University of Southampton

Faculty of Natural and Environmental Sciences

School of Chemistry

The Evaluation of Protein-Ligand Binding Free Energies Using Advanced Potential Energy Functions

by

Noor Asidah Mohamed

Thesis for the degree of Doctor of Philosophy

August 2018

University of Southampton

Abstract

Faculty of Natural and Environmental Sciences

School of Chemistry

Thesis for the degree of Doctor of Philosophy

The Evaluation of Protein-Ligand Binding Free Energies Using Advanced Potential Energy Functions

by

Noor Asidah Mohamed

Electronic polarisation is one of the components that plays an important role in many biomolecular systems. The effects of polarisation will act differently depending on the local environment of the system, such as in DNA, proteins and membranes. Traditionally, molecular mechanical force fields describe electrostatics as the interactions of fixed, atom-centred, point charges. Hence the past decade has seen many additions and improvements to existing force fields to better correlate dynamics with experimental observations. A better description of electrostatics by the inclusion of electronic polarisation is one such improvement. The AMOEBA polarisable force field is one of many possible models that is designed to be capable of capturing this effect. AMOEBA includes mutually polarising induced atomic dipoles at every atomic site, as well as a multipolar representation of fixed electrostatics.

To investigate applications of AMOEBA and where its successes over existing fixed-charge methods may lie, we first evaluate features and performance of the AMOEBA polarisable force field in simple systems based on the evaluation of solvation free energies for small molecules in a range of common organic solvents. Here, we pointed out several challenges and limitations of AMOEBA in this study involving non-aqueous solvents. Then, we further our investigation on more complex systems including protein-ligand interactions. Initially, clear cases of failure in fixed-point-charge force fields were identified by exploring the sensitivity of the calculated free energies to parameter sets and simulation protocols of protein-ligand systems, focusing on binding free energy calculations of the cytochrome c peroxidase protein using the AMBER force field. Finally, we use these results to inform binding free energy calculations for testing of the AMOEBA force field. We discuss the implications of these results for better understanding and improving AMOEBA to aid its full implementation in other biological applications.

Table of Contents

List of Tables

List of Figures

Research Thesis: Declaration of Authorship

Print name:	Noor Asidah Mohamed

Title of thesis:	The Evaluation Of Protein-Ligand Binding Free Energies Using Advanced Potential Energy Functions

I declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. 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;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. 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;
5. I have acknowledged all main sources of help;
6. 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;
7. Parts of this work have been published as:

Publications:

Mohamed N. A.,Bradshaw R. T., Essex J. W., "Evaluation of solvation free energies for small molecules with the AMOEBA polarizable force field", *J. Comp Chem*, **2016**, DOI: 10.1002/jcc.24500

Mohamed N. A., Bradshaw R. T., Essex J. W., "Underlying data for 'Evaluation of solvation free energies for small molecules with the AMOEBA polarizable force field' ", [Data set], Zenodo, **2016**, http://doi.org/10.5281/zenodo.59203

Poster presentations:

Mohamed N. A., Bradshaw R. T., Essex J. W., “Evaluation of solvation free energies for small molecules with the AMOEBA polarisable force field,” 4th Annual CCP-BIOSIM Conference, Frontier of Biomolecular Simulation, Leeds, UK, 2015

Mohamed N. A., Bradshaw R. T., Essex J. W., “Evaluating AMOEBA polarisable force field performance through solvation free energy calculations in non-aqueous environments”, 21st European Symposium on Quantitative Structure-Activity Relationship (EuroQSAR), Verona, Italy, 2016

Mohamed N. A., Bradshaw R. T., Essex J. W., “Evaluating AMOEBA polarisable force field performance through solvation free energy calculations in non-aqueous environments”, Professor Tony Hey’s 70th Birthday Symposium, Southampton, UK, 2016

Mohamed N. A., Bradshaw R. T., Essex J. W., “Evaluating parameter and methodology sensitivity in binding free energy calculations of cytochrome c peroxidase”, Graduate School of Natural and Environmental Sciences Poster Day, Southampton, UK, 2016

Mohamed N. A., Bradshaw R. T., Essex J. W., “Using the AMOEBA polarisable force field in biomolecular simulation”, MGMS Young Modellers' Forum 2016, Greenwich, UK, 2016

Signature:		Date:	

Acknowledgements

I am so indebt to many people who helped me toward accomplishing this project. It is impossible for me to acknowledge every one of them individually, but several in particular deserve recognition.

