this was the reason why DFT took so long to be adopted by the computational chemistry community, because most of the available approximations are derived from the homogeneous electron gas. The most widely used approaches to the exchange-correlation problem within DFT are the local density and generalised gradient approximations (LDA and GGA, respectively). The LDA has been for a long time the most widely used approximation to the exchange-correlation energy. It was proposed in the seminal paper by Kohn and Sham [58]. The main idea is to consider a general inhomogeneous electronic system as locally homogeneous, and then to use the exchange-correlation hole corresponding to the homogeneous electron gas, which is known to an excellent accuracy. In practice, energy terms local in the density are calculated by integrating over the volume of the system, with the corresponding energy density calculated at the values that the electronic density assumes at every point $\mathbf{r}$ in the volume.

One defines the exchange-correlation energy as:

$$E_{\mathrm{XC}} = \frac{1}{2}\int\int \frac{n(\mathbf{r})n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}[g(\mathbf{r},\mathbf{r}')-1]\mathrm{d}\mathbf{r}\mathrm{d}\mathbf{r}' \tag{3.29}$$

where the electron-electron pair distribution, or pair correlation function, represents the probability of finding an electron at $\mathbf{r}$ given that there is another electron at $\mathbf{r}_0$. The presence of this electron discourages other electrons from approaching it because of the Coulomb repulsion. Therefore, the pair distribution function interpolates from zero at $\mathbf{r} = \mathbf{r}_0$ to one at infinite distance. The original definition (3.29) of the exchange-correlation energy, which does not contain kinetic contributions, can be used only if the exact expression for the kinetic energy is known. However, within DFT this is not the case. In Kohn-Sham theory the non-interacting expression for the kinetic energy is used, and then the exchange-correlation term is redefined as:

$$\tilde{E}_{\mathrm{XC}}[n] = E_{\mathrm{XC}}[n] + T[n] - T_R[n] \tag{3.30}$$

which defines a modified exchange and correlation energy $\tilde{E}_{\mathrm{XC}}$, different from the E_{XC} given by (3.29) in that it accounts also for the kinetic correlation ignored in T_R. The kinetic contribution to the exchange term is given by Pauli's principle, and this is already contained in $T_R[n]$ and in the density when adding up the contributions of the N_s, or N, lowest eigenstates according to (3.18) and (3.19). Therefore, the exchange term is not modified by the introduction of the non-interacting reference system.

One can interpret the exchange-correlation energy $\tilde{E}_{\mathrm{XC}}[n]$ as the Coulomb interaction between the electronic density and some displaced charge density. This can be done by defining the exchange-correlation hole in the following way:

$$\tilde{n}_{\mathrm{XC}}(\mathbf{r},\mathbf{r}') = n(\mathbf{r}')[\tilde{g}(\mathbf{r},\mathbf{r}')-1] \tag{3.31}$$

so that the exchange-correlation energy is written:

$$\tilde{E}_{\mathrm{XC}}[n] = \frac{1}{2}\int\int \frac{n(\mathbf{r})\tilde{n}_{\mathrm{XC}}(\mathbf{r},\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\mathrm{d}\mathbf{r}\mathrm{d}\mathbf{r}' \tag{3.32}$$

where $\tilde{g}(\mathbf{r},\mathbf{r}')$ is obtained by averaging the pair correlation function $g(\mathbf{r},\mathbf{r}')$ over the strength of the electron-electron interaction, of which details can be found elsewhere [63]. The XC hole $\tilde{n}_{\mathrm{XC}}$ represents a fictitious charge depletion due to exchange and correlation effects, i.e. due to the fact that the presence of an electron at $\mathbf{r}$ reduces the probability of finding a second electron at $\mathbf{r}'$ in the vicinity of $\mathbf{r}$. It corrects for the fact that the Hartree contribution to the energy completely ignores this depletion. If one separates the exchange and correlation contributions it is easy to see that the displaced electron arises exclusively from the exchange part. This is a consequence of how the electron-electron interaction has been separated.

In the LDA, one re-writes the expression for the (non-local) exchange-correlation hole in the following way [58]:

$$\tilde{n}_{\mathrm{XC}}^{\mathrm{LDA}}(\mathbf{r},\mathbf{r}') = n(\mathbf{r})\left\{\tilde{g}^{\mathrm{HEG}}\left[|\mathbf{r}-\mathbf{r}'|, n(\mathbf{r})\right]-1\right\} \tag{3.33}$$

where $\tilde{g}^{\mathrm{HEG}}$ is the pair correlation function for the homogeneous electron gas. This pair correlation function depends only on the distance between $\mathbf{r}$ and $\mathbf{r}'$ (the system is homogeneous), and must be evaluated for the density n that locally assumes the value $n(\mathbf{r})$. With this definition the exchange-correlation energy can be written as the average of an energy density $\epsilon_{\mathrm{XC}}^{\mathrm{LDA}}[n]$:

$$\tilde{E}_{\mathrm{XC}}^{\mathrm{LDA}}[n] = \int n(\mathbf{r})\tilde{\epsilon}_{\mathrm{XC}}^{\mathrm{LDA}}[n(\mathbf{r})]\mathrm{d}\mathbf{r} \tag{3.34}$$

weighted with the space-dependent electronic density of the system. The expression for the exchange-correlation energy density in terms of the exchange-correlation hole is:

$$\tilde{\epsilon}_{\mathrm{XC}}^{\mathrm{LDA}}[n] = \frac{1}{2}\int \frac{\tilde{n}_{\mathrm{XC}}^{\mathrm{LDA}}(\mathbf{r},\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\mathrm{d}\mathbf{r}' \tag{3.35}$$

While the exchange-correlation energy $E_{\mathrm{XC}}[n]$ should be a functional of n, there is no reason why the energy density should also be so. In fact, in general ϵ_{XC} is not a functional of the density. From its very definition it is clear that it has to be a non-local object, because it reflects the fact that the probability of finding an electron at $\mathbf{r}$ depends on the presence of other electrons in the surroundings, through the exchange-correlation hole. However, in the LDA it becomes a functional of the density because it corresponds to a homogeneous system where ρ is the same everywhere. In 1989, Jones and Gunnarsson discussed thoroughly the LDA, by analysing the performance for different types of systems, in particular atomic and molecular, but also solids. Many of the successes of the approximation can be traced back to two fundamental properties of the LDA exchange-correlation hole:

- It satisfies the sum rule expressing that the exchange-correlation hole contains exactly one displaced electron. This is because for each $\mathbf{r}$, $\tilde{g}^{\mathrm{HEG}}$ is the pair correlation function of an existing system, i.e. the homogeneous gas at density $n(\mathbf{r})$.
- Even if the exact $\tilde{n}_{\mathrm{XC}}$ is not spherically symmetrical, what really matters for the exchange-correlation energy is the spherical average of the hole. This spherical average is reproduced to a good extent by the LDA, whose $\tilde{n}_{\mathrm{XC}}$ is already spherical.

One of the most significant components missing from the LDA is a description of the variation of the electron density from place to place and this approximation can be improved upon by including the gradient of the electron density. Increased numerical accuracy has been demonstrated in the literature for many systems using the so-called generalised gradient approximation (GGA) which has resulted in an entire family of functionals based on this approach. GGA functionals contain an additional term that includes the gradient of the electron density:

$$\tilde{E}_{\mathrm{XC}}^{\mathrm{GGA}}[n(\mathbf{r})] = \int n(\mathbf{r})\tilde{\epsilon}_{\mathrm{XC}}^{\mathrm{GGA}}\left(n(\mathbf{r}), \nabla n(\mathbf{r})\right)\mathrm{d}\mathbf{r} \tag{3.36}$$

with the corresponding exchange-correlation potential equal to:

$$\nu_{\mathrm{XC}}^{\mathrm{GGA}} = \tilde{\epsilon}_{\mathrm{XC}}^{\mathrm{GGA}}\left(n(\mathbf{r})\right) + n(\mathbf{r})\frac{\partial\tilde{\epsilon}_{\mathrm{XC}}^{\mathrm{GGA}}(n\left(\mathbf{r}\right))}{\partial n(\mathbf{r})} + n(\mathbf{r})\frac{\partial\tilde{\epsilon}_{\mathrm{XC}}^{\mathrm{GGA}}(n\left(\mathbf{r}\right))}{\partial\nabla n(\mathbf{r})} \tag{3.37}$$

where the terms with partial derivatives in equation (3.37) are due to the change in the exchange-correlation hole with density. This derivative also appears in the equivalent LDA potential. For an insulator, this derivative is discontinuous across a band gap, due to the fact that the nature of the states change discontinuously as a function of the density. The result is a 'derivative discontinuity' where the Kohn-Sham potential for every electron in a material changes by some constant amount, following the addition of a single electron [64,65]. However, the largest error of this approximation actually arises from the gradient contribution to the correlation term. Provided that the problem of the correlation term can be cured in some way, such as the real space cutoff method of David C. Langreth and M. J. Mehl [66], the biggest problem remains with the exchange energy. One of the main lessons learnt from the early works of Gross and Dreizler [67] and Perdew [68] is that the gradient expansion has to be carried out very carefully in order to retain all the relevant contributions to the desired order. Another important lesson is that these expansions easily violate one or more of the exact conditions required for the exchange and correlation holes, such as the normalisation condition, the negativity of the exchange density, or the self-interaction cancelation. Perdew showed that imposing these conditions to functionals that originally do not verify them results in a remarkable improvement of the quality of exchange energies [68]. On the basis of this type of reasoning, a number of modified gradient expansions have been proposed along the years, mainly between 1986 and 1996. These have been named generalised gradient approximations (GGAs). Normally GGAs improve over some of the drawbacks of the LDA, although this is not always the case. A thorough comparison of different GGAs has been done by Filippi et al. [69].

The specific GGA functional formulated by Perdew, Burke and Ernzerhof (PBE) [70] has proved remarkably successful in many DFT applications. This particular form of the GGA retains the correct features of the LDA and combines them with the inhomogeneity features that are assumed to be energetically the most important ones. It sacrifices a few correct, but less important, features, like the second-order gradient coefficients in the slowly varying limit and the non-uniform scaling of the exchange energy in the rapidly varying density region. The PBE functional is very satisfactory from a theoretical point of view, because it verifies many of the exact conditions for the XC hole, and it does not contain any fitting parameters. It is the PBE functional that is used throughout the DFT calculations presented in this dissertation.

One of the most significant problems with DFT as a theory is that each electron moves in the potential of the total electron density which includes the electron. This self-interaction error is clearly incorrect as a single, isolated electron will not be repelled by itself. Within Hartree-Fock theory, the diagonal Coulomb integrals, corresponding to the self-interaction, are exactly canceled by the corresponding diagonal exchange integrals. Despite being a seemingly simple concept, within DFT a proper mathematical formulation of this problem still remains a challenge. The self-interaction error is the origin of many

qualitative and quantitative failings within DFT and many authors are providing continual efforts toward its removal [71]. One approach to remove this error from DFT is to include the Hartree-Fock expression for the exchange interaction. Formally, this approach scales as the fourth power of the number of orbitals, but a current area of research is to make this cost linear for localised orbitals [72].

The problems discussed for the LDA result in energy differences which are significantly larger than 'chemical accuracy', generally defined as an accuracy of 1 kcal mol^{-1}. The calculated difference between the highest occupied and lowest unoccupied molecular orbitals (HOMO-LUMO gap) is often very much underestimated in the LDA. GGA functionals still suffer from the self-interaction error discussed and still underestimate HOMO-LUMO gaps. In general, DFT approximates both the exchange and correlation energies whilst Hartree-Fock ignores correlation but calculates the exchange exactly. Therefore, adding a fraction of the Hartree-Fock exchange energy to the DFT energy, resulting in so-called hybrid functionals, can significantly improve upon the calculated HOMO-LUMO gap compared to the GGA. However, it is not always clear what fraction of the exact exchange energy should be added to the functional. It has also been shown that whilst hybrid functionals may give an initially larger gap for the same system compared to the gap calculated using a GGA functional, upon increasing the system size this value is shown to decrease. Therefore there are clearly other effects reducing the gap value and these are discussed in Chapter 5 of this dissertation. In addition to the underestimation of HOMO-LUMO gaps, another significant problem with regards to the biomolecular systems described in this dissertation, caused by the self-interaction error, is the generally poor description of transition states. For the general dissociation of any molecule into two fragments, such as the transition state calculated in the next chapter, it has been shown that when one of the fragments has an electron affinity similar to the ionisation energy of the other fragment, the self-interaction error will cause each fragment to have a fractional charge at large separation and, as a result, the total energy is too low [73]. However, despite the faults which have been outlined here, there is ongoing work to improve upon existing exchange-correlation functionals. The methods discussed in this dissertation will also be applicable when more accurate linear-scaling exchange-correlation functionals are available that include a greater proportion of Hartree-Fock exchange.

3.2.2 Basis sets

At the very start of this chapter, Dirac notation was used to express the Schrödinger equation. This is a most elegant approach to working with quantum mechanics without needing to specify a representation. However, when wanting to perform electronic structure calculations, a representation for the operators and wave functions must be chosen. This representation can be fixed by specifying the basis set, which is generally defined as a collection of vectors that spans a space in which a problem is solved. In the same way that $\hat{i}$, $\hat{j}$ and $\hat{k}$ defines a cartesian, three-dimensional linear vector space, within computational

software packages the basis set will often refer to the set of non-orthogonal one-particle functions used to build molecular orbitals. In general, a wave function ψ can be written as:

$$\psi = \sum_i c_i \phi_i \tag{3.38}$$

where, in this deliberately general example, ϕ_i could perhaps represent a set of atomic-like orbitals and c_i would be their associated coefficients. The basis sets used in formal quantum mechanics will be complete in that they perfectly represent any wave function in the space that they span. However, the basis sets chosen to perform electronic structure calculations must be truncated for practical computation and thus a wide variety of approximations are used. Ideally basis sets will use a minimal number of functions, demonstrate systematic convergence upon making the basis set more complete and not impose any assumed property of the system onto the calculations. As the purpose of most computational investigations is to find the equilibrium ionic configuration for a set of atoms, it is also imperative that the basis set must allow accurate calculation of forces. Most often, molecular orbitals are built from a linear combination of atomic orbitals where an orbital is defined as a one-electron function. In the majority of total energy packages available, these atomic orbitals are represented by atom-centred Gaussian type orbitals (GTOs) in the form of:

$$\phi^{\mathrm{GTO}}_{\mathrm{abc}}(x, y, z) = N x^a y^b z^c e^{-\zeta r^2} \tag{3.39}$$

where N is a normalisation constant, a,b and c control the angular momentum L = a + b + c and ζ controls the width of the orbital whereby a high ζ is associated with a tight function and a low ζ gives a diffuse function. Due to the Gaussian product theorem which makes GTOs relatively easy to compute, they are widely used amongst computational chemists. A weakness of GTOs is that they produce less accurate results than Slater type orbitals (STOs) [74], however it takes longer to compute integrals using STOs. Through a linear combination of GTOs, one can approximate an STO, often called an "STO-nG" basis, despite it being a combination of contracted GTOs. An example of one of the simplest, minimal basis sets is STO-3G where each STO is represented by three GTOs. An asterisk after the G would indicate that polarisation functions have been added. A second asterisk will be added if polarisation functions have been applied to the hydrogen atoms. The notation, introduced by Pople, of 3-21G tells us that three GTOs are used for the core, two for the first valence orbital and one for the second valence orbital. This is a double-valence or double-zeta basis set. Triple- and quadruple-zeta basis sets have three and four basis functions for each atomic orbital, respectively. This number can be increased as it has been shown that having different sized functions allows the orbital to adapt according to proximity to other atoms. A disadvantage to atom-centred orbitals is that as the basis set moves with the atoms, so the wave functions will change as the atoms move. This gives rise to Pulay forces [75] which must be calculated as corrections to the Hellmann-Feynman forces, which will be described in Section 3.5.1. Other basis sets

which are also used widely by computational chemists include the correlation-consistent basis sets which can be converged systematically to the complete basis set limit [76] where in this limit the energy can be extrapolated to, in principle, yield an exact solution. The notation for these types of basis sets includes cc-pVTZ to denote a triple-zeta valence set with polarisation functions and 'aug-' will be prepended to indicate the use of additional diffuse functions.

A choice of basis set used very widely amongst condensed matter physicists is plane waves, which are the solution of the Schrödinger equation for free electrons and have the general form of $e^{i\mathbf{G}.\mathbf{r}}$ where $\mathbf{G}$ is some wavevector. By increasing the maximum value of the wavevectors of the plane waves retained in the basis set, the basis set can systematically be made more complete. The basis set is usually defined by an energy cutoff where the kinetic energy of an electron with associated wavevector $\mathbf{G}$ is:

$$E_{\text{cutoff}} = \frac{\hbar^2 \mathbf{G}^2}{2m_e} \tag{3.40}$$

However, large areas of vacuum in a system are computationally expensive when using a plane-wave basis set as plane waves fill all space. Despite this, a significant advantage of the approach is that the basis does not change when the atoms move so there are no associated Pulay forces.

3.2.3 The pseudopotential approximation

The atomic wave functions are eigenstates of the atomic Hamiltonian, therefore they must all be mutually orthogonal. Since the core states are localised in the vicinity of the nucleus, the valence states must oscillate rapidly in this core region in order to maintain this orthogonality with the core electrons. This rapid oscillation results in a large kinetic energy for the valence electrons in the core region, which roughly cancels the large potential energy due to the strong Coulomb potential [77]. Thus the valence electrons are much more weakly bound than the core electrons. In 1934, Hans G. A. Hellmann therefore replaced these effects by a *Zusatzpotential* [78], which is repulsive in the core region and therefore keeps the electrons out of the core (Pauli repulsion). The potential that originates in the atomic nuclei is far from smooth. In the simplest case of hydrogen the potential is $-1/r$, which diverges at the origin. The $1s$ wave function does not diverge, but it exhibits a cusp at the origin, and decays exponentially with distance. For heavier atoms the wave functions associated with core states are even steeper. Therefore, a plane-wave expansion of the wave functions in a real system is a difficult task, because the number of plane-wave components required to represent such steep wave functions is huge. However, it would be desirable to retain the simplicity of the plane-wave approach. In 1937, Slater suggested another possible solution to the problem, where the plane-wave expansion was augmented with the solutions of the atomic problem in spherical regions around the atoms, and the potential was assumed to be spherically symmetric inside the

spheres, and zero outside, in the augmented plane wave (APW) method [79]. In order to overcome this shape approximation of the potential, in 1940, Conyers Herring proposed an alternative method consisting of constructing the valence wave functions as a linear combination of plane wave and core wave functions [80]. By choosing appropriately the coefficients of the expansion, this wave function turns out to be orthogonal to the core states, hence the name of orthogonalised plane wave (OPW) method. Since the troublesome region is taken care of by the core orbitals, the part that must be represented by the plane waves is rather smooth, and a smaller number of plane-wave components is required to reproduce the valence states. One can go a step beyond the OPW approach, and eliminate the core states altogether by replacing their action with an effective potential, or pseudopotential. This pseudopotential, however, cannot just be anything. It has to be constructed carefully in order to reproduce accurately the bonding properties of the true potential.

Electrons are indistinguishable and a separation into pure core, pure valence and mixed core/valence terms for the electronic Hamiltonian is therefore not possible. Nevertheless, it is convenient to classify the electronic states of an atom into: (i) core states, which are highly localised and not involved in chemical bonding, (ii) valence states, which are extended and responsible for chemical bonding, and (iii) semi-core states, which are localised and polarisable, but generally do not contribute directly to chemical bonding. The most common pseudopotential approach consists of not allowing the relaxation of core states according to the environment (frozen core approximation), although some polarisable core approaches have been proposed. In general, this is a very good approximation that reproduces total atomic energies within 0.01 eV [81]. Semi-core states are often treated as part of the frozen core, but when their contribution is important they have to be included in the valence. The valence states, due to orthogonalisation with respect to the core states of the same symmetry, show a marked oscillatory behaviour with a number of nodes equal to $n - l - 1$, n being the principal quantum number and l the angular momentum. Nodeless wave functions ($l = n - 1$) are not oscillatory but, due to the lack of orthogonalisation, they create strongly bound states that are markedly peaked close to the nucleus. This is the case of the $1s$ state in H, the $2p$ states in C, N, O, and F and the $3d$ states in transition metals. When the basis set chosen is that of plane waves, the computation of Hamiltonian matrix elements requires the Fourier decomposition of the wave functions. Features like the above are very stringent for plane waves, because sharp peaks require a very large number of plane waves to achieve convergence in the expansion, and this translates into a vast amount of computational resources (the dimension of the matrix to diagonalise becomes very large). Based on the observations that: (i) core states are not fundamental for the description of chemical bonding, and (ii) a good description of the valence wave functions inside the core region is not strictly necessary, there is no lack of crucial information if the inner solution (inside some cutoff radius) is replaced with a smooth, nodeless pseudo-wave function, which is not a solution to the original atomic

problem. Being nodeless, it now corresponds to the lowest-lying state of an effective, pseudo-atomic problem where the true potential has been replaced by a pseudopotential.

There are two essential steps in pseudopotential theory. The first is that the core electrons are removed from the calculation, and the interaction of the valence electrons with the nucleus plus the core states (including orthogonalisation) is replaced by an effective, screened potential. The screened potential depends on the angular momentum of the valence electrons because of the different orthogonality conditions. For instance, in the C atom, the $2s$ valence state has to be orthogonal to the $1s$ core state, but the $2p$ valence state does not feel the orthogonality constraint of the $1s$ state because they have different angular quantum numbers. Therefore, within the core region, these two states feel very different potentials from the ionic core. At large distances the potential is $-Z_V/r$ independently of the angular momentum, because the ionic core is seen as a point charge of magnitude equal to the valence charge Z_V. For each angular momentum l the pseudopotential should have the valence l-state as the ground state. The second step is that the full ionic core-electron interaction (often called ion-electron interaction), which includes the orthogonality of the valence wave functions to the core states, is replaced by a softer pseudopotential. The solution of the atomic Schrödinger equation for the pseudopotential is a pseudo-wave function different from the true wave function. The pseudopotential, however, is constructed in such a way that its scattering properties and phase shifts are the same as those of the all-electron potential, although the radial pseudo-wave function has no nodes inside the core region.

In 1959, James Charles Philips and Leonard Kleinman [82] showed that one can construct a smooth valence wave function that is not orthogonalised to the core states, by combining the core and the true valence wave functions into a pseudo-wave function that satisfies a modified Schrödinger equation. They then showed that it is possible to construct a pseudo-Hamiltonian with the same eigenvalues of the original Hamiltonian but a smoother, nodeless wave function. The associated potential was called a pseudopotential. This pseudopotential acts differently on wave functions of different angular momentum. When the total pseudopotential acts on the electronic wave function, projection operators select the different angular components of the wave function, which are then multiplied by the pseudopotential corresponding to the angular component. Next, the contributions of all the angular components are added up to form the total pseudopotential contribution to the Hamiltonian matrix elements that enter the Schrödinger equation. Pseudopotentials of this kind are usually called non-local because they act differently on the various angular components of the wave function as a consequence of the exchange with the core. However, as the pseudopotential corresponding to the angular components is a local operator in the radial coordinate, a better name for this type of expression is semi-local or angular-dependent. If all the angular components of the pseudopotential are taken to be the same, then the pseudopotential is said to be local. In principle, local versions can be constructed that verify the required properties for all angular momenta, but they

tend to be quite hard (many plane-wave components are required), and are difficult to construct. That is why it is easier and computationally more effective to use non-local pseudopotentials.

There is an enormous freedom in how pseudopotentials are constructed, the details of which extend beyond the scope of this dissertation. The problem with empirical potentials which were determined primarily through fitting experimental energy bands [83], was that they lacked transferability, namely that a pseudopotential constructed for some specific environment can be used for the same atomic species but in a different environment. The first non-empirical approach to pseudopotentials was the one devised by Philips and Kleinman. This approach, however, had a severe problem: the normalised pseudo-wave function had an amplitude different from that of the all-electron wave function. Outside the core the shapes were the same but the wave functions were only proportional to each other through a normalisation factor. This was not acceptable because it led to an incorrect valence charge distribution, and thus to deviations in the bonding properties. It is important that outside the core region the true and pseudo-wave functions are the same. The construction of a pseudopotential is an inverse problem: given a pseudo-wave function that: (i) beyond some distance decays exactly as the all-electron wave function, and (ii) is an eigenstate of a pseudo-Hamiltonian with the same eigenvalue as the all-electron wave function, the pseudopotential is obtained by inverting the radial Schrödinger equation for that pseudo-wave function. Its solution is uniquely determined by the value of the wave function and its derivative at any given point. These two conditions can be equally realised by specifying the value of the (dimensionless) radial logarithmic derivative of the wave function, together with a normalisation condition, and this can be done for all values of angular momentum l. This involves the phase shifts of the partial waves from scattering theory. Therefore, if the all-electron potential and the pseudopotential are the same outside some cutoff radius r_c, then the all-electron and pseudo-wave functions are proportional if the corresponding logarithmic derivatives are the same. When the pseudo-wave function is further required to preserve the norm inside the cutoff radius this property is called norm-conservation, and it was first introduced in 1979 by Hamann, Schlüter, and Chiang (HSC) [84].

A key result from HSC was to realise that the norm of the wave function also appears in a very important identity related to the Friedel sum rule. The norm-conservation constraint is tightly linked to the concept of transferability through this sum rule and the expression shows that the first order energy variation of the phase shift is proportional to the norm of the wave function in the pseudised region. Therefore, the norm-conservation condition, imposes that, to first-order in the eigenvalue, the logarithmic derivatives of the all-electron and pseudo-wave functions vary in the same way. This implies that a small change in the eigenvalue due to changes in the external potential (the environment) produces only a second-order change in the logarithmic derivative. Therefore, the condition of matching logarithmic derivatives, which by construction is strictly verified only for the

value of the reference energy used to obtain the wave function, becomes approximately valid in a range of eigenvalues around the reference. In this way, pseudopotentials derived from atomic calculations can be exported to other environments. When an atom is part of a molecule or a solid, its electrons feel the influence of the other atoms (the so-called molecular or crystal field). This implies that the electronic eigenvalues are shifted from their atomic values, but the transferability property ensures that the all-electron and pseudo-wave functions still coincide outside the cutoff radius. The norm-conservation constraint guarantees that the pseudopotential is useful, not in every energy range, but at least in environments such that the eigenvalues do not depart significantly from the eigenvalues used in its construction. For example, a pseudopotential for H in the H_2 molecule may not be useful for hydrogen at very high pressures because the energy ranges are completely different, but a pseudopotential for Si constructed with the bulk solid in mind will be useful for the Si surface or for liquid Si under similar external conditions. The straightforward recipe for improving transferability is to reduce the cutoff radius, because in this way the pseudo-wave function becomes closer to the all-electron result. However, the reduction of r_c is limited by the (not strictly necessary) condition of a nodeless pseudo-wave function; the cutoff radius cannot be made smaller than the position of the outermost node of the all-electron wave function. The conditions proposed by HSC for the construction of norm-conserving pseudopotentials are that (i) the eigenvalues of the pseudo-wave functions coincide with those of the all-electron wave functions for a chosen electronic configuration of the atom; (ii) The pseudo-wave function is nodeless, and it is identical to the all-electron wave function outside a suitably chosen cutoff radius r_c; (iii) the norm of the true and pseudo-wave functions inside the pseudized region ($r < r_c$) is the same (the norm-conservation condition); (iv) the logarithmic derivatives of the all-electron and pseudo-wave function agree for $r \geq r_c$.

To illustrate in an example, Figure 3.1 shows a norm-conserving pseudopotential for oxygen, within the PBE approximation to DFT. The pseudopotential has been generated for the neutral configuration $[1s^2]2s^22p^4$, where the $1s$ orbital is a core state and $2s$ and $2p$ are in the valence. Figure 3.1(a) shows the actual $2s$ and $2p$ components of an oxygen pseudopotential, together with the unscreened Coulomb potential $-6/r$. Notice how the pseudopotentials approach the Coulomb potential and merge with it at the cutoff radii. Figure 3.1(b) shows the all-electron and pseudo-wave functions for the two pseudized states, $2s$ and $2p$. The cutoff radii are 0.84 Åand 0.79 Å, respectively. The pseudo-energies are virtually the same. The total energies are different because the pseudo-atom does not contain the $1s$ electrons explicitly. Notice how little pseudisation can do for the $2p$ state, which is already nodeless at the all-electron level. However, the effect is more important for the $2s$ state, where pseudization has eliminated the node, thus making the pseudo-wave function much smoother. Throughout the calculations within this dissertation, norm-conserving pseudopotentials have been used. Despite the robust methods outlined in this section, practical limitations from computational resources dictate that even the largest

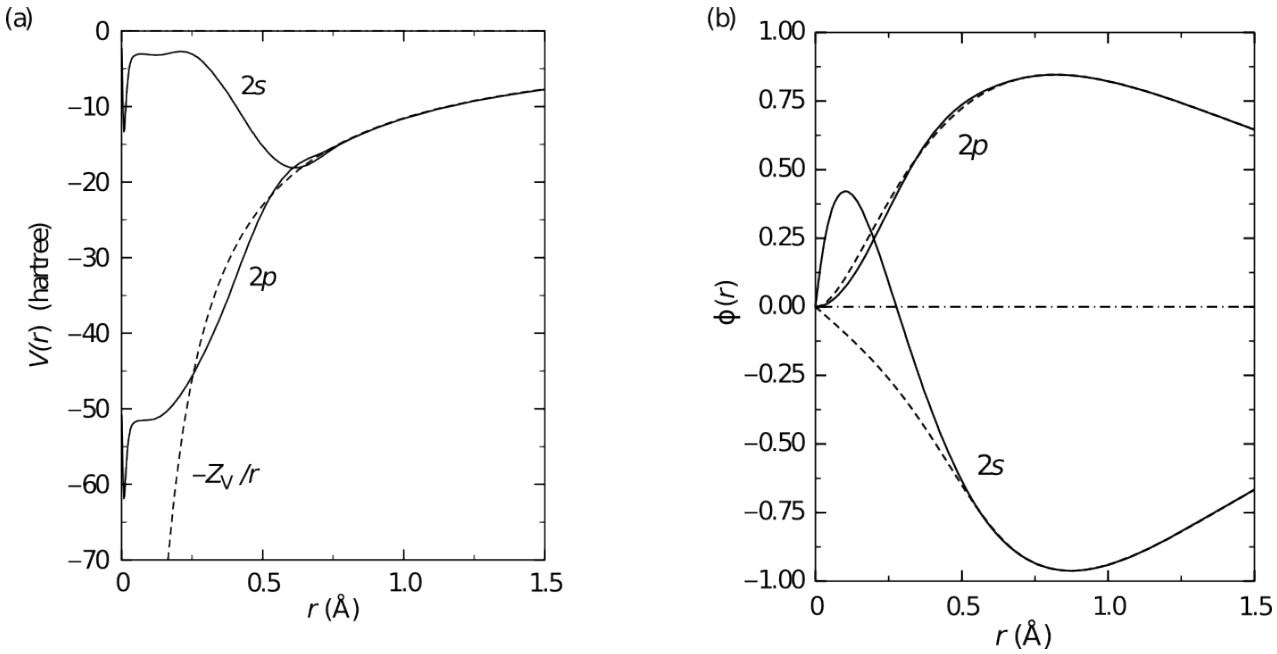


Figure 3.1: (a) Pseudopotentials for the $2s$ and $2p$ states of O (solid lines), and the unscreened $-Z_V/r$ Coulomb potential ($Z_V = 6$). (b) All-electron (solid lines) and pseudo-wave functions (dashed lines) for the O $2s$ (negative values) and $2p$ (positive values) valence states. Figure adapted from Ref. [81].

supercomputers in the world will only allow a system comprising around 500 atoms to be treated with conventional DFT. This is of little use to scientists wishing to explore the realms of biologically relevant systems as a typical biomolecule will contain many more than 500 atoms. One approach toward solving this problem is the focus of the next section.

3.3 Linear-scaling DFT

In recent years, DFT simulations have become increasingly widespread in the simulation of biological systems at the level of individual atoms and electrons. A limiting factor in applying conventional first-principles approaches to large systems is the unfavourable computational requirements which typically increase as the third (or greater) power of the number of atoms in the system. However, methods to overcome this computational bottleneck, resulting in approaches where the computational effort increases linearly with the number of atoms, have been under development for two decades [11] and are still advancing today [12]. These methods provide explicit treatment of the electrons, naturally taking into account the electronic charge transfer and polarisation, for systems containing many thousands of atoms and are transferable to any chemical environment. These advances in linear-scaling first-principles techniques now allow system sizes on the order of tens of thousands of atoms to be routinely accessed [85]. Following this phenomenon

of larger systems becoming accessible, a natural progression has ultimately been to apply density-functional methods to systems of biological interest. The Fritz Haber Institute *ab initio* molecular simulations (FHI-aims) package [86] implements an all-electron/full-potential treatment with computational expense that scales linearly with the size of the system. FHI-aims has, amongst other things, been applied to folding processes in helices and polypeptides with a recently developed DFT+VdW approach [87]. The TeraChem package [88] has recently been used to optimise the structures of more than 50 polypeptides of sizes ranging up to 590 atoms. Large-scale real-space DFT calculations on the electron states of silicon nanowires have been reported [89]. System sizes of 107,292 atoms were treated during the evaluation phase of the K computer using 442,368 cores and the RSDFT code [90]. A report in 2000 demonstrated linear-scaling DFT calculations on a dry DNA model comprising 715 atoms using the SIESTA code [91]. More recent work used the CONQUEST code to perform calculations on a B-DNA decamer system with explicit water molecules and counter ions resulting in a total system size of 3439 atoms [92]. The same code has also been used to calculate total energies and forces of a hydrated ten-mer of DNA using DZP basis sets and comparison to results from the AMBER force field were made [93]. Linear-scaling methods have also been applied to solvent/solute interaction energy studies of drug molecules [94]. Work implementing low-order-scaling approaches applied to DNA has also been reported [95]. The frozen molecular orbital (FMO) approach has been shown to be efficient for biomolecular systems with many published results [96]. One of the largest known systems investigated came from FMO studies of the active sites of influenza A viral haemaglutinin that also used a polarisable continuum model and was applied to a 24,000 atom protein [97]. The ONETEP code [98], which has been used for the majority of the calculations presented in this dissertation, has previously been used by others to perform geometry optimisations characterising binding energetics of small molecules to the metalloprotein myoglobin [99] and also to measure binding of small molecules to T4 lysozyme in solution [100]. There are now a variety of other codes with linear-scaling capabilities [101–107]. Recent developments in the simulation of optical spectroscopy [108], dynamical mean field theory with applications to human respiration [109] and methods to aid interpretation of the electronic structure [110, 111] also broaden the scope of biomolecular simulations. This section outlines the ideas behind density-functional theory with computational time scaling linearly as the number of atoms in the system is increased. A more detailed discussion of the current state of linear-scaling methods is available elsewhere [12].

In order to see the origin of the cubic scaling bottleneck hindering conventional density functional approaches, one must consider equations (3.21) and (**??**). The solutions to (3.21) extend over the entire system (see Figure 3.2) such that the overlap integral in (**??**) requires a computational effort that scales linearly with the number of atoms in the system. However, the number of orbital pairs and the associated number of constraints from (**??**) is proportional to the square of the number of atoms in the system. Therefore

the overall computational effort scales as N^3. The Hohenberg-Kohn theorems, discussed

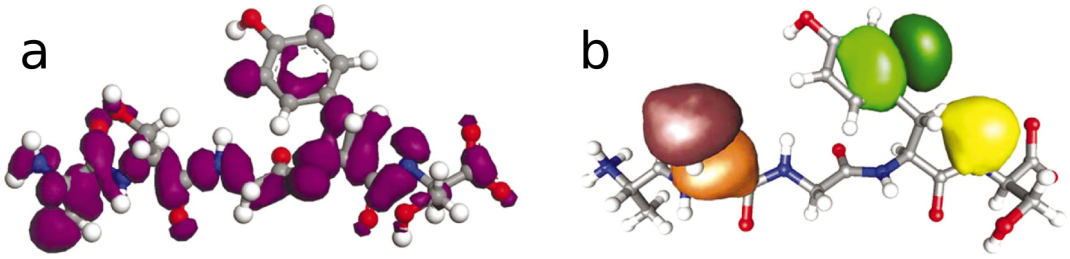


Figure 3.2: (a) One delocalized orbital $\psi_i(r)$ from a conventional DFT calculation with the CASTEP code on a peptide [112]. (b) Three optimised NGWFs $\phi_\alpha(r)$, $\phi_\beta(r)$ and $\phi_\gamma(r)$ from a ONETEP calculation on the same peptide [113].

in Section 3.2, rely on accurate calculation of the electron density in order to provide important information regarding a system. However, if one is interested in generating a linear scaling method, it can be more helpful to not work in terms of the electron density. There are a handful of ways one can go about this, including the Fermi operator expansion [114], divide-and-conquer [115] and orbital minimisation [116] methods. The particular approach that this dissertation will focus on, as is implemented within the ONETEP code that is discussed later, instead works in terms of the density matrix:

$$\rho(\mathbf{r}, \mathbf{r}') = \sum_n f_n \psi_n(\mathbf{r}) \psi^*(\mathbf{r}') \tag{3.41}$$

where f_n is the occupancy and the Kohn-Sham orbitals are $\psi_n(\mathbf{r})$ and $\psi_n(\mathbf{r}')$. The density matrix is idempotent, such that:

$$\rho^2(\mathbf{r}, \mathbf{r}') = \int \mathrm{d}\mathbf{r}'' \rho(\mathbf{r}, \mathbf{r}') \rho(\mathbf{r}'', \mathbf{r}') = \rho(\mathbf{r}, \mathbf{r}') \tag{3.42}$$

which requires the orthonormality of the orbitals from equation (3.21) and the Aufbau principle of singly occupying all states up to the chemical potential, which itself follows from the Pauli exclusion principle. Therefore this ensures the single occupancy of all states up to the chemical potential. The density matrix can then be seen as the position representation of the projection operator onto the space of occupied states $\hat{\rho}$. The charge density $n(\mathbf{r})$ can now be found from the diagonal elements of the density matrix:

$$n(\mathbf{r}) = 2\rho(\mathbf{r}, \mathbf{r}) \tag{3.43}$$

and the total energy of the system can be defined by:

$$E = 2\mathrm{Tr}(\hat{\rho}\hat{H}) \tag{3.44}$$

where $\hat{H}$ is the Hamiltonian from equation (3.21), whose solutions can be found by minimising the energy with respect to the density matrix, subject to idempotency and normalisation constraints:

$$2 \int \mathrm{d}\mathbf{r} \rho(\mathbf{r}, \mathbf{r}) = N_e \tag{3.45}$$

ensuring the density-matrix corresponds to a system of N_e electrons. However, despite this density-matrix reformulation of Kohn-Sham DFT, there still remains the fact that the number of occupied states is directly proportional to N, with each state extending over the entirety of the system so the amount of information in the resultant density-matrix will scale quadratically, as will any associated density-matrix manipulation. If a linear-scaling method is to be found, the nearsightedness of quantum mechanics will need to be exploited. It is Walter Kohn's principle of nearsightedness [117, 118] that tells us the electronic structure of quantum many-body systems is localised. Kohn defines a local volume described in terms of a typical de Broglie wavelength associated with the ground state wave function of the system. Any changes to distant parts of the system (far from all points in the local volume) have a negligible effect on the electronic structure in the local volume. Combined with the consequence of quantum interference effects, the density matrix for systems with a finite band gap is short ranged:

$$\lim_{|\mathbf{r}-\mathbf{r}'|\to\infty} \rho(\mathbf{r},\mathbf{r}') \sim \exp\left(-\gamma|\mathbf{r}-\mathbf{r}'|\right) \to 0 \tag{3.46}$$

where the decay constant γ depends on the energy gap between the highest occupied and lowest unoccupied molecular orbitals (HOMO-LUMO gap), a quantity which is independent of system size and which is the focus of Chapter 5. Therefore the significant information contained in the density matrix scales linearly with the size of the system. The principles discussed in this section are implemented practically in the ONETEP code which is the focus of discussion in the next section.

3.4 The ONETEP code

ONETEP [98] is a linear-scaling DFT package designed for use on parallel computers [119] that uniquely combines near-complete basis set accuracy with a computational cost that scales linearly with the number of atoms. This allows an accurate QM description of systems of thousands of atoms [85], including a range of applications to biomolecular systems [99, 120–122]. Linear scaling is achieved with ONETEP by reformulating conventional Kohn-Sham DFT [56, 58] in order to exploit the "near-sightedness" of the single-particle density matrix in non-metallic systems [117,118]. In terms of Kohn-Sham orbitals the density matrix is expressed as in equation (3.41). However, as the search for the ground state in terms of the density matrix can not be made in terms of the original six-dimensional object, the most common approach is to assume that the density matrix is separable and to work in terms of localised orbitals. Within ONETEP, the density matrix is represented as:

$$\rho(\mathbf{r},\mathbf{r}') = \sum_{\alpha\beta} \phi_\alpha(\mathbf{r}) K^{\alpha\beta} \phi^*_\beta(\mathbf{r}') \tag{3.47}$$

where $\phi_\alpha(\mathbf{r})$ are non-orthogonal generalised Wannier functions (NGWFs) [123] that are localized in real space. In practice, linear scaling arises from enforcing strict localisation

of the NGWFs onto atom-centred regions of fixed radii $\{r_\alpha\}$. The density kernel ($K^{\alpha\beta}$) is a representation of f_n in the duals of the NGWFs and is required to be sparse. This is achieved by discarding elements corresponding to NGWFs centred further apart than some user-defined cutoff r_K. However, a consequence of the non-orthogonality of the NGWFs, combined with the fact that the density kernel is related to the duals of the NGWFs and not the NGWFs themselves, the kernel cutoff r_K is not simply $r_\alpha + r_\beta$. Optimising the NGWFs *in situ* allows for a minimal number of NGWFs to be used whilst maintaining plane-wave accuracy. The basis set underlying the NGWFs consists of periodic cardinal sine (psinc) functions [124] that are related to plane waves by a unitary transformation. The use of a plane-wave basis allows for an unbiased approach to DFT calculations with systematically improvable accuracy through varying a single parameter similar to the energy cutoff in conventional plane-wave DFT packages. The NGWFs are then those functions, when traced with their corresponding optimised density kernel, which reproduce the ground-state density-matrix, whence the ground-state energy [125]:

$$E_0 = \min_n E[n] = \min_{\hat{\rho}} E[\hat{\rho}]_{\hat{\rho}=\hat{\rho}^2} = \min_{\mathbf{K},\phi} E[\mathbf{K},\phi]_{KSK} \tag{3.48}$$

The task is to then extremise the total energy with respect to idempotent density matrices. This is achieved in practice via two nested conjugate gradient variational minimisations. Within the inner loop, for some fixed NGWF expansion, the energy is minimised with respect to the density kernel elements. Then in the outer loop, the total energy is minimised with respect to the coefficients of the NGWF psinc expansion whilst the density kernel remains fixed. In order to impose the idempotency constraint from equation (3.42), a combination of a penalty functional method [126] and the approach of Li, Nunes and Vanderbilt [127] (and independently of Daw [128]), based on McWeeny's purification transformation [129], is used. The purification transformation is defined in terms of some auxiliary matrix $\sigma(\mathbf{r},\mathbf{r}')$:

$$\rho(\mathbf{r},\mathbf{r}') = 3\sigma^2(\mathbf{r},\mathbf{r}') - 2\sigma^3(\mathbf{r},\mathbf{r}') \tag{3.49}$$

Multiple iterations of this transformation will result in any eigenvalues around zero vanishing quadratically while the eigenvalues close to one will converge to that value. In the case of $\rho(\mathbf{r},\mathbf{r}')$, these eigenvalues are the occupation numbers f_n. Therefore, expressing $\rho(\mathbf{r},\mathbf{r}')$ as such, combined with optimising σ, will apply the constraint of idempotency, subject to the occupation numbers remaining in a sensible interval, something that is expected if a system is physically meaningful. General linear-scaling approaches implementing a fixed set of local orbitals will usually use atomic-type functions. Typically, these will be initiated as a solution to the Kohn-Sham equation for atoms inside spherical confinement potentials, obtained using an atom-solver approach as described in the next section. Using a minimal set of orbitals, in order to minimise calculation time, often leads to inaccuracies. This issue can be resolved by simply adding to the set of orbitals used, creating split valence, or multiple-zeta, sets where often additional polarisation functions are included to treat the atomic response to an applied E field. A different approach,

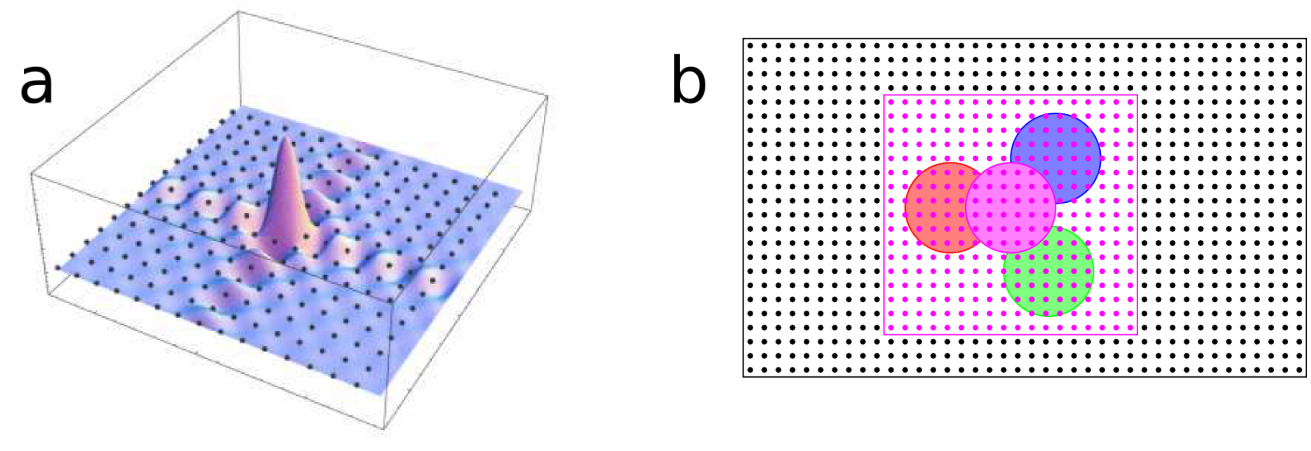


Figure 3.3: (a) A psinc basis function used to expand the NGWFs in ONETEP [131]. (b) An illustration of the 'FFT box' technique used in ONETEP [131].

the one which is used in ONETEP, optimises the orbitals within the environment of the system under calculation, meaning one no longer has to increase the orbital set to ensure transferability. In addition, following this *in situ* optimisation, basis set superposition error corrections are not needed [130].

