UNIVERSITY OF SOUTHAMPTON

SCHOOL OF CHEMISTRY

Protein-ligand Binding Affinities from Large-scale Quantum Mechanical Simulations

by

Stephen John Fox

Thesis for the degree of Doctor of Philosophy

September 2012

Shikin Haramitsu Daikomyo

-in every moment there is the chance of finding the enlightenment we seek

UNIVERSITY OF SOUTHAMPTON

SCHOOL OF CHEMISTRY

Doctor of Philosophy

Protein-ligand Binding Affinities from Large-scale Quantum Mechanical Simulations

by

Stephen John Fox

Abstract

The accurate prediction of protein-drug binding affinities is a major aim of computational drug optimisation and development. A quantitative measure of binding affinity is provided by the free energy of binding, and such calculations typically require extensive configurational sampling of entities such as proteins with thousands of atoms. Current binding free energy methods use force fields to perform the configurational sampling and to compute interaction energies. Due to the empirical nature of force fields and the neglect of electrons, electron polarisation and charge transfer are not accounted for explicitly. This can limit the accuracy with which interactions are calculated and consequently the free energies obtained. Ideally ab initio quantum chemistry approaches should be used as these explicitly include the electrons. However, conventional ab initio approaches are not suitable due to their prohibitively high computational cost and unfavourable scaling.

In this thesis we use large-scale ab initio quantum chemistry calculations within the Density Functional Theory (DFT) method to address the above mentioned limitations of force fields. To obtain quantitative results with ab initio approaches

it is important to converge the calculations with the size of the basis set. For this reason we have used the ONETEP program, which is capable of linear-scaling DFT with near-complete basis set accuracy.

A well known binding free energy approach is the Molecular Mechanics Poisson-Boltzmann Surface Area (MM-PBSA), which obtains free energies from evaluation of the energy of configurations in an implicit solvent model. We present the first application of a “QM-PBSA” approach to a protein-ligand system containing over 2600 atoms. In this QM-PBSA approach the energies of the configurations in vacuum are evaluated with ONETEP. The solvation energies were also obtained with ONETEP using a minimal parameter implicit solvent model within the self-consistent calculation.

Large-scale DFT calculations were also applied within a more theoretically rigorous free energy approach which can, in principle, obtain the full entropic contributions to free energy change. The method performs a mutation from a molecular mechanical (MM) description to an quantum mechanical (QM) description of a system. As a result a QM correction is added to the relative binding free energy obtained from a thermodynamic integration calculation within the MM description. This approach was combined with an electrostatic embedding model within ONETEP and used to calculate the hydration energies of small molecules.

As well as the computation of more accurate energies, large-scale DFT calculation compute the electron density of the entire system. Using electron density analysis approaches, such as the Hirshfeld density analysis, in combination with energy decomposition approaches, such as a second order perturbation estimate of natural bond orbital interactions, both qualitative and quantitative understandings can be gained into the contributions of particular chemical functional groups

that define protein-ligand interactions. These two approaches where applied to study complexes of the Phosphodiesterase type 5 protein and used to rank ligand binding affinities that agree well with then experimentally observed trends.

Contents

List of Figures

List of Tables

Declaration of authorship

I, Stephen John Fox, declare that the thesis entitled "protein-ligand binding affinities from large-scale quantum mechanical simulations" and the work presented in the thesis are both my own, and have been generated by me as the result of my own original research. I confirm that:

- this work was done wholly or mainly while in candidature for a research degree at this University;
- where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- where I have consulted the published work of others, this is always clearly attributed;
- where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- parts of this work have been published as papers.
  - First principles-based calculations of free energy of binding: application to ligand-binding in a self-assembling superstructure S. Fox, H. Wallnoefer, T. Fox, C. Tautermann , and C.-K. Skylaris. *J. Chem. Theor. Comput.* **7** (2011) 1102.

- Electrostatic embedding in large-scale first principles quantum mechanical calculations on biomolecules. S. J. Fox., C. Pittock, T. Fox, C. Tautermann, N. Malcolm, and C.-K. Skylaris. *J. Chem. Phys.* **135** (2011) 224107.

- Large-scale DFT calculations in implicit solvent - a case study on the T4 lysozyme L99A/M102Q protein. J. Dziedzic, S. J. Fox, T. Fox, C. S. Tautermann, and C.-K. Skylaris. *Int. J. Quant. Chem.* (2012).

Signed: . . . .

Date: . . . .

Acknowledgements

Domo Arigato Gozaimasu

“thank you for instructing me”

Many people have offered me support over the last four years:

First I would like to give my sincere gratitude to my academic supervisor, Dr Chris-Kriton Skylaris, for all the help, knowledge, and patience he has given me, and without which I would not have achieved all I have. I would like to thank my Industrial supervisors, Dr Thomas Fox and Dr Christofer S. Tautermann, for all of their expertise, knowledge, insight, and enthusiasm for this project which always renewed my interest whenever it wavered. I very much enjoyed my time spent working with them. I would also like to thank Clara Christ from Boehringer Ingelheim for her co-supervision of the QM corrected TI work, as well as Prof. Jonathan W. Essex for his help with the same project.

None of this work would have been possible without the financial support from BBSRC and Boehringer Ingelheim who provided the studentship. Equally important was the technical support provided by iSolutions at Southampton University for maintaining and managing of the desktops and supercomputers. The computational capacity provided by Iridis 3 has been of great value and without which the calculations of the size carried out in this project would not have been possible.

The daily life in the office has been made fun and enjoyable by other members of the Skylaris group, both old and new, as well as the members of the Essex and Khalid groups, making even the hard and stressful times seem easier.

I have been privileged with good friends I have made outside of studying. In particular, those I climb with: Montse Shelbourne, Nathan Bartlett, George Attard, Rachel Mills, and Adam Wheeler, for many happy weekends stuck to the side of a Dorset sea cliff or precariously balancing on a grit stone slab in the Peak District. The people I have met through martial arts training have been very humbling, showing great character, discipline, dedication, and skill. Without this group of friends, my stress level would be considerably higher, and my pain threshold would not be what it is today. I would like to mention a special thanks to a good friend, Jonathan Flavell, for his friendship, jokes, and the inappropriate links he has sent me over the years.

My family have been very supportive, both financially and emotionally, and have been for all my life so far. So my love and thanks go to my Mum, Dad, Sister, Grandad, Lilly, and Chris and Dawn.

But the person I would like to thank most is Lisa Williams, for making life so much fun over the last five years.

Chapter 1

Introduction

Many illnesses are due to excessive expression of a protein. Drugs (ligands) are designed to block the "active site" of these proteins, and inhibit their activity. Protein-ligand binding was originally thought to follow a "lock and key" mechanism first proposed by Emil Fischer in 1984 (shown in Figure 1.1), with the active site of the protein being the "lock", and the ligand being the "key". Both were believed to be in a fixed geometry and joined neatly without any geometric rearrangements necessary. Based on recent research, this idea has evolved, and now it is believed that the binding of a ligand to a protein often requires dynamic changes between different conformations. This can be by an induced-fit mechanism [8], or through a selected-fit mechanism [9]. An induced-fit mechanism forces the protein conformation to change as the ligand binds to the cavity. Whereas with the selected-fit mechanism, the ligand selects and stabilises a complementary protein conformation from an equilibrium of low-energy and high-energy conformations. These conformational changes can be documented by experimentally determined structures of the protein in the unbound state, and with a ligand bound

[10, 11].

Protein-ligand interactions can be typically classed as van der Waals (vdW), and hydrogen bonding (H-bonding) and more general electrostatic interactions such as dipole-dipole. For a protein/enzyme to function properly, the natural substrate must bind strongly, and stay bound long enough to fulfill its function, but not so strongly that it renders the protein/enzyme useless and unable to perform any further processes. In contrast, a strong interaction that renders the protein useless is necessary for a drug molecule to work. In addition, the protein-ligand interaction must outcompete the natural protein-substrate interaction, thus inhibiting the biological activity of the protein.

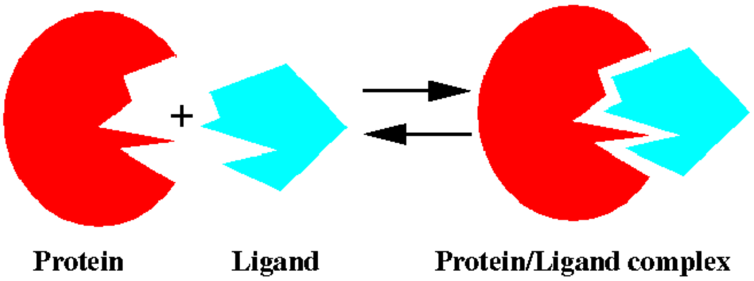


Figure 1.1: A schematic diagram of the lock and key mechanism of inhibitor binding [1]

With the growing accuracy of experimentally determined 3D molecular structures, refined to an atomic resolution, computational molecular modelling is expanding its role in understanding structure/function relationships of biomolecules. There are a variety of simulation techniques available which are based either on classical molecular mechanics (MM), using atomistic [12] or coarse grained models [13], or *ab initio* quantum mechanical calculations (QM) [14], or combinations thereof. These methods allow us to study protein dynamics [15, 16, 17], and estimate the interaction strengths and binding geometries (poses) of ligands [18, 19, 20].

Computational simulations can provide a useful tool for assessing a wide range of potential pharmaceutical drugs whilst also limiting the need for expensive laboratory tests. As well as being cheaper, the capability of computer simulations to evaluate drugs that have not yet been synthesised can significantly speed up the process.

The aim of using computational methods in drug design is to accurately calculate the properties of a molecule, in particular the free energy. These explicitly depend on the movement of electrons upon structural changes caused by ligand binding, resulting in polarisation and charge transfer. The ideal computational method would be a QM approach, which can describe all properties of the system via the wavefunction. With QM we can explicitly account for electrons and therefore calculate all the electronic properties of a molecule. However, with the desire for accurate and reliable QM calculations, two main problems prevent their application in biological systems. Firstly, the system size, which may reach tens of thousands of atoms, and secondly, the required sampling of configurational space. QM calculations from first principles, using conventional Kohn-Sham Density Functional Theory (DFT) [21], are limited to a few hundred atoms due to the scaling of their computational cost with the number of atoms in the system. For this reason, simulations on biomolecules are usually performed using classical MM. MM approaches are based on empirical knowledge, and implicitly account for the electronic charge on atoms by the addition of a partial charge. This limits the transferability of MM approaches and introduces errors: the neglect of electrons in force fields leads to the inability to properly describe polarisation or to account for electron transfer. In some cases, a combination of both approaches can be used in an attempt to take advantage of both. In these hybrid QM/MM approaches, small parts of the active site are simulated using QM, and embedded in a MM described system [22, 23]. In other cases semi-empirical QM methods are used to describe the system [24]. In semi-empirical approaches, the single electron terms are explicitly accounted for, however, the two electron terms are treated empirically. Values for these are taken from high level *ab initio* QM calculations on small molecules in similar ways to entirely empirical MM approaches. Like MM methods however, this also limits the transferability of the approach, whilst still increasing the computational cost.

For the calculation of free energy, or reaction paths, dynamical movement of a molecule (at room temperature) must be accounted for. As mentioned above, QM methods are currently limited in system size, however simulation time scales also present an issue. To simulate chemically interesting phenomena, often "long" time scales are required (tens to hundreds of nano seconds or longer, especially for chemical processes), wherein millions of molecular dynamics time steps are needed to accurately describe such processes. QM calculations are generally 1000 times more computationally expensive than MM, and for simulations of the required time scale would take far too long. So, to simulate dynamical movement of proteins, MM must still be used.

Hybrid QM/MM approaches are a step in the direction of multiscale methods for increasing the use of QM methods, and semi-empirical approaches are allowing faster and larger "QM" calculations. Another direction is the development of linear-scaling DFT, which has the capability of performing calculations of a much larger scale than conventional DFT approaches, and it can be applied to systems containing several thousand atoms. Albeit, dynamic processes over long time scales are still out of reach.

1.1 Research aims

This research project was aimed at utilising large-scale quantum mechanics simulations to study protein-ligand interactions using the linear-scaling density functional code ONETEP in combination with conformation sampling from MM methods.

The first method investigated in this thesis considered the prediction of protein-ligand binding free energies using the Molecular Mechanics Poisson-Boltzmann Surface Area (MM-PBSA) approach. This approach obtains the free energy of binding as a sum of the differences in energies of the complex, receptor and ligand, and the differences in solvation energies, averaged over a structural ensemble taken from a molecular dynamics simulation. Conventional methods use classical force fields to evaluate the energies in vacuum, and the PBSA implicit solvent model for the solvation energies. We have used the ONETEP program to evaluate the energy of the classically derived conformations, and used a minimal implicit solvation model to calculate the solvation energies. This "QM-PBSA" approach was first tested, and validated, on a self assembling dimer, the "Tennis ball" dimer. This system was chosen as it is considered an idealised model of a protein cavity. This approach was then applied to a larger model protein-ligand binding system, the T4 lysozyme double mutant L99A/M102Q.

The MM-PBSA approach is a mid-rigour theoretical free energy approach. We also endeavoured to expand the use of our large-scale QM calculations to more rigorous free energy approaches. With this in mind, we investigated applying a QM correction to a thermodynamic integration method. In principle this approach exactly includes all entropy change, whereas MM-PBSA does not. Thermodynamic integration uses a non-physical pathway to mutate a ligand (A), to another ligand (B). This can be done since thermodynamic functions, such as free energy, are a state property, and so is independent of the path taken to reach that point. The relative free energy can hence be obtained via a thermodynamic cycle. To do this, the mutation is performed for ligand bound to a protein (PA→PB), and free in solution (A→B). This approach was extended by adding a QM correction to the end points of the cycle. The classical and quantum descriptions are different thermodynamic states. To calculate the free energy change from the classical description of the system to a quantum description of the system single step perturbations were performed. Using this extended free energy cycle we computed the relative hydration free energies of several small aromatic molecules.

The final branch of this project moved away from predicting binding free energies and investigated some of the additional information that can be obtained from large-scale QM calculations. We used energy decomposition approaches, and electron density analysis, to study the interactions of several very structurally similar ligands bound to the Phosphodiesterase Type 5 protein. Natural bond orbitals were obtained from ONETEP, and second order perturbation estimates of hydrogens bond strengths between the ligands and the protein were obtained. The electron density was also partitioned via a Hirshfeld analysis to investigate electron redistribution on ligand binding. These approaches provide useful chemical insight s into the interactions at the atomistic level. In combination with the computed binding energies, dispersion interaction, and ligand desolvation energies, qualitative predictions of relative ligand binding affinities can be made.

“However much I study, it is never enough”

Takamatsu Sensei

Chapter 2

Computational Theories

This chapter will detail the computational theories used during this PhD. From the different ways of describing the energy of a molecule and the prediction of the available phase space of a molecule, to the different approaches of accounting for solvent and various ways of estimating binding free energies of host-guest systems.

2.1 Quantum Mechanics

Quantum physics is based on laws discovered in the early 20th century. Unlike classical physics, laws that were developed pre-1900, which explain everyday things in our mesoscopic world, quantum physics explains the microscopic.

At the beginning of the 20th century, De Broglie developed the idea of the dual wave/particle nature of particles. In this theory, all particles that have momentum, can also display wave-like properties described by a wave length (λ), and waves can have particle-like properties.

$$\lambda = \frac{h}{p} \Longleftrightarrow p = \frac{h}{\lambda}. \tag{2.1}$$

The de Broglie hypothesis motivated the discovery of the Schrödinger equation, a fundamental principle which underpins all of chemistry.

The inability of classical physics to describe microscopic particles, such as electrons, protons and neutrons, can be explained by Heisenberg's uncertainty principle [25], which states that the more precise the position (x) of a particle is known, the less precise its momentum will be (p_x),

$$\Delta x \Delta p_x = \frac{\mathrm{h}}{4\pi}, \tag{2.2}$$

where h is Planks constant.

In quantum mechanics (QM), the energy is a quantised property, this can be observed when conducting electronic excitation experiments using electromagnetic waves. From quantum theory we obtain the fundamental laws of chemistry, as well as explanations for the properties of materials. QM can be used to study biological structures and mechanisms for understanding and clarifying their role in several life processes, as well as studying nanostructures and materials.

2.1.1 The Schrödinger equation

In classical mechanics, variables can be directly linked to a physically measurable property (observables), such as momentum or position. This is not the case for QM. Instead these observables are related by "operators" which provide the value of a physical property when they act upon a wavefunction. The postulates of QM assert that microscopic systems are described by wavefunctions (ϕ) that can completely characterise all physical properties of the system. There exist quantum mechanical operators corresponding to each physical observable, which when applied to the wavefunction, allow the prediction of the probability of finding the system exhibiting a particular value (or range of values) for that observable. The (non-relativistic) time independent Schrödinger equation [26] which describes physical properties of the system is given by,

$$\hat{H}\psi = E\psi, \tag{2.3}$$

where ψ is the many-body wavefunction for the N particles in the system and is a function of the particle coordinates. E is the total energy eigenvalue for the system and $\hat{H}$ is the Hamiltonian (energy) operator, which combine the kinetic energy operator ($\hat{T}$) and the potential energy operator ($\hat{V}$), for the system and takes the form (for a molecular system),

$$\begin{aligned}
\hat{H} &= \hat{T} + \hat{V} && (2.4)\\
&= \hat{T}_N\{\mathbf{R_I}\} + \hat{T}_e\{\mathbf{r_i}\} + \hat{V}_{NN}\{\mathbf{R_I}\} + \hat{V}_{Ne}\{\mathbf{R_I}, \mathbf{r_i}\} + \hat{V}_{ee}\{\mathbf{r_i}\} && (2.5)\\
&= -\frac{\hbar^2}{2m_n}\sum \nabla^2_{\mathbf{R_I}} - \frac{\hbar^2}{2m_e}\sum \nabla^2_{\mathbf{r_i}} + \frac{1}{2}\sum_{I=1}^{N}\sum_{J\neq I}^{N} \frac{Z_I Z_J e^2}{4\pi\epsilon_0 \mathbf{R_{IJ}}} \\
&\quad -\frac{1}{2}\sum_{i=1}^{n}\sum_{I=1}^{N} \frac{Z_I e^2}{4\pi\epsilon_0 \mathbf{R_{Ii}}} + \frac{1}{2}\sum_{i=1}^{n}\sum_{j\neq i}^{n} \frac{e^2}{4\pi\epsilon_0 \mathbf{r_{ij}}}. && (2.6)
\end{aligned}$$

The first two terms in Equation 2.5 are the kinetic energy operators for the nuclei (N and $\mathbf{R_I}$) and the electrons (e and $\mathbf{r_i}$). The last three are the potential energy operators, where $\hat{V}_{nn}$ is for nuclear-nuclear repulsion, $\hat{V}_{ne}$ is for nuclear-electron attraction, and $\hat{V}_{ee}$ is for electron-electron repulsion. These are expanded in Equation 2.6. Where N is the number of atoms, n is the number of electrons, Z_I is the atomic charge of atom I, $\mathbf{r_{ij}}$ is the distance between electrons i and j, $\mathbf{R_{Ii}}$ is the distance between nucleus I and electron i, and $\mathbf{R_{IJ}}$ is the distance between nuclei I and J. The above equation for the Hamiltonian takes into account all interactions between particles in the system. Since there are 3N spatial degrees of freedom for atomic positions, equation (2.3) is very complex and difficult, if not impossible to solve analytically for all but the simplest (harmonic oscillator, hydrogen atom, etc) systems.

2.1.2 Born–Oppenheimer approximation

An approximation that is central to quantum chemistry is the Born–Oppenheimer approximation [27]. Electrons are much lighter than nuclei (the mass of a electron is 9.109x10^{-31} kg [28] compared to the mass of a proton of 1.672x10^{-27} kg [28], around $1/2000$ of the mass) and hence move much faster. This means they can be assumed to instantly re-arrange themselves to any nuclear movement. A simple approximation then could be to separate the motion of the nuclei from the electrons, and calculate the wavefunction of the electrons moving in a field of fixed nuclei. Using this approximation, the Hamiltonian becomes the electronic Hamiltonian (the Hamiltonian describing the motion of N electrons in a potential field of point charges).

$$\hat{H}_{\text{elec}} = \hat{T}_e\{\mathbf{r_i}\} + \hat{V}_{Ne}\{\mathbf{R_I}, \mathbf{r_i}\} + \hat{V}_{ee}\{\mathbf{r_i}\}. \tag{2.7}$$

The kinetic energy of the nuclei can now be neglected and the nuclear-nuclear repulsion is a constant. A constant that is added to an operator only adds to the eigenvalues and has no effect on the eigenfunctions. The Schrödinger equation using this approximations is then,

$$\hat{H}_{\text{elec}}\psi_{\text{elec}} = E_{elec}\psi_{\text{elec}}. \tag{2.8}$$

The eigenfunctions are now the electronic wavefunction which explicitly depends on electronic coordinates, and parametrically on the nuclear coordinates, as does the electronic energy. By parametric dependence we mean that, for different arrangements of nuclear coordinates, the eigenfunction is a different function of electronic coordinates.

2.1.3 Potential energy surfaces

The total energy of the system is the electronic energy plus the constant nuclear-nuclear repulsion,

$$E_{\text{PES}} = E_{\text{elec}} + \sum_{A=1}^{M}\sum_{B>A}^{M}\frac{Z_A Z_B}{R_{AB}}. \tag{2.9}$$

By varying the nuclear coordinates and calculating the electronic energy, a potential energy surface (PES) can be obtained. E_{PES} is also parametrically dependent on the nuclear positions. This can be done to find equilibrium bond lengths and stable geometry conformations (reactants and products) which are found at energy minima. As well as transition states, which are found at saddle points (a maximum in one direction and a minimum in all other directions). Figure 2.1 shows how the equilibrium bond length is located by moving two atoms further apart and plotting E$_{PES}$ at each point. PES is also used for chemical reactions as shown in figure 2.2.

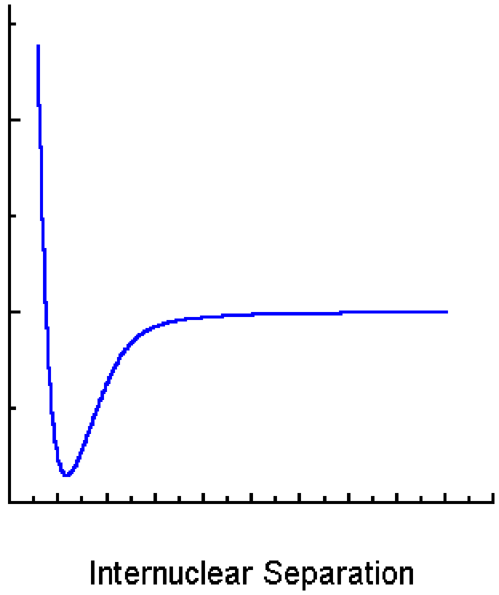


Figure 2.1: PES for a diatomic. Plots E$_{PES}$ as a function of internuclear separation (the distance between nuclei). [2]

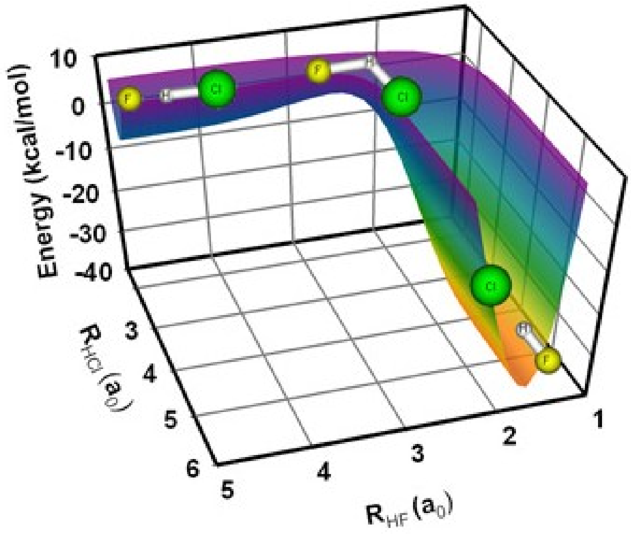


Figure 2.2: PES of a reaction [2].

2.1.4 The variational principle

The variational principle provides an approximation of the ground-state energy, which is the lowest eigenvalue given by the Hamiltonian for a system. It states that any well-behaved, normalised, approximate trial function ϕ, that satisfies the same boundary conditions as the wavefunction ψ, will give an expectation value ε of the Hamiltonian which is greater or equal to the exact ground state ϵ_0. This can be expressed as,

$$\varepsilon = \left\langle \phi | \hat{H} | \phi \right\rangle \geq \epsilon_0. \tag{2.10}$$

The convergence towards ϵ_0 is achieved by varying the parameters of the normalised function (ϕ) in order to minimise $\left\langle \phi | \hat{H} | \phi \right\rangle$ [29].

2.1.5 Wavefunction approaches for approximating the Schrödinger equation

The Hartree-Fock (HF) method is an approximate method for solving the molecular electronic Schrödinger equation (Equation 2.7). It breaks up the many-electron problem into a series of single electron problems. The wavefunction in Hartree-Fock theory is expressed as an anti-symerised product of n (number of electrons) orthonormal spin orbitals ($\chi(\mathbf{x})$), known as a Slater determinant. A spin orbital is a product of a spatial orbital ($\psi(\mathbf{r})$) and a spin function (either $\alpha(\mathbf{s})$ or $\beta(\mathbf{s})$).

The single electron problem, with electron-electron interactions, is expressed using the Fock operator,

$$\begin{aligned}\hat{f}_i &= -\frac{1}{2}\hat{\nabla}_i^2 - \sum_{I=1}^{N_{\text{nuc}}} \frac{Z_I}{R_{iI}} + \hat{\nu}_{HF,i} \\ &= \hat{h}_i + \hat{\nu}_{HF,i}, \end{aligned} \tag{2.11}$$

where $\hat{\nu}_{HF,i}$ is the average potential felt by electron i due to the other electrons. This energy is expressed as a sum of Coulomb integrals,

$$J_{ij} = \int\int \chi_i^*(\mathbf{x_1})\chi_j^*(\mathbf{x_2})r_{ij}^{-1}\chi_i(\mathbf{x_1})\chi_j(\mathbf{x_2})d(\mathbf{x_1})d(\mathbf{x_2}), \tag{2.12}$$

and exchange integrals,

$$K_{ij} = \int\int \chi_i^*(\mathbf{x_1})\chi_j^*(\mathbf{x_2})r_{ij}^{-1}\chi_j(\mathbf{x_1})\chi_i(\mathbf{x_2})d(\mathbf{x_1})d(\mathbf{x_2}). \tag{2.13}$$

The exchange term appears because of asymmetric products in the wavefunction. This arises from the Pauli exclusion principle that states that no two electrons can have identical quantum numbers. This principle applies to all particles with a half integer spin (fermions).

The Fock operator for a one-electron problem in terms of spatial orbitals can be written as,

$$\hat{f}_i(\mathbf{r_i}) = \hat{h}_i + \sum_{j}^{N/2} \left(2\hat{J}_j(\mathbf{r_i}) - \hat{K}_j(\mathbf{r_i})\right), \tag{2.14}$$

with the energy being,

$$E_0 = 2\sum_{i}^{N/2} \int \psi_i^*(\mathbf{r_i})\hat{h}\psi_i(\mathbf{r_i})d\mathbf{r_i} + \sum_{ij}^{N/2} \left(2J_{ij} - K_{ij}\right). \tag{2.15}$$

This method provides a way of approximating the wavefunction, which is limited to be Slater determinants, that minimises the energy, as in Equation 2.10.

Post HF methods

The difference between the exact HF energy and the exact ground state is known as the correlation energy. Many approaches exist that try and improve the energy obtained from HF. These methods improve the energy by attempting to including some of this correlation energy [29]. Examples of such methods are coupled cluster (CC), configuration interaction (CI), and Møller-Plesset perturbation theory (MP2, MP4).

2.1.6 Density functional theory

Density functional theory (DFT) has gained much popularity during the last 20 years in quantum chemistry. Its popularity is owed to its ability to accurately and reliably predict the ground-state properties of many molecular systems with only a small number of well controlled approximations.

For an N-electron system, the single-particle electron density is the square of the wavefunction integrated over $N-1$ electron coordinates multiplied by the number of electrons,

$$n(\mathbf{r}) = N \int |\Psi(\mathbf{r}, \mathbf{r_2}, \mathbf{r_3}, \cdots, \mathbf{r_N})|^2 d\mathbf{r_2} d\mathbf{r_3} \cdots d\mathbf{r_N}, \tag{2.16}$$

which only depends on three coordinates, independent of the system size. In contrast, wave function approaches depend on 3N coordinates. What this means is that as the complexity of a wavefunction increases with the number of electrons, the electron density has a constant number of variables, three. This makes the method very computationally appealing.

The basis of DFT, as given by Hohenberg and Kohn [30], also referred to as "pure" DFT, is that the ground state electronic properties, including the energy, can be completely described by the electron density.

DFT theorems

Hohenberg and Kohn proved that the external potential is determined by the density. Their theorems prove that the exact calculation of the ground state of an N-electron system is possible by only using the electron density. Their formulation can be applied on any stationary, nonrelativistic many-particle system in an external potential $\nu_{\text{ext}}(\mathbf{r})$ and determines all the properties of the ground state.

Hohenberg and Kohn first theorem

States that, "The external potential $\nu_{\text{ext}}(\mathbf{r})$ is a unique functional of the density $n(\mathbf{r})$ (apart from a trivial additive constant)." [30]

Proof:

Assume there are two external potentials, $\hat{\nu_1}(\mathbf{r})$ and $\hat{\nu_2}(\mathbf{r})$ that both give rise to the same ground-state density $n(\mathbf{r})$. This will mean that there are two Hamiltonians $\hat{H}_1$ and $\hat{H}_2$, with two different wavefunctions ψ_1 and ψ_2 but with the same ground state density. Using the variational principle,

$$\begin{aligned} E_1^0 < \left\langle \psi_2 | \hat{H}_1 | \psi_2 \right\rangle &= \left\langle \psi_2 | \hat{H}_2 | \psi_2 \right\rangle + \left\langle \psi_2 | \hat{H}_1 - \hat{H}_2 | \psi_2 \right\rangle \\ &= E_2^0 + \int n(\mathbf{r}) \left[\nu_1(\mathbf{r}) - \nu_2(\mathbf{r}) \right] d\mathbf{r} \qquad (2.17) \\ E_2^0 < \left\langle \psi_1 | \hat{H}_2 | \psi_2 \right\rangle &= \left\langle \psi_1 | \hat{H}_1 | \psi_1 \right\rangle + \left\langle \psi_1 | \hat{H}_2 - \hat{H}_1 | \psi_1 \right\rangle \\ &= E_1^0 + \int n(\mathbf{r}) \left[\nu_2(\mathbf{r}) - \nu_1(\mathbf{r}) \right] d\mathbf{r}, \qquad (2.18) \end{aligned}$$

where E_1^0 and E_2^0 are the ground state energies for $\hat{H}_1$ and $\hat{H}_2$. Adding these inequalities together gives $E_1^0 + E_2^0 < E_2^0 + E_1^0$, which is a contradiction. This result shows that the assumption that a $\nu_2(\mathbf{r})$ exists is wrong, and hence $\nu_{\text{ext}}(\mathbf{r})$ is uniquely determined by $n(\mathbf{r})$. The groundstate wavefunction is, therefore, a functional of $n(\mathbf{r})$, as is the kinetic energy operator $\hat{T}_e$ and the electronic repulsion energy operator $\hat{V}_{ee}$, since $\hat{H}$ is fixed by $n(\mathbf{r})$. The Hohenberg-Kohn functional is defined for these parts of the Hamiltonian by,

$$F_{HK}[n] = \left\langle \psi | \hat{T}_e + \hat{V}_{ee} | \psi \right\rangle . \qquad (2.19)$$

This functional is universal, however its explicit form is unknown. The total energy functional can now be expressed as,

$$E[n] = \int \nu_{\text{ext}}(\mathbf{r}) n(\mathbf{r}) d\mathbf{r} + F_{HK}[n(\mathbf{r})] . \qquad (2.20)$$

The Hohenberg and Kohn second theorem

States that, “The energy functional, $E[n]$, has as its minimum the exact ground state energy associated with $\nu_{\text{ext}}(\mathbf{r})$ if the density is constrained to preserve the number of particles $(N[n] = \int n(\mathbf{r}) dr = N_{elec})$.” [30]

Proof:

For any number of electrons, N_{elec} and external potential $v_{\text{ext}}(\mathbf{r})$ the density functional,

$$E_{(\nu_{\text{ext}},N)}[n] = F_{HK}[n(\mathbf{r})] + \int \nu_{ext}(\mathbf{r})n(\mathbf{r})d^3r, \tag{2.21}$$

obtains its minimal value at the ground-state density. The minimal value of $E_{(\nu_{\text{ext}},N)}[n]$ is then the ground state energy of this system.

This is true since the electronic energy functional of the density (Equation 2.21) is, by definition, equal to the energy functional of the wavefunction,

$$\epsilon_v\left[\psi\right] = \langle\psi|V_{ne}|\psi\rangle + \left\langle\psi|\hat{T}_e + \hat{V}_{ee}|\psi\right\rangle. \tag{2.22}$$

Since $\epsilon_v\left[\phi\right]$ is the ground state energy when ψ is the ground state for our N_e system, then

$$\epsilon_v\left[\phi\right] \geq E_0, \tag{2.23}$$

and hence,

$$E_{(\nu_{\text{ext}},N)}[n] \geq E_0. \tag{2.24}$$

This theorem introduces the variational principle, allowing the density to be used as the variable for minimising the energy. This allows the density to be used as a basic variable in quantum chemistry calculations. It does however restrict the theory to ground states only, since the inequality can only be proved for the ground state.

The electronic Hamiltonian, in atomic units, is

$$\hat{H}_{elec} = -\sum_{i}^{N} \frac{1}{2}\nabla_i^2 + \sum_{i}^{N} \nu_i(\mathbf{r}) + \sum_{i<j}^{N} \frac{1}{r_{ij}}, \tag{2.25}$$

where $\nu_i(\mathbf{r})$ is the external potential (ν_{ext}).

As with wave mechanics approaches, the energy functional can be split into kinetic energy, T$[n]$, electron-nucleus attraction, E$_{ne}[n]$ and the electron-electron repulsion, E$_{ee}[n]$,

$$\begin{aligned} E[n] &= T_e[n] + E_{ne}[n] + E_{ee}[n] && (2.26) \\ &= \int n(\mathbf{r})\nu_{ext}(\mathbf{r})d\mathbf{r} + \langle\psi|\hat{T}_e + \hat{V}_{ee}|\psi\rangle \\ &= \int n(\mathbf{r})\nu_{ext}(\mathbf{r})d\mathbf{r} + F_{HK}[n]. && (2.27) \end{aligned}$$

To minimise the energy with respect to the density, with the constraint $\int n(\mathbf{r})d\mathbf{r} = N_{\text{elec}}$, we need to find,

$$\frac{\delta E[n]}{\delta n(\mathbf{r})} - \mu = 0, \tag{2.28}$$

where $\mu = \nu(\mathbf{r}) + \frac{\delta F[n]}{\delta n(\mathbf{r})}$, is the Euler-Lagrange equation which can be solved for the exact density. E$_{ee}[n]$ can be further divided into a Coulomb and exchange part, J$[n]$ and K$[n]$, and a part for the electron correlation that has no explicit form.

The Coulomb and electron-nucleus terms are both described by their classical representations,

$$E_{ne}[n] = \sum_{\alpha} \int \frac{Z_a n(\mathbf{r})}{|\mathbf{R}_\alpha - \mathbf{r}|} d\mathbf{r} \tag{2.29}$$

$$J[n] = \frac{1}{2} \int \int \frac{n(\mathbf{r}) n(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d\mathbf{r}\mathbf{r}'. \tag{2.30}$$

The problem lies with defining the kinetic energy of the electrons, the exchange energy and the electron correlation energy. The first attempts to create functionals for the kinetic energy, T$[n]$, and exchange energy, K$[n]$, considered a non-interacting uniform electron gas.

Thomas-Fermi theory

In Thomas-Fermi theory the kinetic energy functional is expressed as,

$$T_{TF}[n] = C_F \int n^{\frac{5}{3}}(\mathbf{r}) d\mathbf{r}, \tag{2.31}$$

where,

$$C_F = \frac{3}{10} \left(3\pi^2\right)^{\frac{2}{3}}. \tag{2.32}$$

In Thomas-Fermi-Dirac model, the exchange energy is included and given by,

$$K_D[n] = -C_x \int n^{\frac{4}{3}}(\mathbf{r}) d\mathbf{r}, \tag{2.33}$$

where,

$$C_x = \frac{3}{4} \left(\frac{3}{\pi}\right)^{\frac{1}{3}}. \tag{2.34}$$

This method is inaccurate (total energy errors of $15 - 50\%$) due to the assumption of a uniform electron gas. A serious flaw with this method is its inability to predict bonding in molecules. The error lies with the inadequate description of the kinetic energy. An improvement to these methods is made by improving T$[n]$ and K$[n]$. This can be done by making them dependent not just on the density, but also it derivatives. Although this is an improvement, bonding is now allowed, it still doesn't yield results comparable to wave function methods, or produce chemically useful results.

Kohn-Sham Theory

The Kohn-Sham (KS) [21] reformulation of DFT made it as successful in computational chemistry as it is today. They considered the kinetic energy for a system of non-interacting electrons using molecular wavefunctions, whose expression is known exactly.

The interacting electronic Hamiltonian is given by,

$$\hat{H}_{elec} = \sum_{i=1}^{N_{elec}} -\frac{1}{2}\nabla_i^2 + \sum_{i=1}^{N_{elec}} \left(\sum_{A=1}^{N_{nuclei}} \frac{-Z_A}{|\mathbf{r}_i - \mathbf{R_A}|} \right) + \sum_{i=1}^{N_{elec}} \sum_{j=i+1}^{N_{elec}} \frac{1}{|\mathbf{r}_i - \mathbf{r}_j|}. \tag{2.35}$$

The last term represents the electron-electron interactions which is a two-electron operator. In a non-interacting system this term is replaced by a one-electron operator, $\hat{V}_{av}$, that describes the average effect of the interaction,

$$\hat{H}_{elec} = \sum_{i=1}^{N_{elec}} -\frac{1}{2}\nabla_i^2 + \sum_{i=1}^{N_{elec}} \left(\sum_{A=1}^{N_{nuclei}} \frac{-Z_A}{|\mathbf{r}_i - \mathbf{R_A}|} \right) + \sum_{i=1}^{N_{elec}} \hat{V}_{av}(\mathbf{r}_i). \tag{2.36}$$

$$\begin{aligned}
\hat{H}_{elec} &= \sum_{i=1}^{N_{elec}} -\frac{1}{2}\nabla_i^2 + \sum_{i=1}^{N_{elec}} \hat{V}_{ext}(\mathbf{r}_i) + \sum_{i=1}^{N_{elec}} \hat{V}_{av}(\mathbf{r}_i) \\
&= \sum_{i=1}^{N_{elec}} -\frac{1}{2}\nabla_i^2 + \sum_{i=1}^{N_{elec}} \{\hat{V}_{ext}(\mathbf{r}_i) + \hat{V}_{av}(\mathbf{r}_i)\} \\
&= \sum_{i=1}^{N_{elec}} -\frac{1}{2}\nabla_i^2 + \sum_{i=1}^{N_{elec}} \hat{V}_{eff}(\mathbf{r}_i) \\
&= \sum_{i=1}^{N_{elec}} \{-\frac{1}{2}\nabla_i^2 + \hat{V}_{eff}(\mathbf{r}_i)\} \\
\hat{H}_{elec} &= \sum_{i=1}^{N_{elec}} \hat{h}(\mathbf{r}_i). \qquad (2.37)
\end{aligned}$$

The Hamiltonian is then expressed as a sum of one-electron operators as shown in equation 2.37. In effect making it a Hamiltonian for a non-interacting system of electrons, with eigenfunctions that are Slater determinants of one-electron eigenfunctions (molecular orbitals) and eigenvalues that are a sum of one-electron eigenvalues. This system of non-interacting electrons is constructed in such a way that it has a ground state electron density which is the same as the real system (where electrons do interact). Equation 2.8 can be written for each one-electron Hamiltonian separately,

$$\hat{h}(\mathbf{r})\psi_\alpha(\mathbf{r}) = \varepsilon_\alpha \psi_\alpha(\mathbf{r}), \qquad (2.38)$$

where the eigenvalues are simply the energy of the non-interacting orbitals.

By using the system of non-interacting electrons, the kinetic energy for the real system is thus split into two terms. A term for the non-interacting system that can be solved exactly ($2\sum_{i=1}^{N_{elec}/2}\langle\psi_i|-\frac{1}{2}\nabla_i^2|\psi_i\rangle$), and a small correction term deriving from the interacting nature of electrons. The energy functional can now be divided up as,

$$E[n] = T_e[n] + E_{ne}[n] + E_{ee}[n] \tag{2.39}$$

$$= T_s[n(\mathbf{r})] + E_{ne}[n] + E_H[n] + \Delta T[n] + \Delta E_{ee}[n], \tag{2.40}$$

where is $T_s[n]$ the kinetic energy of the non-interacting system with the density $n(\mathbf{r})$, $E_{ne}[n]$ is the electron-nucleus attraction, $E_H[n]$ is the Coulomb energy (or Hartree energy), $\Delta T[n]$ is the difference between the kinetic energy of the interacting electrons and the non-interacting electrons, and $\Delta E_{ee}[n]$ is the non-classical corrections to the electron-electron repulsion. This equation can be rewritten as,

$$E[n] = 2 \sum_{i=1}^{N_{elec}/2} \langle \psi_i | - \frac{1}{2} \nabla_i^2 | \psi_i \rangle + \int \hat{\nu}_{\text{ext}}(\mathbf{r}) n(\mathbf{r}) d\mathbf{r} + J[n] + E_{xc}[n], \tag{2.41}$$

where the last two terms in equation (2.40) have been combined in the term $E_{xc}[n]$, the exchange-correlation functional. If the exchange-correlation energy is exact, then the exact density and electronic energy for the interacting system are obtained.

2.1.7 Exchange-correlation functional

The exchange-correlation functional, E$_{xc}$, accounts for the difference between classical and quantum electron-electron repulsion, and the difference in kinetic energy between the fictitious non-interacting system and the real interacting system. If the exact exchange-correlation functional was known, KS DFT would provide the exact total energy. This gives DFT the potential to provide the correlation energy, the computationally difficult part, at the computational cost similar to uncorrelated techniques. The problem is that we do not have expressions for the exchange-correlation functional.

It is common to separate the exchange-correlation functional into its two parts,

$$E_{xc}\left[n\right] = E_x\left[n\right] + E_c\left[n\right]. \tag{2.42}$$

The exchange and the correlation parts are then treated separately. There are many different functionals available for the approximation of the exchange-correlation energy, some of which will be discussed below.

Local Density Approximation

The local density approximation (LDA) is the simplest of the approximations used for the exchange energy functional. It uses an expression derived from a uniform electron gas (jellium). This can be considered as a large collection of N electrons in a volume V. The electron density is a constant over the volume and is balanced by a uniform positive background. This approach can be applied to a closed shell, spin-unpolarised system, where the density has the same value (or varies only very slightly) at every position. The exchange energy can be derived from the exact Hartree Fock exchange energy by expressing the orbitals in terms of plane waves and substituting in the density. The end result is an expression for the exchange energy where the density is position dependent,

$$E_x^{LDA}[n] = -\frac{3}{4}\left(\frac{3}{\pi}\right)^{\frac{1}{3}} \int n^{\frac{4}{3}}(\mathbf{r})d\mathbf{r}. \tag{2.43}$$

There are many different correlation functionals that have been made to complement this exchange functional. Of those available, two commonly used are CAPZ (Ceperley, Alder, Perdew and Zunger) [31] and VWN (Vosko, Wilk and Nusair) [32].

LDA methods have been used on many systems that behave similarly to jellium, such as bulk metallic systems, however they often proved disappointing when applied to molecular systems of chemical interest. LDA is surprisingly successful in some case, mainly molecular geometries, but has a serious problem; it significantly overbinds molecules. Due to this it is of limited use in chemistry.

Gradient correction methods or Generalised Gradient Approximations

Generalised Gradient Approximations (GGAs) have the form,

$$E_{XC} = \int F(n, \nabla n) d\mathbf{r}. \tag{2.44}$$

The obvious step to correct for the simplicity of LDA would be to add information about the density gradient. Doing this the exchange functional can be written as,

$$E_X[n] = \int n^{\frac{4}{3}}(\mathbf{r}) f(x(\mathbf{r})) d\mathbf{r}, \tag{2.45}$$

where $x(\mathbf{r}) = \frac{\nabla n(\mathbf{r})}{n^{\frac{4}{3}}(\mathbf{r})}$, and is dimensionless. A simple gradient expansion leads to a divergent exchange-correlation potential in finite systems, so Becke suggested the form [33],

$$f = C_X + \beta \left(\frac{x^2}{(1 + \gamma x^2)} \right), \tag{2.46}$$

which is also used in the PBE functional [34]. Becke made another functional form in 1988,

$$f = \frac{C_X + \beta x^2}{(1 + 6\beta x \operatorname{arcsinh} x)}, \tag{2.47}$$

which is the B88X functional. This functional was largely responsible for the increased interest in DFT in the 1990's. These are just two of the many exchange functionals that are currently available.

For the correlation functional, again there are many forms available. One of these is the Lee-Yang-Parr (LYP) functional [35], and has the form,

$$\begin{aligned} E_C = & -a \int \frac{n}{1 + dn^{-\frac{1}{3}}} d\mathbf{r} - ab \qquad (2.48) \\ & \int \omega n^2 \left[C_F n^{\frac{8}{3}} + |\nabla n|^2 \left(\frac{5}{12} - \delta \frac{7}{12} \right) - \frac{11}{24} |\nabla n|^2 \right] d\mathbf{r}, \qquad (2.49) \end{aligned}$$

where

$$\omega = \frac{\exp(-cn^{-\frac{1}{3}})}{1 + dn^{\frac{1}{3}}} n^{-\frac{11}{3}},$$

and

$$\delta = cn^{-\frac{1}{3}} + \frac{dn^{\frac{1}{3}}}{1 + dn^{\frac{1}{3}}}.$$

This functional was derived from the helium atom and has no relation to a uniform electron gas.

There are many GGA functionals available in the literature. These fall into two categories;

1. Semiempirical (designed to fit experimental data) eg BLYP.

2. Purely theoretical (determined by satisfying exact conditions) eg PBE.

Other functionals

There are two other types of functionals,

- Meta GGAs. These include higher order derivatives of the density, eg

  $$E_{XC} = \int F(n, \nabla n, \nabla^2 n) d\mathbf{r}. \tag{2.51}$$

  The improved accuracy often comes at the expense of numerical instabilities in the calculations.

- Hybrid functionals. These combine GGAs with a fraction of the exact exchange calculated using Kohn-Sham orbitals.

  $$E_{XC} = \int F(n, \nabla n) d\mathbf{r} + \xi E_X^{HF}, \tag{2.52}$$

  with E_X^{HF} being the exact ground state exchange energy. An example of a hybrid functional would be B3LYP [36, 37] which has the general form,

  $$E_{xc}^{B3LYP} = (1 - \alpha_0 - \alpha_x) E_x^{LSDA} + \alpha E_x^{HF} + \alpha E_x^{B88} + (1 - \alpha_c) E_c^{VWN} + \alpha_c E_c^{LYP}. \tag{2.53}$$

  E_x^{LSDA} is an LSDA non-gradient-corrected exchange functional, E_x^{HF} is the KS orbital based HF exchange energy functional, E_x^{B88} is the Becke88 exchange functional, E_c^{VWN} is the Vosko, Wilk, Nusair correlation function, which forms part of the accurate functional for the homogeneous electron gas of the LDA and the LSDA, and E_c^{LYP} is the LYP correlation functional. The parameters α_0, α_x and α_c are those that give the best fit of the calculated energy to molecular energies.

2.1.8 Basis sets

To obtain a numerical solution to the KS equations (and all other QM calculations), the eigenfunctions must be expanded in a set of known functions, a basis set. This is not an approximation if the basis is complete (an infinite number of functions). However, in practice the computational effort of conventional DFT techniques scales with the third power of the number of basis functions, so a compromise is needed between computational time and accuracy. It is desirable to use as few basis functions as possible to describe an unknown function. The more accurately a single basis function is able to reproduce the unknown function, the less basis functions are needed to achieve a given level of accuracy.

Slater-type orbitals

In early calculations Slater-type orbitals [38] (STOs) were often used as they have a form similar to that of the atomic orbital of a hydrogen atom. At the nucleus a cusp forms due to the singularity of the potential on the nucleus with charge +Z, whilst far away from the atom an electron would "see" only a positive charge. STOs display this exponential asymptotic behaviour and have the form,

$$\chi^{STO}(\mathbf{r}) = P(r)e^{-\zeta r}Y_{lm}(\theta, \phi). \tag{2.54}$$

The long range behaviour is only correct for STOs if the smallest component is less than $\sqrt{2I_{min}}$, where I_{min} is the lowest ionisation potential. However, very often smaller values than this are required for accurate results. The major draw back of STOs is that the two electron multi-centred integrals have to be computed numerically. This makes them computationally inefficient, limiting their use to small systems only.

Gaussian-type orbitals

Gaussian-type orbitals [39] (GTOs) were introduced to simplify the multi-centre integrals and allow an analytical solution. This is possible since the product of two Gaussians centred on A and B, is a Gaussian centred at the mid point of the two. GTOs have a the form,

$$\chi^{GTO}(\mathbf{r}) = P(r)e^{-\alpha r^2}Y_{lm}(\theta, \phi). \tag{2.55}$$

GTOs do not show the correct behaviour (cusp) at the nucleus or at long distances that STOs have, decaying much more rapidly as they move away from the nucleus than STOs. This qualitatively wrong behaviour was at first quite disappointing and it was thought that STOs would be the ideal choice if the multi-centred integral problem could be solved. This can be compensated for by the use of contracted Gaussian functions, where a linear combination of GTOs are used to approximate the correct form of the STOs [40]. Recent experience suggests that the Gaussian shape is actually more realistic for a nucleus of finite extension.

Basis Set Superposition Error

When calculating binding energies ($\Delta E_{bind} = E_{A+B} - E_A - E_B$) using atom-centred basis functions (eg. STO or GTO basis sets), there is an increase in the quality of the basis set describing the complex ($A + B$) due to the overlapping of the basis sets from A and B. The result of a larger basis set in the complex compared to the monomers, is to artificially increase the binding energy, since the additional variational freedom provided by the larger basis set reduces the energy of the complex relative to the individual monomers. This error is referred to as basis set superposition error (BSSE) [41]. BSSE is attributed to the use of an incomplete basis set and increasing the number of atom-centred functions used reduces this error. One method of accounting and correcting for BSSE is called the counterpoise correction [41, 42]. In this approach, the energies of both A and B are calculated in the presence of the basis set used in the complex ($A+B$).

Plane waves

Another basis set that is often used, which is very different to STOs and GTOs, are plane waves [3] (PWs). PWs are solutions to the Schrödinger equation for a particle in a periodic box. They have the form (for a box with length l),

$$\Psi_{\mathbf{k}}(\mathbf{r}) = \frac{1}{l^{\frac{2}{3}}} e^{i(k_x x + k_y y + k_z z)} = \frac{1}{V^{\frac{1}{2}}} e^{i\mathbf{k}\cdot\mathbf{r}}. \tag{2.56}$$

Very large numbers of plane wave basis functions are used (tens of thousands or millions, depending on the simulation cell size) when performing DFT calculations, in contrast to the much smaller number of GTOs used in a typical calculation (hundreds or thousands). Plane waves are used more often for studying solids due to their periodic nature. Plane waves exhibit a uniform coverage of the simulation cell, having an advantage over atom-centred basis functions in that they do not suffer from BSSE. However, when using plane waves to study single molecules a supercell approach must be used. In this approach the simulation cell must be large enough to isolate the molecule from its periodic image. The additional "empty" space surrounding the molecule is very computationally expensive. Additionally, using plane waves also loses the connection with atomic orbitals resulting in a final set of molecular orbitals that can be difficult to chemically interpret.

It is common pratice to use another approximation when using plane waves. This approximation is often referred to as the pseudopotential approximation and will be discussed in more detail below.

2.1.9 Pseudopotentials

When heavy atoms are contained within the system, the number of basis functions required can quickly become a very large number. These extra electrons however, are mainly core electrons. They have no participation in chemical processes and so can be represented by a minimal number of basis functions.

A proposed approximation replaced the core electrons with analytical functions to represent the combined nuclear-electronic core to the remaining valence electrons. These are called pseudopotentials [43]. A major advantage of these is that they can be made to include the relativistic effects of the core electrons (relativistic effects become important in very heavy elements where core electrons move at speeds close to the speed of light).

When constructing pseudopotentials it is important to consider how many electrons to include in the “core”. A large-core pseudopotential includes all but the valence electrons, while a small-core pseudopotential only uses up to a shell below. In heavier metals, polarisation of sub-valence shells can be chemically important and it is often worth the extra computational effort to explicitly include these shells. Another important consideration is that the pseudopotential should match the true potential outside of the “core” radius, this is also the case for the pseudowavefunction and true wavefunction (fig 2.3). To do this we impose that the square amplitudes of the wavefunction and pseudowavefunction are identical over the core region, satisfying the condition of "norm-conservation".

The wavefunctions of the pseudopotential are slowly varying and so a much lower kinetic cut-off energy is required for a plane wave basis set. Since only the wavefunctions for the valence electrons (the electrons that are chemically active) need to be considered the efficiency of the calculation is increased even further. When utilising pseudopotentials fewer electrons are considered, so the total energies calculated are much reduced, however, calculated energy differences (binding energies) remain unchanged.

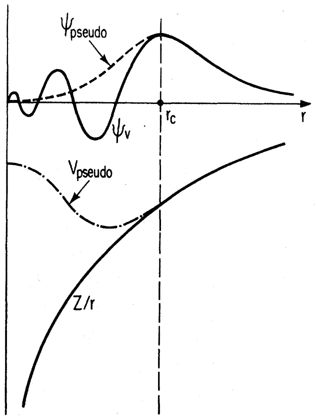


Figure 2.3: Schematic illustration of all-electron(solid lines) and pseudoelectron (dashed lines) potentials and their corresponding wavefunctions [3].

2.1.10 Linear-Scaling Density Functional Theory

The computational cost of QM calculations often limits the feasibility of their application to systems of biological interest, which often contain many hundreds or thousands of atoms. Using conventional wavefunction approaches, the best scaling that can be obtained is to the fourth power, for the HF approach, due to the calculation of the two electron integrals [44]. Many post-HF approaches have much higher scaling, for example CCSD(T) scales to the seventh power.

The computational cost of conventional DFT approaches, detailed earlier, scales with the third power of the number of atoms (number of basis functions). This cubic scaling is due the the orthogonality requirement of the wavefunctions. It is this computational cost that limits the size of a system that can be simulated to a few hundred atoms.

Linear-scaling approaches with respect to system size, also referred to as O(N) methods [45] are essential for the development of large scale *ab initio* calculations on systems containing thousands of atoms. Within DFT this is achieved by reformulating the expression for the energy, the expectation value for the Hamiltonian, of the system in terms of the one particle density matrix. The one particle density matrix is defined as,

$$\rho(\mathbf{r},\mathbf{r}') = 2\sum_{i}^{N_{functions}} f_i\psi_i(\mathbf{r})\psi_i^*(\mathbf{r}'), \qquad (2.57)$$

where the density is the diagonal of the density matrix $n(\mathbf{r}) = n(\mathbf{r},\mathbf{r}')$. This allows for a linear-scaling approach to be developed since the density matrix decays exponentially [46] with $|\mathbf{r}-\mathbf{r}'|$, so $n(\mathbf{r},\mathbf{r}')$ can be truncated for $|\mathbf{r}-\mathbf{r}'|$ being greater than a set threshold.

Using linear-scaling DFT approaches allows simulations of entire biomolecules and large nanostructures, consisting of thousands of atoms to be performed.

2.1.11 ONETEP

The ONETEP [47] program is a linear-scaling DFT code that has been developed for use on parallel computers [48]. ONETEP combines linear-scaling with accuracy comparable to conventional cubic-scaling plane-wave methods, which provide an unbiased and systematically improvable approach to DFT calculations. Its novel and highly efficient algorithms allow calculations on systems containing tens of thousands of atoms [49], as the example shown in Figure 2.4.

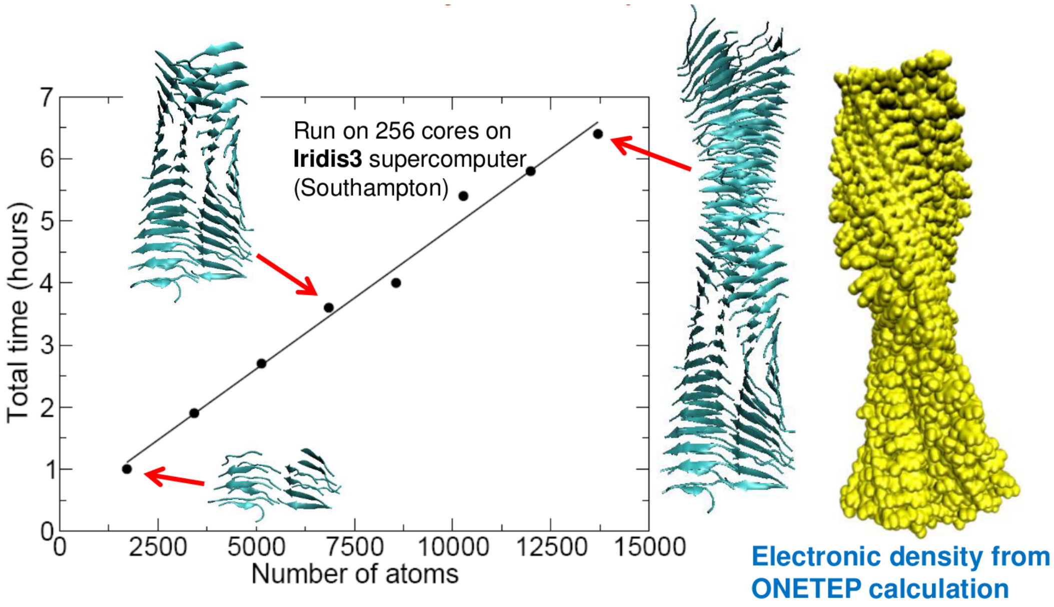


Figure 2.4: ONETEP calculation on an increasing size of an amyloid fibril displaying the linear scaling ability of the code [4].

ONETEP is based on a reformulation of DFT in terms of the one-particle density matrix. The density matrix in terms of Kohn-Sham orbitals is,

$$\rho(\mathbf{r}, \mathbf{r}') = \sum_{n=0}^{\infty} f_n \psi_n(\mathbf{r}) \psi_n^*(\mathbf{r}'), \tag{2.58}$$

where f_n is the occupancy and $\psi_n(\mathbf{r})$ are the Kohn-Sham orbitals. In ONETEP the density matrix is represented as,

$$\rho(\mathbf{r},\mathbf{r}') = \sum_\alpha \sum_\beta \phi_\alpha(\mathbf{r}) K^{\alpha\beta} \phi^*_\beta(\mathbf{r}'), \tag{2.59}$$

where $\phi_\alpha(\mathbf{r})$ are localised non-orthogonal generalised Wannier functions [50] (NGWFs) and $K^{\alpha\beta}$, which is called the density kernel, is the representation of f_n in the duals of these functions. Linear-scaling is achieved by truncation of the density kernel according to,

$$K^{\alpha\beta} = 0, \text{when } r_{\text{cut}} < |R_\alpha - R_\beta|, \tag{2.60}$$

since it decays exponentially for materials with a band gap, and by enforcing strict localisation of the NGWFs onto atomic regions as shown in Figure 2.5. In ONETEP, as well as optimising the density kernel the NGWFs are also optimised, subject to a localisation constraint. Optimising the NGWFs *in situ* allows for a minimum number of NGWFs to be used whilst still achieving plane wave accuracy. The NGWFs (Figures 2.5 and 2.6) are expanded in a basis set of periodic sinc (psinc) functions [51] (Figure 2.7), which are equivalent to a plane-wave basis as they are related by a unitary transformation. Using a plane wave basis set allows the accuracy to be improved by changing a single parameter, equivalent to the kinetic energy cut-off in conventional plane-wave DFT codes. The psinc basis set provides a uniform description of space, meaning that ONETEP does not suffer from basis set superposition error [52]. Unlike planewave basis sets, when using a psinc basis set, the additional vacuum necessary in the super cell approach has no additional computational cost.

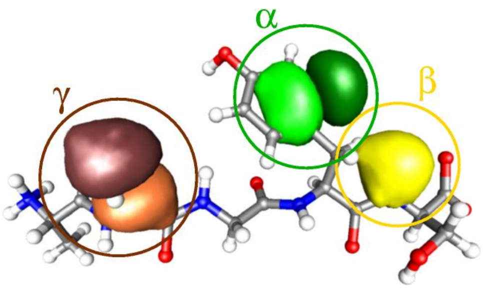


Figure 2.5: NGWFs centred on atoms within localisation spheres.

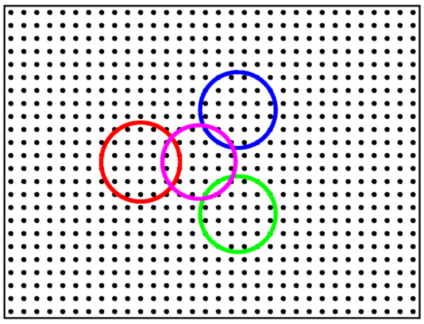


Figure 2.6: NGWF localisation spheres on a regular grid of points.

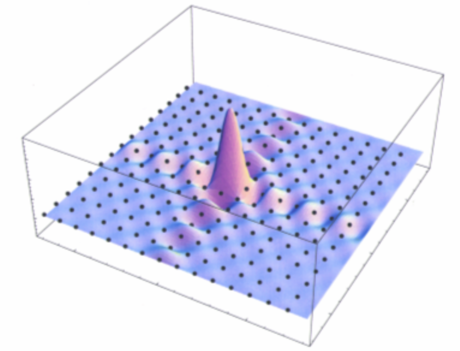


Figure 2.7: A psinc function.

Dispersion in ONETEP

Common DFT functionals do not usually account for dispersion forces (London forces). Dispersion forces are the attractive parts of van der Waals interactions. If an exact exchange correlation functional was used these forces would be described correctly, however, as the form of this is unknown, approximations to this are made. In ONETEP an empirical correction is used to model the dispersion forces [5] following the DFT+D approach of Grimmer *et al* [53]. This correction is in the form of a damped London dispersion term, of the form,

$$E_{disp} = -\sum_{ij,i\neq j} f_{\mathrm{damp}}(r_{ij})\frac{C_{6,ij}}{r_{ij}^{6}}, \tag{2.61}$$

and is added to the total energy equation,

$$E[n] = T_s[n] + E_{ne}[n] + J[n] + E_{xc}[n] + E_{disp}. \tag{2.62}$$

Figure 2.8 shows the importance of accounting for dispersion for the example of a $\pi - \pi$ interaction between two benzene molecules.

This empirical correction was specifically parametrised for the PBE functional for Carbon, Nitrogen, Oxygen, Sulphur and Hydrogen, against CCSD(T) and MP2 binding energies for 60 complexes. Although it was only optimised for the above atoms it is implemented for all atoms used in biological systems.

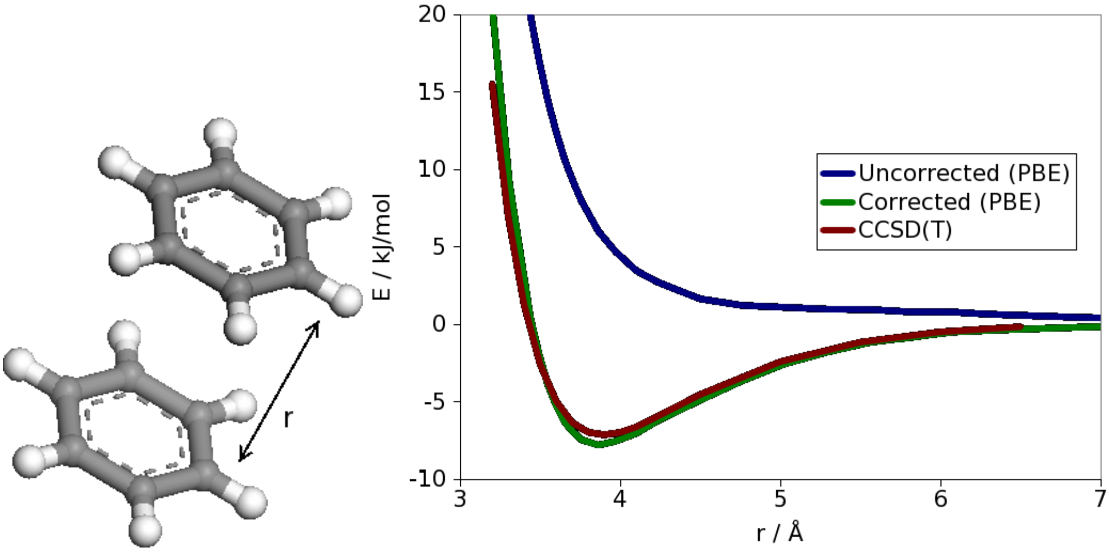


Figure 2.8: Inclusion of dispersion in ONETEP. On the left: the test system of two benzene rings interacting through their π system. On the right: the energy of the uncorrected and corrected DFT energy against a more accurate CCSD(T) calculation. [5]

2.2 Classical Molecular Mechanics

Even with linear-scaling QM approaches, QM remains too computationally demanding for many of the problems we wish to study. A much less computationally demanding approach is to use classical (Newtonian) physics (MM), using empirical force fields to study large systems.

Force field methods do not explicitly include electronic degrees of freedom and calculate the energy as a function of nuclear coordinates only. This is possible due to the Born-Oppenheimer approximation that allows the energy to be written as a function of nuclear positions. These methods are used to perform calculations on systems containing many tens of thousands of atoms and can give reasonably accurate results. However they can not calculate properties that depend explicitly on the electronic distribution in a system. Within this method molecules are modelled as atoms of varying size held together by bonds of varying stiffness, the molecule is thus described as a 'ball and spring' model as shown in figure 2.9.

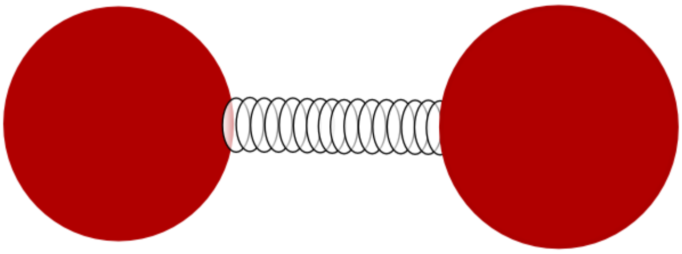


Figure 2.9: Ball and Spring model of a diatomic molecule depicting two atoms of mass M_1 and M_2 connected by a "spring".

Many different force fields exist, they can be tailored to a specific problem (protein residues, DNA) or general (ligands, non-standard residues). This is allowed due to a property intrinsic to chemistry called transferability. Transferability was found to be evident in the 20th century from analysis of spectroscopic data of hundreds of molecules. It was observed that molecules with similar bonds had force constants and equilibrium bond lengths with similar values. For example, all C-H bond lengths are around 1.09Å with similar vibrations (2900-3300 cm^{-1}). This holds true for other 'groups', such as the C=O bond which is approximately 1.21Å with vibrational frequencies around 1700 cm^{-1}. This gives a picture of molecules being made up of structural units, “functional groups”, which behave similarly in different molecules (this forms the basis of organic chemistry). This allows parameters generated from a relatively small number of small molecules to be used to study much larger molecules. Force fields consist of terms that describe intra-molecular forces, as well as interactions between non-bonded parts of the system. Force fields are parametrised from experimental data and high level QM calculations to obtain equilibrium bond lengths, angles, and dihedrals that are found inside molecules. Energy penalties are associated with movement of these structural parameters away from their equilibrium values.

One such functional form (commonly used for biological molecules) for the potential energy is,

$$\begin{aligned}\nu(\mathbf{r^N}) = \sum_{bonds}\frac{k_i}{2}(l_i - l_{i,0})^2 + \sum_{angles}\frac{k_i}{2}(\theta_i - \theta_{i,0})^2 + \sum_{torsion}\frac{V_n}{2}(1+\cos(n\omega-\gamma)) \\ + \sum_{i=1}^{N}\sum_{j=i+1}^{N}\left(4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}}\right). \qquad (2.63)\end{aligned}$$

Each term in Equation 2.63 will be described in detail later in this section, but briefly the terms are defined as: The first term on the right hand side is the interactions between bonded atoms, modelled by a harmonic oscillator. The second

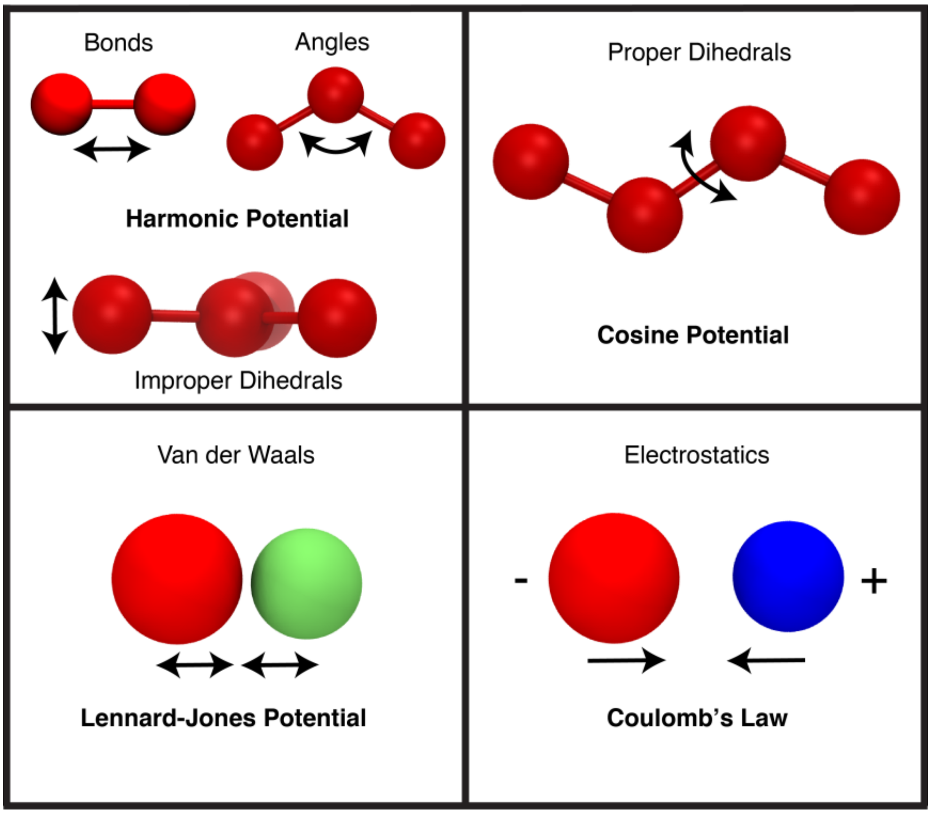


Figure 2.10: Bonded and non-bonded interaction in a force field. Figure provided by Dr Thomas Piggot.

term is over all angles A-B-C, where both A and C are bonded to B, and is also modelled by a harmonic oscillator. The third term is for torsion angles, and the last is the non-bonded terms. This is between atoms that are separated by 4 or more bonds, or in a different molecule. The first part of the non-bonded term is modelled in simple forcefields with a Lennard-Jones potential for describing van de Waals interactions and the second term is a Coulomb potential for describing the electrostatic interactions. A diagram of these bonded and non-bonded terms can be seen in Figure 2.10.

Force fields are comprised of a function (such as Equation 2.63) and a set of parameters for each atom type. There is no 'correct' form for a force field, as two force fields may have the same functional form but different parameters, or may have different functional forms and still give results of comparable accuracy. However, it would not be correct to swap parameters or mix parameters of one model with another.

Force fields, such as Equation 2.63 , can not describe the dissociation of bonds. The energy of a bond is well described by a Morse potential which can describe a wide range of behaviours, including dissociation, and has the form,

$$\nu(l) = D_e \left\{1 - \exp\left[-a\left(l - l_0\right)\right]\right\}^2. \tag{2.64}$$

Where D_e is the depth of the potential energy minimum, $a = \omega\sqrt{\mu/2D_e}$, where μ is the reduced mass and ω is the frequency of the bond vibration, and l_0 is the equilibrium bond length. This potential is not usually used in force fields however. Since there are 3 parameters for each bond it is not very computationally efficient. Also, in biomolecules it is quite rare for bonds to deviate far from the equilibrium bond length. For this reason a much simpler expression is usually used, utilising a Hook's law formula (harmonic potential) were the energy varies with the square of the displacement from l_0.

$$\nu(l) = \frac{k}{2}(l - l_0)^2. \tag{2.65}$$

Force constants are often very large. For example, for a C$_{sp^3}$ - C$_{sp^3}$ bond, the force constant will be around 300 kcal mol$^-1$ Å$^-2$. If the bond were to deviate from l_0 by just 0.2Å the energy would change by 12 kcal mol$^-1$.

A harmonic potential is also often used to describe deviation for angles. Since the distortion of angles away from their equilibrium angle (θ_0) requires considerably less energy than bond stretching, the force constants are considerably smaller, usually around 0.01 kcal mol$^-1$ deg$^-1$.

A torsion angle, or dihedral angle, is associated with the ABCD linkage. It is defined as the angle between the bonds AB and CD when they are projected into the plane bisecting the BC bond. Most force fields express the torsion potential as a cosine series expansion.

$$v(\omega) = \sum_{n=0}^{N} \frac{V_n}{2} \left[1 + \cos\left(n\omega - \gamma\right)\right]. \tag{2.66}$$

Where n is the number of minimum points in the function as the bond is rotated through 360° (2π radians), the multiplicity γ is the phase factor and determines where the angle passes through its minimum value. V_n is the barrier height and is given in kcal mol$^-1$, although this is really only a qualitative indication of the relative barriers of rotation. For example V_n for an amide bond would be larger than for a bond between two sp^3 carbon atoms.

Non-bonded forces play a major role in molecular interactions, whether between independent molecules or in determining the structure of a molecule. They do not have a specific bonding relationship but are interactions through space. As mentioned above, the 1-4 interactions are split into two terms, the van der Waals interactions and the electrostatic interactions.

Electrostatic interactions between two molecules (A and B) are calculated as a sum of interactions between pairs of point charges, using Coulomb's law,

$$V_{\text{elec}} = \sum_{i=1}^{N_A} \sum_{j=1}^{N_B} \frac{q_i q_j}{4\pi\epsilon_o r_{ij}}. \tag{2.67}$$

The point charges (q_i and q_j) are fractional charges, arranged onto atoms and designed in such a ways as to reproduce the electrostatic properties of the molecules. They are often referred to as partial atomic charges as they are restricted to nuclear centres. There have been many suggested methods on ways to calculate the partial charges of a molecule. Since partial charges are a consequence of the electronic distribution in a molecule, it would be reasonable to use quantum mechanics, such as DFT which explicitly calculates the electronic density, to obtain the partial charges. However, the partial atomic charge is not an experimental observable and hence can not be easily and uniquely computed from the wavefunctions. In force fields, the partial charges given to a molecule need to represent how two molecules interact with each other. This has lead to schemes that reproduce charges that are consistent with the electrostatic potential of a molecule.

The electrostatic potential at a point is the force acting on a positive unit of charge placed at that point. It is an observable quantity that can be determined from the wavefunction (density) using,

$$\begin{aligned}\phi(\mathbf{r}) &= \phi_{nuc}(\mathbf{r}) + \phi_{elec}(\mathbf{r}) && (2.68)\\ &= \sum_{A=1}^{N} \frac{Z_A}{|\mathbf{r}-\mathbf{R_A}} - \int \frac{d\mathbf{r}'n(\mathbf{r})}{|\mathbf{r}'-\mathbf{r}}. && (2.69)\end{aligned}$$

To generate charge models for large systems, molecules are broken into fragments. The atomic partial charges are obtained from quantum calculations on fragments that will recreate the immediate local environment that the fragment would “see” in the larger molecule. For the case of proteins, these fragments are chemically well defined units - amino acids. More commonly, the fragments will be ’dipeptides’, which more accurately describes the environment that a single amino acid would be found in.

Van der Waals interactions are a sum of the attractive and repulsive forces that are not due to electrostatic interactions. That attractive force is long range, whereas the repulsive force is short range. The attractive force is due to dispersive forces, as talked about in the ONETEP section. The repulsive forces are due to the Pauli principle, that any two electrons are prohibited from having the same quantum numbers. Because this repulsive force is between electrons with the same spin it is also referred to as exchange force. The effect this has is to reduce electron density between the nuclei as the nuclei approach each other. This reduced shielding leads to repulsion of the two nuclei. In force fields a simple empirical function is used that can quickly evaluate the large number of van der Waals interactions that must be determined in most systems. The most commonly used function is the Lennard-Jones 12-6 function,

$$v(\mathbf{r}) = 4\epsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right]. \tag{2.70}$$

This potential contains only two parameters, the collision diameter σ and the well depth ϵ. It has an attractive part which varies with r^{-6} and a repulsive part that varies with r^{-12}. The repulsive r^{-12} term has little theoretical justification. Quantum mechanically this term is suggested to have an exponential form. The $\mathbf{r}^{-12}$ term best represents noble gas interaction but is too steep for hydrocarbons. Nevertheless it is widely used since it can be rapidly calculated by squaring the r^{-6} term making its calculation in large systems much simpler.

Bond-stretching and angle-bending have much higher energy penalties associated with deformations away form their equilibrium values, they are considered as "hard" degrees of freedom, and these movements tend to be reasonably small. Rotation of torsional angles on the other hand, are known to be fundamental in molecular structural properties. Most of the variation in structures and relative energies are due to torsional and non-bonded interactions.

There are are a few software packages available that can be used for calculation using forcefields. Such as Gromacs, CHARMM and AMBER. During this work only the AMBER [54] package was used, and is briefly detailed below.

2.2.1 AMBER

AMBER10 is a collection of programs designed for carrying out MM and molecular dynamics (MD) calculations. The name "amber" also refers to a number of empirical force fields, however these are not specific to AMBER which can also use a selection of other force fields. The AMBER10 package is capable of running many different types of calculations utilising molecular mechanics, such as dynamics simulation, QM/MM hybrid simulations and various free energy approximation approaches, which will be discussed later in this chapter.

As stated earlier, force field parameters can be highly optimised for specific systems. For instance the ff99 force field [55] (and variants of it) has many different atom types. For example, a carbon in a five membered ring will have a different atom type to that in a 6 membered ring, and a carbon in a protonated histidine would be different to that of an unprotonated histidine. More generic forcefields would just term this an sp^2 carbon, as in the amber-gaff force field. This tailoring of the amber-ff99 forcefield to proteins obtains reasonably realistic trajectories and interaction energies. Parameterising in this way however, limits the transferability of the force field. For ligands and non-standard residues, force fields must be much more general. The general amber force field was designed for this reason and give comparable results for proteins as ff99. A draw back of this generalised parameterisation, is that these force fields can suffer from inaccurate results for more complicated systems, i.e. when they contain halides or transition metals.

Both force fields mentioned above have the functional form,

$$E = \sum_{bonds} k_r(r - r_{eq})^2 + \sum_{angles} k_\theta(\theta - \theta_{eq})^2 +$$

$$\sum_{dihedrals} \frac{\nu_n}{2}[1 + cos(n\theta - \gamma)] + \sum_{i<j} \left\{ \epsilon_{i,j} \left[\frac{r_{0ij}}{R_{ij}^{12}} - \frac{r_{0ij}}{R_{ij}^{6}} \right] + \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} \right\}. \quad (2.71)$$

Here, r_{eq} and θ_{eq} are equilibration structural parameters, k_r , k_θ , v_n are force constants, n is multiplicity and γ is the phase angle for the torsional angle parameters. The r_{0ij} is the equilibrium distance and ϵ is the well depth, q is the charge on atom i or j. These parameters characterise the nonbonded potentials. The differences between these two force fields is the number of specifically parametrised atom types (as mentioned above).

The partial charges in the amber-ff99 force field were calculated using Hatree-Fock and a 6-31G* basis set, which is generally believed to give reasonable results. It is often possible to scale the results obtained using a smaller basis set or lower level of theory, such as semi-empirical calculations, to obtain comparable results. This approach is often used in combination with the gaff force field.

2.2.2 Some available force fields in AMBER

Force fields

- ff99 [55]: Force field developed for proteins.
- ff99SB [12]: "Stoney Brook" modification to ff99. Backbone torsions fitted to *ab initio* calculations to improve performance for protein.
- ff99bsc0 [56]: "Barcelona" modification to ff99SB. Further changes for improved results with nucleic acids.
- ff03 [57]: A variant of ff99. Charges and main-chain torsion potentials have been re-derived based on QM+continuum solvent calculations (for proteins only).
- GLYCAM [58]: Force field for carbohydrates and lipids. Parametrised by fitting to QM data for small molecules.

General force fields

- GAFF [59]: Generalised Amber Force Field: parametrised for use with non-standard residues and small molecules.
- MMFF94 [60]: Merck Molecular Force Field: Created by Merck and parametrised for use with small molecules.

Polarisable force fields

- ff02 [61]: A polarisable variant of ff99. The charges were determined at the B3LYP/cc-pVTZ//HF/6-31G* level, and are more like "gas-phase" charges.
- ff02EP: Modification to ff02. Adds in off-centre charges to mimic electron lone pairs in order to better describe the angular dependence of hydrogen bonds.

2.3 Molecular dynamics

When studying molecular properties it is desirable to sample the configurational space available to the molecule. A phase point (point in phase space) can be defined by the momentum and position of all the particles in the molecule. This means that any phase point can be used to determine the location of the “next” phase point in the trajectory. Since the trajectory is a continuous curve of phase points, the starting geometry can completely determine the forward direction, and since time-dependent Hamiltonians are invariant to time reversal, the entire trajectory.

In molecular dynamics the “next” phase point in the trajectory is generated by integrating Newtons’s laws of motion. This results in a trajectory that will give variations in the positions and momenta of the atoms through time.

The relationship between two positions ($\mathbf{r}$) is,

$$\mathbf{r}(t_2) = \mathbf{r}(t_1) + \int_{t_1}^{t_2} \frac{\mathbf{p}(t)}{m} dt, \tag{2.72}$$

where $\mathbf{p}$ is the momentum (velocity multiplied by mass, $\mathbf{p} = \mathbf{v}m$), and m is the mass. The relationship between two momentum vectors is similarly given as,

$$\mathbf{p}(t_2) = \mathbf{p}(t_1) + m \int_{t_1}^{t_2} \mathbf{a}(t) dt, \tag{2.73}$$

where $\mathbf{a}$ is the acceleration. By solving the differential equations embodied by Newton’s second law ($\mathbf{F} = m\mathbf{a}$) we have a relationship between force and the position and can obtain the trajectory.

$$\frac{d^2x_i}{dt^2} = \frac{F_{x_i}}{m_i}. \tag{2.74}$$

Equation 2.74 describes the motion of a particle with mass m_i along the coordinate x_i, with the force in that direction being given by F_{x_i}.

In real systems, the force felt by any particle will change whenever it moves, or when any particle it interacts with moves. Using a continuous potential, the motions of all the particles are coupled together, however, this leads to a many-body problem that can not be solved analytically. To solve this problem the finite difference method must be used to integrate the equations of motion.

The basic concept is to split the integration into many small intervals, each separated by a constant time δt. The differential form of Equation 2.72 is,

$$\mathbf{r}(t + \delta t) = \mathbf{r}(t) + \frac{\mathbf{p}(t)}{m}\delta t, \tag{2.75}$$

and of Equation 2.73 is,

$$\mathbf{p}(t + \delta t) = \mathbf{p}(t) + m\mathbf{a}(t)\delta t. \tag{2.76}$$

For finite δt, this is called Euler's approximation, and is exact in the limit of $\delta t \to 0$. Thus, using this approach, given the initial positions and momenta, and a formula for calculating the force on each particle at any moment, we have the ability to "simulate" a phase space trajectory.

This approach is often too simple and leads to unstable trajectories. More sophisticated algorithms have been developed based on the assumption that the positions, velocities, and accelerations can be approximated as a Taylor series expansion.

$$\mathbf{r}(t+\delta t) = \mathbf{r}(t) + \delta t\mathbf{v}(t) + \frac{1}{2}\delta t^2\mathbf{a}(t) + \frac{1}{6}\delta t^3\mathbf{b}(t) + \frac{1}{24}\delta t^4\mathbf{c}(t) + \dots \quad (2.77)$$

$$\mathbf{v}(t+\delta t) = \mathbf{v}(t) + \delta t\mathbf{a}(t) + \frac{1}{2}\delta t^2\mathbf{b}(t) + \frac{1}{6}\delta t^3\mathbf{c}(t) + \dots \quad (2.78)$$

$$\mathbf{a}(t+\delta t) = \mathbf{a}(t) + \delta t\mathbf{b}(t) + \frac{1}{2}\delta t^2\mathbf{c}(t) + \dots \quad (2.79)$$

where $\mathbf{r}(t)$ is the position, $\mathbf{v}(t)$ is the velocity (first derivative), $\mathbf{a}(t)$ is the acceleration (second derivative), $\mathbf{b}(t)$ is the third derivative, and so on. One method, first used by Verlet [62], considers the sum of the forward and reverse Δt steps expanded in this way. In the sum, all odd terms cancel since they will have opposite signs, and truncating at the second derivative gives,

$$\mathbf{r}(t+\delta t) = 2\mathbf{r}(t) - \mathbf{r}(t-\delta t) + \delta t^2\mathbf{a}(t). \quad (2.80)$$

Using this equation, the position of each subsequent time step is determined from the current position, the previous position, and the acceleration (calculated from the forces, $\mathbf{a} = \mathbf{F}/m$). For the initial step, where no previous step is available, Equations 2.75 and 2.76 can be used.

An initial structure is often taken from what a chemist thinks to be “reasonable”. When working with a protein system, this can often come from an experimentally resolved structure (from x-ray crystallography or NMR).

Initial momenta are assigned randomly subject to a temperature restraint. Temperature is related to momentum via,

$$T(t) = \frac{1}{(3N-n)k_B} \sum_{i=1}^{N} \frac{|\mathbf{p}_i(t)|^2}{m_i}, \tag{2.81}$$

where N is the total number of atoms and n is the number of constrained degrees of freedom.

It is often desirable to control the simulation temperature. This is accomplished by scaling the particle velocities so that the temperature (Equation 2.81) can be kept constant. However, this is not possible to implement in the Verlet scheme as it has no reference to velocity. To couple the position and velocity vectors a modification is made to Verlet's approach, called the leapfrog algorithm [63]. This approach though has a major disadvantage: the kinetic energy can not be calculated at the same time as the potential energy since the velocity and positions are not synchronised. A further improvement proposed to solve this issue is the velocity Verlet method [64], which calculates the positions, the velocities, and the accelerations at the same time.

$$\mathbf{r}(t+\delta t) = \mathbf{r}(t) + \delta t \mathbf{v}(t) + \frac{1}{2}\delta t^2 \mathbf{a}(t). \tag{2.82}$$

$$\mathbf{v}(t+\delta t) = \mathbf{v}(t) + \frac{1}{2}\delta t \left[\mathbf{a}(t) + \mathbf{a}(t+\delta t)\right]. \tag{2.83}$$

The use of a finite time step produces problems in the practicality of simulating the trajectory. If too large a time step is used then atoms may be rammed into each other, resulting in atomic distances far smaller than the van der Waals contact. This will result in large repulsive forces for the next step and the molecule will be blown apart. At infinitesimally small time steps Equations 2.72 and 2.73 will be recovered. However, due to the computational expense of having to calculate the forces on every particle at every time step, if too small a time step is taken then the simulation will not sample enough of the trajectory to observe anything of chemical interest. The generally accepted time step is one (or two) orders of magnitude smaller than the period of the fastest periodic motion in the system.

The fastest motion in classical (molecular mechanics) systems is the bond vibration of a heavy-atom – hydrogen bond, which is about 10^{-14} s. This limits the time step to a maximum of 1 fs (1x10^{-15} s). A method used for increasing the time step restrains the heavy-atom – hydrogen bonds to a constant length. This results in the fastest motion being heavy-atom – heavy-atom bonds, which have periods that are 2-3 times larger, allowing a time step of 2 fs. A commonly used approach for eliminating these degrees of freedom is the SHAKE algorithm [65].

2.3.1 Thermodynamic ensembles

The thermodynamic state of a molecular system in thermodynamic equilibration is achieved by a set of state variables, namely the temperature, the pressure, the volume, and the number of particles in the system. A thermodynamic ensemble is a surface in phase space that satisfies the conditions of a particular thermodynamic state. There are three such states commonly used during molecular dynamics simulations:

1. The microcanonical ensemble (NVE), where the number of particles (N), the volume (V), and energy (E) is kept constant.
2. The canonical ensemble (NVT), where the number of particles (N), the volume (V), and the temperature (T) are kept constant.
3. The isobaric-isothermal ensemble (NPT), where the number of particles

(N), the pressure (P), and the temperature (T) are kept constant.

To achieve a desired temperature in a system, temperature stabiliser algorithms called "thermostats" are used to imitate heat exchange between the system and its surroundings. Examples of these are the Berendsen thermostat [66], which corrects deviations of T from the set temperature, T_0, by multiplying the velocities by a factor, and the Langevin thermostat [67] which uses the Langevin equation of motion instead of Newton's law of motion.

$$m\mathbf{a} = \zeta\mathbf{v} + \mathbf{f}(\mathbf{r}) + \mathbf{f}', \tag{2.84}$$

where m is the mass, $\mathbf{a}$ is the acceleration, $\mathbf{v}$ is the velocity and ζ is a frictional constant. $\mathbf{f}'$ is a random force randomly determined from a Gaussian distribution to add kinetic energy to the particle, and its variance is the function of the set temperature and time step. This balances the random force with the frictional force.

2.4 Solvation models

When calculating the binding energy of a ligand bound to a protein it is important to take into account the energy of solvation, ΔG_{solv}. This is particularly important in biological molecules when the penalty of taking a ligand out of solution can negate the benefit that arises from binding to a protein. Many models have been proposed for the simulation of liquid water, from explicit solvation models where the molecules are present, to implicit models in which the solvent is represented as a continuum of dielectric permittivity surrounding the solute in a cavity.

2.4.1 Explicit solvation

Explicit water models can be categorised by the number of points used to define the model (additional lone pairs or dummy atoms, Figure 2.11), whether the model is rigid or flexible and whether the model is polarisable.

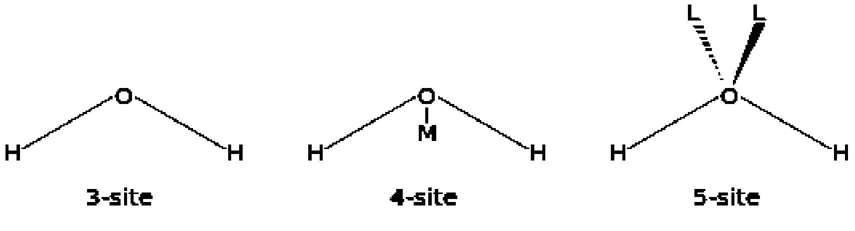


Figure 2.11: Explicit water binding models.

3-site models

The simplest model would be to treat the water molecule as rigid and use only the non-bonded interaction from the force field as given in Equation 2.85.

$$E_{nm} = \sum_{i}^{on\ n} \sum_{j}^{on\ m} \left[\left(\frac{A}{r_{OO}^{12}} - \frac{B}{r_{OO}^{6}} \right) + \frac{q_i q_j}{4\pi\varepsilon r_{ij}} \right], \tag{2.85}$$

where A and B are chosen to give reasonable structural and energetic results for liquid water, and gas phase complexes of water and alcohols. This was the original TIPS model [68]. It was further parametrised for liquid water by Berendsen [69] (SPC model) to produce better energies and a smaller second peak in the OO radial distribution function (shown in Figure 2.12), although the first peak is in worse agreement with experimental x-ray data than TIPS. This 3-point model was later parametrised by Jorgensen to further improve energies and densities of liquid water in the TIP3P [70] model. Berendsen added a term to the SPC potential to account for polarisation, referred to as SPC/E [66], and described by,

$$E_{pol} = \frac{1}{2} \sum_{i} \frac{(\mu - \mu_0)^2}{\alpha_i}, \tag{2.86}$$

where μ is the dipole of the effectively polarised water molecule, μ_0 is the dipole moment of an isolated water molecule, and α_i is an isotropic polarisability constant. Further additions to the SPC model aided to make it flexible. The flexible SPC model is one of the most accurate three-centre water models, and can produce the correct density and dielectric of bulk liquid water during MD simulations.

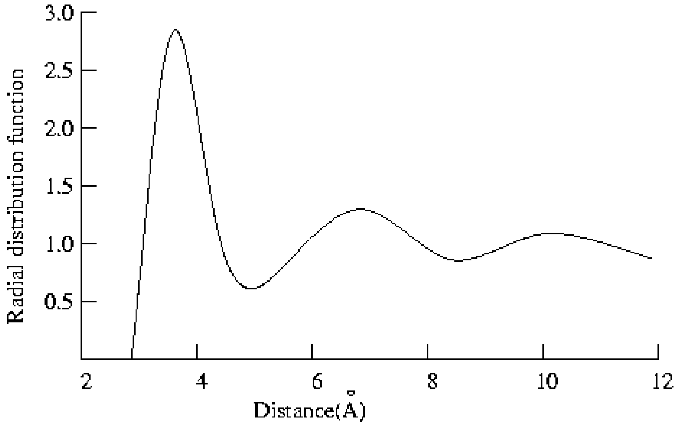


Figure 2.12: Radial distribution function determined from a 100 ps molecular dynamics simulation of liquid water at a temperature of 100 K and a density of 1.396 g/cm^3 [6].

4-site models

The 4-site model was first proposed by Bernal and Fowler [71] (BF model) for calculating the properties of a monomer, a dimer, and ice. The model places a negative charge at the M site in Figure 2.11. This improves the electrostatic distribution around the water molecule. Equation 2.85 still applies but with a little increase in complexity, requiring ten distances to evaluate the function instead of nine for the 3-site model. A different monomer geometry and set of parameters were proposed for the TIPS2 [72] model for liquid water, and another alternative parametrisation for the TIP4P [70] model. TIP4P has been re-parametrised several times to increase accuracy when used with Ewald summation methods (TIP4P-EW [72]), for use with ice (TIP4P/ice [73]), and a more general parametrisation (TIP4P/2005 [74]) .

5-site models

5-site models place negative charges on the L sites in Figure 2.11 to represent the lone pairs. The first model was proposed by Ben-Naim and Stillinger (BNS model) in 1971, and further improved by the ST2 [75] model of Stillinger and Rahman in 1974. The TIP5P [76] model by Mahoney and Jorgensen results in improvements in the geometry for the water dimer, a closer representation of the experimental radial distribution functions, and the temperature of maximum density of water.

2.4.2 Implicit Solvation

Using explicit waters greatly increases the number of molecules in the simulation making it much more complicated and time consuming, even when classical approaches are used. The major contribution to the solvation energy comes from long-range electrostatic interactions. These can be modeled in an implicit way by treating the solvent as a continuum with the same dielectric constant as the bulk solvent. The solvation energy is split into two terms, a polar term describing the electrostatics, and a non-polar term representing the energy of creating a solute-shaped cavity in the solvent.

Polar solvation energy

The theory for calculating the electrostatic component of the solvation free energy is the Poisson-Boltzmann equation,

$$\nabla \cdot (\epsilon(\mathbf{r})\nabla\nu(\mathbf{r})) = -4\pi\{\rho(\mathbf{r}) + \rho_m(\mathbf{r})\}$$

$$\nabla \cdot (\epsilon(\mathbf{r})\nabla\nu(\mathbf{r})) = -4\pi\{\rho(\mathbf{r}) + \sum_j q_j c_j e^{\frac{-E_j(\mathbf{r})}{k_B T}}\}$$

$$\nabla \cdot (\epsilon(\mathbf{r})\nabla\nu(\mathbf{r})) = -4\pi\{\rho(\mathbf{r}) + \sum_j q_j c_j e^{\frac{-z_j q_j \nu(\mathbf{r})}{k_B T}}\}, \qquad (2.87)$$

where $\rho_m(\mathbf{r})$ is the charge density of the mobile charges (ions), and $\rho(\mathbf{r})$ is the charge density (the distribution of charge throughout the system) of the solute. $\epsilon(\mathbf{r})$ is a dielectric constant for the solvent, q is the charge of the proton, c is the concentration of ions and z is the charge of the ion. T is the temperature and k_B is the Boltzmann constant. In this model only the charge of the ion is taken into account (+1, +2, -1, -2), it does not distinguish between atoms (such as K^+ and Na^+).

A cavity in the solvent is created in which the solute is placed. Within the cavity the solute is given a dielectric of 1, and outside the cavity the dielectric is that of the bulk solvent (for example, 80 for water). Different methods differ in their definition of the cavity. Examples are a very simple method using a spherical cavity within the solvent, and a cavity defined by the outside surface of interlocking spheres centred on the solute atoms.

Non-polar solvation energy

The non-polar contribution to the solvation energy (i.e. the energy required to create a solute shaped cavity in the solvent) is often modelled by a term that is dependent on the solvent-accessible surface area of a molecule. This can be obtained by “rolling” a sphere (of solvent) of a certain radius (usually 1.4 Å to represent a water molecule) across the molecule to probe the surface area. The non-polar energy is calculated as,

$$G_{np} = \gamma SA + b, \tag{2.88}$$

where SA is the surface area of the solute, γ is the surface tension and taken to be 0.00542 kcal/Å^2, and b is 0.92 kcal/mol [77].