3.4.1 The periodic cardinal sine function basis set

In order for the NGWFs to be successfully optimised, they must be expanded in terms of a primitive set of functions. This is achieved in ONETEP by using a basis set of periodic cardinal sine (psinc) functions [123, 124]. There is one function centred on each point of a mesh commensurate with the simulation cell and a representation of a single basis function can be seen in Figure 3.3(a). Through varying the fineness of the mesh used for the psincs, the quality of the basis set can be controlled in a manner corresponding to the energy cutoff used to control plane-wave basis sets. These psincs are related to plane waves by a unitary transformation and so give the advantages of both the localised-orbital type and the plane-wave type of basis sets. This relation also allows the efficient calculation of the kinetic energy through the use of fast Fourier transforms (FFTs) [132]. FFTs performed over the entirety of the simulation cell would scale as $\mathcal{O}(M \log M)$ for each NGWF, where M is the number of grid points. In order to achieve linear-scaling behaviour, the 'FFT box' technique is used [133, 134]. In this scheme, the transforms are performed within a box large enough to enclose all overlapping NGWFs and an illustration of this approach can be seen in Figure 3.3(b). The number of grid points in the box is fixed throughout the calculation and depends only on the spacing of the psinc grid and the maximum number of overlapping NGWFs. Therefore, linear-scaling behaviour can be maintained.

For an atom-centred basis set, care must be taken when selecting the basis functions for the system being studied in order to ensure computational accuracy and efficiency. Whilst the *in situ* NGWF optimisation in ONETEP seeks to ameliorate these problems

surrounding careful basis set selection, a sensible choice of initial functions will nevertheless speed up the optimisation process. In order to provide a sensible estimate of the form of the initial NGWFs, resulting in an initial starting wave function closer to the ground state of the system, the atom solver can be used that performs a Kohn-Sham DFT calculation on a pseudoatom in a spherical confinement potential [135]. NGWFs are then initialised as pseudoatomic orbitals, which are obtained through solving the Kohn-Sham equation for a free atom, where the Hamiltonian from equation (3.20) can be decomposed into kinetic, local potential and non-local potential contributions:

$$\hat{H} = -\frac{1}{2}\nabla^2 + V_{\text{loc}}(\mathbf{r}) + \hat{V}_{\text{nl}} \tag{3.50}$$

where $V_{\text{loc}}(\mathbf{r})$ is the local effective potential and and $\hat{V}_{\text{nl}}$ is the nonlocal part of the pseudopotential. The local potential effective potential $V_{\text{loc}}(\mathbf{r})$ is defined as:

$$V_{\text{loc}}(\mathbf{r}) = V_{\text{loc}}^{\text{PS}}(\mathbf{r}) + V_{\text{H}}[n(\mathbf{r})] + V_{\text{xc}}[n(\mathbf{r})] + V_{\text{conf}}(\mathbf{r}) \tag{3.51}$$

where $V_{\text{loc}}^{\text{PS}}(\mathbf{r})$ is the local part of the pseudopotential, $V_{\text{xc}}[n(\mathbf{r})]$ has the same form as (3.37) and $V_{\text{H}}[n(\mathbf{r})]$ is the Hartree potential for a spherical charge distribution. $V_{\text{conf}}(\mathbf{r})$ is a confining potential of the form [86]:

$$V_{\text{conf}}(\mathbf{r}) = S \exp\left[\frac{-w_l}{r - R_c + w_l}\right] (r - R_c)^{-1} \tag{3.52}$$

where the value of this potential is zero for small to medium distances from the atomic centre but it increases rapidly when close to a predetermined confinement radius R_c. This is used to ensure a smooth decay of the atomic orbitals outside R_c. S is the maximum height of the confining potential at $r = R_c$, and w_l is the width of the region over which it is applied. Throughout the calculations in this dissertation, values of $S = 100$ Ha and $w_l = 3.0$ a_0 are implemented for all l-channels, or angular momenta, used. As the atomic eigenstates are solved for the native pseudopotential and exchange-correlation functional of the system being studied, they are a much better starting choice than atom-centred basis sets commonly used in quantum chemistry such as all-electron GTOs.

It is important to note at this stage that, as with all grid-based methods, if the preparation of the calculation is not carefully checked, this approach can suffer from so-called space rippling problems whereby the homogeneity of space is lost by the discretisation. This can result in spurious forces appearing. This effect is most apparent when considering the oscillation of the total energy when a single atom is moved across the simulation box, demonstrating the so-called 'eggbox' effect [136–138]. The problem will diminish for grids that are fine enough and, hence, convergence with respect to the grid spacing, to an acceptable level, must be obtained for all calculations [139, 140]. For all calculations presented in this dissertation, NGWFs were initialised as atomic orbitals obtained using the atom solver approach described in this section to solve the Kohn-Sham equation for atoms in spherical confinement potentials, with a $1s$ configuration for hydrogen, a $2s2p$ configuration for carbon, nitrogen and oxygen and a $3s3p$ configuration for sulphur, when the relevant chemical elements are required.

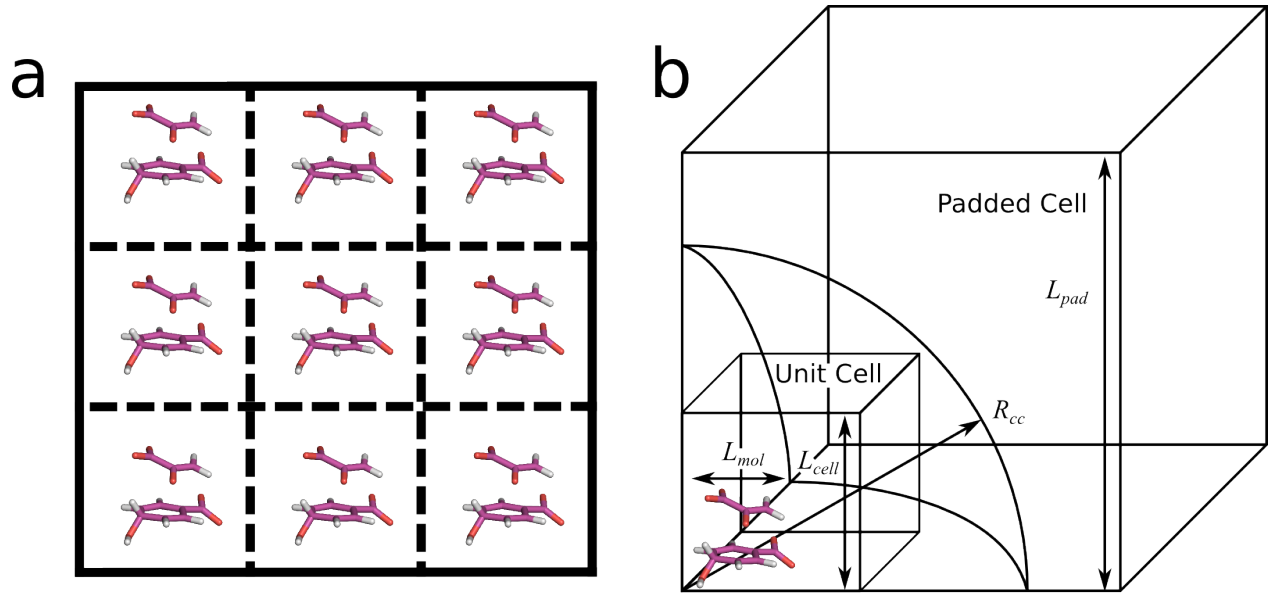


Figure 3.4: (a) Schematic representation of the supercell approach for treating isolated molecules with periodic boundary conditions (PBCs). The molecule is embedded in a repeated unit supercell with boundaries demarcated by dashed lines. (b) Schematic representation of the padded cell in the cutoff Coulomb approach. Figure adapted from Ref. [144].

3.4.2 Cutoff Coulomb interactions

The plane-wave pseudopotential method was developed with crystalline matter in mind, in which periodic boundary conditions (PBC) are required for the calculations. As the Hartree interaction is diagonal in reciprocal space, FFTs are used to calculate the Hartree potential and associated energy. When calculating the properties of bulk solids the influence of periodic images is desirable as it models the true extended bulk system. However, when simulating isolated, finite systems within PBCs the supercell approximation [141–143] must be used. This method replaces a truly isolated system with periodic images with vacuum added around the system to minimise the influence of the periodic replicas on one another. A representation of this approach can be seen in Figure 3.4(a). This approach has been shown to introduce so-called finite size effects whereby the calculated total energy of the system, along with other properties, varies with the size of the supercell [145]. In addition, with particularly large systems there is scope for significant long-range redistribution of the charge due to effects of the periodic images. In order to reach the truly isolated, non-periodic limit, one can simply increase the size of the cell, but this can result in a prohibitively large size for systems such as biomolecules. Even though, in principle, the speed of a ONETEP calculation is independent of cell size, provided enough CPUs are available, in practice, memory requirements limit the cell size. In addition, the decay of the interaction of periodic replicas of a monopole charge goes as $\frac{1}{r}$ so, in practice, the infinite limit is impossible to reach for systems with a non-zero charge. One solution

to this problem is to truncate the Coulomb interaction in real space [146,147] and this approach has been implemented in the ONETEP code [148] and the OCTOPUS code [149,150]. Using a modified form of the Coulomb interaction, the usual FFT approach can still be used along with a periodic supercell, but the Coulomb potential is confined within the primary simulation cell. Essentially, the periodic, background-neutralised Coulomb potential is replaced with a bare counterpart. This replacement interaction is truncated to ensure no part of the simulation cell feels the potential from any of the neighbouring periodic images. It is desirable to maintain the simplicity of a diagonal interaction in reciprocal space. In order to achieve this whilst avoiding the influence of periodic images, the following, 'cutoff', form for the Coulomb potential should be used:

$$V_{\mathrm{CC}}(\mathbf{r}-\mathbf{r}') = \begin{cases} \frac{1}{|\mathbf{r}-\mathbf{r}'|}, & \mathbf{r}-\mathbf{r}' \in \mathcal{R}_1 \\ 0, & \mathbf{r}-\mathbf{r}' \notin \mathcal{R}_1 \end{cases} \tag{3.53}$$

where $\mathcal{R}_1$ is defined as a region of specific shape chosen such that whenever it is centred on some point $\mathbf{r}$, where the Hartree potential is required, it encloses all $\mathbf{r}'$ for which $n(\mathbf{r}+\mathbf{r}') \neq 0$. The region may be anywhere inside the simulation cell or it may just comprise any amount of non-zero density. Calculation of the Hartree potential using the cutoff Coulomb potential in ONETEP is performed on a unit cell padded with vacuum such that, for a spherical cutoff region $\mathcal{R}_1$, the contribution of the electrostatic potential from a periodic image never falls within the cutoff Coulomb radius R_{cc} of the unit cell for all points where the Hartree potential is required. The padded cell can be seen in Figure 3.4(b). L_{mol} is the length scale of the isolated molecule, defined as the largest distance between the edges of the any two NGWFs on the molecule. L_{cell} and L_{pad} are the side lengths of the (cubic) unit and padded cells, respectively. The sphere, of radius R_{cc}, must encapsulate all non-zero density in the unit cell of the molecule for all points where $V_{\mathrm{H}}(\mathbf{r})$ is required. By setting $L_{pad} \geq L_{mol} + R_{cc}$ one can ensure that densities from periodic images never intersect the density within the unit cell. Using this form of the potential has consequences. The corresponding cutoff form of the Coulomb interaction must also be used as a replacement for the long-ranged Coulombic tails of the ion cores in the local pseudopotential form. In a similar vein, when calculating the forces acting on the ion cores, the periodic Coulomb and Ewald terms are replaced with their cutoff Coulomb forms. The spherical variant of the cutoff Coulomb approach has been used throughout the calculations in this dissertation to eliminate all interactions of the molecules with their periodic images.

3.4.3 Implicit solvation

The accurate simulation of the biochemical processes that take place in proteins and enzymes requires careful treatment of solvation effects. Simply including more water atoms in a simulation results in very expensive calculations requiring extensive averaging

over the solvent degrees of freedom. Only a small proportion of the solvent molecules are involved chemically; it is the long range electrostatic effects of the solvent that are most significant. In the implicit solvent approach implemented in ONETEP it is only the atomic details of the solute that are kept. The solute is then placed inside a suitably defined cavity and the solvent environment is represented by an unstructured dielectric continuum outside of this cavity. A plethora of approaches for treating solvation of molecules is available, a review of which is available elsewhere [151]. Many of the models proposed in the literature are based on the self-consistent reaction field (SCRF) mechanism whereby the effect of the electric field due to the dielectric, polarised by the solute, is included within the Hamiltonian self-consistently. Notable variants of the SCRF-type model which are widely used are the polarisable continuum model (PCM) [152] and the conductor-like screening model (COSMO) [153]. Within the many models proposed the shape of the cavity containing the solute has varied. More recent proposals have constructed cavities based on overlapping atomic spheres of varying radii, requiring numerous parameters. Another such example of heavy parameterisation is the SMD model [154], founded in the integral equation formalism of the PCM approach [155]. In this particular solvation model, the 'D' stands for density, denoting the full solute electron density is used without defining partial charges.

In contrast to this, a recent proposal defines the dielectric as a functional of the electronic density of the solute [156]. This was then further developed to include the calculation of the cavitation energy, defining it in terms of the quantum surface of the solute [157]. In other SCRF-type models the solute cavity has a discontinuous boundary. In this formulation a smooth transition of the relative permittivity is defined by:

$$\epsilon[n(\mathbf{r})] = 1 + \frac{\epsilon_\infty - 1}{2}\left(1 + \frac{1 - (n(\mathbf{r})/n_0)^{2\beta}}{1 + (n(\mathbf{r})/n_0)^{2\beta}}\right) \tag{3.54}$$

where $n(\mathbf{r})$ is the electronic density of the solute, ϵ_∞ is the bulk permittivity, β is a parameter controlling the transition of $\epsilon[\mathbf{r}]$ from unity to ϵ_∞ and n_0 is the value of the density for which the permittivity drops to half that of the bulk. However, the original formulation did not include dispersion-repulsion effects and also required an *a posteriori* correction to the energy in vacuum obtained in periodic boundary conditions in order to approximate open boundary conditions. These shortcomings are overcome by Dziedzic *et al.* who include dispersion interactions with the solvent, using appropriate boundary conditions and redetermining the two parameters in the dielectric functional [158, 159]. Another issue with the original formulation is that as the dielectric cavity responds self-consistently to changes in the electronic density the functional derivative of the electrostatic energy now introduces a numerical instability. Dziedzic *et al.* circumvent this instability without loss of accuracy by fixing the dielectric cavity. This is achieved by first solving the homogenous Poisson equation (HPE) for the system in vacuum:

$$\nabla^2 \phi_{\mathrm{HPE}}(\mathbf{r}) = -4\pi\rho_{\mathrm{tot}}(\mathbf{r}) \tag{3.55}$$

in open boundary conditions, where, in principle, one would set up Dirichlet boundary conditions, of the form:

$$\phi_{\mathrm{BC}}^{\mathrm{vac}}(\mathbf{r}) = \int_{\Omega} \frac{n_{\mathrm{tot}}(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} d\mathbf{r}' \quad \forall \quad \mathbf{r} \in \partial\Omega \tag{3.56}$$

over the simulation cell (Ω) on the faces of the simulation cell ($\partial\Omega$). However, in practice, in order to reduce the computational cost, a coarse-grained version of the electron density ($n_{\mathrm{tot}}^{\mathrm{CG}}$) is used. This density is constructed as a set of N_{CG} point charges corresponding to a cubic block of the simulation cell. The magnitude of each point charge is the sum of the charges on the grid points belonging to the block and the charge is positioned at $\mathbf{R}_l$, the centre of charge of the block. The integral in equation (3.56) can now be replaced by a sum over this small number of point charges, so the potential on the faces of the cell is approximated as:

$$\phi_{\mathrm{BC}}^{\mathrm{vac}}(\mathbf{r}) \approx \sum_{l}^{N_{\mathrm{CG}}} \frac{n_{\mathrm{tot}}^{\mathrm{CG}}(\mathbf{R}_l)}{|\mathbf{r}-\mathbf{R}_l|} \quad \forall \quad \mathbf{r} \in \partial\Omega \tag{3.57}$$

The electronic density is represented by $n(\mathbf{r})$ and $n_{\mathrm{tot}}(\mathbf{r})$ is the total density due to the electrons and nuclei (or rather, the ionic cores in the case where pseudopotentials are used). The converged electronic density from vacuum is then used to generate the density-dependent dielectric cavity in solution, obtained from direct solution of the inhomogeneous Poisson equation in real space (the equation is homogeneous in vacuum and inhomogeneous in solution), again under open boundary conditions:

$$\nabla \cdot \big(\epsilon[n(\mathbf{r})]\nabla\phi(\mathbf{r})\big) = -4\pi n_{\mathrm{tot}}(\mathbf{r}) \tag{3.58}$$

However, in solution it is equation (3.58) that needs to be solved but the open boundary conditions can no longer be obtained from equation (3.57). In this instance, the form for the potential is:

$$\phi_{\mathrm{BC}}^{\mathrm{sol}}(\mathbf{r}) \approx \frac{1}{\epsilon_\infty} \sum_{l}^{N_{\mathrm{CG}}} \frac{n_{\mathrm{tot}}^{\mathrm{CG}}(\mathbf{R}_l)}{|\mathbf{r}-\mathbf{R}_l|} \quad \forall \quad \mathbf{r} \in \partial\Omega \tag{3.59}$$

where the dielectric permittivity is assumed to be homogeneous and to have the bulk value of ϵ_∞ everywhere. By constructing the dielectric cavity by application of equation (3.58) to the converged electronic density of the solute obtained from the vacuum calculation, and keeping the cavity fixed throughout the calculation in solvent, the numerical instability is avoided and the associated reduction in accuracy is insignificant. The implicit solvent approach outlined by Dziedzic *et al.* is implemented in ONETEP. In short, when using this approach, through the use of a smeared-ion formalism the molecular Hartree energy is obtained not in reciprocal space, like standard ONETEP, but rather by solving the Poisson equation in real space, as described. To briefly discuss the practicalities of the calculations, the results are achieved via the use of a multigrid approach detailed elsewhere [100, 158]. Within all implicit calculations presented in this dissertation, the ion smearing width is 0.8 a_0 and the values of the solvation parameter β and electronic density threshold n_0

were 1.3 and 3.5×10^{-4} a.u. respectively, as proposed in Ref. [160]. The relative dielectric permittivity of the solvent was set to 80.0 for all implicit solvent calculations. A limitation of multigrid solvers is that every dimension of the grid used in the solver must be a magic number, defined to be of the form $32k+1$, with allowed sizes of 33, 65, 97, 129, 161, 193, 225 and so on. Therefore, there is a certain granularity to the allowed grid sizes for the solver. In ONETEP a fine grid is used in solvation with even dimensions, and thus never magic. To resolve this difficulty, only the subset of the fine grid that is obtained by rounding its dimensions down to the nearest magic number is used in solvation calculations. For a calculation with a psinc grid spacing 0.5, and a cubic cell 42.5 a_0 in size, this will yield a fine grid that is 170×170×170. This value of 170 will be rounded down to the nearest magic number, 161, and only the lower portion of the fine grid, 161×161×161 in size, will be passed to the multigrid. It is then up to the user to ensure that no electron density is contained within the unused margin of 162 to 170. If any NGWFs extend beyond that portion then there will be errors introduced to the calculations. As a rule of thumb, one should have at least about 10 a_0 of vacuum/solvent around the molecule's NGWFs (not atomic positions) on each side of the simulation cell but should also be mindful of minimising the margin, so that as little memory as possible is wasted, as the memory requirement of the solver grows cubically with the grid size.

3.4.4 Calculating the local/partial density of states

The density of states (DoS) is defined as the measure of how many states or, in an electronic structure calculation, how many eigenstates there are at a particular energy or specifically in an energy window. The DoS can be used to analyse the electronic structure of a system and can often be used to compare with experimental techniques such as scanning tunnelling microscopy. In order to provide the required eigenvalues and eigenvectors, the problem to be solved is the generalised eigenproblem:

$$\sum_{\beta} H_{\alpha\beta} M_n^{\beta} = \epsilon_n \sum_{\beta} S_{\alpha\beta} M_n^{\beta} \tag{3.60}$$

where M_n^{β} is a matrix describing the eigenvectors, taking the form of:

$$|\psi_n\rangle = \sum_{\beta} M_n^{\beta} |\phi_{\beta}\rangle \tag{3.61}$$

From the resulting eigenvalues $\{\epsilon_n\}$ and eigenvectors M_n^{β} the total density of states $n(E)$ can be obtained:

$$n(E) = \sum_{n} \delta(E - \epsilon_n) \tag{3.62}$$

where, in practice, the delta function is often replaced with some Gaussian broadening function of user-specified width, typically on the order of 0.1 eV. The physical justification for this lies in the fact that thermal fluctuations will broaden energy levels in a system.

The total DoS of a system may not be particularly useful when considering a system with inherently local features such as surfaces, defects or reaction centres. This quantity is even less useful in large-scale systems studied in this dissertation. In such scenarios it would be more informative to enquire as to what the DoS associated with a particular atom or a certain group of atoms is equal to. Local density of states (LDoS) calculations can provide some of the most valuable sources of information required to interpret and understand electronic structure calculations. The LDoS gives a description that encompasses both the spatial and the energetic distribution of the single-particle eigenstates, simultaneously. By performing a diagonalisation of the Hamiltonian matrix in the basis of the NGWFs, the LDoS decomposition is achieved. This is performed in the local-orbital framework of ONETEP after NGWF and density kernel convergence has been reached. This procedure has a cubic scaling computational cost associated with it, but also has a low prefactor due to the small NGWF basis. In order to calculate the local density of states for a given region, each eigenstate must be projected onto the local orbitals contained within that region. What is obtained is a series of functions for each of the chosen regions, which may be the NGWFs of a single atom, or those of a group of atom types. These may be NGWFs of a single atom, or perhaps a group of atom types. Examples of the LDoS capabilities in ONETEP can be found in the literature [161, 162] and also in Chapter 5 of this dissertation.

3.4.5 Empirical dispersion corrections

Traditional DFT, with commonly used exchange-correlation density functionals, provides an incomplete description of the dispersion interactions required for an accurate description of protein complexes and enzymes, which would be captured by a 'perfect' functional. The DFT energy can be corrected for dispersion by modifying it as such:

$$E_{\mathrm{DFT+D}} = E_{\mathrm{DFT}} + E_{\mathrm{disp}} \tag{3.63}$$

where the dispersion energy correction for N atoms is given by:

$$E_{\mathrm{disp}}(r_{ij}) = -s_6 \sum_{ij,i>j} \frac{C_{6,ij}}{r_{ij}^6} f_{\mathrm{damp}}(r_{ij}) \tag{3.64}$$

where s_6 is a global scaling factor, typically used to adjust the correction to the repulsive behaviour of the chosen density functional [163]. $C_{6,ij}$ is a dispersion coefficient for atom pair ij. $f_{\mathrm{damp}}(r_{ij})$ is a damping function that is equal to unity at large separations (r_{ij}) and 0 at small distances [164, 165]. This damping function is required because electronic structure calculations provide an adequate description of the short-ranged attractions, therefore the empirical correction will become superfluous at small distances. If a damping function is not applied to the dispersion term then the total energy will be distorted, due to the resulting significant artificial strengthening of every covalent bond. All calculations

presented in this dissertation use the form of the damping function due to Elstner *et al.* [166] which takes the form of:

$$f_{\text{damp}}^{\text{Elstner}}(r_{ij}) = \left(1 - exp\left[-c_{\text{damp}}\frac{r_{ij}}{r_{0,ij}}\right]^7\right)^4 \tag{3.65}$$

where c_{damp} is the damping constant and $r_{0,ij}$ is a quantity determined by the van der Waals radii of the atomic pair i and j. In order to improve the description of enzyme systems treated using large scale DFT with ONETEP, dispersion correction schemes have been implemented within the code [165]. Ref. [165] optimised the $C_{6,ij}$ coefficients, the $r_{0,ij}$ and the c_{damp} coefficients against a benchmark set of complexes with dispersion interactions where the binding energies are known for high accuracy. This optimisation was achieved by adjusting the parameters in equation (3.64) in order to minimise the difference between the value of the dispersion energy and the error in the binding energy for each complex. The result of the work presented in Ref. [165] is that ONETEP now has optimised parameters to describe dispersion interactions in four types of damping functions from the literature which can be used with up to six different density functionals present in the code.

3.4.6 Electrostatic embedding and the QM/EE approach

Electrostatic embedding significantly reduces the computational costs associated with large-scale DFT calculations. Within the quantum mechanics/electrostatic embedding (QM/EE) approach implemented in ONETEP, a portion of the total system is represented in terms of highly localised classical charge distributions [167]. By electrostatically coupling the quantum system with classical charge distributions, the effects of the environment in which the quantum system is embedded are accurately represented. The energy of the total embedded system is defined as:

$$E_{\text{QM/EE}} = E_{\text{QM}} + E_{\text{int}} + E_{\text{EE}} \tag{3.66}$$

where E_{QM} is the electronic energy of the quantum system that has its associated charge density and wave functions polarised by the potential due to the embedded charges. The interaction energy between the electrons and nuclei of the quantum system and the embedded charges (E_{int}) is represented as:

$$E_{\text{int}} = \sum_{J}^{N_{\text{at}}}\sum_{a}^{N_{\text{emb}}} Z_J \int \frac{q_a(\mathbf{r}-\mathbf{R}_a)}{|\mathbf{r}-\mathbf{R}_J|}\mathrm{d}\mathbf{r} - \sum_{a}^{N_{\text{emb}}}\int\int \frac{q_a(\mathbf{r}-\mathbf{R}_a)n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\mathrm{d}\mathbf{r}\mathrm{d}\mathbf{r}' \tag{3.67}$$

for an environment of N_{emb} atomistic charge distributions $q_a(\mathbf{r}-\mathbf{R}_a)$ localised around the point $\mathbf{R}_a$ and evaluated at position $\mathbf{r}$. The first term on the right hand side of equation (3.67) represents the Coulomb interaction energy between N_{at} nuclei of atomic number Z_J and the second term represents the Coulomb interaction energy between the charge

density $n(\mathbf{r})$ and the embedded charges. The energy of interaction between the embedding charges, E_{EE}, is represented as:

$$\sum_{a,b>a}^{N_{\mathrm{emb}}} \int\int \frac{q_a(\mathbf{r}-\mathbf{R}_a)q_b(\mathbf{r}'-\mathbf{R}_b)}{|\mathbf{r}-\mathbf{r}'|}\mathrm{d}\mathbf{r}\mathrm{d}\mathbf{r}' \tag{3.68}$$

Within DFT calculations, the embedded charges are present throughout the QM calculation via an SCF approach and the resulting charge density is polarised by the classical charge distribution. In order for self-consistent embedding to occur, the Kohn-Sham Hamiltonian previously defined in equation (3.21) requires an additional term to describe the potential each electron will experience from the embedded charges:

$$\hat{H}_{\mathrm{KS/EE}} = \hat{T} + \hat{V}_{\mathrm{H}} + \hat{V}_{\mathrm{xc}} + \hat{V}_{\mathrm{ext}} + \hat{V}_{\mathrm{emb}} \tag{3.69}$$

where the potential due to the embedded charges, $\hat{V}_{\mathrm{emb}}$, is represented by:

$$\hat{V}_{\mathrm{emb}}(\mathbf{r}) = \sum_{a}^{N_{\mathrm{emb}}} \int \frac{q_a(\mathbf{r}'-\mathbf{R}_a)}{|\mathbf{r}-\mathbf{r}'|}\mathrm{d}\mathbf{r}' \tag{3.70}$$

Throughout this dissertation the TIP3P model [168] for the charge distribution of water has been used where $q(O) = -0.834e$ and $q(H) = 0.417e$.

3.4.7 Natural bond orbital analysis

NBO analysis provides a chemical picture of bonding in terms of localised Lewis-type bond and lone pair orbitals. Such an analysis is helpful as state-of-the-art first-principles electronic structure calculations, whilst able to provide an accurate description of the system under study in terms of the total electron density, often do not provide a very good description of the qualitative chemical information available. ONETEP has been interfaced with the NBO 5 analysis program [169] in order to provide chemical insights into subregions of large systems by studying effects such as electronic delocalisation [170]. An example of this type of analysis for methylamine, showing the expected double occupancy of Lewis-type bonding orbitals and vacancy of their antibonding counterparts, can be seen in Figure 3.5. This approach works by transforming optimised NGWFs from ONETEP into atom-centred, orthogonal natural atomic orbitals (NAOs) [171], then into natural hybrid orbitals (NHOs) [172]. NHOs are the individual atom-centred hybrids that then constitute a two-centred natural bond orbital (NBO). NBO analysis can be performed within a localised region (such as an enzyme active site) in such a way as to ensure the results are in fact identical to a calculation on the entire system.

One of the most interesting and biologically relevant effects that can be studied with this combined approach is electronic delocalisation. Delocalised charge transfer represents a deviation from the ideal Lewis description in the NBO formalism. The variational energetic lowering due to charge transfer from bonding to anti-bonding NBOs [173] can

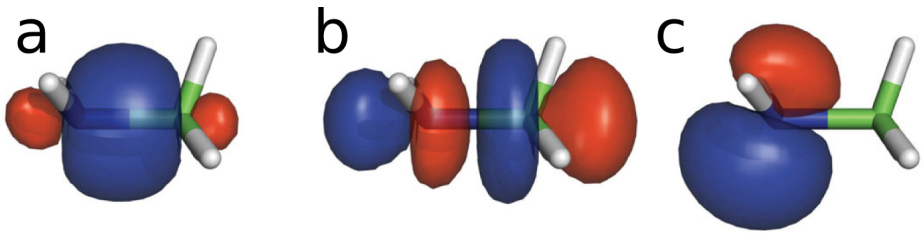


Figure 3.5: Examples of NBOs representing the (a) C-N σ bond, (b) C-N σ^* antibond and the nitrogen n lone pair of methylamine, obtained in ONETEP from the final optimised NGWFs. NBOs have been normalized to unity and plotted with an isosurface value of ± 0.05 a.u. (red is positive and blue is negative). Figure adapted from Ref. [170].

be estimated through the use of second-order perturbation theory [174–176]. In the instance of a $\sigma_i \rightarrow \sigma_j^*$ donor-acceptor interaction, the stabilising effects can be estimated by inspecting the elements of the first nonzero energetic correction due to off-diagonal couplings:

$$\Delta E^{(2)}_{i \rightarrow j^*} = \sum_{j \neq i} f_i \frac{\langle \sigma_i | \hat{H} | \sigma_j^* \rangle^2}{\epsilon_{j^*} - \epsilon_i} \tag{3.71}$$

where $\epsilon_i = \langle \sigma_i | \hat{H} | \sigma_i \rangle$ and $\epsilon_{j^*} = \langle \sigma_j^* | \hat{H} | \sigma_j^* \rangle$ are the orbital energies of the donor and acceptor NBOs, respectively. These $\Delta E^{(2)}$ values can be thought of as an intermolecular analogue of the stabilising intramolecular donation of electron density in hyperconjugation [177,178]. An example is the $n \rightarrow \sigma^*$ secondary hyperconjugation interaction of a hydrogen bond involving intermolecular delocalisation (charge transfer) between a lone pair donor n and an antibond acceptor σ^* of an adjacent molecule. It has been shown that such interactions are prevalent within biological systems, acting to stabilise protein and nucleic acid structures and also regulating their interactions with their environment [174, 175]. It has also been shown that the energetic lowering due to charge transfer that is calculated by second-order perturbation theory is correlated with the strengths of hydrogen bonds [176]. Thus the framework gives a means to qualitatively assess hydrogen-bond strength from a single large-scale DFT calculation. In Chapter 6 of this dissertation, an investigation of the interactions within an enzyme-substrate complex, using the natural bond orbital (NBO) analysis approach outlined here, has been performed, detailing the interactions between lone pair and antibonding orbitals between the substrate and active-site residues. By inspecting the changes in $\Delta E^{(2)}$ energies at the stationary points in an enzyme-catalysed reaction, the stabilisation effects on the TS arising from donor-acceptor interactions between the substrate and active-site residues can be estimated. NBOs are ultimately constructed from natural atomic orbitals (NAOs) [171] and these form the basis of natural population analysis (NPA), a widely used method for assigning atomic partial charges. These NPA charges are shown to be less basis set dependent than the often used Mulliken charges.

3.4.8 Density derived electrostatic and chemical method for computing net atomic charges

It has been shown in this dissertation that, by virtue of the Hohenberg-Kohn theorems and Kohn-Sham ansatz, the ground state electronic density is sufficient to derive all the information about a system. However, the electron density, by itself, is often not very useful for understanding chemical reactions and it can be conceptually more convenient to assign electrons to individual atoms or fragments. Net atomic charges (NACs) can be used to both understand the chemical states of the atoms in some material and also to accurately represent the electrostatic potential of the material outside of the region occupied by its electron distribution. As will be discussed in Section 3.6, the treatment of molecules as a collection of point charges following the laws of classical mechanics is the basis for most molecular mechanics methods. The same section will also discuss how, for commonly used force fields, the partial charges are fitted to reproduce the QM electrostatic potential of small molecules in an ESP scheme. The resultant ESP charges are well-suited for force fields, reproducing *ab initio* multipole moments and electrostatic interactions between molecular fragments. However, there is no unique method available for the partitioning of the rigorously calculated quantum mechanical (QM) electron density among the individual atoms. Furthermore, different charge derivation schemes can often lead to very different results. In order to aid in the interpretation of QM simulations, atomic point charges should respond in a chemically intuitive manner to their environment. In addition, within biological systems the correct treatment of electrostatics is vital in the accurate determination of the associated molecular interactions.

Electron density-based atoms-in-molecule (AIM) charge partitioning, which is based on the Hirshfeld approach [179], differs from the ESP technique in that the NACs are assigned by dividing a converged electron density into a union of overlapping basins. Shortcomings of the original Hirshfeld method included an arbitrariness in the choice of reference atomic densities used to define these overlapping atomic basins. Such problems are addressed in recently proposed iterative extensions to the Hirshfeld method, in which reference densities are successively improved until self-consistency is achieved in the iterative Hirshfeld (IH) scheme [180] so the resultant atomic densities closely resemble the reference densities of free ions in vacuum, giving chemically meaningful properties. An alternative to the IH scheme is termed iterated stockholder atoms (ISA) where the spherical average of the partitioned atomic density is used as a reference density. This scheme is argued to be less empirical than the IH approach, producing a better fit to the electrostatic potential by constraining the atomic densities to be as close to spherical as possible, ensuring there are no higher order multipoles and allowing the charges to reproduce the desired electrostatic properties. The density derived electrostatic and chemical charges (DDEC) method [181] combines the IH and ISA approaches to assign atomic charges from the electron density. DDEC charges simultaneously reproduce the chemical states of atoms in a material and the electrostatic potential surrounding the material's electron

density distribution. The formal mathematical details of the DDEC approach can be found in Ref. [181] and references therein. Previously, the use of a DDEC AIM scheme was always limited by the computational expense of the underlying QM method. This is a problem as the DDEC approach has many features making it suitable for designing flexible force fields [182]. To resolve this issue, recent work has implemented a DDEC method into ONETEP and this has been shown to generate system-specific charges for a range of large-scale biomolecular systems [183]. An accurate description of the electrostatic potential is critical in understanding enzymatic reaction mechanisms and so, where indicated, the DDEC method has been used in this dissertation to derive accurate atomic partial charge values from a converged electron density using full-DFT. For all DDEC analysis presented in this work, the mixing parameter χ is set to a value of $\frac{3}{14}$ which is shown to give optimum balance between minimising atomic multipoles whilst also maximising chemical accuracy [184]. Vacuum reference densities for Hirshfeld analysis were generated internally by solving the Kohn-Sham equation for free atoms in the presence of a charge compensation sphere [183, 185] and were conditioned to the chemical environment via the method described in Ref. [184]. It has been shown that this approach is suitable for biomolecular systems [183], such as those studied in Chapters 5 and 6. In addition, these charges can be compared with NPA values to ensure consistency of results.

3.5 Structural optimisation

It is the stationary points on the potential energy surface that hold the most important information regarding chemical reactions. These points are defined as the nuclear configurations at which the energy gradient is zero as the forces on the system vanish, specifically:

$$\frac{\partial E(\mathbf{X})}{\partial X_\alpha} = 0 \quad \text{for} \quad 1 \leq \alpha \leq 3N \tag{3.72}$$

The aim of geometry optimisation is to locate these stationary points on the potential energy surface. One of the main objectives within this dissertation is finding the stationary points that correspond to local energy minima and the transition states that connect them. This will allow an enhanced understanding of these chemical reactions in the gas and solution phase along with those catalysed by enzymes.

3.5.1 Calculation of forces

The general idea of the force conjugate to any parameter in the Hamiltonian was first formulated by Ehrenfest in 1927 [186]. It was he who first recognised that this relation is crucial for the correspondence of classical and quantum mechanics. What Ehrenfest showed was an expression for force equal to the expectation value of the operator that corresponds to acceleration $\langle \frac{d^2\hat{x}}{dt^2} \rangle$. These ideas were then implicit in other works by Born and Fock [187] that followed in 1928 and later made explicit by Güttinger [188] in 1931.

These formulae were included in later work by Pauli [189, pp. 83-272] in 1933 but it was Hellmann who reformulated them as a variational principle in a form ready to apply to molecules [190] in 1937. In 1939, Feynman derived the force theorem [191], explicitly pointing out that the force on a nucleus is independent of the electron kinetic energy, exchange and correlation, depending only on the charge density. It is the term "Hellmann-Feynman theorem" that has appeared to stick, now being widely used amongst the community of computationalists implementing force calculations. The force theorem gives us the force conjugate to any parameter describing our system of interest, in this case the parameter is the position of a nucleus $\mathbf{R}_I$, and the force on this nucleus can be written as the negative total derivative of the energy with respect to this parameter:

$$\mathbf{F}_I = -\frac{\mathrm{d}E}{\mathrm{d}\mathbf{R}_I} \tag{3.73}$$

and in the limit of a complete basis set the theorem holds for self-consistent solutions [192–194]. The Hellmann-Feynman theorem has been shown to hold in DFT [195]. In practice, DFT calculations employ a finite number of basis functions. When the basis functions depend explicitly on the positions of the ions, corrections to the Hellmann-Feynman forces must be calculated, in the form of Pulay forces [75]. Within ONETEP, the *in situ* optimisation of the NGWFs with respect to the psinc functions, a systematically improvable basis set independent of the position of the atoms, should, in principle, eliminate the correction due to Pulay forces from the total ionic forces. It has been shown that for strict localisation constraints, especially with small localisation regions, there can be non-negligible Pulay forces that must be calculated as a correction to the Hellmann-Feynman forces in the ground state [140]. Geometry optimization calculations, which rely heavily upon accurate evaluation of the total ionic forces, show much better convergence when Pulay forces are included. In ONETEP, equation (3.73) is redefined in order to include the terms of the implicit dependency of the density kernel and NGWFs on the nuclear coordinates, as well as the explicit dependency of the energy on the nuclear coordinates:

$$-\frac{\mathrm{d}E}{\mathrm{d}\mathbf{R}_I} = -\frac{\partial E}{\partial \mathbf{R}_I} - \frac{\partial E}{\partial K^{\beta\alpha}}\frac{\partial K^{\alpha\beta}}{\partial \mathbf{R}_I} - \int \mathrm{d}\mathbf{r}\frac{\delta E}{\delta \phi_\alpha(\mathbf{r})}\frac{\partial \phi_\alpha(\mathbf{r})}{\partial \mathbf{R}_I} \tag{3.74}$$

However, due to the LNV algorithm described in Ref. [127], the total energy is converged with respect to the density kernel to a very high tolerance, giving the condition of:

$$\frac{\partial E}{\partial K^{\alpha\beta}} = 0 \quad \forall\ \alpha\beta \tag{3.75}$$

and this can be routinely achieved. However, the task of achieving energy convergence with respect to the expansion of the NGWFs with their underlying psinc basis set is somewhat more difficult, resulting in $\frac{\partial E}{\partial \phi_\alpha} \neq 0$ so the last term in equation (3.74) remains, such that:

$$\mathbf{F}_{\mathrm{Pulay}} = \int \mathrm{d}\mathbf{r}\frac{\delta E}{\delta \phi_\alpha(\mathbf{r})}\frac{\partial \phi_\alpha(\mathbf{r})}{\partial \mathbf{R}_I} \tag{3.76}$$

needs to be calculated along with the Hellmann-Feynman term $\frac{\partial E}{\partial \mathbf{R}_I}$. Obtaining the desired tight convergence of energy with respect to the psinc expansion of the NGWFs can be very difficult due to the kinetic energy operator, which has the effect of spreading the NGWFs across the cell. With the added constraint of strict localisation within a sphere, the resulting NGWF energy gradient converges, but often to a small non-zero value. In this instance the total energy will converge quadratically with respect to the KS states but the associated forces will converge at a slower rate. Therefore the residual NGWF energy gradient has an insignificant effect on the ground state energy but the $\mathbf{F}_{\text{Pulay}}$ term is non-negligible and needs to be retained in the calculation of forces. It is expected that Pulay forces will be more significant in the description of the forces in weakly bonded systems such as biomolecules, especially when calculated using localised orbitals of small radii. The Pulay corrections applied to the Hellmann-Feynman forces calculated in ONETEP [140] lead to an improvement in the consistency between ground state energies and the associated forces acting on the system for any size of localisation region. Biomolecular systems which have significant numbers of weakly bound components are much better described when Pulay corrections to the forces are included and, as such, these corrections are used throughout the calculation of forces within this dissertation. Many algorithms have been suggested for the problem of locating local energy minima on the potential energy surface. One of the fastest methods that can be applied to large systems of biomolecular interest is the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm and this is implemented in ONETEP. This implements a quasi-Newton approach and constructs an approximate inverse Hessian matrix of second derivatives from energy gradients calculated at a specified number of previous points. These approaches are used to structurally optimise energy minima for large-scale biomolecular systems investigated in this dissertation.

3.5.2 Normal mode analysis

The characterisation of atomic structures via the calculation of normal modes is an important aspect of biomolecular simulation that should not be overlooked. The process of normal mode analysis (NMA) is primarily used for identifying and characterising the slowest, or lowest frequency, macromolecular motions that would be otherwise inaccessible using other methods such as short timescale molecular dynamics. NMA can, in principle, be applied to system sizes ranging from small protein-ligand complexes up to the ribosome. NMA is defined as the study of harmonic potential wells by some analytical means. To begin studying normal modes, a stable configuration that represents the minimum of the potential energy surface of the system is required. Figure 3.6 illustrates a two-dimensional ($\mathbf{r}_i$) representation of a harmonic potential well. The directions $\mathbf{e}_i$ are the associated normal modes. It can be instructive to imagine the potential well as a bowl in which a classical sphere moves around. If the sphere is moved along one of the normal modes, it will move back and forth in this direction, whereas in any other

direction it would be deflected by the potential along some perpendicular direction. It is only the normal mode directions that are independent for the system. As such, oscillations of the sphere along either of the normal modes will have a defined frequency related to the curvature of the bowl, or potential, along this direction of motion. In order

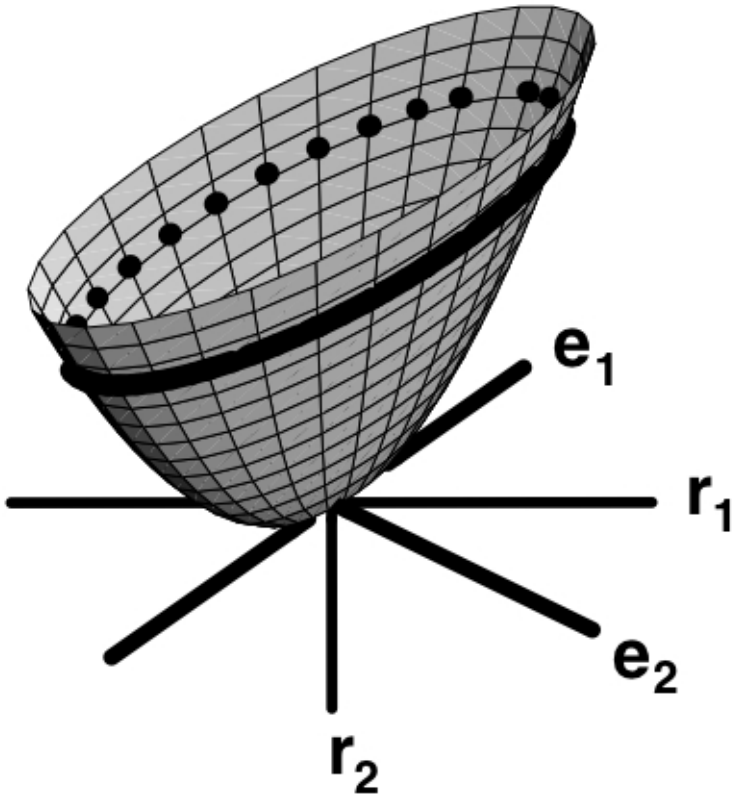


Figure 3.6: A two-dimensional harmonic potential well. The two Cartesian coordinate axes of the system are $\mathbf{r}_1$ and $\mathbf{r}_2$, the two normal mode directions are $\mathbf{e}_1$ and $\mathbf{e}_2$ [196].

to understand this framework mathematically one must first consider an N-dimensional potential $V = V(X_1, X_2, \ldots, X_N)$ with generalised coordinates X_i. Assuming that the energy of the system at some initial position X_i^0 is given by $V(X_i^0)$ then the energy at a new position $X_i = X_i^0 + h_i$ may be approximated through the use of a Taylor series expansion, up to second order [197]:

$$V(X_i) \approx V(X_i^0) + \sum_j \left.\frac{\partial V}{\partial X_j}\right|_{X_{ij}=X_{ij}^0} h_j + \frac{1}{2}\sum_{i,j} h_i \left.\frac{\partial^2 V}{\partial X_i \partial X_j}\right|_{X_{ij}=X_{ij}^0} h_j \tag{3.77}$$

which can be written much more elegantly using matrix notation:

$$V(\mathbf{X}) \approx V(\mathbf{X}^0) + \mathbf{G}^{\mathrm{T}}\mathbf{h} + \frac{1}{2}\mathbf{h}^{\mathrm{T}}\mathbf{H}\mathbf{h} \tag{3.78}$$

where $\mathbf{G}$ is the gradient vector and $\mathbf{H}$ is the Hessian matrix of second derivatives. It is now helpful to recall that the eigenvectors $\mathbf{e}_i$ of our symmetric Hessian can be chosen to be mutually orthogonal with eigenvalues λ_i, expressed as:

$$\mathbf{H}\mathbf{e}_i = \lambda_i \mathbf{e}_i \tag{3.79}$$

and it is the eigenvalues λ_i that describe the curvature of the potential along the normal mode directions. When considering the physical interpretation of normal modes, it is the eigenvalues λ_i that describe the energetic cost of displacing the system of one unit of length along eigenvector $\mathbf{e}_i$. So in fact, NMA is classifying the potential deformations of a protein by their energetic cost. In the presence of a realistic potential, low frequency modes will be associated with collective and delocalised deformations whereas high frequency modes will correspond to local deformations. This can be thought of as arising from the non-linearity of the interaction terms. Short-ranged interactions, such as bond stretching, are stronger, and more rapidly varying with position, than long-ranged interactions such as electrostatics. In practice, when NMA is applied to an isolated system the first six eigenvalues should be zero, but in reality, this test is often failed due to errors in the calculated forces and Hessian matrix. The six zero eigenvalues are due to the fact that they describe the six rigid-body movements of the system. These movements are translation along three independent axes and rotation about three independent axes. These incur no energy cost and as such are ignored in the analysis. Therefore, in practice, the 'non-zero' modes are usually taken to refer to the lowest energy modes possessing non-zero energies. An energy minima on the potential energy surface will have Hessian eigenvalues that are all positive, whereas a transition state will have one negative eigenvalue. In the next chapter, NMA will be used to confirm that this is the case for the calculated energy minima and transition states.

3.5.3 Transition state searching

In comparison to local minima, locating transition states is often much more difficult. One defines a transition state (TS) as a stationary point on the potential energy surface where one of the eigenvalues of the Hessian matrix of second derivatives is negative [198]. A TS therefore corresponds to a local energy maximum in one eigendirection but a minimum in all others. A TS, by definition, has a negative force constant and thus an imaginary vibrational frequency. This tells us that the corresponding motion described by the associated normal coordinate will lower the energy, hence showing that the current TS is not a stable structure in that eigendirection. The calculation of transition states is vital in the prediction of activation energies for enzymatic reactions as it is the TS that must be passed through in order to make the transition from reactant to product. As discussed earlier in Chapter 2, it is transition states that are stabilised by enzymes in order to allow a reaction to occur at a faster rate. A large range of methods exist for calculating transition states. Many have been proposed on the basis of the so-called eigenvector-following technique which is described in Section 3.5.5. This is a so-called single-ended method as it only requires an initial structure, acting as a TS approximant, to start off the calculation. Other methods which are double-ended require two equilibrium geometries between which it is expected a transition state lies that connects the two.

3.5.4 Linear and quadratic synchronous transit methods

The linear synchronous transit (LST) method performs a series of single point calculations on a set of linearly interpolated structures between given reactant (initial) and product (final) atomic configurations as illustrated in Figure 3.7. The path may be defined by:

$$r^i_{\mathrm{ab}}(f) = (1-f)r^{\mathrm{R}}_{\mathrm{ab}} - f r^{\mathrm{P}}_{\mathrm{ab}} \tag{3.80}$$

where $r^{\mathrm{R}}_{\mathrm{ab}}$ and $r^{\mathrm{P}}_{\mathrm{ab}}$ are the internuclear distances between atoms a and b in the reactant and product, respectively. In addition, the interpolation parameter f runs from 0 to 1. A significant drawback with equation (3.80) is that it over-specifies the geometry of the

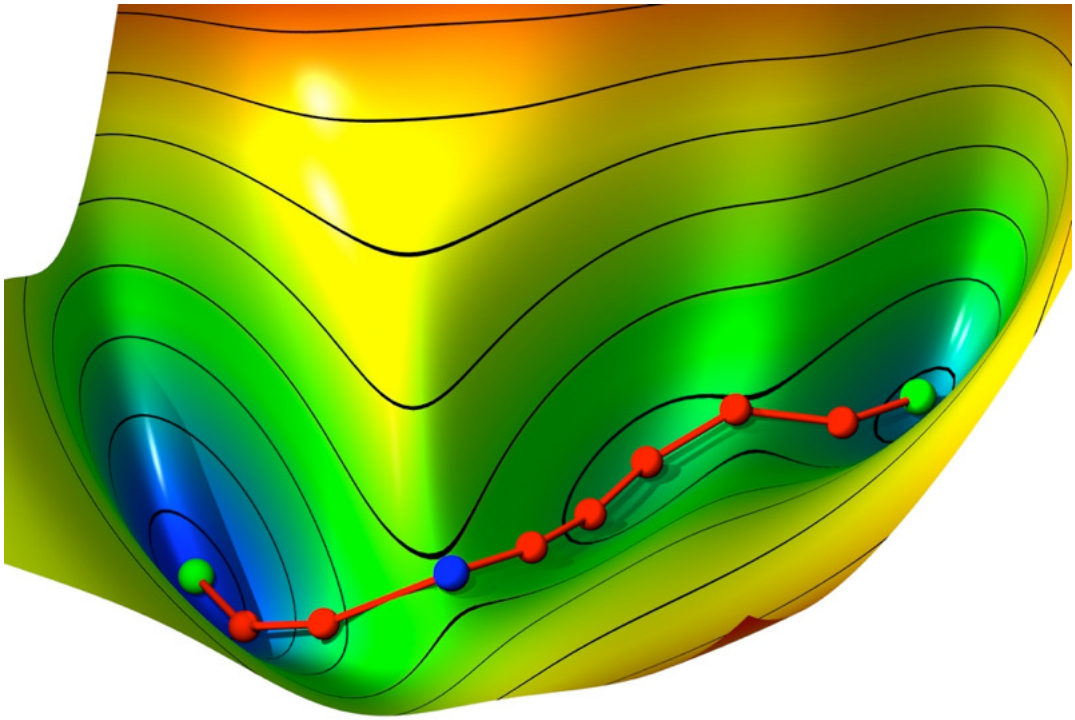


Figure 3.7: Single-point projections of idealised structures (red) between minima (green) and transition state (blue) structures [199].

system. This can be understood from the fact that the number of distinct internuclear separations for a molecule comprising N atoms is equal to $N(N-1)/2$ which for a molecule of $N > 7$ is greater than the $3N$ Cartesian degrees of freedom for the system. Therefore the transit path must instead be defined through the use of geometries with internuclear separations as close as possible to the idealised values, found by minimising the following function [200]:

$$S(f) = \frac{1}{2}\sum_{\mathrm{a}\neq\mathrm{b}} \frac{(r_{\mathrm{ab}} - r^i_{\mathrm{ab}}(f))^2}{r^i_{\mathrm{ab}}(f)^4} + 10^{-6} \sum_{\zeta=x,y,z} \sum_{a} (\zeta_a - \zeta^i_a(f))^2 \geq 0 \tag{3.81}$$

where the interpolated Cartesian position of an atom is represented by ζ_a^i and the actual coordinate is ζ_a. By construction, equation (3.81) is identical to equation (3.80) at the reactant and product geometries, when the interpolation parameter (f) is equal to either zero or one, respectively. However, equation (3.80) can give multiple structures that satisfy the constraints on internuclear separations at, for example, the LST maximum. So by minimising S, for fixed f, using the coordinates from (3.80), one can solve for the new coordinates, which makes a better interpolation, to a first approximation.

The LST maximum estimate is further improved by minimising the geometry, with respect to the generalised reaction coordinate of the structure, defined by:

$$p = \frac{d_{\mathrm{R}}}{d_{\mathrm{R}} + d_{\mathrm{P}}} \tag{3.82}$$

where $d_{\mathrm{R(P)}}$ is equal to the distance between the reactant (product) and any other geometry of the molecule such that:

$$d_{\mathrm{R(P)}}^2 = \frac{1}{N} \sum_a (\zeta_a - \zeta_a^{\mathrm{R(P)}})^2 \tag{3.83}$$

The value of the reaction coordinate p from equation (3.82) runs from 0 at the reactant to 1 at the product state. So far, this transit path has been constructed purely on the basis of geometric analysis alone, without the use of energy calculations. The maximum energy structure along this pathway provides the first estimate of the TS structure. A conjugate gradient (CG) refinement is then performed on this maximum in directions conjugate to the reaction pathway, with the resultant structure used as an intermediate to define the quadratic synchronous transit (QST) pathway, defined by:

$$r_{\mathrm{ab}}^i(f) = (1-f) r_{\mathrm{ab}}^{\mathrm{R}} - f r_{\mathrm{ab}}^{\mathrm{P}} + \gamma f(1-f) \tag{3.84}$$

where γ ensures that the QST pathway includes this newly calculated intermediate structure. An example of LST/QST searching in practice is illustrated in Figure 3.8 where the activation barrier converges with calculated reaction coordinate from equation (3.82). Firstly, the single-point energies of the reactant and product state geometries are calculated, with a reaction coordinate of 0 and 1 respectively. Using these two energy points, the energy maximum along the LST path is located (black line: LST 1). A CG optimisation of this structure, in directions conjugate to the reaction pathway, is then performed (red triangles: CG 1). Naive application of the CG saddle-point algorithm discussed by Govind *et al.* will usually prove unsuccessful because the system will have a tendency to fall to a point on the energy surface below the saddle and repeated application of CG minimisations will optimise the system to one of the local minima in the vicinity of the transition state rather than the saddle point itself. A tendency of the optimisation process to veer away from the saddle towards a minimum will manifest itself as a build up of the gradient in the direction of negative curvature $\mathbf{s}_0$. In the practical scheme devised by Govind *et al.* the conjugate gradient process must be restarted with a new maximisation

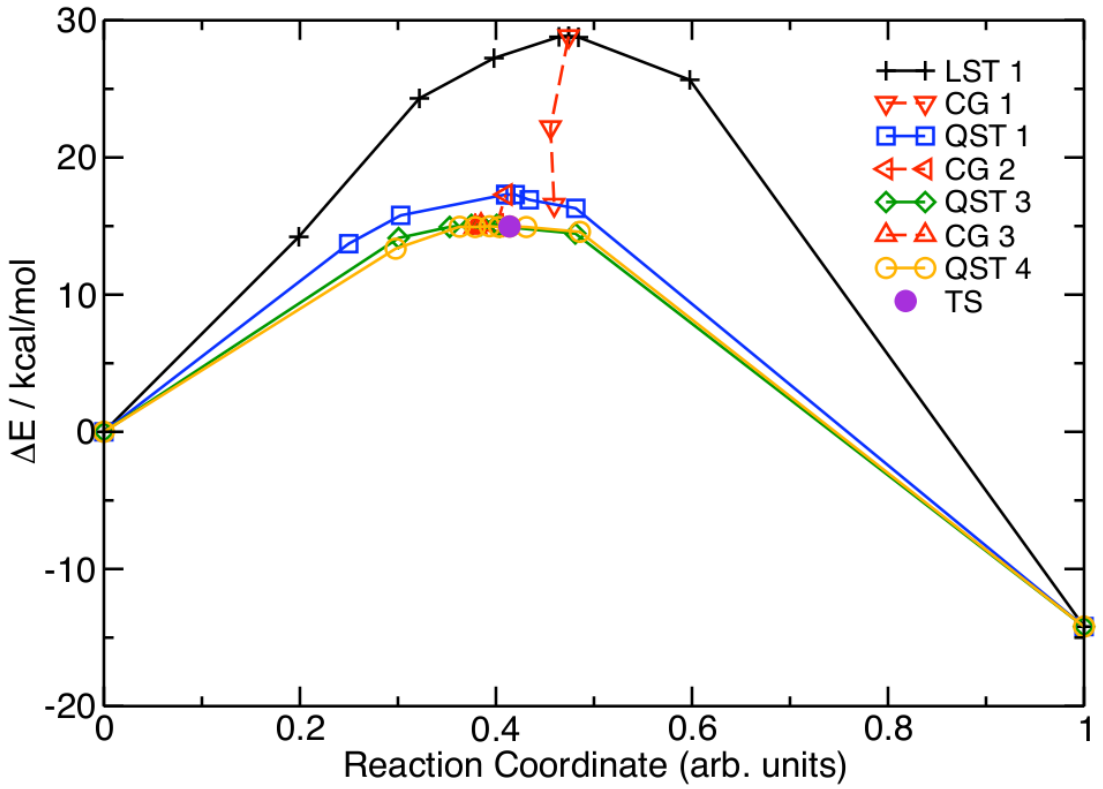


Figure 3.8: Linear and quadratic synchronous transit searching in practice, leading to the resultant transition state. The activation energy (ΔE) converges toward the reaction coordinate (p) of the transition state conformation.

step if the gradient in the direction $\mathbf{s}_0$ becomes too large. In this scheme the gradient in the direction of $\mathbf{s}_0$ is monitored, and the conjugate gradient process is terminated when this build up extends beyond some tolerance factor. A new maximum is then searched for along the QST pathway connecting the reactant, product and the best transition state structure (blue line: QST 1). A new CG cycle is then initiated (red triangles: CG 2). If the residual forces on the TS approximant fall below some user-specified tolerance, the calculation is considered to have converged. The LST/QST algorithm, based on the original inception by Halgren and Lipscomb [201] and modified by Govind *et al.* [200] has been successfully applied using implementations in the CASTEP code [202–205] and the DMol software package [206–217]. The LST and QST approaches discussed here are implemented in the ONETEP code and have been used throughout this dissertation to calculate activation energies.

3.5.5 The eigenvector-following approach

The principle of eigenvector-following lies in maximising the energy in one eigendirection whilst simultaneously minimising the energy in all other eigendirections. Recalling equation (3.79) we can write an arbitrary vector $\mathbf{y}$ as a linear combination of the Hessian

eigenvectors:

$$\mathbf{y} = \sum_i a_i \mathbf{e}_i \tag{3.85}$$

where a_i are scalar coefficients. We are now confronted with the fact that for large biomolecular systems comprising thousands of atoms it is undesirable, or sometimes unfeasible, to calculate the Hessian matrix of second derivatives. In this instance it is advisable to use a variational approach in order to find the smallest eigenvalue and its corresponding eigenvector. This particular eigenvector is that of the 'softest' mode, as higher-order modes can be found using eigenvector-following but these often correspond to transition states higher in energy than the true TS. Now if one considers taking a step in some arbitrary direction $\mathbf{y}$ the expected value for $\mathbf{y}$ can be defined as the Rayleigh-Ritz ratio [20]:

$$\lambda(\mathbf{y}) = \frac{\mathbf{y}^{\mathrm{T}}\mathbf{H}\mathbf{y}}{\mathbf{y}^2} \tag{3.86}$$

and by defining $\mathbf{y}$ in terms of the Hessian eigenvectors from equation (3.79) as in equation (3.85) and recalling that the eigenvalues of these mutually orthogonal eigenvectors are λ_i then a lower bound for $\lambda(\mathbf{y})$ can be found:

$$\lambda(\mathbf{y}) = \frac{\sum_i a_i^2 \lambda_i}{\sum_j a_j^2} = \frac{\sum_i a_i^2 (\lambda_i - \lambda_{\min})}{\sum_j a_j^2} + \lambda_{\min} \geq \lambda_{\min} \tag{3.87}$$

and differentiating with respect to a_α yields:

$$\frac{\partial \lambda}{\partial a_\alpha} = \frac{2a_\alpha}{\sum_j a_j^2} \left(\lambda_\alpha - \frac{\sum_i a_i^2 \lambda_i}{\sum_j a_j^2} \right) \tag{3.88}$$

There are nontrivial turning points that exist for $\lambda(\mathbf{y})$. In practice it has been found that minimising $\lambda(\mathbf{y})$ with respect to $\mathbf{y}$ ensures $\lambda(\mathbf{y})$ becomes the smallest eigenvalue of the Hessian whilst $\mathbf{y}$ becomes the corresponding eigenvector. In order to avoid the explicit calculation of the Hessian, the numerical second derivative of the energy is used as an approximation to $\lambda(\mathbf{y})$:

$$\lambda(\mathbf{y}) \approx \frac{V(\mathbf{X}_0 + \xi\mathbf{y}) + V(\mathbf{X}_0 - \xi\mathbf{y}) - 2V(\mathbf{X_0})}{(\xi\mathbf{y})^2} \tag{3.89}$$

where $\xi \ll 1$ and differentiating whilst keeping $|\mathbf{y}| = 1$ gives:

$$\frac{\partial \lambda}{\partial \mathbf{y}} = \frac{\nabla V(\mathbf{X}_0 + \xi\mathbf{y}) - \nabla V(\mathbf{X}_0 - \xi\mathbf{y})}{\xi} \tag{3.90}$$

Once the smallest eigenvalue and its corresponding eigenvector are known, an uphill step can be taken in the direction of the eigenvector in order to find the transition state. The magnitude of this step is derived in detail elsewhere [218]:

$$h = \frac{2F}{|\lambda_{\min}| \left(1 + \sqrt{1 + 4F^2/\lambda^2} \right)} \tag{3.91}$$

where F is the component of the gradient along the eigenvector $\mathbf{e}_{\min}$ corresponding to the smallest eigenvalue $\lambda_{\min}$. A hybrid eigenvector-following/minimisation approach is then used that combines this uphill step along $\mathbf{e}_{\min}$ with minimisation in all the orthogonal directions, generating the transition state when the gradient drops below a certain tolerance. In practice, TS conformations obtained from LST/QST calculations may be more accurately refined using the gradient-only version of hybrid eigenvector-following [20,219,220]. The procedure detailed here is repeated until a stationary point with a negative eigenvalue and a maximum magnitude of energy gradient below 0.01 eV/Å per atom is obtained. The hybrid eigenvector-following technique described here is implemented in the `OPTIM` code which has been interfaced with ONETEP. This approach has been used in this dissertation, where indicated, in order to compare activation energies to less rigorous, but computationally less expensive, LST/QST approaches.

3.6 Classical force fields

In order to accurately explore the conformational space of large, complex systems, within the limits of reasonable computational resources, a representative potential energy function, or force field, is required to approximate the atomic interactions. Force fields should be simple and easily differentiable, whilst preserving the characteristic features of the more accurate, yet time consuming, *ab initio* methods that are necessary for describing the electronic structure of the system. They are a common tool for studying macromolecules of biological interest. Amongst other things, force fields allow the structure-activity relationships of macromolecules to be studied in atomic detail. Whilst the quantum mechanical techniques outlined previously can be applied to systems up to tens of thousands of atoms, empirical approaches can be routinely applied to systems comprising hundreds of thousands of atoms. In fact, a total system size of 320 billion atoms has been studied using molecular dynamics, representing a cubic piece of metal with an edge length on the micrometer scale [221]. On a more practical length scale of calculations, the dynamics of systems can also be investigated, up to the nanosecond time regime and beyond. The essential job of any force field is to map the structure $\mathbf{R}$ onto the energy $U(\mathbf{R})$ of a system of interest. A force field is usually expressed as sums of two-, three- and, sometimes, four-body particle interactions. Some of the common and more important contributions to force fields are illustrated in Figure 3.9. Ideally, a minimal set of functions will be used to describe the molecular structure. A harmonic treatment is usually applied to the bonds, angles and out of plane distortions (improper dihedrals). The torsional and dihedral terms are described by a sinusoidal expression. On the non-bonded side, a Coulombic term is used to treat the electrostatics. This is usually combined with a Lennard-Jones term to describe the atomic repulsion and dispersion interaction. The form used in the

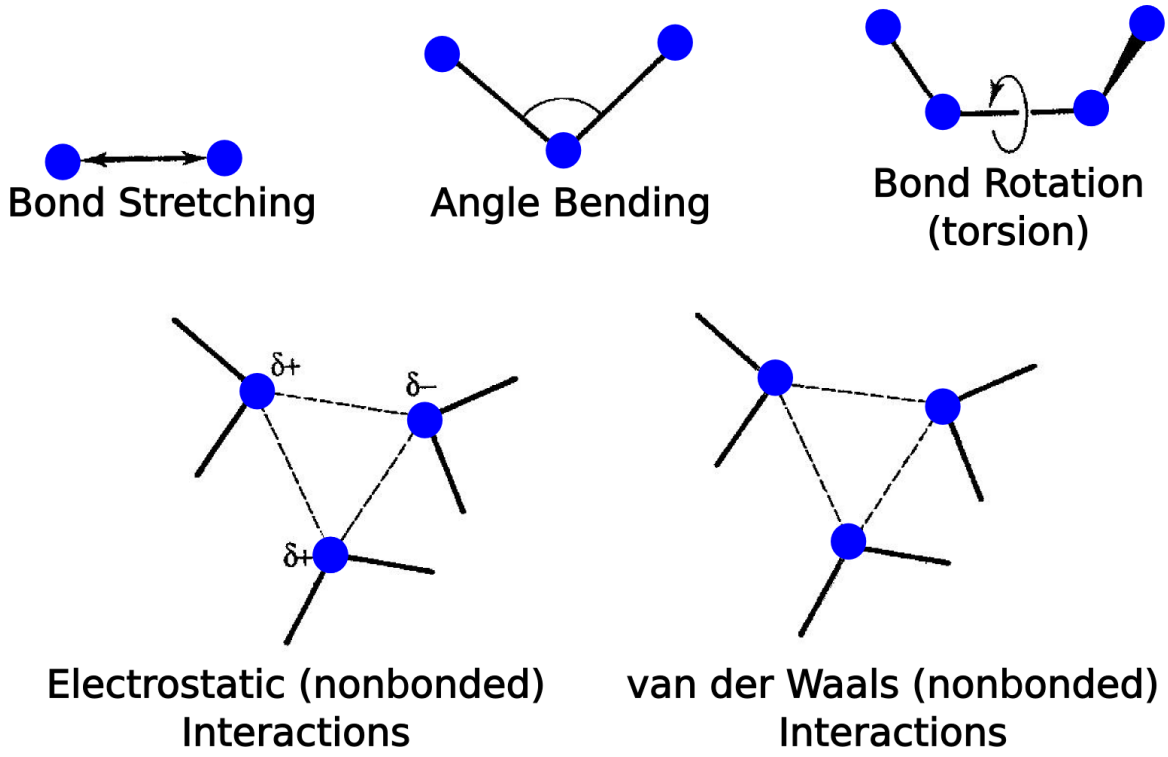


Figure 3.9: Key contributions to biomolecular force fields.

most commonly found biomolecular force fields is as follows:

$$U(\mathbf{R}) = \sum_{\text{bonds b}} K_{\text{b}}(r_{\text{b}} - r0)^2 + \sum_{\text{angles a}} K_{\theta}(\theta_{\text{a}} - \theta_0)^2 + \sum_{\text{dihedrals d}} K_{\chi}\left(1 + cos(n\chi_{\text{d}} - \sigma)\right)$$
$$+ \sum_{\text{impropers } \eta} K_{\eta}(\phi_{\eta} - \phi_0)^2 + \sum_{\text{nonbonded pairs ij}} \left(\left[\frac{C_{ij}^{(12)}}{r_{ij}^{12}} - \frac{C_{ij}^{(6)}}{r_{ij}^{6}} \right] \right) + \sum_{i<j} \frac{q_i q_j}{r_{ij}} \quad (3.92)$$

where r_{b} is the bond length, θ_{a} is the valence angle, χ_{d} is the dihedral angle, ϕ_{η} is the improper angle and r_{ij} is the separation between nonbonded atom pairs i and j. The intramolecular parameters are parameterised against experimental and *ab initio* observables. The associated equilibrium values are denoted by a 0 subscript. K_b and K_θ are the force constants for bond length and valence angle, respectively. The dihedral force constant, multiplicity and phase angle are represented by K_χ, n and σ respectively. The improper force constant and improper equilibrium angle are represented by K_η and θ_η. The $C_{ij}^{(6)}$ and $C_{ij}^{(12)}$ are the van der Waals terms and the final summation treats the electrostatic interactions.

Classical force fields have been established as an invaluable tool for the investigation of many systems of biological relevance. Extensive development of force fields and the parameterisation thereof has provided a vast toolkit containing a multitude of classical approaches to treat the common amino acids found in enzymes. The generic terms presented in equation (3.92) are commonly used within many biomolecular force fields, including CHARMM [222], AMBER [223,224], GROMOS [225] and OPLS [226]. Within which,

the canonical ensemble (NVT) and the isothermal-isobaric ensemble (NPT) are two key approaches that are used when simulating biomolecular systems. However, all protein force fields have a significant drawback in that the calculated results can sometimes depend strongly on the choice of parameters used for interatomic potentials. Force fields can not accurately describe electron bond cleavage and formation, electronic states and by association, spectroscopy. In addition, the inclusion of transition metals or unusual ligands or functional groups that are difficult to correctly parameterise within force fields are also more accurately treated within a DFT framework. The atomic partial charges used within the majority of standard force fields are often fit to the electrostatic potentials of small molecules calculated via expensive quantum chemical calculations. Whilst this process generates accurate partial charges for small molecules such as amino acids in the gas-phase, using these values for an entire protein will neglect long-range electronic polarisation and give only an average picture. A recent DFT investigation of an entire protein in water demonstrated that net charges of residues can vary by up to 0.5e from their putative integer values [170], while the electrostatic potential generated by force field charges may differ significantly from accurate *ab initio* simulations [183, 227]. In addition, force fields can not include charge transfer effects, so they have issues with transferability and accuracy. A number of validation studies have gathered an overall consensus that, whilst fixed charge models offer computationally tractable descriptions and are robust enough for calculating the equilibrium properties of homogeneous systems, away from these ambient conditions significant discrepancies between experiment and simulation have been seen [228, 229]. These discrepancies caused through the use of the fixed charge approximation have an effect on dynamical properties and heterogeneous systems in general. The problems outlined here can be partially overcome by implementing a mixture of QM within classical force fields, in hybrid QM/MM approaches, which are discussed in the next section.

3.7 Hybrid quantum mechanics/molecular mechanics approaches

The extensive conformational sampling which is required to accurately treat the entirety of a biomolecule is unfeasible with conventional quantum mechanical (QM) approaches due to computational demands. Due to the problems with force fields, described in the previous section, combined with the fact that many enzyme-catalysed reactions involve the breaking and forming of electron bonds, methods using a combination of classical molecular mechanics (MM) force fields and QM methods are often implemented. In these so-called QM/MM approaches [13, 230–232], a small part of a system, usually a region of chemical interest in which important changes such as covalent bond breaking and forming take place, is treated with QM in order to describe the electronic structure of specific fragments. The remainder of the system is then treated with a comparatively simple

empirical potential in order to describe the protein and water environment surrounding the QM region. The level of QM theory used must be balanced against computational expense which results in less rigorous methods often being used in comparison to conventional quantum chemistry calculations on small molecules. Particular applications that have proved sturdy test-beds for QM/MM approaches include citrate synthase [233], $P450_{cam}$ [234] and lysozyme [13, 235, 236], while many more are reported in the literature [237].

There are two predominant schools of thought concerning how to approach QM/MM methods. The additive approach describes the total energy as:

$$E_{\text{tot}} = E_{\text{QM}} + E_{\text{MM}} + E_{\text{QM-MM}} \tag{3.93}$$

where E_{QM} is the energy of the QM region calculated with the QM method. The energy of the MM region calculated with the force field approach is represented by E_{MM}. The interaction energy between the two regions is represented by $E_{\text{QM-MM}}$. Whereas in the subtractive approach the total energy is described as:

$$E_{\text{tot}} = E_{\text{MM-tot}} + E_{\text{QM}} - E_{\text{MM-QM}} \tag{3.94}$$

where $E_{\text{MM-tot}}$ and $E_{\text{MM-QM}}$ are the energies of the total and QM systems, respectively, calculated by the MM approach. In studies of enzyme catalysis it is mainly the additive method which is used, treating the substrate undergoing a reaction with a QM approach and the surrounding protein and solvent with MM. The use of atomistic simulations to model enzyme-catalysed reactions, starting from the pioneering works of Warshel, Levitt and others [13, 44] has risen to prominence in recent years, leading to the new field of computational enzymology [45, 46].

Due to the computational expense of conventional QM approaches, the QM region in a QM/MM simulation is often restricted to include only the reactive groups, an approach rationalised mainly by chemical intuition. The validity of this approach and how the QM region size impacts calculated properties has been investigated in a handful of systems but no definitive conclusions have been reached. A QM/MM study of the dependence of the central QM atom force error on the size of the QM region demonstrated that convergence was only achieved at a size of between 300 and 500 atoms, depending on how the region is chosen, when describing the QM system using PM3 and AM1 semiempirical methods [238]. Similarly, a QM/MM study of a twin arginine pair in adenovirus Ad11, implementing the HF/SVP level of theory for the QM atoms, demonstrated a QM region of 437 atoms was required to generate converged isomerisation energies [239]. Further, to achieve spectral convergence of QM/MM excited states of photoactive yellow protein, a QM region of 723 atoms was required in which the QM atoms were described at the TD-ωPBE/6-31G(d) level [240]. Results such as these suggest that inconsistencies in the treatment of long-range electrostatic interactions between QM and MM regions [238] may occur, which can critically affect results due to the polar nature of many enzyme active sites.

QM/MM can also suffer from inaccuracies due to the coupling scheme used to link the two regions, for which numerous approaches exist. The results from simulations can depend strongly on this choice and it may not always be straightforward to converge the results or test the methodology applied. In explicit solution models or enzyme systems whereby the substrate and any co-factors are not covalently bound to the enzyme, this partitioning does not prove to be a problem. However, in the case of a QM region describing a ligand covalently bound to a protein, it would be unrealistic to simply truncate the QM region, therefore treating the bond as being homolytically or heterolytically cleaved. A QM/MM boundary that slices through a covalent bond will first need to address the dangling bonds of any cleaved QM atoms, followed by ensuring the QM region is not over-polarised by the neighbouring MM charges. In addition, the bonded MM terms involving atoms from each subsystem must be selected as to avoid double-counting. There are three broad ways in which this can be achieved. The first involves link-atom schemes [241, 242] that introduce an additional atomic centre into the QM region, typically a hydrogen atom, that is not part of the real system but acts to saturate any dangling bonds. Secondly, boundary-atom schemes [243] replace the MM atom at the border between regions with a boundary atom that then appears in each region. In the QM region it mimics the severed bond and the MM residue bonded to it, whilst the MM region 'sees' a normal MM atom. Lastly, localised-orbital schemes [13] place hybrid orbitials at the boundary, keeping some of them frozen, in order to cap the QM region.

Another potential problem encountered by QM/MM calculations is that of electron leakage. This is the phenomenon whereby positively charged classical point charges can act as traps for the electron density from the QM region [231, 244–250]. This problem emerges due to a lack of Pauli repulsion from the electron cloud that would normally surround a positively charged atom, resulting in over-polarisation of the electron density at short range by an incorrect and purely attractive potential. A further problem with QM/MM approaches is the error associated with the choice of force field. It has been shown that different force fields, or even different versions of the same force field, can produce qualitatively different results [251–253]. One systematic study of the folding behaviour of non-amyloid peptides indicated that AMBER99 favours helical structure, GROMOS96 [254] may overestimate β conformations but OPLS-AA [255] results in balanced α and β structures [253]. It has also been shown that nonpolarisable force fields have a tendency to overestimate the coordination number and rigidity of the solvation shell of typical cations and ions [256, 257]. However, advances are being made in the development of polarisable approaches [258], such as the AMOEBA force field [259], where the fixed partial charge model is improved upon by the use of atomic multipole-based electrostatics and explicit treatment of dipole polaristation. These approaches have not been used in QM/MM studies due to a lack of accurate paramaterisation and the need for a framework to interface polarisable force fields with QM calculations. In general, QM/MM investigations of enzymatic reactions require a significant investment in the setup and preparation

of the system in question before the calculations can even begin [231]. Any errors introduced at this stage through unsuitable choices will propagate throughout the study and cannot be recovered at a later phase of the calculations [232].

In order for computational enzymology to have impact in other communities it is important to make simulation methods and their associated results accessible to non-specialists [237]. This is a central theme of this dissertation and is evidenced by the removal of the additional complexity of the selection of force field parameters and QM/MM boundary partition scheme in later chapters. An alternative to the hybrid schemes outlined in this section, is to perform QM calculations on an entire system, or a significant part of it. Unbiased *ab initio* calculations of peptides, harnessing the power of GPUs, have been shown to be able to predict protein structure at the same level as empirical force fields that have been extensively paramaterised specifically for that purpose [260]. The same calculations demonstrated that *ab initio* approaches can predict the structures of proteins with regions of disorder to a much greater accuracy than that of force fields. Other investigations have also shown that DDEC charges generated from large-scale DFT calculations perform better than standard AMBER ff99SB atomic partial charges in replicating protein dynamics when incorporated within classical force fields [183]. Using the linear-scaling first principles approach described in Section 3.4, explicit treatment of the electrons can be performed, taking into account the electronic charge transfer and polarisation, for systems containing many thousands of atoms and the methods are transferable to any chemical environment. These types of investigations are the subjects of Chapters 5 and 6.

3.8 Summary

This chapter has outlined the capabilities of the various tools used within this dissertation and the approaches taken when simulating systems of biological interest. It is the combination of the techniques described in this chapter that will allow the study of the properties of complex biomolecules such as enzymes, described in the previous chapter. Best practices and efficient methodologies for the general treatment of biomolecular systems using the methods outlined in this chapter, from classical to QM/MM to full-DFT, will be the focus of Chapter 5. Building upon these methodologies, an investigation into the *Bacillus subtilis* chorismate mutase enzyme will be discussed in Chapter 6. Before extending to much larger systems of biomolecular interest, the next chapter will validate the methods described in this chapter through investigations of well-studied small molecules with the goal of reproducing known results.

Chapter 4

Validation studies

"Aristotle maintained that women have fewer teeth than men; although he was twice married, it never occurred to him to verify this statement by examining his wives' mouths"

Bertrand Russell, *Impact of Science on Society - Ch. 1 (1952)*

There are few principles that should be prioritised over the process of validation. To ensure one can rely on the results that emerge from proof-of-principle calculations, such as those presented in Chapter 6, one must be able to reproduce, or closely approximate, results seen both in experiment and more advanced levels of theory and computation, whilst keeping in mind any technical issues present in the approaches. The calculations presented in this chapter are designed to provide confidence in the methods available in the ONETEP and `OPTIM` codes and the efficient interface that exists between the two for searching for stationary points on the potential energy surface. A primary method of interest within this dissertation is structural optimisation, comprising both energy minimisation and transition state searching. To ensure the reactants and products, along with the transition states that connect them, can be trusted when investigating enzyme catalysed reactions, the methods must first be validated against the structures of small molecules *in vacuo*. The simple ethene molecule has a well-defined structure and accurately observed normal mode frequencies. Therefore structural optimisation procedures and normal mode analysis methods will be used to attempt to reproduce what is seen in experiment. Ethene will be the focus of Section 4.1. The well-studied alanine dipeptide has proved to be a particularly good example for verifying new methods or extensions to existing approaches in order to validate the particular advances made. In addition, the structural properties this molecule possesses are very similar to those found in the backbone of many proteins. Therefore, if the structural properties of dialanine energy minima and transition states, *in vacuo*, can be accurately reproduced, this will provide confidence of the ability of the methods to tackle the protein structures found in real systems. This molecule is the subject of Section 4.2. The focus of the chapter then shifts to the pericyclic rearrangement of chorismate to prephenate in Section 4.3. In order to be able to consider the complexities involved in describing enzyme reactions, the structural optimisation procedures in ONETEP must be validated against a relevant small-scale reaction.

This work will not only validate the structural optimisation procedures in ONETEP but will also test the interface that exists between ONETEP and `OPTIM` in order to further refine the linear and quadratic synchronous transit (LST/QST) transition state candidates with rigorous hybrid eigenvector-following methods. The chapter concludes with a brief summary outlining the findings and their significance to the remainder of the dissertation.

4.1 Ethene

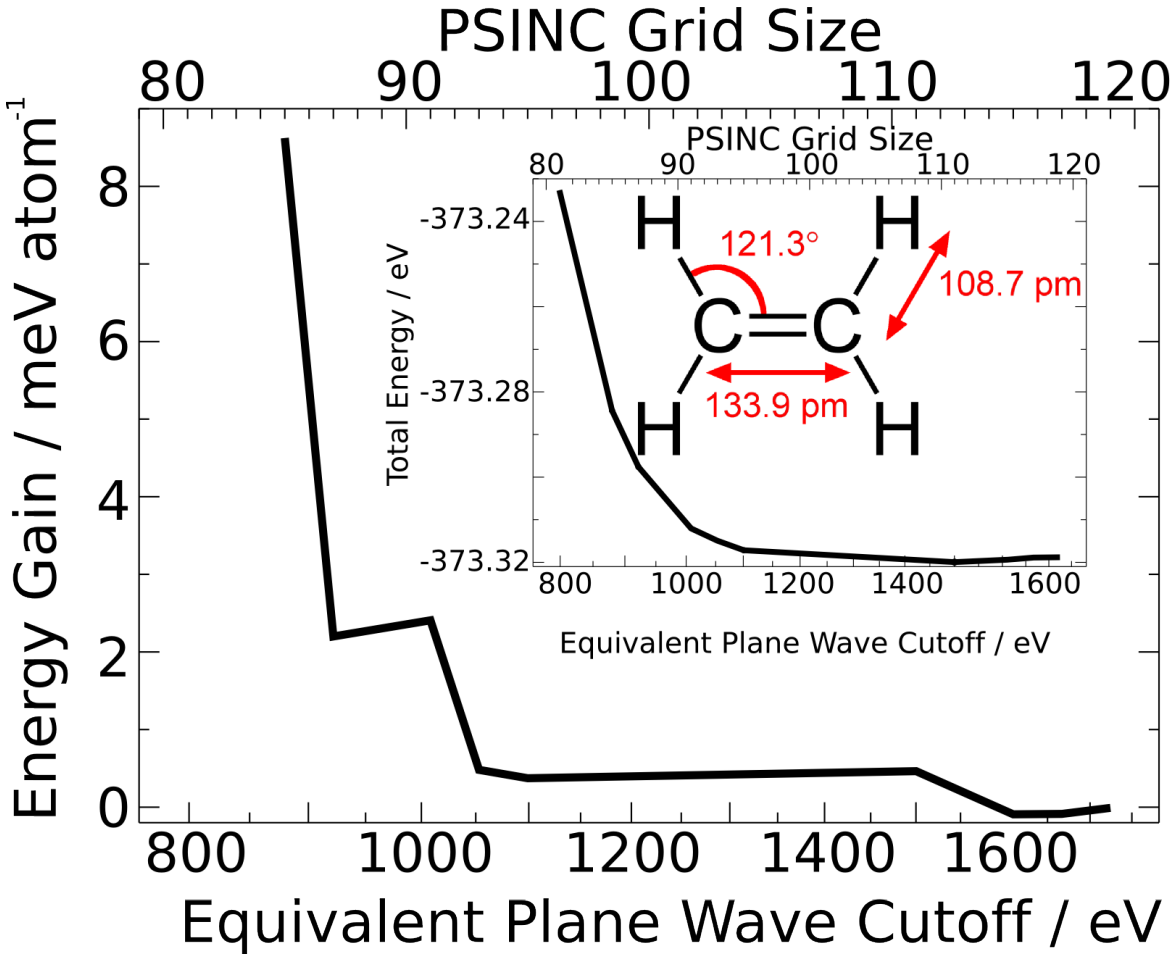


Figure 4.1: Convergence of ethene energy gain per atom. Inset: convergence of total energy with increasing equivalent plane wave cutoff energy and a skeletal representation of the ethene molecule with experimental measurements determined via microwave spectroscopy [19]. The spacing of the psinc grid was explicitly set in order to compare with the cutoff used in an equivalent plane wave calculation.

Ethene is a simple, unsaturated hydrocarbon with the chemical formula $H_2C{=}CH_2$. The molecule is widely used in industry and its production exceeds that of any other organic compound. It is also used within the agricultural sector to accelerate the ripening of fruit. A representation of the molecule can be seen in Figure 4.1. The main objective of

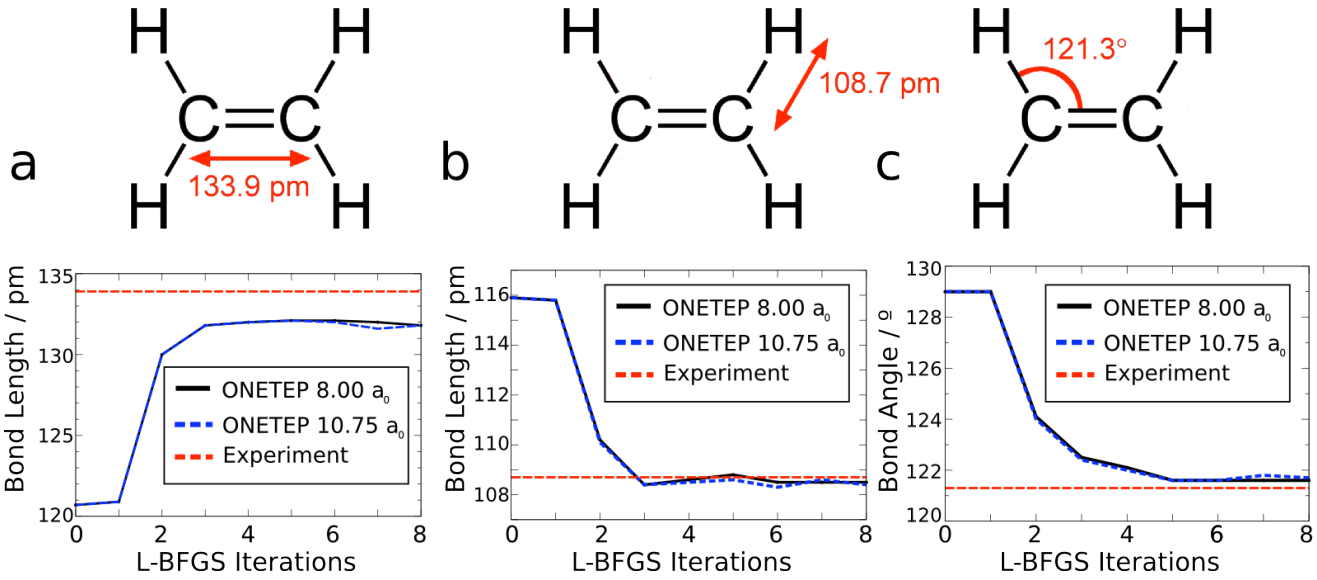


Figure 4.2: Convergence of ethene (a) carbon-carbon double-bond length, (b) carbon-hydrogen bond length and (c) carbon-hydrogen angle during ONETEP structural optimisation. The NGWFs used for the optimisation were localised in radii of $8a_0$ (black line) and $10a_0$ (blue line). A comparison is made with experimental values (red line) [19].

studying ethene is to validate the ONETEP/`OPTIM` interface for a small molecule. Preliminary convergence tests were performed, using a reasonable starting guess of coordinates, in order to demonstrate the convergence of calculated properties. Figure 4.1 shows total energy convergence with respect to equivalent cutoff energy in ONETEP. For increasing equivalent plane wave cutoff, the total energy for the molecule decreases, as is expected for variational methods. Therefore there is a gain in energy upon increasing the cutoff. This energy gain, per atom, is shown to converge to the meV level.

Figure 4.2 compares the calculated bond lengths and angles angles at each step through the optimisation procedure with those obtained from microwave spectroscopy [19]. It is the PBE functional [70] combined with dispersion corrections due to Elstner *et al.* [166] that is used throughout the DFT calculations presented in this chapter and the remainder of this dissertation. The figure shows that the calculations localising the NGWFs in a radius of 8.0 a_0 yielded converged bond lengths and angles in the ONETEP structural optimisation. Using this value with the converged equivalent plane wave cutoff energy of 1050 eV, as suggested from Figure 4.1, the geometry of the structure was then optimised using ONETEP until the maximum calculated force on any atom decreased below 0.01 eV/Å. The geometry optimisation yielded a C=C double-bond length 1.57% shorter than that observed via spectroscopy and a shorter C−H bond by 0.33%, along with a 0.28% wider H−C=C angle. The magnitude of these errors are typical of those found in density-functional studies [262]. A normal mode analysis was then performed on this structurally optimised conformation. By definition, for the zero-frequency modes, the geometry of

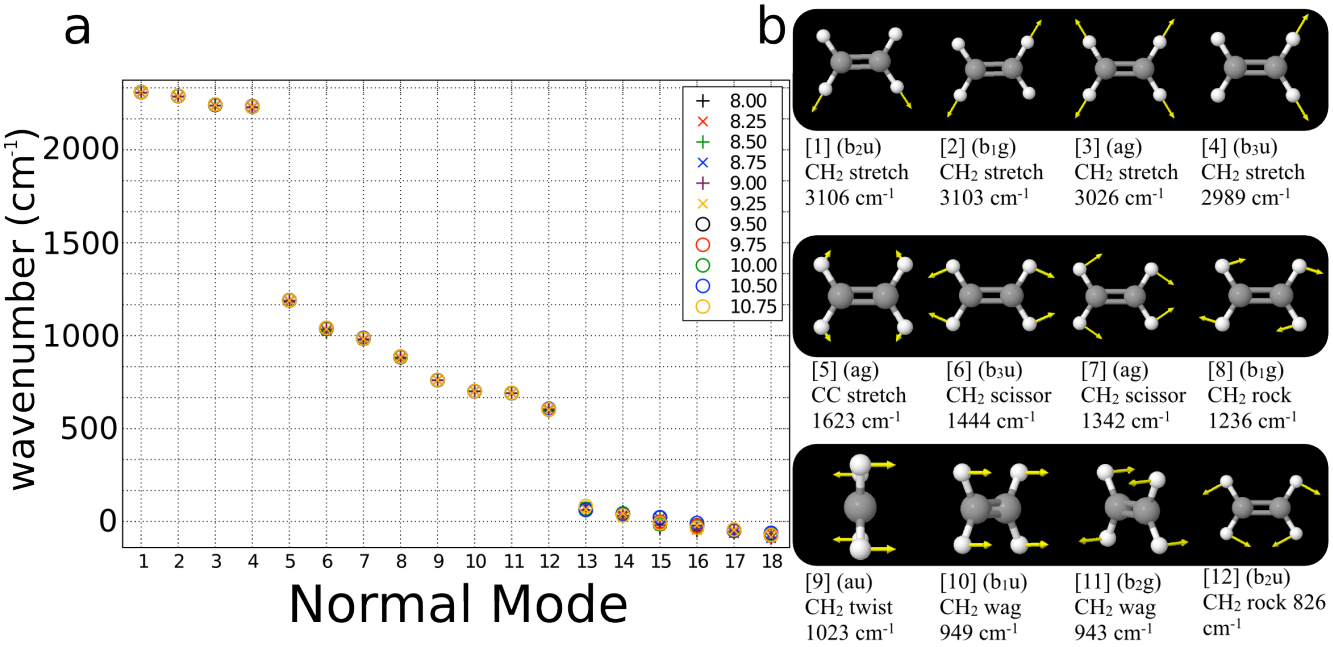


Figure 4.3: (a) Analysis of normal modes performed using the ONETEP/OPTIM interface (using NGWFs of radii ranging from $8.00a_0$ to $10.75a_0$) compared with (b) experimental observation. [261]

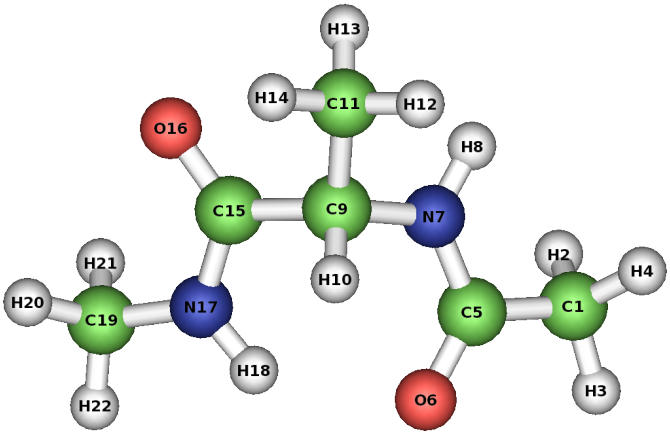
the molecule is not altered. Non-linear molecules have three zero-frequency rotational modes, hence $3N - 6$ normal modes. The calculations in Figure 4.3(a), corresponding to the normal modes illustrated in Figure 4.3(b), yield a systematic error with experiment of 26%. Whilst these sorts of errors are toward the upper bound of those expected within DFT implementing PBE functionals, previous authors have demonstrated underestimates in calculated frequencies using density-functional methods [263], and many DFT methods have been shown to yield inaccurate estimates for C−H frequencies [264, 265]. However, Figure 4.3(a) shows that qualitative agreement has been found between calculated and experimentally observed normal modes. Six 'zeroes' along with a band of eight finite frequencies followed by a band of four frequencies at approximately twice as large a frequency have been calculated. These match qualitatively with what is observed experimentally [261], shown in Figure 4.3(b). Previous authors have demonstrated that ONETEP is able to achieve plane wave accuracy [139, 266] upon increasing NGWF radii and fineness of the psinc grid. Temperature effects have not been considered but these could potentially be important. Normal mode analysis assumes harmonic potential wells and the harmonic frequencies have been calculated at 0K, whereas the experimental observations were performed at room temperature.