2.4.3 Implicit solvation in ONETEP

A minimum parameter implicit solvent model has recently been developed within ONETEP [78]. In this model the total potential of the solute is obtained by solving the nonhomogeneous Poisson equation within the self-consistent calculation in ONETEP,

$$\nabla \cdot (\epsilon[\rho]\nabla\phi) = -4\pi\rho_{tot}(\mathbf{r}). \tag{2.89}$$

Where $\rho_{tot}(\mathbf{r})$ is the total charge density and is calculated as a sum of the electronic density $\rho(\mathbf{r})$ and the density of the atomic cores. The solute cavity is constructed directly from isosurfaces of the electronic density of the solute, which reduces the number of parameters required to only two. The model includes a smooth transition of the relative permittivity, shown by the graph in Figure 2.13, according to the following expression,

$$\epsilon(\mathbf{r}) = 1 + \frac{\epsilon_\infty - 1}{2}\left(1 + \frac{1 - (\rho(\mathbf{r})/\rho_o)^{2\beta}}{1 + (\rho(\mathbf{r})/\rho_o)^{2\beta}}\right), \tag{2.90}$$

where ϵ_∞ is the bulk permittivity, β controls the smoothness of the transition of $\epsilon(\mathbf{r})$ from 1 to ϵ_∞, and ρ_0 is the density value for which the permittivity drops to half that of the bulk. Figure 2.14 displays the approach that this method uses to calculate the solvation energy. The model in ONETEP defines the shape of the

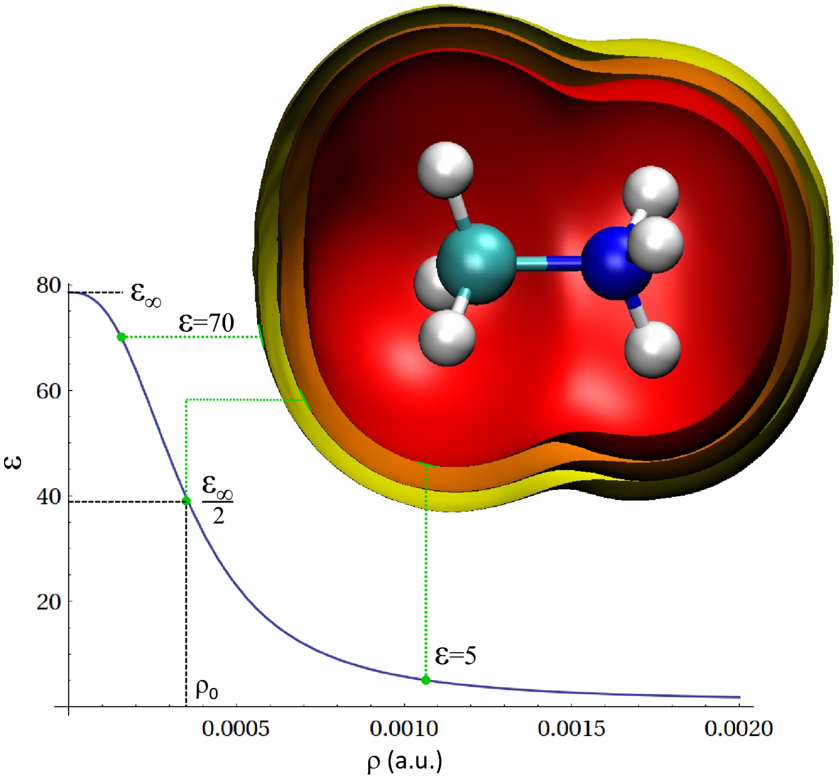
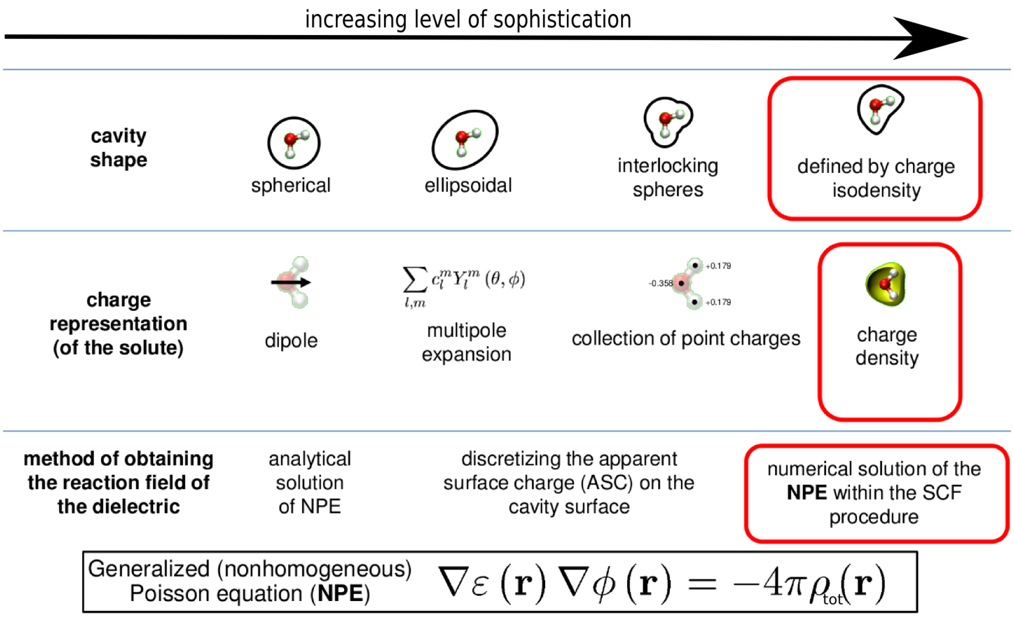


Figure 2.13: Depiction of density dependent dielectric.



cavity shape	spherical	ellipsoidal	interlocking spheres	defined by charge isodensity
charge representation (of the solute)	dipole	$\sum_{l,m} c_l^m Y_l^m(\theta,\phi)$ multipole expansion	collection of point charges	charge density
method of obtaining the reaction field of the dielectric	analytical solution of NPE	discretizing the apparent surface charge (ASC) on the the cavity surface		numerical solution of the **NPE** within the SCF procedure

Generalized (nonhomogeneous) Poisson equation (**NPE**) $\nabla\varepsilon(\mathbf{r})\nabla\phi(\mathbf{r}) = -4\pi\rho_{tot}(\mathbf{r})$

Figure 2.14: Implicit solvation computation methods in *ab initio* quantum chemistry approaches.

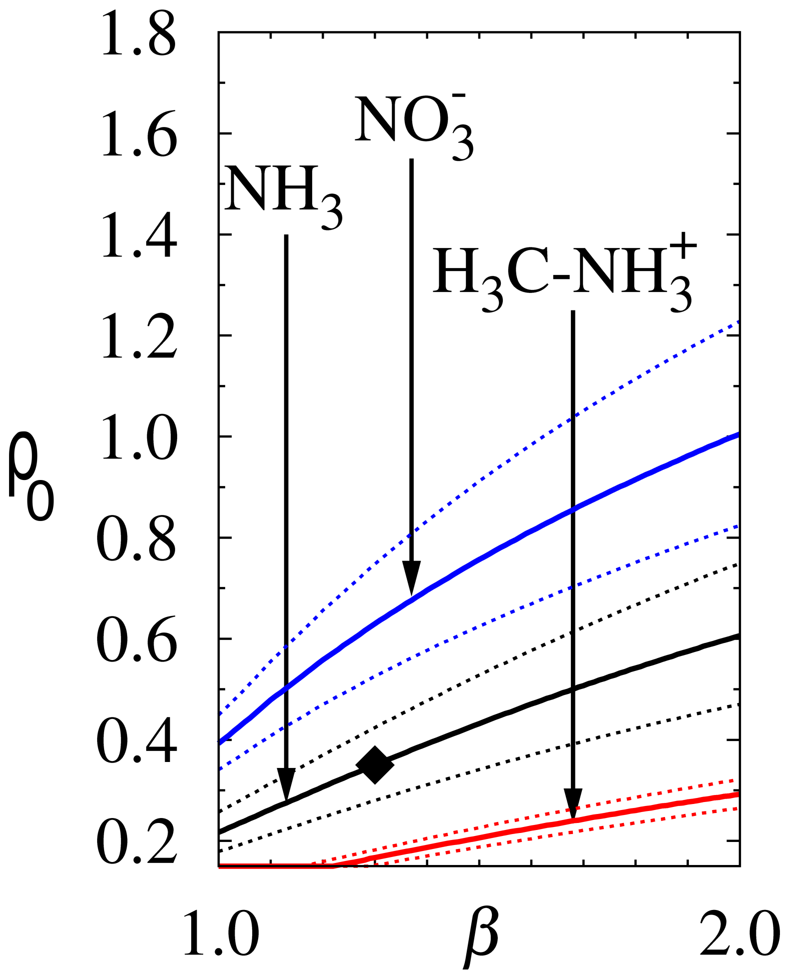


Figure 2.15: Graph of β plotted against ρ_0 for the three molecules. The diamond shows the choice for both parameters.

solute cavity from the charge isodensity. The charge representation of the solute is taken from the computed charge density of the molecule, and the reaction field of the dielectric is computed by the numerical solution of the nonhomogeneous Poisson equation (NPE) within the SCF procedure. This makes it the most *ab initio* implicit solvation model available.

The parameters for ρ_0 and β were taken from the best choices from three molecules, a neutral molecule (NH_3), a positively charged molecule ($H_3C\text{-}NH_3^+$), and a negatively charge molecule (NO_3^-). These are shown in Figure 2.15.

The non-polar term (cavitation energy) uses Equation 2.88. The surface area is calculated by the electronic density at ρ_0, γ is scaled by a factor of 0.281 to take account of dispersion, and b is set to zero.

This model was validated on two test sets, one of 60 small molecules (20 neutral, 20 cationic and 20 anionic) and the second of 71 larger molecules; On which the solvation energies obtained had a root mean square (rms) error with respect to experiment of 3.8 kcal/mol (when dispersion is included). While the polarisable continuum model (PCM) in Gaussian 03 [79] showed an rms error of 10.9 kcal/mol and the highly parametrised state-of-the-art SMD model [80] in Gaussian 09 [81] had a rms error of 3.4 kcal/mol.

2.5 Free Energy of Binding

A central problem in drug discovery is the prediction of ligand-receptor binding free energies. There are many approaches available, from purely empirical based methods such as QSAR, to much more theoretically rigorous approaches such as free energy perturbation [82]. An important consideration when choosing the binding free energy approach to use is the computational time required for the calculations. Amongst the many approaches available for free energy calculations, scoring [18] methods, commonly used in conjunction with docking, are amongst the least computationally expensive, and therefore the fastest, but also most approximate. In these methods ligand orientations (poses) are assigned scores, and the quality of the fit is expressed by an empirical function, the scoring function. These scores are used to rank each pose relative to other poses and other ligands. Methods with a higher level of statistical mechanics rigour include Molecular Mechanics Poisson-Boltzmann Surface Area (MM-PBSA) [83] and Molecular Mechanics Generalised Born Surface Area (MM-GBSA) [84]. These methods estimate absolute free energies of bound and unbound reference states using molecular dynamics (MD) simulations to sample phase space. Free energies of binding are obtained as averages of interaction energies over snapshots from the MD simulations with entropic contributions calculated from vibrational frequency calculations and solvation free energy contributions from an implicit solvent model. Although this approach has found extensive usage, especially for the calculation of relative free energies of binding, its accuracy is limited by the approximate nature of including entropy and solvation effects, as well as the force field which is required to reproduce structures and energies with high accuracy. At the most theoretically rigorous end of the spectrum, and most computationally expensive, we have methods such as alchemical free energy perturbation approaches [85]. Another example of an alchemical method is Thermodynamic Integration (TI). It follows an unphysical pathway, where one ligand is "mutated" to another. It estimates relative binding free energies and their gradual change during the mutation which happens in small steps, and fully includes the entropic and solvation contributions which are heavily approximated with the less rigorous approaches. In principle, alchemical free energy calculations allow the exact prediction of relative binding free energies, at very high computational cost. However, inadequacies in the force fields used and insufficient sampling introduce errors into the calculated free energies. These errors are exacerbated by ligands that cause changes which are difficult to capture by classical force fields such as charge transfer and polarisation, or cause conformational change on binding which may require extremely long simulations to describe sufficiently.

This section will give an overview of some of the more commonly used binding free energy approaches. The two main methods that will be discussed are TI and MM-PBSA. The differences between these two methods are shown in Figure 2.16. Details of the advancements made during this PhD towards more accurate QM binding free energies will be detailed in later chapters.

2.5.1 Free Energy Perturbation

The free energy difference between two states is formally obtained from the Zwanzig equation [86].

$$\Delta A = A_Y - A_X = -\beta^{-1} \ln \left\langle e^{-\beta \Delta V} \right\rangle_X , \tag{2.91}$$

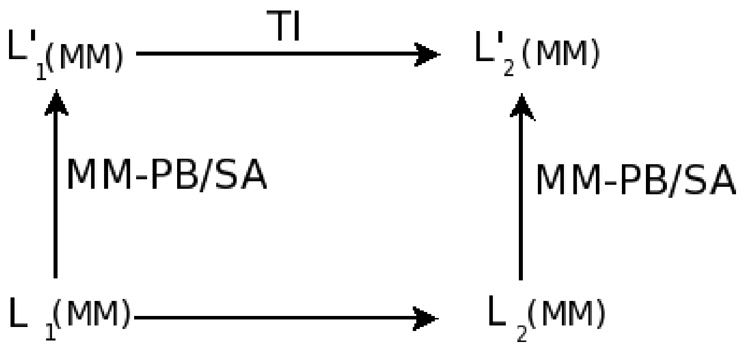


Figure 2.16: Free energy cycle. Moving from bottom to top follows the realistic "physical" route from the unbound to the bound states. Moving from left to right follows the "unphysical" mutation of L_1 to L_2 in the bound (top) and unbound (bottom) states.

where $\beta = 1/k_B T$, k_B is Boltzmann's constant and T is the temperature, and $\langle\rangle_X$ denotes an an ensemble average of $\Delta V = V_Y - V_X$ that is sampled using the potential of V_X. Equation 2.91 calculates the Helmholtz free energy,

$$\Delta A = \Delta U - T\Delta S, \tag{2.92}$$

where U is the internal energy and S is the vibrational entropy. Gibbs free energy is related to Helmholtz free energy by,

$$\Delta G = \Delta A + P\Delta V, \tag{2.93}$$

where P is the pressure and V is the volume. When doing this mutations, ΔV is often very small, so small is can often be neglected. Given this approximation, the Gibbs free energy is equivalent to the Helmholtz free energy, and hence the Zwanzig equation in 2.91 can be written as,

$$\Delta G = G_Y - G_X = -\beta^{-1} \ln \left\langle e^{-\beta\Delta V} \right\rangle_X . \tag{2.94}$$

For this equation to be practical, the structural configuration sampled on the potential V_X, should also have a high probability of occurrence on V_Y. This means that the potential energy surfaces of X and Y should overlap very well. If this is not the case then convergence will be very slow. In order to solve the case where potentials do not overlap well, a path between the states X and Y is adopted in a multistep approach, in which state X is slowly mutated into state Y.

2.5.2 Thermodynamic Integration

In Thermodynamic Integration (TI), a set of intermediate to X and Y potential energy functions are introduced. These are usually constructed as a linear combination of the initial (X) and final (Y) state potentials.

$$V_m = (1 - \lambda_m)V_X + \lambda_m V_Y, \tag{2.95}$$

where λ varies from 0 to 1 and V_m goes from V_X to V_Y. The free energy change can now be calculated by summing over the intermediate states,

$$\begin{aligned} \Delta G = & \quad \Delta G_1 + \Delta G_2 + \Delta G_3 + \Delta G_4 + \cdots \Delta G_{n-1} \\ = & \quad G_Y - G_X = -\beta^{-1} \sum_{m=1}^{n-1} \ln \left\langle e^{-\beta(V_{m+1} - V_m)} \right\rangle_X . \end{aligned} \tag{2.96}$$

The exponent in Eq. 2.96 can be written as $V_{m+1} - V_m = \frac{\partial V_m}{\partial \lambda_m} \Delta\lambda_m$, as long as the λ steps are sufficiently small. Eq. 2.96 then takes the form,

$$\Delta G = G_Y - G_X = -\beta^{-1} \sum_{m}^{n-1} \ln \left\langle e^{-\beta \frac{\partial V_m}{\partial \lambda_m} \Delta\lambda_m} \right\rangle_X . \tag{2.97}$$

This equation can be linearised, for small steps in λ, by retaining only the leading terms in the Taylor expansion of the exponent to give us,

$$\Delta G = \sum_{m}^{n-1} \left\langle \frac{\partial V_m}{\partial \lambda_m} \right\rangle_m \Delta\lambda_m. \tag{2.98}$$

This can be written as an integral over λ with $\lambda \rightarrow 0$

$$\Delta G = \int_0^1 \left\langle \frac{\partial V}{\partial \lambda} \right\rangle d\lambda. \tag{2.99}$$

The use of a linear combination for intermediate potentials means that only the two end point potentials need to be used and the forces and energies are simply scaled by the appropriate λ_m coefficient.

There are a large variety of algorithms available to preform TI simulations The "slow growth" method involves a single topology, were λ is changed during the MD simulation, changing state X into state Y. Another method is the dual topology approach, where a separate trajectory is obtained for every λ value. This approach has the advantage of allowing equilibration at each point, as well as the ability to add additional λ points at any time to increase the smoothness of the transition to optimise and improve convergence. This allow this approach to be efficiently parallelised since each λ window is ran independently of the others, and the accuracy to be systematically improvable.

The TI calculations often use soft-core potentials, which generally use a modified

Lenard Jones (LJ) equation, of the form,

$$
\begin{aligned}
V_{ij} = 4\epsilon_{ij}\left(1-\lambda\right)^{t}&\left[\alpha_{LJ}\lambda^{s}+\left(\frac{r_{ij}}{\sigma_{ij}}\right)^{n}\right]^{-\frac{12}{n}} \\
&- 4\epsilon_{ij}\left(1-\lambda\right)^{t}\left[\alpha_{LJ}\lambda^{s}+\left(\frac{r_{ij}}{\sigma_{ij}}\right)^{n}\right]^{-\frac{6}{n}},
\end{aligned}
\tag{2.100}
$$

in which ϵ_{ij} and α_{ij} are common LJ parameters, r_{ij} is the interatomic distance, α_{LJ} adjusts the softness of the potential and t,s, and n are set to 1, 2, and 6 respectively in the original formulation of the potential function. The use of soft-core potentials allows oppositely charged particles to come to close to each other, since at low λ values the repulsive force in Equation 2.100 is weakend, leading to numerical instabilities in the simulation. Since as $\mathbf{r}_{ij}$ decreases the Coulomb potential can give rise to numerical singularities. This introduces an additional practical problem when running such calculations. A solution to this is that the vdW mutation from X→Y is done without the partial charges on the atoms (zero charge atoms). To remove this problem the additional steps are added, of removing the partial charges (CR), mutating the atoms (VDW), and adding the new partial charges (CA), following the free energy cycle in Figure 2.17. This is performed for both the bound and unbound systems following the TI route in Figure 2.16.

2.5.3 Molecular Mechanics Poisson-Boltzmann Surface Area

Ligand binding affinity is calculated as,

$$\Delta G_{bind} = G_{complex} - G_{receptor} - G_{ligand}. \tag{2.101}$$

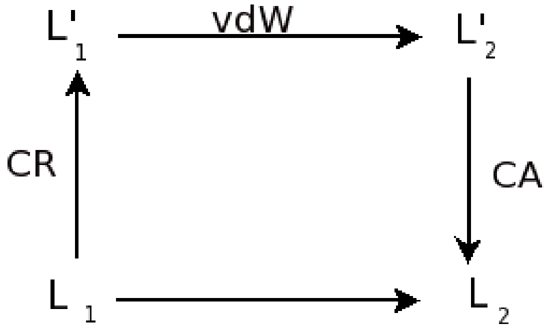


Figure 2.17: Free energy cycle. Removing partial charges going up, mutating the atoms going right, adding the new partial charges going down.

Molecular Mechanics Poisson-Boltzmann Surface Area (MM-PBSA) is a method of computationally predicting ligand binding affinity. The approach is based on the postprocessing of a molecular dynamics trajectory, typically ran in explicit solvent and counterions in a periodic box. The free energy of binding is computed by extracting a representative structural ensemble of “snapshots” from the trajectory. Solvent molecules and counterions are removed, then MM is used to calculate the gas phase interaction energy and a continuum solvent model (PBSA) to calculate the solvation energies.

The free energy of binding is then obtained as the average over the ensemble of structures as shown in Figure 2.18, with the interaction energies from each snapshot being calculated using the free energy cycle in Figure 2.19.

$$\begin{aligned}
\Delta G_{bind,solv} &= \Delta G_1 + \Delta G_2 + \Delta G_3 \cdots + \Delta G_n && (2.102)\\
&= \langle \Delta G_{bind,vac} \rangle + \langle \Delta G_{solv3} - \Delta G_{solv1} - \Delta G_{solv2} \rangle && (2.103)\\
&= \left\langle \Delta E^{MM} \right\rangle + \langle \Delta G_{PBSA} \rangle - T \left\langle \Delta S^{MM} \right\rangle . && (2.104)
\end{aligned}$$

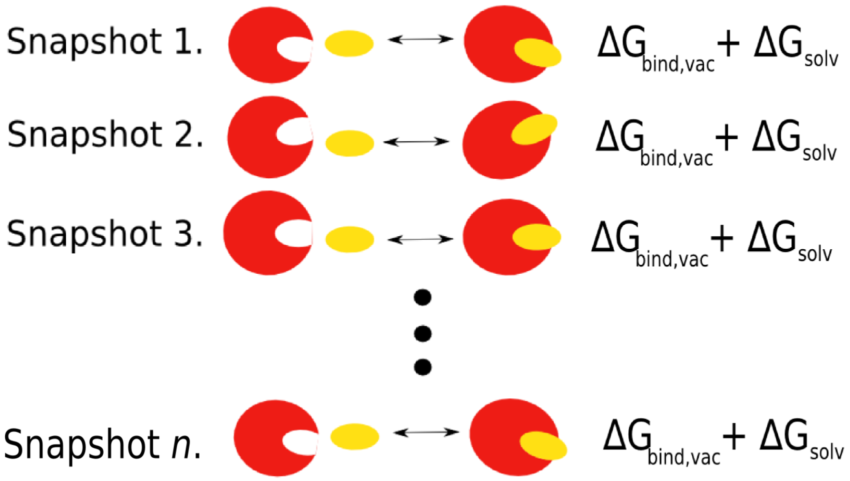


Figure 2.18: Binding free energies obtained from averaging of interaction energies over structural ensemble taken from MD simulation.

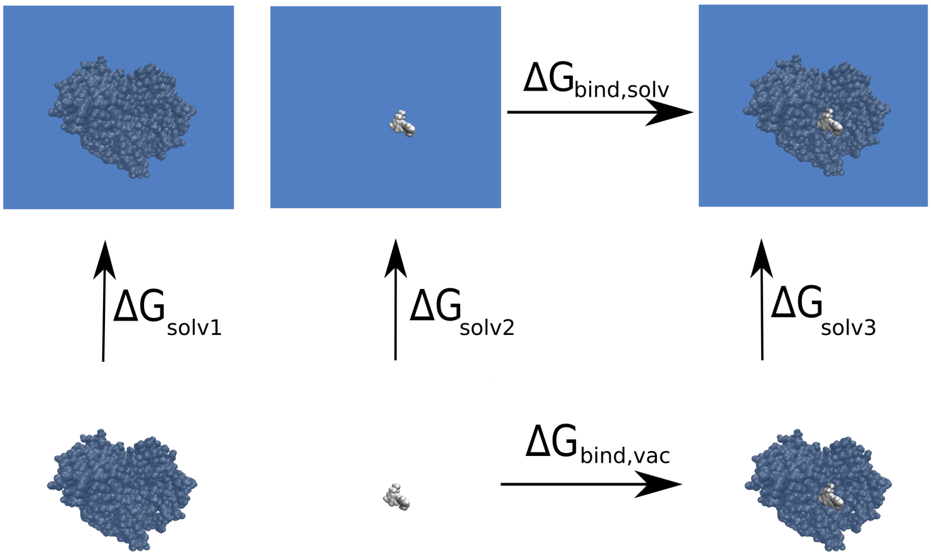


Figure 2.19: MM-PBSA free energy cycle.

In 2.104 $\left\langle \Delta E^{MM} \right\rangle$ is the interaction energy in vacuum, $\langle \Delta G_{PBSA} \rangle$ is the difference in solvation energy on binding, and $\left\langle \Delta S^{MM} \right\rangle$ is the change in entropy, which can be obtained using normal mode analysis. By using a continuum solvent model, the problem is simplified since we are implicitly integrating out all the solvent coordinates which results in more rapid convergence with the number of snapshots.

The obvious way to calculate this is the three trajectory approach, where separate simulations are carried out for the complex, receptor and ligand. However, it has been found that the one-trajectory approach, where only the complex simulation is ran and receptor and ligand configurations are extracted from the complex geometries, is more accurate due to error cancellation [87]. It is also 2-3 times faster, since the most computationally demanding part is the MD simulations and now only a single trajectory is required. However, this approach assumes that there are no conformational changes to the ligand or receptor upon binding. Furthermore, using the single trajectory approach, $\left\langle E^{MM} \right\rangle$ is the difference in non-bonded terms only, since all bonded terms will cancel.

Calculation of the entropy in a consistent and accurate manner is challenging. This makes calculation of the absolute binding free energies difficult. An approximation often used is to calculate, instead, the relative binding free energies of similar ligands. In this case, entropy change is assumed to be comparable for the two ligands and, can be assumed to cancel. Although this may seem a poor assumption, calculating the entropy of the ligand from structures taken from the complex trajectory may well be an equally poor simplification; Since the ligand geometry is extracted from the complex the structural ensemble is constrained and would be expected to have many more degrees of freedom when free in solution. The relative free energy is hence calculated as,

$$\Delta\Delta G_{A\rightarrow B} = \Delta\langle\Delta E^{MM}\rangle_{A\rightarrow B} + \Delta\langle\Delta G_{solv}\rangle_{A\rightarrow B}. \tag{2.105}$$

A significant source of error in MM-PBSA can be the accuracy of the interaction energies computed for each snapshot, as this accuracy depends on the parameterisation and transferability of the selected force field. Apart from the obvious approach of using more advanced force fields, a related direction for improvement is to replace either part or all of the force field description by a quantum description of the system. This would be expected to be more accurate and transferable due to explicitly accounting for the electronic effects, which are the source of all the interactions.

2.5.4 Some other binding free energy approaches

Scoring functions

Scoring functions are empirically trained functions using a large data base of known binders. They are most commonly used in combination with docking to obtain quick and predictive estimates of ligand affinities.[88, 89, 90, 91, 92] There are many different scoring functions available, they are however very simple and often not very accurate for molecules that do not fit the training set very well.

Linear interaction energy / Linear response approximation

In the linear interaction energy [93, 94] (LIE) method two MD simulations are ran: one for the ligand in the solution, and an other for the ligand in the protein binding site. Snapshots are extracted from the trajectories to represent Boltzmann ensembles of structural conformations. Boltzmann-averaged electrostatic and van der Waals interaction energies are then computed for the ligand with its surrounds, in the bound and unbound states. The binding free energy is estimated as a linear combination of the differences in potentials between the bound and unbound states over Boltzmann averages.

“Don’t have a fixed idea in your head. Use everything you’ve learned until now.”

Soke Hatsumi Massaki

Chapter 3

T4 lysozyme

Lysozymes are enzymes that act as a natural form of protection from pathogens, forming part of the innate immune system. They destroy bacteria by attacking the carbohydrate chains which are one of the main structural component of the bacterial cell wall ("skin") that supports their delicate membranes against the cell's high osmotic pressure. The process involves catalysing hydrolysis of 1,4-beta-linkages between N-acetylmuramic acid and N-acetyl-D-glucosamine residues. The lysozyme binds to the bacterial cell wall and destroys its structural integrity so that the bacteria burst under their own internal pressure.

There has been a great deal of research into protein stability, folding and design by looking at mutations of the lysozyme from the bacteriophage T4. T4 lysozyme can only hydrolyse substrates which have peptide side chains bonded to the polysaccharide backbone. Two well studied mutants of T4 lysozyme are Leu99Ala (L99A) [95, 96, 97] and Leu99Ala/Met102Gln (L99A/M102Q) [98, 99]. These mutations create a small buried apolar and polar cavity respectively, which are capable of encapsulating small aromatic ligands. Both the T4 lysozyme single mutant L99A and the double mutant L99A/M102Q have been used to compare and validate binding free energy methods and to develop docking procedures [98, 100]. Some of these have been briefly summarised below.

Wei *et al* [98] used the single mutation, L99A, of T4 lysozyme to investigate how the atomic charges and solvation energies affect the molecular docking and the quality of scoring functions. To further investigate the new atomic charges for their docking model and qualify their predictive ability, the apolar Met102 residue in the binding pocket was mutated to a polar Glutamine (making a double mutant L99A/M102Q). They repeated their docking procedure with this double mutant and tested seven molecules that, on the basis of the simulations, bound preferentially to the polar site (over the apolar site) experimentally via isothermal titration calorimetry (ITC) to verify their results. From this work they concluded that better treatment of the atomic charges and desolvation energies can lead to better distinction between binders and non-binders. They later (2004) [101] used the L99A to evaluate the ability of a new flexible-receptor docking algorithm, using around 200,000 molecules from the Available Chemical Directory. They found that larger ligands bound more favourably, but after adding an energy correction to account for the formation of the larger cavity they obtained improved ranking. To test their method they then used the L99A/M102Q mutant of T4 lysozyme. They predicted 18 new binders and tested these experimentally. Of the 14 experimentally confirmed binders, the bound structures of 7 were determined from x-ray crystallography. In conclusions this work found that improved enrichment of docking can be obtained from sampling receptor flexibility, however, it is important to account for receptor conformational energy.

Graves *et al* [102] also used both T4 lysozyme mutants, as well as β-lactamase, to test their docking and scoring functions. They looked at using decoy databases to improve protein structure algorithms. By using these simple cavities, the decoys can be used to highlight weaknesses in their scoring functions. They used a mixture of geometric decoys and "hit list" decoys, which after being ranked high by a number of docking algorithms were tested experimentally and confirmed as decoys.

Boyce *et al* [100] used binding free energy methods to predict the binding affinities of previously untested ligands for the T4 lysozyme double mutant L99A-/M102Q. A large library of small organic molecules were docked, from which thirteen ligands were chosen. The binding free energy was obtained computationally and experimentally using ITC. In addition 6 phenol derivatives were chosen, and the relative binding energies calculated relative to phenol and catechol. X-ray co-complex structures were obtained and the bound geometries compared to the computational comformations to help understand, at the atomistic level, the obtained computational results and the variation seen from experiment. It was concluded that it is important to start from near native binding poses and that unexpected binding modes, protein conformational changes and multiple ligands binding all proved challenging to the computational free energy methods.

Gallicchio *et al* [99] used the two cavities to present their new binding free energy approach, the Binding Energy Distribution Analysis Method (BEDAM). This approach is based on a statistical mechanics theory of molecular association. The binding constant in BEDAM is computed by a weighted integral of the probability distribution of the binding energy from the canonical ensemble, in which the ligand is positioned in the binding site, but both the receptor and the ligand interact only with the solvent continuum. In the paper it is shown that the binding energy distribution encodes all of the physical effects of binding. This method successfully distinguished between known binders and non-binders. They conclude their paper by stating that in these two systems the binding affinities are reflected in the contributions from multiple conformations over a wide range of binding energies.

Deng *et al* [103] presented a review of results on 5 systems, including the two mutants of T4 lysozyme, using two binding free energy approaches; alchemical double decoupling, were the environment surrounding the ligand is turned off and the potential of mean force method, were the ligand is physically separated from the receptor. Restraining potentials are activated and released during the simulations to increase configurational sampling, but the bias added by doing this must be rigorously accounted for when calculating the binding free energies. They observe that it is difficult to account for induced conformational changes on binding without biasing the sampling, but this requires prior knowledge of the relevant degrees of freedom. This is only a part of the overall problem though, the accuracy of the computations is determined by the force field. Binding energies calculated from force fields can benefit from fortuitous error cancellation. Dependable results require accurate representation of the ligand, receptor and solvent. They conclude by stating that “while there is still much to be done, the methods are already bearing fruits and the path towards progress is very clear.”

For our study we have chosen to investigate small aromatic ligands binding to the polar cavity of the double mutant L99A/M102Q. The relative simplicity and small size of this system make it attractive for validating computational studies. Coupled with the abundance of literature, it is a good choice for our benchmark calculations. Below describes the MD simulation set up and equilibration . Further use of this system is detailed in the next chapter.

3.1 T4 lysozyme double mutant L99A/M102Q

The first mutation to this protein changes leucine 99 into an alanine. This creates a buried, hydrophobic pocket capable of encapsulating small molecules. When this pocket is empty, it is completely dry [98] under normal conditions (1 atmosphere of pressure). The second mutation changes methionine 102 to a glutamine. This mutation adds a polar binding site within the cavity. Under normal conditions it is believed that the apo-protein contains a single water molecule in the cavity, hydrogen bonded to the glutamine. These mutations can be seen in Figure 3.1.

3.1.1 Molecular dynamics simulations

The lysozyme structure was protonated with the MOE2010 [104] program using Pronate3D and visually check the His, Gln, and Asn residues. MM simulations were carried out using the AMBER10 [54] package, with the ff99SB [105] forcefield used for the protein and the generalised AMBER forcefield [59] (gaff) used to model the ligands. Ligand charges were calculated with the AM1-BCC method with antechamber (part of AMBER). The system was explicitly solvated in the TIP3P water model [70] and the charge neutralised by Cl^- ions.

The system was equilibrated using the following protocol. Hydrogens were relaxed with restraints placed on all heavy atoms in the complex and solvent, before relaxing the solvent with restaints only on the complex. The system was heated to 300 K over 200 ps, still restraining the heavy atoms of the complex, with the NVT ensemble. Then ran for a further 200 ps with the NPT ensemble at 300 K in order to equilibrate the solvent density. This was cooled over 100 ps to 100 K and a number of relaxations were ran, reducing the restraints on the heavy atoms in stages $(1000, 500, 100, 50, 20, 10, 5, 2, 1, 0.5$ kcal mol^{-1}Å$^{-2})$. Finally the system was reheated to 300 K with no restraints over 200 ps and then for a further 200 ps at 300 K with the NPT ensemble. At the end of this it was confirmed that the water density in the box (Figure 3.2), the energies (Figure 3.4) and the protein structure were stable, as measured by the root mean squared deviation of the protein backbone atoms (converged to 0.8Å relative to the starting frame).

Since the binding modes of the ligands are all very similar, only catechol bound in the pocket (PDB: 1XEP) was equilibrated. All other ligands were mutated from the catechol at the end point of the equilibration. Production simulations were run for 20 ns with the NVT ensemble at 300 K, with the first 1 ns being considered as further equilibration of the ligand in the pocket. All MD simulations used the Langevin thermostat [67], the particle mesh Ewald sum (PME) for the electrostatic interactions and the SHAKE algorithm [65] to constrain hydrogen-containing bonds allowing a time-step of 2 fs.

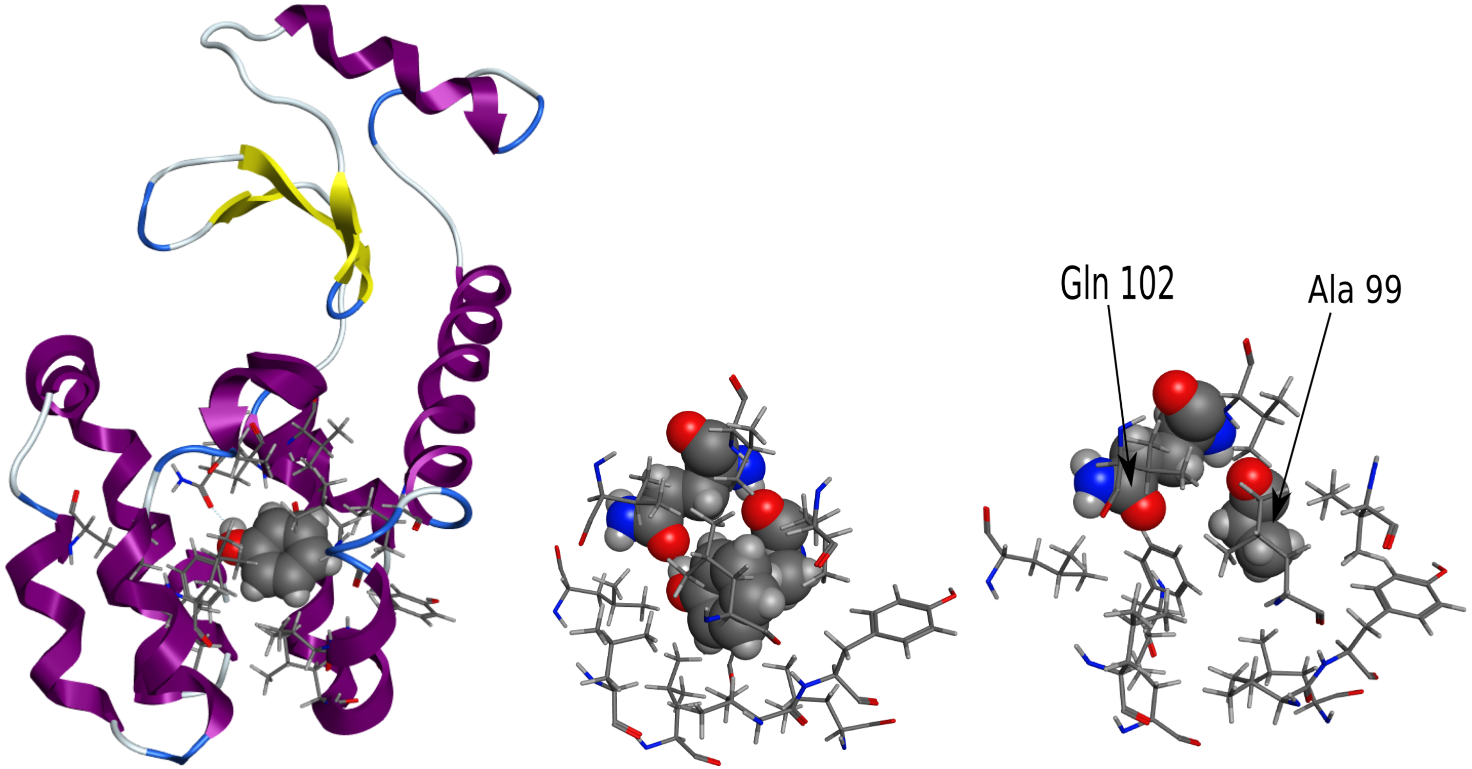


Figure 3.1: Phenol bound in the cavity of T4 lysozyme L99A/M102Q. Displaying the specific mutations in the cavity with a space filling depiction of the residues with the phenol ligand (centre picture), and without the phenol (right picture).

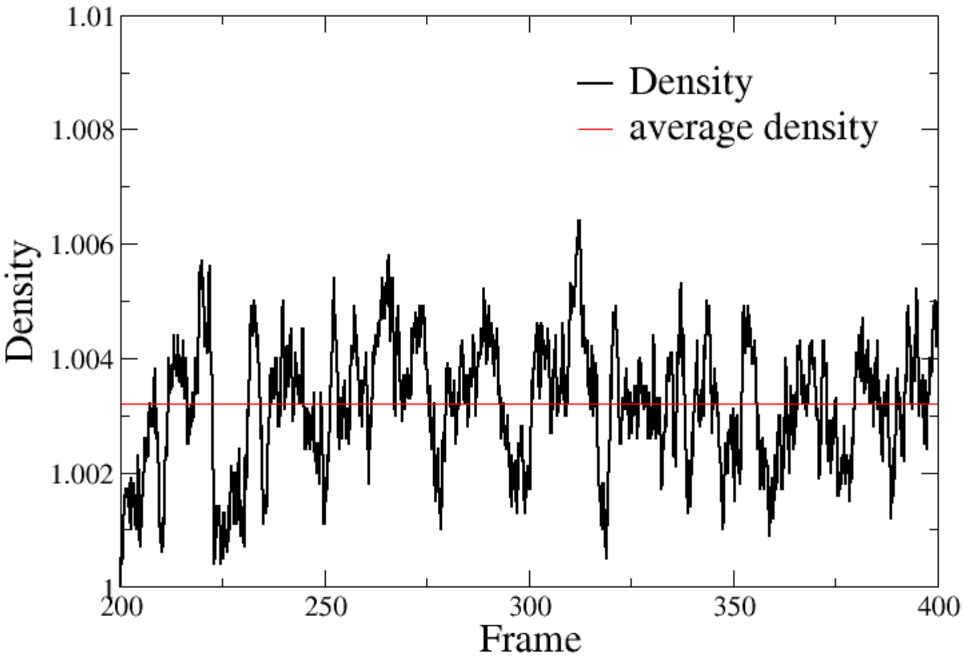


Figure 3.2: The density of the box for the final step of the equilibration. Frames were recorded every 0.5 ps.

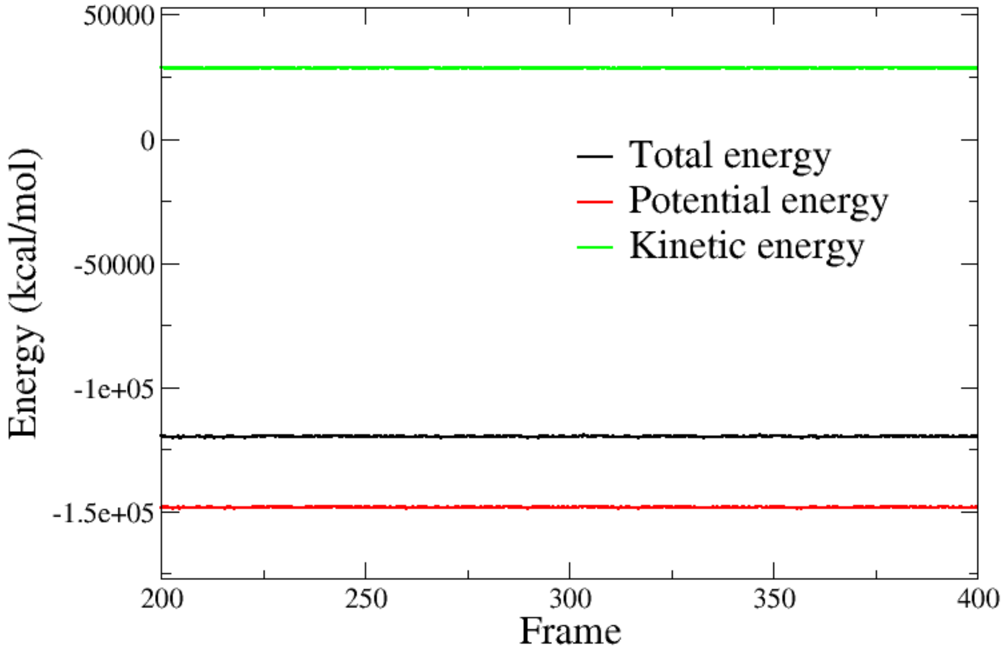


Figure 3.3: The total, potential and kinetic energies for the final step of the equilibration. Frames were recorded every 0.5 ps.

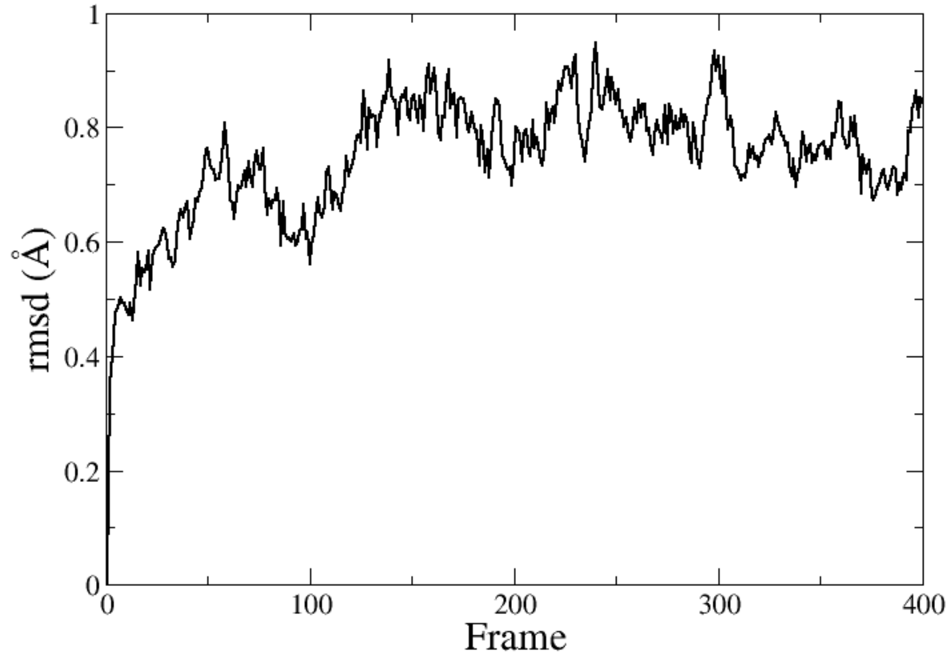


Figure 3.4: The rmsd of the backbone for the final step of the equilibration. Frames were recorded every 0.5 ps.

Chapter 4

QM-PBSA

The accurate prediction of drug binding affinities is an ongoing goal within computational drug optimisation and development. A quantitative measure of binding affinity is provided by the free energy of binding. Such calculations typically require configurational sampling of entities such as proteins with thousands of atoms and are beyond the reach of conventional ab initio quantum chemistry approaches. The sampling of configurations and energies is usually carried out with force fields, using a variety of approaches. One such approach is Molecular Mechanics Poisson-Boltzmann Surface Area (MM-PBSA), which obtains free energies from evaluation of the energy of configurations in an implicit solvent model. A limitation of MM-PBSA is the force field, which can potentially lead to large errors due to the restrictions in accuracy imposed by its empirical nature.

Kaukonen *et al* [106] presented a QM/MM-PBSA approach and compared it to MM-PBSA for the purpose of studying reactions in proteins. The QM system consisted of 46 atoms with the MM part consisting of 12132 atoms. The aim was to study the stability of two states with a shared proton. The QM calculations were performed with DFT using the BP86 exchange-correlation functional with a 6-31G* basis set and DZP for metal ions for one system, and the B3LYP functional for another system using the same basis sets. This method showed improved results for the proton transfer and was in good agreement with more rigorous approaches (QTCP [107]) with median absolute deviations (MADs) of 4-22 kJ/mol.

Wang *et al* [108] used the SIESTA DFT code [109] combined with the implicit solvation model in the UHBD software [110] in a QM/MM-PBSA approach. They used a fixed geometry (single structure) approximation, were only minimised crystal structures are used and no protein configurational sampling was performed. The only differences between the ligands in the pocket was a single chemical functional group, with all common atoms being in identical positions. This was done to improve the odds of cancellation of systematic errors when comparing binding free energies. Relaxed structures were generated in three different ways. The first was geometry optimised with SIESTA [109] and the second and third in a QM/MM approach with the ONIOM method in GAUSSIAN 03 [79].

Diaz *et al* [15] replaced $\Delta\langle\Delta E^{MM}\rangle$ with energies from linear-scaling semi-empirical QM calculations on an ensemble of structures from an MM MD simulation in a QM-PBSA type model. Prior to the single-point energy calculations, QM/MM geometry optimisations of a subsystem of the enzyme were performed, keeping the rest of the enzyme fixed. Single point energy calculations were performed with AM1 [24] and PM3 [111] using the divide and conquer (D&C) approach on the subsystem of the enzyme. The DivCon99 [112] program was used to perform the D&C calculations. They found that the resulting QM/MM geometry optimised structures were similar to the MD representations generated from the force fields, and using semi-empirical QM D&C gave comparable relative binding free energies to MM-PBSA.

Cole *et al* [113] have recently extended the MM-PBSA approach to a full QM-PBSA approach, with sampling of protein motion, where the calculation of the interaction energies in vacuum by the force field is replaced by DFT calculations on the entire molecule for an ensemble of snapshots taken from an MD simulation. The energy of each snapshot is obtained as $E^{QM} = E_{\text{DFT}} + E_{\text{disp}}$, where E_{disp} is the dispersion correction [5] to the total DFT energy, E_{DFT}. In previous work [113, 114], the free energy of solvation in the QM calculation, G^{QM}_{solv} was obtained by scaling the classical solvation energy by the QM electrostatic energy, giving the free energy of binding as,

$$\begin{aligned} \Delta G_{tot} &= \langle \Delta E^{QM} \rangle + \langle \Delta G_{\text{PB}} \left(\frac{\Delta E_{\text{DFT}}}{\Delta E_{\text{EL}}} \right) + \Delta G_{\text{SA}} \rangle && (4.1) \\ &= \langle \Delta E^{QM} \rangle + \langle \Delta G^{QM}_{\text{solv}} \rangle, && (4.2) \end{aligned}$$

where ΔE_{EL} is the electrostatic contribution to the binding energy from the MM calculation, ΔG_{PB} is the polar term from the solvation energy and ΔG_{SA} is the non-polar term. The first application of QM-PBSA with ONETEP has been on protein-protein interactions [113]. The results obtained were in good agreement with MM-PBSA, most likely because the force field employed has been extensively and carefully parametrised for protein systems and improved over a number of years.

This chapter will detail the work we have been doing towards a more accurate QM-PBSA approach. This approach was applied to a model of a host-guest system [114] where the force fields are much more general and harder to parametrise, and a model protein-ligand system, where force fields can be well parametrised to describe the protein.