4.2 Alanine dipeptide

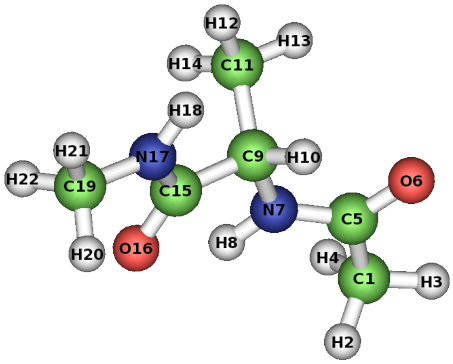
The alanine dipeptide, terminally blocked by methyl groups, is regarded as a prototype of non-proline/non-glycine protein residues [267]. This is due to the full ϕ/ψ sampling that dialanine allows, without the added complexity of the sidechain degrees of freedom. If the calculated potential energy surfaces for the vacuum and the solvated dialanine conformations are combined, it can be shown that the molecule is able to adopt every (ϕ, ψ) dihedral angle combination observed for α-helix and β-sheet structures within protein complexes [268, 269]. The dialanine molecule is deceptively simple but the effective potential energy surface of the dipeptide displays of the order of five minima. However, this precise number depends on the method of solvation (or lack thereof) and the particular form of the effective potential model used [270]. Dialanine is a very good candidate for validating any previously untested computational approaches, as many studies exist in the literature, starting from Peter Rossky and Martin Karplus' pioneering 1979 publication [271], that investigated the molecule from the perspective of kinetics [270, 272], thermodynamics [273–277] and spectroscopy [278–282]. Dialanine along with the alanine tri- or tetrapeptides [283] are often simulated at the *ab initio* level in order to parameterise the amino acid backbone force fields used within molecular mechanics calculations [284, 285]. It is normally the case in standard empirical molecular mechanics force fields that bonded and non-bonded terms derived from dialanine calculations will be used to model the backbone of all non-glycine and non-proline residues [286]. However, there are a few particular instances where this is not the case, notably the AMBER force field where the backbone atomic partial charges may vary for each residue [287].

Conformational changes from one dialanine energy minimum to another involve rotation about the backbone dihedral angles, passing through transition state structures. At least four such transitions occur in a force field description of dialanine. As such, this small molecule provides an extremely good test for transition state searching techniques. The molecule in the gas phase has a conformational space with well-defined minima and transition states that are present without the need for surrounding solvent or protein. Therefore, the inclusion of additional protein scaffolds such as enzyme active sites are not required in the calculations, making them inexpensive and a good starting point for further investigations involving enzyme-catalysed transition state searching on substrates embedded within a large protein matrix such as the calculations presented in Chapter 6 of this dissertation. In principle, when repeated over multiple residues, β structures, with a ϕ range from -60° to -170° and a ψ range from 120° to 170°, correspond to extended β-sheet secondary structures. $C7_{ax}$ structures, with a ϕ of around 50° and ψ of around -130°, are associated with the formation of turns and loops within protein secondary structure.

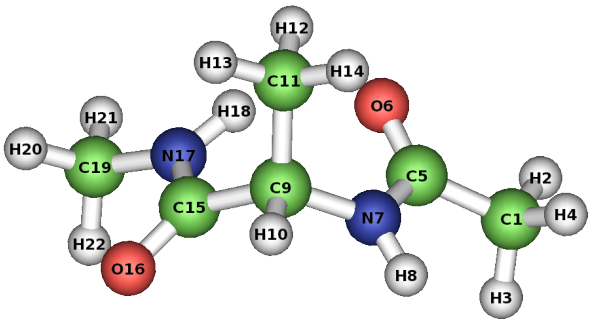
Previous authors have calculated minima and transition states of dialanine using the CHARMM22 force field in vacuum (C22VAC) [251] and the resulting structures are shown in Figure 4.4. The minima and transition states from Ref. [251] are simple geometrically defined objects, which are temperature independent and the calculated harmonic



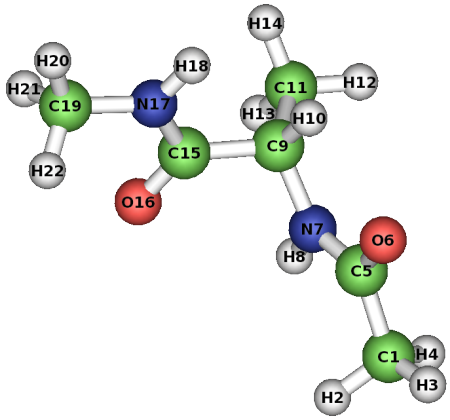
(a) $C7_{eq}$



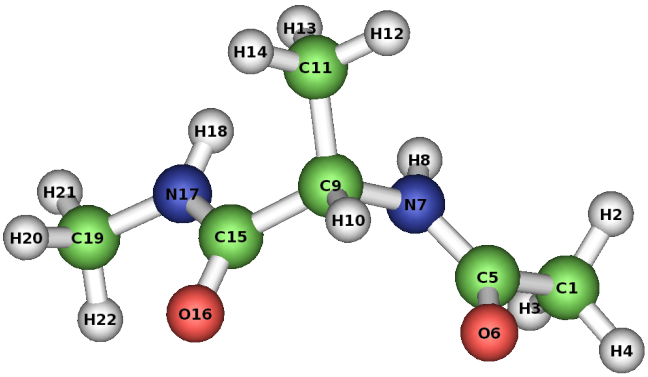
(b) β



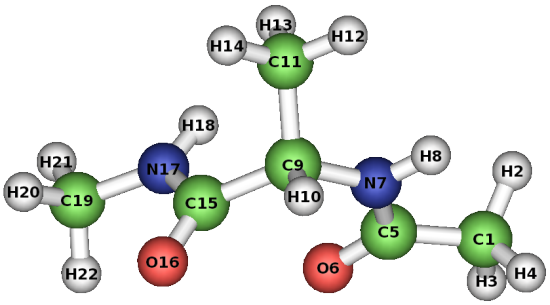
(c) C7ax



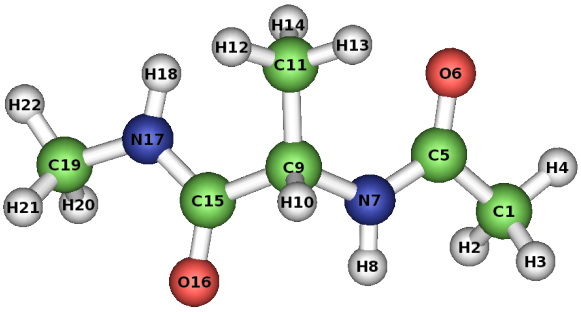
(d) $(C7_{eq} \leftrightarrow \beta)^{\ddagger}_1$



(e) $(C7_{eq} \leftrightarrow \beta)^{\ddagger}_2$



(f) $(C7_{eq} \leftrightarrow C7_{ax})^{\ddagger}$



(g) $(\beta \leftrightarrow C7_{ax})^{\ddagger}$

Figure 4.4: Dialanine TS minima (a-c) and transition state (d-g) conformations yielded from the *in vacuo* C22VAC potental in Ref. [251]

	$C7_{eq}$		β		$C7_{ax}$	
	C22VAC [251]	ONETEP	C22VAC [251]	ONETEP	C22VAC [251]	ONETEP
ϕ	-81.4	-84.3	-151.4	-157.6	69.7	73.6
ψ	70.5	71.3	170.6	160.3	-67.6	-54.3
C1-C5=O6	121.3	122.7	121.2	123.0	120.6	121.9
C1-C5-N7	116.6	115.2	116.4	114.7	115.9	114.6
O6=C5-N7	122.0	122.1	122.4	122.3	123.5	123.5
C5-N7-H8	120.4	119.4	121.5	123.0	118.2	116.9
C5-N7-C9	123.3	123.1	122.8	122.1	125.9	127.1
C9-N7-H8	116.3	117.5	115.6	114.7	115.8	115.9
C9-C15-N17	116.8	113.0	117.6	115.4	117.9	115.6
C9-C15=O16	121.9	122.9	120.8	121.7	120.6	120.2
O16=C15-N17	121.3	124.0	121.5	122.8	121.4	124.3
H18-N17-C15	118.5	117.2	120.4	118.9	117.9	117.7
C15-N17-C19	122.4	121.5	121.6	122.0	122.6	121.5
C19-N17-H18	119.0	120.4	118.1	119.1	119.3	120.8
C1-H2	1.11	1.09	1.11	1.09	1.11	1.09
C1-H3	1.11	1.09	1.11	1.09	1.11	1.09
C1-H4	1.11	1.09	1.11	1.09	1.11	1.09
C11-H12	1.11	1.09	1.11	1.09	1.11	1.09
C11-H13	1.11	1.09	1.11	1.09	1.11	1.09
C11-H14	1.11	1.09	1.11	1.09	1.11	1.09
C9-H10	1.08	1.09	1.08	1.09	1.08	1.09
C19-H20	1.11	1.09	1.11	1.09	1.11	1.09
C19-H21	1.11	1.09	1.11	1.09	1.11	1.09
C19-H22	1.11	1.09	1.11	1.09	1.11	1.09
N7-H8	0.99	1.01	1.00	1.02	0.99	1.01
N17-H18	1.00	1.02	0.99	1.01	1.00	1.02
C1-C5	1.48	1.51	1.48	1.51	1.48	1.51
C9-C11	1.54	1.51	1.54	1.53	1.55	1.52
C9-C15	1.53	1.54	1.52	1.52	1.53	1.54
C5-N7	1.34	1.36	1.34	1.36	1.34	1.36
C9-N7	1.45	1.46	1.44	1.44	1.46	1.47
C15-N17	1.35	1.35	1.35	1.35	1.35	1.35
C19-N17	1.44	1.45	1.44	1.45	1.44	1.45
C5=O6	1.22	1.23	1.22	1.23	1.22	1.23
C15=O16	1.23	1.23	1.23	1.23	1.23	1.23

Table 4.1: Dialanine geometrical parameters from C22VAC and ONETEP minimum energy bond lengths in Ångströms, indicated by N1-N2 labels, and bond angles in degrees (°), indicated by N1-N2-N3 labels. The dihedral angles ϕ and ψ are defined as those involving C5−N7−C9−C15 and N7−C9−C15−N17, respectively.

	$(C7_{eq} \leftrightarrow \beta)_1^{\ddagger}$	$(C7_{eq} \leftrightarrow \beta)^{\ddagger}$	$(C7_{eq} \leftrightarrow \beta)_2^{\ddagger}$	$(C7_{eq} \leftrightarrow C7_{ax})^{\ddagger}$		$(\beta \leftrightarrow C7_{ax})^{\ddagger}$	
	C22VAC [251]	ONETEP	C22VAC [251]	C22VAC [251]	ONETEP	C22VAC [251]	ONETEP
ϕ	-104.8	-124.1	-99.3	-1.1	71.0	125.7	71.8
ψ	139.8	109.8	-72.3	-69.8	-48.2	-120.1	-56.9
C1-C5=O6	121.2	122.1	120.9	120.1	121.9	119.7	121.9
C1-C5-N7	116.4	115.4	116.2	115.4	114.7	114.9	114.6
O6=C5-N7	122.45	122.50	122.9	124.6	123.4	125.4	123.5
C5-N7-H8	120.4	120.7	119.2	116.0	116.6	116.8	117.0
C5-N7-C9	123.2	121.1	123.3	130.4	127.6	130.0	127.3
C9-N7-H8	116.1	117.8	117.5	113.6	115.6	113.2	115.8
C9-C15-N17	117.4	114.7	117.7	118.3	116.2	118.4	115.3
C9-C15=O16	121.2	122.0	121.5	120.5	119.0	120.3	120.0
O16=C15-N17	121.4	123.3	120.8	121.2	124.8	121.3	124.6
H18-N17-C15	120.4	118.5	120.4	118.5	117.4	119.5	117.4
C15-N17-C19	121.5	122.3	118.2	119.5	121.9	118.5	121.9
C19-N17-H18	118.0	119.2	121.4	121.7	120.6	121.9	120.7
C1-H2	1.11	1.09	1.11	1.11	1.09	1.11	1.09
C1-H3	1.11	1.09	1.11	1.11	1.09	1.11	1.09
C1-H4	1.11	1.09	1.11	1.11	1.09	1.11	1.09
C9-H10	1.08	1.09	1.08	1.08	1.09	1.08	1.09
C11-H12	1.11	1.09	1.11	1.11	1.09	1.11	1.09
C11-H13	1.11	1.09	1.11	1.11	1.09	1.11	1.09
C11-H14	1.11	1.09	1.11	1.11	1.09	1.11	1.09
C19-H20	1.11	1.09	1.11	1.11	1.09	1.11	1.09
C19-H21	1.11	1.09	1.11	1.11	1.09	1.11	1.09
C19-H22	1.11	1.09	1.11	1.11	1.09	1.11	1.09
N7-H8	0.99	1.01	0.99	0.99	1.01	0.99	1.01
N17-H18	0.99	1.01	0.99	0.99	1.02	0.99	1.02
C1-C5	1.48	1.50	1.48	1.48	1.51	1.48	1.51
C9-C11	1.54	1.53	1.54	1.54	1.52	1.55	1.52
C9-C15	1.52	1.52	1.53	1.53	1.54	1.52	1.54
C5-N7	1.34	1.36	1.34	1.34	1.36	1.34	1.36
C9-N7	1.45	1.46	1.45	1.46	1.47	1.46	1.47
C15-N17	1.35	1.36	1.35	1.35	1.35	1.35	1.35
C19-N17	1.44	1.45	1.44	1.44	1.45	1.44	1.45
C5=O6	1.22	1.23	1.22	1.22	1.23	1.22	1.23
C15=O16	1.23	1.23	1.23	1.23	1.23	1.23	1.23

Table 4.2: Dialanine geometrical parameters from C22VAC and ONETEP transition state structures in Ångströms, indicated by N1-N2 labels, and degrees (°), indicated by N1-N2-N3 labels. The dihedral angles ϕ and ψ are defined as those involving C5-N7-C9-C15 and N7-C9-C15-N17 respectively.

frequencies are temperature independent and correspond to normal mode analysis for the stationary points. Therefore the geometries can be directly compared with ONETEP geometry optimised minima and LST/QST TS structures. The investigation in Ref. [251] calculated the dialanine free energy surfaces using a superposition of partition functions based upon harmonic densities of states sampled at local energy minima and transition states. The relative free energies of the stationary points, which are calculated from the harmonic vibrational densities of states at the relevant temperature, are quoted in Ref. [251] at room temperature. As such, the resulting ΔF values represent the harmonic free energy at 298K, relative to the $C7_{eq}$ minimum energy conformation. The investi-

Structural	ΔE / kcal mol^{-1}		GRC / arb. units	
Conformation	C22VAC	ONETEP	C22VAC	ONETEP
$C7_{eq}$	0.00	0.00	-	-
β	0.91	1.14	-	-
$C7_{ax}$	2.05	1.31	-	-
$(C7_{eq} \leftrightarrow \beta)^{\ddagger}$	-	5.22	-	0.39
$(C7_{eq} \leftrightarrow \beta)^{\ddagger}_1$	1.50	-	0.43	-
$(C7_{eq} \leftrightarrow \beta)^{\ddagger}_2$	4.88	-	0.42	-
$(\beta \leftrightarrow C7_{ax})^{\ddagger}$	7.78	1.73	0.74	0.84
$(C7_{eq} \leftrightarrow C7_{ax})^{\ddagger}$	8.48	1.58	0.53	0.83

Table 4.3: Comparison of energy differences (ΔE / kcal mol^{-1}) and generalised reaction coordinates (GRC / arb. units) [201] between C22VAC potential minima and transition state structures [251], recalculated to generate temperature-independent potential energies, and those in the present chapter. The PBE functional [70] with empirical dispersion corrections [165] have been used for the ONETEP calculations presented.

gation presented in this chapter has taken the minima and transition state structures from Ref. [251] and calculated the temperature independent potential energies, using the classical C22VAC potential, that can be compared directly with ONETEP activation and reaction energies. Such a comparison is made in Table 4.3.

Tables 4.1 and 4.2 compare the structural parameters of dialanine calculated from Ref. [251] and ONETEP simulations. First considering the minima structures in Table 4.1, the bond lengths and angles calculated using ONETEP structural optimisation is in good agreement with the classical potential, yielding differences of between 1 and 2%. In addition, similar differences are present in the bond lengths and angles calculated using the LST/QST TS searching approach for the transition state structures present in Table 4.2. The ONETEP-calculated torsion angles for the minima, presented in Table 4.1, display a difference of between 1 and 6 % with the classical potential, except for the ψ angle for $C7_{ax}$ which yields a discrepancy of 20% with the classical approach. However, the torsion (ϕ,ψ) angles, for the transition state structures, presented in 4.2 show significant discrepancies in the structures structurally optimised using ONETEP compared to the classical potential. This reflects on the lack of a rigorous treatment for dispersion interactions in DFT and the empirical corrections that have been applied in these calculations. Therefore it is unsurprising that the torsion angles from DFT calculations do not agreed entirely with those from a classical potential. However, another source of the discrepancies for the transition state structures are the errors that are present in the classical approach, such as the fact that the force field is unlikely to be parameterised for structures in their transition state, but instead in their energy minima.

Table 4.3 compares the relative enthalpies calculated using ONETEP structural optimi-

sation with the temperature-independent potential energies calculated using the C22VAC classical potential for the structures presented in Ref. [251]. The generalised reaction coordinate, as explained in Ref. [201], was calculated for the converged transition state structures, yielded from C22VAC and ONETEP simulations, using equations (3.82) and (3.83). Table 4.3 reveals that the energetic ordering of the C22VAC energy minima is reproduced in ONETEP calculations. It is fairly well agreed upon that the equatorial C7 conformation, $C7_{eq}$, is that of the global energy minimum of the *in vacuo* dialanine potential energy surface. This structure gained its name as a result of the seven-atom central ring structure present and the equatorial orientation of the alanine sidechain with respect to this structure [288]. Raman experiments, NMR and depolarised Rayleigh scattering observations also suggest this conformation is the most energetically favoured in both aqueous and non-aqueous solution [289]. It is stated in Ref. [251] that the β conformation is entropically favoured as it allows for more conformational flexibility than the $C7_{eq}$ conformation. However, the entropic contribution, included within the harmonic free energies calculated in Ref. [251] (ΔF values shown in Table 1 of the paper), is not significant enough to change this ordering. This is also shown to be the case when re-calculating the temperature-independent potential energy differences (ΔE), as has been done in this investigation. An important structural characteristic that ONETEP describes well is the hydrogen bond between the C5=O6 and N17−H18 groups of the two peptide links. As these hydrogen-bonding interactions are very important in the stability of protein structures in general, and also in the catalytic rate enhancement produced by enzyme active sites, this is a very encouraging result.

The TS structures presented in Table 4.2 were calculated using the LST/QST approach, using only the $C7_{eq}$, β and $C7_{ax}$ energy minima structurally optimised in ONETEP calculations. No additional information was required from the classical simulations in order to obtain the transition states. By the nature of the LST/QST approach, only one transition state can be found between two structures. However, with the TS searching performed in Ref. [251], two TS structures were found for the $C7_{eq} \leftrightarrow \beta$ transition. As a sanity check, the $(C7_{eq} \leftrightarrow \beta)_1^{\ddagger}$ and $(C7_{eq} \leftrightarrow \beta)_2^{\ddagger}$ transition state structures, located from C22VAC simulations, were used in separate QST calculations as an initial approximant for the mid-point between the energy minima calculated using ONETEP. In each instance, the final structure, generated by ONETEP QST simulations, adopted the identical conformation to when no initial guess for the mid-point was supplied, indicating the robust nature of the LST/QST TS searching approach.

The conformational space accesible to dialanine can be illustrated in two dimensions by plotting the ϕ angles against the ψ angles, in a so-called Ramachandran plot named after its inventor G. N. Ramachandran [290]. Such plots for dialanine described by a variety of classical force fields are presented in Figure 3 of Ref. [251]. On inspecting the Ramachandran plot for C22VAC vacuum simulations, it is clear that the $(C7_{eq} \leftrightarrow \beta)_1^{\ddagger}$ structure requires the least amount of atomic rearrangement in order to pass from the

initial to final energy minima via this transition state conformation. Therefore, it is the $(C7_{eq} \leftrightarrow \beta)_1^{\ddagger}$ transition state structure that is most likely to be found in each instance of LST/QST calculations, which provide a less rigorous, though computationally less expensive method compared to the reaction path Hamiltonian superposition approach used in Ref. [251]. In a similar manner to the discrepancies seen in the torsion angles, the ONETEP-calculated GRC values also display significant differences compared to the classical potential. Again this is likely to be due to the lack of parameterisation of the classical force field at a transition state. However, these differences then leads to the underestimate of the activation energy for the $(\beta \leftrightarrow C7_{ax})^{\ddagger}$ and $(C7_{eq} \leftrightarrow C7_{ax})^{\ddagger}$ TS structures from DFT compared to the classical potential.

Throughout this analysis it must be kept in mind that the results from ONETEP simulations are fully quantum mechanical and the results from C22VAC simulations are from a classical force field. It would appear that one needs to move to a higher level of QM theory in order to clarify the energetics of the small molecule around the transition state. From the current level of QM, the DFT and classical results are in good agreement close to the energy minima and, crucially, this is where protein backbones are. It is not always entirely clear whether there are perhaps spurious transition state structures which appear on the classical potential energy surface which will not appear on the equivalent quantum mechanical surface. In the Ramachandran plots in Figure 3 of Ref. [251] where there are either three or four minima present, depending on which version of the CHARMM force field is used and there are either four or five transition states present. In addition, depending on which type of implicit solvent model is used, within the same CHARMM version, there are five or six minima present and either six or eight transition states, depending on the solvation model used. The Ramachandran (ϕ,ψ) plots in vacuum also show quantitative differences for different versions of the CHARMM forcefield and qualitative differences when compared to the AMBER forcefield. In addition, the energetic ordering shown from classical methods may in fact be skewed by inaccuracies present in the force field. The main aim of this section was to validate the ONETEP LST/QST transition state searching capabilities against a well-studied molecule, dialanine, but in order to provide some clarity into the issue discussed in Ref. [251], where it is shown that different classical force fields result in some qualitative and quantitative differences in their associated potential energy surfaces, further investigations, that remain outside the scope of this dissertation, will need to be undertaken. This still remains an ongoing issue but the potential energy surfaces of small molecules is not the primary concern of this dissertation. The focus of this chapter shall now move to another small molecule, the understanding of which may have far-reaching consequences within enzymology-related calculations in general.

4.3 Pericyclic chorismate rearrangement

Before one can even begin to consider the complexities introduced by explicitly including the active-site residues and associated protein scaffold in an enzyme reaction, or the solution surrounding the substrate in the equivalent reaction in water, the structural optimisation procedures in ONETEP must be validated. The calculations presented here are intended to validate the geometry optimisation methods and the LST/QST transition state searching algorithm on the chorismate to prephenate rearrangement in vacuum. This

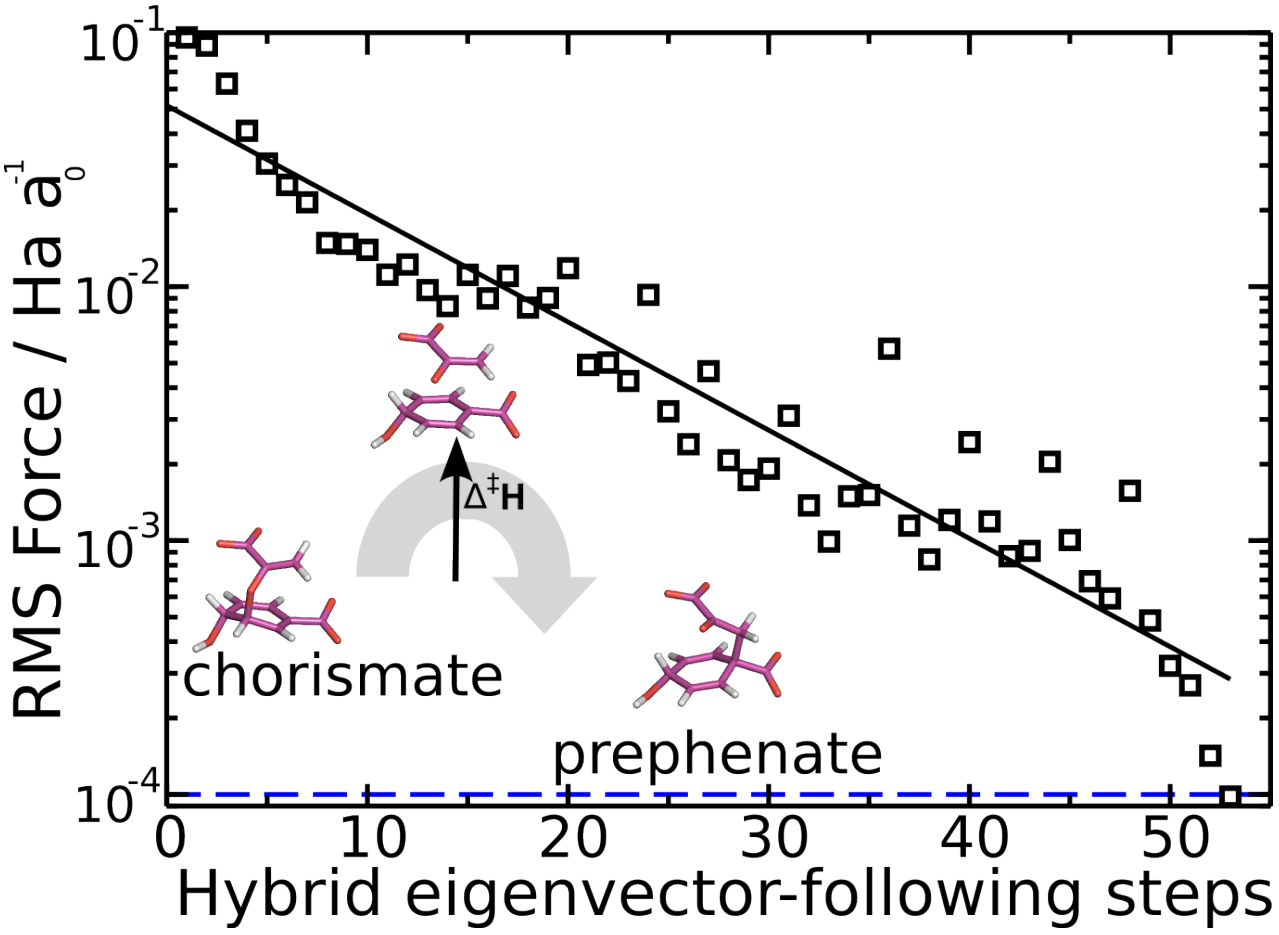


Figure 4.5: Decreasing calculated RMS force on the transition state structure during the eigenvector-following optimisation procedure. Inset: The pericylic Claisen rearrangement of chorismate to prephenate.

particular rearrangement takes place both in solution and in the presence of the *Bacillus subtilis* chorismate mutase (CM) enzyme but the focus of this section is on the reaction in the gas phase. The reaction is illustrated in Figure 4.5. Calculations by other authors on the substrate in the gas phase have found that many local minima are stable *in vacuo* but are subsequently unstable in the presence of the CM active site [291]. A much more detailed description of CM is given in Chapter 6. However, the purpose of this chapter is to determine the accuracy of the algorithms used in later investigations in this dissertation.

Therefore, rather than searching for global minima, the optimisation of the reactant and product state structures is initiated with conformations close to the configurations found within the CM enzyme. As discussed in the previous section detailing calculations on dialanine, ONETEP does not include entropic contributions to transition state activation energies. However, in the rearrangement of chorismate to prephenate it has been shown experimentally that the entropic contributions are orders of magnitude smaller than the enthalpic contributions. Therefore it is likely that the ONETEP calculations will yield an accurate description of the reaction.

The associated energy of reaction, yielded following geometry optimisation in ONETEP, is -17.0 kcal mol^{-1} and, following LST and QST searches for the transition state, the activation energy was determined to be 29.7 kcal mol^{-1}. The energy of reaction for the chorismate rearrangement was shown to converge with an energy cutoff of 1020 eV, corresponding to a psinc grid spacing of 0.45 a_0, with the NGWFs localised to a radius of 5.3 Å. Repeating the optimisation process with a 1687 eV cutoff, corresponding to a grid spacing of 0.35 a_0, and NGWFs with a 7.4 Å radius, showed the energy of reaction is converged to within 0.05 kcal mol^{-1} and the resultant activation energy was converged to within 0.1 kcal mol^{-1}. The transition state structure obtained from LST/QST calculations was then further refined using the gradient-only version of hybrid eigenvector-following [20, 219, 220] using the interface that exists between ONETEP and `OPTIM`, as discussed in Chapter 3. However, using the parameters shown to converge the reaction and activation energies (psinc grid spacing: 0.45 a_0, NGWF radii: 5.3 Å), a geometry with a single negative Hessian eigenvalue could not be found. This is due to the 'eggbox' effect [136–138] which, as discussed in Chapter 3, is important to avoid and the effect is detrimental for the gradient-only variant of hybrid eigenvector-following used in the ONETEP/`OPTIM` calculations. The activation energy calculated for the reaction varied significantly upon translating the atomic coordinates by 0.2 a_0, a fraction of the psinc grid size in use, despite previous successful convergence testing for the size of NGWF radii and fineness of the psinc grid. However, upon using an energy cutoff of 1687 eV, corresponding to a psinc grid spacing of 0.35 a_0, with the NGWFs localised to a 7.4 Å radius, the activation energy changed by only 0.01 kcal mol^{-1} upon translating the atomic coordinates by 0.2 a_0.

Following convergence testing, the higher cutoff and NGWF radii values were used in order to converge the eigenvalue and minimise the RMS force on the substrate. The forces generated by using the more rigorous parameters are a much closer representation of the true derivatives of the total energy. The RMS force on the transition state is shown to decrease during the eigenvector-following procedure in Figure 4.5. This additional eigenvector-following optimisation altered the activation barrier by less than 0.1 kcal mol^{-1}. The resultant structure following eigenvector-following yielded an RMSD value of less than 0.01 Å compared to the structure found using the LST/QST approach. Therefore, this work on the vacuum structure confirms that the LST/QST approach gives

essentially the identical transition state to eigenvector-following. Furthermore, it is clear that this computationally less expensive method can be used with confidence on the large systems tackled in Chapter 6. It has previously been found that the LST/QST approach can provide reasonable starting points for accurate refinement of transition states in small molecule systems [202–205] and this has been confirmed in the TS searching on the chorismate substrate in this chapter. What has not been shown in the literature is whether the TS approximant found from the LST/QST approach can provide a converged activation energy. This will be the focus of Chapter 6 The accuracy achieved by LST/QST for the current problem probably reflects the simplicity of the pathway. The LST/QST and eigenvector-following converged transition state was additionally characterised via normal mode analysis in `OPTIM` and found to have one imaginary frequency. The gas phase minima were found to have only real frequencies.

4.4 Summary

The task of describing complex chemical processes, such as those involved in enzyme catalysis, at the same high level of accuracy as rigorous quantum chemical calculations on small molecules in the gas phase is a highly active area of research. A first step toward this goal is the validation of the theoretical methods developed by applying them to smaller systems that approximate the properties of interest of the target real systems. In addition, the delicate balance of quantum accuracy along with computational efficiency must also be obtained. This chapter has demonstrated the use of density-functional methods for the study of systems of organic and biological interest, namely ethene, dialanine and the chorismate to prephenate rearrangement. The results obtained from calculations on ethene yield bond angles and bond lengths in agreement with experiment. It would be prudent to note that it is only the internal energy of the molecules that is being calculated here. There is no consideration of entropy in these calculations. In principle it would be possible to calculate the entropy and subsequent free energy of these molecules by calculating the electronic, rotational and vibrational partition functions. Computing the vibrational frequencies required to evaluate the vibrational free energy contribution is much more computationally expensive compared to calculating the electronic energy alone, though of course an approximation is possible based on the approximate Hessian generated during the minimisation procedure. When considering different conformations of large biomolecular complexes such as proteins or enzymes, there can be large differences in the contribution of the entropy to differences in the free energy, so it is important to carefully choose the system under study such that the entropic contributions are small or that the only comparisons made with experiment are the changes in enthalpy.

From the present work on dialanine, ONETEP DFT structural optimisation yield bond lengths and angles that agree well with the C22VAC classical potential parameterised using high-level QM calculations on small molecules. However, relatively large discrepancies

were found in the torsion angles of the transition states, which is to be expected due to the lack of a rigorous treatment for dispersion interactions in DFT but is also linked to the deficiencies present in classical force fields in describing regions away from where they are parameterised. This shows that the resultant coordinates from a parameterised force field can in some instances produce better torsion angles, but it is also important to use a method incorporating minimal parameters in the model. This not only allows the model to be transferable but also ensures that multiple potential energy surfaces are not found from the use of models with differing parameters, as shown in Ref. [251]. Ultimately, the work on dialanine has shown the validity of using an approach, such as that utilised in Chapter 6, where the atomic coordinates calculated from a classical potential are used to act as a protein matrix to surround an enzyme active site. This will allow QM-based methods to be used to break and form electron bonds, something that can not be described classically, and to perform TS searching, whilst the surrounding protein remains in the classical coordinates and torsion angles.

The work on the chorismate to prephenate rearrangement obtained converged transition states, found using LST/QST, that were further refined using eigenvector-following. This additional refinement resulted in the activation barrier changing by less than 0.1 kcal mol^{-1}. It is therefore sensible to use the computationally less expensive LST/QST method for the large-scale DFT calculations that are the subject of Chapter 6. The systems focussed on in that chapter, and also Chapter 5, will be water clusters, benchmark proteins and a significant portion of an enzyme.

Chapter 5

Explaining the closure of calculated HOMO-LUMO gaps in biomolecular systems

"Discovery consists of seeing what everybody has seen and thinking what nobody has thought."

Albert Szent-Györgyi (1893-1986)

It is quite alarming that several publications exist in the literature raising serious questions about the applicability of DFT techniques to large-scale systems such as water clusters and proteins [260, 292–296]. These investigations have demonstrated that calculated energy differences between the highest occupied and lowest unoccupied molecular orbitals (HOMO-LUMO gaps) are vanishingly small, or in some cases non-existent, for biomolecular systems. Such systems should ideally display insulating behaviour, if the calculations are to be trusted. These unphysical results are generally attributed to the treatment of exchange in the density functional used. If true, this would be a serious impediment to the further application of state-of-the-art DFT techniques to systems of biological relevance. The work discussed in this chapter indeed confirms that DFT electronic structure calculations are hindered by vanishing HOMO-LUMO gaps in large water and protein clusters. However, in contrast to the investigations by previous authors, the work presented in this chapter conclusively demonstrates that the issue results not from the choice of density functional used within the calculations but from improper treatment of the interface between the system under simulation and the vacuum in which it is contained. It is shown that this produces large, spurious electrostatic fields, thus closing the HOMO-LUMO gap.

5.1 Introduction

In order to address the issue of vanishingly small HOMO-LUMO gaps, the work in this chapter provides practical and realistic solutions for ensuring the HOMO-LUMO gap is

maintained as the number of atoms in the system is increased. Some of the approaches used here in an attempt to prevent the closure of the HOMO-LUMO gap include structural optimisation of water/vacuum interfaces using classical methods; the screening of molecular dipole moments through the use of implicit solvation; and embedding the quantum mechanical system in the potential of classical point charges representing the water environment. The biomolecular systems used in this process consist of protein molecules containing up to 2386 atoms. I believe that the practical solutions demonstrated here should allow the continued investigation of complex biomolecular systems through the use of Kohn-Sham DFT. This work has important implications for the use of large-scale density-functional theory in the simulation of biomolecular systems, especially where there is crucial dependance on the accurate calculation of the energy differences between molecular orbitals, such as in the simulation of photoemission, optical absorption and electronic transport. In addition, the work outlined in this chapter also addresses a widely-held misconception about the unsuitability of applying Kohn-Sham DFT to such systems.

5.2 Vanishing HOMO-LUMO gaps

Despite the considerable interest in the use of *ab initio* simulations for the study of complex biomolecular systems, as mentioned above, there still remains a growing concern that DFT, in conjunction with pure exchange-correlation functionals (those defined as containing no Hartree-Fock exchange), may be inappropriate for the simulation of large molecular clusters. Two of the recent reports that show unphysical vanishing HOMO-LUMO gaps in systems such as proteins [292] and even water clusters [293], additionally blame poor convergence during the self-consistent electronic structure optimisation procedure on this vanishing gap. Poor self-consistent field (SCF) convergence has been shown both for BLYP and for B3LYP DFT functionals [297] and also for LDA and PBE functionals when simulating large Glu-Ala helices [102]. In addition, Grimme and co-workers have experienced SCF convergence problems when performing generalised gradient approximation calculations on large protein fragments including cation-ion pairs, due to the self-interaction error creating a vanishing HOMO-LUMO gap [296]. Investigations carried out using the TeraChem package, with BLYP and B3LYP functionals, to optimise polypeptide structures *in vacuo* reported a lack of SCF convergence, caused by self-interaction and delocalisation errors giving vanishing HOMO-LUMO gaps, for many of the peptides that were simulated [260]. Similar problems have been observed when using molecular fractionation with conjugated caps to compute *ab initio* binding energies in vacuum for protein-ligand complexes of between 1000 and 3000 atoms [298].

The SCF convergence issues that have been seen by many in the literature are widely blamed upon the well-known phenomenon of pure functionals underestimating the HOMO-LUMO gap, within Kohn-Sham DFT, of semiconductors and insulators [294, 299, 300]. In biological systems Rudberg *et al.* have shown in Ref. [292] that these issues result in

complete closure of the gap. However, while the lack of the derivative discontinuity of the exchange-correlation potential at integer particle numbers, discussed in Chapter 3, and errors in the single-particle eigenvalues resulting from the approximate nature of the functional itself will indeed act to reduce the gap, there is no inherently obvious reason why the effect should get worse as the system size increases [299], as reported by Rudberg *et al.* Furthermore, approximations to energy-dependent electron self-energies [301], which lie beyond the conventional Kohn-Sham formalism, need to be improved in order to recover the gap [302]. In addition, problems in predicting the gap have been linked to the incorrect description of systems that will contain fractional charges at large separation [303] as discussed in Chapter 3. This prediction of metallic behaviour, in systems that should generally demonstrate insulating characteristics, is concerning. It is clear from previous authors' work related to the band gap issue that the use of the Hartree-Fock approach overestimates the gap whilst hybrid functionals, defined as those that include a certain portion of Hartree-Fock exchange, often get quite close to experimental values for the gap, but do not solve the underlying problem of gap closure. In addition, as the size of the water clusters investigated is increased, all types of functional, both pure and hybrid, demonstrate a decrease in the gap until eventually it closes entirely. Previous authors have also shown that recovery of a sizeable gap and consequently robust self-consistent convergence of the electronic energy levels is possible. Refs. [292] and [295] achieved this by including embedded electrostatic point charges in the system to represent water molecules around an inner cluster treated with quantum mechanics whilst Ref. [298] simulated the system within a dielectric medium. The results therefore point to the possibility that the vanishing gap is in fact actually a surface effect and, thus, not an inherent difficulty when performing pure Kohn-Sham DFT calculations.

Within this chapter, the ONETEP software package has been used to investigate the HOMO-LUMO gap of water clusters and protein systems. In agreement with previous work on similar systems performed in Ref. [292] it has been found that, indeed, the gap often vanishes *in vacuo*. However, whilst these calculations show what others have seen before, this work provides conclusive evidence that the issue is actually not related to the use of a pure exchange-correlation functional at all. It is, in fact, a result of the approach used to prepare the system in the first place. In the following sections a number of practical measures for preparing large systems are outlined. These approaches counter the vanishing HOMO-LUMO gaps seen in previous works, opening the way for the continued investigation of systems of biological interest with Kohn-Sham DFT.

5.3 Water clusters

The correct treatment of water is crucial for accurate and realistic simulations of biomolecular environments. However, recent large-scale density-functional simulations in Ref. [292] using pure functionals have encountered SCF convergence problems when simulating iso-

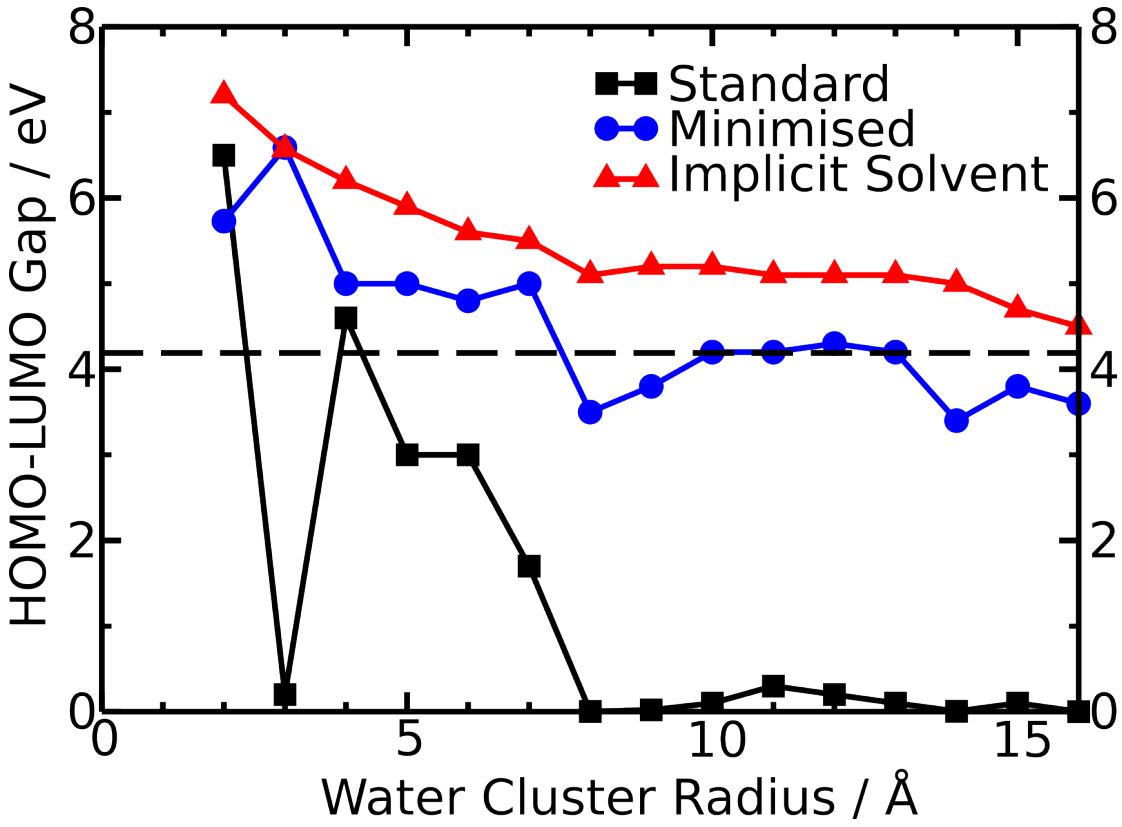


Figure 5.1: DFT HOMO-LUMO gaps of water clusters of increasing radius extracted from a larger 50 Å cube of water equilibrated at 300 K using classical molecular dynamics. Black line: Extracted straight from bulk water. Blue line: After classical minimisations are performed on each extracted cluster. Red line: Simulating the extracted clusters in an implicit solvent model. Dashed line: HOMO-LUMO gap of bulk water.

lated clusters of water, due to the HOMO-LUMO gap decreasing to zero when the cluster radius becomes larger than approximately 10 Å. This vanishing gap phenomenon is discussed in the present chapter and the results from simulations performed for the current work are summarised in Figure 5.1. Through a combination of AMBER and ONETEP, the HOMO-LUMO gap of a 2010-atom system of bulk water was calculated. A periodic supercell of water molecules was generated using the `tleap` module of the AMBER package with the TIP3P force field. The Coulomb interactions were then treated using the Particle Mesh Ewald sum, with a real space cutoff of 10 Å. The cutoff length for the Lennard-Jones interactions was also set to this distance of 10 Å. The system was minimised in the NVT ensemble before being heated to a temperature of 300K, in six stages, in the NPT ensemble. A production run of 5 ns at a temperature of 300 K was then performed from which snapshots were generated and saved at equal picosecond time intervals. Subsequent calculations on the last snapshot with the ONETEP code showed a clearly defined band gap of 4.2 eV. This value is consistent with previous DFT calculations of water employing the PBE gradient-corrected functional and is what one would expect to see for a bulk insulator. In addition to bulk water, isolated spherical water clusters of increasing size

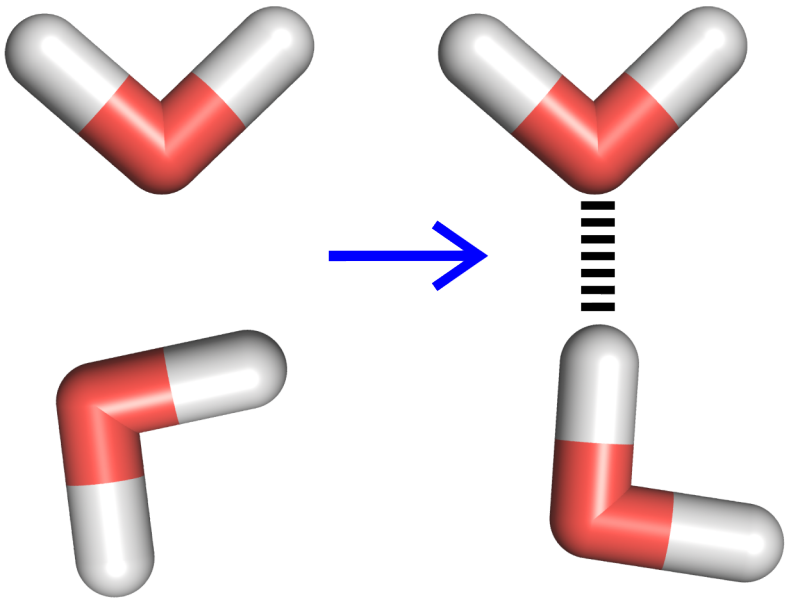


Figure 5.2: Rearrangement of water molecule orientation to maximise hydrogen-bonding and to minimise the electrostatic energy.

have also been investigated. These clusters were extracted from an initial 50 Å cube of water comprising 14,289 atoms that had previously been prepared in an identical manner to the 2010-atom system described above.