4.1 The Tennis ball dimer

This host-guest system was chosen for evaluating the use of first principles calculations in combination with a classical force field to simulate host-guest interactions. The system we have selected to study is a model for a protein ligand-binding cavity based on a self-assembling superstructure, the "tennis ball" dimer (Figure 4.1). The QM free energy of solvation was calculated in a slightly different way to that proposed by Cole *et al*, using the following,

$$\Delta G^{QM}_{solv} = \Delta G^{MM}_{solv} \left(\frac{\Delta E^{QM}}{\Delta E^{MM}} \right), \tag{4.3}$$

where ΔE^{QM} is the total QM energy, ΔE^{MM} is the total binding energy from the MM force field, and, as in usual MM-PBSA, is averaged over the snapshots and added to the total DFT energy to give the free energy of binding as,

$$\Delta G_{tot} = \langle \Delta E^{QM} \rangle + \langle \Delta G^{QM}_{solv} \rangle. \tag{4.4}$$

In this system the scaling method in Equation 4.1 does not work since dispersion interactions are responsible for most of the binding energy. The result of this is that the MM electrostatic component of the binding energy, in the denominator, is very close to zero, effectively multiplying the solvation energy by a very large number leading to a numerical error. We found that the simpler form shown in Equation 4.3 produces reasonable solvation energies. We have chosen this model

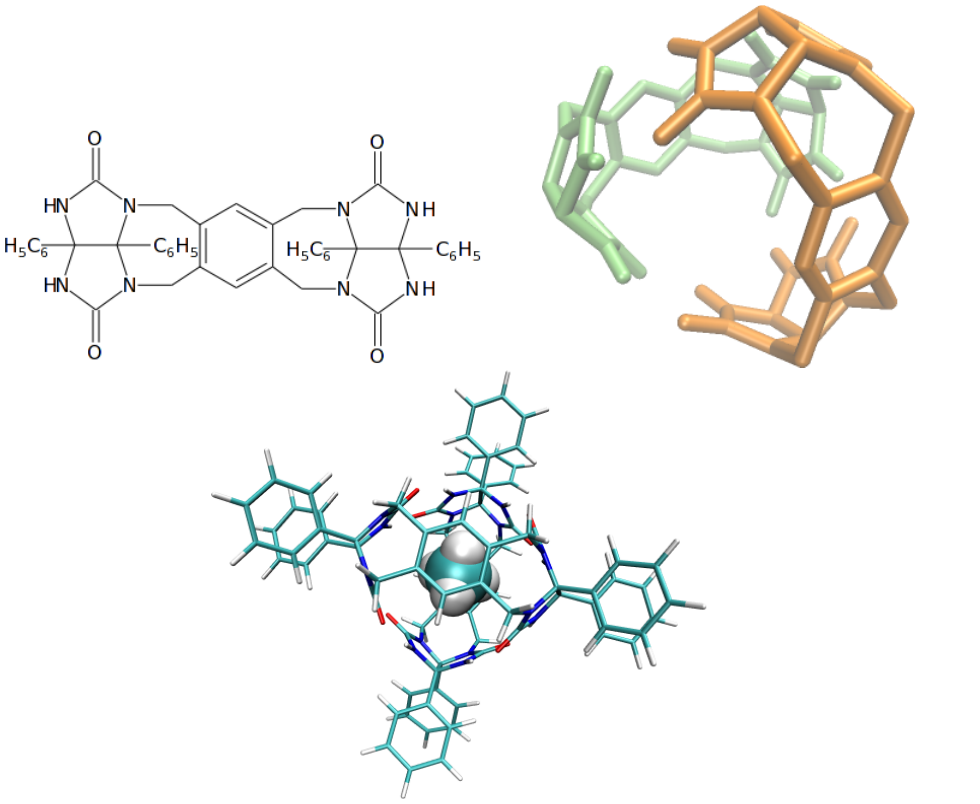


Figure 4.1: 2D diagram of the monomer (left), Truncated structure of the "tennis ball" depicting the shape of the cavity (middle), Encapsulation of a methane molecule in the whole dimer (right).

as it combines simplicity with realism and also because there are previous computational [115] studies and experimental [116] data to compare with. We first compare structure optimisation with a force field and first principles approaches in terms of the structural parameters. We then introduce dynamic effects through molecular dynamics simulations and compare binding energies calculated from MM-PBSA and QM-PBSA to experimental values.

4.1.1 Simulation details

The tennis ball structure was built and loosely minimised with the MOE [104] program. MM simulations were carried out using the AMBER10 [54] package. The tennis ball was modelled using the generalised AMBER force field [59] (gaff) and solvated with the $CHCl_3$ explicit solvent model (as implemented in AMBER10) in a periodic box.

To equilibrate the system, the hydrogens were relaxed keeping all heavy atoms restrained in the host and solvent, then relaxing the solvent with restraints still on the host. The system was heated to 300 K still restraining the host for 200 ps with the NVT ensemble and ran for a further 200 ps with the NPT ensemble at 300 K in order to equilibrate the solvent density. This was cooled to 100 K over 100 ps and minimisation's carried out reducing the restraints on the host heavy atoms in stages ($500, 100, 50, 20, 10, 5, 2, 1, 0.5$ kcal mol^{-1}Å^{-2}). Finally the system was heated to 300 K with no restraints over 200 ps and then ran for a further 200 ps at 300 K with the NPT ensemble, at the end of which the root mean squared deviation of the C,N and O atoms was converged and less than 0.8Å relative to the starting frame. Production simulations were run for 2 ns with the NPT ensemble at 300 K. All MD simulations used the Langevin thermostat, the particle mesh Ewald sum (PME) for the electrostatic interactions, a time-step of 2 fs and the SHAKE algorithm [65]. For the MM-PBSA calculation an infinite non-bonded cutoff was used with a dielectric constant of 4.5 to represent the chloroform solvent. All ONETEP single point energies were converged to 0.0002 Hartree ($\sim$0.1 kcal mol^{-1}). 4 NGWFs were used to describe carbon, oxygen and nitrogen, 1 for hydrogens and 9 NGWFs for the halogen atoms. A kinetic energy cutoff of 800 eV for the psinc basis set was used, with the GGA exchange-correlation functional

PBE [34] combined with our implementation of the DFT+D approach to account for dispersion parametrised specifically for this functional [5].

4.1.2 Results and Discussion

Validation tests

In cases where two different approaches are used to explore the conformational space, the compatibility of the methods used is an important consideration [17]. Firstly, it is desirable that the minima on the potential energy surface between the QM and the MM approach are as close as possible. To investigate this we have carried out geometry optimisations of the three complexes using ONETEP and AMBER. We have also carried out further validation of the QM approach by doing the same geometry optimisations with the Gaussian [81] program which can perform all-electron DFT calculations with Gaussian basis sets. For these all-electron calculations we used a correlation consistent split valence basis set (cc-pVDZ [117]) and the B97 exchange-correlation functional [118] with the DFT+D approach for including dispersion contributions as parametrised by Grimme *et al* [53]. The structural parameters between the optimised geometries by the three methods were compared. Bond lengths vary by less than 0.03 Å and internal angles, such as those within rings, vary by less than 0.5°, with the more flexible angles differing by 2-3°. Hydrogen bonds from ONETEP (Gaussian) are shorter than these from the AMBER optimised structure by 0.2 Å (0.1 Å), and the distance separating the monomers differs by as much as 0.5 Å between the ONETEP and AMBER structures. All the methods predict hydrogen bonds which are longer by around 0.2 Å for the $CHCl_3$ complex compared to the tennis ball complexed with

CH_4 or CF_4 and the empty dimer, which is to be expected as the $CHCl_3$ is slightly larger than the size of the empty cavity.

As we are interested in properties at finite temperatures (usually room temperature), using only equilibrium geometries is not sufficient as dynamical motion causes the molecules to visit many configurations which can differ from the relaxed structures. Thus, MD simulations are run for time-scales which are long enough (ns) to sample the dynamical behaviour of this system, using the classical force field approach. The importance of accounting for dynamic motion for the tennis ball system is shown in Figure 4.2. Here we examine hydrogen bond lengths in the CH_4 and $CHCl_3$ complexes throughout the 2ns MD simulations. During the simulation the hydrogen bonds in the CH_4 complex are stable, staying at around 2 Å. In contrast, the hydrogen bonds in the $CHCl_3$ complex are intermittent: we observe that the dimer opens at a point, to around 4 Å, then moves back into position, re-establishes the hydrogen bond and breaks at another point. This happens due to the size of the chloroform ligand; it is too large to fit comfortably between the monomers causing the cavity to open and close during the simulation. Figure 4.2 demonstrates that the $CHCl_3$ complex has one hydrogen bond broken most of the time. In this case the minimum energy structure which has all the hydrogen bonds intact, albeit elongated, will not provide an adequate description of the ensemble of structures encountered at room temperature. We can demonstrate this further by noting that the binding energy for the $CHCl_3$ complex as calculated with ONETEP on the optimised structure is 2.6 kcal mol^{-1} while when taking into account 200 snapshots extracted from the MD ensemble it is -7.1 kcal mol^{-1}, in close agreement with the experimental value of -7 kcal mol^{-1}.

As a dynamical ensemble of structures is necessary for this study we also need

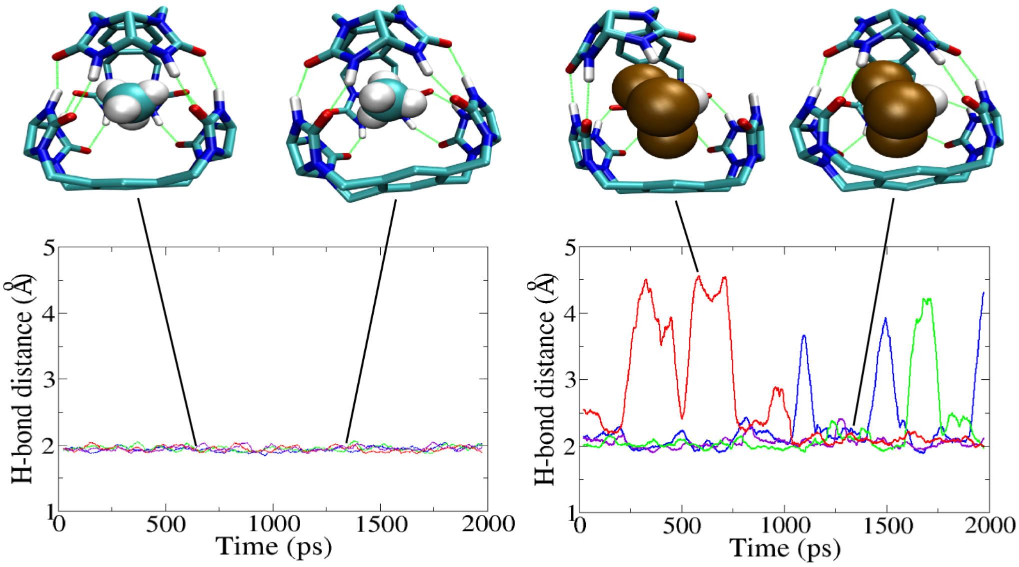


Figure 4.2: Plots of the hydrogen bond lengths from four H-bonding positions (C=O's of top monomer to H-N's of bottom monomer) in the CH_4 complex (left) and $CHCl_3$ complex (right). Structures taken as "snapshots" at two points of the simulations are shown, the green dashed lines represent the hydrogen bonds present at each snapshot. In the graphs, the four coloured lines correspond to the four hydrogen bonds measured in each complex. (Phenyl-rings not displayed)

Table 4.1: Average (maximum) of forces on atoms from AMBER and ONETEP from 10 snapshots. Values in kcal mol^{-1} Å^{-1}.

Complex	ONETEP	AMBER
CH_4	29.3 (153.0)	29.9 (107.0)
$CHCl_3$	29.3 (147.8)	29.7 (105.0)
CF_4	30.1 (150.8)	30.1 (128.1)

to confirm that the conformations sampled by the force field are not unphysical as far as the QM potential energy surface is concerned. To explore this issue, we have compared forces on atoms calculated from ONETEP and AMBER on several of the snapshots. An indication that the compatibility of the two approaches is good in this case is given by the values reported in Table 4.1 which presents the average (maximum) of the absolute values of the force on all atoms, over 10 equally-spaced snapshots. Even though these agree extremely well between the two approaches, if we look in more detail at individual atoms the agreement is not so good. The largest difference between the QM and MM forces on any single atom is ~80 kcal mol^{-1} Å^{-1}, but for most atoms it is less than 20 kcal mol^{-1} Å^{-1}. Fig. 4.3 compares the forces between QM and MM between individual atoms for a single snapshot of the CH_4 complex, coloured according to the type of element. We can observe that for hydrogen and carbon atoms both ONETEP and AMBER forces agree reasonably well. The large differences are on the oxygen and nitrogen atoms are as expected, this is because the parametrisation of the ligand is done in group-wise fashion, so an urea-group will have a charge of one, but it is not clear how the charges are distributed over the atoms, thus the charges for heteroatoms will strongly differ from ONETEP causing a strong difference in gradient. This behaviour is representative of all snapshots for the three systems.

Our comparisons show that there is substantial variability in the forces obtained with the two approaches, however the forces in both cases are within expected

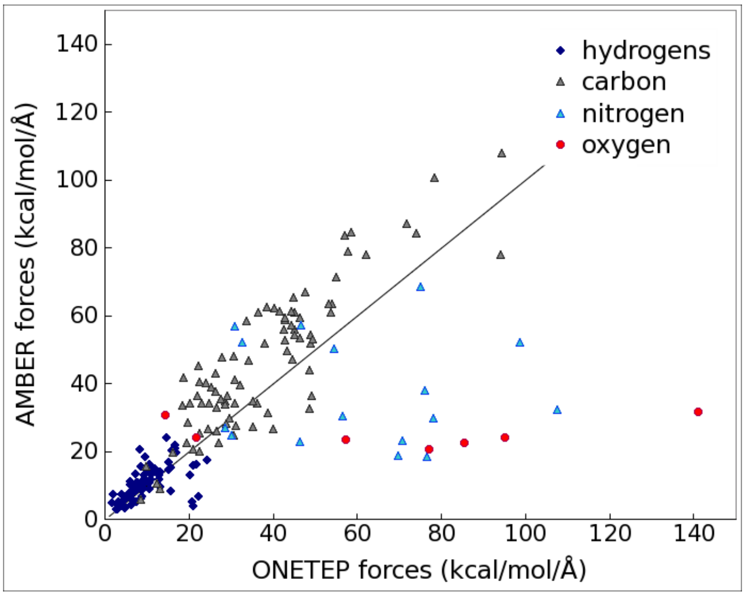


Figure 4.3: Correlation between $|\mathrm{F}_{QM}|$and $|\mathrm{F}_{MM}|$ for a single snapshot of the CH_4 complex. Other snapshots and complexes show similar behaviour.

ranges and the average forces are comparable. This suggests that no unphysical conformations are visited by the force field.

Having established the importance of taking into account the dynamical behaviour of this system, we finally tested the convergence of PBSA energies as a function of the number of MD snapshots. An increasing number of snapshots was used, obtained by sampling uniformly through the 2ns production simulations. 50, 100, 160 and 200 equally-spaced snapshots were extracted from each simulation to study the convergence. We found that the variation in the binding free energies calculated in ONETEP or AMBER when going from 50 snapshots to 200 snapshots was less than 0.2 kcal mol^{-1} for all systems studied. All MM-PBSA and QM-PBSA results we report here were obtained using 200 snapshots.

Free Energies of Binding

The energies of binding that were obtained with the MM-PBSA and QM-PBSA approaches for all the complexes are presented in Table 4.2. The table shows the enthalpies of binding (ΔH) computed from either the force field or the DFT calculations with ONETEP and the free energy of binding (ΔG) which includes solvation contributions. We obseve that for CH_4 AMBER predicts a ΔH that agrees well with experiment (to <0.3 kcal mol^{-1}) however it overestimates the ΔH for the halogen containing ligands to over twice the experimental value. This suggests that the force field does not capture well the interaction energies of the halogen atoms with the cavity. ONETEP produces ΔH values that are in close agreement (within 0.1 kcal mol^{-1}) to the experimentally determined ΔH values, which supports further our earliest observation that the ensemble of structures provided by the force field has a high overlap with the QM ensemble. The larger standard errors in the calculated energy differences for the $CHCl_3$ complex, 0.27 kcal mol^{-1} compared to 0.04 kcal mol^{-1} for CH_4, are expected since this structure shows considerably more fluctuation than the other complexes, as we saw in Figure 4.2. Standard error is calculated as the standard deviation over the square root of the number of data points.

Since AMBER overestimates the interaction energies for the halogen containing ligands, the calculated $\Delta\Delta G$'s predict a more favourable interaction than was found experimentally and in the previous computational study. Our improvements by the QM calculations refer to the enthalpic part of the binding energies and indeed we can observe that the $\Delta\Delta H$ values are a very good match to experiment.

As the enthalpy is accounted for so well, and the free energy of solvation in this case makes a minimal contribution due to the non-aqueous solvent, the large discrepancy in the free energy differences ΔG can be attributed to the neglect of configurational entropy. When considering the $\Delta\Delta G$ values a large fraction of this error is cancelled and we obtain reasonably good agreement with experiment (2.7 kcal mol^{-1} versus 5.2 kcal mol^{-1} for $\Delta\Delta$G($CH_4 \longrightarrow CHCl_3$) and 3.2 kcal mol^{-1} versus 2.8 kcal mol^{-1} for $\Delta\Delta$G($CH_4 \longrightarrow CF_4$)) for ONETEP while the AMBER values show discrepancies of more than 5 kcal mol^{-1}, precisely due to the bad estimation of enthalpy.

Previous computational results by Fox *et al* [115] were obtained from TI calculations. Simulations were performed with AMBER4.1 using the all atom force field by Cornell *et al* [55] and partial charges obtained from a multiple molecule RESP fit. Table 4.3 compares their results to TI free energies we obtained with AMBER10 using the gaff force field and with our MM-PBSA and QM-PBSA results. We observe that both TI approaches obtain comparable relative binding free energies (7.8 kcal mol^{-1} versus 7.2 kcal mol^{-1} for $CH_4 \rightarrow CHCl_3$ and 0.9 kcal mol^{-1} versus 0.1 kcal mol^{-1} for $CH_4 \rightarrow CF_4$) and considerably better than MM-PBSA ($\Delta\Delta G$($CH_4 \rightarrow CHCl_3$) of 7.8 kcal mol^{-1} rather than -6.4 kcal mol^{-1}). While TI is a more rigorous approach which fully accounts for entropic effects, we observe that our QM-PBSA energies achieve improved agreement with experiment. So at least in this system the accurate description of interaction energies that is provided by the DFT calculations is critical for the correct calculation of free energy differences.

As we have mentioned in the introduction, force fields are significantly less computationally demanding than first principles quantum calculations, and this is reflected in our timings. For example, a single point energy force field calculation on

Table 4.2: MM-PBSA and QM-PBSA results presenting binding free energies separated into differences in enthalpy and free energy and relative differences between ligands using CH_4 as a reference.

	ΔH			ΔG		
	MM-PBSA	QM-PBSA	Exp. [116]	MM-PBSA	QM-PBSA	Exp. [116]
CH_4	-8.7 ± 0.05	-9.0 ± 0.04	-9	-8.5 ± 0.05	-8.9 ± 0.05	-3.0
$CHCl_3$	-16.8 ± 0.19	-7.1 ± 0.27	-7	-14.9 ± 0.16	-6.2 ± 0.23	2.2
CF_4	-12.6 ± 0.09	-6.0 ± 0.08	N/A	-11.9 ± 0.08	-5.7 ± 0.08	-0.2[115]

	$\Delta\Delta H$			$\Delta\Delta G$		
	MM-PBSA	QM-PBSA	Exp. [116]	MM-PBSA	QM-PBSA	Exp. [116]
$CH_4 \longrightarrow CHCl_3$	-8.1	2.1	2	-6.4	2.7	5.2
$CH_4 \longrightarrow CF_4$	-3.9	3.0	N/A	-3.4	3.2	2.8 [115]

ΔH is the energy in vacuum. ΔG is the vacuum energy plus the solvation energy (for MM-PBSA and QM-PBSA this term does not include conformational entropy).

Table 4.3: Relative binding free energies (kcal mol^{-1}) obtained via TI by Fox *et al*, TI results using the generalised amber force field and results from our QM-PBSA approach.

	TI [115]	TI with gaff	QM-PBSA	MM-PBSA	Exp. [116]
$CH_4 \longrightarrow CHCl_3$	7.8	7.2	2.8	−6.4	5.2
$CH_4 \longrightarrow CF_4$	0.9	0.1	3.2	−3.4	2.8

one of our complexes takes about 0.35 core-seconds on an Intel CORE2 machine, while the same calculation with DFT takes about 24 core-hours on the same computational platform. Therefore in terms of throughput, the force field calculations have a clear advantage. However, the point is, that in several cases the unbiased and accurate description that is provided by the first principles calculations can be indispensable. For example, electronic polarisation, or halogen-pi interactions, which are poorly described by available force fields. We therefore expect that large-scale first principles quantum calculations will be a valuable tool in the final stages of computational drug design where careful refinement is required. The linear-scaling formalism makes it feasible to extend the application of these calculations to biomolecules with thousands of atoms, especially in combination with new HPC technologies such as GPUs and peta-scale supercomputers.

4.1.3 Conclusions

In this work we have presented an approach for reducing some of the limitations of the MM-PBSA method. Towards this aim we have used the ONETEP program to calculate the QM interaction energies with solvation contributions extracted from a traditional MM-PBSA method and scaled to match the QM energies. Conformational space was sampled with classical force field molecular dynamics simulations and the compatibility of the structural ensemble, with respect to the potential energy surface, was checked by comparing forces on atoms between the two methods. These showed that although there was substantial variation in the forces obtained with the two approaches, the forces in both cases were within expected ranges and no unphysical conformations appear to be visited by the force field. This QM-PBSA approach obtained energies which are significantly improved over the MM computed energies, with enthalpic energies agreeing with experimental ΔH values to within 0.1 kcal mol^{-1}. The neglect of entropy leads to poor agreement with experimental absolute binding free energy values, however, relative binding free energies show considerable improvement agreeing well with experiment. These even show an improvement over the more rigorous TI method.

While the model we have studied is relatively simple and small (for biomolecular standards), it does include important and difficult to capture interactions such as halogen-pi interactions which are not at all well described by force fields and even hydrogen bonds whose accurate description by non-quantum methods is reasonable, but cannot be taken for granted. Therefore this is a small but important step towards modelling some of the crucial interactions in real protein-ligand systems.

4.2 T4 lysozyme L99A/M102Q

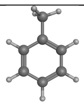
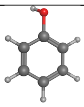
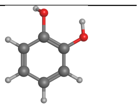
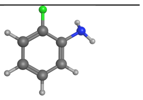
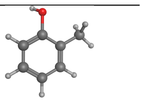
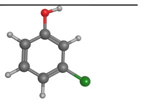
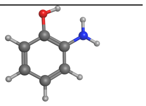
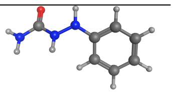
In this work we are presenting the first QM-PBSA study of a protein-ligand system where the entire protein of 2602 atoms has been described by DFT calculations. These calculations have been performed with ONETEP, and in contrast to our previous QM-PBSA studies in section 4.1, the solvation free energy has been computed within ONETEP with a newly implemented self-consistant minimal parameter implicit solvation model [78]. In this study we are comparing the free energy of binding from the conventional MM-PBSA approach with our QM-PBSA approach on 8 ligands bound the the T4 lysozyme double mutant L99A/M102Q. Rather than going back to the tennis dimer with a full QM solvent, this system was chosen. Since the solvent in the tennis ball work was chloroform, which has a small dielectric and little effect on the binding free energies, very little would be gained by including the solvent in the QM calculation. Water has a much higher dielectric and is the solvent most biological molecules will be in. This system is the next step towards protein-ligand systems of interest.

The 8 chosen ligand are displayed in Table 4.4. These ligands were chosen as they comprise of a variety of chemical and physical properties (polarity, inclusion of halides, size and binder/non-binder).

4.2.1 Simulation details

The protein-ligand system was equilibrated as previously discussed. Since the binding modes of the ligands are all very similar, only catechol bound in the pocket (PDB: 1XEP) was equilibrated. All other ligands were mutated from the catechol at the end point of the equilibration. Production simulations were

Table 4.4: Chosen ligands for study in the T4 lysozyme double mutant L99A/M102Q. Experimentally measured free energies of binding (ΔG_{exp}) are given in kcal mol^{-1}.

L99A/M102Q ligands	ΔG_{exp}	PDB code	structure
Toluene	-5.2 [98]		
Phenol	-5.5 [98]	1LI2	
Catechol	-4.4 [100]	1XEP	
2-fluoroaniline	-5.5 [98]	1LGW	
2-methylphenol	-4.4 [98]	3HT6	
3-chlorophenol	-5.8 [98]	1LI3	
2-aminophenol	non-binder		
1-phenylsemicarbazide	non-binder		

run for 20 ns with the NVT ensemble at 300 K, with the first 1 ns being considered as further equilibration of the ligand in the pocket. All MD simulations used the Langevin thermostat [67], the particle mesh Ewald sum (PME) for the electrostatic interactions and the SHAKE algorithm [65] to constrain hydrogen-containing bonds allowing a time-step of 2 fs. For the MM-PBSA calculation an infinite non-bonded cutoff was used with a dielectric constant of 80 to represent the water solvent. The 1-phenylsemicarbazide ligand was treated slightly differently. Since it is a lot bigger than catechol this ligand was built into the protein pocket using the structure of benzyl acetate (PDB:3HUK), which is structurally similar to 1-phenylsemicarbazide. This structure was then equilibrated in the same way as catechol followed by a 20ns NVT production simulation, the last 19ns of which were used to generate the snapshots for this study.

In the ONETEP calculations, 4 NGWFs were used to describe carbon, oxygen and nitrogen atoms, 1 NGWF for hydrogens and 9 NGWFs for the halogen atoms. A kinetic energy cutoff of 800 eV for the psinc basis set was used, with the GGA exchange-correlation functional PBE [34] combined with our implementation of the DFT+D approach to account for dispersion, parametrised specifically for this functional [5]. The QM implicit solvation model parameters were chosen after validation in previous work involving the same protein [119].

It is important to note that the QM total energies of the complex and host are of the order of millions of kcal mol^{-1}, in contrast the binding energies are only a few tens of kcal/mol, a minute value in comparison. For the accurate calculation of the binding energies the total energies have to be very well converged, and for systems of this size (+2600 atoms) this can be challenging. The parameters chosen for our ONETEP calculation were previously compared against a Gaussian basis set code on a small system and shown to very accurately calculate binding energies [4]. We further tested these parameters on our large system to observe total energy convergence for some indication of the reliability of the calculated binding energies. The results for two snapshots of phenol bound in the pocket is displayed in Table 4.5. For all our ONETEP calculations on this system we see very good convergence with errors less than 0.1 kcal mol^{-1}.

Snapshots were taken in two ways.

The first way aims to select a single snapshot in such a way that it will provide a binding energy in solvent that is as close as possible to the result obtained by the MM-PBSA or QM-PBSA calculation. To achieve this we assume that the median of the MM binding energy in vacuum agrees closely with the mean of these binding energies. This can be justified by considering that in the MM-PBSA approach the free energy of binding is obtained as the mean of the binding energies in the solvent. If we approximate the distribution of binding energies with a Gaussian, then the mean will be equal to the median. Fig 4.4 shows the interaction energy distribution for 1000 MM snapshots for 2-methylphenol in vacuum with a fitted Gaussian overlayed (R^2=0.98). Here the mean is 28.3 kcal mol^{-1} and the median is 28.4 kcal mol^{-1}, showing that this could be a good approach to use for reducing the number of snapshots required to obtain free energies. An additional requirement of the single snapshot approach, is that the energy distributions in vacuum correlates well with the energy distribution in solvent, so that the median chosen in vacuum remains also a very good approximation to the median in solvent. This would be ideally done using the full solvated binding energies. However, if we look at Figure 4.5b, which shows the QM solvated binding energies against the MM solvated binding energies, the R“2 value is 0.17. Showing no correlation.

Table 4.5: Total energies and binding energies from ONETEP for two snapshots of phenol bound in the cavity of T4 lysozyme L99A/M102Q, and the SCF convergences errors of the calculations. Energies in kcal mol^{-1}

Snapshot	Complex	Receptor	Ligand	Binding energy
1	-7359829.2 ± 0.07	-7325916.1 ± 0.07	-33881.9 ± 0.000007	-31.2
2	-7360291.3 ± 0.07	-7326381.5 ± 0.07	-33882.9 ± 0.000007	-26.9

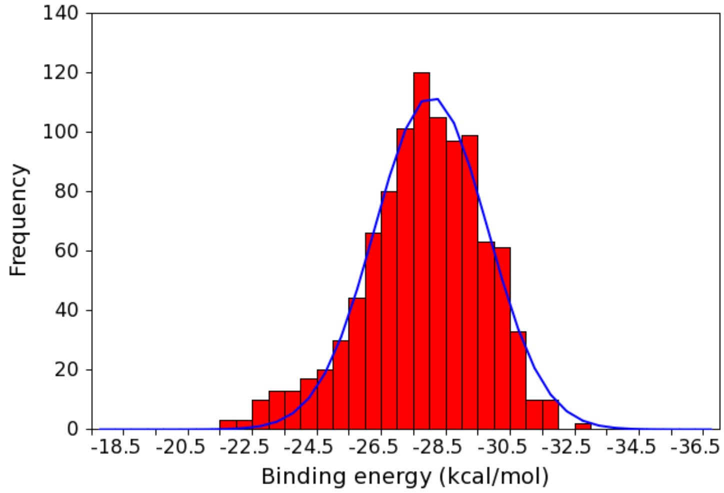


Figure 4.4: 2-methylphenol binding energy distribution from 1000 MM snapshots with a fitted Gaussian. Energy in kcal mol^{-1}.

Figure 4.5 shows that there is good correlation between $\Delta G_{\text{vac}}^{MM}$ and $\Delta G_{\text{vac}}^{QM}$ (Fig 4.5a), $\Delta G_{\text{vac}}^{MM}$ and $\Delta G_{\text{solv}}^{QM}$(Fig 4.5c), and $\Delta G_{\text{vac}}^{QM}$ and $\Delta G_{\text{solv}}^{QM}$ (Fig 4.5d), with R^2 values of 0.95, 0.67, and 0.73 respectively. This would suggest that if we select the MM median vacuum energy snapshot the QM energy of that snapshot in solvent will be close to the average QM binding energy in solvent, or in other words the QM-PBSA results that would be obtained by averaging many snapshots.

The second way involves taking snapshots at constant time intervals from the production trajectory. For the MM-PBSA the binding free energies were calculated with an increasing number of snapshots up to a total of 1000 snapshots from the 19ns production simulations. Figure 4.6 displays the convergence of the energies as more snapshots are included in the ensemble, taking the value at 1000 snapshots as the fully converged value (standard errors less than 0.08 kcal mol^{-1}). The maximum error observed using 5 snapshots is 1.15 kcal mol^{-1}, for the 2-methylphenol ligand. Using 50 snapshots the maximum error is reduced to less than 0.5 kcal mol^{-1} (with catechol having the largest error of 0.41 kcal mol^{-1}).

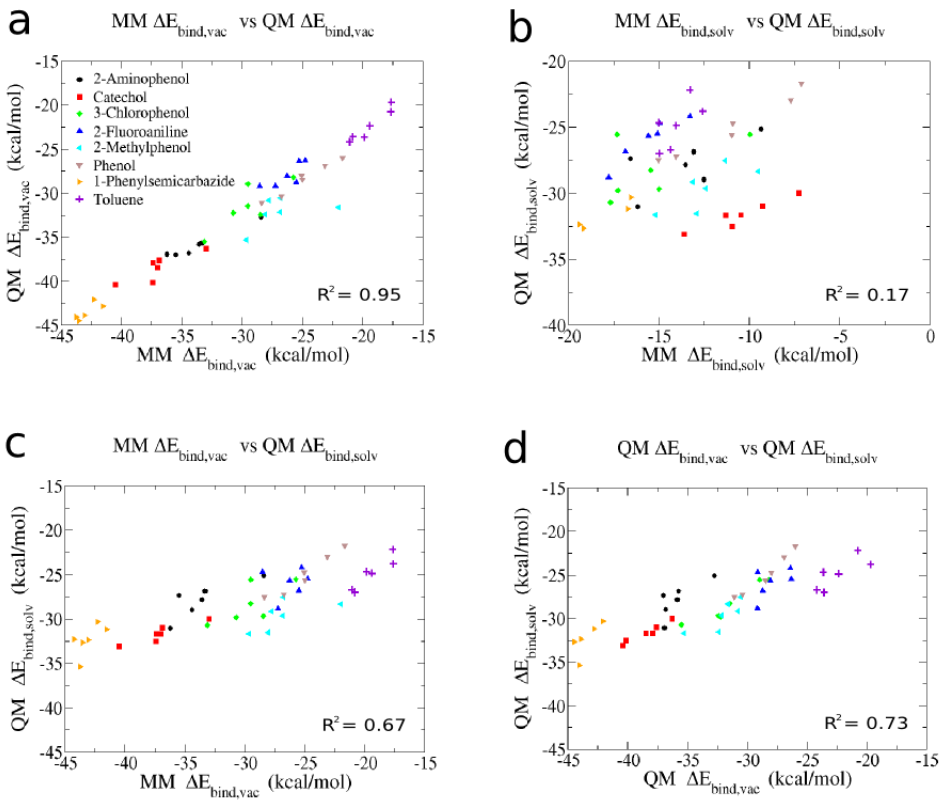


Figure 4.5: Correlation of MM and QM binding energies for the 8 ligands. a. MM vacuum binding energies against QM vacuum binding energies. b. MM solvated binding energies against QM solvated binding energies. c. MM vacuum binding energies against QM solvated binding energies. d. QM vacuum binding energies against QM solvated binding energies.

Assuming equally good convergence for the QM system, in this study we have chosen to use 5 snapshots to calculate the QM-PBSA energies. We can further examine the standard errors for the 5 snapshots for the MM and QM binding energies (both in solvent) to support the assumption of similar convergence rates. For the case of phenol the MM standard error over 5 snapshots is 1.4 kcal mol^{-1}, while the QM value is 1.0 kcal mol^{-1}. For catechol the MM error is 1.0 kcal mol^{-1} compared to a QM value of 0.5 kcal mol^{-1}, and for toluene the MM standard error is 0.4 kcal mol^{-1} compared to a QM value of 0.9 kcal mol^{-1}. MM and QM standard error values are similar to each other, suggesting that the QM

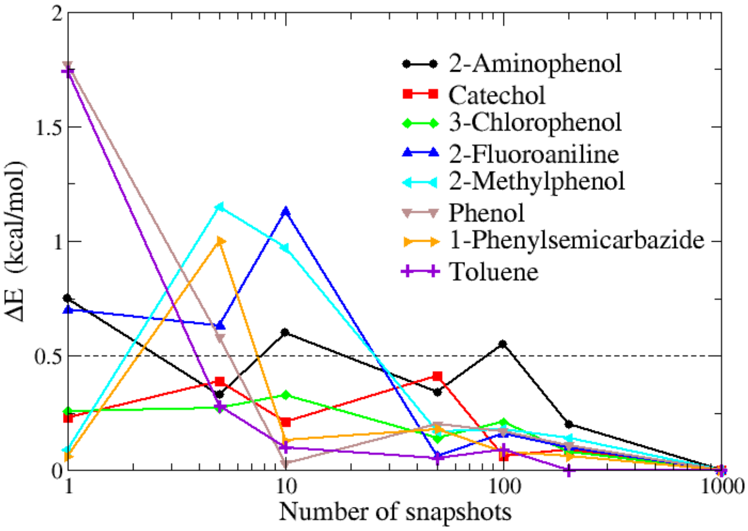


Figure 4.6: Absolute deviations of the binding energies of our ligands as a function of snapshots included in the MM-PBSA calculation, taking 1000 snapshots as the converged value.

Table 4.6: Components of binding energy for a single snapshot of phenol. Energies in kcal mol^{-1}.

QM	Binding energy		MM	Binding energy
Purely QM	97.62		Purely MM	0.0
Electrostatic terms	-107.57		Electrostatic terms	-10.1
Dispersion terms	-23.80		Dispersion terms	-19.1
Total binding energy	-33.8		Total binding energy	-29.2

binding energies would converge at a similar rate as MM binding energies shown in Figure 4.6.

To better understand the contributions to the binding energies, and the differences seen between QM and MM, the components of the binding energies in vacuum for a single snapshot of phenol bound to the protein are presented in Table 4.6. From this we can see further evidence that the binding energies in vacuum agree very well between the QM and the highly parametrised force field (R^2=0.95 for Figure 4.5a).

Another issue for the T4 lysozyme L99A/M102Q is that of Val111 in the binding pocket, which is known to have two rotamers with a χ_1 angle of $\sim$ 180° and $\sim$ $-60°$, shown in Fig 4.7. Studies have shown that using the wrong rotamer can have an effect around 4 kcal/mol on the calculated binding free energies [120, 121]. Explicit modelling of this has been shown to improve the agreement with experimental binding free energies [121]. The starting structure for all our MD simulations is the catechol bound crystal structure. Taking this fact into consideration requires two median snapshots in total, which correspond to the two rotamer distributions. The obtained binding energy in solution is the weighted sum of these two snapshots. The weights are computed as the fraction of time the simulation spent on each rotamer.

4.2.2 Results and Discussion

We computed the binding free energy of the 8 ligands shown in Table 4.4 to the T4 lysozyme double mutant L99A/M102Q using MM-PBSA and QM-PBSA. These ligands were chosen as they comprise of a variety of chemical and physical properties (polarity, inclusion of halides, size and binder/non-binder). Tables 4.7 and 4.8 present the computed binding free energies of these ligands, relative to phenol, for the case of the median energy snapshots and for the mean of five snapshots respectively. Since the cavity of the apo-protein will always contain a single water (bound to Gln102) [98], we also calculated the binding free energy using only the ligand solvation energy added to the binding energy in vacuum. The approximation made in this case, is that the complex and receptor solvation energies should be very similar and assumed to cancel each other, so only the desolvation of the ligand would make a significant contribution to the free energy of binding. Figures

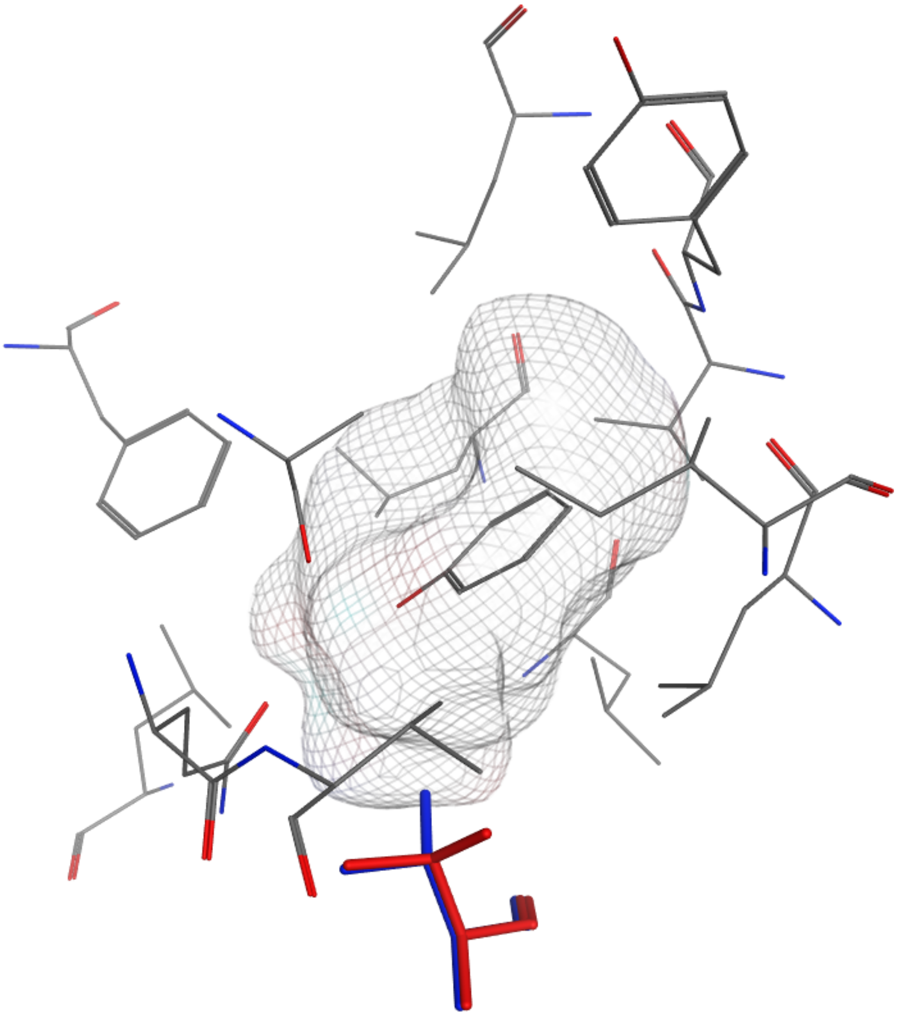


Figure 4.7: Molecular surface depiction of the binding pocket and the the rotamers of Val111. Red: -60°. Blue: -180°.

4.8c and 4.8d show that this approximation works quite well in this system, as the binding solvation energies are very similar to the negative of the ligand solvation energies. This approximation works better for the QM solvation energies than the MM solvation energies, as the complex and host solvation energies happen to be much closer when calculated by QM. If we compare the smallest difference between Figure 4.8c and 4.8d for the QM energies, which is seen to be phenol, the QM solvation energies of the complex and host only differ by 0.1 kcal mol^{-1}, whereas the MM solvation energies differ by 4.2 kcal mol^{-1}. The largest difference observed between Figure 4.8c and 4.8d is for the 1-phenylsemicarbazide ligand. In this case the complex and host solvation energies differ by 2.1 kcal mol^{-1} in the QM calculations, and 7.9 kcal mol^{-1} in the MM calculations. The other ligands have differences ranging between 0.2 to 1.2 kcal mol^{-1} for the QM calculations and 4.0 to 5.9 kcal mol^{-1} for the MM calculations, and are shown in Table 4.9. This observation can be explained by considering that in the QM calculation the dielectric is density dependent, so the pocket will have a smaller dielectric than the bulk value due to residual density of the protein in the cavity. This is in contrast to the MM calculation which will have the full solvent dielectric in the pocket. This will make a significant contribution to the solvation energy of the host leading to the larger differences seen in MM. The largest differences in both MM and QM is observed for the 1-phenylsemicarbazide ligand. It is much larger than the others ligands and causes the pocket to enlarge to accommodate it, and hence has a larger solvation contribution to the host.

When relative binding free energies are calculated an approximation that can be used, if ligands are of a similar size, is that the changes in entropy of binding between different ligands are comparable and will cancel. To investigate the effect of this approximation for this system, we have also calculated the entropy using normal mode analysis in AMBER and the results with and without entropy are shown. In Table 4.7 the entropy is calculated from the median energy snapshots and weighted in the same way, in Table 4.8 the entropy is calculated as the average from 50 snapshots, instead of 5 snapshots, taken at constant time intervals from the trajectories to improve the convergence of the entropy. This was done to improve convergence for the computed entropies. The standard error using the 5 snapshots to calculate the entropy (TΔS) for phenol bound is 0.94 kcal mol^{-1}, with an average TΔS of -14.64 kcal mol^{-1}. Using different sets of 5 snapshots, the average entropies range from -13.4 kcal mol^{-1} to -15.2 kcal mol^{-1} with standard errors of up to 1.9 kcal mol^{-1}. Using 50 snapshots the average entropy (TΔS) is -14.22 kcal mol^{-1} with a standard error of 0.38 kcal mol^{-1}.

Figures 4.8a, 4.8b and 4.8c display the QM and MM binding energy in vacuum, solvent and solvation energies for the eight ligands. We observe very good correlation of the binding energies in vacuum as can be seen in 4.5a (and 4.8a). This is to be expected since the ff99SB force field is well parametrised for proteins, but there is hardly any correlation when comparing the solvated binding energies as shown in 4.5b (and 4.8b). In figure 4.8c we see just the binding solvation energies (the difference between $\Delta G_{\mathrm{bind,solv}}$ and $\Delta G_{\mathrm{bind,vac}}$). The solvation energy in MM-PBSA is a combination of the polar term, G_{polar} from the PB equation and the non-polar term, $G_{\mathrm{non-polar}}$ from the solvent accessible surface area (SASA) calculation which are added to the vacuum energy. In the QM solvation calculation the reaction field and the cavitation energy are explicitly included into the Hamiltonian and affect the density, and hence the final ground state energy. Thus the polar and non-polar parts of solvation between the two methods are very different. For example, for catechol, the polar QM part is 15.5 kcal mol^{-1} while the polar MM part is 28.8 kcal mol^{-1}. The non-polar QM is -9.3 kcal mol^{-1} while for MM it is -2.7 kcal mol^{-1}. Since these partitions are so different, only the total solvation energy should be compared. The experimentally obtained hydration energies of catechol, phenol and toluene are -9.4 kcal mol^{-1}, -6.6 kcal mol^{-1} and -0.8 kcal mol^{-1} respectively. Hydration energies obtained from QM-PBSA averaged over our five snapshots are -8.0 kcal mol^{-1}, -3.7 kcal mol^{-1} and 1.5 kcal mol^{-1} in contrast to MM-PBSA which gives -20.6 kcal mol^{-1}, -9.9 kcal mol^{-1} and -1.4 kcal mol^{-1}. These hydration energies are shown in Table 4.10 for easy comparison. We can clearly see that the MM-PBSA hydration energies are less accurate and this is expected to impact the quality of the free energy calculations. The QM-PBSA energies on the other hand have a smaller error and the relative hydration energies are substantially closer to experimentally obtained values (errors less than 1.5 kcal mol^{-1} for QM compared to 7.9 kcal mol^{-1} for MM).

To test the robustness of the median snapshot approach, the snapshots either side of the median energy snapshot (eg. left and right of the median in Figure 4.4) were taken to calculate the QM binding free energy. The three MM binding energies in vacuum differ by less than 0.1 kcal mol^{-1}, however, there is a difference of 1.2 kcal mol^{-1} in the QM binding energies. In solution the binding energies from MM-PBSA differ by 2.7 kcal mol^{-1}, whereas from QM-PBSA they differ by 1.0 kcal mol^{-1}. This suggests that using just one snapshot (the median) is not likely to produce converged binding energies. The median snapshot binding energies are presented in Table 4.7.

Looking at the binding free energies averaged over five snapshots in Table 4.8, the relative binding free energies from MM-PBSA are not very close to the experimental values. Catechol is the exception and has a predicted relative binding free energy with an error of 0.7 kcal mol^{-1} compared to the experimental value. The smallest error for the other ligands is 2.3 kcal mol^{-1}, which is for 2-methylphenol. The computed ligand rankings from MM-PBSA is, however, very similar to experimental ranking, with an R^2 of 0.93 for the ligand binders. The binding free energies from QM-PBSA are slightly improved, with two ligands having errors less than 0.8 kcal mol^{-1} from experiment (2-fluoroaniline and toluene). Catechol however, in contrast to MM-PBSA, has the largest error of the known binders of 7.8 kcal mol^{-1}. The overall result is a worse trend for QM-PBSA compared to experiment than for MM-PBSA, with an R^2 of 0.41. Given that the vacuum binding energies from MM and QM correlate very well, the difference observed in the binding energies in solvent is due to the solvation energies. For a simpler comparison, we will look at only the desolvation energy of the ligands, using the approximation that complex and host solvation energy cancels, when comparing relative binding energies. We will use catechol as an example, since the MM-PBSA relative binding free energy is very close to the experimental value (a difference of 0.7 kcal/mol) and the QM-PBSA value is not (with a difference of 7.8 kcal mol^{-1}). The experimental hydration energy of catechol is -9.4 kcal mol^{-1}. The QM hydration energy averaged over the five snapshots is -8.0 kcal mol^{-1}, in contrast to the MM hydration energy averaged over the same five snapshots which is -20.6 kcal mol^{-1}. This shows that the MM solvation energy is substantially overestimated, however, MM-PBSA still produces a very good relative binding free energy compared to experiment. Since MM and QM binding energies in vacuum are so close to each other, this suggests that both MM and QM overbind the catechol in the pocket in vacuum, however the excessively large solvation energy from MM-PBSA cancels out the overbinding in vacuum to give a final relative free energy of binding that does agree closely with experiment.

As we have seen, the MM-PBSA solvation energies seem to be inconsistent for the different ligands. Not every ligand appears to have such a ideal error cancellation between an overestimated solvation energy and the overbinding in vacuum that catechol appears to have to produce such a good relative binding free energy. The more accurate QM solvation energy does not lead to the same error cancellation, resulting in worse relative QM-PBSA binding free energy for catechol, but does see improvements for some of the other ligands. Thus, it appears that the error in the QM results may be mainly due to the inherent approximations in DFT, such as the exchange-correlation functional chosen.

Both methods predict the experimental non-binders as good binders. 1-phenylsemicarbazide has a very strong binding energy in vacuum, this value is reduced when including the desolvation energy of the ligand, which is higher for this ligand than the others, but remains the strongest predicted binder. Due to the larger size of 1-phenylsemicarbazide, when it is placed in the pocket it forces the pocket to expand so that it can fit. This causes around a 4.1 kcal mol^{-1}difference in the calculated binding entropies compared to phenol. The approximation of entropy cancellation in this case is not valid. All other ligands have entropies of binding much closer to phenol, between 0.4 kcal mol^{-1} and 1.6 kcal mol^{-1}, with the largest value being for 2-methylphenol. Even though this is quite a small difference, we observe a small improvement in agreement with experiment when entropy is included in the calculation of the relative binding free energies averaged over 5 snapshots.

Since MM-PBSA appears to produce improved rankings for the binding free energies as a result of error cancellation due to the overestimation of the solvation energy, the QM binding free energies have been calculated scaling the MM solvation energy as previously proposed by Cole *et al* [113] (Equation 4.1). The

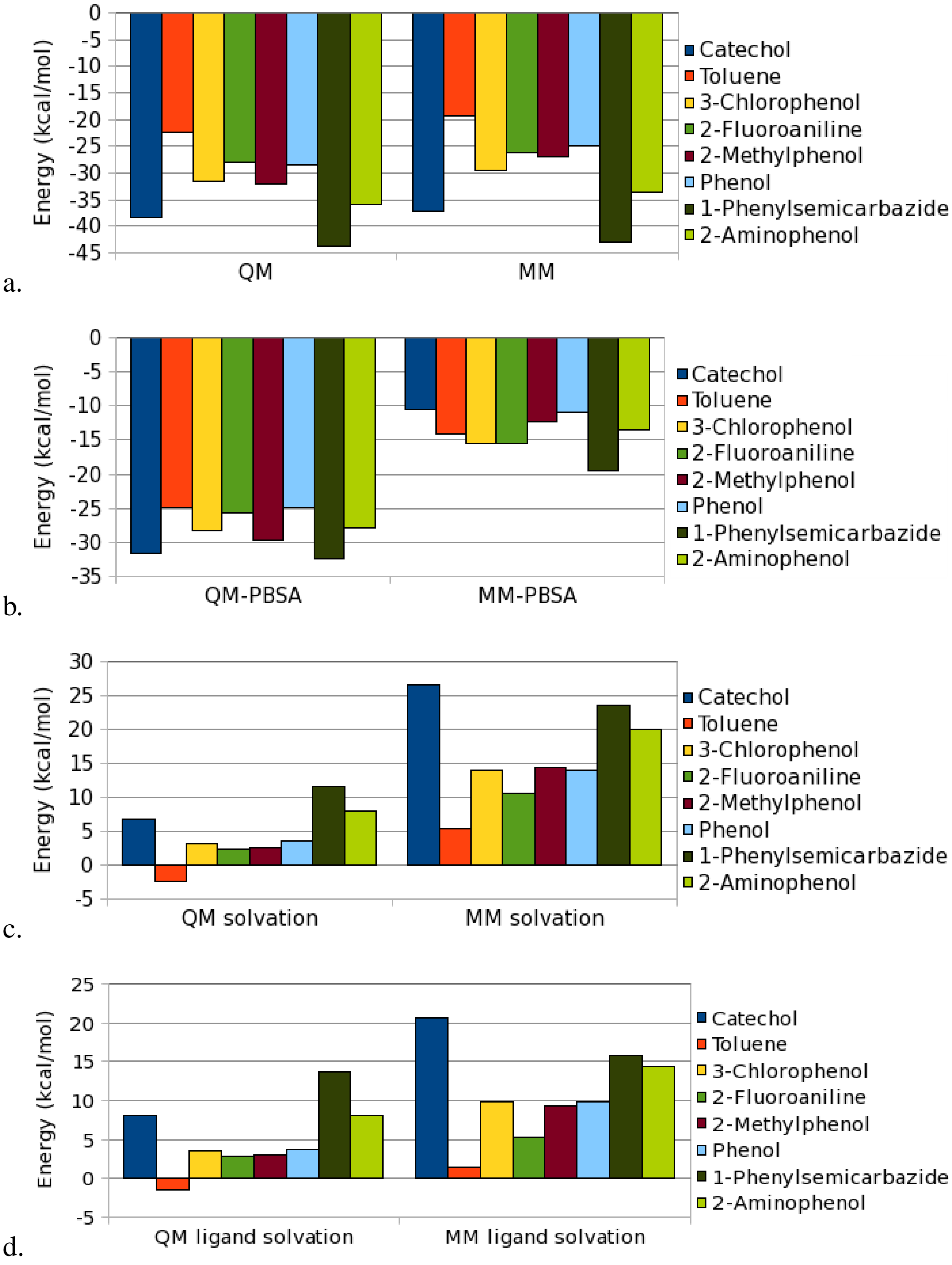


Figure 4.8: Energies averaged over 5 snapshots for all eight ligands compared between QM and MM calculations. a. Binding energies in vacuum. b. Binding energies in solvent. c. Binding solvation energies. d. The negative of the ligand solvation energies.

results are shown in Table 4.11. The largest difference between the scaled solvation energy approach and the full QM solvation energy approach is observed for catechol bound in the cavity. The MM-PBSA relative binding free energy is very close to the experimental value and the MM and QM vacuum binding energies so similar, that when the MM solvation energy is scaled and added to the QM vacuum binding energy, the QM-PBSA energy is much closer to the experimental value: With an error of 0.2 kcal mol^{-1} off instead of 7.8 kcal mol^{-1}. The other ligands are effected very little, except for toluene, whose binding free energy is made much worse. This is due to the scaling method used in Equation 4.1. For toluene, ΔE_{DFT} is a positive value, while ΔE_{EL} is negative, so the resulting QM solvation energy is a large negative number. This results in a relative binding free energy that is too strong in favour of toluene. Due to the form of Equation 4.1, this method would only work when the binding energy from DFT, and the electrostatic part of the binding energy from MM, are less than -1.0 kcal mol^{-1}. If this is not the case the scaled solvation energy will be meaningless. Using this approach to calculate the QM solvation energy is not a reliable approach, it is much more preferable to use a full QM solvation approach, such as the method implemented in ONETEP, since this result is more “correct”.

4.2.3 Conclusions

We have presented a QM-PBSA approach in which large-scale DFT calculations with a near-complete basis set were performed to evaluate the energy of the configurations in place of the force field that is used in the conventional MM-PBSA technique. The solvent in the DFT calculations was described by a minimal parameter self-consistent implicit solvent model. We applied the QM-PBSA approach to compute the relative binding free energies of eight small aromatic ligands bound in the polar cavity of the T4 lysozyme mutant L99A/M102Q protein which contains more than 2600 atoms, and have compared our results to the traditional MM-PBSA method.

Due to the high cost of the quantum calculations, and the limited computer resources we currently have, the ensemble of structures we used here is not large enough to obtain converged results, but the trends we see when comparing approaches are converged.

The MM and QM binding energies correlate very well for all ligands in vacuum, and the relative binding free energies obtained by the two methods in vacuum are very similar. However, there is very little correlation between binding energies in solvent. The QM ligand hydration energies have systematic errors of about 1.5 kcal mol^{-1} and relative errors of less than 1.5 kcal mol^{-1} with respect to experimentally measured values while the MM hydration energies for the ligands show inconsistencies and errors of up to -11 kcal mol^{-1}. Nevertheless, for this system, the MM-PBSA calculations replicate the experimental trend in binding affinities better than our QM-PBSA approach. This appears to be due to fortuitous error cancellation between the vacuum and solvent energies in the MM model, something that is not observed for the more accurate QM solvation energies. An alternative, earlier approach, which involves obtaining the solvation energy of the QM calculation by the scaled MM solvation energy tends to imitate the MM-PBSA results for the polar ligands but fails spectacularly for the non-polar ones such as toluene.

Thus, while the QM-PBSA approach is more rigorous than MM-PBSA, in the sense that interaction energies are obtained from a calculation that explicitly includes electronic polarisation, in this case they appear to over-estimate the binding energies in vacuum which results in errors in the free energies of binding in solution. This limitation could be overcome in future studies by using a more accurate exchange correlation functional such as for example hybrid functionals and/or a functional which explicitly includes dispersion interactions.

There are many sources of error in the MM-PBSA approach. These include the sampling of the phase space which can often be incomplete, the use of the implicit solvation model, the approximation of entropy, as well as those inherent in the force field. The last of these is reduced for the calculated interaction energies when it is evaluated using a QM potential, however the systematic errors of the force field in the structure generation are still present. The calculation of entropy using normal mode analysis with the force field is approximate at best, and assuming cancellation of entropy when considering relative binding free energies, and hence neglecting it, could well provide less error when observing small perturbations. For this system most of the experimental binding free energies are within k_BT (0.6 kcal mol^{-1}) of each other, which makes them very hard to distinguish between. When combining all these errors it is impressive that for this system the approach can obtain reasonable results, both with MM and QM.

The types of ligands and protein considered here were common enough to be well-described by the force field so the MM-PBSA approach performs very well. This QM-PBSA method would be expected to perform better than MM-PBSA approaches on systems which force fields would not describe very well (e.g. ligands with unconventional functional groups). As well as more accurate energies, there are also other advantages of large-scale ab initio quantum chemistry calculations that have not been explored in this work, such as the ability to visualise localised orbitals, densities and potentials that are responsible for specific interactions and the quantitative estimation of these interactions with energy decomposition approaches.

Table 4.7: QM-PBSA and MM-PBSA binding free energies for median energy snapshots relative to phenol. $\Delta G^{QM}_{\text{bind,vac}}$ is the binding energy in vacuum, $\Delta G^{QM}_{\text{bind,solv}}$ is the binding energy in solvent, $\Delta G^{QM}_{\text{lig,solv}}$ is the solvation energy of the ligand, and ΔS is the vibrational entropy. All energies in kcal mol^{-1}.

			Weighted sum of median energies			
Ligand	$\Delta G^{QM}_{\text{bind,vac}}$	$\Delta G^{QM}_{\text{bind,solv}}$	$\Delta G^{QM}_{\text{bind,vac}} - \Delta G^{QM}_{\text{lig,solv}}$	$\Delta G^{QM}_{\text{bind,solv}}$ - $T\Delta S$	$\Delta G^{QM}_{\text{bind,vac}} - \Delta G^{QM}_{\text{lig,solv}}$ - $T\Delta S$	ΔG^{exp}
Catechol	-12.6	-9.0	-7.7	-6.6	-5.3	-4.4
Toluene	2.4	-3.3	-2.6	-1.5	-0.8	-5.2
2-fluoroaniline	-1.4	-1.1	-1.9	-1.2	-2.0	-5.5
3-chlorophenol	-6.7	-6.4	-6.4	-2.0	-2.0	-5.8
2-methylphenol	-6.6	-6.7	-7.4	-0.5	-1.2	-4.7
2-aminophenol	-8.1	-7.3	-3.9	-5.6	-2.3	non-binder
1-phenylsemicarbazide	-18.1	-7.8	-6.3	-3.5	-3.7	non-binder
Phenol	-5.6	-5.6	-5.6	-5.6	-5.6	-5.6
Max error	12.6	4.6	3.6	4.3	4.5	
rms error	6.7	2.8	2.4	3.3	3.3	

			Weighted sum of median energies			
Ligand	$\Delta G^{MM}_{\text{bind,vac}}$	$\Delta G^{MM}_{\text{bind,solv}}$	$\Delta G^{MM}_{\text{bind,vac}} - \Delta G^{MM}_{\text{lig,solv}}$	$\Delta G^{MM}_{\text{bind,solv}}$ - $T\Delta S$	$\Delta G^{MM}_{\text{bind,vac}} - \Delta G^{MM}_{\text{lig,solv}}$ - $T\Delta S$	ΔG^{exp}
Catechol	-14.5	-2.3	-3.7	-0.5	-1.6	-4.4
Toluene	1.3	-4.3	-6.6	-3.3	-5.6	-5.2
2-fluoroaniline	-5.3	-7.9	-8.8	-7.2	-10.0	-5.5
3-chlorophenol	-8.9	-7.2	-8.5	-2.4	-3.7	-5.8
2-methylphenol	-7.6	-5.7	-7.6	-0.2	-2.1	-4.7
2-aminophenol	-12.4	-6.2	-8.9	-5.2	-7.9	non-binder
1-phenylsemicarbazide	-22.2	-10.6	-7.1	-6.4	-5.5	non-binder
Phenol	-5.6	-5.6	-5.6	-5.6	-5.6	-5.6
Max error	16.8	4.6	3.3	5.3	4.5	
rms error	8.5	2.2	2.5	3.4	2.6	

Table 4.8: QM-PBSA and MM-PBSA binding free energies for five snapshots relative to phenol. $\Delta G^{QM}_{\text{bind,vac}}$ is the binding energy in vacuum, $\Delta G^{QM}_{\text{bind,solv}}$ is the binding energy in solvent, $\Delta G^{QM}_{\text{lig,solv}}$ is the solvation energy of the ligand, and ΔS is the vibrational entropy. $\Delta G^{MM}_{\text{bind,solv}}$(1000) is the MM-PBSA value averaged over 1000 snapshots.All energies in kcal mol^{-1}.

Ligand	$\Delta G^{QM}_{\text{bind,vac}}$	$\Delta G^{QM}_{\text{bind,solv}}$	$\Delta G^{QM}_{\text{bind,vac}} - \Delta G^{QM}_{\text{lig,solv}}$	$\Delta G^{QM}_{\text{bind,solv}}$ - $T\Delta S$	$\Delta G^{QM}_{\text{bind,vac}} - \Delta G^{QM}_{\text{lig,solv}}$ - $T\Delta S$	ΔG^{exp}
Catechol	-15.4	-12.2	-11.1	-11.8	-10.7	-4.4
Toluene	0.2	-5.9	-5.0	-5.6	-4.7	-5.2
2-fluoroaniline	-4.6	-6.3	-5.4	-5.1	-4.3	-5.5
3-chlorophenol	-8.7	-9.1	-8.8	-9.5	-9.2	-5.8
2-methylphenol	-8.9	-10.1	-9.6	-8.5	-8.1	-4.7
2-aminophenol	-12.6	-8.2	-8.2	-7.2	-7.1	non-binder
1-phenylsemicarbazide	-20.1	-12.2	-10.1	-8.1	-6.1	non-binder
Phenol	-5.6	-5.6	-5.6	-5.6	-5.6	-5.6
Max error	14.5	7.8	6.7	7.4	6.3	
rms error	7.9	4.7	3.9	3.6	3.1	

Ligand	$\Delta G^{MM}_{\text{bind,vac}}$	$\Delta G^{MM}_{\text{bind,solv}}$	$\Delta G^{MM}_{\text{bind,vac}} - \Delta G^{MM}_{\text{lig,solv}}$	$\Delta G^{MM}_{\text{bind,solv}}$ - $T\Delta S$	$\Delta G^{MM}_{\text{bind,vac}} - \Delta G^{MM}_{\text{lig,solv}}$ - $T\Delta S$	ΔG^{exp}
Catechol	-17.6	-5.1	-6.8	-4.7	-6.4	-4.4
Toluene	0.0	-8.6	-8.4	-8.3	-8.0	-5.2
2-fluoroaniline	-6.8	-10.2	-11.4	-9.0	-10.2	-5.5
3-chlorophenol	-10.1	-10.0	-10.2	-9.3	-9.4	-5.8
2-methylphenol	-7.4	-7.0	-8.1	-5.4	-6.5	-4.7
2-aminophenol	-14.1	-8.1	-9.6	-7.1	-8.6	non-binder
1-phenylsemicarbazide	-23.6	-14.0	-17.7	-8.0	-11.7	non-binder
Phenol	-5.6	-5.6	-5.6	-5.6	-5.6	-5.6
Max error	18.1	8.4	12.2	4.6	8.1	
rms error	9.5	4.4	5.9	2.9	4.3	

Table 4.9: QM and MM complex and host solvation energies for the phenol ligand averaged over 5 snapshots. Energies in kcal mol^{-1}.

	QM			MM		
Ligand	$G_{\mathrm{solv,com}}$	$G_{\mathrm{solv,host}}$	$\lvert\Delta G_{solv}\rvert$	$G_{\mathrm{solv,com}}$	$G_{\mathrm{solv,host}}$	$\lvert\Delta G_{solv}\rvert$
Catechol	-2416.0	-2414.8	1.2	-2409.0	-2414.9	5.9
Toluene	-2475.8	-2474.8	1.0	-2493.8	-2497.8	4.0
2-fluoroaniline	-2463.7	-2462.8	1.0	-2431.7	-2437.0	5.4
3-chlorophenol	-2476.6	-2476.2	0.4	-2512.2	-2516.5	4.3
2-methylphenol	-2450.8	-2450.3	0.5	-2464.6	-2469.9	5.3
2-aminophenol	-2448.7	-2448.5	0.2	-2446.9	-2452.6	5.7
1-phenylsemicarbazide	-2451.5	-2449.4	2.1	-2454.5	-2462.5	7.9
Phenol	-2429.9	-2429.8	0.1	-2442.5	-2446.7	4.2

Table 4.10: Hydration energies for catechol, phenol and toluene from experiment, QM-PBSA and MM-PBSA in kcal/mol. Hydration energies averaged over the 5 chosen snapshots.

Molecule	$\Delta G^{\text{exp}}_{lig,solv}$	$\Delta G^{\text{QM}}_{lig,solv}$	$\Delta G^{\text{MM}}_{lig,solv}$
Catechol	-9.4	-8.0	-20.6
Phenol	-6.6	-3.7	-9.9
Toluene	-0.8	1.5	-1.4
Relative hydration energies			
Catechol - phenol	-2.8	-4.3	-10.7
Toluene - phenol	5.7	5.2	8.5

Table 4.11: QM binding free energies with the solvation energy calculated via Equation 4.1. Energies in kcal mol^{-1}.

Ligand	$\Delta G^{QM}_{\text{bind,vac}}$	$\Delta G^{QM}_{\text{bind,solv}}$	$\Delta G^{QM}_{\text{bind,solv}}$ - $T\Delta S$	ΔG_{exp}
Catechol	-15.6	-4.6	-4.2	-4.4
Toluene	0.5	-17.6	-17.2	-5.2
2-fluoroaniline	-5.0	-6.7	-5.6	-5.5
3-chlorophenol	-8.6	-8.7	-9.1	-5.8
2-methylphenol	-9.2	-9.6	-8.0	-4.7
2-aminophenol	-12.9	-9.9	-8.9	0.0
1-phenylsemicarbazide	-21.0	-14.9	-10.8	0.0
Phenol	-5.6	-5.6	-5.6	-5.6

Chapter 5

QM corrected Thermodynamic Integration

The work presented in this chapter has been done in collaboration with Chris Pittock.

An approach for accurately estimating relative binding free energies using a fast Hamiltonian, and then using efficient sampling approaches to approximate the difference between a QM/MM Hamiltonian and the fast Hamiltonian, was first proposed by Warshel and and co-workers in 2002 [122]. This approach was proposed as a method to overcome the limitations of force fields to describe polarisation and charge transfer, and implemented for the most theoretically rigorous binding free energy approach available, thermodynamic integration (TI). This extension is presented in Figure 5.1.

This method first uses a reference Hamiltonian (often a pure MM force field) to estimate the free energy. This estimate is then corrected by calculating the free energy necessary to change the reference Hamiltonian to the QM/MM Hamiltonian.

$$\begin{tikzcd}
A'_{QM} & A'_{MM} \arrow[l, "\overset{\text{FEP}}{\Delta G_1}"'] \arrow[r, "\overset{\text{TI}}{\Delta G_2}"] & B'_{MM} \arrow[r, "\overset{\text{FEP}}{\Delta G_3}"] & B'_{QM} \\
A_{QM} \arrow[u, "\Delta G_8"] & A_{MM} \arrow[u] \arrow[l, "\Delta G_7"] \arrow[r, "\Delta G_6"'] & B_{MM} \arrow[u] \arrow[r, "\Delta G_5"'] & B_{QM} \arrow[u, "\Delta G_4"']
\end{tikzcd}$$

Figure 5.1: Extended free energy cycle. Using Thermodynamics to mutate from A to B (A′ or B′) and free energy perturbation to change from the MM description of the system to the QM description.

Theoretically this method can calculate the exact QM/MM free energy change, however, since the computational cost of sampling phase space using a QM potential is still very high, sampling is done with only the cheaper and faster reference potential. The free energy change from the MM to QM/MM Hamiltonian is then done using only the structural ensemble generated from the reference Hamiltonian via a single step free energy perturbation. This approach will converge quickly if the energies from the MM and QM/MM have good overlap. However, in practice, there can be large fluctuations between the energies of the reference Hamiltonian and the QM/MM Hamiltonian, which leads to poor convergence of the free energy.

Woods *et al* [123] used two approaches to calculate the MM to QM energy. The first involved using the same approach as Warshel with further development aimed at creating approximate Hamiltonians that were a good match to the target QM/MM Hamiltonian to improve the overlap. The second method involved using an approximate Hamiltonian to speed up the sampling of phase space described by a QM/MM Hamiltonian. Using this second method the ensembles produced will be correct for the QM/MM Hamiltonian used, and the ensemble can be used directly with FEP.

To solve the issue of the large differences in the magnitudes between the MM and QM/MM energies, Beierlein *et al* [124] proposed to correct only the solute/solvent Coulomb interaction energy differences. In this way the polarisation of the solute (ligand) would be accounted for in the different states A and B whilst sampling would be done using a relatively cheap MM method. The Zwanzig equation is then written as,

$$\Delta G_{\mathrm{MM}\rightarrow\mathrm{QM/MM}} = -k_B T \ln\langle e^{-\frac{U_{QM/MM}-U_{QM,vac}-U_{charges,MM}-U_{Coul,solute-solv,MM}}{k_B T}}\rangle_{\mathrm{MM}} \tag{5.1}$$

where $U_{QM/MM}$ is the energy of the QM/MM system, $U_{QM,vac}$ is the energy of just the QM region, $U_{QM/MM}$ is the energy of the MM region in the QM/MM system, and $U_{Coul,solute-solv,MM}$ is the electrostatic part of the all MM system. The aim of their work was to obtain reproducible, converged free energies, without the need for closely coupled MM and QM programs. Beierlein *et al* [124] demonstrated, via a series of careful tests, that free energies obtained from a classical forcefield (MM) can be converted to free energies that would have been obtained if a quantum description was used for the solute, and a classical description for the surrounding atoms.

We propose that instead of total energies, or just correcting the Coulomb energies as in the approach by Beierlein, we use the complete interaction energies, defined as,

$$\Delta E_{AB} = E_{AB} - E_A - E_B. \tag{5.2}$$

Where, for example, A is a solvent, B is a ligand, and AB is the protein-ligand complex. The MM to QM energy will now be calculated using,

$$\Delta G_{\mathrm{MM}\rightarrow\mathrm{QM}} = -k_B T \ln\langle e^{-\frac{\Delta E^{\mathrm{QM}}-\Delta E^{\mathrm{MM}}}{k_B T}}\rangle_{\mathrm{MM}}, \tag{5.3}$$

where ΔE^{QM} is the interaction energy in the quantum description with $E^{\mathrm{QM}} = E^{DFT} + E^{disp}$ and so including all interaction energies. ΔE^{MM} is the interaction energy in the force field description. The notation $\langle\cdots\rangle_{MM}$ signifies an ensemble average over the structures obtained from the MD simulation with the MM force field. To save on computational time we will employ a QM/MM approach to do this. As long as no covalent bonds are involved in the QM/MM boundary, interactions between the quantum and classical parts can be described by non-bonded terms only. This approach is commonly referred to as "electrostatic embedding".

5.1 Electrostatic embedding in ONETEP

The energy of the entire embedded system is composed of the following terms,

$$E_{\mathrm{QM/q}} = E_{\mathrm{QM}} + E_{\mathrm{int}} + E_{\mathrm{q}}, \tag{5.4}$$

where E_{QM} is the electronic energy of the quantum system (with its density/wavef-

unctions polarised by the potential due to the embedding charges), E_{int} is the energy of interaction of the electrons and nuclei of the quantum system with the embedding charges, and E_{q} is the electrostatic energy of the embedding charges.

The interaction of the QM with the MM, in atomic units, is,

$$E_{\mathrm{int}} = \sum_{J=1}^{N_{\mathrm{at}}} \sum_{a=1}^{N_{\mathrm{emb}}} Z_J \int \frac{q_a(\mathbf{r} - \mathbf{R_a})}{|\mathbf{r} - \mathbf{R}_J|} d\mathbf{r} - \sum_{a=1}^{N_{\mathrm{emb}}} \int \int \frac{q_a(\mathbf{r} - \mathbf{R}_a)\, n(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d\mathbf{r}' d\mathbf{r}, \tag{5.5}$$

where the first term on the righthand side is the Coulomb (i.e. electrostatic) energy of interaction between N_{at} nuclei (with atomic number Z_J) and the $\mathbf{N}_{\mathrm{emb}}$ embedding charges q_a, and the second term is the Coulomb energy of interaction between the electronic density $n(\mathbf{r})$ and the embedding charges. We also have,

$$E_{\mathrm{q}} = \sum_{a=1}^{N_{\mathrm{emb}}} \sum_{b>a}^{N_{\mathrm{emb}}} \int\int \frac{q_a(\mathbf{r}-\mathbf{R}_a)\, q_b(\mathbf{r}'-\mathbf{R}_b)}{|\mathbf{r}-\mathbf{r}'|} d\mathbf{r}' d\mathbf{r}, \tag{5.6}$$

which is the energy of interaction between the embedding charges.

In ONETEP the external potential due to the ionic cores, which are represented by norm-conserving pseudopotentials, in the Kleinman and Bylander representation has local and non-local parts:

$$\hat{V}_{\mathrm{ext}}(\mathbf{r}) = \hat{V}_{\mathrm{ext,loc}}(\mathbf{r}) + \hat{V}_{\mathrm{ext,nl}}(\mathbf{r}) = \sum_{p=1}^{N_{at}} \left[\hat{v}^{p}_{\mathrm{ps,loc}}(\mathbf{r}-\mathbf{R}_p) + \hat{v}_{\mathrm{ps,nl}}(\mathbf{r}-\mathbf{R}_p)\right], \tag{5.7}$$

where N_{at} is the total number of atoms and $\mathbf{R}_p$ is the position of atom p. The potential due to the embedding charges is of similar form to the local part of the external potential,

$$\hat{V}_{\mathrm{emb}} = \sum_{a=1}^{N_{\mathrm{emb}}} \hat{v}^{(a)}_{\mathrm{emb}}(\mathbf{r}-\mathbf{R}_a). \tag{5.8}$$

We therefore generate directly the sum of the two [4], which can be written in the following form,

$$\begin{aligned}\hat{V}_{\mathrm{ext,loc}}(\mathbf{r}) + \hat{V}_{\mathrm{emb}}(\mathbf{r}) = &\sum_{k=1}^{N_{\mathrm{species}}} \sum_{I=1}^{N_k} \hat{v}^{(k)}_{\mathrm{ps,loc}}(\mathbf{r}-\mathbf{R}_{k,I}) \\ &+ \sum_{j=1}^{N_{\mathrm{emb-species}}} \sum_{L=1}^{N_j} \hat{v}^{(j)}_{\mathrm{emb}}(\mathbf{r}-\mathbf{R}_{j,L}),\end{aligned} \tag{5.9}$$

where $\hat{v}^{(k)}_{\mathrm{ps,loc}}(\mathbf{r}-\mathbf{R}_{k,I})$ is the local pseudopotential for a particular “species” of atomic core (e.g. oxygen) which is centred at position, $\mathbf{R}_{k,I}$. $\hat{v}^{(j)}_{\mathrm{emb}}(\mathbf{r}-\mathbf{R}_{j,L})$ is the electrostatic potential due to a particular type of embedding charge distribution which is centred at position $\mathbf{R}_{j,L}$. If the Fourier transform of the potential of each species is provided, the Fourier transform of the total local potential can be obtained as follows,

$$\begin{aligned}\tilde{V}_{\mathrm{ext,loc}}(\mathbf{g})+\tilde{V}_{\mathrm{emb}}(\mathbf{g}) &= \sum_{j=1}^{N_{\mathrm{species}}} \tilde{v}^{(j)}_{\mathrm{ps,loc}}(\mathbf{g}) \sum_{J=1}^{N_j} e^{-i\mathbf{g}\cdot\mathbf{R}_{j,J}} \\ &+ \sum_{p=1}^{N_{\mathrm{emb.species}}} \tilde{v}^{(p)}_{\mathrm{emb}}(\mathbf{g}) \sum_{P=1}^{N_p} e^{-i\mathbf{g}\cdot\mathbf{R}_{p,P}} \\ &= \sum_{j=1}^{N_{\mathrm{species}}} \tilde{v}^{(j)}_{\mathrm{ps,loc}}(\mathbf{g})\, S^{(j)}_{\mathrm{ps}}(\mathbf{g}) \\ &+ \sum_{p=1}^{N_{\mathrm{emb.species}}} \tilde{v}^{(p)}_{\mathrm{emb}}(\mathbf{g})\, S^{(p)}_{\mathrm{emb}}(\mathbf{g}), \qquad (5.10)\end{aligned}$$

where the terms $S^{(j)}_{\mathrm{ps}}(\mathbf{g})$ and $S^{(p)}_{\mathrm{emb}}(\mathbf{g})$ as defined by the above equation are the structure factors [125] for each species of pseudopotential and embedding potential respectively. The incorporation of the embedding potentials to the Kohn-Sham Hamiltonian can be done with minimal additional cost by building them into the Fourier transform of the local part of the external potential. The above form gives us the flexibility to use any functional form for the embedding charge distribution $q^{(p)}_{\mathrm{emb}}(\mathbf{r})$, since if its Fourier transform $\tilde{q}^{(p)}_{\mathrm{emb}}(\mathbf{g})$ can be obtained, it is possible to obtain an expression for its potential, $\tilde{v}^{(p)}_{\mathrm{emb}}(\mathbf{g})$. The incorporation of the embedding potentials into the electronic Hamiltonian through equation 5.10 ensures that the second term in equation 5.5 is obtained as part of the interaction of the electrons with the external potential (now augmented by the embedding potentials).

Electrostatic energies are computed for periodically repeated charge distributions and uniform background charges to ensure charge neutrality. Dispersion interactions (only amongst the quantum atoms) are included with the DFT+D approach as implemented in ONETEP [5]. To avoid the unphysical effect of "charge spilling"

[126] (shown in Figure 5.2), where electron density from the QM region is pulled onto the classical atoms with positive charge, $\tilde{v}_{\mathrm{emb}}$ is described as the potential of a Gaussian function [127], rather than a point charge.

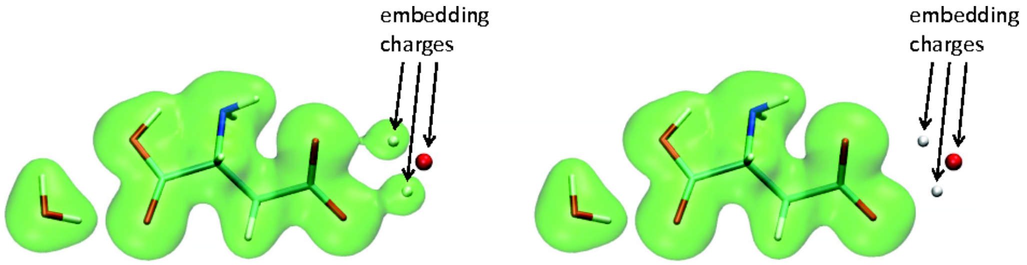


Figure 5.2: Electron density isosurfaces of the diaqua ASP complex, with the molecular structure overlayed, for the case q_O=-4.0 e^- (isosurface value=0.2 e^2 a_0^{-3}). Spilling of electronic density occurs when the embedding is done with point charges (left), but not when Gaussian charge distributions are used (right) This is for quite extreme charges and is used to show the effect which with normal charges is still present and effect the energy but is not as obvious to visualise. [4]

5.1.1 Interaction energies

Our goal is to extend the QM corrected TI approach to a quantum mechanical treatment of the ligand and a large portion of the surrounding atoms, large enough to ensure that the interaction energies will be converged to chemical accuracy. Although this could be achieved by simply increasing the size of the QM region, using the electrostatic embedding should allow for faster convergence with a much smaller QM region. The aim is to obtain interaction energies via a QM EE approach with no appreciable change from the energies obtained from a calculation when the entire system is described by QM. This was investigated for a number of different ligands in water, and for a protein-ligand system where all of the ligand, the protein, and a number of waters surrounding the pocket are treated in a quantum way.

Solvent-ligand interactions

As ligand test molecules we have used toluene, bromobenzene, phenol, thiophenol, catechol (2-hydroxyphenol), cysteine terminated with ACE and NME groups, cystine zwitterion, and serine zwitterion.

Each ligand was generated in the MOE program [104] and solvated with explicit water in a cubic box with periodic boundary conditions in the AMBER Version 10 [54] package. To equilibrate the ligand in a waterbox, the system was heated from 100 K to 300 K with the NVT ensemble over 300 ps then for 200 ps at 300 K with the NPT ensemble in order to adjust the volume of the simulation cell and consequently the density of the water. Then the equilibration was completed with the NVT ensemble for 200 ps again at 300 K. The production calculation was with NVT at 300 K for 1 ns.

All MD simulations used the Langevin thermostat, the particle mesh Ewald sum (PME) for the long range electrostatics, and a time-step of 2 fs with the SHAKE algorithm [65]. The AM1-BCC method was used to obtain partial charges for the ligands with antechamber in the AMBER package. The TIP3P model [70] was used for the water solvent and the generalised amber forcefield (gaff) [59] for the ligands. The small systems (ligands in water and amino acid pairs) were solvated in such a way that the simulation cells were cubes with faces at least 15Å away from the atoms of the initial structure of the solute. This resulted in a total of between 1500 and 1600 water molecules in each simulation cell.

For each of these ligands, production MD simulations were ran and 2 (or 3) snapshots were taken. Interactions energies were obtained for each snapshot with an increasing number of solvent molecules from the simulation cell surrounding the ligand. The waters that are kept are those closest to the ligand, described as "solvation shells" of increasing radius, as demonstrated in Figure 5.3. The interaction energies were then calculated in three ways:

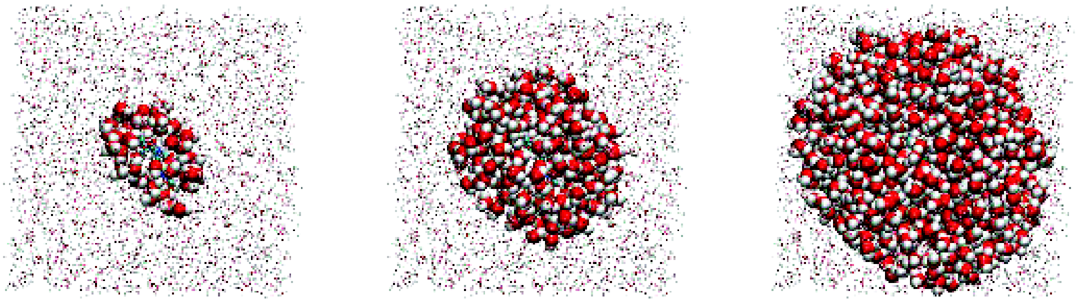


Figure 5.3: Separation of the phenol-water system into quantum and embedding atoms. From left to right, 50, 250 and 750 water molecules closest to the ligand are treated as quantum atoms within the ONETEP calculation. The remaining water molecules of the simulation (which was carried out in a waterbox of about 1600 water molecules) are treated as classical embedding charges. [4]

1. MM[1] force field calculations for the ligand and the surrounding water solvation shell (MM).
2. ONETEP DFT calculations for the ligand and the surrounding water solvation shell (QM).
3. ONETEP DFT calculations for the ligand and the surrounding water solvation shell, including the remaining water molecules of the waterbox via electrostatic embedding (QM EE).

The interactions energies as a function of increasing number of water molecules obtained from these three approaches are shown in Figure 5.4. A measure for deciding how many quantum waters to include can be provided by the radii of the solvation shells. For the case of the phenol molecule in Figure 5.4, the energies are obtained for shells of water with approximate distances from the atoms

[1]

of the phenol of 3.4 Å (first solvation shell), 5 Å (second solvation shell), 9 Å (200 water molecules), 12 Å (400 water molecules), 14 Å (700 water molecules), and 17 Å (1000 water molecules). The embedded QM (QM EE) calculations gave the smoothest and most rapid convergence for all the cases. These results indicate that the combination of a number of quantum waters surrounding the solute, with electrostatic embedding to represent the remaining water in the simulation cell, can adequately capture all the charge polarisation of the ligand and the back-polarisation of its surroundings that is characteristic of the quantum description. This approach also correctly describes the long range electrostatic interactions as they emerge from the periodic boundary conditions that apply to both our MD and quantum calculations. From these calculations it would suggest that with a QM region consisting of around 200-400 water molecules (radii of 9-12Å) combined with electrostatic embedding is enough to reproduce interaction energies that are very similar to those obtained from full QM calculations with errors less than 0.5 kcal mol^{-1}. Although treating the entire system is possible, reducing the number of QM atoms in the simulation dramatically reduces the computational cost. For example, the full quantum system took 2800 core hours, compared to 336 core hours for the 400 water QM EE calculations.

A further measure of the performance of the embedding approach can be provided by investigation of its effect on atomic charges, which are indicators of the chemical environment that the atoms experience. In Figure 5.5 we investigate the Mulliken atomic charges of the atoms of the cysteine zwitterion for increasing sizes of solvation regions for the QM and QM EE approaches. Taking the fully quantum calculation as the benchmark, we observe that for the first solvation sphere the QM EE approach produces for most atoms less than half the error of the QM calculation. However for the case of 400 quantum waters or more the differences

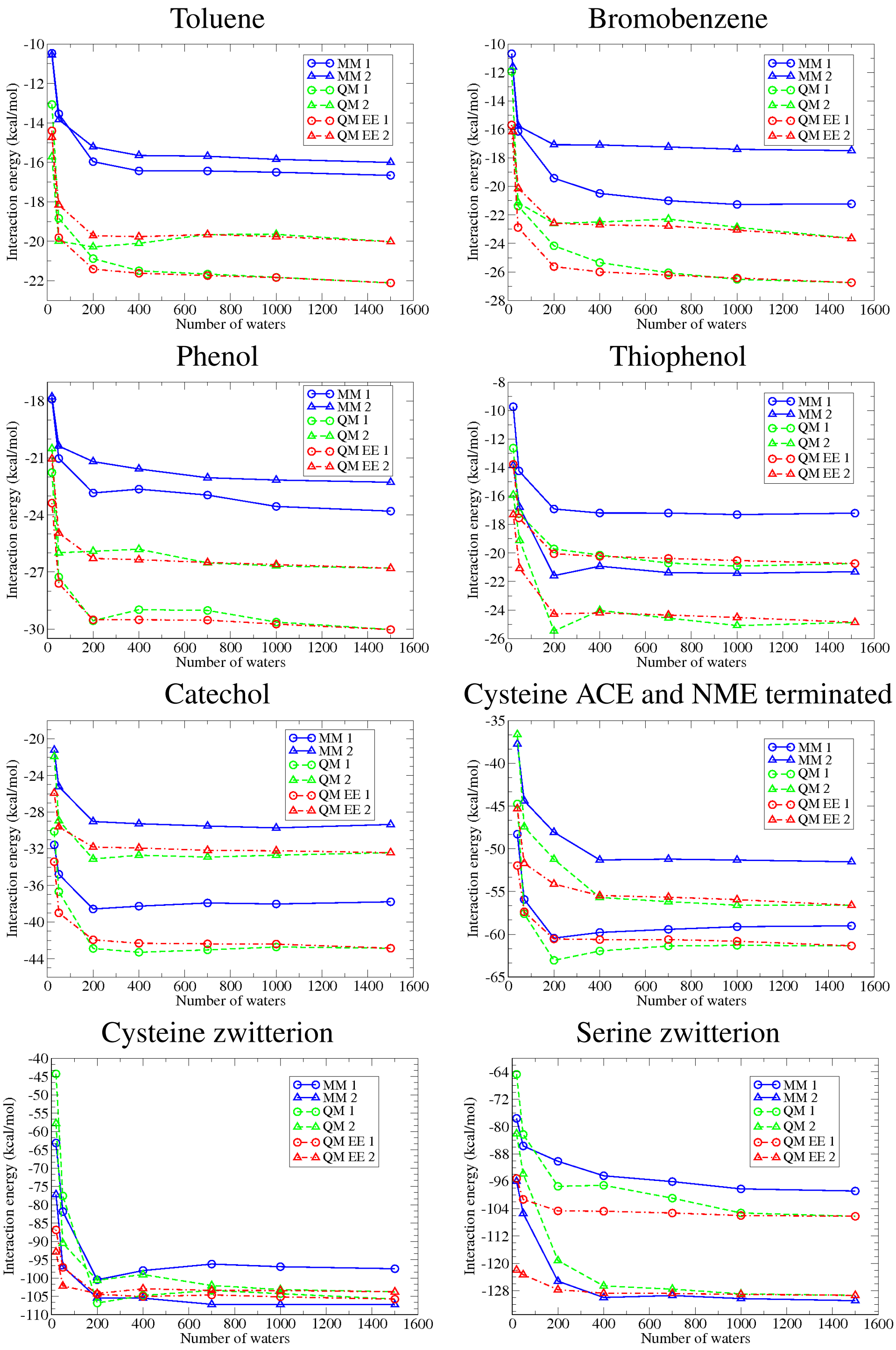


Figure 5.4: Interaction energies between ligands and water as a function of increasing number of water molecules. [4]

between QM and QM EE diminish as the errors become small (less than 0.01 eV), and for this case the difference between the QM and QM EE interaction energies is small.

Receptor-ligand interactions

Initially we considered the interaction of two amino acids: a serine-lysine complex with a net charge of +1, and a serine-aspartate complex with a net charge of -1. Electrical neutrality is imposed by the presence of a counterion (either Na^+ or Cl^-) and is always treated as a classical charge. Three snapshots were taken: snapshot 1 with the counterion at 10 Å form the complex, snapshot 2 with the counterion at 17 Å from the complex, and snapshot 3 with the counterion at 24 Å from the complex. The interactions for the three snapshots for these two receptor-ligand complexes are shown in Figure 5.6 as a function of increasing number of water molecules. As with the solvent-ligand calculations, the interaction energies have been calculated with MM, QM and QM EE. Interaction energies converge following patterns similar to those of the neutral ligands in water (Figure 5.4) for the electrostatic embedding approach.

Finally this approach was applied to phenol bound in the buried polar cavity of T4 lysozyme L99A/M102Q. For the protein ligand complex, the X-ray crystal structures were checked and protonated with the MOE program [104], then solvated with explicit water in a rectangular box with periodic boundary conditions in the AMBER Version 10 [54] package. The following equilibration procedure was employed: the hydrogens were relaxed keeping all heavy atoms fixed with harmonic restraints in the protein and solvent, then the solvent was relaxed with the protein atoms still fixed. The system was heated gradually to 300 K while still restraining

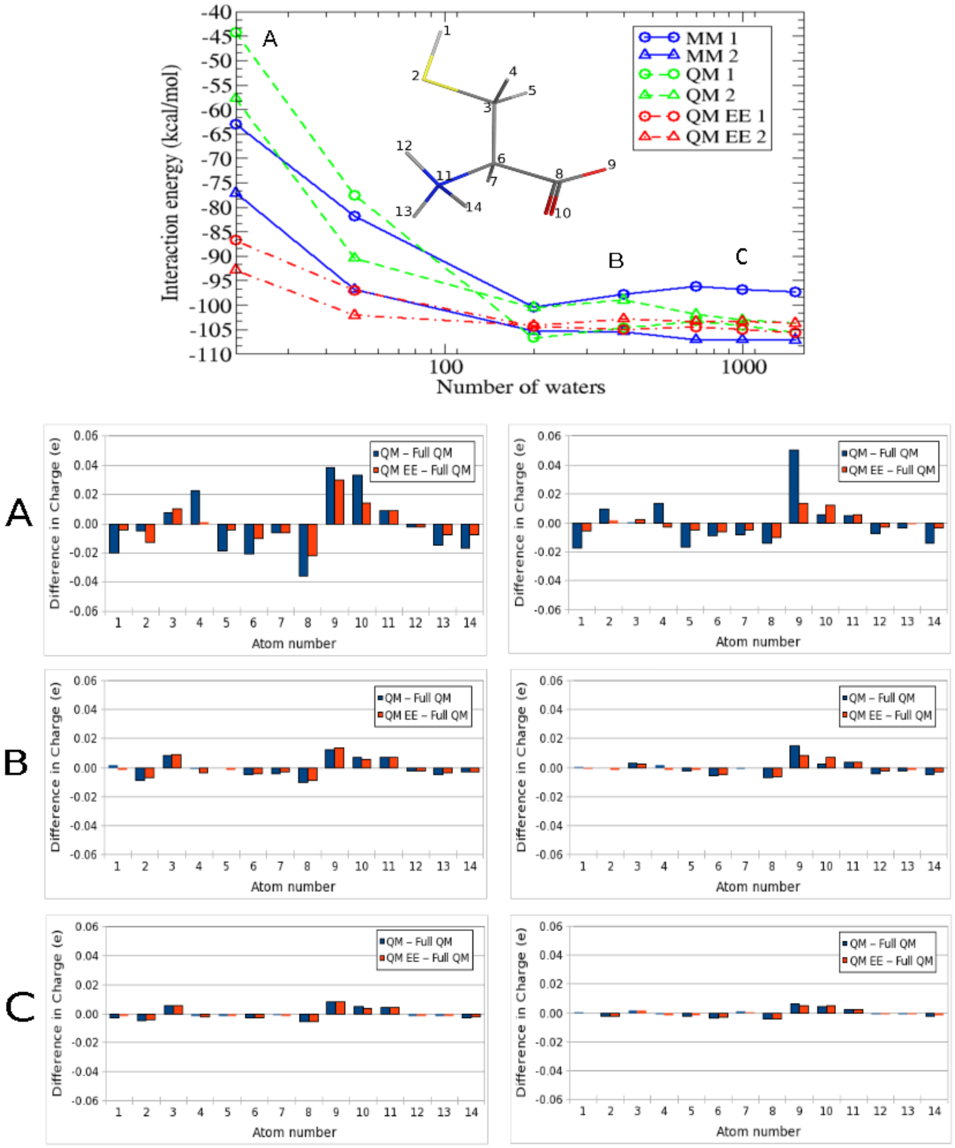


Figure 5.5: Variation of atomic charges from the quantum calculation on the cysteine molecule as a function of the thickness of the solvation shell, for snapshot 1 (left) and snapshot 2 (right) for (A) 20 quantum waters, (B) 400 quantum waters and (C) 1000 quantum waters. For each solvation shell, the difference of the charge on each atom from the charge obtained from the full QM calculation (including all waters in the simulation cell in the quantum description) is given for the QM and QM EE approaches.[4]

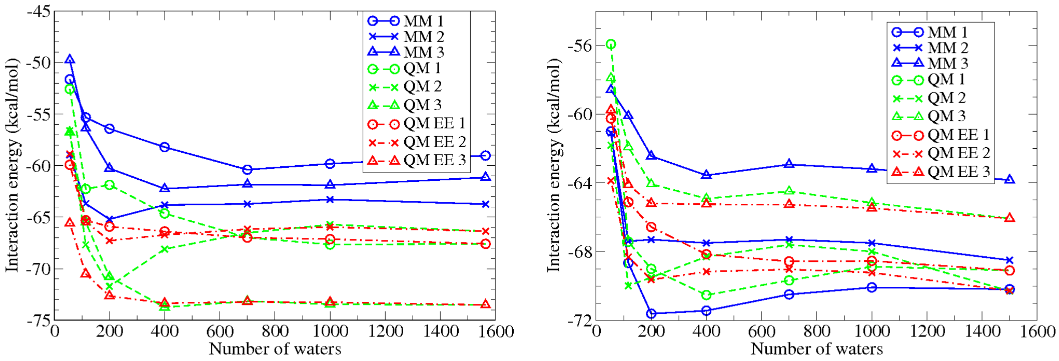


Figure 5.6: Interaction energies between a serine (SER) and a Lysine (LYS) in water (left) and between serine and an aspartate (ASP) in water (right). [4]

the protein for 200 ps with the NVT ensemble and ran for a further 200 ps with the NPT ensemble at 300 K. This was cooled to 100 K over 100 ps and a series of minimisations was carried out reducing the restraints on the protein heavy atoms in stages ($500, 100, 50, 20, 10, 5, 2, 1, 0.5$ kcal mol^{-1}Å^{-2}). Finally the system was heated to 300 K with no restraints over 200 ps and then ran for a further 200 ps at 300 K with NPT, at the end of which the energy and the density of water in the simulation cell were stabilised and so was the internal structure of the protein as measured by the root mean squared deviation of the backbone atoms from the starting structure, which was 0.75 Å. Production simulations were run for 10 ns with the NVT ensemble at 300 K. All MD simulations used the Langevin thermostat, the particle mesh Ewald sum (PME) for the long range electrostatics, and a time-step of 2 fs with the SHAKE algorithm. The AM1-BCC method was used to obtain partial charges for the ligands with antechamber in the AMBER package. The ff99SB forcefield [105] was used for the protein with the TIP3P model [70] for the water solvent and the generalised amber forcefield (gaff) [59] for the ligands. The total charge of this protein is +8, so 8 Cl^{-} counterions were included to impose charge neutrality. The system contained 9053 water molecules.

In this case the receptor is now the entire 2601 atom protein. DFT calculations have been performed for zero quantum waters (where only the 2601 protein atoms plus the 13 ligand atoms are described by DFT), and by including shells with thickness of 3.4 Å, 5.0 Å, 7.5 Å, 9.0 Å, 10.5 Å, up to 12.0 Å which results in 10151 atoms in total being treated by DFT. The interaction energies of two snapshot shots calculated in the same manner as the previous two example is shown in Figure 5.7.