Recalling the quantum confinement effect, whereby, if the diameter of a material is of the same magnitude as the wavelength of the electronic wave function, then the electronic and optical properties can deviate substantially from those of bulk materials, one would generally expect to see the HOMO-LUMO gaps of these clusters as being larger than that of bulk water. One might also expect the HOMO-LUMO gap value to tend toward the bulk value as the size of the system was steadily increased. However, in agreement with Ref. [292], Figure 5.1 (black line) shows that the HOMO-LUMO gap quickly approaches zero for systems containing more than around 200 atoms. It is this observation that has led many to question the applicability of pure DFT functionals to large systems, such as water clusters and proteins [260, 292, 295, 296, 298]. However, the results demonstrating the insulating nature of very large periodic supercells of bulk water indicate that the problem is not the size of the system in itself but rather, the small HOMO-LUMO gap is in fact caused by the interface between the water and the surrounding vacuum introduced by the cluster. To understand this further, one must look more closely at the properties of bulk water. Bulk water consists of a continuous hydrogen-bonded network of water molecules; there is a dipole moment associated with each water molecule. The process for preparing the water clusters, which involves extracting a cluster, freezing the atomic positions and surrounding with vacuum, can potentially result in a large surface dipole being created. As a rough estimate, if one considers the dipole moment of a single isolated water molecule to be on the order of 0.73 e.a_0 then this will produce a potential difference

of around 0.4 V between opposing points of some sphere of radius 5 Å with the water molecule located in the centre of the sphere. In general, the molecular dipoles within the cluster are orientated in such a way as to mostly cancel out any long-ranged effects, as is illustrated in Figure 5.2, however, those dipoles on the surface of the sphere are not compensated by their neighbours. Therefore, depending on the orientation of these dipoles around the surface of the entire cluster, a large cluster may retain a large net dipole moment. In order to test this hypothesis, clusters were extracted from the 14,289-atom cube and their associate classical dipole moments were calculated. A TIP3P point charge model was used for the water and Figure 5.3(a) (green line) shows the classical dipole moment of the water clusters averaged over the 5000 snapshots from the molecular dynamics simulations. These results reveal that the dipole moment increases with system size as the created surfaces become larger in area. A similar trend in the dipole moment with cluster size is observed for QM calculations of single snapshots as can be seen in Figure 5.3(b).

To see whether these large dipole moments can cause the closure of the gap, one must also consider the potential due to the dipoles which is generated from the effective surface charge produced. In order to comprehend this, one must consider a uniform array of identical dipoles between two surfaces such as that illustrated in Figure 5.4. Internally, the heads and tails of the dipoles will be adjacent and thus will cancel, however, at the bounding surfaces, no such cancellation occurs. Instead, on one surface the dipole heads create a positive surface charge, whilst at the opposite surface the dipole tails create a negative surface charge. These two opposite surface charges create a net electric field in a direction opposite to the direction of the dipoles. To illustrate this effect further, the electrostatic potential calculated from density-functional simulations is plotted on a plane behind the 16 Å water cluster in Figure 5.5(a). The DFT-calculated electrostatic potential clearly reveals a dipolar potential. Now that one can see that water clusters that are extracted from equilibrated bulk periodic calculations display large multipole moments, as measured both by MM and QM, the natural next question to ask is what effect does this have on the computed HOMO-LUMO gap ? One approach to answering this question would involve calculating the density of states (DoS) of the water cluster. More specifically, the local density of states (LDoS) will give us a more localised picture of what is happening. Figure 5.5(a) also shows the LDoS for a 16 Å water cluster, alongside the DFT electrostatic potential. In order to determine the LDoS, the water clusters are nominally divided into 10 slabs in a direction perpendicular to the dipole moment. Then the slab local density of states is defined as the sum of the contributions to the total DoS from the local orbitals centered on the atoms within each slab. The plot in Figure 5.5(a) displays a clear shift in the LDoS as a function of the position along the dipole moment vector. This shift is due to the electric field pushing some states higher in energy and some states lower in energy. This effect can be considered as analogous to the concept of Fermi-level pinning found in polar semiconductor nanorods [304, 305]. In that

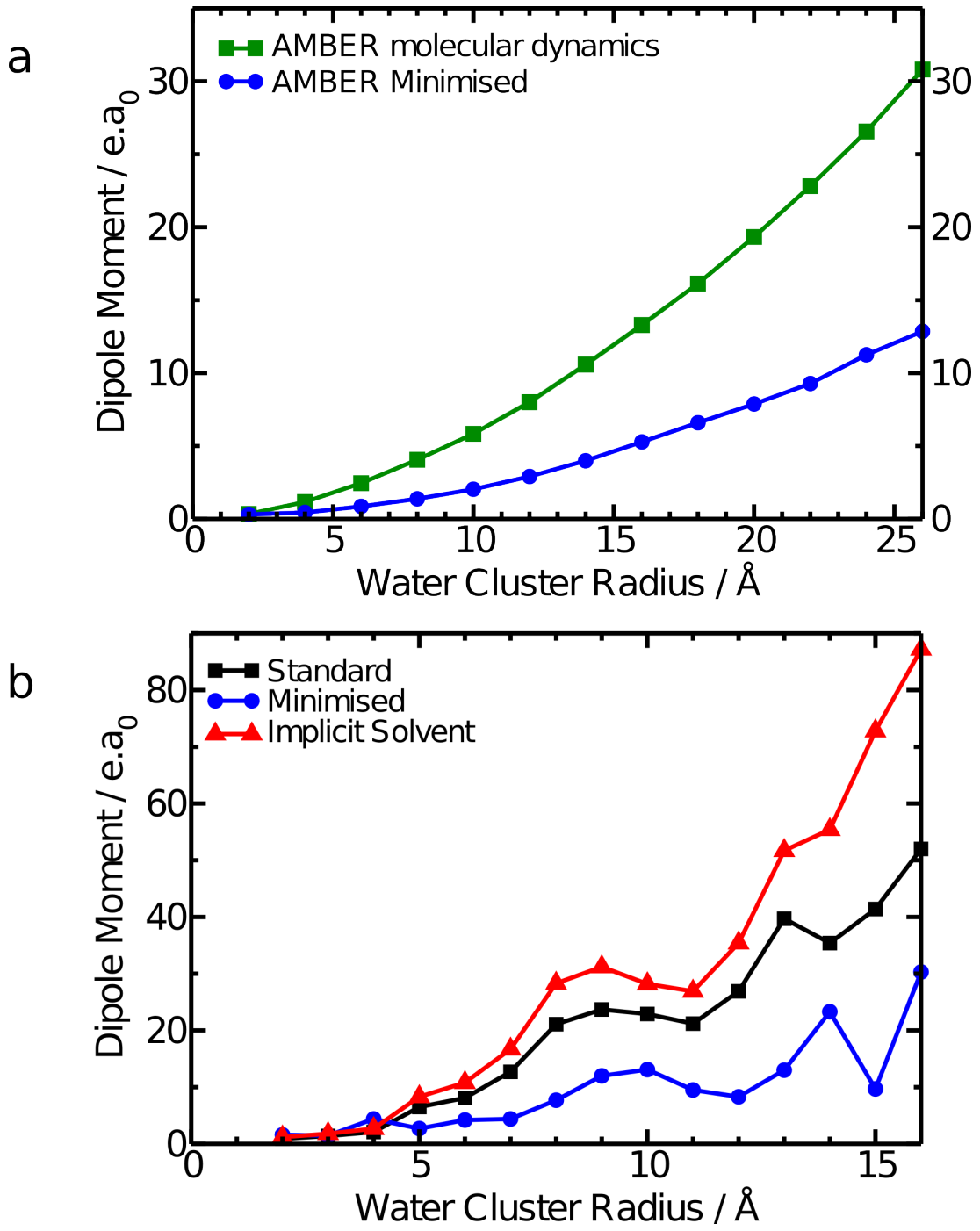


Figure 5.3: (a) Average dipole moments of water clusters of increasing radius calculated using the TIP3P point charge model. Green line: averaged over 5000 snapshots extracted from a larger 50 Å cube of water. Blue line: averaged over 1400 snapshots extracted from the bulk and minimised using an MM force field. (b) Quantum mechanically calculated total dipole moment of water clusters of increasing radius. Black Line: ONETEP-calculated QM dipole moment from the final molecular dynamics snapshot at each radius. The dipole moment increases with radius in the same manner as in our classical simulations. Blue line: ONETEP-calculated dipole on the same snapshot after classical minimisation, which reduces the dipole moment across the cluster. Red line: ONETEP implicit solvent calculations. In this case, the dielectric medium supports a higher dipole moment, although the net potential is screened at large distances.

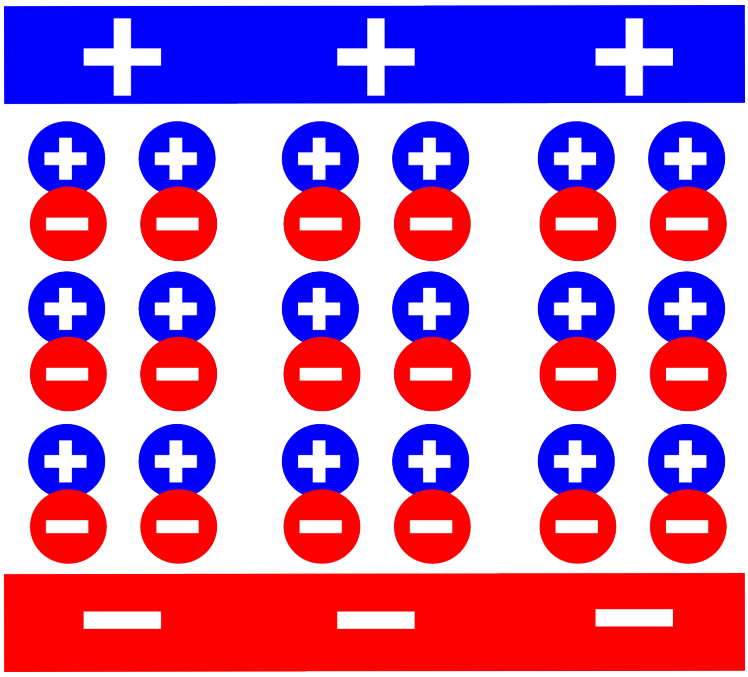


Figure 5.4: Schematic figure to illustrate that a uniform array of identical dipoles is equivalent to a surface charge.

particular case, the Fermi energy coincides with a finite density of states at either end of the rods. In the case of the water clusters the Fermi energy coincides with a non-zero DoS on opposite surfaces of the extracted sphere. When the radius of the water cluster increases such that the surface potential is of sufficient magnitude to bridge the HOMO-LUMO gap, it would be expected that the gap will disappear completely. Given the apparent electrostatic nature of the vanishing HOMO-LUMO gap, it would make sense to postulate that there is no fundamental problem in the use of pure density-functionals in the simulation of large systems, but simply that the issue manifests itself at smaller system sizes than it would do for hybrid functionals which have an inherently larger gap. It would therefore be expected that any method that corrects for these surface effects will also restore the HOMO-LUMO gap, which is consistent with observations made by other authors. By embedding the system in a set of classical point charges outside the electron distribution to represent, for example, the aqueous environment of the water or protein cluster, work in Ref. [292] has shown that the HOMO-LUMO gap may be restored. In these particular cases there were no significant changes seen to occur to the electronic density of the inner water molecules. In the work presented in Ref. [295], the only significant changes in electronic density were observed on the water molecules that were close to the surface of the cluster. The investigation reported in Ref. [298] also found that the use of a dielectric medium with permittivity $\epsilon = 4$ leads to robust self-consistent field convergence of proteins in vacuum, therefore this approach implies that screening of the surface dipole is sufficient to restore the HOMO-LUMO gap. In the remainder of this chapter a number of methodologies for setting up a QM cluster are tested and

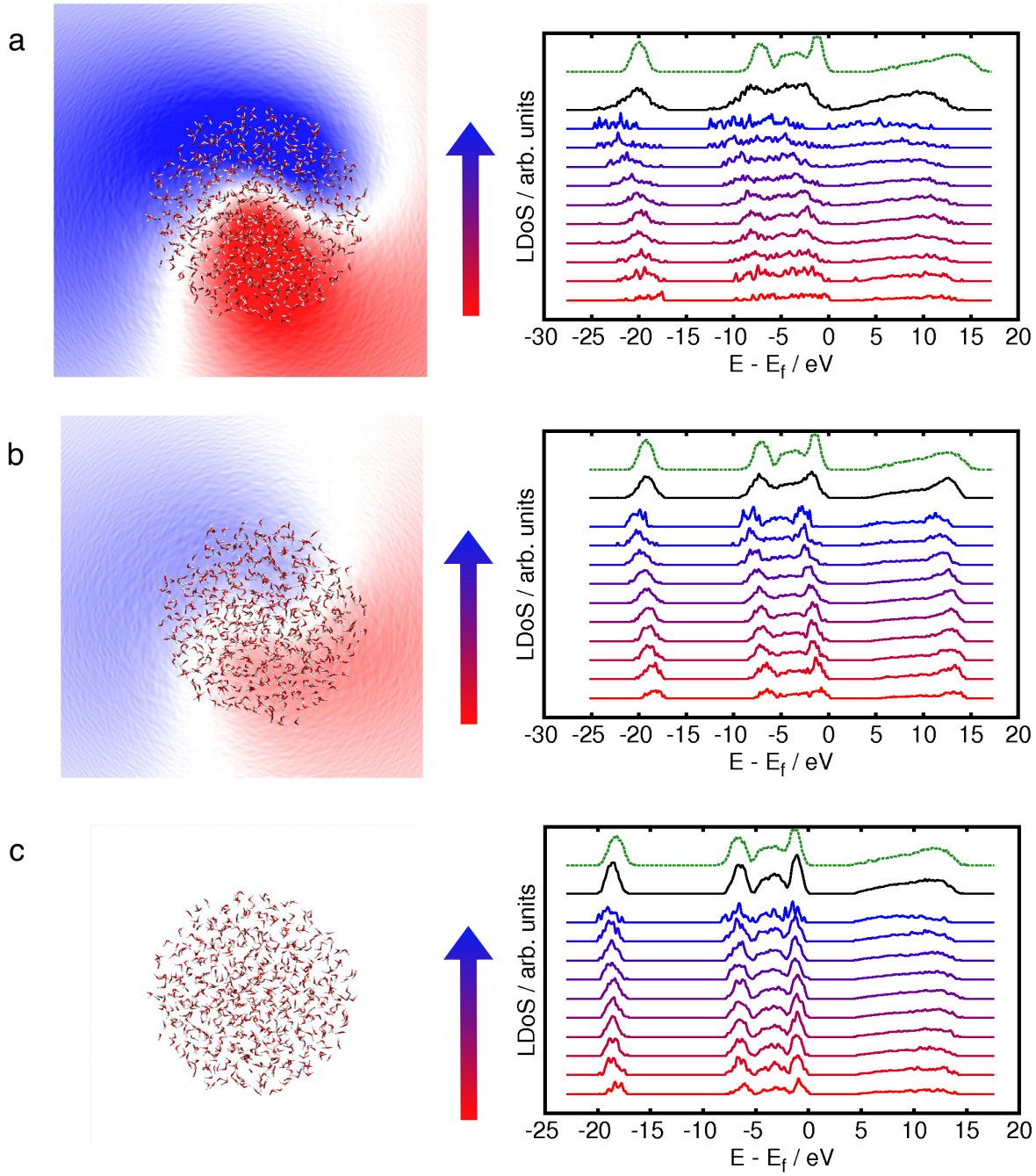


Figure 5.5: Electrostatic potential for a water cluster of 16 Å radius and local density of electronic states (LDoS) for groups of atoms as a function of position along the dipole moment vector of the cluster. The dipole moment vector (coloured arrow) runs from the red line to blue. The black line is the total density of states (DoS) and the green dashed line is the DoS for bulk water. Each line in the LDoS plot is normalised by the number of molecules contained in the slab. The electrostatic potential ranges from -0.3 V (red) to +0.3 V (blue). The slice is 24.6 Å behind the water cluster. (a) Snapshot extracted from bulk water. The dipole moment is high, the LDoS is strongly dependent on position relative to the dipole moment vector, and the total range of states is much wider than for bulk water. (b) After classical minimisation of the same snapshot. (c) Simulated using the ONETEP implicit solvent model. In both cases, the dipole moment is reduced and the DoS closely resembles the bulk.

are shown to allow density-functional calculations to be performed using either pure or hybrid functionals, without closure of the gap, so that the ensuing results are accurate and realistic. The electric field across a water cluster will be reduced if the atomic positions are allowed to relax, either by some form of structural optimisation or a simulated annealing procedure. The particular form of geometry optimisation applied to the water clusters was fast conjugate gradient (CG) optimisation followed by Newton-Raphson (NR) minimisation until the root mean square force decreased below 10^{-4} kcal mol^{-1} Å^{-1} during CG minimisation and below 10^{-10} kcal mol^{-1} Å^{-1} for NR minimisation. This substantially reduced the average dipole moment of the extracted water clusters, as measured using classical TIP3P point charges. The effects of the optimisation can be seen in Figure 5.3(a) (blue line). In addition, Figure 5.1 (blue line) reveals that clusters that have undergone MM minimisation all have their HOMO-LUMO gaps restored to values close to the bulk water value of 4.2 eV. It is also expected that an implicit solvation model will reduce this observed shift in electronic states on opposite surfaces of the sphere by screening the electrostatic potential across the entirety of the cluster. It can be seen in Figure 5.1 (red line) that when the extracted water clusters are simulated with implicit solvent using ONETEP, the HOMO-LUMO gap is again restored to the value found in bulk water. In panels (b) and (c) of Figure 5.5 it is shown that following classical minimisation the dipole moment of the 16 Å water cluster is significantly reduced and the associated electrostatic potential is negligible when simulated in the dielectric medium. For both the case of classical minimisation and the use of implicit solvation, the local density of states is much less dependent on position along the dipole moment vector and it much more closely resembles the bulk density of states, shown in Figure 5.5 (green dashed line), as should be expected for a large water cluster.

5.4 Protein systems

The problems of vanishing HOMO-LUMO gaps are not only found in water but also in biomoleclar systems Therefore, the next area of focus must be on proteins, to ensure that the computational approaches used are reliable and that the results being produced are accurate and trustworthy. In order to investigate the claims made in Ref. [292], six protein conformations: methionine-enkephalin (1PLW) [306], the RGD peptide (1FUL) [307], transthyretin (1RVS) [308], the third intradiskal loop of bovine rhodopsin (1EDW) [309], the seventh transmembrane helical domain of bovine rhodopsin (1FDF) [310] and ubiquitin (1UBQ) [311], were accessed from the Brookhaven National Laboratory Protein Data Bank (PDB) [312] as starting configurations for the calculations. With regards to the specific ONETEP parameters used in both these and the water cluster calculations presented in this chapter, an energy cutoff of 916 eV was used, corresponding to a psinc grid spacing of 0.475 a_0. The NGWFs were localised in real space with radii of 5.3 Å. For the case of the 1PLW protein, upon increasing the NGWF radii from 5.3 Å to 6.4 Å , the

			HOMO-LUMO Gap / eV			
PDB ID	Atoms	Charge	*in vacuo*	QM water	Implicit Solvent	QM/EE
1PLW	75(456)	0	0.0	3.7	3.7	3.5
1FUL	135(453)	-1	n/a	2.7	2.6	2.6
1RVS	172(670)	0	n/a	3.4	3.7	2.9
1EDW	399(978)	-1	n/a	3.1	3.7	2.6
1FDF	419(1526)	3	n/a	1.9	3.3	1.6
1UBQ	1231(2386)	0	n/a	2.6	3.4	2.4

Table 5.1: HOMO-LUMO gaps for a range of proteins from the PDB. Atom number in parentheses includes a 5 Å solvation shell of water used in classical minimisation and QM/EE simulations. Systems that did not converge are indicated by n/a. Vacuum calculations and implicit solvent simulations did not include any explicit water molecules.

calculated HOMO-LUMO gap was found to be converged to within 3 meV. Repeating these calculations at an energy cutoff of 1020 eV, corresponding to a grid spacing of 0.45 a_0, the calculated HOMO-LUMO gap was converged to within 6 meV. The NGWFs in ONETEP are optimised *in situ* to represent the valence states, however, previous experience shows that these NGWFs describe the conduction states well for at least the first 1 to 2 eV above the LUMO and thus produce the same level of gap as equivalent plane-wave calculations. At energies beyond this point, however, work reported in Ref. [108] shows that the density of states is not well-represented by NGWFs and should be discounted.

In the instance that a set of starting configurations that had been obtained from the PDB had more than one conformation available, the structure labelled as 'model 1' was used in each case. It is worth noting that these coordinates are resolved from NMR investigations and so give the positions of the hydrogen atoms. Therefore, with no prior preparation, the coordinates extracted from the PDB were placed straight into a calculation in ONETEP *in vacuo*. The computed HOMO-LUMO gaps for these structures can be seen in Table 5.1. As would be expected for this method of system preparation (or lack thereof !), the simulations did not produce finite HOMO-LUMO gaps and so therefore did not converge for any of the proteins apart from the smallest system that was studied. In plotting the DFT electrostatic potential far from the 1UBQ protein, a strong dipole moment is revealed, as can be seen in Figure 5.6(a). It is also clear from the local density of states in Figure 5.6(b) that a number of electronic states for the protein are close to the Fermi level. Clearly the problem of the vanishing HOMO-LUMO gap is very similar in nature to that found in water clusters. Other authors have also since shown that the number of protein electronic states close to the Fermi level can be reduced by simulating the protein in a dielectric medium [122]. Perhaps some clarity can be gained from stepping back a little and thinking about protein characteristics from a broader viewpoint.

When considering the general properties of protein secondary structure, all the back-

bone hydrogen bonds in an α-helix point in the same direction. This is due to the fact that all the peptide units are aligned in the same orientation along the helical axis. A peptide unit has a well-defined dipole moment, arising from the differing polarity of the N−H and the C′−O groups and the partial double bond character of the N−C bond [313]. The accepted value for the dipole moment of the individual peptide units is around 0.9 to 1.3 e.a_0 [314]. Around 97% of the peptide dipole moments point in the direction of the helical axis and this percentage is insensitive to the dihedral angles [315]. The amino terminus has a partial positive charge and the carboxyl terminus has a partial negative charge. Therefore, the overall effect on the resulting α-helix is a significant net dipole, which to a first approximation one can deduce as being equal to $N\times1.1$ e.a_0 for N residues [316]. The most common location for an α-helix in a protein is on the outside of the structure, where one side will face the hydrophobic interior of the protein and the other will face the solution [317]. The α-helices that are not part of an enzyme active site, or a protein binding site in general, will often have a negatively charged side-chain at the amino terminus or a positively charged residue at the carboxyl terminus. These dipole-compensating residues act to stabilise the helical form of the peptide in solution. In addition to these stabilising residues, for a protein solvated in aqueous solution, the effective dipole of an α-helix will be reduced as a result of solvent screening of the peptide group charges. The solvent generates a reaction field that acts against the field generated by the vacuum dipole, leading to the screening and effective lowering of the dipole moment [318]. When calculating these structures *in vacuo*, due to a lack of electrostatic reaction field that was generated by the solvent, the strength of the α-helix dipole, when compared to aqueous solution, may increase drastically [319]. The most significant factor in this increase in dipole moment is likely to be charged side chains such as solvent-exposed Arg, Lys, Glu and Asp. As described in Chapter 2, these residues are generally charged in physiological conditions at pH 7. When simulated *in vacuo* these residues are unscreened and, without correction, the α-helix dipole moment due to these untreated side chains will be large, causing an undesired shift in the surface electronic states. It has also been found that when in aqueous solution the effective dipole moment is found to have a strong dependence on the position and orientation of the helix with respect to the solvent. It is of note that additional secondary structure motifs such as β-sheets carry comparatively little dipole moment. However, in calculations, the surrounding solvent must be treated with care otherwise spurious and unphysical effects are likely to arise.

In an attempt to recover the expected HOMO-LUMO gaps in proteins, similar techniques to those described in the previous section have been used. The protein structures were solvated in a 50 Å water cube using the TIP3P force field and all of the protein interactions were described using the AMBER ff99SB biomolecular force field. NVT minimisation was initially performed on the system before equilibrating in the NPT ensemble up to a temperature of 300 Kelvin in six equal steps. A 5 ns NVT production run was then performed in order to generate the final structures. Throughout the minimisation, equi-

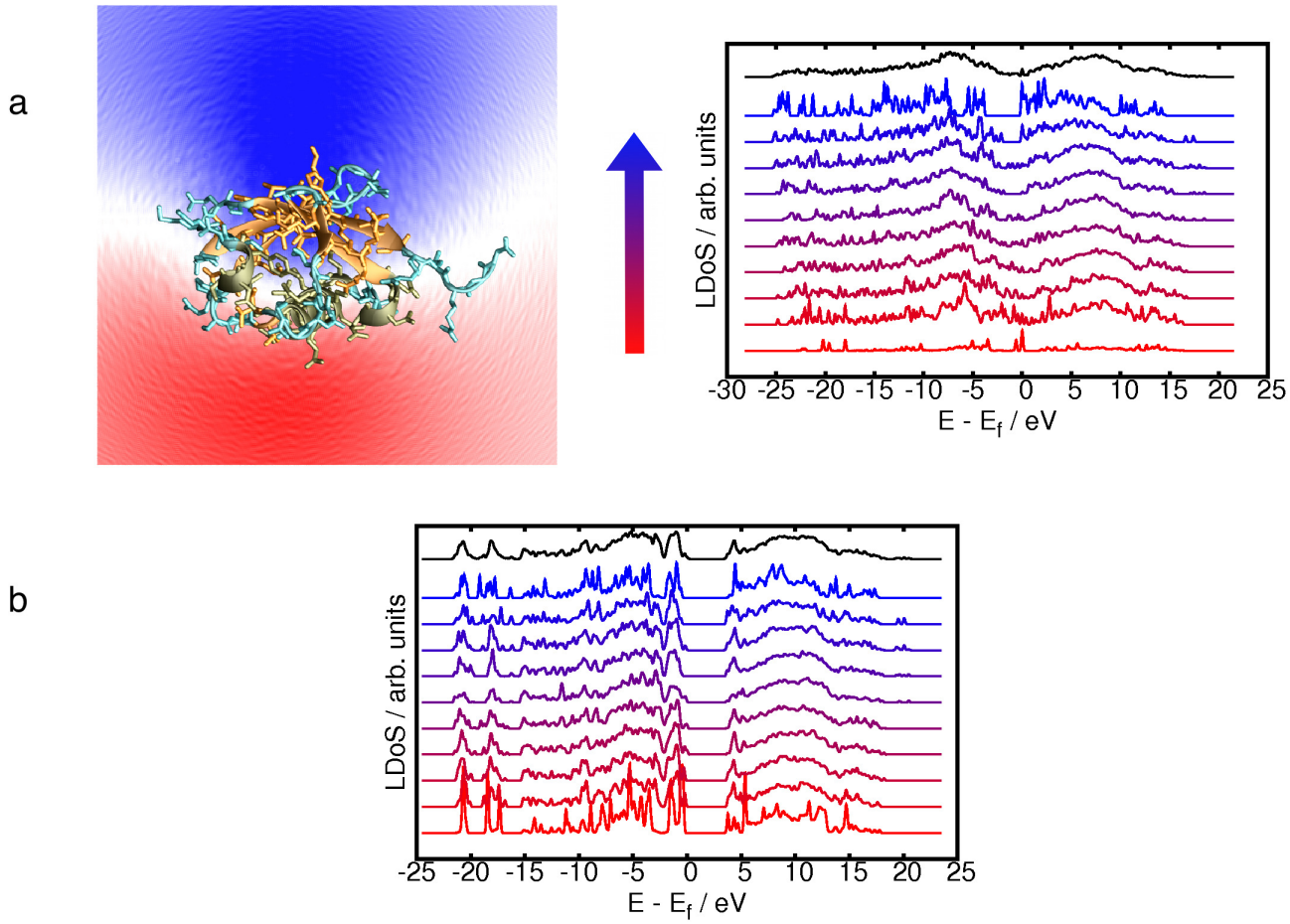


Figure 5.6: Electrostatic potential and local density of states (LDoS) for groups of atoms as a function of position along the dipole moment vector (coloured arrow) of ubiquitin. The dipole moment vector runs from red to blue. The black line is the total density of states. Each line in the LDoS plot is normalised by the number of molecules contained in the slab. Panel (a) shows the experimental structure with no solvation and the electrostatic potential ranges from -0.2 V (red) to +0.2 V (blue). The slice is 42.9 Å behind the protein. Panel (b) shows the LDoS along the dipole moment of the same protein structure, this time after having been simulated with implicit solvent.

libration and production runs, harmonic constraints of 100 kcal mol^{-1} Å^{-2} were imposed upon the protein structure. Following these runs in a 50 Å water cube, the majority of the water molecules were stripped from the system. 5 Å of the surrounding water molecules were retained from the simulations for each protein. The solvent geometry was then optimised via fast conjugate gradient (CG), followed by Newton-Raphson (NR) minimisation, until the root mean square force decreased below 10^{-4} kcal mol^{-1} Å^{-1} during CG minimisation and below 10^{-10} kcal mol^{-1} Å^{-1} for NR minimisation. During this process the protein residues remained constrained. The resulting protein configurations were used as the starting vacuum conformations for ONETEP calculations.

Following classical minimisation on the 1FDF structure *in vacuo*, the electronic struc-

ture calculation again failed to converge. The calculations thus have shown that whilst classical optimisation is able to restore the HOMO-LUMO gap for water clusters, this approach is unsuitable for proteins. This result can be understood from the fact that the protein residues are fixed in their secondary structure conformations. This constraint results in much less opportunity for the structural mobility of proteins in general, especially when compared to water. In order to proceed beyond the methods used to successfully recover the HOMO-LUMO gap of the water systems, a more effective strategy will be to include the effects of the protein environment through the use of explicit water molecules or, the computationally less intense, implicit solvation. Either approach should screen the effect of the charged residues within the system.

The first step is to take each protein structure that has been solvated by the classically minimised 5 Å layer of water and simulate them in ONETEP using full DFT for the entire system, up to a maximum system size of 2386 atoms, in order to re-calculate the electronic structure. Calculations using implicit solvent have also been performed for the protein structures in their vacuum configurations. The strategies of using either explicit or implicit solvation both restore the HOMO-LUMO gaps to similar values. It can be seen in Figure 5.6(b) that the density of states for the implicitly solvated system resembles more closely that of an insulator. As mentioned in Chapter 3, by representing the explicit water layer by embedded point charges through the use of a TIP3P charge distribution and removing the water molecules from the calculation, the computational costs associated with a calculation can be dramatically reduced. To explore whether this approach retains the correct gap the explicit water layer surrounding the protein conformation has also been removed and, instead, represented by embedded point charges with a TIP3P charge distribution. In this instance, the calculated HOMO-LUMO gap is restored to a value very similar to that of the simulation using a full QM water layer. This result shows that classical charges can indeed reproduce the correct electrostatic environment. Overall, Table 5.1 reveals that the use of implicit solvent largely produces a HOMO-LUMO gap that is of greater magnitude than when an explicit water layer is considered or embedded charge distributions are used. The significant outliers are shown in the QM water and QM/EE gap values for the 1FDF protein. It is likely that these particular methods struggle with the net charge of $+3e$, indicating that in such systems where there is a significant net charge, an implicit solvent approach must be used. Furthermore, these discrepancies in the gap values are more pronounced as the system sizes increase. The use of explicit QM water manages to produce significant gap values up to the mid-range of the systems in this study but Table 5.1 shows that for larger systems it is necessary to use the computationally more expensive implicit solvent model. As an additional viable alternative to restore the calculated gap, charged residues within the 1FDF protein are mutated to alanine, and any charged groups of the N- and C-terminus are hydrogen-capped. Alanine is used because of its non-bulky, chemically inert, methyl functional group that adequately mimics the secondary structure preferences that

many other amino acids possess. The calculated HOMO-LUMO gap increases by 1.3 eV, using this technique, after the original calculation on the unmodified structure failed to generate sensible eigenvalue occupancy. This is achieved by reducing the spurious dipole moment caused by the vacuum-exposed charged residues.

5.5 Summary

This chapter has confirmed recent findings from Ref. [292] that DFT electronic structure optimisation can be hindered by vanishing HOMO-LUMO gaps in large water and protein clusters, systems that should, in fact, display insulating behaviour. This problem has been shown to manifest itself in clusters prepared with improper treatment of the interface between the system and the surrounding vacuum. From the examples presented in this chapter it has been shown that unequilibrated vacuum/water interfaces combined with X-ray protein crystal structures taken straight from experimental repositories can exhibit strong molecular dipole moments. The present work on water has shown that starting from a continuous polar substance, where each molecule has a fairly large dipole moment, the randomly arranged network of dipoles will then rearrange in order to minimise the electrostatic energy. The protocol used in this chapter involved extracting a cluster from a larger classical simulation, or in the case of the proteins, simulating an entire structure using the experimental coordinates obtained from NMR or X-ray. This process of extracting a smaller cluster, freezing the atomic positions and surrounding with vacuum, results in a large surface dipole being exposed. Larger and larger extracted clusters have surfaces further apart which results in increasing net dipole moments due to the larger surface areas. The spurious electric fields associated with these unphysical dipole moments will reduce the HOMO-LUMO gap by raising the energies of the electronic states on one side of the cluster and lowering the energies of those on the other side. Depending on the value of the local electric field, a large enough cluster will have the HOMO-LUMO gap closed completely

By investigating the local density of states of these systems, decomposed into slabs along the direction of the molecular dipole moment, it has been proved that the energies of the electronic states are shifted by the electric field generated across the cluster. This then results in the Fermi energy being pinned by states on opposite surfaces of the water cluster, leading to the HOMO-LUMO gap closing, something that should not happen for these structures. Whilst, in the literature, this effect is widely associated with the use of the PBE gradient-corrected functional, the results presented here emphasise that this effect should not be particular to the PBE functional used in these calculations. Previous authors have shown that hybrid functionals tend to have an intrinsically wider gap and Refs. [260] and [298] have demonstrated that calculations, for systems comprising thousands of atoms, implementing those functionals do converge. However, it is expected that, even for functionals containing Hartree-Fock exchange, the HOMO-LUMO gap is

still likely to close upon increasing system size at the time when DFT methodological advances allow such access to larger systems. The development of linear-scaling functionals with accurate Hartree-Fock exchange is a current area of research [72]. With such methods available it would be prudent to test the HOMO-LUMO gap dependence on increasing system size for the particular cases outlined in this chapter.

Practical solutions for restoring the HOMO-LUMO gap in water clusters and protein systems have been demonstrated in this chapter. The methodologies used have ranged from classical structural optimisation of the interfaces between water and vacuum, to the screening of molecular dipole moments through the implicit solvation of protein structures. It has been shown that implicit solvation seems to give the best correspondence between the HOMO-LUMO gaps of large isolated explicit water clusters and that of bulk water obtained in periodic calculations. The use of implicit solvation techniques also restores larger HOMO-LUMO gaps for proteins to a greater extent than when 5 Å of the surrounding water molecules, retained from bulk periodic simulations, are explicitly simulated. The systems investigated here comprised up to 2386 atoms and the practical solutions demonstrated in this chapter have implications for the remainder of the dissertation as they show that the proposed methodologies for treating biomolecular structures will generate sensible and reliable results. It has also been shown that the use of classical charges can reproduce the correct electrostatic environment, and hence restore the HOMO-LUMO gap, whilst also significantly reducing the computational cost of the simulation, compared to using explicit QM water. This has positive implications for future DFT studies of biomolecular systems, as the computational costs can be reduced by such approaches. However, of more immediate importance is the fact that the next chapter relies heavily on this approach in order to reduce the computational costs for large clusters of water molecules undergoing full-DFT structural optimisation. Therefore, the results presented in this chapter instil confidence in the approach used in the next chapter. The calculations presented here could be further extended to systems such as the myoglobin protein (PDB ID: 1A6N) to investigate whether other purely quantum mechanical phenomena, such as the spin states on an iron ion, will be better described compared with experimentally resolved structures as a result of following the classical structural optimisation procedures discussed in this chapter. Another potential application area of impact could be the spectroscopy of proteins, where the HOMO-LUMO energy levels of central pigments are crucial [96, 122]. In general, I am hopeful that the insights from the investigation presented in this chapter will be a small contributor toward allowing the continued modelling and simulation of biomolecular systems through the use of Kohn-Sham DFT. One such system, where there is additional complexity to the protein structure, is an enzymatic reaction in which a small molecule is undergoing some chemical reaction catalysed by the surrounding protein structure. Such a system is the focus of the next chapter.

Chapter 6

A density-functional perspective on the chorismate mutase enzyme

"You can never cross the ocean unless you have the courage to lose sight of the shore"

Christopher Columbus (1451-1506)

The last two decades have witnessed the continual and concerted effort toward the development of powerful tools which allow DFT calculations to be efficiently performed for systems containing thousands of atoms. The challenge of performing quantum mechanical simulations on such large systems is not just the computational cost of a calculation for, say, a 1000-atom system, but also the fact that such large systems have complex free energy landscapes thus significantly increasing the number of calculations needed to extract meaningful predictions of the properties of such systems. The efficiency of the linear-scaling DFT code ONETEP, along with significant associated computing resources, allow real science to be performed, rather than simply allowing a small number of single-point energy calculations to be performed, which would be the case if conventional cubic-scaling DFT codes were used. A key aim in this chapter is to calculate an activation and reaction energy for the conversion of chorismate to prephenate, catalysed by the *Bacillus subtilis* chorismate mutase (CM) enzyme, which is fully converged with respect to the size of the system. In doing so, a powerful proof-of-principle demonstration of the predictive power of DFT calculations in biology will be demonstrated, which will, hopefully, in turn, provide a very powerful push for the adoption of first-principles modelling techniques within biologically-relevant disciplines. In this context, it is worth noting that it was successful Grand Challenge applications in the early 1990's that led to the widespread adoption of DFT within the physical sciences. It is my impression, garnered from experts in the field with much more experience than I, that we are very close to a similar tipping point for the adoption of DFT in biology. However, for this to take place it will require the successful demonstration of proof-of-principle Grand Challenge applications such as accurate simulations of an entire enzyme. It is my hope that the calculations presented in this dissertation will serve as a modest foundation for the further pursuit of such

milestones within biomolecular simulation.

On the specifics of the work included in this chapter, a benchmark study on a large portion of the CM enzyme using linear-scaling density-functional theory is discussed. As outlined in Chapter 3, treating the entirety of an enzyme with conventional QM approaches is largely unfeasible due to computational demands, so hybrid QM/MM methods are often applied instead. A recent QM/MM study has identified reaction pathways for the rearrangement of chorismate to prephenate in solution and catalysed by CM [320]. However, due to the advances in linear-scaling density-functional methods outlined in Chapter 3, it is now possible to apply these approaches to accurately predict transition state geometries and energetics through treating a system of thousands of atoms at the fully quantum mechanical level. QM/MM may suffer from inaccuracies introduced by using classical force fields and from the coupling scheme used to link the two regions. However, a full-DFT approach will allow a comparison to be made with hybrid methods to investigate these inaccuracies. Through the use of the ONETEP code, large-scale DFT calculations are performed on structures taken from the CM pathways in Ref. [320], in order to address the convergence of energies of activation and reaction with respect to the total size of the fragment considered.

6.1 Introduction

The CM enzyme is relatively simple, but has still managed to generate much controversy and debate amongst enzymologists, despite just catalysing a one-step pericyclic reaction. The enzyme catalyses the Claisen rearrangement of chorismate to prephenate, the gas phase version of which has been discussed in Chapter 4. Within the larger scheme of the biological process, the reaction is situated at a branch point in the shikimate metabolic pathway [321]. This particular pathway is crucial for generating the aromatic amino acids phenylalanine, tyrosine and tryptophan. In terms of practical applications, it has been shown that herbicides that inhibit the biosynthesis of amino acids prove to be very useful tools within the weed management industry. The particular success of these types of herbicides has been due to their low toxicity in mammals, in other words, these herbicides inhibit pathways that are lacking in mammals. There are now several types of herbicides used within the industry with primary targets, or sites of action, that are associated with the targeted and specific inhibition of enzymatic activity within biosynthetic pathways for amino acids [322]. As discussed in Chapter 3, synthesised molecules which act as transition state analogues are competitive inhibitors of enzyme activity, binding more tightly to the active site than the natural or expected substrate in the reaction. A major difference with designed transition state analogues is that the dissociation rate will be orders of magnitude slower. Therefore, once the synthesised molecule binds, the enzyme is essentially inactivated. Such an analogue has been synthesised for the CM enzyme [323] and has been used to crystallise the enzyme [324]. The deregulation of the shikimate

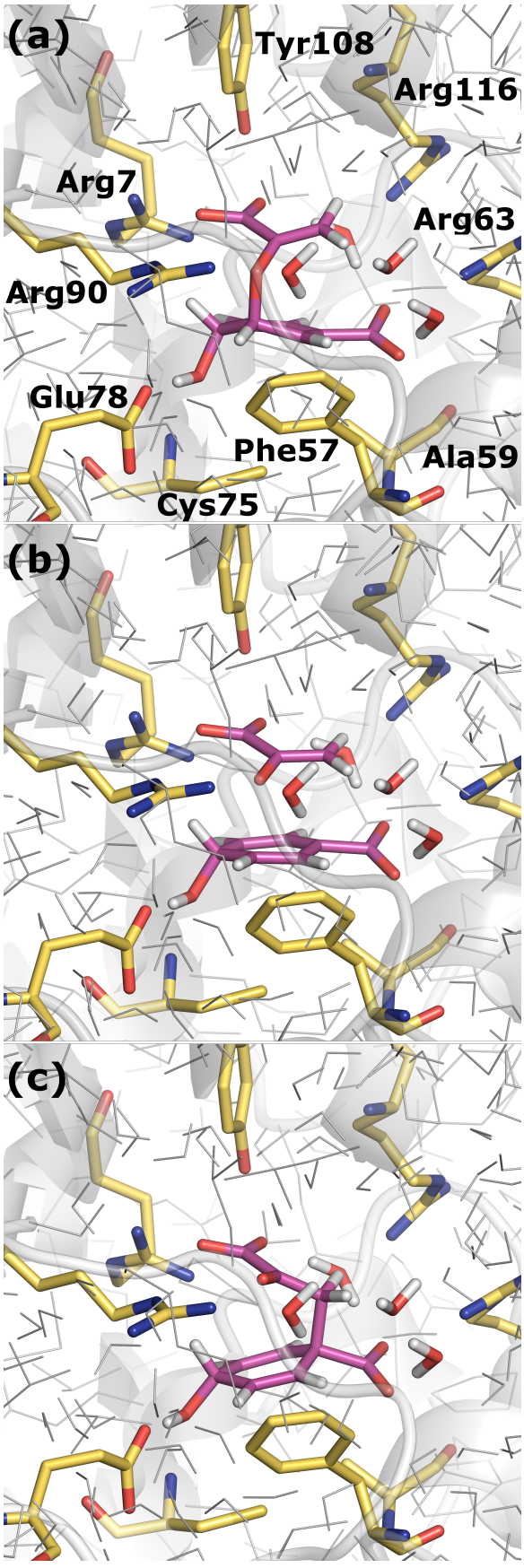


Figure 6.1: Rearangement of the substrate (magenta) from chorismate to prephenate within the CM active site (yellow) and surrounding protein (grey) in the (a) reactant, (b) transition state and (c) product conformations from DFT-optimised structures.

pathway results in the accumulation of very high levels of shikimate and shikimate-3-phosphate, and in some plant species this accumulation can account for up to around 16% of plant dry weight in sink tissues, where the products of photosynthesis are used or stored. Important building blocks for other metabolic pathways are also reduced by uncontrolled carbon flow through the shikimate pathway and reduced levels of aromatic amino acids cause a significant reduction in protein synthesis. It has been shown that the shikimate pathway exists only in fungi, bacteria and higher plants [325]. Therefore CM inhibitors may be useful in the development of herbicides, fungicides and antibacterial therapeutics [323] with low toxicity. Work investigating catalysis within CM may then be able to elucidate more general principles of catalysis and TS analogue binding that can then be harnessed in order to understand further a variety of other enzymes.

A key factor in why CM has undergone much study, both through computation [320, 326–329] and experiment [324, 330–337], is the fact that there is no covalent bonding between the substrate and the enzyme active-site residues [330–334]. This can be seen in Figure 6.1. It is this characteristic of the enzyme that has led the majority of researchers in the field who are treating the enzyme computationally, to do so using QM/MM whereby the substrate is treated with a quantum mechanical method and the active site and surrounding enzyme residues and water molecules are treated with a classical molecular mechanics approach. In addition, it has been shown that the reaction also takes place in aqueous solution, with a similar mechanism [335]. Such an observation allows a direct comparison of the reaction in the two environments. Therefore, the catalytic enhancement of the enzyme can be calculated via simulation and directly compared with experimental observations. Experimental investigations have found the enthalpy of activation to be lowered from 20.71 $\pm$ 0.35 kcal mol^{-1} in a water environment [335] to 12.7 $\pm$ 0.4 kcal mol^{-1} in the presence of the CM enzyme [333]. This lowering of the activation barrier, in going from one environment to another, translates to a catalytic enhancement to the reaction rate of approximately 10^6. The enthalpy of reaction in water has been shown, via calorimetric measurements, to be equal to -13.2 $\pm$ 0.5 kcal mol^{-1} [334]. The QM/MM study detailed in Ref. [320], of the reaction both in enzyme and solution environments at the B3LYP/6-31G(d)/CHARMM27 level of theory, yielded activation energies of 17.4 $\pm$ 1.9 kcal mol^{-1} in water and 11.3 $\pm$ 1.8 kcal mol^{-1} in enzyme. To have a greater understanding of the accuracy of these simulations and to place them within historical context, these results are in much closer agreement with experiment than the first reported QM/MM study of the CM enzyme [326]. This early work of Ref. [326] gave an activation barrier of 17.8 kcal mol^{-1} in enzyme at the AM1/CHARMM27 level of theory, giving somewhat of an overestimate of the reaction barrier. However, more recent work detailing a QM/MM investigation of the reaction in enzyme yielded an activation energy of 1.4 kcal mol^{-1} [338]. This study used the same initial X-ray structure as Ref. [320] and prepared it in the same way. The QM region comprised 24 atoms and was treated at the B3LYP/6-31G(d) level, again identical to the approach adopted in Ref. [320]. Where the investigations differ is in

the classical methods used to describe the MM region of the simulations. Ref. [338] treated 4117 atoms of the surrounding enzyme and water molecules with the AMBER 4.0 force field, whereas in Ref. [320] the CHARMM27 force field was used to describe 7053 atoms of the environment. Results such as those in Ref. [338] demonstrate the range of estimates available for QM/MM calculations. This underestimation of the experimentally observed barrier by an order of magnitude highlights the importance of careful path sampling and potentially indicates the differences in results that can be obtained from the use of differing parameterisation sets available in different force fields. This is only a whistle-stop tour of the notable CM investigations present in the literature but a broad survey of CM simulations can be found elsewhere [232, 327].

As outlined in Chapter 3, a central theme of this dissertation is to attempt to make simulation approaches and their associated results more accessible to non-specialists. This will allow computational enzymology to have impact in other scientific communities. I feel that the present exploratory work contained within this chapter also contributes to this aspiration as it removes the additional complexity inherent in the selection of force field parameters and of choosing a QM/MM boundary partitioning scheme. The motivation for treating atoms beyond that of the CM substrate with QM methods stems, in part, from mutagenesis experiments that demonstrate the significant role the Arg90 residue, illustrated in Figure 6.1, plays in catalysis within the CM enzyme [339]. The experimental findings from Ref. [339] corroborated predictions from prior theoretical work [326, 340] and agreed with the expected outcomes from previous experimental proposals [333, 341]. In addition to these investigations, computational studies have been shown to demonstrate the significant involvement of the Glu78 and Tyr108 sidechains [342] along with the Arg7 charged residue [343] within CM catalysis. Activation energies have been shown to change by just 1 kcal mol^{-1} upon including the charged residues Glu78 and Arg90, along with the substrate, in the QM region, compared to only including the substrate, at the PBE/DVZP/AMBER level [328]. However, within the same calculations, a much more significant difference in reaction energy of 5.7 kcal mol^{-1} was observed. This result not only displays the difficulties involved in calculating a converged reaction energy within this reaction but it also indicates that a larger QM region will be needed at that level of theory in order to converge the calculated reaction energy. Furthermore, studies at the AM1/CHARMM27 level of theory that included the same additional charged residues as Ref. [328], found a change of 3.1 kcal mol^{-1} in the activation energy [343]. It has been proposed that polarisation in the neighbouring charged residues and the associated charge transfer from the substrate to the active site may be important for catalytic activity in CM [342, 343]. As the fixed charge approximation is generally assumed within force field approaches, and QM/MM may encounter the problem of electron leakage, as discussed in Chapter 3, the behaviour proposed in Refs. [342] and [343] may not be accurately described by the previous approaches used by other authors. In addition, coupled-cluster calculations performed in Ref. [329] on the active site of CM have demonstrated that by

increasing the total size of the system from only the substrate to also include 4 active-site residues – namely Arg7, Arg63, Glu78 and Arg90 – surrounding the substrate, changes the activation barrier by around 0.7 kcal mol^{-1}. However, as this is the largest system size accessible with coupled-cluster approaches, the study provides a very limited test of convergence of energies with respect to the size of the QM region. Ultimately, it is unclear whether the computed value of the barrier will continue to change upon the addition of further active-site residues.

Chapter 3 discusses the fact that conventional QM methods incur a computational cost that typically increases as the third, or greater, power of the number of atoms in the system. Nevertheless, an increasingly viable alternative approach to QM/MM schemes is to perform QM calculations on a significant portion of an enzyme. Previous authors have taken CM enzyme structures, which have been optimised at the RHF/6-31+G(d,p)/AMBER level of theory, and have performed single-point energy calculations at the all-electron quantum chemical level, using the fragment molecular orbital (FMO) method [344]. The Effective FMO (EFMO) method has also been used to investigate CM [345], yielding averaged enthalpies of activation and reaction. The activation barrier overestimates experiment by 5.5 kcal mol^{-1}. The energy of reaction was found to be strongly basis set-dependent, varying from -5.5 to a positive value of 0.8 kcal mol^{-1}, in contrast with many predictions from other levels of theory of a very exothermic reaction. Within the EFMO approach, an active region is defined as the active site of the enzyme, in a similar manner to that of QM/MM simulations. Ref. [345] reports a doubling in computational costs upon increasing the number of atoms in the EFMO active region from 129 to 241. Following this increase, the activation barrier changes only by 0.2 kcal mol^{-1} but the reaction energy changes by 2.3 kcal mol^{-1}. In contrast to QM/MM calculations, within EFMO simulations the atoms outside the active region remain frozen. The electrons of the fixed fragments are kept in place by using frozen orbitals across any bonds to the active region. The system in Ref. [345] was prepared using the approach outlined in Ref. [320], optimising the geometry of the transition state analog previously used in Ref. [323] to crystallise the enzyme, generating an initial conformation for the reactant state. However, Ref. [320] uses a fully flexible model for both the substrate and the enzyme allowing the entire protein to adjust, contrary to Ref. [345] where active fragments have been pre-chosen.

Whilst the authors of Ref. [345] claim their work can be better considered an approximation to a full-QM calculation, the work detailed in this chapter uses the linear-scaling DFT code ONETEP to perform completely quantum mechanical calculations on a CM fragment. It is through the use of ONETEP in the work in this chapter that I hope to avoid the inherent errors that can be encountered using hybrid methods discussed in this dissertation. By applying ONETEP, the entire enzyme fragment chosen can be treated at the same quantum mechanical level of theory. The investigation presented in this chapter takes CM reaction pathways that have been previously optimised at the

B3LYP/6-31G(d)/CHARMM27 level in Ref. [320]. From these pathways, protein fragments ranging up to 1999 atoms have been extracted. Each protein fragment has a well-defined optimisation region, centred on the substrate, which is structurally optimised in ONETEP whilst the remainder of the fragment is kept fixed. Reaction energies are calculated as the total energy difference between the optimised reactant state and product state configurations, following geometry optimisation. Activation energies are calculated as the total energy difference between the optimised reactant state and transition state conformations, following transition state searching in the LST/QST formalism, using the optimised reactant and product state structures as end-point conformations. Following an exhaustive literature search I am confident that the work in the present chapter is the only reported study taking optimum CM structures from a QM/MM level of theory and re-optimising with full-DFT, thus requiring no further input from classical approximations to generate a QM-only transition state conformation and associated activation barrier.

The next section outlines the preparation process applied to the systems considered, along with their associated optimisation procedures. The results are presented through Sections 6.3 to 6.7. Further synoptic analysis and a discussion of the results is presented in Section 6.8 and the chapter is brought to a close in Section 6.9.

6.2 General preparation and optimisation of systems

In this chapter, CM structures in their reactant and product state configurations have been extracted from QM/MM pathways optimised in Ref. [320]. These minimum energy stationary point conformations have then been re-optimised using density-functional theory in ONETEP. To briefly outline the protocol used to generate the QM/MM pathways presented in Ref. [320], Claeyssens and co-workers took the CM structure reported in Ref. [324] with Protein Data Bank ID 2CHT, which, in its crystal structure, has a transition state analogue bound to the enzyme active site [323]. The chorismate substrate was then optimised separately in the gas phase at the RHF/6-31G(d) level. This molecule was then used to replace the transition state analogue bound to the active site [327,346]. Multiple structures were then generated through semi-empirical QM/MM molecular dynamics at the SCCDFTB/CHARMM22 level. This was achieved by constraining the substrate to be close to the transition state, as defined by a reaction coordinate described by the difference in C-O bond breaking and C-C bond forming distance. The substrate was treated with B3LYP/6-31G(d) and the surroundings for both environments were treated with the CHARMM27 and TIP3P force fields. Following initial equilibration purely based on molecular dynamics, reaction pathways were generated via restrained optimisation for fixed reaction coordinate moving from the transition state forward in the reaction coordinate to the product state conformation and backward to the reactant state configuration. This process yielded 28 snapshots for each enzyme pathway and 30 snapshots for each pathway in water. In total, 16 pathways were generated in enzyme by this process and 24 in water.

It is clear from the literature that molecular dynamics has been shown to be a very useful tool for providing initial structures for the study of enzymatic reactions [347–349], so the initial CM structures should be reliable. A much more detailed description of the preparation of the system and the protocol followed for the QM/MM simulations is available elsewhere [327, 346, 350, 351].

6.2.1 Specific preparation of the enzyme system

An initial spherical cluster, that was centred on the substrate, was extracted from one of the QM/MM pathways, detailed in Ref. [320], in the reactant state configuration. This extract of the total system contained the substrate, the 57 nearest residues to the substrate and the 41 closest water molecules to the substrate. In total the structure contained 999 atoms and any terminated peptide bonds were protonated accordingly, using the Open Babel software package [352]. Within this protein fragment, an optimisation region was defined to comprise the substrate, the four nearest water molecules and three nearest active-site residues, namely Arg7, Glu78, and Arg90, which are illustrated in Figure 6.1. All other residues and water molecules were defined to comprise the outer, or fixed, region. From the same pathway, all of the same residues and water molecules were extracted in their product state conformation. The residues from the product outer region were then replaced by the residues from the reactant outer region. The individual reactant and product optimisation regions were then structurally optimised whilst the outer region remained fixed. This type of optimisation scheme ensures that the calculated total energy differences are directly attributable to any local changes in the active site whilst accounting for the long-ranged polarisation and steric constraints of the surrounding protein scaffold of the enzyme. This model assumes that there are no significant changes in the structure of the outer region when moving from the reactant to product conformation of the enzyme. These assumptions agree with the experimental observation that there are no large-scale changes in the enzyme conformation during the course of the reaction [324,336,337]. These assumptions are also consistent with the QM/MM approach used in Ref. [320] whereby the outer 5 Å of the outer structure is frozen with all other atoms free to move. As a result there are no significant structural changes in the outer region of the system when comparing the reactant and product state enzyme conformations along the pathways from QM/MM. Transition state searching has also been performed using the ONETEP-optimised reactant and product state structures as end point conformations. At this point it must be emphasised that no information regarding the transition state structure was taken from the QM/MM calculations.

The aim of Chapter 4 of this dissertation was to demonstrate that validation is an important aspect of any computational investigation. It is of vital importance to ensure that energies of reaction and activation are converged with respect to the size of the atomic region undergoing structural optimisation. Therefore for the 57-residue system, the optimisation region was also increased to include the substrate, the four closest water

molecules and the nine nearest active-site residues, namely Arg7, Phe57, Ala59, Arg63, Cys75, Glu78, Arg90, Tyr108 and Arg116, which are illustrated in Figure 6.1. A larger structure in the same conformation, from the same pathway, was also extracted. This structure in total comprises the substrate, the 99 nearest residues and the 129 nearest water molecules to the substrate. This fragment contains 1999 atoms and can be seen in Figure 6.2. The same three-residue optimisation region as is contained within the

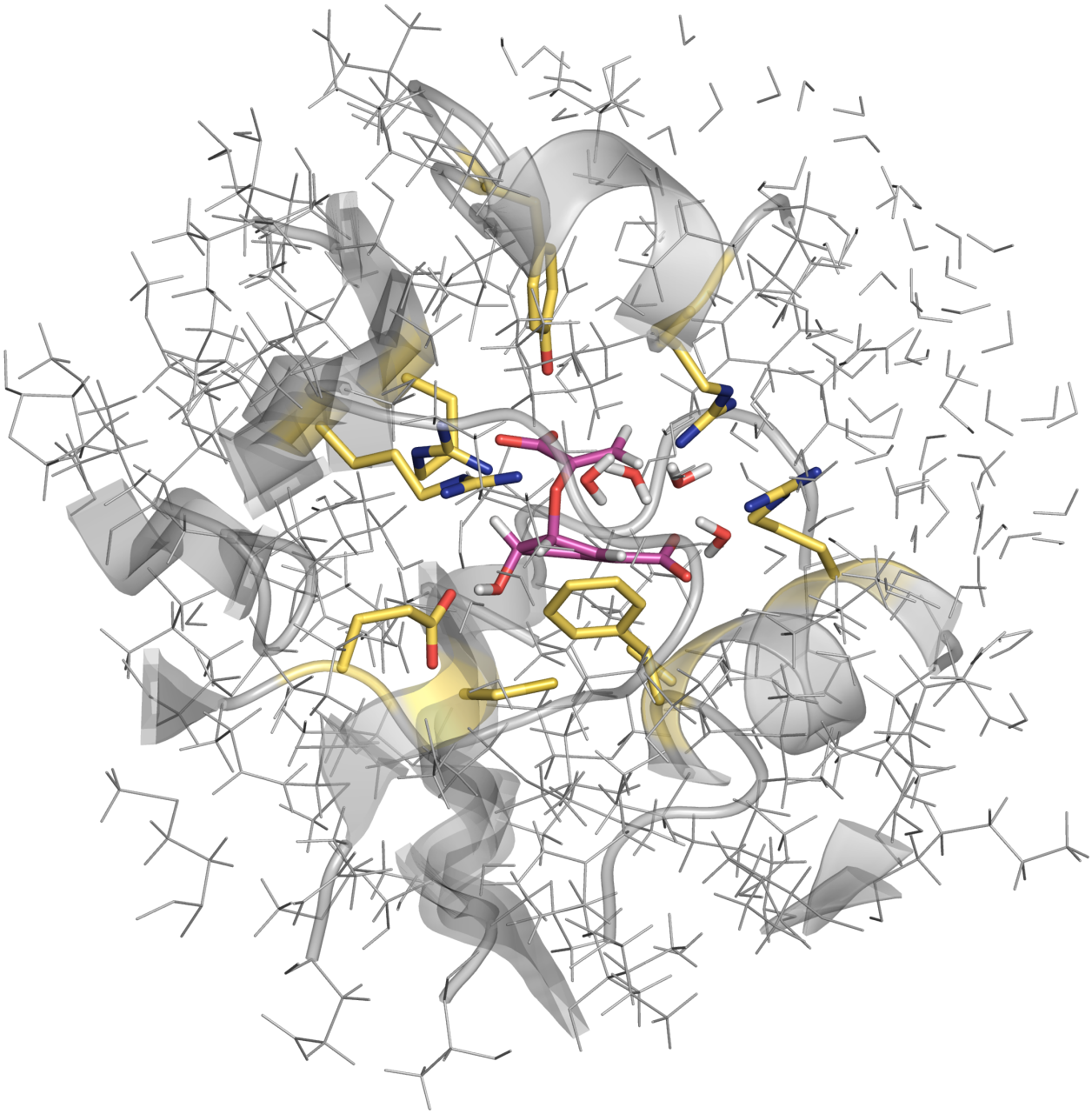


Figure 6.2: Exemplar 1999-atom CM fragment. The substrate, nine active-site residues and four water molecules are shown in colour and the remainder of the residues and water molecules are shown in grey.

57-residue system was allowed to optimise for the 99-residue fragment.

The work carried out in Ref. [292] reports vanishing HOMO-LUMO gaps for systems such as proteins and even water clusters, leading to poor convergence of the self-consistent electronic structure optimisation procedures. This resultant occupation of the lowest unoccupied molecular orbitals is unphysical as proteins should generally display insulating behaviour with large HOMO-LUMO gaps. However, in Chapter 5, it was shown that vanishing HOMO-LUMO gaps are the result of large surface dipoles being created across an extracted water or protein cluster surrounded by vacuum. In addition, the work presented within that chapter demonstrated practical solutions for reducing the dipole moment, recovering the HOMO-LUMO gap in these systems through means of classical minimisation, implicit solvation and/or the use of embedded classical point charges. Chapter 5 also showed that classical minimisation performed in solution prior to electronic structure calculations will allow the system to retain any major structural features obtained during a fully solvated molecular dynamics simulation and reduce any electrostatic artefacts that may arise from performing the subsequent electronic structure calculations *in vacuo*. Following on from the work in Chapter 5, that showed mutating charged residues – on the surface of the extracted protein, to alanine, helped to significantly increase the calculated gap of the 1FDF protein, potentially dangling charged residues on the outer shell of the CM structure were identified as leading to a vanishing HOMO-LUMO gap if not correctly treated. Ordinarily, these residues would form salt bridges or would be solvent-exposed in their real environment but in the model presented here, they are vacuum-exposed. The specific residues, within this category in the present studies, were His58 and Glu19 in the 57-residue fragment and Lys24, Lys185, Arg203 and Glu208 in the 99-residue fragment. The best course of action was therefore to mutate the Glu19 to Ala and de-protonate the His58 in the 57-residue fragment. Similarly, with the 99-residue fragment the four vacuum-exposed charged residues were mutated to alanine and the dangling histidine residue was de-protonated, thereby neutralising all vacuum-exposed charged residues. The overall effect of these changes was to reduce the number of protein states close to the Fermi level, ultimately improving the convergence of the electronic structure optimisation procedures, and ensuring a significant HOMO-LUMO gap was maintained for the system. Following these mutations, the resulting improvements seen in the occupation eigenvalues surrounding the Fermi Energy are in agreement with the findings of other authors who implemented similar approaches [99].

First turning attention to the 57-residue system, five initial configurations were taken from the optimised pathways calculated in Ref. [320]. Starting the optimisation procedure from a range of QM/MM reactant and product state configurations ensures that the DFT-optimised structures give a representative sample of the reactive conformations, and their associated enthalpies of activation and reaction, found in the enzyme system at room temperature. Therefore, building on the robust approach presented in Ref. [320] where multiple pathways have been investigated for the reaction in the CM enzyme, the average

reaction and activation energy will need to be calculated in order to take into consideration local minima present along the pathway. Once optimised structures for the reactant and product state conformations were generated and their associated reaction energies and forces were shown to converge, those structures were then used as the starting point for transition state searching performed using the linear and quadratic synchronous transit approaches described in Chapter 3. The energies calculated in this investigation do not include entropy, but as the experimental values available include enthalpies of activation and reaction, combined with the fact that entropic effects in the enzyme are relatively small and are not considered to be significant [333], this is a valid comparison to make. Within the ONETEP calculations presented in this chapter, the PBE functional, including dispersion corrections, is used to describe the entire system at the full-DFT level. With regards to the specific parameters used for the simulations performed in ONETEP, the NGWFs have an equivalent plane wave energy cutoff of 1020 eV, corresponding to a psinc grid spacing of 0.45 a_0. An increase in NGWF radii from 5.3 Å to 6.4 Å led to a change in reaction energy of 0.3 kcal mol^{-1} for the chorismate to prephenate reaction in the largest protein system discussed here. Repeating the calculation with an energy cut-off of 1290eV changed the reaction energy by 0.1 kcal mol^{-1}.

QM/MM calculations on the optimised enzyme pathways detailed in Ref. [320] have been repeated, in order to compare the results from the large-scale density-functional calculations, performed in this chapter, directly with QM/MM. These new QM/MM calculations make use of the PBE density functional with empirical dispersion corrections based on the formulation by Grimme *et al.* [353] to describe the QM region. In comparison, Ref. [320] made use of the B3LYP density functional, which was not dispersion corrected, in order to calculate the QM parts of the calculation. It has been shown that the inclusion of dispersion within QM/MM calculations has a significant effect on B3LYP-calculated energies and geometries of transition states and encounter complexes, in the case of cytochrome P450, and is also argued to be important in modelling reactions catalysed by other enzymes [354,355]. The 6-31G(d) basis set used for the new set of QM/MM calculations is the same, along with the associated calculation protocol, as implemented within Ref. [320]. In comparison to these new QM/MM calculations, the difference with the full-DFT calculations lies in the basis set used. Therefore it may be instructive, at this point, to compare these basis sets more closely. The so-called 6-31G basis is a split-valence double-zeta basis set; the core orbital is a contracted Gaussian-type orbital made of 6 Gaussians, and the valence is described by two orbitals – one contracted Gaussian-type orbital made of 3 Gaussians, and one single Gaussian. 6-31G(d) is then a 6-31G basis set with added *d* polarisation functions on non-hydrogen atoms. Previous work in the CM literature [329], has shown that, for the reaction, the computed LCCSD(T0) barrier heights agree with the full CCSD(T) values at the basis-set limit to within 1kcal mol^{-1}. Here, the acronyms refer to coupled-cluster (CC) theory with single (S) and double (D) excitations combined with an (approximate (T0)) triples (T) correction and local

approximations (L). However, these calculations are prohibitively expensive and are not routinely used. For reasons of computational convenience, the 6-31G(d) basis set is used. With regards to basis set used in the current study, the psinc basis, that used within ONETEP, is a systematic basis set, the accuracy of which may be tuned with a single adjustable parameter. In addition, the activation and reaction energies for the system have been converged with respect to the spacing of the psinc grid. The ONETEP calculations describe the entire enzyme fragment at the full-DFT level. As a result, this yields a more accurate description of the surrounding active site and associated protein scaffold compared to the force field description in the QM/MM calculations presented here.

6.2.2 Specific preparation of system in solution

In a manner identical to that of the enzyme system preparation, an initial structure, centred on the substrate, was extracted from one of the QM/MM pathways in solution, in the reactant state configuration. The solution system comprises a total of 2025 atoms divided into three regions and a schematic representation of this can be seen in Figure 6.3. Within this total structure, the chorismate substrate and the 76 closest water molecules were chosen as the region that undergoes structural optimisation. The optimisation region is then surrounded by 199 water molecules that are fixed in their positions from QM/MM calculations but are now treated with QM. The remaining 392 water molecules are represented as classical TIP3P electrostatic point charges fixed in their positions from QM/MM calculations. This use of classical point charges aids in convergence of the ONETEP density kernel and is necessary to restore the HOMO-LUMO gap, as shown in Chapter 5. The chorismate substrate and the subsequent 76 closest water molecules from the same pathway in the product state configuration were also extracted and then surrounded by the RS fixed atoms and electrostatically embedded point charges. Adopting the principles outlined in Chapter 4, two additional systems were prepared to check the convergence of calculated properties. The first additional structure defines the substrate and the closest 123 water molecules to be those undergoing structural optimisation. 248 water molecules then surround this region, fixed in their QM/MM positions and treated with QM. The remaining 296 water molecules are treated as embedded classical point charges. The second additional structure defines the optimisation region as the substrate and the closest 170 water molecules. Surrounding this region are 303 water molecules fixed in QM/MM positions that are treated with QM. The 194 water molecules that remain are treated as TIP3P point charges. A total system size of 2025 atoms was maintained for each of the water spheres extracted in each of the three-region models used. Further QM/MM calculations were also performed on the water pathways calculated in Ref. [320]. These calculations again use the PBE functional with empirical dispersion corrections based on the Grimme formulation.

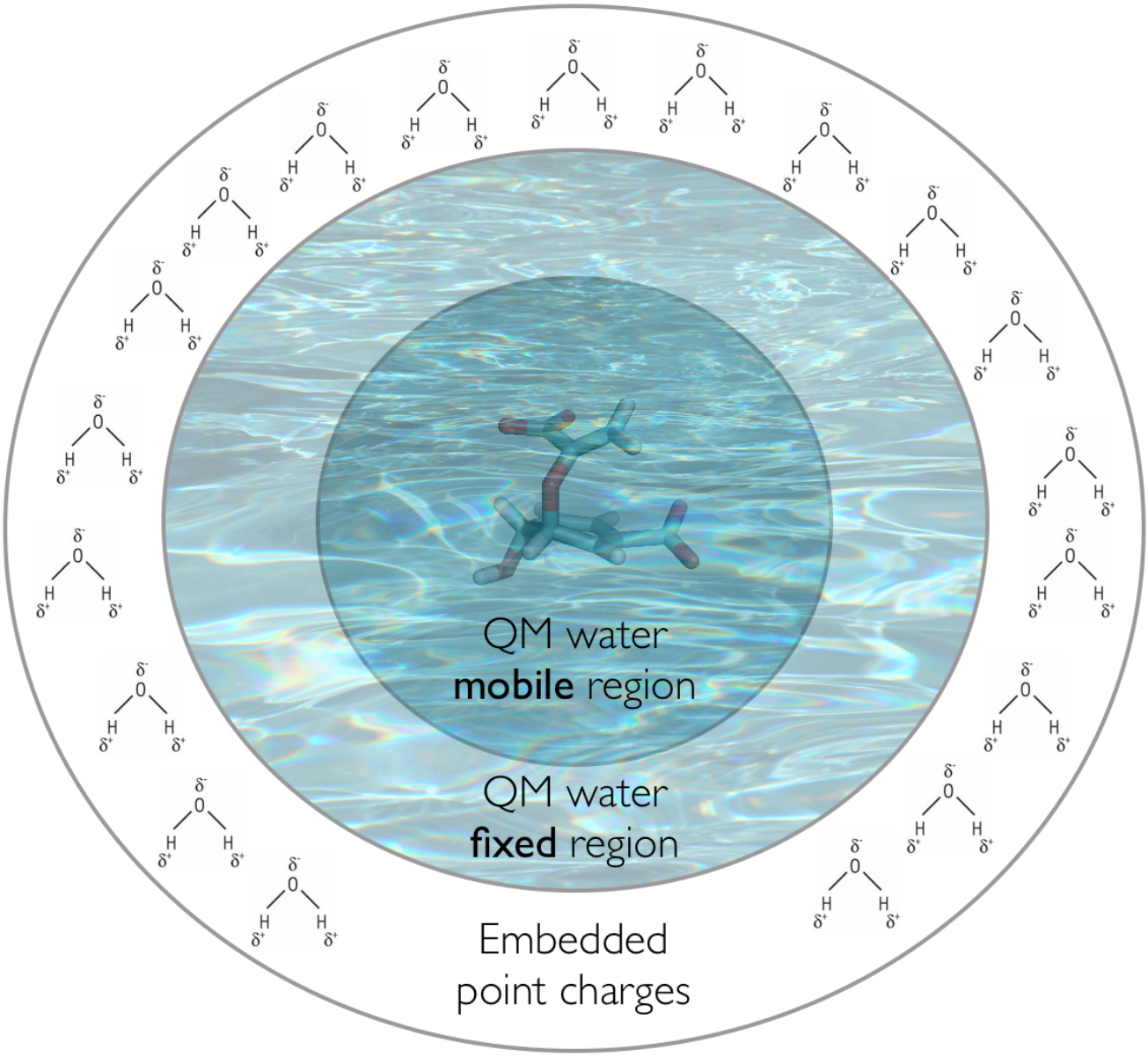


Figure 6.3: Schematic representation of the three-region optimisation model for the water systems considered.

6.3 Rearrangement in Enzyme

By treating the entirety of the protein fragment at the full-DFT level within ONETEP, the DFT-optimised structures and total energies of the reactant and product state of the 999-atom system can be computed. The resulting optimised reactant and product structures can be seen in Figure 6.1. In addition, the figure also shows the transition state conformation solved using first-principles transition state searching. These resultant structures are in excellent agreement with the corresponding conformation from the QM/MM-calculated pathways in Ref. [320], which are not shown here. The 999-atom protein fragment with an associated 98-atom optimisation region yields energies of activation and reaction of

13.4 and −7.7 kcal mol^{-1}, respectively. These can be seen in Table 6.1. The table also

# Mobile : Frozen Atoms	Energies / kcal mol^{-1}	
	$\Delta^{\ddagger}E_{\mathrm{tot}}$	ΔE_{tot}
98 : 901	13.4	-7.7
211 : 788	13.5	-8.0
98 : 1901	13.3	-7.9
98 : 901 (Implicit Solvation)	13.6	-8.2

Table 6.1: Energies of activation ($\Delta^{\ddagger}E_{\mathrm{tot}}$) and reaction ($\Delta E_{\mathrm{tot}}$) for increasing size of optimisation region and total fragment.

shows that the activation and reaction energies for the 999-atom system change by 0.1 and 0.3 kcal mol^{-1}, respectively, when the number of atoms in the optimisation region is increased from 98 to 211 atoms, thus increasing the total number of that are structurally optimised from three to nine. The resultant change of activation and reaction energies, after increasing the total size of the fragment from 999 to 1999 atoms whilst maintaining an optimisation region of 98 atoms, amount to 0.1 and 0.2 kcal mol^{-1}, respectively.

As discussed in Chapter 5, performing large-scale density-functional calculations with a cluster geometry *in vacuo* can, in some instances, lead to a large surface dipole moment. If incorrectly treated, this may then lead to poor convergence of the density kernel occupancies and may potentially have an effect on the energetics of the substrate in the centre of the extracted cluster. In order to test that this effect has been minimised here, additional implicit solvent calculations have been performed on the three optimised stationary point structures along the reaction pathway for the 999-atom system. The same implicit solvation approach has been used as in the previous chapter to screen any surface dipole moment and increase the HOMO-LUMO gap in problem cases of water clusters and protein fragments. Following this approach, the HOMO-LUMO gaps of the DFT-optimised protein clusters are always greater than 0.6 eV and increase by 0.3 eV upon including implicit solvent. Table 6.1 also reveals that through the use of implicit solvation on the optimised structures, the energies of activation and reaction are changed by just 0.2 and 0.5 kcal mol^{-1}, respectively. One should therefore be convinced that the calculated properties of interest are well and truly converged for the smallest cluster studied, comprising 999 atoms, with an associated optimisation region of 98 atoms. It should hopefully also be evident that the use of an implicit solvent model to additionally include the effects of the environment is not necessary for the systems in the present chapter.

In order to compute an average for the calculated energies of activation and reaction for the chorismate to prephenate rearrangement in CM, five pathways were selected from Ref. [320]. For each pathway, the structures of the end points were extracted, truncated to form the 999-atom cluster described previously and re-optimised using large-scale density-functional approaches with an associated optimisation region comprising 98 atoms. Table

6.2 compares the averaged values calculated using ONETEP with averages from QM/MM calculations and also experimental enthalpies of activation and reaction. The averaged energies of activation and reaction, calculated using ONETEP, are equal to 13.6 ± 1.3 and -7.8 ± 0.5 kcal mol^{-1}, respectively. Here the results are presented in the form of $\mu \pm \frac{\sigma}{\sqrt{n}}$, where μ is the sample mean, $\frac{\sigma}{\sqrt{n}}$ is the standard error of the mean, σ is the sample standard deviation and n is the sample size. Table 6.2 also shows that the calculated value for the activation energy of 13.6 ± 1.3 kcal mol^{-1} is in excellent agreement with both the experimental value of 12.7 ± 0.4 kcal mol^{-1} and the B3LYP/CHARMM27 calculations conducted by other authors. However, when comparing the averaged reaction energy, calculated with ONETEP, to that of QM/MM calculations in Ref. [320], it is evident that the averaged full-DFT values predict a significantly less exothermic reaction for the chorismate to prephenate rearrangement in the presence of CM. However, following an exhaustive literature search, no experimental information regarding this reaction energy in the presence of CM could be found.

In order to further elucidate the effect on the calculated energies of treating an entire enzyme fragment with QM, additional QM/MM calculations using PBE+D/CHARMM27 have been performed. As a result, the exchange-correlation functional used to describe the QM region is directly comparable with that used in ONETEP. However, it is important to note that the 6-31G(d) basis set used in the QM/MM calculations is less accurate than the ONETEP psinc basis approach, which has been shown to approach the complete basis set limit [124]. Table 6.2 reveals that the calculated activation energy is significantly underestimated by PBE+D/CHARMM27, which can be ascribed directly to inaccuracies in the molecular mechanics force field and the QM basis set. The underestimation of activation barriers using a PBE functional within QM/MM calculations has been previously reported in the literature [328, 356]. The calculated reaction energy in the QM/MM approach is again significantly more exothermic compared to that predicted by calculations performed in ONETEP. It has also been shown in Ref. [328] that the use of a PBE functional in a QM/MM calculation can overestimate the energy of reaction. However, once again, the lack of experimental reaction enthalpy data precludes any conclusions from being drawn from this particular result. The above results should, however, instil confidence that all

	Energies / kcal mol^{-1}	
	$\Delta^{\ddagger}E_{\text{tot}}$	ΔE_{tot}
ONETEP	13.6 ± 1.3	-7.8 ± 0.5
B3LYP/CHARMM27 [320]	11.3 ± 1.8	-18.2 ± 1.3
PBE+D/CHARMM27	7.5 ± 0.4	-19.7 ± 0.5
Experiment [333]	12.7 ± 0.4	–

Table 6.2: Comparison of energies of activation ($\Delta^{\ddagger}E_{\text{tot}}$) and reaction ($\Delta E_{\text{tot}}$), from the literature, to those in the present work.

other remaining discrepancies between ONETEP calculations and experiment or QM/MM are due to the approximations within full-DFT such as the PBE exchange-correlation functional, the pseudopotential approach and the use of the underlying psinc basis set.

In order to examine the differences between the full-QM non-bonded interactions and the QM/MM approximations, Table 6.3 compares the interaction energies between active site and substrate for the optimised enzyme systems with results from the literature and also with additional QM/MM simulations carried out for the present chapter. Within

	Relative Energies / kcal mol^{-1}					
	$\Delta^{\ddagger}E_{\text{int}}$	ΔE_{int}	$\Delta^{\ddagger}E_{\text{env}}$	ΔE_{env}	$\Delta^{\ddagger}E_{\text{sub}}$	ΔE_{sub}
ONETEP	-2.1 ± 0.3	12.7 ± 0.9	0.6 ± 0.2	-0.5 ± 0.1	15.1 ± 1.2	-20.0 ± 1.2
B3LYP/CHARMM27 [320]	-7.3 ± 2.0	7.0 ± 2.5	–	–	–	–
PBE+D/CHARMM27	-8.5 ± 0.8	7.5 ± 0.5	2.8 ± 0.2	-3.8 ± 0.5	13.2 ± 0.3	-23.4 ± 0.3

Table 6.3: Comparison of relative interaction energies in enzyme, from the literature, to those in the present work. The energies of interaction are defined relative to the RS, measured at the TS ($\Delta^{\ddagger}E_{\text{int}}$) and the PS ($\Delta E_{\text{int}}$). Also shown are the components from the environment ($\Delta^{(\ddagger)}E_{\text{env}}$) and from the substrate ($\Delta^{(\ddagger)}E_{\text{sub}}$) as defined in equation (6.1).

this chapter the interaction energy is defined as:

$$E_{\text{int}} = E_{\text{tot}} - (E_{\text{sub}} + E_{\text{env}}) \tag{6.1}$$

where E_{int} is calculated as the total energy of the whole system E_{tot} less the components of energy of the substrate E_{sub} and that of the environment E_{env}, which is the reorganisation energy due to the enzyme in the protein environment or due to the water molecules in the solution environment. The latter two terms are single-point energies from calculations on sub-systems extracted from the optimised structures. No further re-optimisation is used to calculate these interaction terms. The interaction energy should ideally be stabilising at the transition state due to favourable Coulombic interactions between the dianionic substrate and the surrounding positively charged active-site residues.

From Table 6.4 it can be seen that all interaction energies, calculated using full-DFT, are converged to within 0.5 kcal mol^{-1} with respect to the size of the optimisation region and the total size of fragment simulated. The calculated interaction energies from

# Mobile : Frozen Atoms	Relative Energies / kcal mol^{-1}					
	$\Delta^{\ddagger}E_{\text{int}}$	ΔE_{int}	$\Delta^{\ddagger}E_{\text{env}}$	ΔE_{env}	$\Delta^{\ddagger}E_{\text{sub}}$	ΔE_{sub}
98 : 901	-2.1 ± 0.3	12.7 ± 0.9	0.6 ± 0.2	-0.5 ± 0.1	15.1 ± 1.2	-20.0 ± 1.2
211 : 788	-1.8	13.0	0.7	-0.7	14.6	-20.3
98 : 1901	-2.2	12.9	0.4	-0.5	15.1	-20.3

Table 6.4: Convergence of calculated interaction energies with respect to the size of the total fragment considered and the associated optimisation region.

QM/MM approaches at the transition state are an overestimate compared to those calculated using full-DFT within ONETEP. In QM/MM, the strain energy in the enzyme is calculated using an entirely classical approach and, as such, is expected to be strongly dependent on the accuracy of the force field being used. This can be understood from the present simulations, as the CHARMM27 force field also overestimates the changes in the enzyme strain energy, at both the transition and product states, when compared to a full-DFT approach. Table 6.3 shows that the cost in reorganisation energy for the enzyme to pass from the chorismate reactant to the transition state structure is 0.6 kcal mol^{-1} on average, which is around 30% of the gain in substrate-enzyme interaction energy. It is of note that the strain energy stored in the enzyme is less than 1 kcal mol^{-1} over the entire course of the reaction, which may be a favourable design feature in these astonishing natural catalysts. Table 6.3 shows that the $\Delta^{\ddagger}E_{\text{sub}}$ energy calculated with ONETEP and PBE+D/CHARMM27 are both significantly smaller than the activation energy calculated *in vacuo* of 29.7 kcal mol^{-1}.

6.4 Natural bond orbital analysis

As discussed in Chapter 3, although density-functional calculations can provide a very accurate description of a system in terms of its total electron density, problems can often arise when trying to decompose intermolecular interactions into chemically intuitive local quantities in order to generate a more qualitative description of the electronic behaviour. In the particular case of CM catalysis, it is the contributions of individual active-site residues to transition state stabilisation that are of interest. For this reason, a natural bond orbital (NBO) analysis has been performed on a subsystem comprising the substrate and surrounding in the optimisation region. Such an analysis allows the electron density to be re-described in terms of localised Lewis-type bonds and anti-bonds, along with lone pair orbitals. It is the delocalisation of electronic density from filled to vacant NBOs that causes a variational lowering of the total energy. This phenomenon is particularly important for hydrogen bonds. However, whilst it is important to keep in mind that no quantitative conclusions can realistically be drawn from just the charge transfer component, this has been shown previously to be strongly correlated with hydrogen-bonding strength in simple systems [174–176]. The three sets of NBOs that are estimated, via second order perturbation theory, to provide the strongest stabilisation energy to the substrate at the transition state are shown in Figure 6.4. As is expected for these types of interactions, each of them involves the delocalisation of electronic density between the substrate and neighbouring charged active-site residues. Specifically, they are all interactions from lone pairs (n) to anti-bonding (σ^*) orbitals.

With regards to the details of the interactions, the analysis of the substrate $n_{\text{O}} \rightarrow \sigma^*_{\text{NH}}$ interaction with the Arg90 charged residue reveals a favourable change in second-order perturbation energy in going from the reactant to transition state conformation $\Delta\Delta^{\ddagger}E^{(2)}$

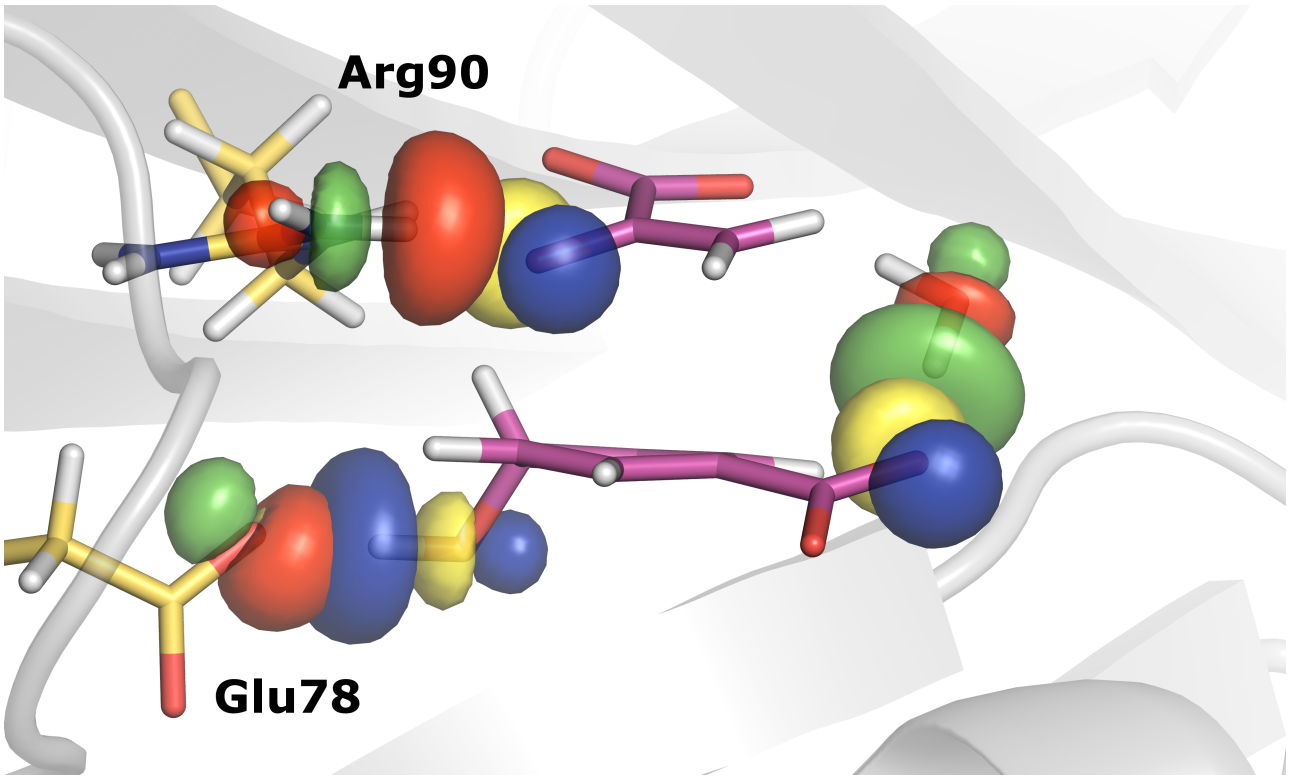


Figure 6.4: The three most stabilising NBO interactions at the transition state. The red/green isosurface represents positive/negative NBOs on the enzyme and the blue/yellow isosurface represents those on the substrate.

of -11.4 ± 3.1 kcal mol^{-1}. The same interaction shows an unfavourable change in going from the reactant to product state conformation $\Delta\Delta E^{(2)}$ of 7.2 ± 2.9 kcal mol^{-1}. This particular interaction indicates that the Arg90 charged residue is both helpful in stabilising the transition state and also in destabilising the product, leading toward the eventual unbinding and dissociation of the substrate. Focussing attention on the crystallographically observed water molecule, there exists an $n_{\mathrm{O}} \rightarrow \sigma^{*}_{\mathrm{OH}}$ interaction between this molecule and the substrate. This interaction gives $\Delta\Delta^{\ddagger}E^{(2)} = -8.6 \pm 0.8$ kcal mol^{-1}, stabilising the transition state, whilst a separate observed water molecule is involved in the same type of interaction giving $\Delta\Delta E^{(2)} = 3.4 \pm 0.9$ kcal mol^{-1}, destabilising the product. There also exists a substrate $n_{\mathrm{O}} \rightarrow \sigma^{*}_{\mathrm{NH}}$ interaction with the Arg7 charged residue, giving $\Delta\Delta^{\ddagger}E^{(2)} = -2.4 \pm 0.4$, stabilising the transition state combined with $\Delta\Delta E^{(2)} = 1.7 \pm 0.3$, destabilising the product. The NBO analysis shows that the charged Glu78 active-site residue stabilises the transition state with $\Delta\Delta^{\ddagger}E^{(2)} = -3.1 \pm 1.0$. Therefore, this analysis suggests that overall the active-site structure has evolved to an extent that the charged residues are accurately positioned to provide optimal orbital overlap with the substrate at the transition state, thereby strengthening the intermolecular interaction, as seen in Table 6.3, and ultimately lowering the activation energy barrier compared to the equivalent reaction in solution, which is discussed further in Section 6.6.

6.5 Structural analysis

Active Site Residue	H-N-H bond angle χ / °					
	Reactant		Transition State		Product	
	χ^{RS}_{ONETEP}	$\chi^{RS}_{QM/MM}$	χ^{TS}_{ONETEP}	$\chi^{TS}_{QM/MM}$	χ^{PS}_{ONETEP}	$\chi^{PS}_{QM/MM}$
Arg7	118.9 ± 0.1	121.6 ± 0.1	118.9 ± 0.1	121.5 ± 0.1	118.6 ± 0.1	121.2 ± 0.1
Arg90	118.1 ± 0.1	122.0 ± 0.1	118.0 ± 0.1	122.0 ± 0.2	117.6 ± 0.1	121.6 ± 0.1
Arg63	116.9	118.8	116.9	119.2	116.1	119.4

Table 6.5: Comparison of selected H-N-H bond angles of three arginine residues hydrogen-bonded with the substrate in the CM active site. The QM/MM values are from calculations presented in Ref. [320] and the ONETEP values are those calculated following structural optimisation. The Arg7 and Arg90 bond angles are averaged over the five pathways in which the residues were structurally optimised. The Arg63 residue was optimised in the system comprising 211 mobile atoms over a single selected pathway.

No information regarding the QM/MM structure of the transition state was used in the current LST/QST simulations performed with ONETEP. Despite this, there exists very little difference between the structures computed using full-DFT when compared with the original QM/MM structures. Indeed, the hydrogen-bonding network is identical in the two sets of structures, at all three stationary points of the reaction. There are, however, some more subtle deformations in the three charged arginine active-site residues that are hydrogen-bonded to the substrate. A structural analysis has been made of the H-N-H angles for the part of the guanidinium groups of each charged arginine active-site residue that is hydrogen-bonded with the substrate. The results of this analysis can be seen in Table 6.5. Following the analysis one can assert that, other than the fact that the H−N−H bond angles are consistently smaller in the structures calculated with ONETEP, compared to QM/MM, it is not particularly clear if the identified NBO interactions are correlated with any change in ionic structure. Therefore, selected dihedral angles, defined using the atomic positions highlighted in Figure 6.5, were also measured for the three arginine residues in the initial QM/MM structures and following optimisation in ONETEP. In all three cases it is clear that there is a significant distortion of the arginine guanidinium group away from planarity. This can be seen in Table 6.6. *Ab initio* simulations of Arg radicals have shown that the planarity of this guanidinium group can be affected by its charge state [357, 358]. In addition, it has been shown that the environment of the arginine residue can also affect its planarity [359]. Whilst the classical force field will indeed allow some flexibility of the dihedral angle, the particular arginine distortion identified in the ONETEP-optimised structures is not accurately treated in conventional QM/MM calculations as the residue is not contained within the QM region. Many authors have discussed the importance of the Arg90 residue in stabilising the transition state within

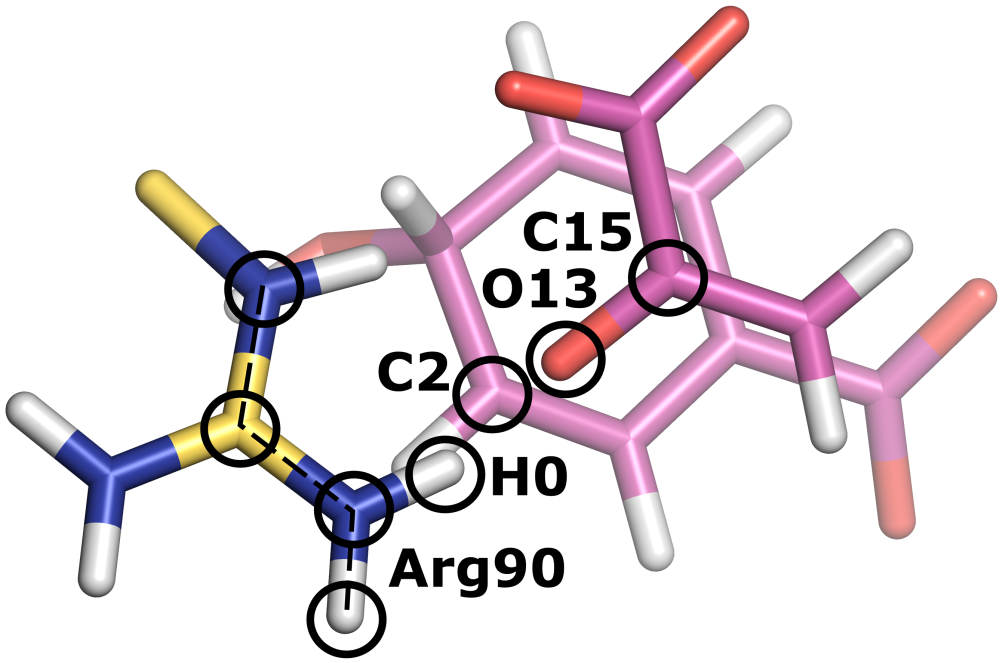


Figure 6.5: Atomic positions (connected by a dashed line) used in the definition of the dihedral angles presented in Table 6.6 for the three arginine residues hydrogen-bonded to the substrate, calculated before and after structural optimisation. The two carbon atoms highlighted, C2 and C15, display the most significant charge redistribution during the reaction in enzyme. A hydrogen-bond is formed between H0 of Arg90 and O13 of the substrate. The H−N−H angles shown in Table 6.5 are calculated with H0 and its neighbouring nitrogen and hydrogen atoms.