The inclusion of water into the calculation of the binding energy has a very small effect on the obtained value. The embedded calculation shows marginally better convergence with respect to the quantum calculation without embedding, but the advantage of embedding is almost negligible with variations which are less than 1 kcal mol^{-1}. This is a result of the fact that the cavity of T4 lysozyme (L99A/M102Q) is completely buried and shielded from the solvent. In complexes with solvent accessible cavities the inclusion of water is expected to have a larger effect. The regions and thickness of quantum water layers that need to be included will vary from one protein to another and need to be determined on a case by case basis. It is interesting however to observe that in the case of this protein the use of electrostatic embedding with no quantum waters leads to the largest errors as the embedding atoms in contact with the quantum atoms of the protein appear to overpolarise it. In fact Figure 5.7 shows that the DFT calculation with no water at all would be in this case the best compromise between accuracy and efficiency as the errors that result are of the order of 0.5 kcal mol^{-1} which is comparable to other errors intrinsic in DFT calculations (such as the choice of exchange-correlation functional and the basis set). Another interesting observation is that the QM and QM EE curves, for snapshots 1 and 3, do not coincide, even for the largest calculations with 2517 quantum water molecules. As the total energies in our calculations were converged to 0.1 kcal mol^{-1} it is unlikely that this is due to numerical noise but most likely it is a manifestation of the long-range nature of electronic polarisation which appears to not be completely converged for these structures even with this large number of water molecules.

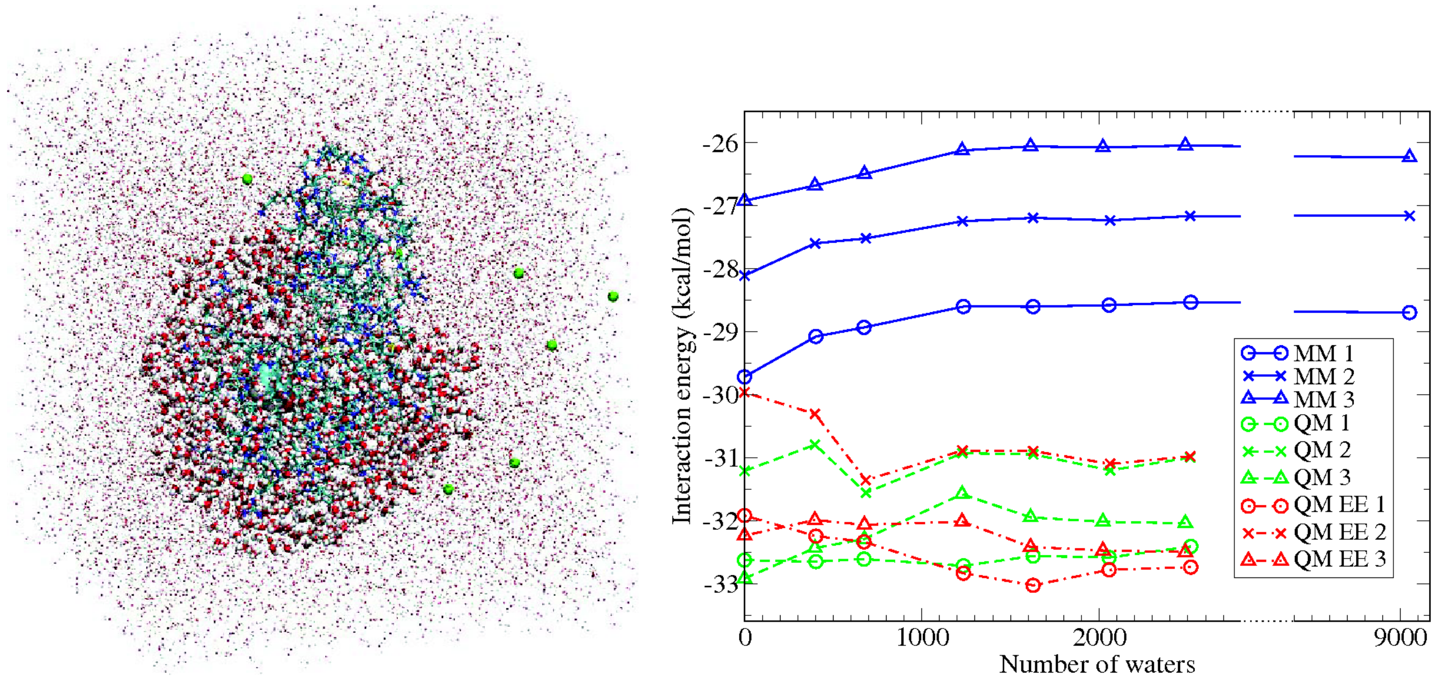


Figure 5.7: Left: The complex of L99A/M102Q T4 Lysozyme and phenol in water. The second solvation sphere around the ligand binding cavity is shown in ball and stick representation while the rest of the waters are shown as dots. Right: Interaction energies between phenol and L99A/M102Q T4 Lysozyme in water for increasing numbers of water molecules within a sphere around the binding pocket.

5.2 QM corrected TI ligand hydration energies

Relative hydration energies were computed using the electrostatic embedding approach described in section 5.1 and the scheme in Figure 5.8. The molecules for which hydration energies were obtained are toluene, phenol, catechol, 2-fluoroaniline, 3-chlorophenol, thiophenol, and 2-methylphenol (shown in Figure 5.9)

5.2.1 MM simulation setup

For the setup of the MD simulations we started from a catechol molecule (generated in the MOE program [104]) placed in a water box containing 1545 explicit waters in a cubic box with periodic boundary conditions in the AMBER Version 10 [54] package.

To equilibrate the system, the positions of the Hydrogens were relaxed before heating the system from 100 to 300 K with the NVT ensemble over 200 ps with positional restraints of 1000 kcal mol^{-1} Å$-^{2}$ on the catechol molecule. Then we switched to the NPT ensemble for 200 ps keeping the positional restraints on catechol. The system was then ran for a further 200 ps with no restraints in the NVT ensemble and again switched to NPT for 200 ps at 300 K in order to add a final adjustment to the volume of the simulation cell, and consequently the density of the water. The simulation cell was constrained to remain cubic and its final volume had sides of 36.222 Å.

At this point it was confirmed that the water density, kinetic energy, and potential energy had only small oscillations around a constant value so the system was deemed to be equilibrated. The catechol ligand was manually mutated in the MOE program to the six other ligands in Figure 5.9. Production MD simulations were

$$\begin{tikzcd}
(\mathrm{L}_1)_{\mathrm{aq,QM}} & (\mathrm{L}_1)_{\mathrm{aq,\,MM}} \arrow[l, "\Delta G_{\mathrm{L1,\,MM\text{-}>QM,\,aq}}"'] \arrow[r, "\Delta G_{\mathrm{L1\text{-}>L2,\,MM,\,aq}}"] & (\mathrm{L}_2)_{\mathrm{aq,MM}} \arrow[r, "\Delta G_{\mathrm{L2,\,MM\text{-}>QM,\,aq}}"] & (\mathrm{L}_2)_{\mathrm{aq.QM}} \\
(\mathrm{L}_1)_{\mathrm{vac,\,QM}} \arrow[u, "\Delta G_{\mathrm{hyd1,\,QM}}"'] & (\mathrm{L}_1)_{\mathrm{vac,\,MM}} \arrow[l, "\Delta G_{\mathrm{L1,\,MM\text{-}>QM,\,vac}}=0"] \arrow[r, "\Delta G_{\mathrm{L1\text{-}>L2,\,MM,\,vac}}"] \arrow[u, dashed, "\Delta G_{\mathrm{hyd1,\,MM}}"'] & (\mathrm{L}_2)_{\mathrm{vac,\,MM}} \arrow[r, "\Delta G_{\mathrm{L2,\,MM\text{-}>QM,\,vac}}=0"] \arrow[u, dashed, "\Delta G_{\mathrm{hyd2,\,MM}}"'] & (\mathrm{L}_2)_{\mathrm{vac,\,QM}} \arrow[u, "\Delta G_{\mathrm{hyd2,\,QM}}"']
\end{tikzcd}$$

$$\Delta\Delta G_{\mathrm{hyd,QM}} = \Delta G_{\mathrm{hyd2,QM}} - \Delta G_{\mathrm{hyd1,QM}}$$

$$= \underbrace{\Delta G_{\mathrm{L1\to L2,aq,MM}} - \Delta G_{\mathrm{L1\to L2,vac,MM}}}_{\text{TI}} + \underbrace{\Delta G_{\mathrm{L2,MM\to QM,aq}} - \Delta G_{\mathrm{L1,MM\to QM,aq}}}_{\text{MM->QM steps}}$$

Figure 5.8: MM → QM corrected ligand hydration free energy cycle. Going from L_1 to L_2 both described by QM via a MM alchemical mutation from left to right. The top line depicts the ligand in solvent $(\mathrm{L})_{aq}$, and the bottom line the ligand in vacuum $(\mathrm{L})_{vac}$.

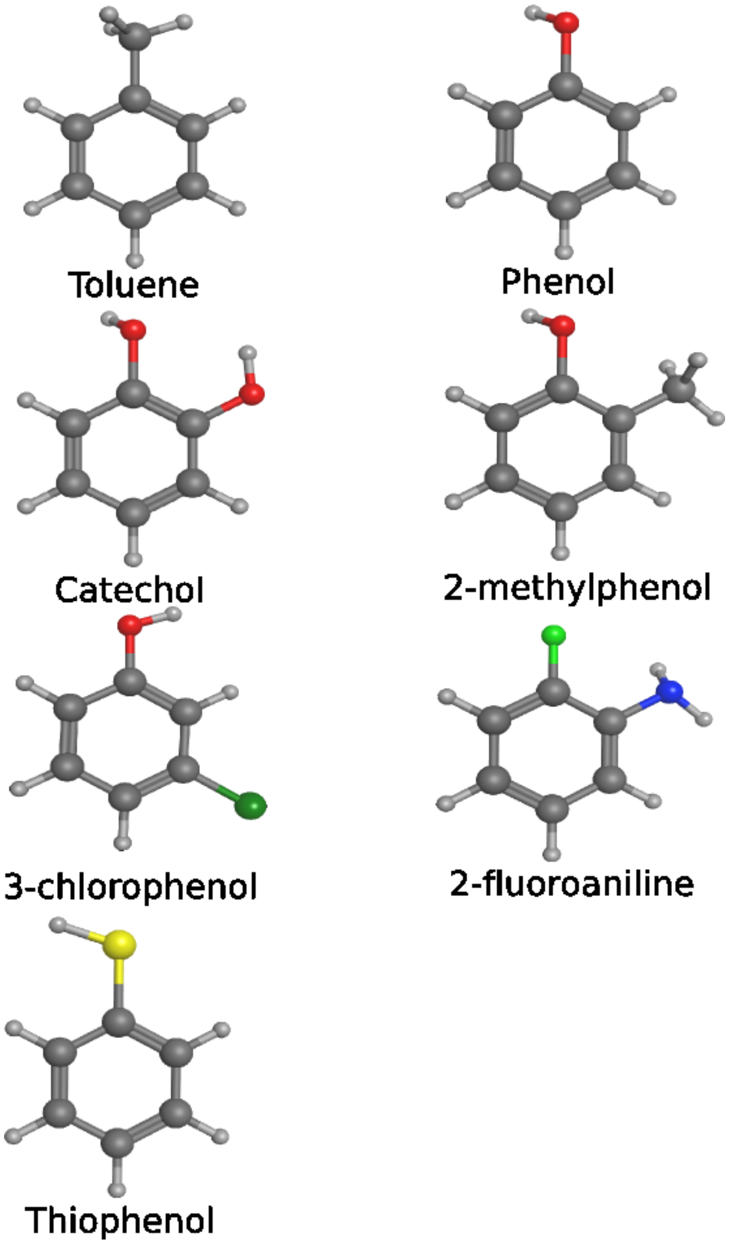


Figure 5.9: Chosen ligands for study.

started from the catechol and these new structures (with randomly assigned initial velocities), each containing a ligand in a water box, and ran in the NVT ensemble at 300 K for 20 ns. Snapshots were taken from the last 18 ns of the trajectory treating the first 2 ns as further equilibration.

For thermodynamical consistency, we have ensured that the same number of water molecules (1545) was used in all simulations. Furthermore, a cubic simulation cell of length 36.222 Å was used in all simulations to ensure identical basis sets for all subsequent ONETEP calculations. For our MD simulations we used the Langevin thermostat [67] with the default parameters in AMBER10, the particle mesh Ewald sum (PME) for the long range electrostatics, a non-bonded cutoff of 8.0 Å, and a time-step of 2 fs with the SHAKE algorithm [65]. The AM1-BCC method was used to obtain partial charges for the ligands with the antechamber tool in the AMBER package. The TIP3P model [70] was used to describe the water solvent and the generalised amber forcefield (gaff) [59] to describe the ligands.

5.2.2 MM to QM calculation set up

Each QM region was defined as the ligand (solute) and the closest 200 waters (roughly equivalent to a 9.0 Å solvation shell). All the remaining water molecules were treated as classical embedding charges. The charge given to the classical Oxygens was -0.834 e and for the classical Hydrogens 0.417 e, as they are in the TIP3P model. NGWF radii of 8.0 a_0 were used for all atoms, with 4 NGWFs on Carbon, Oxygen, and Nitrogen, 9 NGWFs on Sulphur, Fluorine, and Chlorine, and 1 NGWF on Hydrogen. A kinetic energy cut-off of 800 eV was used along with the PBE exchange-correlation functional [34] and with the DFT+D approach [5] accounting for dispersion interactions. All simulation cells were cubic and had identical sizes with a side length of 68.450 a_0 (equivalent to 36.222 Å) to ensure identical psinc basis sets.

For the MM single point energy calculations a non-bonded cutoff of 13.0 Å was used in a periodic cubic box with side lengths 36.222 Å. Full Ewald was used to accurately calculate the electrostatic interactions.

5.2.3 TI calculations

TI calculations were performed with the AMBER program. Ligand starting geometries were taken from the starting geometries for the MD simulations. Perturbations were in the direction phenol $\rightarrow$ new ligand. 39 λ windows were performed ($\lambda = 0.025$). Each λ step involved the relaxation of the entire system, an equilibration rising the temperature from 100 K to 300 K in the NVT ensemble over 50 ps, and finally a 200ps production step in the NPT ensemble at 300 K.

Convergence tests were performed using 9 λ windows, 19 λ windows, then finally 39 λ windows. The difference in the calculated $\Delta\Delta G$ using 9, 19, or 39 windows is very small. For example, the difference between using 19 windows or 39 windows to calculate the $\Delta\Delta G$ for the phenol→catechol mutation is 0.34 kcal mol^{-1}, with the error between the forward and reverse calculations being reduced from 0.04 kcal mol^{-1} for 19 windows to 0.01 kcal mol^{-1} when using 39 windows. When using only 9 windows the $\Delta\Delta G$ differs form 39 windows by 0.20 kcal mol^{-1}, however the error between the forward and reverse paths is larger at 0.08 kcal mol^{-1}. The maximum difference between 9 and 39 windows is for the phenol→2-methylphenol mutation, which is 0.66 kcal mol^{-1}. This is reduced to 0.09 kcal mol^{-1} when using 19 windows. We have chosen to use 39 windows for the small improvement in convergence that is seen.

5.2.4 Interaction energy distributions

A one step energy perturbation uses the form of the Zwanzig equation we have proposed in Equation 5.3 to calculate the energy change from the MM description of the system to the QM EE description of the system. Snapshots are taken at constant time intervals from the last 18 ns of the production trajectories, and the interaction energies are computed with MM and with QM EE.

As well as using interaction energies in the one step perturbation, the interaction energy distributions were fitted to a Gaussian curve and the resulting function used in the Zwanzig equation to calculate the MM → QM free energy change [128]. This method of analysis is done to minimise systematic and random error. The form of the Gaussian used is,

$$C \exp\left(\alpha(E - E_0)^2\right). \tag{5.11}$$

An example of a Gaussian of this form fitted to the histogram of $\Delta\Delta E$ of the phenol QM interaction energies minus the MM interaction energies can be seen in Figure 5.10.

The Zwanzig equation is now written as,

$$\Delta G = E_0 - \left(\frac{1}{4\alpha k_B T}\right). \tag{5.12}$$

This form has been derived in the following way:

If the probability distribution of the energies is given by a function W(E), we wish to obtain the average of the Zwanzig equation in the form of an integral as follows,

$$\left\langle e^{\frac{-E}{k_BT}} \right\rangle = \frac{\int_{-\infty}^{\infty} e^{\frac{-E}{k_BT}} W(E) dE}{\int_{-\infty}^{\infty} W(E) dE} = \frac{I_1}{I_2}, \tag{5.13}$$

where the function W(E) is, in our case, a Gaussian function as presented in Equation 5.11.

If we first look at I_1,

$$\begin{aligned} I_1 &= \int_{-\infty}^{\infty} e^{\frac{-E}{k_BT}} e^{-\alpha(E-E_0)^2} dE \\ &= \int_{-\infty}^{\infty} e^{-\beta E} e^{-\alpha(E-E_0)^2} dE \\ &= \int_{-\infty}^{\infty} e^{-\alpha(E-E_0)^2-\beta E} dE. \end{aligned} \tag{5.14}$$

where $\beta = \frac{1}{k_BT}$. If we work on the exponent,

$$\begin{aligned} &-\alpha\left(E-E_0\right)^2 - \beta E \\ &\quad = -\left\{\alpha E^2 + \alpha E_0^2 - 2\alpha E E_0 + \beta E\right\} \\ &\quad = -\left\{\alpha E^2 + E(\beta - 2\alpha E_0) + \alpha E_0^2\right\} \\ &\quad = -\alpha\left\{E^2 + E\left(\frac{\beta}{\alpha} - 2E_0\right) + E_0^2\right\} \\ &\quad = -\alpha\left\{E^2 + 2E\left(\frac{\beta}{2\alpha} - E_0\right) + \left(\frac{\beta}{2\alpha} - E_0\right)^2 - \left(\frac{\beta}{2\alpha} - E_0\right)^2 + E_0^2\right\} \\ &\quad = -\alpha\left\{\left[E + \left(\frac{\beta}{2\alpha} - E_0\right)\right]^2 - \frac{\beta^2}{4\alpha^2} + \frac{\beta E_0}{\alpha}\right\} \\ &\quad = -\alpha\left[E + \left(\frac{\beta}{2\alpha} - E_0\right)\right]^2 - \frac{\beta^2}{4\alpha} + \beta E_0. \end{aligned} \tag{5.15}$$

Therefore I_1 becomes,

$$\begin{aligned} I_1 &= \int_{-\infty}^{\infty} e^{-\alpha(E-E_0)^2-\beta E} dE \\ &= \int_{-\infty}^{\infty} e^{-\alpha\left[E+\left(\frac{\beta}{2\alpha}-E_0\right)\right]^2-\frac{\beta^2}{4\alpha}+\beta E_0} dE \\ &= e^{\frac{\beta^2}{4\alpha}+\beta E_0} \int_{-\infty}^{\infty} e^{-\alpha\left[E+\left(\frac{\beta}{2\alpha}-E_0\right)\right]^2} dE. \end{aligned} \tag{5.16}$$

If we set $x = \sqrt{\alpha}\left[E+\left(\frac{\beta}{2\alpha}-E_0\right)\right]$, then $dx = \sqrt{\alpha}dE \therefore dE = \frac{1}{\sqrt{\alpha}}dx$. The integral then becomes,

$$I_1 = \frac{1}{\sqrt{\alpha}} e^{\frac{\beta^2}{4\alpha}+\beta E_0} \int_{-\infty}^{\infty} e^{-x^2} dx, \tag{5.17}$$

if $\int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi}$, then

$$I_1 = \sqrt{\frac{\pi}{\alpha}} e^{\frac{\beta^2}{4\alpha}+\beta E_0}. \tag{5.18}$$

If we now look at I_2

$$I_2 = \int_{-\infty}^{\infty} e^{-\alpha(E-E_0)^2} dE \tag{5.19}$$

If we set $y = \sqrt{\alpha}(E-E_o)$, then $dy = \sqrt{\alpha}dE \therefore dE = \frac{1}{\sqrt{\alpha}}dy$.

$$I_2 = \frac{1}{\sqrt{\alpha}} \int_{-\infty}^{\infty} e^{-y^2} dy = \sqrt{\frac{\pi}{\alpha}}. \tag{5.20}$$

Combining I_1 and I_2 gives,

$$\frac{I_1}{I_2} = \left\langle e^{-\beta E} \right\rangle = e^{-\frac{\beta^2}{4\alpha}+\beta E_0}. \tag{5.21}$$

So if $E = \Delta E^{QM} - \Delta E^{MM}$ then,

$$\begin{aligned}
\Delta G &= -\beta^{-1} \ln \langle e^{-\beta E} \rangle \\
&= -\beta^{-1} \ln \frac{I_1}{I_2} \\
&= -\beta^{-1} \ln \left[e^{-\frac{\beta^2}{4\alpha} + \beta E_0} \right] \\
&= -\beta^{-1} \left(\frac{\beta^2}{4\alpha} - \beta E_0 \right) \\
&= E_0 - \frac{\beta}{4\alpha}. \qquad (5.22)
\end{aligned}$$

Another function that can be used is a “universal” probability function (UPDF) [128], and has the form,

$$P(y) = K e^{a\left[b(y-s) - e^{b(y-s)}\right]}, \qquad (5.23)$$

where K, a, b, and s are adjustable parameters. This function has the advantage of being capable of modelling data that is both non-Gaussian and “heavily-tailed”.

Snapshots were taken from the 18 ns trajectories at constant time intervals. Convergence of the Zwanzig equation using interaction energies are shown in Figure 5.11. This shows the result for the MM to QM energy as the number of snapshots are increased, up to a total of 180 snapshots. It can be seen that a few snapshots result in jumps in the free energy change, however the energy is seen to be converging with the last 10 snapshots changing the energy by less than 0.5 kcal mol^{-1}. The free energy change was also calculated using a “leave one out” approach, where the free energy change using the Zwanzig equation is calculated using all but the current snapshot. This data is presented in Figure 5.12 against the energy difference between the MM and QM interaction energies for energy snapshot that is “left out”. You can see that the snapshots with a large negative energy difference between the MM and QM interaction energies have the largest effect on the total free energy change. In most cases this difference is less than 0.2 kcal mol^{-1}. Thiophenol is a noticeable exception with the largest difference being 0.73 kcal mol^{-1}, and phenol with a difference of 0.48 kcal mol^{-1}.

The 180 snapshots were split into 4 sets of 90 snapshots in the ways depicted in Figure 5.13. The standard error over the 4 sets of 90 snapshots was estimated to present an idea of convergence of the computed free energy. For the interaction energies being used in the Zwanzig equation the standard errors are reasonably small, less than 0.3 kcal mol^{-1} for most of the ligands. Except for thiophenol which has a standard error of 0.69 kcal mol^{-1}. This suggests that more sampling should ideally be done for thiophenol, whereas the other ligands are reasonable well converged by 90 snapshots. This is not the case when using the fitted Gaussian approach, with generally larger errors for the ligands, and substantially larger for thiophenol, at 1.53 kcal mol^{-1}. The standard errors can be seen in Table 5.1.

Since interaction energies are being used, and not total energies, the right and left sides of the bottom line in Figure 5.8 will be zero since there will be no change in the interaction energy of the ligand with a vacuum between the MM and QM EE descriptions. In the future, when protein-ligand binding energies are calculated, this will not be the case, as the bottom row will depict the ligand in solvent, and sampling would have to be performed at these points.

An overview of the simulations performed is depicted in Figure 5.14.

Table 5.1: Standard errors over the 4 sets of 90 snapshots.

	Catechol	Toluene	3-chlorophenol	2-fluoroaniline	2-methylphenol	Thiophenol
Zwanzig	0.17	0.21	0.29	0.15	0.19	0.69
average $\Delta\Delta$	0.28	0.05	0.15	0.03	0.1	0.16
Gaussian fit	0.48	0.27	0.31	0.10	0.48	1.53

5.2.5 Results

Relative binding free energies obtained from TI have been corrected using the Zwanzig equation, using the averages of the interaction energies, using a Gaussian fit (as in Equation 5.12), and using implicit solvation rather than explicit waters. The calculated relative hydration free energies for the corrected and uncorrected results are presented in Table 5.2 using 90 snapshots for the MM $\rightarrow$ QM perturbation (the odd set as depicted in Figure 5.13), and in Table 5.3 using 180 snapshots in the MM $\rightarrow$ QM perturbation.

When the experimental hydration value wasn't available, the computed hydration energy is compared to the solvation energy obtained from a Gaussian SMD implicit solvation [80] calculation. The Gaussian calculations were performed with Gaussian09 [81] at the m052x/6-31G(d) SCRF(IEFPCM,Solvent=Water,SMD) level. The average $\Delta\Delta E$ approach simply consists of using the average of the MM and the QM EE interaction energies in Equation 5.3. This results in ΔG values being the difference in just the average of the MM interaction energies and the average of the QM EE interaction energies. The implicit solvation calculations were performed on the 90 snapshots using the implicit solvation model in ONETEP [78] and the MM-PBSA approach in AMBER. The difference in the resulting solvation energies was used in the Zwanzig equation.

For a small mutation in the ligand, as with the molecules in this study, standard TI predicts the relative binding free energies very well, and has a very good correlation with experiment with an R^2 value of 0.98. The largest error seen for the MM TI results is 1.4 kcal mol^{-1} (for toluene) with a mean error of 0.9 kcal mol^{-1} and an rms error of 0.9 kcal mol^{-1}. TI is the most theoretically rigorous free energy approach available and the results from standard TI show that for these cases the

Table 5.2: Relative binding free energies normalised on the Phenol experimental hydration energy using 90 snapshots in the MM → QM perturbation. Energies presented in kcal mol^{-1}.

Normalised on Phenol						
	Zwanzig	average $\Delta\Delta E$	Gaussian fit	ONETEP IS	AMBER-TI	Experiment
Catechol	-9.3	-8.4	-12.7	-9.3	-8.7	-9.4
Toluene	-0.8	-2.3	-1.5	-1.5	-2.3	-0.9
3-chlorophenol	-6.6	-7.4	-8.0	-6.6	-6.8	-6.7†
2-fluoroaniline	-5.1	-5.3	-4.8	-6.1	-5.8	-4.5†
2-methylphenol	-6.5	-7.2	-9.0	-6.1	-6.5	-6.3†
Thiophenol	-4.9	-2.1	-11.3	-2.7	-3.4	-2.6
Phenol (ref)	-6.6	-6.6	-6.6	-6.6	-6.6	-6.6

† energies calculated with Gaussian SMD implicit solvent model.

Table 5.3: Relative binding free energies normalised on the Phenol experimental hydration energy using 180 snapshots in the MM → QM perturbation. Energies presented in kcal mol^{-1}.

Normalised on Phenol						
	Zwanzig	average $\Delta\Delta E$	Gaussian fit	ONETEP IS*	AMBER-TI	Experiment
Catechol	-9.6	-9.1	-11.6	-9.3	-8.7	-9.4
Toluene	-1.0	-2.1	-0.7	-1.5	-2.3	-0.9
3-chlorophenol	-7.3	-7.5	-7.3	-6.6	-6.8	-6.7†
2-fluoroaniline	-5.1	-5.1	-4.7	-6.1	-5.8	-4.5†
2-methylphenol	-6.8	-7.3	-7.8	-6.1	-6.5	-6.3†
Thiophenol	-4.8	-1.7	-6.1	-2.7	-3.4	-2.6
Phenol (ref)	-6.6	-6.6	-6.6	-6.6	-6.6	-6.6

* using only 90 snapshots.

† energies calculated with Gaussian SMD implicit solvent model.

force field describes the systems well.

The best QM corrected result is obtained with the Zwanzig equation (column 2 of Tables 5.2 and 5.3. For 90 snapshots we see an improvement over the MM TI, when one is possible, and very little change when the MM TI result is already very good. The exception is thiophenol whose value is made worse when the QM correction is applied. The correlation using the Zwanzig approach to add a QM correction shows an improvement with an R^2 of 0.99 when not including the result for thiophenol (and 0.91 when it is). The max error is 0.6 kcal mol^{-1} (2.3 kcal mol^{-1}) with an average error of 0.2 kcal mol^{-1} (1.0 kcal mol^{-1}) and an rms error of 0.3 kcal mol^{-1} (1.0 kcal mol^{-1}). The results from the Zwanzig equation using 180 snapshots agree less well with the experimental values, with an rms error of 0.5 kcal mol^{-1} (1.0 kcal mol^{-1}), however the correlation with experiment is further improved with an R^2 of 1.0 when not including the result for thiophenol.

The Gaussian fit shows no improvement over standard TI using 90 snapshots. The total max error of this approach is 8.8 kcal mol^{-1} (for thiophenol), with an rms error of 4.0 kcal mol^{-1}. This approach does retain a high correlation however, with an R^2 value of 0.97., which is improved when using 180 snapshots, now with an R^2 of 0.99. When the 180 snapshots are used the max error is reduced to 3.6 kcal mol^{-1} and the rms error is reduced to 1.8 kcal mol^{-1}. These improvements are likely due to the increased sampling performed when using 180 snapshots which will produce coefficients for the Gaussian curve that better describe the distribution.

The implicit solvation does surprisingly well, and substantially better for thiophenol, reducing the error in the QM corrected value from the experimental value from 2.3 kcal mol^{-1} to 0.1 kcal mol^{-1}. The problem ligand is now 2-fluoroaniline, which goes from an improvement in accuracy of 0.7 kcal mol^{-1}, to a deterioration in accuracy of 0.3 kcal mol^{-1}. The max error here is 1.6 kcal mol^{-1}, with an average error of 0.7 kcal mol^{-1} and an rms error of 0.7 kcal mol^{-1}.

5.2.6 Conclusions

In this study we have used an extended free energy cycle to calculate the relative hydration energies using a quantum Hamiltonian. We have used an electrostatic embedding approach to reduce the number of quantum waters required in the simulation, describing a solvation sphere of waters around the solute with quantum mechanics and the remaining waters in the simulation cell with classical charges. Interaction energies with a differing numbers of quantum water molecules were calculated to obtain the optimal number to use in combination with electrostatic embedding so that no discernable difference is observable from the full QM calculation. We have seen that quite a significant number of water molecules described by QM are necessary to see convergence towards a full QM calculation. Using 200 waters is seen to be a good compromise between accuracy and computational cost, and this number was used for the perturbation from the MM to QM description.

Using the Zwanzig equation with the interaction energies gives the best results compared to experiment from the approaches investigated when 90 snapshots are used. Increasing the number reduces the improvement, this is possibly due to increased sampling of the high energy differences between the QM and MM energies. This is seen when the snapshot generated by the MM potential would not be visited by the QM potential, leading to errors in the energies, and large interaction energy differences between the two methods. Since the exponential of the value is used in the Zwanzig equation this can lead to a few unduly influential snapshots being responsible for the majority of the computed free energy change. This can be avoided by using large structural ensembles, or by removing high energy snapshots from the sample. Using a Gaussian fit was another possibility for correcting for this phenomenon.

Although an improvement towards the experimental relative hydration energy is ideal, this approach is actually obtaining a relative hydration energy that will be closer to the value that would have been obtained if the mutation had been performed completely by a QM approach. Errors from the results form this approach and the experimental results can be due to approximations made in the QM calculation, which can be reduced by increasing the quality of the QM calculations. For the case of our DFT calculations, improvements could be made by using a higher quality exchange-correlation functional.

The improvements that we see with respect to the R^2 values are very marginal (0.97 to 0.99) and with such a small sample it would be hard to say that it is actually an improvement and not just noise. Ideally more ligand perturbations would be performed to better judge the advantage of this method. Standard TI also does quite well for these test systems. It would be interesting to apply this method to small systems where TI in known to break.

This approach could be applied to protein-ligand complexes, however the computational cost is still quite prohibitive. Not only does it require the computational cost of the TI approach, it also require long MD simulations to sample the phase space of the entire molecule, and many expensive QM calculations at either end of the free energy cycle in Figure 5.1.

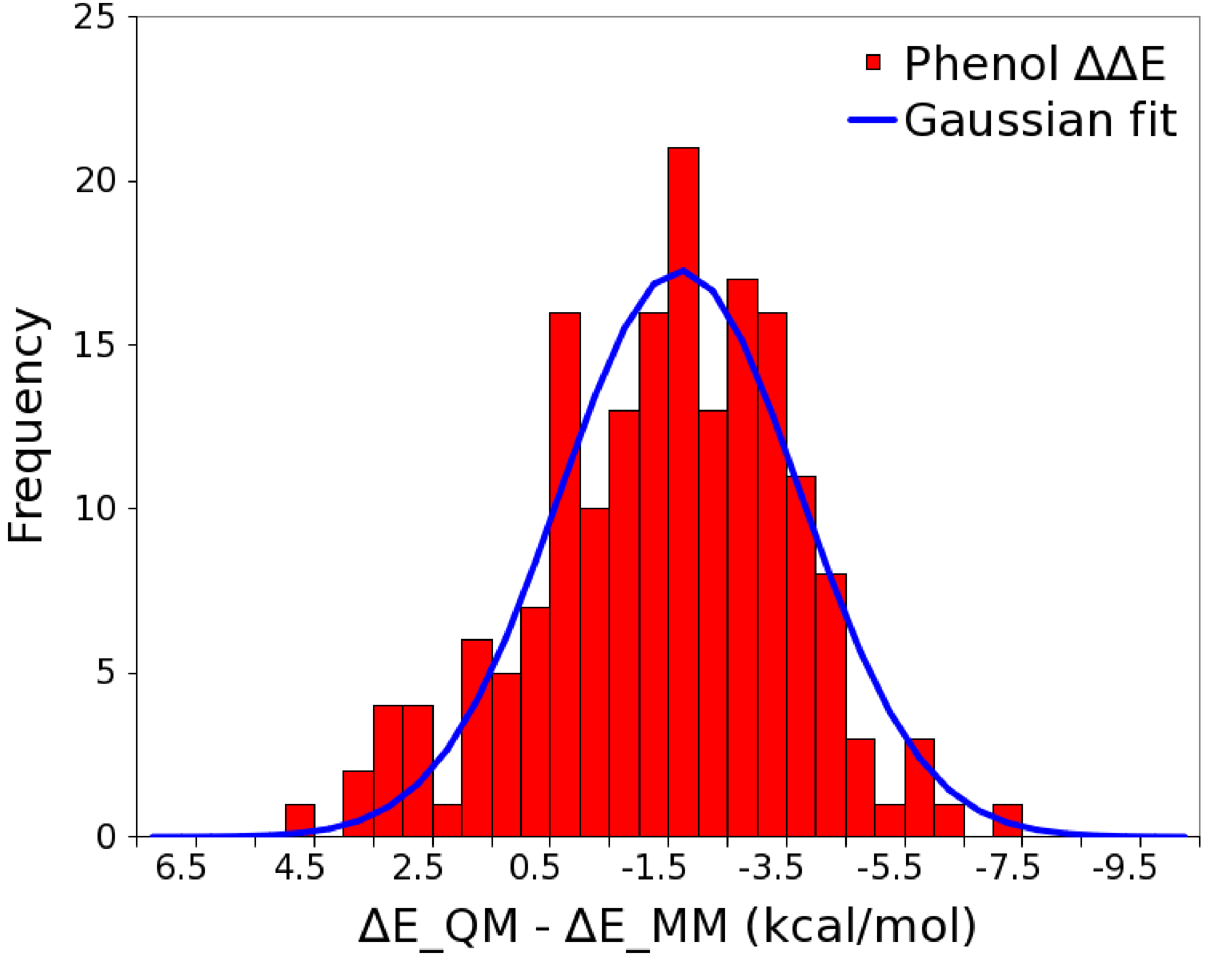


Figure 5.10: Histogram of $\Delta\Delta E$ of the phenol QM interaction energies minus the MM interaction with a fitted Gaussian curve overlayed.

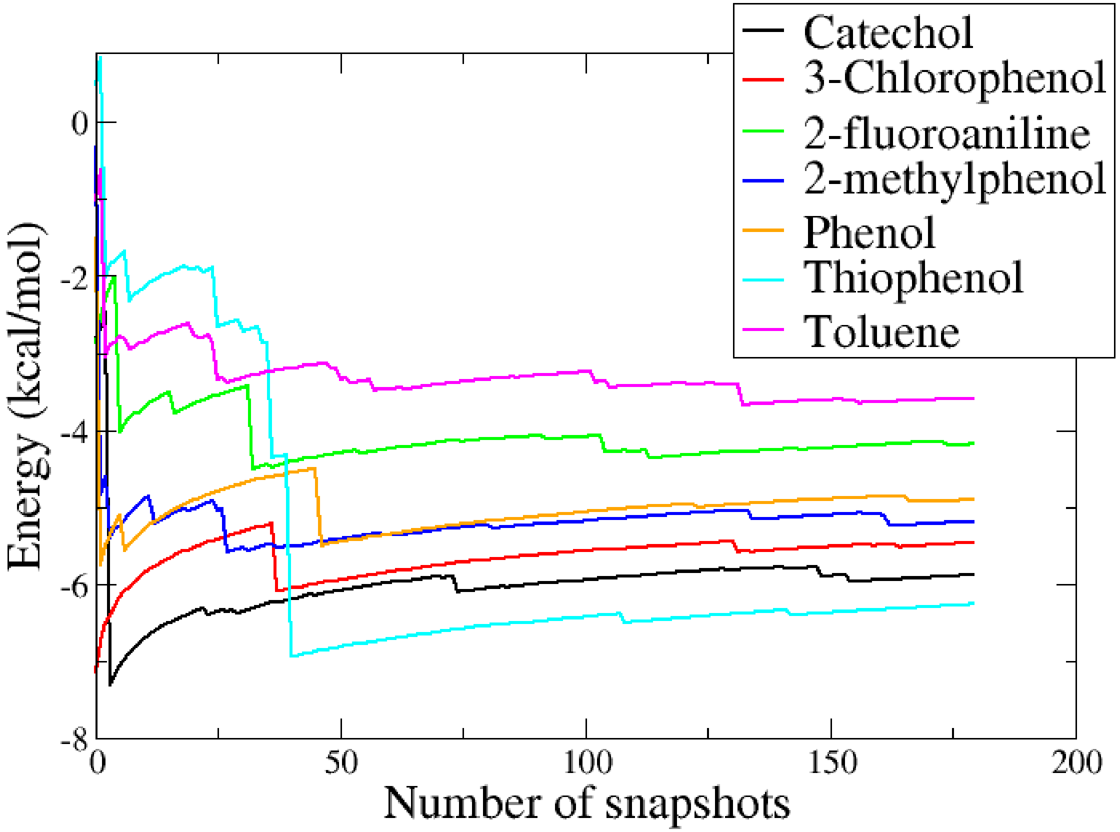


Figure 5.11: Convergence of the MM to QM free energy change as a function of the number of snapshots.

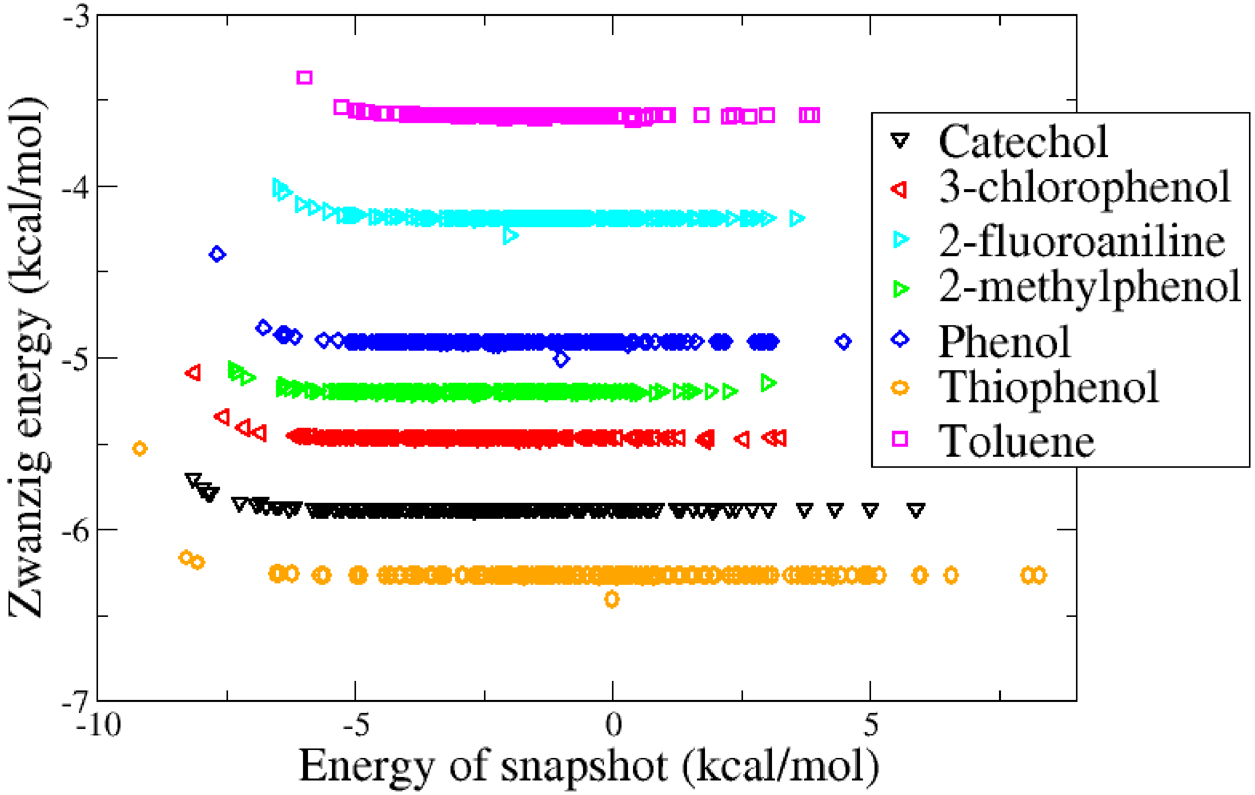


Figure 5.12: The “Leave one out” approach for the MM to QM free energy change.

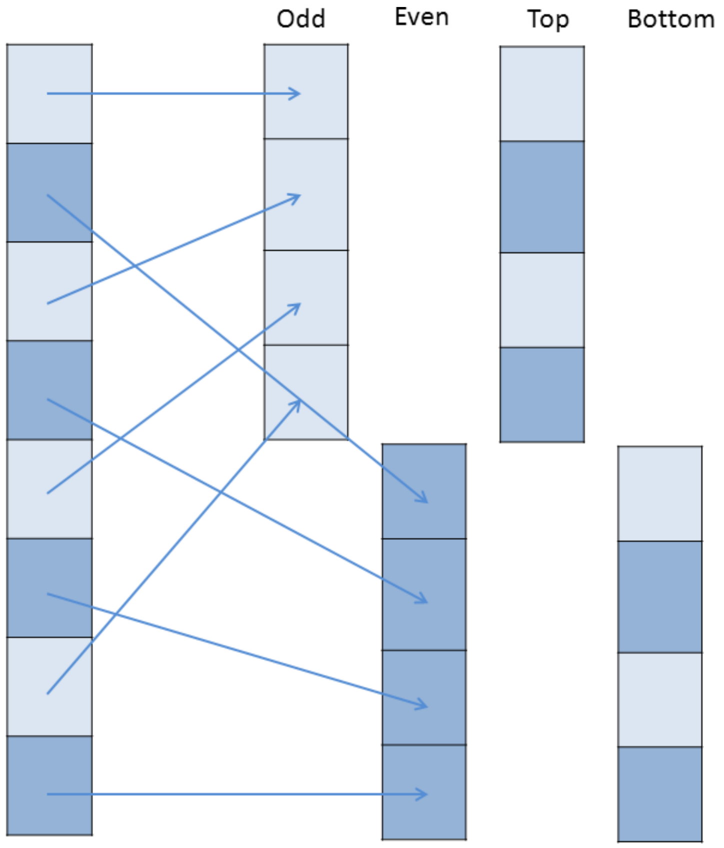


Figure 5.13: Schematic showing the four different ways the 180 snapshots were divided into sets of 90 to estimate errors.

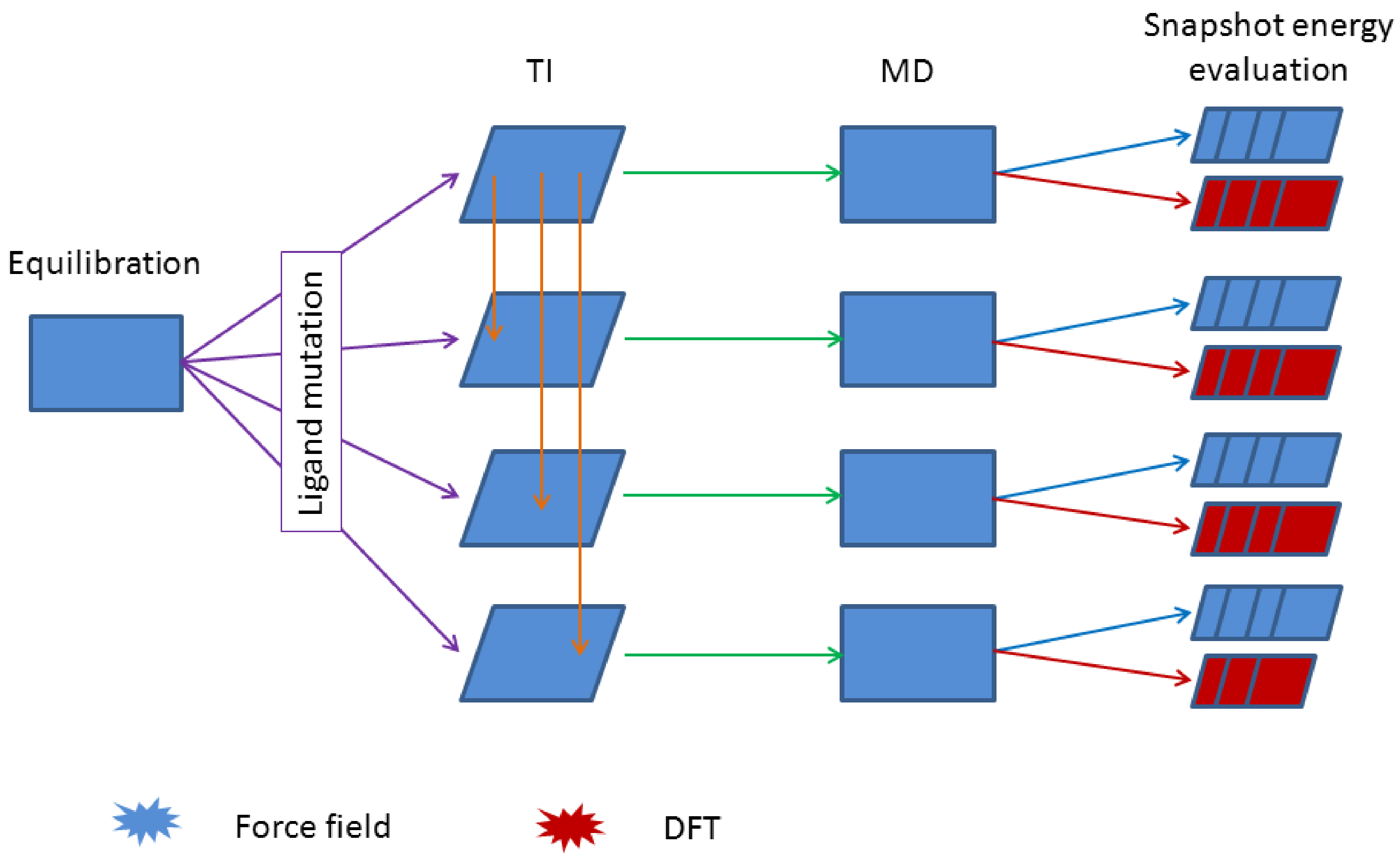


Figure 5.14: Schematic of the workflow of calculations performed in this study.

Chapter 6

Extracting binding information from the electron density and molecular orbitals

When performing large-scale quantum mechanics calculations, the explicit treatment of the electrons allows the calculation of properties dependent on the electronic distribution. Examples include the polarisation of a molecule on binding, charge redistribution, or a breakdown of the contributions to the binding energy into specific interactions.

This chapter will detail two such approaches, Natural bond orbital analysis, Hirshfeld density analysis, and their application to inhibitors of the cGMP-specific Phosphodiesterase type 5 protein.

6.1 Energy decomposition approaches (EDA)

6.1.1 Natural bond orbitals

Molecular orbitals (MOs) obtained from DFT are delocalised over the entire molecule. Since each MO contains two electrons this means that the electrons are also delocalised, and hence do not have any chemical interpretation. A natural bond orbital (NBO) [129, 130, 131] is a localised orbital providing an optimal representation of a chemical bond between two atoms. The NBOs are one of a sequence of natural localised orbital sets that include “Natural Atomic Orbitals” (NAO), “Natural Hybrid Orbitals” (NHO), “Natural Bonding Orbitals” (NBO) and “Natural (semi-)Localised Molecular Orbitals” (NLMO). These natural localised sets are intermediates between atomic orbitals and molecular orbitals.

NBO theory allows the construction of hybrid atomic orbitals which are numerically optimised to give the best description of the chemical environment. These hybrid atomic orbitals are called natural hybrid orbitals (NHOs), they are generated from a linear combination of atomic orbitals (AOs) of the atom on which they are centred. For example, the hybrid orbital on a Carbon pointing towards a Hydrogen, $h_{C_1 \rightarrow H_1}(\mathbf{r})$, would be,

$$h_{C_1 \rightarrow H_1}(\mathbf{r}) = \sum_{i=1}^{N_{C_1}} c_i \chi_i(\mathbf{r}), \tag{6.1}$$

where χ_i are the AOs on the Carbon, and c_i are the coefficients, summed over all atomic orbitals on the Carbon.

Unlike AOs, NHOs are directional and point towards the atom they are bonded to. In traditional chemistry a C-H bond would consist of an ‘sp^3’ hybrid orbital

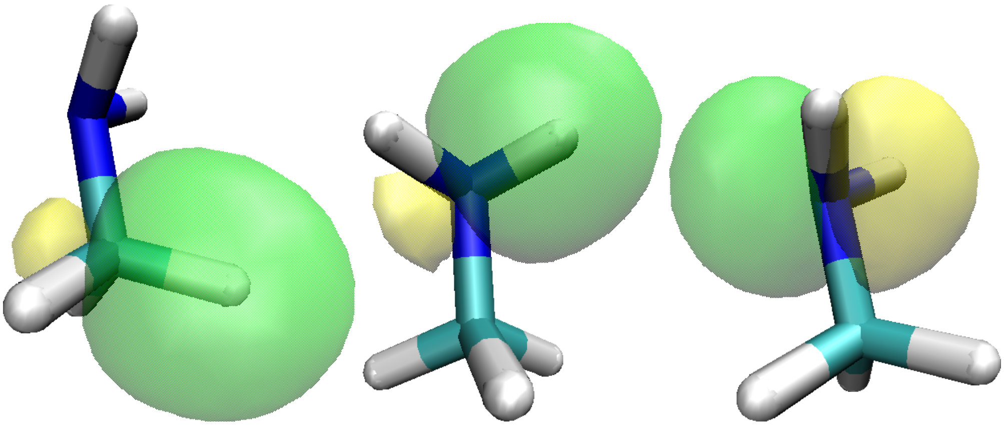


Figure 6.1: Examples of NBOs of methylamine. The left picture shows a C-H bond, where the NHO on the Carbon is an sp$^{2.84}$. The middle shows an N-H bond, where the NHO on the Nitrogen is an sp$^{2.55}$. The right shows a Nitrogen lone pair with an sp$^{6.72}$ hybrid orbital.

based on the Carbon and an ‘s’ AO on the Hydrogen. However, in NBO theory the hybridisation is optimised to take into account the chemical environment of that bond (now derived from c_i), and so moves away from idealised Lewis structure. An example of these hybrid and natural bond orbitals for methylamine are shown in Figure 6.1. The C-H NBO shown on the left in Figure 6.1 is made from an ‘sp$^{2.84}$’ on the Carbon and an ‘s’ on the Hydrogen. The N-H NBO shown in the middle of Figure 6.1 is made from an ‘sp$^{2.55}$’ on the Nitrogen and an ‘s’ on the hydrogen. The electron lone-pair of the Nitrogen shown on the right of Figure 6.1 is a ‘sp$^{6.72}$‘ hybrid orbital.

NBOs are then a Linear combination of these NHOs from only two atoms involved in a chemical bond. For example, the NBO for the C-H bond shown in Figure 6.1 is made by,

$$\Omega(C_1 - H_1) = 0.77843 h_{C_1 \rightarrow H_1}(\mathbf{r}) + 0.6204 h_{H_1 \rightarrow C_1}(\mathbf{r}). \tag{6.2}$$

NBO theory gives the best possible NBOs in the sense that they minimise the energy. They are however, by definition, approximations to the exact orbitals, which are the delocalised MOs. NBO-based properties have been found to converge rapidly to well-defined numerical limits, independently of the basis set used to approximate the wavefunction [129].

Natural Lewis structures

The "Lewis-structure" model [132] of a molecule is the traditional way of viewing a molecule, made up of chemical bonds sharing electron pairs, and lone-pairs of electrons. The description of the electronic structure in terms of NBOs is consistant with the Lewis picture of a chemical bond. The NBO approach produces two valence-shell NBOs: a Lewis-type "in-phase" NBO, and a corresponding non-Lewis "out-of-phase" NBO (which is unoccupied in the Lewis-structure picture). Lewis-type NBOs include one-centre core, lone pair, and two-centre bond orbitals, whilst non-Lewis sets include unoccupied lone pair and Rydberg orbitals as well as valance antibonds.

EDA by 2$^{\text{nd}}$-order perturbation theory involving NBOs

It is possible to obtain quantitative estimates of the strength of specific acceptor-donor NBO interactions. This analysis is carried out by examining all possible interactions between "filled" (donor) Lewis-type NBOs and "empty" (acceptor) non-Lewis NBOs, and estimating their energetic importance by 2nd-order perturbation theory. Since these interactions lead to donation of electrons from the localised NBOs of the idealised Lewis structure into the empty non-Lewis orbitals (and thus, to departures from the idealised Lewis structure description), they are referred to as ”delocalisation” corrections to the zeroth-order natural Lewis structure. For each donor NBO (i) and acceptor NBO (j), the stabilisation energy $E(2)$ associated with delocalisation is estimated as,

$$E(2) = \Delta E_{ij} = q_i \frac{F(i,j)^2}{\epsilon_i - \epsilon_j}, \tag{6.3}$$

where q_i is the donor orbital occupancy, ϵ_i and ϵ_j are diagonal elements (orbital energies), and $F(i,j)$ is the off-diagonal NBO Fock matrix element.

NBO calculations with ONETEP

A recent development has interfaced the ONETEP program with the NBO 5 analysis package [133] in order to perform NBO analysis of large systems containing thousands of atoms [134]. In this approach the NGWFs are transformed into orthogonal natural atomic orbitals (NAOs), the linear combination of which create NHOs, and then finally NBOs which are obtained from the NBO 5 analysis package. Using ONETEP, NBO analysis can be performed within a localised region of the system in such a way that the results are identical to an analysis on the full system. In this manor, interactions in a particular region of chemical interest, such as the active site of a protein, can be investigated, whilst fully accounting for long-range electrostatic effects from the entire system.

6.1.2 Density analysis

The view of molecules as combinations of atoms being held together by chemical bonds is prominent and successful in all fields of chemistry. One of the most straightforward and clear-cut schemes for partitioning the electron-density is the

Hirshfeld population analysis.

The Hirshfeld approach and iterative Hirshfeld approach

This approach was originally proposed to obtain reasonable partial charges on atoms [135]. This was done by generating a "promolecule", which is a sum of the atomic densities of the individual atoms, and using it to determine the contribution of individual atoms to the density in the molecule. Another use of this approach is to find the change in electronic density when atoms (or molecules) combine to gain a qualitative understanding on the electron redistribution.

For a protein-ligand system, the electronic density of a promolecule is defined as,

$$n^{pro}(\mathbf{r}) = n^{rec}(\mathbf{r}) + n^{lig}(\mathbf{r}), \tag{6.4}$$

where $n^{rec}(\mathbf{r})$ is the density of the receptor, and $n^{lig}(\mathbf{r})$ is the density of the ligand. A sharing function is then defined as,

$$w^{X}(\mathbf{r}) = \frac{n^{X}(\mathbf{r})}{n^{pro}(\mathbf{r})}, \tag{6.5}$$

where X denotes either the ligand or receptor density. The density of the bonded fragment is then,

$$n^{b.X}(\mathbf{r}) = w^{X}(\mathbf{r}) n^{com}(\mathbf{r}), \tag{6.6}$$

where $n^{com}(\mathbf{r})$ is the density of the complex. Density deformations can be calculated by,

$$\Delta n^{X}(\mathbf{r}) = n^{b.X}(\mathbf{r}) - n^{X}(\mathbf{r}). \tag{6.7}$$

The charge gain/loss (q) on the fragment can then be calculated by integrating over space,

$$q = \int \Delta n^X(\mathbf{r}) d\mathbf{r}. \tag{6.8}$$

A positive value of q would indictate charge gain upon binding, and a negative value would indicate charge loss.

An issue with the Hirshfeld approach is the arbitrariness in the choice of the promolecule, which effects the partial charges produced by this method. An extension to the Hirshfeld approach to solve this issue is the Iterative Hirshfeld approach [136]. In this approach the aim is to obtain “pro-atom” densities that have the same number of electrons as the atomic partitions in the molecule. This method has been shown to obtain atomic charges that are less basis set dependent than the original approach.

Voronoi deformation density

A similar approach to Hirshfeld’s is the Voronoi deformation density (VDD) [137],

$$Q_A = - \int_{\text{Voronoi cell } A} \Big(n(\mathbf{r}) - \sum_B n_B(\mathbf{r}) \Big) d\mathbf{r}, \tag{6.9}$$

where Q_A is the charge on atom A of the molecule. The Voronoi cell of atom A is defined as the compartment of space bounded by the bond midplanes on and perpendicular to all bond axes between nucleus A and its neighboring nuclei. It is therefore the region of space that is closer to atom A than any other atom. $n(\mathbf{r})$ is the electron density of the molecule and $\sum_B n_B(\mathbf{r})$ is equivalent to the promolecule. Q_A has a straightforward interpretation, it is the amount of charge that flows into ($Q_A < 0$) or out of ($Q_A > 0$) the Voronoi cell of atom A due to chemical interaction in the molecule.

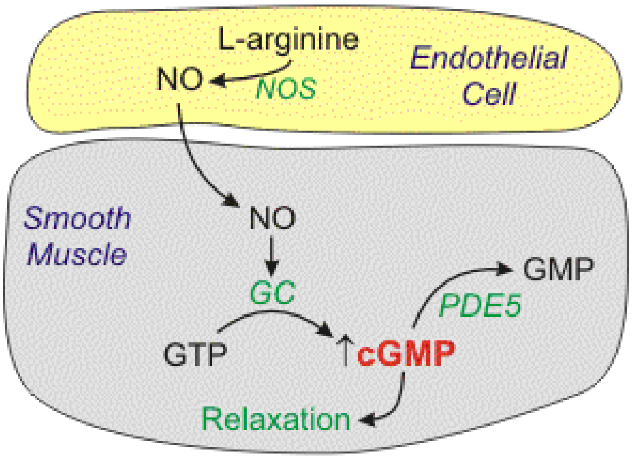
6.2 Phosphodiesterases

Phosphodiesterases (PDEs) comprise a large family of enzymes that catalyse the hydrolysis of cyclic adenosine mono-phosphate (cAMP) or cyclic guanosine monophosphate (cGMP) and are implicated in various diseases. cAMP and cGMP are ubiquitous second messengers that mediate biological responses to a variety of extracellular cues, including hormones, neurotransmitters, chemokines, and cytokines. Increased concentration of these cyclic nucleotides results in the activation of protein kinase A and protein kinase G. These protein kinases phosphorylate a variety of substrates, including transcription factors and ion channels, which regulate a myriad of physiological processes, such as immune responses, cardiac and smooth muscle contraction, visual response, glycogenolysis, platelet aggregation, ion channel conductance, apoptosis, and growth control. There are at least 11 members of the Phosphodiesterase superfamily [138]. Drugs for this family can be non-selective PDE inhibitors such as caffeine and pentoxifylline, or highly specific to a certain PDE isotype, such as Vinpocetine for PDE1, Ibudilast for PDE4, Sildenafil for PDE5, and Papaverine for PDE10.

Card *et al* [139] reported the cocrystal structures of PDE4B, PDE4D, and PDE5A chimera in complex with ten known inhibitors to try to define some of their common and selective features. Through this study they revealed two common features of binding in PDEs which define all known PDE inhibitors. They found that selectivity of inhibitors towards different members of the PDE family can be achieved by exploiting the differences in shape of the hydrophobic binding cavity near the glutamine. The inhibitors generated in this study, of drastically different chemotypes, have a highly conserved binding mode. They share a core binding site that can be characterised by H-bonds to an invariant glutamine, and a planar ring held by a hydrophobic clamp. This clamp is formed by a pair of conserved hydrophobic residues: A phenylalanine with an off-set face to face interaction with the main aromatic ring of the inhibitor, and a valine/leucine/isoleucine that interacts at the centre of the ring from the opposite side (These interaction are depicted in Figure 6.3). They conclude that all inhibitor scaffolds have the same interactions that can be split into three parts: the interactions with the metal ions through a water network, the H-bond interactions with the glutamine, and the hydrophobic interactions with residues lining the pocket. They state that the design of new drugs for PDEs should take advantage of these three interaction. By the addition of new functional groups onto ligand scaffolds that meet these binding criteria, they believe greater selectivity can be introduced for inhibition of specific PDEs.

6.3 PDE5

PDE5 (cGMP-specific Phosphodiesterase type 5) specifically targets cyclic guanosine monophosphate (cGMP), which is a purine second messenger, and is regulated by the synthesis and degradation of GMP. The PDE5 isoform is expressed in smooth muscle tissue, including the rod and cone photoreceptor cells of the retina, but most prominently the corpus cavernosum found in the penis/clitoridis. The inhibition of the PDE5 enzyme causes the concentration of cGMP to increase which leads to a reduction in the amount of calcium, resulting in smooth muscle relaxation, and increased sexual arousal. This cycle is shown in Figure 6.2. The PDE5 inhibitor Sildenafil (Viagra) provides an effective treatment for erectile dysfunction [140]. Sildenafil has also been shown to have positive effects in the



Abbreviations: NO, nitric oxide; NOS, nitric oxide synthase; GC, guanylyl cyclase; PDE5, cGMP-dependent phosphodiesterase (type 5)

Figure 6.2: Cycle of cGMP in smooth muscle [7].

treatment of heart failure, systemic hypertension and Vascular disease.

Despite the clear utility of these compounds, one potential drawback is cross-reactivity with the closely related PDE6 and PDE11. Interest in PDEs as molecular targets of drug action has grown with the development of isozyme-selective PDE inhibitors. These offer potent inhibition of the selected isozymes without the side-effects that can be caused by nonselective inhibitors.

PDE5 is composed of 3 functional domains: an N-terminal cyclin fold domain, a linker helical domain and a C-terminal helical bundle domain. The active site is a deep pocket at the junction of the 3 domains and is lined with highly conserved residues between the different isotypes of PDE. The active site can be split into three pockets; A metal binding pocket, a solvent-filled pocket, and a pocket containing a hydrophobic clamp (VAL250 and PHE 288) and a selective glutamine (GLN 285), shown in Figure 6.3. The metal site is at the wider end of the pocket. It is a binuclear metal centre that contains highly conserved polar and hydrophobic

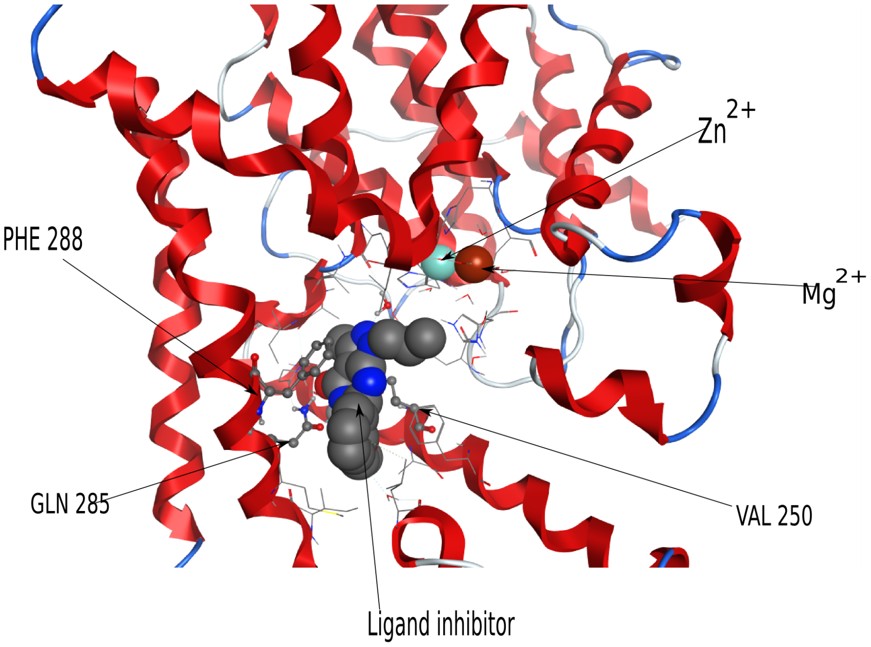


Figure 6.3: PDE5 cavity highlighting the three regions in the binding cavity. A metal binding pocket (Zn^{2+} and Mg^{2+}), a solvent-filled pocket, and a pocket containing a hydrophobic clamp (made from VAL250 and PHE 288) and a selective glutamine (GLN 285).

residues that coordinate to these metal ions. The first metal is a Zinc ion (Zn^{2+}) and the second is a Magnesium ion (Mg^{2+}). The Zinc is coordinated to two histidines, two apartates and two water molecules. The Magnesium is coordinated to five water molecules and one of the aspartates the Zinc binds with. One of the water molecules bridges the two ions and is believed to be a hydroxide ion [141] (HO^{-}).

The pharmacological interest in this target protein has lead to a number of computational studies attempting to gain greater insight into the reaction mechanisms of the protein with its natural substrate, and towards reducing the expense involved in drug discovery projects.

O'Brien *et al* [142] used all-quantum hybrid calculations with ONIOM(B3LYP/6-

31G(d):PM3MM) [142] to accurately describe the interactions of PDE5 with cGMP. They concluded that the preference of cGMP over cAMP for PDE5 is due a number of factors. This preference comes mainly from the fixed orientation of a consevered glutamine reside (Gln 817) together with the fixed orientation of a nonconserved glutamine residue (Gln 877). cGMP has stronger hydrophobic interaction, having a near parallel alignment between Phe 820 and the guanine base suggesting favourable pi stacking, in contrast to cAMP that is off by 23°. The deselection of cAMP is enhanced by an energy penalty that arises due to a steric clashes with Gln 775 and Gln 817 at the back corner of the pocket.

Zagrovic *et al* [143] studied the thermodynamics and mechanics of PDE5 selective inhibitors Sildenafil and Vardenafil. Through the use of molecular dynamics of PDE5 with the inhibitors bound and unbound, they suggest a mechanism in which two loops surrounding the binding pocket execute sizable conformational changes, clamping the ligand into place. They used the GROMOS package for their MD simulations. The complex was solvated with the SPC water model [69] and the GROMOS 45A3 force field [144] was used to describe the system. They noted that there were changes in the coordination's of the divalent ions and changes to the motions of PDE5 when the ligand was bound. They went on to perform thermodynamic integration (TI) and single-step perturbation (SSP) to calculate the relative binding free energies of the two ligands plus demethyl-vadenafil. TI was used to check the quality of the force field for describing the energy of this system. Their TI results accurately predicted the experimental trend in ligand binding affinities. However, the results obtained from SSP were at odds with both the experimental and the TI results, suggesting poor convergence of the SSP approach.

Work on modelling novel tetrahydro-β-carboline derivatives with PDE5 inhibitory and anticancer properties was carried out by Mohamed *et al* [145]. As well as synthesis of the novel compounds, docking was carried out with GOLD. Three independent docking experimenters were carried out using three different scoring functions in GOLD (GoldScore, ChemScore, and ASP) [90, 91]. 10 poses for each compound were generated for each function, then rescored with the other two functions and further scored using DrugScore$^{\mathrm{CSD}}$ [92] and DrugScore$^{\mathrm{PDB}}$. The final score for each pose was a consensus calculated from the mean of the 5 scoring methods. They concluded that docking accurately differentiated between active and inactive analogues and revealed conformational, steric, and lipophilic requirements for potent inhibition of PDE5. Many of the derivatives they found showed some low potency inhibition of the growth of MDA-MB-231 breast tumour cell line.

Niinivehmas *et al* [146] presented an improvement to the negative image-based (NIB) screening approach. PDE5 was chosen as the target protein to improve the method with electrostatic information since the binding site contains both polar groups and coordinated water molecules. The ligand-shape and the protein structure-based virtual high throughout screening (vHTS) methods were compared using this protein. The top 5% of the ranked results were rescored using MM-GBSA to estimate the binding free enegies. MD simulations were ran in AMBER using the ff03 force field to describe the protein, gaff to describe the ligand, and used the TIP3P water model to solvate the system. Starting structures for the ligands were taken from poses generated by GLIDE [147]. They conclude, that with the improvements made to the NIB screening approach, in combination with MM-GBSA, the enrichment is taken to a level that is desired to keep the costs of drug discovery projects within reasonable limits.

The PDE5 inhibitors chosen in this study were first presented by Haning *et al* [148]. They synthetically prepared molecules based on a number of different ligand scaffolds to characterise the structure-activity relationship (SAR) trends. The scaffolds they used were the known 3H-imidazo[5,1-f][1,2,4]triazin-4-ones and pyrazolopyrimidinones, and a new iosomeric imidazo[1,5-a][1,3,5]triazin-4(3H)-ones, which they identified as a new PDE5 inhibitor with oral efficacy.

6.3.1 MD simulations

The 1XP0 crystal structure of the PDE5 complex was checked and protonated in the MOE [104] program. MD simulations were carried out using the AMBER10 [54] package, with the ff99SB [105] forcefield used for the protein and the generalised AMBER forcefield [59] (gaff) used to model the ligands, the metal ions and the hydroxide ion. The gaff parameters for the Zn^{2+} ion where obtained through personal correspondence [149]. Ligand charges were calculated with the AM1-BCC method with the antechamber tool in AMBER10. The charges of the metals and the hydroxide where set to their formal charges, +2 and -1 respectively. The system was explicitly solvated in the TIP4P water model [70], keeping all crystallographic waters. This resulted in a total of 13333 water moleclues in the system.

The system was equilibrated using the following protocol. Hydrogens were relaxed with restraints placed on all heavy atoms in the complex and solvent, before relaxing the solvent with restaints only on the complex. The system was heated to 300 K over 200 ps, still restraining the heavy atoms of the complex, with the NVT ensemble. Then ran for a further 200 ps with the NPT ensemble at 300 K in order to equilibrate the solvent density. This was cooled over 100 ps to 100 K and

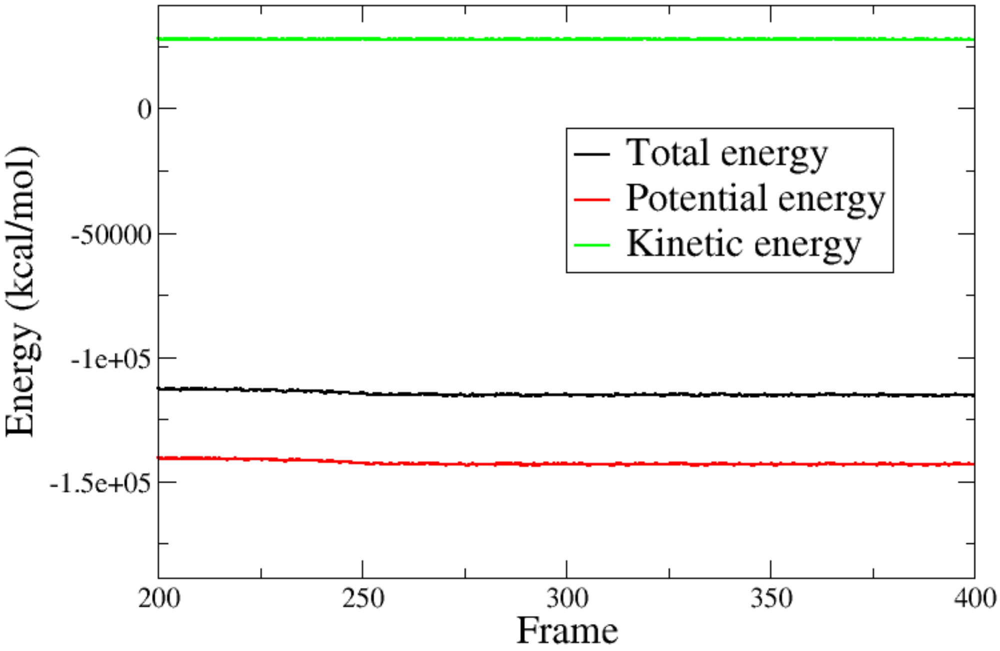


Figure 6.4: PDE5 energies from the last equilibration step. Frames recorded every 0.5 ps.

a number of relaxations were ran, reducing the restraints on the heavy atoms in stages $(1000, 500, 100, 50, 20, 10, 5, 2, 1, 0.5$ kcal mol^{-1}Å$^{-2})$. Finally the system was reheated to 300 K with no restraints over 200 ps and then for a further 200 ps at 300 K with the NPT ensemble. At the end of this it was confirmed that the water density in the box was stable and energies converged to oscillate about a constant value (Figure 6.4). MD simulations used the Langevin thermostat [67], the particle mesh Ewald sum (PME) for the electrostatic interactions and the SHAKE algorithm [65] to constrain hydrogen-containing bonds allowing a time-step of 2 fs.

This system is more complicated than the T4 lysozyme previously mentioned. It is over twice the size containing almost 5800 atoms. It has two metal ions in the binding cavity and a host of structurally important water molecules that must be accounted for in the MD simulation, one of which is a hydroxide ion bridging the metal ions. A production MD simulation was ran for 50 ns in the NPT ensemble using the Langevin thermostat, PME, and a timestep of 2 fs.

During the course of the 50 ns MD simulation the ligand was seen to move very little. The two residues that make the hydrophobic clamp (Phe288 and Val250) also showed very little movement. The RMSD of Phe288 stays around a value of 0.4 Å form the first frame (Figure 6.5). Val250 has two distinct structures shown by the RMSD, one at around 1 Å and another at around 1.5 Å (Figure 6.6). These correspond to a swap of the γ_1 and γ_2 Carbons, having very little change in the actually position of the side chain. The Gln285 residue, in which two hydrogen bonds are formed with the ligand, showed more movement than the other important pocket residues. It has two distinct structures, one with an RMSD of around 0.3 Å, and the other with an RMSD of around 1.1 Å, as shown in Figure 6.7. These two structures show the C=O and the N-H of the glutamine both pointing at the ligand, generating two hydrogen bonds, and a switch of side chain that points the C=O in the opposite direction, leaving the NH_2 group pointing at the ligand. This is shown in Figure 6.8.

During the simulation the interaction of the Nitrogen lone pair with a water molecule remains fairly constant. However, the water molecule involved in the interaction is not constant. The water molecules seem to be quite free to move in and out of this end of the cavity, although there is always a water molecule in this position. This might suggest that this water molecule could be easily displaced with an additional mutation of the ligand to take advantage of the polar interactions that a

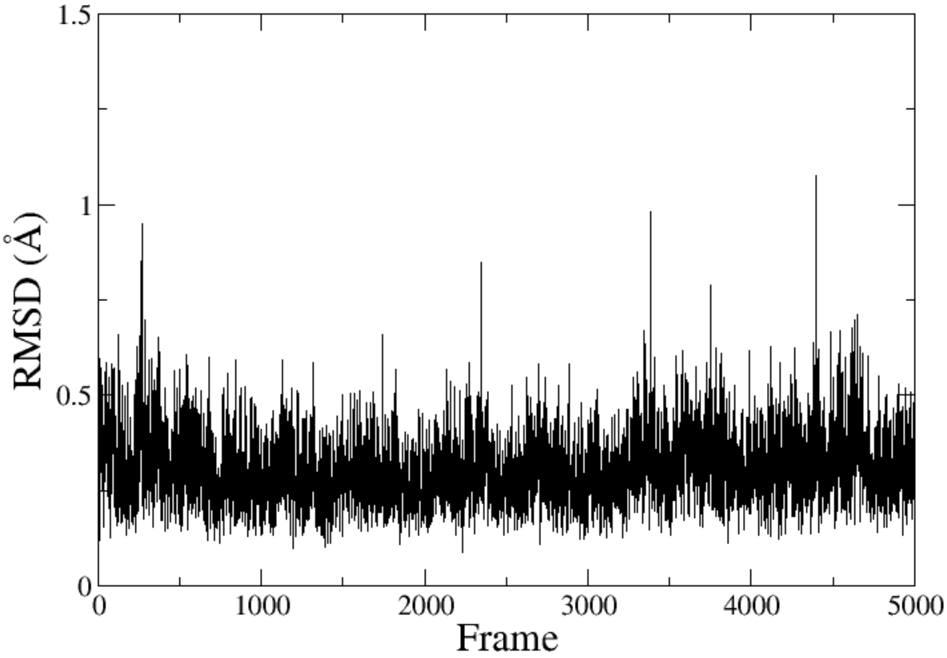


Figure 6.5: RMSD of Phe288. Frames recorded every 10 ps.

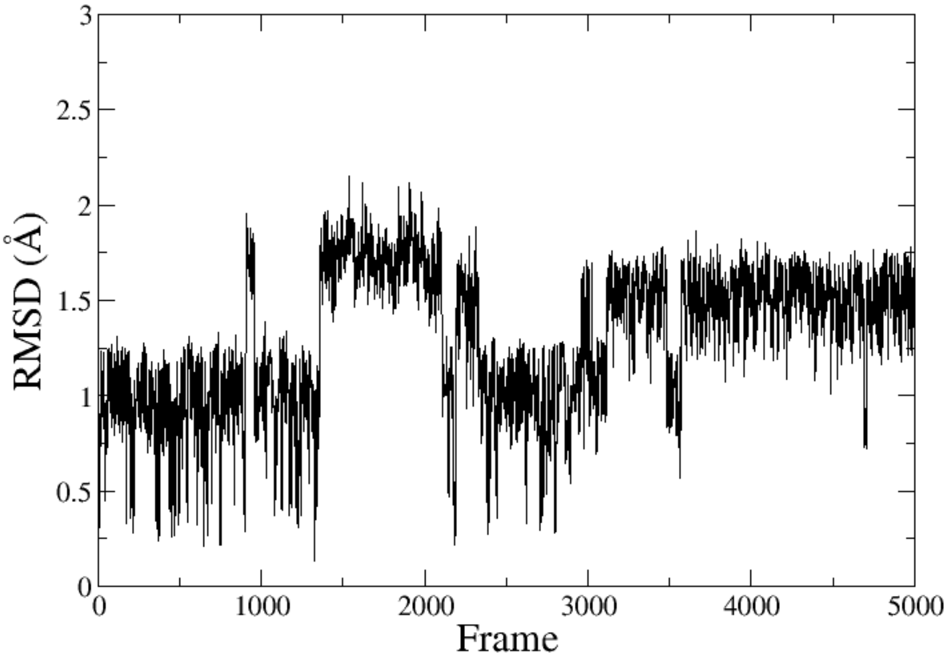


Figure 6.6: RMSD of Val250. Frames recorded every 10 ps.

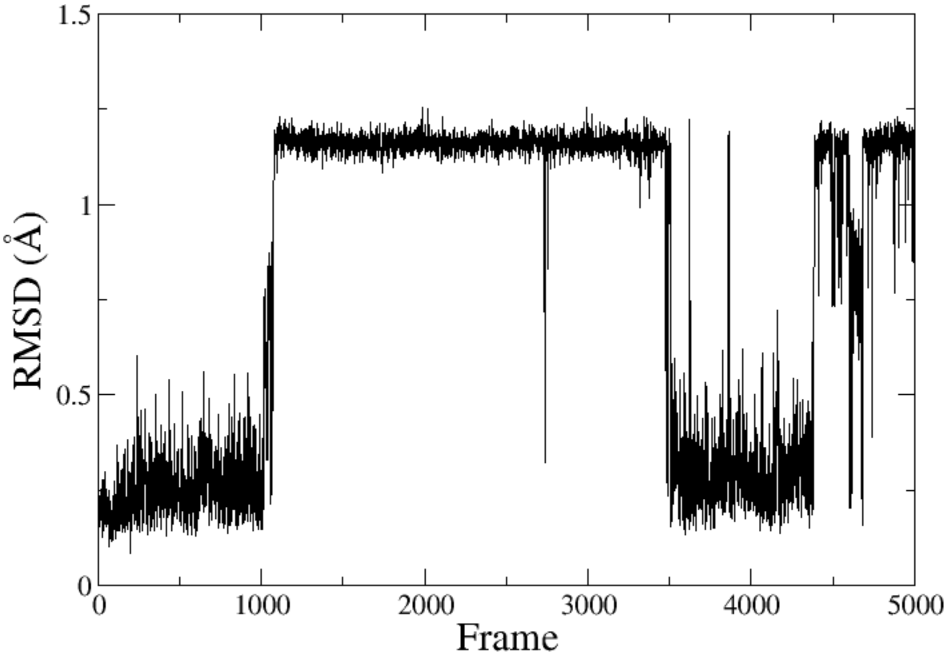


Figure 6.7: RMSD of Gln285. Frames recorded every 10 ps.

water in this position benefits from.

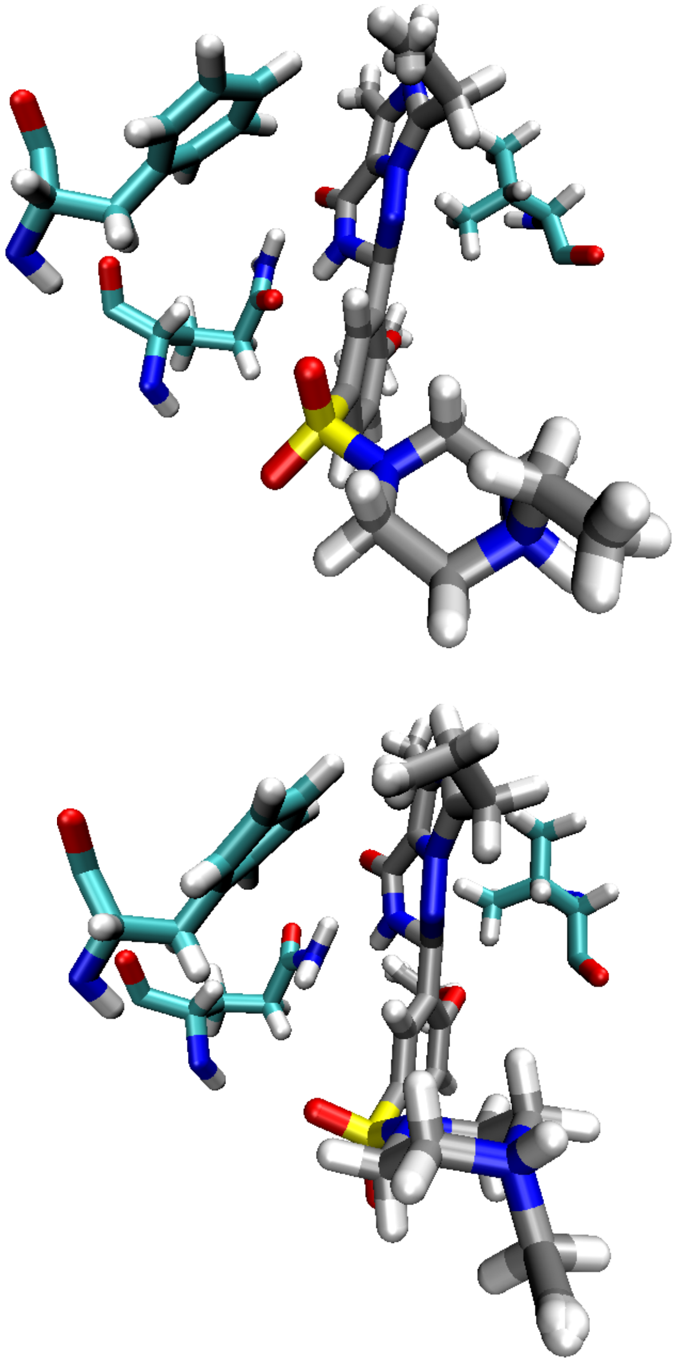


Figure 6.8: The ligand and interacting residues of the protein, showing the two orientations of Gln285. The ligand Carbons are coloured grey and the protein residue Carbons in cyan.

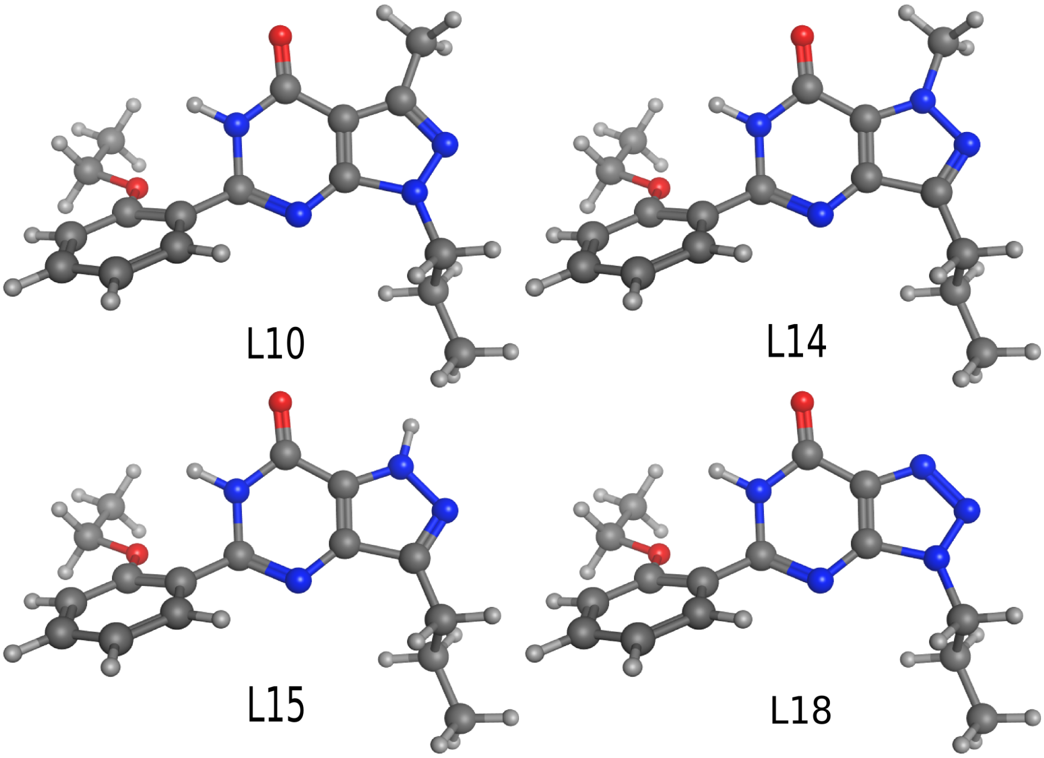


Figure 6.9: PDE5 inhibitors based on a 2-ethoxyphenyl heterocyclic scaffold. Top left: L10 - IC_{50} 1 nM. Top right: L14 - IC_{50} 27 nM. Bottom left: L15 - IC_{50} 50 nM. Bottom right: L18 - IC_{50} 300 nM.

6.4 Binding interactions in PDE5

Four ligands were selected from the ligands described by Hanning *et al* [148] based on a 2-ethoxyphenyl hetrocyclic scaffold. They were chosen for their range of IC_{50} values (varying from 1 nM to 300 nM) and their similar structures, and are displayed in Figure 6.9. The ligand labels have been taken from the numbering system in Table 1 of the Hanning *et al* paper.

6.4.1 DFT calculation set up

For the NBO and electron density analysis calculations initial structures were taken from the crystal structure of 1XP0. The PDB was checked and protonated with the MOE [104] program and side chains and ligand relaxed using the MMFF94 force field [60]. From this relaxed structure the ligand was manually mutated into the other ligands, keeping all common atoms in identical positions. ONETEP calculations used a kinetic energy cut-off of 800 eV with the PBE exchange-correlation functional. NGWF radii were 7.0 a_0 for all atoms, with 1 NGWF for Hydrogen, 4 NGWFs for Carbon, Oxygen, Nitrogen and Magnesium, and 9 NGWFs for Sulphur and Zinc. The atomic solver in ONETEP [150] was used to generate better starting NGWFs for the metal ions and standard STOs where used to initialise the NGWFs for all other atoms. Although the entire system was included in the calculation, the NBOs were generated only for the atoms of a subsystem containing the active site (pocket residues and ligand).

NBO calculations were also carried out using the GAMESS-UK program. This was performed on a much smaller system containing only the ligand and an important water molecule hydrogen bonding to the 5 membered hetrocyclic ring the mutations occur in (shown in Figure 6.10). The GAMESS-UK calculations were done at PBE/TVZP and B3LYP/TVZP levels of theory.

6.4.2 Results

Hirshfeld density analysis was performed on the four ligands bound in the pocket. Fig. 6.11 shows density deformations of ligand 14 and the receptor over the entire system, and focused on the pocket, showing the hydrogen bonds in green. This

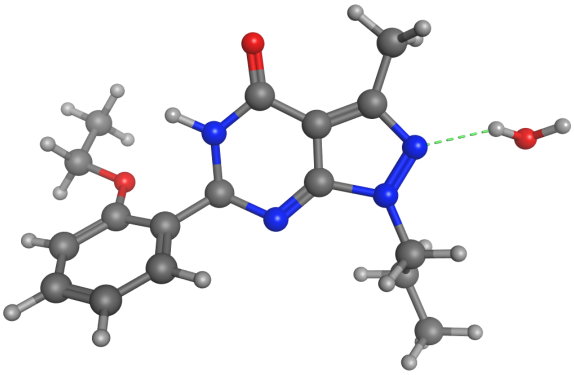


Figure 6.10: Minimal system for NBO calculations with GAMESS-UK (example shown is for L10).

depicts the charge redistribution when the ligand binds, showing the extent of polarisation that the ligand and receptor experience. It is interesting to note that this redistribution of charge in not localised to the pocket but is seen on peripheral charged residues, some at a distance from the binding cavity greater than 10 Å.

Integrating the density deformations can give a qualitative insight of charge transfer that occurs on ligand binding. The unbound ligands have a charge of 0. The charges displayed in Table 6.1 suggest that the ligands are, overall, electron acceptors (an increase in the total number of electrons on the ligand). The structural differences between these ligands occur around the H-bond between the water and the lone-pair of electrons on the common Nitrogen in the 5 membered hetrocyclic ring. Assuming all other interactions are the same between the different ligands and the receptor, we can focus on just this interaction. For this H-bond the ligand acts as an electron donor (giving electrons). Since overall the ligand is an electron acceptor (gaining electrons), the stronger this H-bond, the lower the integrated density deformations (charge of the ligand) would be. This is due to more charge being donated from the ligand to the receptor, and this is indeed the trend we ob-

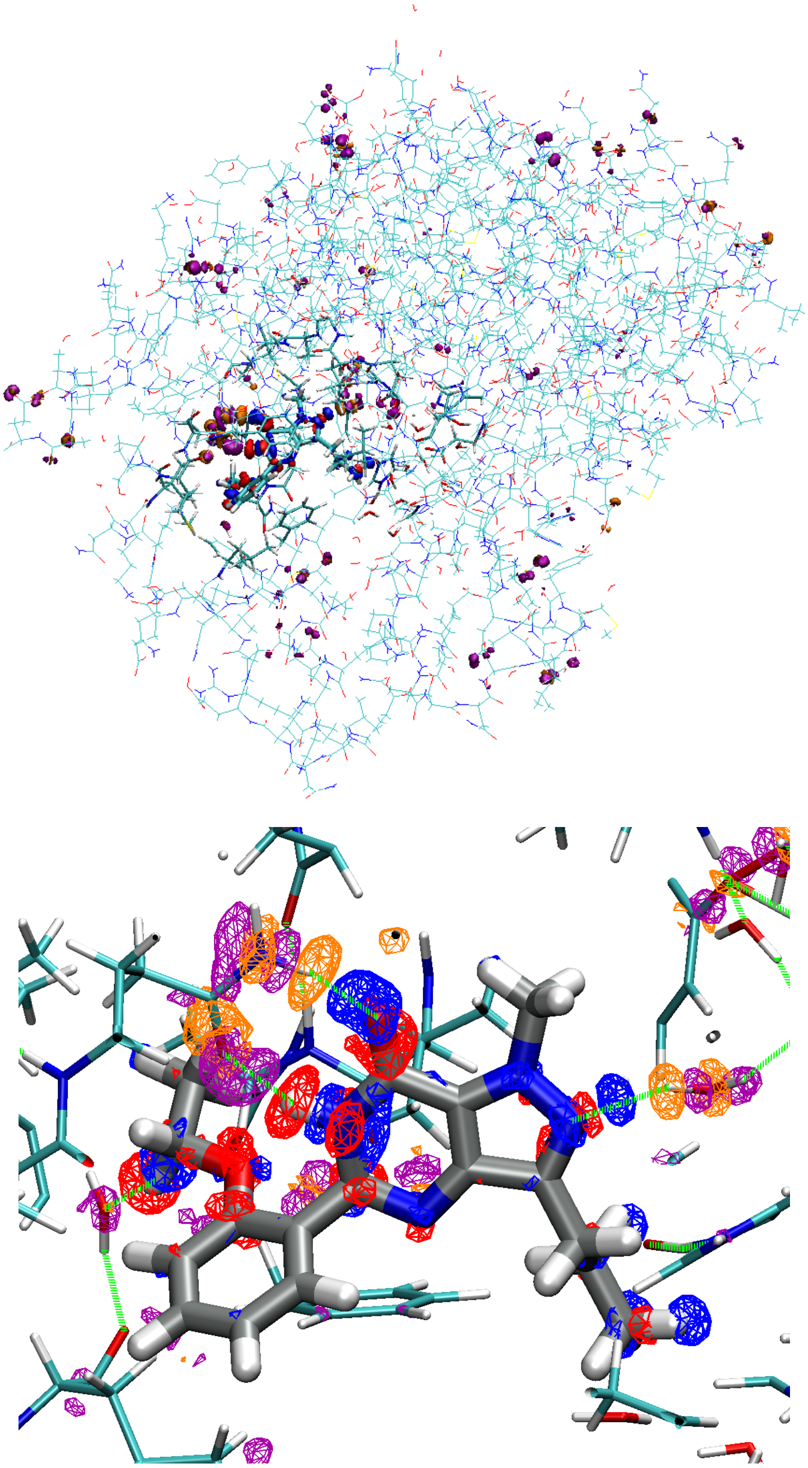


Figure 6.11: Density denformation on the ligand and receptor (example is for L14 bound in the cavity). The top panel shows the deformation density over the entire protein. The bottom panel shows a close up of the pocket.

Table 6.1: Intergrated density deformations giving a qualitative overview of electron gain/lose on the ligand (positive is electron gain).

Ligand	L10	L14	L15	L18
$\int \Delta n^{lig}(\mathbf{r})d\mathbf{r}$	0.016	0.017	0.029	0.042

Table 6.2: Strengths of the hydrogen bond between the water molecule and the Nitrogen lone pair in the ligands (LigN–HOH). Energies in kcal mol^{-1}.

Ligand	L10	L14	L15	L18
GAMESS-UK (B3LYP/TZVP)	5.0	5.2	5.1	4.3
GAMESS-UK (PBE/TZVP)	4.7	4.9	4.9	4.0
ONETEP (PBE / 800 eV)	13.7	14.0	13.7	11.9
ONETEP (PBE / 1200 eV)	13.8	14.1	13.8	12.0

serve in Table 6.1; with the strongest binder, having the smallest charge, and the weakest binder, having the largest charge.

NBO calculations were first performed on the minimal system of the ligand hydrogen bonding with a water molecule, as shown in Figure 6.10. The strength of the interaction of the nitrogen lone-pair with the H-O anti-bonding orbital, shown in Figure 6.12, was obtained by the 2$^{\text{nd}}$ order perturbation estimate and are presented in Table 6.2. The results from GAMESS-UK are fairly independent of the exchange-correlation functional, or the basis set used in the ONETEP calculation. We see a reasonable agreement with the more qualitative overview from the Hirshfeld analysis, with the weakest binder having the weakest bond. The results from GAMESS-UK are closer to the expected strength of a hydrogen bond, at around 5 kcal mol^{-1}, however ONETEP shows an excellent qualitative correlation with these results, with an R^2 value of 0.97 between the GAMES-UK PBE/TZVP results and the ONETEP PBE/800 eV results.

The strength of the other hydrogen bonds between the ligand and the protein were obtained with ONETEP only, since the entire system is far too large to use

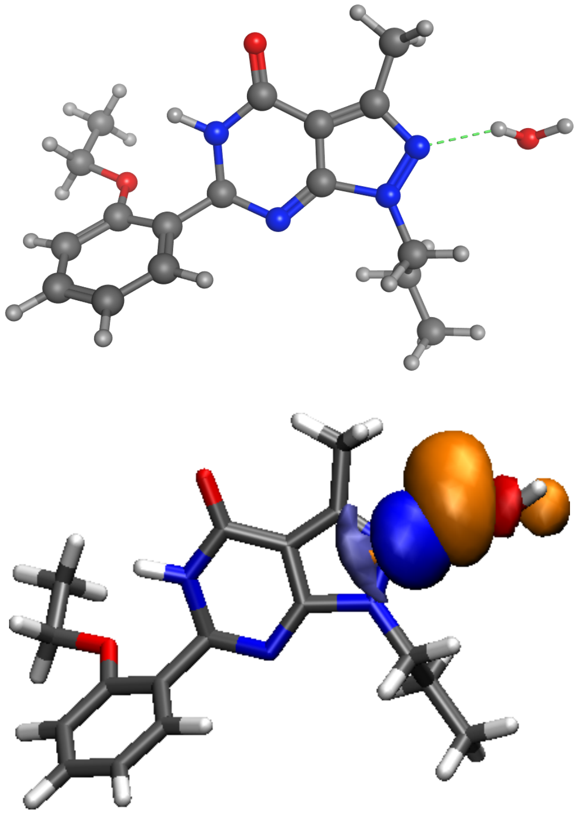


Figure 6.12: NBOs of the hydrogen bond between the ligand and a water molecule, as obtained form ONETEP calculations. The blue isosurface indicates the electron donor: the lone pair on the Nitrogen. The orange and red isosurfaces indicate the electron acceptor: the H-O antibonding NBO.

Table 6.3: Estimated E(2) energies of the hydrogen bonds between the ligands and the receptor. Energies in kcal mol^{-1}.