Active Site	Dihedral Angle ϕ / °					
Residue	Reactant		Transition State		Product	
	$\phi^{\mathrm{RS}}_{\mathrm{ONETEP}}$	$\phi^{\mathrm{RS}}_{\mathrm{QM/MM}}$	$\phi^{\mathrm{TS}}_{\mathrm{ONETEP}}$	$\phi^{\mathrm{TS}}_{\mathrm{QM/MM}}$	$\phi^{\mathrm{PS}}_{\mathrm{ONETEP}}$	$\phi^{\mathrm{PS}}_{\mathrm{QM/MM}}$
Arg7	-14.8 ± 1.7	-7.3 ± 0.6	-13.9 ± 1.5	-6.7 ± 0.7	-11.7 ± 1.5	$-5.7 \pm .6$
Arg90	-11.9 ± 1.2	-3.2 ± 0.7	-13.8 ± 1.3	-4.0 ± 0.6	-14.4 ± 1.5	-5.7 ± 0.5
Arg63	-13.7	-1.3	-16.7	-5.1	-18.5	-6.6

Table 6.6: Comparison of selected dihedral angles of three arginine residues hydrogen-bonded with the substrate in the CM active site. The QM/MM values are from calculations presented in Ref. [320] and the ONETEP values are those calculated following structural optimisation. The Arg7 and Arg90 dihedral angles are averaged over the five pathways in which the residues were structurally optimised. The Arg63 residue was optimised in the system comprising 211 mobile atoms over a single selected pathway.

the CM active site, yet its inclusion within a larger QM region, than is commonly used to treat this system, has not been reported. Therefore, the work presented in this chapter is the first observation of the resultant changes in dihedral angle for the Arg90 residue following structural optimisation at the full-DFT level.

6.6 Rearrangement in solution

A review on catalysis penned by Nobel laureate Arieh Warshel [360] states that in order to generate a more quantitative understanding of catalysis one must ask the question *"catalysis relative to what ?"* and it is almost immediately apparent how one can go about answering this. One must investigate the uncatalysed version of the reaction in water. As outlined in Section 6.2.2, the equivalent rearrangement of chorismate to prephenate in solution has also been investigated. It is shown in Table 6.7 that both the energies of

Mobile Atoms	Frozen Atoms	Point Charges	Energies / kcal mol^{-1}	
			$\Delta^{\ddagger}E_{\text{tot}}$	ΔE_{tot}
252	597	1176	29.8	−4.7
393	744	888	23.5	−10.0
534	909	582	23.9	−9.7

Table 6.7: Energies of activation ($\Delta^{\ddagger}E_{\text{tot}}$) and reaction ($\Delta E_{\text{tot}}$) for increasing size of optimisation region, along with number of frozen atoms and electrostatic point charges.

activation and of reaction are converged with respect to the size of the optimisation region used. However, it should be made clear that a much larger optimisation region, comprising more than 300 atoms, is required in the liquid phase compared with the simulation bound to the relatively structured enzyme. The best course of action, following these convergence tests, is to proceed with investigating the QM/MM pathways using a 534-atom optimisation region in order to make a comparison with experiment and additional QM/MM simulations. These findings are presented in Table 6.8. The averaged activation

	Energies / kcal mol^{-1}	
	$\Delta^{\ddagger}E_{\text{tot}}$	ΔE_{tot}
ONETEP	24.1 ± 1.1	-9.4 ± 2.2
B3LYP/CHARMM27 [320]	17.4 ± 1.9	-16.7 ± 2.2
PBE+D/CHARMM27	20.8 ± 3.9	-23.4 ± 0.9
Experiment [334]	20.71 ± 0.35	-13.2 ± 0.5

Table 6.8: Comparison of energies of activation ($\Delta^{\ddagger}E_{\text{tot}}$) and reaction ($\Delta E_{\text{tot}}$) in water, from the literature, to those in the present work.

	Relative Energies / kcal mol^{-1}					
	$\Delta^{\ddagger}E_{\text{int}}$	ΔE_{int}	$\Delta^{\ddagger}E_{\text{env}}$	ΔE_{env}	$\Delta^{\ddagger}E_{\text{sub}}$	ΔE_{sub}
ONETEP	-1.2 ± 1.2	14.0 ± 1.5	6.8 ± 0.7	-1.2 ± 0.5	18.5 ± 0.8	-22.2 ± 1.6
PBE+D/CHARMM27	-7.0 ± 2.4	-1.6 ± 3.2	4.9 ± 2.3	-6.5 ± 1.3	22.9 ± 3.5	-15.3 ± 1.6

Table 6.9: Comparison of interaction energies in solution, from PBE/MM, to those in the present work. The interaction energies are split into components relative to the transition state ($\Delta^{\ddagger}E_{\text{int}}$) and relative to the product state (ΔE_{int}) and comprise the components from the water environment ($\Delta^{(\ddagger)}E_{\text{env}}$) and from the substrate ($\Delta^{(\ddagger)}E_{\text{sub}}$).

and reaction energies are equal to 24.1 ± 1.1 and -9.4 ± 2.2 kcal mol^{-1}, respectively. These values are in agreement with experiment to within 4 kcal mol^{-1}. In a similar manner to the result in the enzyme, the activation barrier has again been overestimated. The calculated reaction energy in solution is similar to the value in enzyme of -7.8 kcal mol^{-1}. However, both of these values are underestimated with respect to the experimentally observed heat of reaction. It can be seen in Table 6.8 that the magnitude of the error, relative to experimental results, of the B3LYP/CHARMM27 calculations (-3.3 kcal mol^{-1}) is similar to the full-DFT approach in ONETEP (3.4 kcal mol^{-1}), although B3LYP/MM provides an underestimate and full-QM provides an overestimate. A PBE+D/CHARMM27 approach yields a good agreement with the experimentally observed activation energy, which is perhaps more likely to be a fortuitous result. However, the simulations still have their flaws as the averaged reaction energy is very much overestimated.

It is prudent to note at this point that the reaction conformation in water found in this chapter, which is based on the global minimum found in enzyme, is very likely to be one amongst many local minima present in the system in water. Previous authors have shown that the global minimum energy structure in the enzyme has a different conformation to the optimal structure found in solution [361]. It has been proposed that an associated free energy difference, estimated in Ref. [361] as 1.2 kcal mol^{-1}, is likely to contribute to the overall barrier to the non-enzymatic reaction in aqueous solution. Another exhaustive literature search could not find an associated correction to the enthalpy barrier but this fact is unlikely to significantly affect the conclusions drawn from the present calculations.

In an identical manner to the enzyme system, the total energies calculated for the water system have additionally been decomposed into interaction energy between the substrate and water, internal energy of the substrate and associated energy of the water environment. These decomposed energies can be seen in Table 6.9. The convergence of these component energies with respect to system size can be seen in Table 6.10. It can be seen that, again, the energy components are converged to within 0.5 kcal mol^{-1} with respect to the size of the optimisation region used in the simulation. As is expected for the water environment, which is less rigid than the protein system, the standard errors for

the individual components of the energy are much larger than those found in enzyme. It can now be seen that the discrepancy in calculated reaction energies between the full-DFT and QM/MM approaches, that both implement the PBE functional, is in fact dominated by the interaction energy between the substrate and the surrounding environment. This result could well be due to the electron leakage effect. Such an effect has been reported by other authors where the electron density is over-polarised by point charges [244, 248, 249]. There may also be charge transfer between the substrate and the surrounding water that can not be incorporated within existing classical force field approaches; this hypothesis is investigated in the following section.

6.7 DDEC and NPA charge analysis

The overall charge distributions of the enzyme-substrate complex and the solvated substrate have been investigated using both natural population and DDEC atoms-in-molecule analysis. The total charge on the substrate in the three stationary point conformations during the reaction, along with the associated charge redistribution, are given in Table 6.11. The results from both types of charge analysis show that the net charge on the substrate is, on average, 0.63 e and 0.79 e, lower in magnitude than the formal charge of -2 e assigned to the molecule, for the NPA and DDEC approaches, respectively. This is an indicator of significant charge transfer to the surroundings. This is likely to have implications for the contributions of the internal and the interaction energies to the relative interaction and activation energies, as demonstrated in Table 6.3. Although Table 6.11 shows the NPA charge analysis approach yields a slightly more negative total charge for the substrate, the redistribution of charge across the reaction is in fact very similar for both analyses, showing a consistent increase in negative charge on the substrate over the course of the reaction. Whilst one should be cautious to jump to conclusions based on the 0.013 e to 0.025 e net charge redistributions over the entirety of the substrate shown in Table 6.11, it is more instructive to look at the local atom-specific charge redistributions. Upon doing so, one can see that the most significant charge redistribution during the course of the reaction in the enzyme is found to be located on the C2 and C15 carbon atoms on the substrate, following the labelling convention in Ref. [320] and that shown

Mobile Atoms	Relative Energies / kcal mol^{-1}					
	$\Delta^{\ddagger}E_{\text{int}}$	ΔE_{int}	$\Delta^{\ddagger}E_{\text{env}}$	ΔE_{env}	$\Delta^{\ddagger}E_{\text{sub}}$	ΔE_{sub}
252	-0.8	17.8	12.1	-0.8	18.5	-21.7
393	-1.6	14.5	6.5	-1.9	18.6	-22.6
534	-1.1	14.1	6.7	-1.4	18.4	-22.1

Table 6.10: Convergence of energy components with regard to the size of the optimisation region used.

in Figure 6.5. The associated change in charge for C2 is -0.24 ± 0.003 e. The change in charge for C15 is equal to 0.27 ± 0.012 e. Both of these results agree with the natural population analysis previously performed in Ref. [320]. The charge values from those analyses are corroborated by DDEC charges which reveal a similar change in charge over the course of the reaction of -0.29 ± 0.013 e for C2 and 0.28 ± 0.011 e for C15.

The analyses have also indicated that there is a redistribution of charge on the Arg90 active-site residue. There is a hydrogen-bond between the H0 and O13 atoms highlighted in Figure 6.5. The result of this is a charge redistribution on H0, relative to the reactant, equal to 0.02 ± 0.001 e at the transition state and 0.03 ± 0.001 e at the product. Whilst these values are an order of magnitude less than the observed charge redistributions for the C2 and C15 atoms on the substrate, following the charge analysis of surrounding active site and protein scaffold, the charge redistribution localised to the Arg90 hydrogen bond is an order of magnitude larger than any other redistribution over the course of the reaction. This result indicates that the enzyme is relatively unaffected by the reaction of the substrate. The net charge, derived from natural population analysis, for the substrate in water is shown in Table 6.12. The table reveals that the charge on the substrate is less negative than the equivalent charge in enzyme, thereby again indicating significant charge redistribution. For the substrate in solution, the net charge again becomes more negative during the course of the reaction, as was seen to occur in the equivalent enzyme reaction. The partial charges, for both the NPA and DDEC schemes, are converged to within 0.01 e with respect to the size of the optimisation region used, as can be seen in Table 6.13. Once again it is the C2 and C15 carbon atoms on the substrate that are found to have the most significant charge redistribution during the reaction, matching the prior natural population analysis from Ref. [320]. In short, overall the enzyme is relatively unaffected by the substrate reaction, as the charge changes are small. However, the charge changes are statistically significant (evidenced by the error bars) and the NPA and DDEC methods agree.

It is important to note at this point that one of the main aims of implementing the DDEC atoms-in-molecules scheme within ONETEP, as discussed by its main developers in Ref. [183], is to ultimately replace the partial charges in standard force field approximations with those derived from an optimised ground state electron density. Previous authors have incorporated the DDEC charges, derived from ONETEP, of three proteins

Method	Charge Redistribution / e		Total Charge on Substrate / e		
	$\Delta^{\ddagger} q_{\mathrm{sub}}$	Δq_{sub}	$q_{\mathrm{sub}}^{\mathrm{RS}}$	$q_{\mathrm{sub}}^{\mathrm{TS}}$	$q_{\mathrm{sub}}^{\mathrm{PS}}$
NPA	-0.014 ± 0.003	-0.025 ± 0.009	-1.36 ± 0.02	-1.37 ± 0.02	-1.39 ± 0.01
DDEC	-0.013 ± 0.005	-0.028 ± 0.008	-1.19 ± 0.01	-1.21 ± 0.01	-1.22 ± 0.01

Table 6.11: Charge redistribution ($\Delta^{(\ddagger)} q_{\mathrm{sub}}$) and total charge on the substrate (q_{sub}) in enzyme.

Method	Charge Redistribution / e		Total Charge on Substrate / e		
	$\Delta^{\ddagger}q_{\mathrm{sub}}$	Δq_{sub}	$q_{\mathrm{sub}}^{\mathrm{RS}}$	$q_{\mathrm{sub}}^{\mathrm{TS}}$	$q_{\mathrm{sub}}^{\mathrm{PS}}$
NPA	-0.022 ± 0.013	-0.058 ± 0.024	-1.16 ± 0.014	-1.18 ± 0.013	-1.22 ± 0.014
DDEC	-0.018 ± 0.026	-0.040 ± 0.023	-1.13 ± 0.014	-1.14 ± 0.006	-1.17 ± 0.010

Table 6.12: Charge redistribution ($\Delta^{(\ddagger)}q_{\mathrm{sub}}$) and total charge on the substrate (q_{sub}) in water.

into a classical force field and run molecular mechanics simulations to compute NMR order parameters and scalar couplings [183]. The bonded and Lennard-Jones parameters were taken directly from the AMBER ff99SB force field but the atom-centered point charges were replaced by the DDEC/ONETEP charges. DDEC AIM charges performed better than mean field force field charges in providing a suitable electrostatic environment that maintained protein stability throughout a 10 ns trajectory while remaining dynamically consistent with experimental observations. The study also compared the NMR scalar coupling which provides a measure of hydrogen bond dynamics within a protein. DDEC/ONETEP charges performed at least as well as AMBER charges, illustrating that backbone N-H and C=O bond polarisation is also suitably described by the DDEC electron density partitioning approach to charge derivation. The only difference between the simulation protocols was in the point charges and the improvement in the calculated order parameters was due to the inclusion of native state polarisation in their calculations As an exploratory step, the protein-specific charges derived from the optimised electronic density of the CM fragments considered in this chapter were incorporated within the CHARMM27 force field and QM/MM simulations were run in a PBE+D/CHARMM27+DDEC approach. It must be noted that this method still does not reduce error due to electron leakage. Due to the fact that the protein fragments were extracted from a larger protein cluster in QM/MM pathways presented in Ref. [320], and in order to reduce the complexity of incorporating the protein-specific charges – the DDEC partial charges were used for the

Method	Mobile	Charge Redistribution / e		Total Charge on Substrate / e		
	Atoms	$\Delta^{\ddagger}q_{\mathrm{sub}}$	Δq_{sub}	$q_{\mathrm{sub}}^{\mathrm{RS}}$	$q_{\mathrm{sub}}^{\mathrm{TS}}$	$q_{\mathrm{sub}}^{\mathrm{PS}}$
NPA	252	-0.028	-0.055	-1.10	-1.15	-1.21
	393	-0.025	-0.058	-1.18	-1.20	-1.22
	534	-0.021	-0.056	-1.18	-1.21	-1.24
DDEC	252	-0.019	-0.041	-1.06	-1.09	-1.11
	393	-0.020	-0.036	-1.08	-1.13	-1.15
	534	-0.017	-0.045	-1.11	-1.15	-1.16

Table 6.13: Convergence of charge redistribution of the substrate in water, with respect to the size of the optimisation region used.

inner 1999 atoms of the protein structure and the CHARMM27 values were used for the remainder of the enzyme. The resultant energy of activation and of reaction were 3.3 ± 0.2 and -8.3 ± 0.9 kcal mol^{-1}, respectively. Both the energy of reaction and the activation barrier have been significantly underestimated by this approach. This indicates that there exists a fundamental flaw in this particular method of incorporating the protein-specific charges. In principle, one could combine the DDEC method of charge derivation with a much more accurate QM method, perhaps toward the MP2 level, where affordable, to describe the substrate in a QM/MM scheme. One of the benefits of generating partial charges derived from a single DFT calculation of an entire biomolecule is that the environmental polarisation is naturally included. By only calculating the charges for the smaller extract and incorporating this alongside existing force field partial charges, with no adequate partitioning method, the error on the activation energy is very pronounced. One reason why the approach could have failed is the possibility that the rest of the force field will need re-parameterising. Whilst the results of Ref. [183] are one piece of evidence that the DDEC scheme behaves well with existing force fields, much more work needs to be done to investigate the scheme, and specifically its use within QM/MM simulations, and these initial results support this. Whilst the immediate next steps are beyond the scope of this chapter, investigations such as these should be the focus of future work.

6.8 Discussion

The work presented in this chapter comprises a benchmark study which compares the results from full-DFT calculations of a large section of the *Bacillus subtilis* chorismate mutase enzyme with state-of-the-art QM/MM simulations from Ref. [320] and also with experimental investigations from Refs. [333] and [334]. For the total system comprising 999 atoms, full-DFT calculations implementing structural optimisation and LST/QST transition state searching methods, that have been validated against hybrid eigenvector-following techniques, have yielded an energy of activation of 13.6 kcal mol^{-1}. This result is in very good agreement with both QM/MM investigations and experimental studies. However, this test of large-scale DFT does not include any explicit considerations of entropic effects and, as a result, comparisons have been made only to experimental enthalpies of activation and of reaction. It is important to note that while it is fortunate that a direct comparison with experimental enthalpies of activation and reaction can be made in the particular case of the CM enzyme, large-scale DFT calculations may be less applicable to reactions that are expected to be strongly entropy-dependent. The main focus of this work lies primarily on the retention of the reactant and product structures in local energy minima nearest to those which were taken from optimised QM/MM calculations.

The activation energy has been shown to converge upon increasing the number of active-site residues treated in the optimisation region from 3 to 9. In addition, the activation energy has been shown to be converged on increasing the total protein fragment

	$\Delta\Delta^{\ddagger}E_{\text{tot}}$	$\Delta\Delta^{\ddagger}E_{\text{int}}$	$\Delta\Delta^{\ddagger}E_{\text{env}}$	$\Delta\Delta^{\ddagger}E_{\text{sub}}$
ONETEP	-10.5	-0.9	-6.2	-3.4
Experiment [333, 334]	-8.0 ± 0.4	$--$	$--$	$--$
	$\Delta\Delta E_{\text{tot}}$	$\Delta\Delta E_{\text{int}}$	$\Delta\Delta E_{\text{env}}$	$\Delta\Delta E_{\text{sub}}$
ONETEP	1.6	-1.3	0.7	2.2

Table 6.14: Changes in total energies, along with their components (kcal mol^{-1}), in going from a water to protein environment.

size from 999 atoms to 1999 atoms. An implicit solvent model has also been used to demonstrate the robustness of this calculated energy of activation. As well as being converged with respect to the size of the optimisation region and the total size of the protein fragment considered, the activation energy of 13.6 kcal mol^{-1}, calculated here, is in good agreement with QM/MM calculations and with experimental studies. The calculated energy of reaction, from the present study, underestimates experimental values from investigations in water from Ref. [334]. A decomposition of the calculated energies into components comprising substrate, interaction and reorganisation has been performed. By treating all of these components at the same QM level of theory, additional insight into the mechanism of rate enhancement in CM is provided. This decomposition reveals that the use of classical force fields within the framework of QM/MM simulations results in an overestimation of the interaction terms between the substrate in the QM region and the associated active-site residues in the MM region. The method of embedding the systems in a set of electrostatic point charges has also been used to investigate the equivalent reaction in a water environment. Embedding TIP3P charges within the full-DFT system comprising 2025 atoms gives a calculated energy of activation of 24.1 kcal mol^{-1}. Upon comparison to experimental work in Ref. [334], this value has been shown to be an overestimate. The activation energy is, however, converged with respect to the size of optimisation region considered. An identical decomposition procedure was followed for the solution system to investigate the non-bonded interactions between the substrate and surrounding water environment. A direct comparison of the activation barrier and reaction energy from enzyme with their corresponding values in solution allows the catalytic rate enhancement in CM to be studied further. These calculated relative energy changes have been collected together in Table 6.14. Combining Tables 6.2 and 6.8, it can be seen that the calculated energy barrier for the reaction decreases from the uncatalysed value of 24.1 kcal mol^{-1} in water to the catalysed value of 13.6 kcal mol^{-1} in the presence of CM. This reduction of activation energy by 10.5 kcal mol^{-1} in the enzyme compares extremely favourably with the reduction in the heat of activation barrier determined experimentally in Refs. [333] and [334]. From the decomposition of the total energies it can clearly be seen that the most significant component in the reduction of the activation energy by CM ($\Delta\Delta^{\ddagger}E_{\text{tot}}$) arises due to the more favourable reorganisation energy for the reaction

in the enzyme compared to in solution. Comparing Tables 6.3 and 6.9 shows that the reorganisation energy value decreases from 6.8 kcal mol^{-1} in the water environment to 0.6 kcal mol^{-1} in the presence of CM, giving a relatively large and negative value for $\Delta\Delta^{\ddagger}E_{\mathrm{env}}$. This observation is consistent with the notion that the CM active-site residues are favourably orientated to interact strongly with the substrate in its transition state conformation, thereby not introducing any significant strain into the structure of the enzyme. In addition, the gas phase water-optimised substrate has an activation energy of 18.5 kcal mol^{-1} compared to the enzyme-optimised barrier of 15.1 kcal mol^{-1} and the *in vacuo* water-optimised reaction energy is -22.2 kcal mol^{-1} compared to the enzyme-optimised reaction energy of -20.0 kcal mol^{-1}. Therefore, a secondary contributor to the catalytic rate enhancement is the fact that the change in the internal energy of the substrate is lower in the enzyme than in solution. This effect is reminiscent of the so-called near-attack conformation theory of enzyme catalytic rate enhancement [362–364], but it only contributes around 30% of the total barrier lowering found in this investigation. However, the details of this debate are outside the scope of the present chapter. It has also been demonstrated from the *in vacuo*-optimised conformation that both the enzyme and solution environments stabilise the substrate at the transition state in comparison to the gas phase. Table 6.14 also shows the calculated changes to the heat of reaction in going from a solution to enzyme environment. Although there is no experimental data available to enable a comparison to be made, it is perhaps not surprising that the reaction energy is similar in the two environments as, in general, enzymes are able to catalyse a reaction thereby speeding up their rate, but they do not change the standard free energy change of the reaction overall [18]. In this case, the interaction and internal substrate energy differences are of similar magnitude, but opposite in sign.

In addition to the energetics of the reaction, an investigation into the natural bond orbitals of the system has been described in this chapter. Such an analysis allows one to break down the contributing factors to catalysis further still, so that the importance and catalytic significance of individual active-site residues can be discussed. The particular charged active-site residues that have been shown to be vital in stabilising the transition state at the substrate were Arg7, Glu78 and Arg90. It is these residues which, combined also with a crystallographically observed water molecule, are expected to be important in lowering the value of $\Delta^{\ddagger}E_{\mathrm{tot}}$ in the enzyme relative to the equivalent reaction in water. These particular active-site residues have also been shown to destabilise the substrate in its product state conformation, the result of which will eventually allow the substrate to dissociate and be released to allow it to continue along the shikimate catalytic pathway [321]. The most favourable NBO interaction with the substrate was shown to be with the Arg90 charged active-site residue. A significant deviation from planarity has been observed in the guanidinium group of Arg90, which has not previously been observed in QM/MM calculations in Ref. [320]. In addition, a shortening in length of its hydrogen bond with the substrate, in comparison to QM/MM calculations from Ref. [320], has been observed.

This prediction from full-DFT, of the significant contribution of Arg90 to CM catalysis, is in qualitative agreement with experiments implementing site-specific mutagenesis [339] and also matches theoretical predictions of previous authors [326, 327, 340, 346, 365] and proposals devised following experimental investigations [333, 341]. However, despite the clear presence of a discussion of the catalytic significance of Arg90 in the literature, the work presented in this chapter is the first report of the link between the residue's structure, following structural optimisation, and associated catalytic effect. This chapter has also shown that the Arg7 charged active-site residue demonstrates a favourable NBO interaction with the substrate, combined with a shortening of its hydrogen bond length with the substrate at the transition state. It has also been observed that, following structural optimisation, Arg7 displays significant changes in dihedral angle compared to its value in QM/MM. The last of the charged active-site residues predicted to have catalytic significance following optimisation using full-DFT is Glu78. The favourable NBO interaction between this residue and the substrate at the transition state is in qualitative agreement with experimental investigations demonstrating the importance of the residue in the CM-catalysed reaction [366].

6.9 Summary

Large-scale density-functional calculations are very much a necessity if a substrate is covalently bound to an enzyme active site and one wishes to avoid the additional complexities involved in partitioning individual QM and MM regions through chemical bonds. A recent review on linear-scaling methods in Ref. [12] outlines many areas where large-scale DFT calculations are expected to play an important role. However, the review ultimately concludes that the applications of such approaches are still rather limited. In addition, the outcome of this survey suggested that the accuracy and efficiency of the techniques involved still require further investigation, and that it is not obvious as to what quantities can be accurately calculated by large-scale DFT studies. Nevertheless, treating entire proteins with quantum mechanics is becoming more widespread [99, 120, 121, 367] and will potentially increase not only the accuracy, but also the range, of problems open to investigation in fields ranging from small molecule therapeutics to molecular biology, enzymology and biomimetics.

Despite the work of previous authors and the encouraging results from the exploratory investigation presented in this chapter, it is evident from the investigations of the previous and present chapter that the issues involved in the accurate treatment of enzyme systems can not be tackled with large-scale QM approaches alone. Instead, a combined strategy utilising the relative strengths of MM, QM/MM and large-scale QM methods is required. In the present chapter it has been shown that, using the capabilities of the ONETEP code, one can start from QM/MM calculated reactant and product state structures and accurately predict the correct transition state in what is a fairly straightforward reaction.

However, when water molecules are involved – as is the case here, a preliminary reaction path is needed such that a reasonable mapping from water positions in the reactant to their positions in the product state can be obtained. This can be achieved from less computationally expensive semi-empirical approaches, on the condition that the reaction path is qualitatively correct. Whilst care has been taken to ensure that all calculated energies are converged with respect to the total size of the system and the number of atoms undergoing structural optimisation, it is likely that the optimisation region in less rigid protein structures will need to be extended in order to ensure the convergence of elastic energy with system size. However, the process of deciphering the allosteric role of the protein scaffold within enzyme catalysis remains outside of the scope of the present chapter. In order to improve upon the current description of the QM optimisation and constrained regions, improved linear-scaling density functionals that contain a more rigorous treatment of electron exchange and correlation must be used. The improvement of such functionals is a current active area of research and the method developments presented in Ref [72] are expected to be the foundation of many key future improvements to the ONETEP code. Ultimately, such advancements should enable one to be able to run Hartree-Fock and hybrid-DFT calculations on systems of the size discussed in the present chapter. In addition to this, large-scale DFT+U approaches [99,368] and methods harnessing dynamical mean field theory [369] ought to be used when treating strongly correlated transition metals within systems of biological interest such as the active sites of organometallic enzymes.

The present chapter comprises a proof-of-principle demonstration of the power of linear-scaling density-functional methods applied to large-scale systems of biomolecular interest. This work has combined the methodological development that was presented in Chapter 5, and has applied the validation techniques that were discussed in Chapter 4, to a real-world reaction of biomolecular relevance taking place in water and also catalysed by the CM enzyme. One of the final aims of this dissertation is to be able to proceed from the work outlined here and to use large-scale DFT calculations to improve the *de novo* computational design of enzymes and also to allow a range of biomimetic design principles to be drawn from the biological catalysts that are seen in nature, in order to utilise their properties in advancing industrial catalytic processes along with biomedical applications.

Chapter 7

Concluding remarks

"Have no fear of perfection - you'll never reach it"

Salvador Dali (1904-1989)

The proof-of-principle investigations reported in this dissertation have demonstrated the ability of large-scale density-functional calculations, combined with molecular mechanics and hybrid quantum/classical approaches, to accurately predict the electronic structure of enzymes and the energetics of their associated catalysed reactions. With the advent of new and powerful linear-scaling methods opening up potential applications in the biosciences, new lessons need to be learned about how it is best to apply them. In this dissertation, it has been shown how to adequately prepare systems to enable large-scale electronic optimisation. In addition, it has been shown that by working in combination with hybrid quantum mechanics/molecular mechanics methods to perform sampling of the conformational space, reliable and accurate results can be obtained for systems of biomolecular relevance which both match well with experiment and also improve upon the description available at other levels of theory. Therefore the overarching theme of the dissertation is that there is not necessarily one catchall approach to biomolecular simulation but that in order to sufficiently sample the conformational space for, and accurately treat the electronic structure of, a system, a combined strategy of MM, QM/MM and full QM is needed.

7.1 Summary of dissertation

The first three chapters of this dissertation provided the historical, biological and computational background and laid out the reasons why there is much to be pursued in the field of computational enzymology. Chapter 4 demonstrated the ability of these computational methods to accurately treat small molecules that have biological relevance. Simulating ethene, dialanine and the chorismate to prephenate transformation, structural optimisation and analysis techniques were tested in ONETEP and `OPTIM` and compared with experimental observations. Overall, the investigations presented in the chapter provided

the necessary validation of the approaches discussed and gave confidence that the results generated by them later in the dissertation would be reliable, particularly that the transition state searching methods can be trusted. The chapter also detailed the first reported use of the LST/QST algorithm in ONETEP and the first time this has been combined with the powerful eigenvector-following techniques available in the interface with the efficient `OPTIM` code.

Chapter 5 provided a roadmap of system preparation to enable convergence of the electronic structure of inhomogeneous systems – a subject that had been causing increasing debate over the last few years. Recent findings from Ref. [292] that the calculated HOMO-LUMO gap in water and protein systems does vanish under particular system preparation conditions, were confirmed; this characteristic is unphysical. However, it was shown that unequilibrated vacuum/protein interfaces arising from using X-ray crystal structures taken straight from experimental repositories can exhibit strong molecular dipole moments, and that these are ultimately responsible for this phenomenon. The work in the chapter then demonstrated general practical solutions for restoring the gap in systems comprising up to 2386 atoms. One of the approaches involved mutating protein charged surface residues to alanine if they were identified as potentially causing a closure of the HOMO-LUMO gap. This opened up the calculated gap, from negligible to 1.3 eV, for the 1FDF protein. Other solutions to the gap closure problem included classical structural optimisation of the interfaces between water and vacuum, to the screening of molecular dipole moments through the implicit solvation of protein structures, combined with the use of embedded classical point charges. The implications for the remainder of the dissertation were also discussed.

Chapter 6 investigated the enzymatic mechanisms of the *Bacillus subtilis* chorismate mutase enzyme and the chorismate to prephenate rearrangement that it accelerates relative to the equivalent reaction in water. The work in the chapter demonstrated that by combining the powerful methods available for performing large-scale DFT with the robust methodologies and best practices developed during this dissertation, calculated values for energies of activation and reaction could be converged with respect to the size of the fragment extracted from the full protein. The calculated activation energy barrier was found to be lowered by 10.5 kcal mol^{-1} in the presence of the enzyme, compared to the uncatalysed reaction in solution, a result which is in good agreement with experiment. Furthermore, due to the full-DFT nature of the simulations, additional information can be obtained from these calculations that is previously not discussed in QM/MM investigations or studies using methods based purely on classical approaches such as molecular dynamics. The catalytic rate enhancement provided by the enzyme is attributed to strong overlap between orbitals on the substrate and several charged active-site residues which results in strong intermolecular hydrogen bonding at the transition state whilst inducing negligible strain in the enzyme. One specific example from the work is the observation linking the structure of the Arg90 residue in the CM active site with its catalytic effect.

The work presented in Chapter 6 is the first such investigation to do so, emphasising the benefits of large-scale QM simulations and demonstrating how they can complement additional molecular dynamics and quantum/classical simulations. In addition, this proof-of-principle demonstration of powerful large-scale DFT methods shows their relevance in studying systems of genuine biological interest and, I hope, will produce new insight into enzymatic principles from an atomistic perspective.

One of the key aims of this dissertation has been to establish a computational methodology, allowing the full quantum-mechanical treatment of systems in enzymology. A suggested three-point plan, encapsulating the procedures developed in this dissertation is discussed in the following:

1. **Initial starting coordinates from experiment (PDB) or theory**

   One should start with experimentally resolved coordinates that have been archived in a repository such as the Brookhaven National Laboratory Protein Data Bank (PDB) [312]. Ideally, the method of crystallisation should be of a form where the hydrogen atoms have been explicitly indicated such as in solution NMR, as with every structure studied in Chapter 5 apart from 1UBQ. However it is not feasible to expect solution NMR structures to be available for enzymes. Typically, this method of crystallisation is only applicable to relatively small protein structures. In the case of 1UBQ in Chapter 5 the structure was resolved using X-ray diffraction. In this case, a resolution of 1.8 Å or better should only be accepted, as is the case with 1UBQ. However, as is often the case, the hydrogen atoms may not be explicitly indicated and the resolution of the diffraction resolved structure may be worse than 1.8 Å . In such instances, then software packages, such as Molprobity [370], should be used to add hydrogens, assign tautomeric states and generally refine Xray structures or NMR ensembles before simulations should begin.

2. **Refine with hybrid quantum mechanics/molecular mechanics (QM/MM)**

   Minimisation techniques should then be used in order to adequately sample the configuration space of the system in a computationally less expensive, though less rigorous, manner, through the use of classical potentials. However, within many systems in enzymology it is likely that structures will contain some substrate or transition state analogue, used to crystallise the enzyme, that has not been parameterised for the particular classical potential being used. Therefore it may not always be feasible to perform sampling in a purely classical manner. In addition, classical potentials can not describe the cleavage and formation of electron bonds. This is a key feature of many enzyme-catalysed reactions that are important to study in enzymology. Therefore, one must use a level of QM in simulations that allow electron bond rearrangement. This can be achieved using the hybrid QM/MM approach that treats a user-defined region of the system with QM, allowing the electronic structure of the system to be accurately treated, whilst using a less rigorous classi-

cal potential to describe the protein matrix and surrounding solvent. In addition, techniques that combine classical sampling within a QM/MM framework, such as those presented in Ref. [320], are then advisable to use. By doing so, one can then generate different reactant structures, acting as starting points, that will allow multiple reaction pathways to be constructed. This will then allow temperature effects to be modelled, in an approximate manner, which are important to consider. Such conformational sampling is not feasibly done at the full-QM level due to the restriction of computational resources. Therefore this QM/MM stage is very difficult to avoid at the moment.

3. **Further refinement with full-quantum mechanics (QM)**

It important to ensure that the calculated properties of interest for the system are converged with respect to the QM region, which is difficult to do within QM/MM due to the multitude of different approaches for interfacing the QM and MM regions. The investigations discussed in this dissertation instead perform full-QM calculations on a significant portion of a system using DFT. In addition, for protein systems the accurate treatment of the solvent is crucial. Chapter 5 demonstrated that structural optimisation of water/vacuum and protein/vacuum interfaces using classical methods are required to prevent the closure of the HOMO-LUMO gaps of water clusters and protein molecules. However, one can only ensure these gap values are accurately maintained using full-QM approaches. Such approaches used in this dissertation have included the screening of molecular dipole moments through the use of implicit solvation, surrounding the system with explicit water layers and embedding the quantum mechanical system in the potential of classical point charges representing the water environment. In the case of significant net charge ($+3e$) in a protein, the work in Chapter 5 has shown that an implicit solvent will be necessary to accurately treat the system. Implicit solvation yields the closest agreement between the HOMO-LUMO gaps of large isolated explicit water clusters and that of bulk water. The work in Chapter 5 also demonstrated that the use of implicit solvation techniques restores larger HOMO-LUMO gaps for proteins to a greater extent than when 5 Å of the surrounding water molecules, retained from bulk periodic simulations, are explicitly simulated. Activation energies are often calculated in a QM/MM framework by proceeding from reactant to product by using some simply defined reaction coordinate. However, the work presented in Chapter 6 of this dissertation uses the more rigorous LST/QST algorithm to accurately calculate the full-QM transition state for an enzyme-catalysed reaction, both in water and in the presence of the enzyme. The DFT-predicted reduction in activation barrier from water to enzyme, of 10.5 kcal mol^{-1} is in good agreement with the experimentally observed reduction of 8.0 kcal mol^{-1}.

7.2 Suggestions for further work

A recurring theme throughout this dissertation is that there is not just one computational method that can be simply applied to biomolecular systems. Therefore the most effective way to proceed will be to ensure a robust strategy can be formulated to enable accurate and reliable investigations to be performed using multiple approaches, with each method complementing the others. As mentioned in Chapter 6, it has been shown that replacing standard AMBER ff99SB atomic partial charges with those calculated from large-scale DFT simulations incorporating the DDEC approach generates a force field that is better at replicating protein dynamics. This arises because of the error in the original force field description of the protein electrostatic potential. This is an artefact due mostly to the atomic partial charges for the force field being fitted to the electrostatic potential of small molecules, neglecting the important long-range electronic polarisation present in the protein. A future goal, leading out of the work presented in this dissertation, will be to extend upon the work in Ref. [183] and to treat CM in a force field where DDEC charges replace the standard atomic partial charges. Thereby, the active site will be treated with an accurate QM method and a DDEC-augmented force field will describe the surrounding environment.

Transition metals have a distinct presence at the reactive centres in molecules active in biological catalytic cycles [371]. Within such systems, mid-row $3d$ transition metals facilitate reactions as diverse as methane-to-methanol conversion at an antiferromagnetically coupled dimetal center [372,373], unactivated alkane halogenation by a high-energy, high-spin ferryl-oxo center in the SyrB2 halogenase [374], and oxygen binding at iron porphyrins in haemoglobin [375]. Therefore, the development of techniques that allow one to progress beyond DFT, employing large-scale DFT+U approaches [99, 368] and methods harnessing dynamical mean field theory [369], are of particular importance. In addition, a more rigorous treatment of electron exchange and correlation may improve the agreement between density-functional calculated activation barriers and experiment. The development of such functionals is a current topic of research [72] and a first step to test what will be a key future improvement to the ONETEP code will be to observe the dependence of calculated HOMO-LUMO gap on system size for larger proteins and water clusters than the ones studied in Chapter 5.

I strongly believe that one of the Grand Challenge areas in large-scale computational biology will be to understand how the electronic structure of biomolecular systems informs their function. Further, as the length scale on which accurate QM-optimised structures are calculated is increased, the aim will be to decipher the allosteric role of the protein scaffold within enzyme catalysis using these fast and accurate computational methods. The investigations carried out in this dissertation provide confidence and proof-of-concept results for studying large, biologically relevant systems from a DFT-based perspective. The continued understanding of biological systems at an atomic and electronic level will allow a detailed picture of the mechanisms of enzyme active sites to be constructed.

Such a picture will ultimately be of use in biomimetic approaches attempting to solve many important problems such as hydrogen storage and carbon capture. Exploratory computational work helping to inform on how to redesign hydrogen-abstracting enzymes toward alternative products is already underway [376], along with the application of first-principles catalyst design to carbon capture through biomimetic means [377, 378]. In addition, experimental investigations have recently shown the existence of an enzyme that efficiently hydrogenates carbon dioxide to produce liquid formate that can be safely transported and used as a high energy-density power source for hydrogen fuel cell devices [379]. As such, there remains still a lot to learn from biology as to how one can most efficiently develop catalytic solutions for some of the most challenging global problems. I feel that biomimetic first-principles based design will be a significant factor in the success of these solutions. The ability to predict the properties and function of an enzyme that has yet to be experimentally characterised and facilitating the design of new enzymes, leading to their subsequent synthesis – for both industrial and biomedical purposes, are significant goals in the field. It is the anticipation of achieving milestones such as these that shall fuel the continual development and application of computational techniques to problems in enzymology.

Bibliography

"Lesser artists borrow, great artists steal"

Igor Stravinsky (1882-1971)

[1] F. H. Westheimer and J. E. Mayer. *J. Chem. Phys.* **14**, 733 (1946).

[2] T. L. Hill. *J. Chem. Phys.* **14**, 465 (1946).

[3] J. Drostovsky, E. D. Hughes and C. K. Ingold. *J. Chem. Soc.* **173** (1946).

[4] N. L. Allinger, M. A. Miller, L. W. Chow, R. A. Ford and J. C. Graham. *J. Am. Chem. Soc.* **87**, 3430 (1965).

[5] B. J. Alder and T. E. Wainwright. *J. Chem. Phys.* **27**, 1208 (1957).

[6] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller and E. Teller. *J. Chem. Phys.* **21**, 1087 (1953).

[7] S. Lifson and A. Warshel. *J. Chem.Phys* **49**, 5116 (1968).

[8] M. Levitt and S. Lifson. *J. Mol. Biol.* **46**, 269 (1969).

[9] R. Car and M. Parrinello. *Phys. Rev. Lett.* **55**, 2471 (1985).

[10] W. Yang. *Phys. Rev. Lett.* **66**, 1438 (1991).

[11] S. Goedecker. *Rev. Mod. Phys.* **71**, 1085 (1999).

[12] D. R. Bowler and T. Miyazaki. *Rep. Prog. Phys.* **75**, 036503 (2012).

[13] A. Warshel and M. Levitt. *J. Mol. Biol.* **103**, 227 (1976).

[14] A. L. Bowman, I. M. G. and Mulholland, A. J. *Chem. Commun.* **37**, 4425 (2008).

[15] P. Nozières. *Annu. Rev. Condens. Matter Phys.* **3**, 1 (2012).

[16] T. Takano. *J. Mol. Biol.* **110**, 569 (1977).

[17] F. Sanger and H. Tuppy. *Biochem. J.* **49**, 463 (1951).

[18] B. Alberts, D. Bray, K. Hopkin, A. Johnson, J. Lewis, M. Raff, K. Roberts and P. Walter. *Essential Cell Biology, third edition.* Garland Science, (2010).

[19] Structural information (determined by microwave spectroscopy) from CRC Handbook of Chemistry and Physics, 88th edition.

[20] D. J. Wales. *Energy Landscapes.* Cambridge University Press, Cambridge, (2003).

[21] W. T. Astbury. *Trans. Faraday Soc.* **29**, 193 (1933).

[22] L. Pauling, R. B. Corey and H. R. Branson. *Proc. Natl. Acad. Sci. USA* **37**, 205 (1951).

[23] A. Bairoch. *Nucleic Acids Res.* **28**, 304 (2000).

[24] E. Fischer. *Ber. Chem. Ges. Berl.* **27**, 2985 (1894).

[25] H. Eyring and M. Polanyi. *Z. Phys. Chem. Abt. B* **12**, 279 (1931).

[26] M. G. Evans and M. Polanyi . *Trans. Faraday Soc.* **31**, 875 (1935).

[27] L. Pauling. *Am. Sci.* **36**, 50 (1948).

[28] M. M. Mader and P. A. Bartlett. *Chem. Rev.* **97**, 1281 (1997).

[29] R. A. Copeland, J. P. Davis, G. A. Cain, W. J. Pitts and R. L. Magolda. *Biochem.* **35**, 1270 (1996).

[30] V. L. Schramm. *Annu. Rev. Biochem.* **80**, 703 (2011).

[31] P. Kollman, B. Kuhn and M. Peräkylä. *J. Phys. Chem. B* **106**, 1537 (2002).

[32] G. Hou, G. Hou and Q. Cui. *J. Am. Chem. Soc.* **134**, 229 (2011).

[33] A. Fersht. *Enzyme Structure and Mechanism.* W. H. Freeman and Company, New York, (1985).

[34] A. Radzicka and R. Wolfenden. *Science* **267**, 90 (1995).

[35] N. J. Mulder, P. Kersey, M. Pruess and R. Apweiler. *Mol. Biotechnol.* **38**, 165 (2008).

[36] J. P. Overington, B. Al-Lazikani and A. L. Hopkins. *Nat. Rev. Drug Discov.* **5**, 993 (2006).

[37] A. L. Hopkins and C. R. Groom. *Nature Rev. Drug Discovery* **1**, 727 (2002).

[38] A. J. Mulholland. *Drug Discovery Today* **10**, 1393 (2005).

[39] M. Pirmohamed and B. K. Park. *Toxicology* **192**, 23 (2003).

[40] L. Ridder and A. J. Mulholland. *Curr. Top. Med. Chem.* **3**, 1241 (2003).

[41] M. Dantus, R. M. Bowman, J. S. Baskin and A. H. Zewail. *Chem. Phys. Lett.* **159**, 402 (1989).

[42] S. Pedersen, L. Bañares and A. H. Zewail. *J. Chem. Phys.* **97**, 8801 (1992).

[43] J. C. Polanyi and A. H. Zewail. *Acc. Chem. Res.* **28**, 1992 (1995).

[44] S. Scheiner and W. N. Lipscomb. *Proc. Natl. Acad. Sci. USA* **73**, 432 (1976).

[45] M. A. Cunningham and P. A. Bash. *Biochimie* **79**, 687 (1997).

[46] T. C. Bruice and K. Kahn. *Curr. Opin. Chem. Biol.* **4**, 540 (2000).

[47] R. Lonsdale, J. N. Harvey and Adrian J. Mulholland. *Chem. Soc. Rev.* **41**, 3025 (2012).

[48] J. Locke. *An Essay Concerning Human Understanding.* London, (1689).

[49] N. Lambert, Y.-N. Chen, Y.-C. Cheng, C.-M. Li, G.-Y. Chen and F. Nori. *Nature Physics* **9**, 10 (2013).

[50] E. Schrödinger. *Phys. Rev.* **28**, 1049 (1926).

[51] W. Heitler and F. London. *Z. Phys.* **44**, 455 (1927).

[52] F. Bloch. *Z. Phys.* **52**, 555 (1928).

[53] M. Born and J. R. Oppenheimer. *Ann. Physik* **84**, 457 (1927).

[54] L. H. Thomas. *Proc. Cambridge Phil. Soc.* **23**, 542 (1927).

[55] E. Fermi. *Rend. Accad. Naz. Lincei* **6**, 602 (1927).

[56] P. Hohenberg and W. Kohn. *Phys. Rev.* **136**, B864 (1964).

[57] M. Levy. *Phys. Rev. A* **26**, 1200 (1982).

[58] W. Kohn and L. J. Sham. *Phys. Rev.* **140**, A1133 (1965).

[59] J. Sun, J. P. Perdew and M. Seidl. *Phys. Rev. B* **81**, 085123 (2010).

[60] W. Pauli. *Z. Phys.* **31**, 765 (1925).

[61] W. Heisenberg. *Z. Phys.* **38**, 411 (1926).

[62] P. Dirac. *Proc. R. Soc. A* **112**, 661 (1926).

[63] R. O. Jones and O. Gunnarsson. *Rev. Mod. Phys.* **61**, 689 (1989).

[64] J. P. Perdew and M. Levy. *Phys. Rev. Lett.* **51**, 1884 (1983).

[65] L. J. Sham and M Schlüter. *Phys. Rev. Lett.* **51**, 1888 (1983).

[66] D. C. Langreth and M. J. Mehl. *Phys. Rev. Lett.* **47**, 446 (1981).

[67] E. K. U. Gross and R. M. Dreizler. *Z. Phys. A* **302**, 103 (1981).

[68] J. P. Perdew. *Phys. Rev. Lett.* **55**, 1665 (1985).

[69] C. Filippi, C. J. Umrigar and M. Taut. *J. Chem. Phys.* **100**, 1290 (1994).

[70] J. P. Perdew, K. Burke and M. Enzerhof. *Phys. Rev. Lett.* **77**, 3865 (1996).

[71] P. Mori-Sánchez, A. J. Cohen and W. Yang. *J. Phys. Chem. Lett.* **125**, 201102 (2006).

[72] J. Dziedzic, Q. Hill and C.-K. Skylaris. *J. Chem. Phys.* **139**, 214103 (2013).

[73] Y. Zhang and W. Yang. *J. Chem. Phys.* **109**, 2604 (1998).

[74] J. C. Slater. *Phys. Rev.* **36**, 57 (1930).

[75] P. Pulay. *Mol. Phys.* **17**, 197 (1969).

[76] T. H. Dunning. *J. Phys. Chem. A* **104**, 9062 (2000).

[77] W. H. E. Schwarz, E. M. van Wezenbeck, E. J. Baerends and J. G. Snijders. *J. Phys. B* **22**, 1515 (1989).

[78] H. Hellmann. *J. Chem. Phys.* **3**, 61 (1935).

[79] J. C. Slater. *Phys. Rev.* **51**, 846 (1937).

[80] C. Herring. *Phys. Rev.* **57**, 1169 (1940).

[81] J. Kohanoff. *Electronic Structure Calculations for Solids and Molecules: Theory and Computational Methods*. Cambridge University Press, (2006).

[82] J. C. Philips and L. Kleinman. *Phys. Rev.* **116**, 287 (1959).

[83] J. A. Appelbaum and D. R. Hamann. *Phys. Rev. B* **8**, 1777 (1973).

[84] D. R. Hamann, M. Schlüter and C. Chiang. *Phys. Rev. Lett.* **43**, 1494 (1979).

[85] N. D. M. Hine, P. D. Haynes, C.-K. Skylaris and M. C. Payne. *Comp. Phys. Commun.* **180**, 1041 (2009).

[86] V. Blum, R. Gehrke, F. Hanke, P. Havu, V. Havu, X. Ren, K. Reuter and M. Scheffler. *Comp. Phys. Comm.* **180**, 2175 (2009).

[87] A. Tkatchenko, M. Rossi, V. Blum, J. Ireta and M. Scheffler. *Phys. Rev. Lett.* **106**, 118102 (2011).

[88] I. S. Ufimtsev and T. J. Martínez. *Comp. Sci. Eng.* **10**, 26 (2008).

[89] Y. Hasegawa, J.-I. Iwata, M. Tsuji, D. Takahashi, A. Oshiyama, K. Minami, T. Boku, F. Shoji, A. Uno, M. Kurokawa, H. Inoue, I. Miyoshi and M. Yokokawa. *Proc. 2011 Int. Conf. for High Perf. Comp., Net., St. and An.* **1**, 1 (SC 2011).

[90] J. Iwata, D. Takahashi, A. Oshiyama, T. Boku, K. Shiraishi, S. Okada and K. Yabana. *J. Comp. Phys* **229**, 2339 (2010).

[91] P. J. de Pablo, F. Moreno-Herrero, J. Colchero, J. Gómez Herrero, P. Herrero, A. M. Baró, P. Ordejón, J. M. Soler and E. Artacho. *Phys. Rev. Lett.* **85**, 4992 (2000).

[92] T. Otsuka, T. Miyazaki, T. Ohno, D. R. Bowler and M. J. Gillan. *J. Phys.: Condens. Matter* **20**, 294201 (2008).

[93] T. Miyazaki, D. R. Bowler, M. J. Gillan, T. Otsuka and T. Ohno. *AIP Conf. Proc.* **1148**, 685 (2009).

[94] L. Bondesson, E. Rudberg, Y. Luo and P. Sałek. *J. Phys. Chem. B* **111**, 10320 (2007).

[95] T. Ozaki. *Phys. Rev. B* **82**, 075131 (2010).

[96] D. G. Fedorov and K. Kitaura. *J. Phys. Chem. A* **111**, 6904 (2010).

[97] T. Sawada, D. G. Fedorov and K. Kitaura. *J. Phys. Chem. B* **114**, 15700 (2010).

[98] C.-K. Skylaris, P. D. Haynes, A. A. Mostofi and M. C. Payne. *J. Chem. Phys.* **122**, 084119 (2005).

[99] Cole, D. J., O'Regan, D. D., and Payne, M. C. *J. Phys. Chem. Lett.* **3**, 1448 (2012).

[100] J. Dziedzic, S. J. Fox, T. Fox, C. S.Tautermann and C.-K. Skylaris. *Int. J. Quantum Chem.* (2012).

[101] D. R. Bowler, T. Miyazaki and M. J. Gillan. *J. Phys.: Condens. Matter* **14**, 2781 (2002).

[102] E. Rudberg, E. H. Rubensson and P. Sałek. *J. Chem. Theory Comput.* **7**, 340 (2011).

[103] E. Tsuchida. *J. Phys. Soc. Japan* **74**, 034708 (2007).

[104] N. Bock, M. Challacombe, C. K. Gan, G. Henkelman, K. Nemeth, A. M. N. Niklasson, A. Odell, E. Schwegler, C. J. Tymczak and V. Weber. *FREEON*. Los Alamos National Laboratory (LA-CC 01-2; LA-CC-04-086), copyright University of California.

[105] T. Ozaki. *Phys. Rev. B* **74**, 245101 (2006).

[106] L. Hung, C. Huang, I. Shin, G. S. Ho, V. L. Lignères and E A Carter. *Comput. Phys. Commun.* **181**, 2208 (2010).

[107] P. Ordéjon. *Phys. Status Solidi b* **217**, 335 (2000).

[108] L. E. Ratcliff, N. D. M. Hine and P. D. Haynes. *Phys. Rev. B* **84**, 165131 (2011).

[109] C. Weber, D. D. O'Regan, N. D. M. Hine, P. B. Littlewood, G. Kotliar and M. C. Payne. *Phys. Rev. Lett.* **10**, 110 (2013).

[110] L. P. Lee, D. J. Cole, M. C. Payne and C.-K. Skylaris. *J. Comp. Chem.* **34**, 429 (2013).

[111] L. P. Lee, D. J. Cole, C.-K. Skylaris, W. L. Jorgensen and M. C. Payne. *J. Chem. Theory Comput.* **9**, 2981 (2013).

[112] M. D. Segall, P. J. D. Lindan, M. J. Probert, C. J. Pickard, P. J. Hasnip, S. J. Clark and M. C. Payne. *J. Phys.: Condens. Matter* **14**, 2717 (2002).

[113] C.-K. Skylaris, P. D. Haynes, A. A. Mostofi and M. C. Payne. *J. Chem. Phys.* **122**, 084119 (2005).

[114] S. Goedecker and L. Colombo. *Phys. Rev. Lett.* **73**, 122 (1994).

[115] W. Yang. *Phys. Rev. Lett.* **66**, 1438 (1991).

[116] E. Tsuchida. *Phys. Soc. Japan* **76**, 034708 (2007).

[117] W. Kohn. *Phys. Rev. Lett.* **76**, 3168 (1996).

[118] E. Prodan and W. Kohn. *Proc. Natl. Acad. Sci. USA* **102**, 11635 (2005).

[119] C.-K. Skylaris, P. D. Haynes, A. A. Mostofi and M. C. Payne. *Phys. Stat. Sol. (b)* **243**, 973 (2006).

[120] D. J. Cole, C.-K. Skylaris, E. Rajendra, A. R. Venkitaraman and M. C. Payne. *Euro. Phys. Lett.* **91**, 37004 (2010).

[121] D. J. Cole, E. Rajendra, M. Roberts-Thomson, B. Hardwick, G. J. McKenzie, M. C. Payne, A. R. Venkitaraman and C.-K. Skylaris. *PLoS Comp. Bio.* **7**, e1002096 (2011).

[122] Cole, D. J., Chin, A. W., Hine, N. D. M., Haynes, P. D., and Payne, M. C. *J. Phys. Chem. Lett.* **4**, 4206 (2013).

[123] C.-K. Skylaris, A. A. Mostofi, P. D. Haynes, O. Dieguez and M. C. Payne. *Phys. Rev. B* **66**, 035119 (2002).

[124] A. A. Mostofi, P. D. Haynes, C.-K. Skylaris and M. C. Payne. *J. Chem. Phys.* **119**, 8842 (2003).

[125] D. D. O'Regan, M. C. Payne and A. A. Mostofi. *Phys. Rev. B* **85**, 193101 (2012).

[126] P. D. Haynes and M. C. Payne. *Phys. Rev. B* **59**, 12173 (1999).

[127] X. -P. Li, R. W. Nunes and D. Vanderbilt. *Phys. Rev. B* **47**, 10891 (1993).

[128] M. S. Daw. *Phys. Rev. B* **47**, 10895 (1993).

[129] R. Mc Weeny. *Rev. Mod. Phys.* **32**, 335 (1960).

[130] P. D. Haynes, C.-K. Skylaris, A. A. Mostofi and M. C. Payne. *Chem. Phys. Lett.* **422**, 345 (2006).

[131] P. D. Haynes, C.-K. Skylaris, A. A. Mostofi and M. C. Payne. *Phys. Stat. Sol. (B)* **243**, 2489 (2006).

[132] C.-K. Skylaris, O. Diéguez, P. D. Haynes, and M. C. Payne. *Phys. Rev. B* **66**, 073103 (2002).

[133] C.-K. Skylaris, A. A. Mostofi, P. D. Haynes, C. J. Pickard, and M. C. Payne. *Comput. Phys. Commun.* **140**, 315 (2001).

[134] A. A. Mostofi, C.-K. Skylaris, P. D. Haynes and M. C. Payne. *Phys. Rev. B* **147**, 788 (2002).

[135] Á. Ruiz-Serrano and C.-K. Skylaris. *J Chem. Phys.* **139**, 164110 (2013).

[136] J. Bernholc, E. L. Briggs, D. J. Sullivan, C. J. Brabec, M. B. Nardelli, K. Rapcewicz, C. Roland, and M. Wensell. *Int. J. Quantum Chem.* **65**, 531 (1997).

[137] E. Artacho, E. Anglada, O. Dieguez, J. D. Gale, A. Garcia, J. Junquera, R. M. Martin, P. Ordejon, J. M. Pruneda, D. Sanchez- Portal, and J. M. Soler. *J. Phys.: Conden. Matt.* **20**, 064208 (2008).

[138] J. L. Fattebert and F. Gygi. *Comput. Phys. Commun.* **162**, 24 (2004).

[139] N. D. M. Hine, M. Robinson, P. D. Haynes, C.-K. Skylaris, M. C. Payne and A. A. Mostofi. *Phys. Rev. B* **83**, 195102 (2011).

[140] Á. Ruiz-Serrano, N. D. M. Hine and C.-K. Skylaris. *J Chem. Phys.* **136**, 234101 (2012).

[141] M. L. Cohen, M. Schlüter, J. R. Chelikowsky and S. G. Louie. *Phys. Rev. B* **12**, 5575 (1975).

[142] J. Ihm, A. Zunger and M. L. Cohen. *J. Phys. C* **12**, 4409 (1979).

[143] M. C. Payne, M. P. Teter, D. C. Allan, T. A. Arias and J. D. Joannopoulos. *Rev. Mod. Phys.* **64**, 1045 (1992).

[144] N.D.M. Hine, J. Dziedzic, P. D. Haynes and C.-K. Skylaris. *J. Chem. Phys.* **135**, 204103 (2011).

[145] G. Makov and M. C. Payne. *Phys. Rev. B* **51**, 4014 (1995).

[146] M. R. Jarvis, I. D. White, R. W. Godby and M. C. Payne. *Phys. Rev. B* **56**, 14972 (1997).

[147] C. A. Rozzi, D. Varsano, A. Marini, E. K. U. Gross and A. Rubio. *Phys. Rev. B* **73**, 205119 (2006).

[148] N. D. M. Hine, J. Dziedzic, P. D. Haynes and C.-K. Skylaris. *J. Chem. Phys.* **135**, 204103 (2011).

[149] M. A. L. Marques, A. Castro, G. F. Bertsch and A. Rubio. *Comput. Phys. Commun.* **151**, 60 (2003).

[150] A. Castro, H. Appel, M. Oliveira, C. A. Rozzi, X. Andrade, F. Lorenzen, M. A. L. Marques, E. K. U. Gross and A. Rubio. *Phys. Status Solidi B* **243**, 2465 (2006).

[151] J. Tomasi, B. Mennucci and R. Cammi. *Chem. Rev.* **105**, 2999 (2005).

[152] J. Tomasi and M. Persico. *Chem. Rev.* **94**, 2027 (1994).

[153] A. Klamt and G. Schüürmann. *J. Chem. Soc. Perkin Trans. 2* , 799 (1993).

[154] A. V. Marenich, C. J. Cramer and D. G. Truhlar. *J. Chem. Theor. Comput.* **5**, 2447 (2009).

[155] B. Mennucci, E. Cancè and J. Tomasi. *J. Phys. Chem. B* **101**, 10506 (1997).

[156] J.-L. Fattebert, F. Gygi. *J. Comp. Chem.* **23**, 662 (2002).

[157] D. Scherlis, J. Fattebert, F. Gygi, M. Cococcioni and N. Marzari. *J. Chem. Phys.* **124**, 074103 (2006).

[158] J. Dziedzic, H. H. Helal, C.-K. Skylaris, A. A. Mostofi and M. C. Payne. *Europhysics Letters* **95**, 43001 (2011).

[159] J. Dziedzic, S. J. Fox, T. Fox, C. S. Tautermann and C.-K. Skylaris. *Int. J. Quantum Chem.* **113**, 771 (2013).

[160] D. A. Scherlis, J. Fattebert, F. Gygi, M. Cococcioni and N. Marzari. *J. Chem. Phys.* **124**, 074103 (2006).

[161] P. W. Avraam, N. D. M. Hine, P. Tangney and P. D. Haynes. *Phys. Rev. B* **85**, 115404 (2012).

[162] N. D. M. Hine, P. W. Avraam, P. Tangney and P. D. Haynes. *J. Phys. Conf. Ser* **367**, 012002 (2012).

[163] S. Grimme. *J. Comput. Chem.* **25**, 1463 (2006).

[164] S. Grimme. *J. Comput. Chem.* **27**, 1787 (2006).

[165] Q. Hill and C.-K. Skylaris. *Proc. R. Soc. A* **465**, 669 (2009).

[166] M. Elstner, P. Hobza, T. Frauenheim, S. Suhai and E. Kaxiras. *J. Chem. Phys.* **114**, 5149 (2001).

[167] S. J. Fox and C. Pittock and C. S. Tautermann and N. Malcolm and C.-K. Skylaris. *J. Chem. Phys.* **135**, 224107 (2011).

[168] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey and M. L. Klein. *J. Chem. Phys.* **79**, 926 (1983).

[169] E. D. Glendening, J. K. Badenhoop, A. E. Reed, J. E. Carpenter, J. A. Bohmann, C. M. Morales and F. Weinhold. *NBO 5.* Theoretical Chemistry Institute, University of Wisconsin, Madison, (2001).

[170] L. P. Lee, D. J. Cole, M. C. Payne and C.-K. Skylaris. *J. Comp.Chem.* **34**, 429 (2013).

[171] A. E. Reed, R. B. Weinstock and F. Weinhold. *J. Chem. Phys.* **83**, 735 (1985).

[172] J. P. Foster and F. Weinhold. *J. Chem. Phys.* **102**, 7211 (1980).

[173] A. E. Reed, L. A. Curtiss and F. Weinhold. *Chem. Rev.* **88**, 899 (1988).

[174] R. Ludwig. *J. Mol. Liq.* **84**, 65 (2000).

[175] A. Mohajeri and F. F. Nobandegani. *J. Phys. Chem. A* **112**, 281 (2008).

[176] G. J. Bartlett, A. Choudhary, R. T. Raines and D. N. Woolfson. *Nature Chem. Biol.* **6**, 615 (2010).

[177] I. Fleming. *Frontier Orbitals and Organic Chemical Reactions.* Wiley, New York, (1976).

[178] Alabugin, I. V., Gilmore, K. M., and Peterson, P. W. *WIREs Comput. Mol. Sci.* **1**, 109 (2011).

[179] F. Hirshfeld. *Theor. Chim. Acta* **44**, 129 (1977).

[180] P. Bultinck, C. V. Alsenoy, P. W. Ayers and R. Carbó-Dorca. *J. Chem. Phys.* **144111**, 126 (2007).

[181] T. A. Manz and D. S. Sholl. *J. Chem. Theory Comput.* **6**, 2455 (2010).

[182] T. A. Manz and D. S. Sholl. *J. Chem. Theory Comput.* **8**, 2844 (2012).

[183] L. P. Lee, D. J. Cole, C. -K. Skylaris, W. L. Jorgensen and M. C. Payne. *J. Chem. Theory Comput.* **9**, 2981 (2013).

[184] T. A. Manz and D. S. Sholl. *J. Chem. Theory Comput.* **8**, 2844 (2012).

[185] T. A. Manz and D. S. Sholl. *J. Chem. Theory Comput.* **6**, 2455 (2010).

[186] P. Ehrenfest. *Z. Phys.* **45**, 455 (1927).

[187] M. Born and V. Fock. *Z. Phys.* **51**, 165 (1928).

[188] P. Güttinger. *Z. Phys.* **73**, 169 (1931).

[189] W. Pauli. *Handbuch der Physik*. Springer, Berlin, (1933).

[190] H. Hellmann. *Einfuhrung in die Quantumchemie*. Franz Duetsche, Liepzig, (1937).

[191] R. P. Feynman. *Phys. Rev.* **56**, 340 (1939).

[192] A. Hurley. *Proc. R. Soc. London Ser. A* **226**, 179 (1954).

[193] M. Schefer, J. P. Vigneron and G. B. Bachelet. *Phys. Rev. B* **31**, 6541 (1985).

[194] M. Di Ventra and S. Pantelides. *Phys. Rev. B* **61**, 16207 (2000).

[195] J. Harris, R. O. Jones and J. E. Müller. *J. Chem. Phys.* **75**, 3904 (1981).

[196] K. Hinsen. *Normal mode theory and harmonic potential approximations.* Contribution to: Normal Mode Analysis: Theory and Applications to Biological and Chemical Systems. edited by Q. Cui and I. Bahar, (2005).

[197] J. C. Mauro, R. J. Loucks and J. Balakrishnan. *J. Phys. Chem. A* **109**, 9578 (2005).

[198] J. N. Murrell and K. J. Laidler. *Trans. Faraday Soc.* **67**, 371 (1968).

[199] J. Kästner, J. M. Carr, T. W. Keal, W. Thiel, A. Wander and P. Sherwood. *J. Phys. Chem. A* **113**, 11856 (2009).

[200] N. Govind, M. Petersen, G. Fitzgerald, D. King-Smith and J. Andzelm. *Comp. Mat. Sci.* **28**, 250 (2003).

[201] T. A. Halgren and W. N. Lipscomb. *Chem. Phys. Lett.* **49**, 225 (1977).

[202] O. R. Inderwildi, D. Lebiedz, O. Deutschmann and J. Warnatz. *J. Chem. Phys.* **122**, 034710 (2005).

[203] W. A. Abdallah and A. E. Nelson. *J. Phys. Chem. B* **109**, 10863 (2005).

[204] D. L. Geatches, H. C. Greenwell and S. J. Clark. *J. Phys. Chem. A* **115**, 2658 (2011).

[205] Z. Zhao, Z. Li and Z. Zou. *J. Phys. Chem. C* **116**, 7430 (2012).

[206] M. Sun, A. E. Nelson and J. Adjaye. *J. Catal.* **233**, 411 (2005).

[207] B. Blom, G. Klatt, J. C.Q. Fletcher and J. R. Moss. *Inorg. Chim. Acta* **360**, 2890 (2007).

[208] A. Simperler, A. Kornherr, R. Chopra, W. Jones, W. D. S. Motherwell and G. Zifferer. *Phys. Chem. Chem. Phys.* **9**, 3999 (2007).

[209] W. Gao, M. Zhao and Q. Jiang. *J. Phys. Chem. C* **111**, 4042 (2007).

[210] B. Liu, M. T. Lusk, J. F. Ely, A. C. T. van Duin and W. A. Goddard III. *Mol. Sim.* **34**, 967 (2008).

[211] J. D. Zheng, C. H. Lu, B. Z. Sun and W. K. Chen. *Acta Phys. -Chim. Sin.* **24**, 1995 (2008).

[212] B. Liu, M. T. Lusk and J. F. Ely. *J. Phys. Chem. C* **113**, 13715 (2009).

[213] F. Zhang, L. Li and A. Tian. *Acta Phys. -Chim. Sin.* **25**, 1883 (2009).

[214] P. Peng, G. F. Li, Z. A. Tian, K. J. Dong and R. S. Liu. *Comp. Mat. Sci.* **44**, 881 (2009).

[215] Y. L. Yang, W. K. Chen, X. Guo, Y. Li and Y. F. Zhang. *Chinese J. Struct. Chem.* **29**, 1021 (2010).

[216] D. Wang, P. Zhu and L. J. Hu. *Adv. Mat. Res.* **183**, 800 (2011).

[217] X. Zhao, L. Song, J. Fu, P. Tang and F. Liu. *Surf. Sci.* **605**, 1005 (2011).

[218] D. J. Wales. *J. Chem. Phys.* **101**, 3750 (1994).

[219] L. J. Munro and D. J. Wales. *Phys. Rev. B* **59**, 3969 (1999).

[220] Y. Kumeda, L. J. Munro and D. J. Wales. *Chem. Phys. Lett.* **341**, 185 (2001).

[221] K. Kadau, T. C. Germann and P. S. Lomdahl. *Int. J Mod. Phys. C* **17**, 1755 (2006).

[222] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan and M. Karplus. *J. Comput. Chem.* **4**, 187 (1983).

[223] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R.Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell and P. A. Kollman. *J. Am. Chem. Soc.* **117**, 5179 (1995).

[224] D. A. Case, T. A. Darden, T. E. Cheatham III, C. L. Simmerling, J. Wang, R. E. Duke, R. Luo, R. C. Walker, W. Zhang, K. M. Merz, B. Roberts, B. Wang, S. Hayik, A. Roitberg, G. Seabra, I. Kolossvry, K. F. Wong, F. Paesani, J. Vanicek, J. Liu, X. Wu, S. R. Brozell, T. Steinbrecher, H. Gohlke, Q. Cai, X. Ye, J. Wang, M. -J. Hsieh, G. Cui, D. R. Roe, D. H. Mathews, M. G. Seetin, C. Sagui, V. Babin, T. Luchko, S. Gusarov, A. Kovalenko and P. A. Kollman . *AMBER 11.* University of California, San Francisco, (2010).

[225] W. F. van Gunsteren. *Groningen Molecular Simulation Program Package; University of Groningen* (1987).

[226] W. L. Jorgensen and J. TiradoRives. *J. Am. Chem. Soc.* **110**, 1657 (1988).

[227] P. Söderhjelm and U. Ryde. *J. Comput. Chem.* **30**, 750 (2009).

[228] J. W. Ponder and D. A. Case. *Adv. Protein Chem.* **66**, 27 (2003).

[229] A. D. MacKerell Jr. *J. Comput. Chem.* **25**, 1584 (2004).

[230] J. L. Gao and D. G. Truhlar. *Annu. Rev. Phys. Chem.* **53**, 467 (2002).

[231] H. M. Senn and W. Thiel. *Top. Curr. Chem.* **268**, 173 (2007).

[232] H. M. Senn and W. Thiel. *Angew. Chem. Int. Ed.* **48**, 1198 (2009).

[233] A. J. Mulholland, P. D. Lyne and M. Karplus. *J. Am. Chem. Soc.* **122**, 534 (2000).

[234] R. A. Friesner and V. Guallar. *Annu. Rev. Phys. Chem.* **56**, 389 (2005).

[235] C. B. Post and M. Karplus. *J. Am. Chem. Soc.* **108**, 1317 (1986).

[236] A. L. Bowman, I. M. Grant and A. J. Mulholland. *Chem. Commun.* , 4425 (2008).

[237] R. Lonsdale, K. E. Ranaghan and A. J. Mulholland. *Chem. Commun.* **46**, 2354 (2010).

[238] I. Solt, P. Kulhànek, I. Simon, S. Winfield, M. C. Payne, G. Csànyi and M. Fuxreiter. *J. Phys. Chem. B* **113**, 5728 (2009).

[239] C. V. Sumowski, B. B. T. Schmitt, S. Schweizer and C. Ochsenfeld. *Angew. Chem. Int. Ed.* **49**, 9951 (2010).

[240] C. M. Isborn, A.W. Götz, M. A. Clark, R. C. Walker and T. J. Martínez. *J. Chem. Theory Comput.* **8**, 5092 (2012).

[241] M. J. Field,P. A. Bash and M. Karplus. *J. Comput. Chem.* **11**, 700 (1990).

[242] U. C. Singh and P. A. Kollman. *J. Comput. Chem.* **7**, 718 (1986).

[243] Y. Zhang, T.-S. Lee and W. Yang. *J. Chem. Phys.* **110**, 46 (1999).

[244] A. Laio, J. V. Vondele and U. Rothlisberger. *J Chem. Phys.* **116**, 6941 (2002).

[245] M. Schmidt am Busch and E. -W. Knapp. *Chem. Phys. Chem,* **5**, 1513 (2004).

[246] M. Schmidt am Busch and E. -W. Knapp. *J. Am. Chem. Soc.* **127**, 15730 (2005).

[247] A. S. Galstyan, S. D. Zarić and E. -W. Knapp. *J. Biol. Inorg. Chem.* **10**, 343 (2005).

[248] M. Dułak and T. A. Wesołowski. *J Chem. Phys.* **124**, 164101 (2006).

[249] J. Neugebauer. *Chem. Phys. Chem.* **10**, 3148 (2009).

[250] T. Renger and F. Müh. *Phys. Chem. Chem. Phys.* **15**, 3348 (2013).

[251] B. Strodel and D. J. Wales. *Chem. Phys. Lett* **466**, 105 (2008).

[252] Z. Cao, L. Liu, L. Zhao and J. Wang. *Int. J. Mol. Sci.* **12**, 8259 (2011).

[253] D. Matthes and B. L. de Groot. *Biophysical. J.* **97**, 599 (2009).

[254] W. F. Van Gunsteren, S. R. Billeter, A. A. Eising, P. H. Hunenberger, P. Krüger, A. E. Mark, W. R. P. Scott and I. G. Tironi. *Biomolecular Simulation: The GROMOS96 Manual and User Guide.* vdf Hochschulverlag AG an der ETH Zürich and BIOMOS b.v., Zürich, Switzerland and Groningen, The Netherlands, (1996).

[255] W. L. Jorgensen, D. S. Maxwell, J. Tirado-Rives. *J. Am. Chem. Soc.* **18**, 11225 (1996).

[256] M. Probst, T. Radnai, K. Heinzinger, P. Bopp and B. Rode. *J. Phys. Chem.* **89**, 753 (1985).

[257] A. Tongraar, K. R. Liedl and B. M. J. Rode. *Phys. Chem. A* **101**, 6299 (1997).

[258] T. A. Halgren and W. Damm. *Curr. Opin. Struct. Biol.* **11**, 236 (2001).

[259] J. W. Ponder, C. Wu, P. Ren, V. S. Pande, J. D. Chodera, M. J. Schnieders, I. Haque, D. L. Mobley, D. S. Lambrecht, R. A. DiStasio Jr., M. Head-Gordon, G. N. I. Clark, M. E. Johnson and T. Head-Gordon. *J Phys Chem B* **114**, 2549 (2010).

[260] H. J. Kulik, N. Luehr, I. S. Ufimtsev and T. J. Martínez. *J. Phys. Chem. B* **116**, 12501 (2012).

[261] NIST Chemistry WebBook, NIST Standard Reference Database Number 69, Eds. P.J. Linstrom and W.G. Mallard.

[262] N. Biswas and S. Umapathy. *J. Phys. Chem. A* **101**, 5555 (1997).

[263] B. Karthikeyan. *Spectrochimica Acta Part A* **64**, 1083 (2006).

[264] M. Nonella and P. Tavan. *Chem. Phys.* **199**, 19 (1997).

[265] A. A. El-Azhary and H. U. Suter. *J. Phys. Chem.* **100**, 15056 (1996).

[266] C.-K. Skylaris and P. D. Haynes. *J. Chem. Phys.* **127**, 164712 (2007).

[267] M. Feig. *J. Chem. Theory Comput.* **4**, 1555 (2008).

[268] C. L. Brooks III and D. Case. *Chem. Rev.* **93**, 2487 (1993).

[269] P. E. Smith, B. M. Pettitt and M. J. Karplus. *J. Phys. Chem.* **97**, 6907 (1993).

[270] D. S. Chekmarev, T. Ishida and R. M. Levy. *J. Phys. Chem. B* **108**, 19487 (2004).

[271] P. Rossky and M. Karplus. *J. Am. Chem. Soc.* **101**, 1913 (1979).

[272] W. C. Swope, J. W. Pitera, F. Suits, M. Pitman, M. Eleftheriou, B. G. Fitch, R. S. Germain, A. Rayshubski, T. J. C. Ward, Y. Zhestkov, and R. Zhou. *J. Phys. Chem. B* **108**, 6582 (2004).

[273] P. E. Smith. *J. Chem. Phys.* **111**, 5568 (1999).

[274] A. N. Drozdov, A. Grosseld and R. V. Pappu. *J. Am. Chem. Soc.* **126**, 2574 (2004).

[275] Z. X. Wang and Y. Duan. *J. Comput. Chem.* **25**, 1699 (2004).

[276] M. Feig. *J. Chem. Theory Comput.* **3**, 1734 (2007).

[277] K. Kwac, K. K. Lee, J. B. Han, K. I. Oh and M. Cho. *J. Chem. Phys.* **128**, 105106 (2008).

[278] R. J. Lavrich, D. F. Plusquellic, R. D. Suenram, G. T. Fraser, A. R. Hight Walker and M. J. Tubergen. *J. Chem. Phys.* **118**, 1253 (2003).

[279] C. F. Weise and J. C. Weisshaar. *J. Phys. Chem. B* **107**, 3265 (2003).

[280] Y. S. Kim, J. P. Wang and R. M. Hochstrasser. *J. Phys. Chem. B* **109**, 7511 (2005).

[281] J. Grdadolnik, S. G. Grdadolnik and F. Avbelj. *J. Phys. Chem. B* **112**, 2712 (2008).

[282] V. Parchaňský, J. Kapitán, J. Kaminský, J. Šebestík and P. Bouř. *J. Phys. Chem. Lett.* **4**, 2763 (2013).

[283] M. D. Beachy, D. Chasman, R. B. Murphy, T. A. Halgren, and R. A. Friesner. *J. Am. Chem. Soc.* **119**, 5908 (1997).

[284] A. D. MacKerell Jr., D. Bashford, M. Bellott, J. D. Dunbrack, M. J. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin and M. Karplus. *J. Phys. Chem. B* **102**, 3586 (1998).

[285] A. D. MacKerell Jr., M. Feig and C. L. Brooks III. *J. Am. Chem. Soc.* **126**, 698 (2004).

[286] A. D. MacKerell. *J. Comput. Chem.* **25**, 1584 (2004).

[287] D. A. Case, T. E. Cheatham III, T. Darden, H. Gohlke, R. Luo, K. M. Merz Jr., A. Onufriev, C. Simmerling, B. Wang and R. J. Woods. *J. Comput. Chem.* **26**, 1668 (2005).

[288] M. Avignon, P. V. Huong and J. Lascombe. *Biopolymers* **8**, 69 (1969).

[289] M. Avignon, C. Garrigou and P. Botherel. *Biopolymers* **12**, 1651 (1973).

[290] G. N. Ramachandran and V. Sasissekharan. *Adv. Prot. Chem.* , 283 (1968).

[291] S. Martí, J. Andrés, V. Moliner, E. Silla, I. Tuñón and J. Bertrán. *J. Phys. Chem. B* **104**, 11308 (2000).

[292] E. Rudberg. *J. Phys.: Condens. Mat.* **24**, 072202 (2012).

[293] E. H. Rubensson and E. Rudberg. *J. Comput. Chem.* **32**, 1411 (2011).

[294] J. P. Perdew and A. Zunger. *Phys. Rev. B* **23**, 5048 (1981).

[295] P. C. do Couto, S. G. Estácio and B. J. C. Cabral. *J. Chem. Phys.* **123**, 054510 (2005).

[296] S. Grimme, W. Hujo and B. Kirchner. *Phys. Chem. Chem. Phys.* **14**, 4875 (2012).

[297] C. M. Isborn, N. Luehr, I. S. Ufimtsev and T. J. Martínez. *J. Chem. Theory Comput.* **7**, 1814 (2011).

[298] J. Antony and S. Grimme. *J. Comput. Chem.* **33**, 1730 (2012).

[299] R. W. Godby and M Schlüter and L. J. Sham. *Phys. Rev. B* **37**, 10159 (1988).

[300] A. Seidl, A. Görling, P. Vogl, J. A. Majewski and M. Levy. *Phys. Rev. B* **53**, 3764 (1996).

[301] L. J. Sham and W. Kohn. *Phys. Rev.* **145**, 561 (1966).

[302] J. P. Perdew and M. Levy. *Phys. Rev. Lett.* **51**, 1884 (1983).

[303] P. Mori-Sánchez, A. J. Cohen and W. Yang. *Phys. Rev. Lett.* **100**, 146401 (2008).

[304] P. W. Avraam, N. D. M. Hine, P. Tangney and P. D. Haynes. *Phys. Rev. B* **85**, 115404 (2012).

[305] N. D. M. Hine, P. W. Avraam, P. Tangney and P. D. Haynes. *J. Phys. Conf. Ser.* **367**, 012002 (2012).

[306] I. Marcotte, F. Separovic, M. Auger and S. M. Gagne. *Biophys. J.* **86**, 1587 (2004).

[307] N. Assa-Munt, X. Jia, P. Laakkonen and E. Ruoslahti. *Biochemistry* **40**, 2373 (2001).

[308] C. P. Jaroniec, C. E. MacPhee, V. S. Bajaj, M. T. McMahon, C. M. Dobson and R. G. Griffin. *Proc. Natl. Acad. Sci. USA* **101**, 711 (2004).

[309] P. L. Yeagle, A. Salloum, A. Chopra, N. Bhawsar, L. Ali, G. Kuzmanovski, J. L. Alderfer and A. D. Albert. *J. Pept. Res.* **55**, 455 (2000).

[310] P. L. Yeagle, C. Danis, G. Choi, J. L. Alderfer and A. D. Albert. *Mol. Vis.* **6**, 125 (2000).

[311] S. Vijay-Kumar, C. E. Bugg and W. J. Cook. *J. Mol. Biol.* **194**, 531 (1987).

[312] F. C. Bernstein, T. F. Koetzle, G. J. Williams, E. E. Meyer Jr., M. D. Brice, J. R. Rodgers, O.Kennard, T. Shimanouchi and M.Tasumi. *J. Mol. Biol.* **112**, 535 (1977).

[313] L. Pauling. *The Nature of the Chemical Bond.* Cornell University Press, Ithaca, New York, (1960).

[314] W. G. J. Hol, P. T. Van Duijnen and H. J. C. Berendsen. *Nature* **273**, 443 (1978).

[315] A. Wada. *Adv. Biophys.* **9**, 1 (1976).

[316] W. G. J. Hol . *Prog. Biophys. Molec. Biol.* **45**, 149 (1985).

[317] C. -I. Brändén and J. Tooze. *Introduction to Protein Structure.* Garland Science, New York, NY, USA and Abingdon, Oxford, UK, (1998).

[318] R. R. Gabdoulline and R. C. Wade. *J. Phys. Chem.* **100**, 3868 (1996).

[319] D. Sengupta, R. N. Behera, J. C. Smith and G. M. Ullmann. *Structure* **13**, 849 (2005).

[320] F. Claeyssens, K. E. Ranaghan, N. Lawan, S. J. Macrae, F. R. Manby, J. N. Harvey and A. J. Mulholland. *Org. Biomol. Chem.* **9**, 1578 (2011).

[321] E. Haslam. *Shikimic Acid: Matabolism and Metabolites.* John Wiley & Sons, (1993).

[322] A, Zulet , M. Gil-Monreal, J. G. Villamor, A. Zabalza, R. A. L. van der Hoorn and M. Royuela. *PLoS ONE* **8**, e73847 (2013).

[323] P. A. Bartlett and C. R. Johnson. *J. Am. Chem. Soc.* **45**, 6856 (1985).

[324] Y. M. Chook, H. Ke and W. N. Lipscomb. *Proc. Natl. Acad. Sci. USA* **90**, 8600 (1993).

[325] U. Weiss and J. M. Edwards. *The Biosynthesis of Aromatic Amino Compounds.* Wiley, New York, (1980).

[326] P. D. Lyne, A. J. Mulholland and W. G. Richards. *J. Am. Chem. Soc.* **11345**, 17 (1995).

[327] K. E. Ranaghan, L. Ridder, B. Szefczyk, W. A. Sokalski, J. C Hermann and A. J. Mulholland. *Mol. Phys.* **101**, 2695 (2003).

[328] A. Crespo, D. A. Scherlis, M. A. Marti, P. Ordejon, A. E. Roitberg and D. A. Estrin. *J. Phys. Chem. B* **107**, 13728 (2003).

[329] F. Claeyssens, J. N. Harvey, F. R. Manby, R. A. Mata, A. J. Mulholland, K. E. Ranaghan, M. Schütz, S. Thiel, W. Thiel and H. -J. Werner. *Angew. Chem. Int. Ed.* **45**, 6856 (2006).

[330] W. J. Guildford and J. R. Knowles. *J. Am. Chem. Soc.* **109**, 5103 (1987).

[331] J. V. Gray, D. Eren and J. R. Knowles. *Biochemistry* **29**, 8872 (1990).

[332] J. V. Gray and J. R. Knowles. *Biochemistry* **29**, 8872 (1994).

[333] P. Kast, M. Asif-Ullah and D. Hilvert. *Tetrahedron Lett.* **37**, 2691 (1996).

[334] P. Kast, Y. B. Tewari, O. Wiest, D. Hilvert, K. N. Houk and R. N. Goldberg. *J. Phys. Chem. B* **101**, 10976 (1997).

[335] P. R. Andrews, G. D. Smith and I. G. Young. *Biochemistry* **12**, 3492 (1973).

[336] Y. M. Chook, J. V. Gray, H. Ke and W. N. Lipscomb. *J. Mol. Biol.* **240**, 476 (1994).

[337] A. Eletsky, A. Kienhöfer, D. Hilvert and K. Pervushin. *Biochemistry* **44**, 6788 (2005).

[338] R. J. Hall, S. A. Hindle, N. A. Burton and I. H. Hillier. *J. Comp. Chem.* **21**, 1433 (2000).

[339] S. T. Cload, D. R. Liu, R. M. Pastor and P. G. Schultz. *J. Am. Chem Soc.* **118**, 1787 (1996).

[340] C. R. W. Guimaraēs, M. Udier-Blagović, I. Tubert-Brohman and W. L. Jorgensen. *J. Am. Chem. Soc.* **125**, 6892 (2003).

[341] H. Gorisch. *Biochemistry* **3700**, 17 (1978).

[342] S. E. Worthington, A. E. Roitberg and M. Krauss. *J. Phys. Chem. B* **103**, 7087 (2001).

[343] Y. S. Lee, S. E. Worthington, M. Krauss and B. R. Brooks. *J. Phys. Chem. B* **106**, 12059 (2002).

[344] T. Ishida, D. G. Fedorov and K. Kitaura. *J. Phys. Chem. B* **110**, 1457 (2006).

[345] C. Steinmann, D. G. Fedorov and J. H. Jensen. *PLoS One* **8**, e60602 (2013).

[346] K. E. Ranaghan, L. Ridder, B. Szefczyk, W. A. Sokalski, J. C Hermann and A. J. Mulholland. *Org. Biomol. Chem.* **2**, 968 (2004).

[347] J. Jitonnom, A. J. Mulholland, P. Nimmanpipug and V. S. Lee. *Maejo Int. J. Sci. Technol.* **5**, 47 (2011).

[348] J. Jitonnom, V. S. Lee, P. Nimmanpipug, H. A. Rowlands and A. J. Mulholland. *Biochemistry* **50**, 4697 (2011).

[349] V. S. Lee, K. Kodchakorn, J. Jitonnom, P. Nimmanpipug, P. Kongtawelert and B Premanode. *J. Comput. Aided. Mol. Des.* **24**, 879 (2010).

[350] K. E. Ranaghan and A. J. Mulholland. *Chem. Commun.* , 1238 (2004).

[351] F. Claeyssens, K. E. Ranaghan, F. R. Manby, J. N. Harvey and A. J. Mulholland. *Chem. Commun.* **45**, 6856 (2006).

[352] N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch and G. R. Hutchinson. *J. Cheminf.* **3**, 33 (2011).

[353] S. Grimme, J. Antony, S. Ehrlich and H. Krieg. *J. Chem. Phys.* **132**, 154104 (2010).

[354] Lonsdale, R., Harvey, J. N., and Mulholland, A. J. *J. Phys. Chem. Lett..* **1**, 3232 (2010).

[355] Lonsdale, R., Harvey, J. N., and Mulholland, A. J. *J. Chem. Theory Comput.* **8**, 4637 (2012).

[356] K. E. Ranaghan and A. J. Mulholland. *Int. Rev. Phys. Chem.* **29**, 65 (2010).

[357] A. Gobbi and G. Frenking. *J. Am. Chem. Soc.* **115**, 2362 (1993).

[358] M. Wu and Å Strid and L. A. Eriksson. *Chem. Phys. Lett.* **584**, 188 (2013).

[359] Z. B. Maksić and B. Kovacêvić. *J. Chem. Soc., Perkin Trans.* **2**, 2623 (1999).

[360] M. H. M. Olsson and W. W. Parson and A. Warshel. *Chem. Rev.* **106**, 1737 (2006).

[361] S. D. Copley and J. R. Knowles. *J. Am. Chem. Soc.* **109**, 5008 (1987).

[362] S. Hur and T. C. Bruice. *J. Am. Chem. Soc.* **125**, 1472 (2003).

[363] S. Hur and T. C. Bruice. *J. Am. Chem. Soc.* **125**, 5964 (2003).

[364] S. Hur and T. C. Bruice. *J. Am. Chem. Soc.* **125**, 10540 (2003).

[365] B. Szefczyk, A. J. Mulholland, K. E. Ranaghan and W. Andrzej Sokalski. *J. Am. Chem. Soc.* **126**, 16148 (2004).

[366] P. Kast, J. D. Hartgerink, M. Asif-Ullah and D. Hilvert. *J. Phys. Chem. B* **118**, 3069 (1996).

[367] I. S. Umfistev, N. Luehr and T. J. Martinez. *J. Phys. Chem. Lett.* **2**, 1789 (2011).

[368] D. D. O'Regan, N. D. M. Hine, M. C. Payne and A. A. Mostofi. *Phys. Rev. B* **85**, 085107 (2012).

[369] C. Weber, D.J.Cole, D.D. O'Regan and M.C.Payne. *Proc. Natl. Acad. Sci. USA* **111**, 5790 (2014).

[370] V. B. Chen, W. B. Arendall III, J. J. Headd, D. A. Keedy, R. M. Immormino, G. J. Kapral, L. W. Murray, J. S. Richardson and D. C. Richardson. *Acta Crystallographica* **D66**, 12 (2010).

[371] H. J. Kulik and N. Marzari. *Fuel Cell Science: Theory, Fundamentals, and Biocatalysis, Chapter 14: Catalytic activity of transition-metal complexes.* Wiley, (2010).

[372] K. Yoshizawa. *J. Inorg. Biochem.* **78**, 23 (2000).

[373] K. Yoshizawa and T. Yumura. *J. Inorg. Biochem.* **96**, 257 (2003).

[374] L. C. Blasiak, F. H. Vaillancourt, C. T. Walsh and C. L. Drennan. *Nature* **440**, 368 (2006).

[375] S. P. D. Visser, F. Ogliaro, Z. Gross and S. Shaik. *Chem. Eur. J.* **7**, 4954 (2001).

[376] H. J. Kulik and C. L. Drennan. *J. Biol. Chem.* **288**, 11233 (2013).

[377] S. E. Wong, E. Y. Lau, H. J. Kulik, J. H. Satcher Jr., C. A. Valdez, M. Worsely, F. C. Lightstone and R. D. Aines. *Energy Procedia* **4**, 817 (2011).

[378] H. J. Kulik, S. E. Wong, S. E. Baker, C. A. Valdez, J. H. Satcher Jr., R. D. Aines and F. C. Lightstone. *Acta Crystallographica C: Structural Chemistry (Special issue on computational materials discovery)* (2013).

[379] K. Schuchmann and V. Müller. *Science* **342**, 1382 (2013).