Interaction	L10	L14	L15	L18
LigN—HOH	14.0	14.2	13.9	12.4
LigO—H-N(GLN)	17.9	17.7	17.9	18.2
LigN-H—O(GLN)	16.4	16.4	16.5	16.4

GAMESS-UK (5799 atoms). Figure 6.13 displays the electron density of the entire 5799 atom system and the NBOs for the hydrogen bonds between L10 and the PDE5 cavity. The strengths of these interactions are presented in Table 6.3. When the rest of the protein is taken into account when the NBOs are generated, the strengths of the hydrogen bond between the Nitrogen lone-pair and the water molecule increase by a small amount (less than 0.3 kcal mol^{-1}), but retain the same trend. The other hydrogen bonds between the ligand and the receptor (the Gln 285 shown in Figure 6.3) have very similar strengths for the four ligands, varying by a maximum of 0.5 kcal mol^{-1}. In contrast to the first hydrogen bond that varies by up to 1.8 kcal mol^{-1}.

To estimate ligand binding affinities we need more than just a measure of the strength of the polar interaction in the pocket. We also require an estimate of the hydrophobic interactions, the desolvation energies of the ligand, and the entropies of binding. Since the ligands in this study are so structurally similar, entropy can be assumed to cancel. Hydrophobic interactions are estimated by calculating the binding energy of the empirical dispersion correction in ONETEP [5]. The solvation energies of the four ligands were calculated using our implicit solvent model [78] in ONETEP. These results, along with the H-bond strengths for the Nitrogen–water interaction from Table 6.3, and the experimentally derived IC$_{50}$ values, are displayed in Table 6.4. The IC$_{50}$ value is a measure of the effectiveness of a compound in inhibiting biological function. The value represents the amount

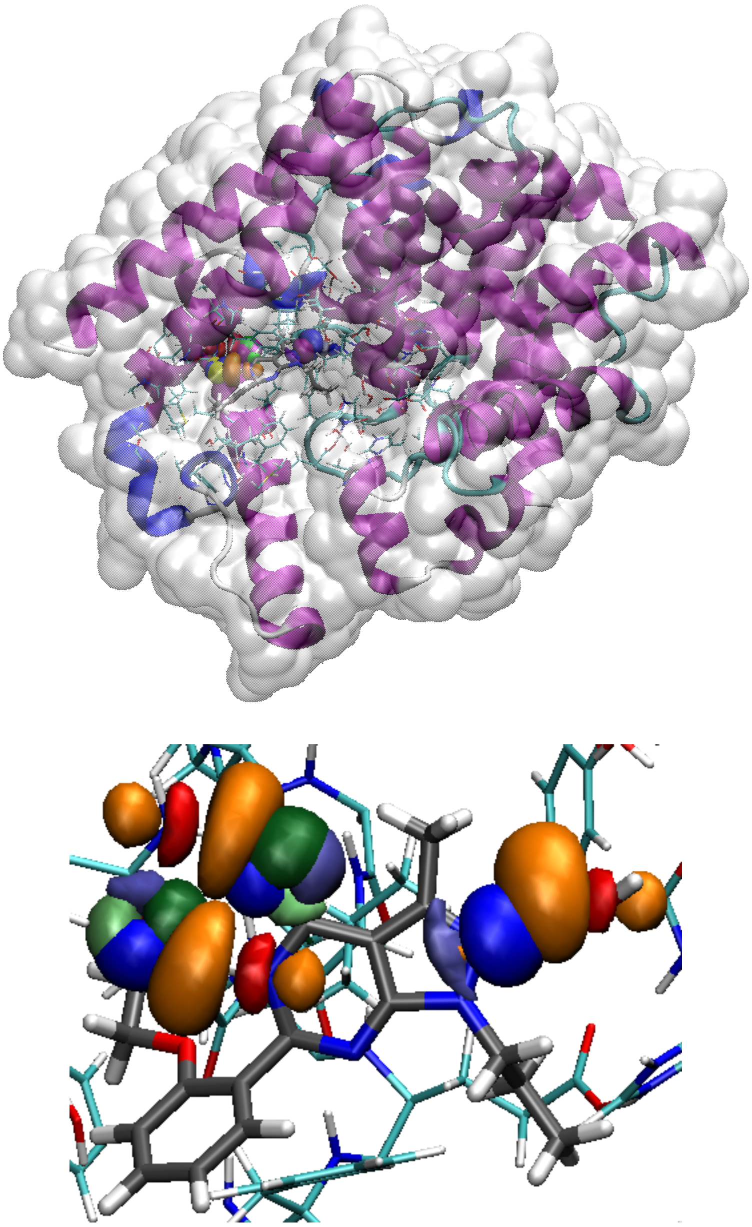


Figure 6.13: NBOs of the hydrogen bonds between the ligand and the receptor. In the top picture the protein structure is shown as cartoon, with the subregion in which the NBOs were calculated shown as lines, with coloured isosurfaces indicating some of the NBOs. The transparent contour is the electron density of the entire system as obtained with ONETEP. The bottom picture shows the NBOs involved in the hydrogen bonds between the ligand and the glutamine residue, and the ligand and the water molecule. The blue and green isosurface indicates the electron donor: the lone pairs on the Oxygen. The orange and red isosurfaces indicate the electron acceptor: the H-N antibonding NBO.

Table 6.4: Solvation energies (kcal mol^{-1}), E(2) (kcal mol^{-1}) and IC$_{50}$ values (μM) of ligands.

Ligand	L10	L14	L15	L18
Solvation energy of ligand	-5.2	-5.0	-8.6	-11.6
$\Delta E_{\text{dispersion}}$	-49.8	-49.8	-45.9	-45.0
E(2) LigN–HOH	14.0	14.2	13.9	12.4
IC$_{50}$[148]	1	27	50	300

of a drug needed to inhibit the target by 50%. The smaller this value the more effective the drug.

The hydrophobic interactions, measured by $\Delta E_{\text{dispersion}}$, for ligands 10 and 14 are the same strength, while ligand 15 is weaker by 3.9 kcal mol^{-1}, and ligand 18 is weaker by 4.9 kcal mol^{-1}. The change in hydrophobic interaction strength could be from the removal of the methyl at the top of the 5 membered hetrocyclic ring. In the case of ligand 15 this is mutated to a Hydrogen, whereas for ligand 18 this is removed entirely. This would suggest some additional favourable interactions from this site. Looking at the solvation energies, L10 and L14 are very similar, however, L15 has a solvation energy that is larger by 3.6 kcal mol^{-1}, and L18 is much larger again with a value of -11.6 kcal mol^{-1}, 6.6 kcal mol^{-1} larger.

Using a combination of the deformation densities (with a chemical interpretation of the results), the strengths of the individual H-bond interactions that the ligands have in the pocket, as well as the hydrophobic interactions, and the ligand solvation energies, expected trends can be derived for ligand relative binding affinities. For these four ligands, with the amount of data we have collected, the predicted ligand binding affinities would be ranked,

$$L10 \sim L14 > L15 >> L18,$$

which agrees well with the experimental IC_{50} values in table 6.4.

These calculations were only performed on a static structure, with side-chains and ligand position relaxed from the crystal structure using the MMFF94 force field in MOE. To study the change of these interactions through time, MD simulations must be ran and multiple conformations analysed at different points in time.

6.4.3 Conclusions

We ran large-scale QM calculations on the PDE5 protein with 4 different ligands bound in the cavity. Calculations were started from the relaxed crystal structure of PDB 1XP0 and the ligand was mutated into three very chemically similar ligands. The electron density was analysed using the Hirshfeld approach to gain a qualitative measure of the interactions and charge redistributions when a ligand binds. This showed charge redistribution across the entire protein when the ligand bound, not just localised to the binding cavity. We also gained a qualitative view of the charge transfer upon ligand binding, with charges calculated on the bound ligands indicating that all the ligands always gain charge when they bind. To gain a more quantitative understanding of the polar interactions of the ligands with the receptor, NBOs were generated and 2^{nd} order perturbation estimates of the hydrogen bond strengths were performed. The ligands all made two hydrogen bonds with Gln285 in the cavity, these were found to have equivalent strengths for the four ligands. The differences were most evident for the hydrogen bond made between the lone pair on the common Nitrogen of the hetrocyclic ring where the ligand mutations occurred, and a water molecule. For this hydrogen bond, we saw that the weakest binder also had the weakest bond, smaller by 1.8 kcal/mol. Looking at the hydrophobic binding energy, as calculated in ONETEP, L15 and

L18, where a methyl group has been removed, have smaller binding energies. Reduced by 3.9 kcal mol^{-1} when a Hydrogen is present instead of the methyl, and reduced by 4.9 kcal mol^{-1} when nothing replaces the methyl. The calculation of the ligand solvation energies (calculated in ONETEP) further distinguished the trends in ligand binding affinity, with the weakest experimental binder having the largest computed desolvation energy. Using the data acquired from these two approaches, in combination with the binding dispersion energies and the ligand solvation energies, the computed ligand rankings reproduced the experimentally observed ligand rankings.

In this study only a small test set was used, however, these preliminary results look quite promising. Overall this method gives a qualitative view of the ligand binding affinities, but has the potential to lead to a greater understanding and an explanation of the trends observed. Most importantly these methods are based on *ab initio* calculations and were able to give us information on the PDE example without any prior empirical parameterisation. Therefore we expect that this approach is fully transferable to any biomolecular system, and as a result, could be very useful towards drug design and optimisation.

Chapter 7

Conclusions

This research project was aimed at utilising large-scale quantum mechanics simulations to study protein-ligand interactions using the linear-scaling density functional code ONETEP in combination with conformation sampling from MM methods. The applications of these approaches to various problems have been described in Chapters 4, 5, and 6.

From this work we have shown that large-scale DFT calculations can be used to obtain information on protein-ligand interactions. We have performed the first large-scale full QM-PBSA calculations for protein-ligand binding free energy estimation. While the QM-PBSA approach is more rigorous than MM-PBSA, in the sense that interaction energies are obtained from a calculation that explicitly includes electronic polarisation, they appear to over-estimate the binding energies in vacuum, which results in errors in the free energies of binding in solution. Although, as we have seen, ligand solvation energies are much more accurately computed via the QM implicit solvation model. Thus, better MM-PBSA results seem to come from fortuitous error cancellation between the overbound ligands and overestimated solvation energies. The overbinding from the QM calculations could be overcome in future studies by using a more accurate exchange correlation functional, such as for example hybrid functionals and/or a functional which explicitly includes dispersion interactions. Due to the high cost of the quantum calculations, and the limited computer resources we currently have, the ensemble of structures we used here is not large enough to obtain converged results, but the trends we see when comparing approaches are converged. The types of ligands and protein considered here were common enough to be well-described by the force field so the MM-PBSA approach performs very well. However, this QM-PBSA method would be expected to perform better than MM-PBSA approaches on systems which force fields would not describe very well (e.g. ligands with unconventional functional groups).

The QM corrected thermodynamic integration calculations gave better relative hydration energies compared with the experimental values than standard TI. Although an improvement towards the experimental relative hydration energy is the ideal result, this approach (in the limit of infinite sampling) is actually obtaining a relative hydration energy that would have been obtained if the mutation had been performed completely in a QM description. Errors seen in the trends compared to experiment from these approaches appear to be mainly due to the inherent approximations in DFT, such as the exchange-correlation functional chosen. For the case of our DFT calculations, improvements could be made by using a higher quality exchange-correlation functional.

As well as more accurate energies, there are also other advantages of large-scale quantum calculations that have been explored in this work. Such as the ability to visualise natural bond orbitals (NBOs), densities, and potentials that are responsible for specific interactions and the quantitative estimation of these interactions with energy decomposition approaches. Such approaches were applied to the phosphodiesterase type 5 protein. The results from the NBO analysis of the hydrogen bond interactions of the ligands with the protein, and the subsequent energy decomposition using a second order perturbation estimate, were used in combination with the binding dispersion energies and the ligand solvation energies. The computed ligand rankings reproduced the experimentally observed ligand rankings. Overall this method gives a qualitative view of the ligand binding affinities, and leads to an understanding and explanation of the trend seen in ligand binding.

This work, to our knowledge, has presented:

- The first application of large-scale full QM-PBSA calculations on a protein-ligand system.
- The first QM EE calculations using hundreds of atoms in the QM region combined with an extended QM corrected free energy cycle for the prediction and improvement of relative hydration energies.
- The first use of large-scale QM calculations to generate NBOs of protein-ligand interactions and full protein-ligand complex electron density analysis.

The application of large-scale QM calculations to the three problems has shown great potential for an increased understanding and increased accuracy in the computational prediction of protein-ligand binding affinities. But most importantly, these methods are based on *ab initio* calculations and were able to give us information on the protein-ligand examples without any prior empirical parameterisation.

We therefore expect that these approaches are fully transferable to any biomolecular system, and as a result, could be very useful tools towards drug design and optimisation.

“always be able to kill your students”

Soke Hatsumi Massaki

Bibliography

[1] L. Heady. *Inhibiting CDK2: A Computational Study with Ab-Initio and Classical Methods*. PhD thesis, Trinity College, Cambridge, 2005.

[2] C.-K. Skylaris. Chem3023: Lecture 3.

[3] M. P. Teter, D. C. Allan, T. A. Arias, J. D. Joannopoulos, and M. C. Payne. Iterative minimisation techniques for the bi initio total-energy calculations - molecular-dynamics and conjugate gradients. *Rev. Mod. Phys.*, 64(4):1045–1097, 1992.

[4] S. Fox, C. Pittock, T. Fox, C. Tautermann, N. Malcolm, and C.-K. Skylaris. Electrostatic embedding in large-scale first principles quantum mechanical calculations on biomolecules. *J. Chem. Phys.*, 135:224107, 2011.

[5] Q. Hill and C.-K. Skylaris. Including dispersion interactions in the ONETEP program for linear-scaling density functional theory calculations. *Proc. R. Soc. A.*, 465:669–683, 2009.

[6] A. R. Leach. *Molecular Modelling: Principles and Application*. Prentice Hall, 2 edition edition, 2001.

[7] http://cvpharmacology.com/vasodilator/pdei.htm.

[8] D. E. Koshland. Application of a theory of enzyme specificity to protein synthesis. *Proc. Natl. Acad. Sci. USA*, 1958(44):98–104, 1958.

[9] C. J. Tsai, S. Kumar, B. Ma, and R. Nussinov. Folding funnels, binding funnels, and protein function. *Protein Sci.*, 8:1181–1190, 1999.

[10] M. Gerstein and W. Kerbs. A database of macromolecular motions. *Nucleic Acids Res*, 26:4280–4290, 1998.

[11] C.-S. Goh, D. Milburn, and M. Gerstein. Conformational changes associated with protein-protein interactions. *Curr. Opin. Struct. Biol.*, 14:104–109, 2004.

[12] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, and C. Simmerling. Comparison of multiple amber force fields and development of improved. *Proteins*, 65:712–725, 2006.

[13] S. J. Marrink, A. H. de Vries, and A. E. Mark. Coarse grained model for semiquantitative lipid simulations. *J. Phys. Chem. B.*, 108(2):750–760, 2004.

[14] R. A. Friesner and B. D. Dunietz. Large-scale ab initio quantum chemical calculations on biological systems. *Acc. Chem. Res.*, 34(5):351–358, 2001.

[15] N. Diaz, D. Suárez, K. M. Merz, and T. L. Sordo. Molecular dynamics simulations of the tem-1 β-lactamase complexed with cephalothi. *J. Med. Chem.*, 48:780–791, 2005.

[16] P. A. Bash, M. Karplus, and M. J. Field. A combined quantum mechanical and molecular mechanical potential for molecular dynamics simulations. *J. Comp. Chem.*, 11:700, 1990.

[17] M. Robinson and P. D. Haynes. Dynamical effects in ab initio NMR calculations: Classical force fields fitted to quantum forces. *J. Chem. Phys.*, 133:084109, 2010.

[18] I. Halperin, B. Y. Ma, H. Wolfson, and R. Nussinov. Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins-Structure Function and Genetics.*, 47(4):409–443, 2002.

[19] I. V. Khavrutskii and A. Wallqvist. Computing relative free energies of solvation using single reference thermodynamic integration augmented with hamiltonian replica exchange. *J. Chem. Theory. Comput.*, 6:3427–3441, 2010.

[20] S. Huo, I. Massova, and P. A. Kollman. Computational alanine scanning of the 1:1 human growth hormone-receptor complex. *J. Comp. Chem.*, 23(1):15–27, 2002.

[21] W. Kohn and L. J. Sham. Self-consistent equations including exchange and correlation effects. *Phys. Rev.*, 140(4A):A1133–A1138, 1965.

[22] S. C. L. Kamerlin, M. Haranczyk, and A. Warshel. Progress in Ab Initio QM/MM Free-Energy Simulations of Electrostatic Energies in Proteins: Accelerated QM/MM Studies of pK(a), Redox Reactions and Solvation Free Energies. *J. Phys. Chem. B.*, 113(5):1253–1272, 2009.

[23] D. Riccardi, P. Schaefer, Y. Yang, H. B. Yu, N. Ghosh, X. Prat-Resina, P. Konig, G. H. Li, D. G. Xu, H. Guo, M. Elstner, and Q. Cui. Development of effective quantum mechanical/molecular mechanical (QM/MM) methods for complex biological processes. *J. Phys. Chem. B.*, 110(13):6458–6469, 2006.

[24] M. J. S. Dewar, E. G. Zoebisch, E. F. Healy, and J. J. P. Stewart. Development and use of quantum mechanical molecular models. AM1: a new general purpose quantum mechanical molecular model. *J. Am. Chem. Soc.*, 107(13):3902–3909, 1985.

[25] W. Heisenburg. Uber den anschaulichen inhalt der quantentheoretischen kinematik und mechanik. *Zeitschrift fur Physik*, 43:172–198, 1927.

[26] E. Schrödinger. An undulatory theory of the mechanics of atoms and molecules. *Phys. Rev.*, 28(6):1049–1070, 1926.

[27] M. Born and R. Oppenheimer. Zur quantentheorie der molekeln. *Annalen der Physik*, 389:457–464, 1927.

[28] P. J. Mohr, B. N. Taylor, and D. B Newell. Codata recommended values of the fundamental physical constants. *Rev. Mod. Phys.*, 80:633–730, 2006.

[29] A. Szabo and N. S. Ostlund. *Modern Quantum Chemistry*. Dover, 1996.

[30] P. Hohenberg and W. Kohn. Inhomogeneous electron gas. *Phys. Rev.*, 136(3B):B864–B871, 1964.

[31] D. M. Ceperley and B. J. Alder. Ground state of the electron gas by a stochastic method. *Phys. Rev. Lett.*, 45(7):566–569, 1980.

[32] S. H. Vosko, L. Wilk, and M. Nusair. Accurate spin-dependent electron liquid correlation energies for local spin density calculations: a critical analysis. *Can. J. Phys.*, 58(8):1200–1211, 1980.

[33] A. D. Becke. Density functional calculations of molecular bond energies. *J. Chem. Phys.*, 84(8):4524–4529, 1986.

[34] J. P. Perdew, K. Burke, and M. Ernzerhof. Generalized gradient approximation made simple. *Phys. Rev. Lett.*, 77(18):3865–3868, 1996.

[35] C. Lee, W. Yang, and R. G. Parr. Development of the colle-salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B.*, 37(2):785–789, 1988.

[36] A. D. Becke. Density-functional thermochemistry. III. the role of exact exchange. *J. Chem. Phys.*, 98(7):5648–5652, 1993.

[37] A. D. Becke. A new mixing of Hartree–Fock and local density-functional theories. *J. Chem. Phys.*, 98(2):1372–1377, 1993.

[38] E. R. Davidson and D. Feller. Basis set selection for molecular calculations. *Chem. Rev.*, 86(4):681–696, 1986.

[39] S. F. Boys. Electronic wave functions. A general method of calculation for the stationary states of any molecular system. *Proceedings of the Royal Society of London A.*, 200(1063):542–554, 1950.

[40] C. J. Cramer. *Essentials of Computational Chemistry*. Wiley. 2nd edition edition, 2004.

[41] S. F. Boys and F. Bernardi. The calculation of small molecular interactions by the differences of separate total energies. some procedures with reduced errors. *Mol. Phys.*, 19(4):553–&, 1970.

[42] K. N Kirschner, J. B. Sorensen, and J. P. Brown. Calculating interaction energies using rst principle theories: Consideration of basis set superposition error and fragment relaxation. *J. Chem. Ed.*, 84:1225, 2007.

[43] J. C. Phillips. Energy-band interpolation scheme based on a pseudopotential. *Phys. Rev.*, 112(3):685–695, 1958.

[44] L. Füsti-Molnár and P Pulay. Gaussian-based first-principles calculations on large sytems using the fourier transform coulomb method. *J. Mol. Struc. (Theochem).*, 666-667:25–30, 2003.

[45] S. Goedecker. Linear scaling electronic structure methods. *Rev. Mod. Phys.*, 71(4):1085–1123, 1999.

[46] S. Ismail-Beigi and T. A. Arias. Locality of the density matrix in metals, semiconductors, and insulators. *Phys. Rev. Lett.*, 82(10):2127–2130, 1999.

[47] C.-K. Skylaris, P. D. Haynes, A. A. Mostofi, and M. C. Payne. Introducing ONETEP: Linear-scaling density functional simulations on parallel computers. *J. Chem. Phys.*, 122:084119, 2005.

[48] C.-K. Skylaris, P. D. Haynes, A. A. Mostofi, and M. C. Payne. Implementation of linear scaling plane wave density functional theory on parallel computers. *Phys. Stat. Sol.*, 243(5):973–988, 2006.

[49] N. D. M. Hine, P. D. Haynes, A. A. Mostofi, C.-K. Skylaris, and M. C. Payne. Linear-scaling density-functional theory with tens of thousands of atoms: Expanding the scope and scale of calculations with onetep. *Comp. Phys. Comm.*, 180(7):1041–1053, 2009.

[50] C.-K. Skylaris, A. A. Mostofi, P. D. Haynes, O. Diéguez, and M. C. Payne. Nonorthogonal generalized wannier function pseudopotential plane-wave method. *Phys. Rev. B.*, 66(3):035119, 2002.

[51] A. A. Mostofi, P. D. Haynes, C.-K. Skylaris, and M. C. Payne. Precondi-

tioned iterative minimization for linear-scaling electronic structure calculations. *J. Chem. Phys.*, 119(17):8842–8848, 2003.

[52] P. D. Haynes, C.-K. Skylaris, A. A. Mostofi, and M. C. Payne. Elimination of the basis set superposition error in linear-scaling density-functional calculations with local orbitals optimized in situ. *Chem. Phys. Lett.*, 422(4-6):345–349, 2006.

[53] S. Grimme. Semiempirical GGA-type density functional constructed with a long-range dispersion correction. *J. Comp. Chem.*, 27:1787–1799, 2006.

[54] D. A. Case, T.A. Darden, T.E. Cheatham, C.L. Simmerling, J. Wang, R.E. Duke, R. Luo, M. Crowley, S. Hayik A. Roitberg, G. Seabra, I. Kolossváry, K. F. Wong, F. Paesani, J. Vanicek, X. Wu, S. R. Brozell, T. Steinbrecher, H. Gohlke, L. Yang, C. Tan, J. Mongan, V. Hornak, G. Cui, D.H. Mathews, M.G. Seetin, C., Sagui, V. Babin, and P. A. Kollman. Amber10, 2008.

[55] W. D. Cornell, P.Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Fergusson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.*, 117:5179–5197, 1995.

[56] A. Perez, I. Marchan, D. Svozil, J. Sponer, T. E. Cheatham. III., C. A. Laughton, and M. Orozco. Refinement of the amber force field for nucleic acids: Improving the description of (α/γ) conformers. *Biophys J.*, 92:3817–3829, 2007.

[57] Y. Duan, C. Wu, S. Chowdhury, M. C. Lee, G. Xiong, W. Zhang, R. Yang, P. Cieplak, R. Lou, and T. Less. A point-charge force field for molecu-

lar mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J. Comp. Chem.*, 24:1999–2012, 2003.

[58] K. N Kirschner, A. B. Yongye, S. M. Tschampel, J. Gonzales-Outeirino, C. R. Daniels, B. L. Foley, and R. J. Woods. A generalizable biomolecular force field. carbo- hydrates. *J. Comp. Chem.*, 29:622–655, 2008.

[59] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case. Development and testing of a general amber force field. *J. Comp. Chem.*, 25:1157–1174, 2004.

[60] T. A. Halgren. Merck molecular force field. Basis, form, scope, parameterization, and performance of mmff94. *Journal of Computational Chemistry*, 17(5-6):490–519, 1996.

[61] P. Cieplak, J. Calderwell, and P. Kollman. Molecular mechanical models for organic and biological systems going beyond the atom centered two body additive approximation: aqueous solution free energies of methanol and n-methyl acetamide, nucleic acid base, and amide hydrogen bonding and chloroform/water partition coefficients of the nucleic acid bases. *J. Comp. Chem.*, 10:1048–1057, 2001.

[62] L. Verlet. Computer “experiments” on classical fluids. Thermodynamical properties of lennard-jones molecules. *Phys. Rev.*, 159:98–103, 1967.

[63] R. W. Hockney. Potential calculation and some applications. *Methods Comput. Phys.*, 9:135–211, 1970.

[64] W. C. Swope, H. C. Andersen, P. H. Berens, and K. R. Wilson. A computer simulation method for the calculation of equilibrium constants for

the formation of physical clusters of molecules: Application to small water clusters. *J. Chem. Phys.*, 76:637–649, 1982.

[65] J.-P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.*, 23:327–341, 1977.

[66] H. J. C. Berendsen, J. R. Grigera, and T. P Straatsma. The missing term in effective pair potentials. *J. Phys. Chem.*, 91:6269–6271, 1987.

[67] S. A. Alderman and J. D. Doll. Generalized langevin equation approach for atom-solid-surface scattering - general formulation for classical scattering off harmonic solids. *J. Chem. Phys.*, 64:2375–2388, 1976.

[68] W. L. Jorgensen. Quantum and statistical mechanical studies of liquids. 10. transferable intermolecular potential functions for water, alcohols, and ethers. application to liquid water. *J. Am. Chem. Soc.*, 103:335–340, 1981.

[69] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, and J. Hermans. *Intermolecular Forces*. Reidel, Dordrecht, 1981.

[70] W. L. Jorgensen, J. Chandrasekhar, and J. D. Madura. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, 79:926–935, 1983.

[71] J. D. Bernal and R. H. Fowler. A theory of water and ionic solution, with particular reference to hydrogen and hydroxyl ions. *J. Chem. Phys.*, 1:515, 1933.

[72] W. L. Jorgensen. Revised tips for simulations of liquid water and aqueous solutions. *J. Chem. Phys.*, 77:4156–4163, 1982.

[73] J. L. F. Abascal, E. Sanz, R. Garcia Fernandez, and C. Vega. A potential model for the study of ices and amorphous water: TIP4P/Ice. *J. Chem. Phys.*, 122:234511, 2005.

[74] J. L. F. Abascal and C. Vega. A general purpose model for the condensed phases of water: TIP4P/2005. *J. Chem. Phys.*, 123:234505, 2005.

[75] F. H. Stillinger and A. Rahman. Improved simulation of liquid water by molecular dynamics. *J. Chem. Phys.*, 60:1545–1557, 1974.

[76] M. W. Mahoney and W. L. Jorgensen. A five-site model liquid water and the reproduction of the density anomaly by rigid, non-polarizable models. *J. Chem. Phys.*, 112:8910–8922, 2000.

[77] D. Sitkoff, K. A. Sharp, and B. Honig. Accurate calculation of hydration free energies using macroscopic solvent models. *J. Phys. Chem.*, 98:1978–1988, 1994.

[78] J. Dziedzic, H. H. Helal, C.-K. Skylaris, A. A. Mostofi, and M. C. Payne. Minimal parameter implicit solvent model for ab initio electronic structure calculations. *Europhysics Letters*, 95:43001, 2011.

[79] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, Jr. J. A. Montgomery, T. Vreven, K. N. Kudin, J. C. Burant, J. M. Millam, S. S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega, G. A. Petersson, H. Nakatsuji, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, M. Klene, X. Li, J. E. Knox, H. P. Hratchian, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski,

P. Y. Ayala, K. Morokuma, G. A. Voth, P. Salvador, J. J. Dannenberg, V. G. Zakrzewski, S. Dapprich, A. D. Daniels, M. C. Strain, O. Farkas, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. V. Ortiz, Q. Cui, A. G. Baboul, S. Clifford, J. Cioslowski, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, C. Gonzalez, and J. A. Pople. Gaussian 03, Revision C.02. Gaussian, Inc., Wallingford, CT, 2004.

[80] A. V. Marenich, C. J. Cramer, and D. G. Truhlar. Universal solvation model based on solute electron density and on a continuum model of the solvent defined by the bulk dielectric constant and atomic surface tensions. *J. Phys. Chem. B.*, 113(18):6378–6396, 2009.

[81] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, Jr. J. A. Montgomery, J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, Ö. Farkas, J. B. Foresman, J. V.

Ortiz, J. Cioslowski, and D. J. Fox. Gaussian 09 Revision A.1. Gaussian Inc. Wallingford CT 2009.

[82] C. D. Christ, A. E. Mark, and W. F. van Gunsteren. Basic ingredients of free energy calculations: A review. *J. Comp. Chem.*, 31:1569–1582, 2010.

[83] P. A. Kollman, I. Massova, C. Reyes, B. Kuhn, S. Huo, L. Chong, M. Lee, T. Lee, Y. Duan, W. Wang, O. Donini, P. Cieplak, J. Srinivasan, D. A. Case, and T. E. Cheatham. Calculating structures and free energies of complex molecules: Combining molecular mechanics and continuum models. *Acc. Chem. Res.*, 33(12):889–897, 2000.

[84] W. C. Still, A. Tempczyk, R. C. Hawley, and T. Hendrickson. Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.*, 112:6127–6129, 1990.

[85] J. Michel and J.W. Essex. Prediction of protein–ligand binding affinity by free energy simulations: assumptions, pitfalls and expectations. *J. Comput. Aided. Mol. Des.*, 24:639–658, 2010.

[86] R. Zwanzig. High-temperature equation of state by a perturbation method. i. nonpolar gases. *J. Chem. Phys.*, 22(8):1420–1426, 1954.

[87] I. Massova and P. A. Kollman. Computational alanine scanning to probe proteinprotein interactions: A novel approach to evaluate binding free energies. *J. Am. Chem. Soc.*, 121(36):8133–8143, 1999.

[88] G. Jones, P. Willett, and R. C. Glen. Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation . *J. Mol. Biol.*, 245:43–53, 1995.

[89] G. Jones, P. Willett, R. C. Glen, A. R. Leach, and R. Taylor. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.*, 267:727–748, 1997.

[90] M. L. Verdonk, J. C. Cole, M. J. Hartshorn, C. W. Murray, and R. D. Taylor. Improved protein-ligand docking using gold. *Proteins*, 52:609–623, 2003.

[91] J. C. Cole, J. W. M. Nissink, and R. Taylor. *Protein-Ligand Docking and Virtual Screening with GOLD*. irtual Screening in Drug Discovery. Taylor & Francis CRC Press, Boca Raton, Florida, USA, 2005.

[92] H. E. Velec, H. Gohlke, and G. Klebe. Drugscore(csd)-knowledge-based scoring function derived from small molecule crystal data with superior recognition rate of near-native ligand poses and better affinity prediction. *J. Med. Chem.*, 48:6296–6303, 2005.

[93] J. Aqvist and J. Marelius. The linear interaction energy method for predicting ligand binding free energies. *Comb Chem High Throughput Screen.*, 4:613–626, 2001.

[94] Y. Su, E. Gallicchio, K. Das, E. Arnold, and R. M. Levy. Linear interaction energy (LIE) models for ligand binding in implicit solvent: Theory and application to the binding of NNRTIs to HIV-1 reverse transcriptase. *J. Chem. Theory. Comput.*, 3:256–277, 2007.

[95] W. A. Baase, X. J. Zhang, D. W. Heinz, M. Blaber, E. P. Baldwin, B. W. Matthews, and A. E. Eriksson. Responce of a protein-structure to cavity-creating mutations and its relation to the hydrophobic effect. *Science.*, 255(5041):178–183, 1992.

[96] A. Morton, W.A Baase, and B. W. Matthews. Energetic origins of specificity of ligand binding in an interior nonpolar cavity of T4 Lysozyme. *Biochemistry.*, 34:8564–8575, 1995.

[97] A. Morton. Specificity of ligand binding in a buried nonpolar cavity of T4 lysozyme: Linkage of dynamics and structural plasticity. *Biochemistry.*, 34:8576–8588, 1995.

[98] B. Q. Wei, W. A .Baase, L. H. Weaver, B. W. Matthews, and B. K. Shoichet. A model binding site for testing scoring functions in molecular docking. *J. Mol. Biol.*, 322:339–355, 2002.

[99] E. Gallicchio, M. Lapelosa, and R. M. Levy. Binding energy distribution analysis method (BEDAM) for estimation of proteinligand binding affinities. *J. Chem. Theory. Comput.*, 6:2961–2977, 2010.

[100] S. E. Boyce, D. L. Mobley, G. J. Rocklin, A. P. Graves, K. A. Dill, and B. K. Shoichet. Predicting lingand binding affinity with alchemical free energy methods in a polar model binding site. *J. Mol. Biol.*, 394:747–763, 2009.

[101] B. Q. Wei, L. H. Weaver, A. M. Ferrari, B. W. Matthews, and B. K. Shoichet. Testing a flexible-receptor docking algorithum in a model binding site. *J. Mol. Biol.*, 337:1161–1182, 2004.

[102] A. P. Graves, R. Brenk, and B. K. Shoichet. Decoys for docking. *J. Med. Chem.*, 48:3714–3728, 2005.

[103] Y. Deng and B. Roux. Computations of standard binding free energies with molecular dynamics simulations. *J. Phys. Chem. B.*, 113:2234–2246, 2009.

[104] MOE2009.10. *Chemical Computing Group Inc., Montreal*, 2009.

[105] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, and C. Simmerling. Comparison of multiple amber force fields and development of improved protein backbone parameters. *PROTEINS: Structure, Function, and Genetics*, 65:712–725, 2006.

[106] M. Kaukonen, P. Soderhjelm, J. Heimdal, and U. Ryde. QM/MM-PBSA method to estimate free energies for reactions in proteins. *J. Phys. Chem. B.*, 112:12537–12548, 2008.

[107] T. H. Rod and U. Ryde. A method for calculating high-level QM/MM free energies for enzymatic reactions: methyl transfer catalyzed by catechol O-methyltransferase. *Phys. Rev. Lett.*, 94:198302, 2005.

[108] M. Wang and C. F. Wong. Rank-ordering protein-ligand binding affinity by a quantum mechanics/molecular mechanics/poisson-boltzmann-surface area model. *J. Chem. Phys.*, 126:026101, 2007.

[109] J. M. Soler, E. Artacho, J. D. Gale, A. Garcia, J. Junquera, P. Ordejon, and D. Sanchez. The Siesta method for ab initio order-n materials simulation. *J. Phys. Cond. Mat.*, 14:2745–2779, 2002.

[110] M. E. Davis, J. D. Madura, B. A. Luty, and J. A. McCammon. Electrostatic and diffusion of molecules in solution: Simulations with the university of houston brownian dynamics program. *Comp. Phys. Comm.*, 62:187–197, 1991.

[111] J. J. P. Stewart. Optimization of parameters for semiempirical methods i. method. *J. Comp. Chem.*, 10(2):209, 1989.

[112] S. L. Dixon, A. van der Vaart, B. Wang, V. Gogonea, J. J. Vincent, E. N. Brothers, D. Suarez, L. M. Westerhoff, and Jr K. M. Merz. Divcon, the pennsylvania state university, university park, pa, 16802, 2004.

[113] D. J. Cole, C.-K. Skylaris, E. Rajendra, A. R. Venkitaraman, and M. C. Payne. Protein-protein interaction from linear-scaling first-principles quantum-mechanical calculations. *Europhysics Letters*, 91:37004, 2010.

[114] S. Fox, H. G. Wallnoefer, T. Fox, C. S. Tautermann, and C.-K. Skylaris. First principles-based calculations of free energy of binding: Application to ligand binding in a self-assembling superstructure. *J. Chem. Theory. Comput.*, 7:1102–1108, 2011.

[115] T. Fox, B. E. Thomas IV, M. McCarrick, and P. A. Kollman. Application of free energy perturbation calculations to the “tennis ball” dimer: Why is CF4 not encapsulated by this host? *J. Phys. Chem.*, 100(25):10779–10783, 1996.

[116] N. Branda, R. Wyler, and J. Rebek. Encapsulation of methane and other small molecules in a self-assembling superstructure. *Science.*, 263(5151):1267–1268, 1994.

[117] T. H. Dunning. Gaussian basis sets for use in correlated molecular calculations. The atoms boron through neon and hydrogen. *J. Chem. Phys.*, 90:1007–1023, 1989.

[118] A. D. Becke. Density-functional thermochemistry. v. systematic optimization of exchange-correlation functionals. *J. Chem. Phys.*, 107(20):8554–8560, 1997.

[119] J. Dziedzic, S. J. Fox, T. Fox, C. Tautermann, and C.-K. Skylaris. Large-scale DFT calculations in implicit solvent - A case study on the T4 lysozyme L99A/M102Q protein. *Int. J. Quantum Chem.*, 2012.

[120] Y. Deng and B. Roux. Calculation of standard binding free energies: Aromatic molecules in the T4 lysozyme L99A mutant. *J. Chem. Theory. Comput.*, 2(5):1255–1273, 2006.

[121] D. L. Mobley, A. P. Graves, J. D Chodera, A. C. McReynolds, B. K. Shoichet, and K. A. Dill. Predicting absolute ligand binding free energies to a simple model site. *J. Mol. Biol.*, 371:1118–1134, 2007.

[122] M. Štrajbl, G. Hong, and A. Warshel. Ab initio QM/MM simulation with proper sampling: First principle calculations of the free energy of the autodissociation of water in aqueous solution. *J. Phys. Chem. B.*, 106(51):13333–13343, 2002.

[123] C. J. Woods, F. R. Mandy, and A. J. Mullholland. An efficient method for the calculation of quantum mechanics/molecular mechanics free energies. *J. Phys. Chem.*, 128:014109, 2008.

[124] F. R. Beierlein, J. Michel, and J. W. Essex. A simple QM/MM approach for capturing polarization effects in protein-ligand binding free energy calculations. *J. Phys. Chem. B.*, 115(17):4911–4926, 2011.

[125] Richard M. Martin. *Electronic Structure. Basic Theory and Practical Methods*. Cambridge University Press, 2004.

[126] Bernstein N., Kermode J. R., and G. Csanyi. Hybrid atomistic simulation methods for materials systems. 72:026501, 2009.

[127] J.-L. Fattebert, R. J. Law, B. Bennion, E. Y. Lau, E Schwegler, and F. C. Lightstone. Quantitative assessment of electrostatic embedding in density functional theory calculations of biomolecular systems. 5:2257–2264, 2009.

[128] H. Nanda, N. Lu, and T. B. Woolf. Using non-gaussian density functional fits to improve relative free energy calculations. *J. Chem. Phys.*, 122:134110, 2005.

[129] F. Weinhold and C. Landis. *Valency and Bonding: A natural Bond Orbital Donor-Acceptor Perspective*. 2005.

[130] A. E. Reed, L. A. Curtiss, and F. Weinhold. Intermolecular interactions from a natural bond orbital, donor-acceptor viewpoint. *Chem. Rev.*, 88:899–926, 1988.

[131] A. D. MacKerell, B. Brooks Jr., C. L. Brooks III, L. Nilsson, B. Roux, Y. Won, and M. Karplus. *Encyclopedia of Computational Chemistry*. John Wiley & Sons, 1998.

[132] G. N. Lewis. The atom and the molecule. *J. Am. Chem. Soc.*, 38:762–785, 1916.

[133] E. D. Glendening, J. K. Badenhoop, A. E. Reed, J. E. Carpenter, J. A. Bohmann, C. M. Morales, and F. Weinhold. *The NBO 5.9 Manual*.

[134] L. P. Lee, D. J. Cole, M. C. Payne, and C.-K. Skylaris. Natural bond orbital analysis in the onetep code: Applications to large protein systems. *J. Comp. Chem.*, 2012.

[135] F. L. Hirshfeld. Bonded-atom fragments for describing molecular charge densities. *Theoret. Chim. Acta*, 44:129–138, 1977.

[136] P. Bultinck, C. Van Alsenoy, P. W. Ayers, and R. Carbo-Dorca. Critical analysis and extension of the hirshfeld atoms in molecules. *J. Chem. Phys.*, 126:144111, 2007.

[137] C. Fonseca Guerra, J. W. Handgraaf, E. J. Baerends, and F. M. Bickelhaupt. Voronoi deformation density (VDD) charges: Assessment of the mulliken, bader, hirshfeld, weinhold, and VDD methods for charge analysis. *J. Comp. Chem.*, 25:189–210, 2004.

[138] J Bingham, S Sudarsanam, and S Srinivasan. Profiling human phosphodiesterase genes and splice isoforms. *Biochemical and Biophysical Research Communications*, 350:25–32, 2006.

[139] G. L. Card, B. P. England, Y. Suzuki, D. Fong, B. Powell, B. Lee, C. Luu, M. Tabrizizad, S. Gillette, P. N. Ibrahim, D. R. Artis, G. Bollag, M. V. Milburn, S.-H. Kim, J. Schlessinger, and K. Y. J. Zhang. Structural basis for the activity of drugs that inhibit phosphodiesterases. *Structure*, 12:2233–2247, 2004.

[140] H. A Ghofrani, I. H. Osterloh, and F. Grimminger. Sildenafil: from angina to erectile dysfunction to pulmonary hypertension and beyond. *Nature Reviews Drug Discovery*, 5:689–702, 2006.

[141] K.-Y. Wong and J. Gao. Insight into the phosphodiesterase mechanism from combined qm/mm free energy simulations. *FEBS J.*, 278:2579–2595, 2011.

[142] K. A. O'Brien, E. A. Salter, and A. Wierzbicki. Oniom qunatum chemistry study of cyclic nucleotide recognition in phosphodiesterase 5. *Int. J. Quantum Chem.*, 107:2197–2203, 2007.

[143] B. Zagrovic and W. F. van Gunsteren. Computational analysis of the mechanism and theromdynamics of inhibition of phosphodiesterase 5a by synthetic ligands. *J. Chem. Theory. Comput.*, 3:301–311, 2007.

[144] T.A. Soares, X. Daura, C. Oostenbrink, L. J. Smith, and W. F. van Gunsteren. Validation of the GROMOS force-field parameter set 45A3 against nuclear magnetic resonance data of hen egg lysozyme. *J. Biomolecular NMR*, 30:407–422, 2004.

[145] H. A. Mohamed, N. M. R. Girgis, R. Wilcken, M. R. Bauer, H. N. Tinsley, B. D. Gary, G. A. Piazza, F. M. Boeckler, and A. H. Abadi. Synthesis and molecular modeling of novel tetrahydro--carboline derivatives with phosphodiesterase 5 inhibitory and anticancer properties. *J. Med. Chem.*, 54(2):495–509, 2011.

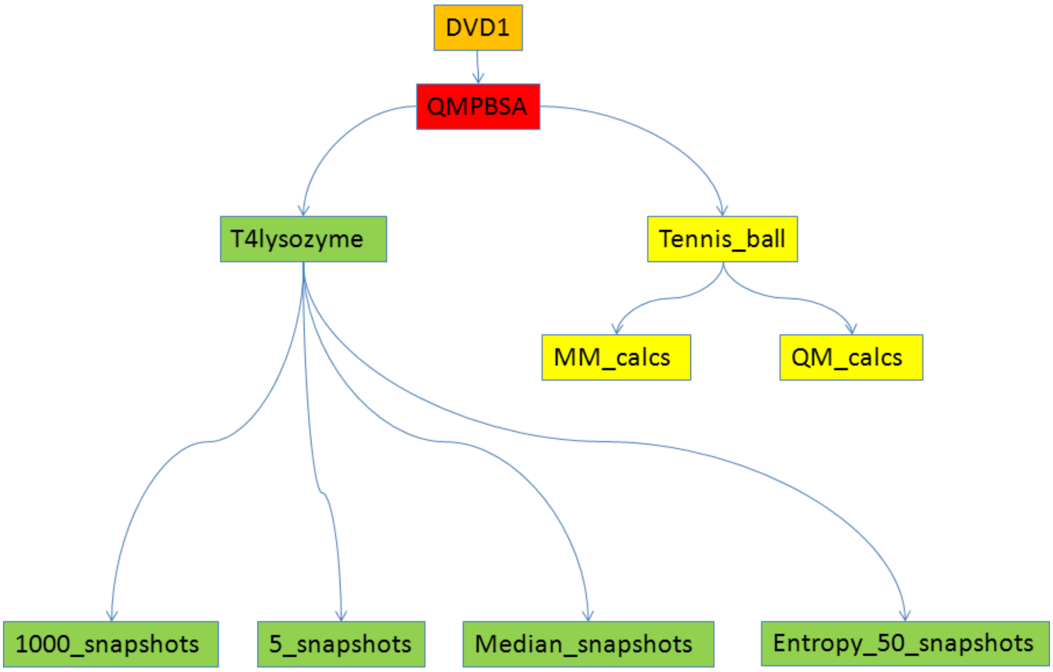
[146] S. P. Niinivehmas, S. I. Virtanen, J. V. Lehtonen, P. A. Postila, and O. T. Pentikainen. Comparison of virtual high-throughput screening methods for the identification of the phosphodiesterase-5 inhibitors. *J. Chem. Inf. Model*, 51:1353–1363, 2011.

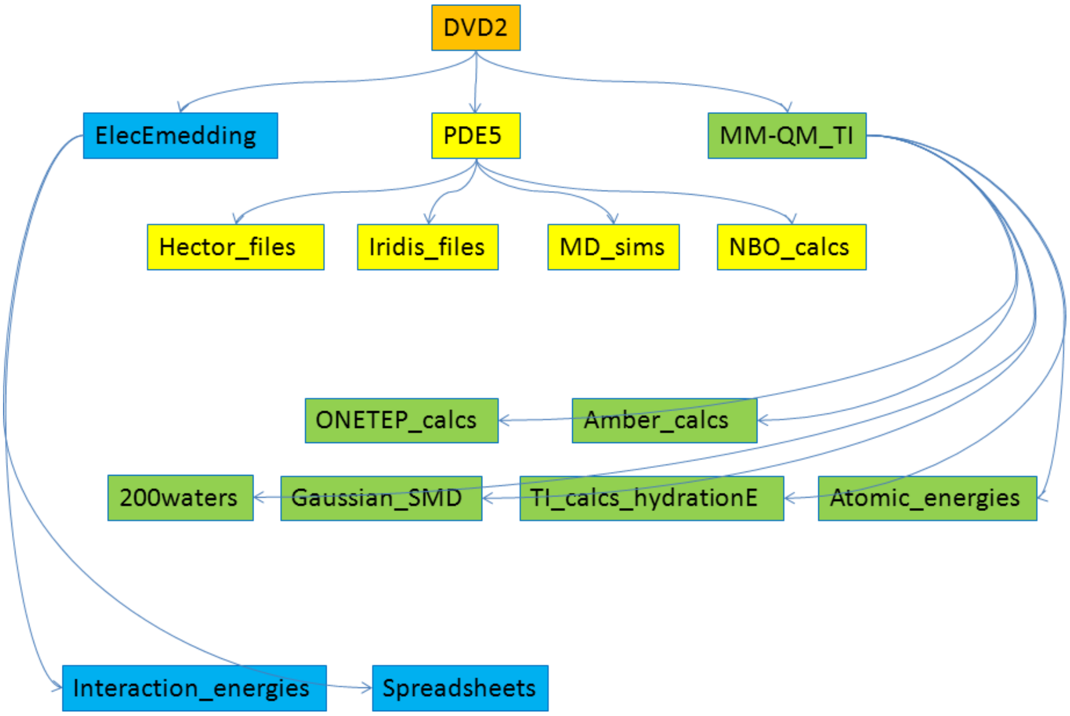
[147] R. A. Friesner, J. L. Banks, R. B. Murphy, T. A. Halgren, J. J. Klicic, D. T. Mainz, M. P. Repasky, E. H. Knoll, M. Shelley, J. K. Perry, D. E. Shaw, P. Francis, and P. S. Shenkin. Glide: A new approach for rapid, accurate docking and scoring. 1. method and assessment of docking accuracy. *J. Med. Chem.*, 47(7):1739–1749, 2004. PMID: 15027865.

[148] H. Haning, U. Niewohner, T. Schenke, T. Lampe, A. Hillisch, and E. Bischoff. Comparison of different heterocyclic scaffolds as substrate analog pde5 inhibitors. *Bioorg. Med. Chem. Lett*, 15:3900–3907, 2005.

[149] T. Fox. Personal correspondence.

[150] A. Ruiz-Serrano, N. D. M. Hine, and C.-K. Skylaris. Pulay forces from localized orbitals optimized in situ using a psinc basis set. *J. Chem. Phys.*, 136:234101, 2012.


DVD1
QMPBSA
T4lysozyme
Tennis_ball
MM_calcs
QM_calcs
1000_snapshots
5_snapshots
Median_snapshots
Entropy_50_snapshots


DVD2
ElecEmedding
PDE5
MM-QM_TI
Hector_files
Iridis_files
MD_sims
NBO_calcs
ONETEP_calcs
Amber_calcs
200waters
Gaussian_SMD
TI_calcs_hydrationE
Atomic_energies
Interaction_energies
Spreadsheets