I wish to express deepest appreciation and thanks to my supervisor Prof Jonathan W. Essex for his invaluable concern, sustained guidance and unstinting support that enables me to bring this work to completion.

I also would like to extend my heartiest gratitude to my teammate, Richard, whose critical eye and enlightened monitoring were both instrumental and inspiring. His continuous review, guidance, ideas and suggestion have been invaluable to this piece of work.

Thanks should also go to all Essex group's members for their help and encouragement and making 2011 lab such an enjoyable place to work.

Not forgetting, special thanks are due to my family, especially my husband, Fadlee and my son, Iman for their inspiring words that spurred me on to work harder to make this thesis reality. Their enduring and unselfish support and understanding during the completing of this thesis is been invaluable for me.

Definitions and Abbreviations

3D	Three-Dimensional
AM1-BCC	Semi-empirical (AM1)-Bond Charge Correction
AMBER	Assisted Model Building and Energy Refinement
AMOEBA	Atomic Multipole Optimized Energetics for Biomolecular Application
BAR	Bennett's Acceptance Ratio
CFF	Consistent Force Field
CHARMM	Chemistry at Harvard Molecular Mechanics
DMSO	Dimethyl sulfoxide
FEP	Free Energy Perturbation
GAFF	General AMBER Force Field
GB	Generalised Born
GDMA	Gaussian Distributed Multipole Analysis
GROMOS	Groningen Molecular Simulation
HT	High Throughput
HTS	High Throughput Screening
ITC	Isothermal Titration Calorimetry
MD	Molecular Dynamics
MM	Molecular Mechanics
MM-GBSA	Molecular Mechanics Generalised Born Surface Area model
MM-PBSA	Molecular Mechanics Poisson-Boltzmann Surface Area
MUE	Mean Unsigned Error
NMR	Nuclear Magnetic Resonance

NPT	Isobaric/isothermal ensemble (constant Number/Pressure/ Temperature)
NVT	Canonical ensemble (constant Number/Volume/Temperature)
OPLS	Optimized Potential for Liquid Simulations
OPLS-AA	Optimized Potential for Liquid Simulations-All Atom
PB	Poisson-Boltzmann
PME	Particle Mesh Ewald
QM	Quantum Mechanics
SE	Standard Error
SPR	Surface Plasmon Resonance
TI	Thermodynamic Integration
VdW	Van der Waals

Chapter 1: Introduction

Biological systems are governed by biomolecular interactions between DNA, proteins and ligands for numerous biological functions. Protein-ligand binding is one of the interactions that occurs virtually in all cellular processes and is crucial for metabolism, signalling and development in living systems.

In the current post-genomics era, explanation of the molecular recognition of complexes requires a comprehensive understanding of the specific interactions of the biomolecule units and how they cooperate to express their functions. Although structural knowledge can arise from a variety of experimental techniques, structural knowledge can also arise from computational techniques to predict protein-ligand complex structure and protein binding sites. Free energy calculations provide great opportunities for understanding the action and regulation mechanism of protein-ligand interactions. However, free energy estimation methods traditionally suffer because of challenges in sampling and approximations in empirical potential energy functions. Sampling issues have been tackled using a variety of methods for decades.[1,2] Likewise, the issues associated with force fields are also well known over the past few decades, and have resulted in many systematic improvements to existing force fields.[3–5] However, issues with force fields still remain problematic in free energy evaluations.

This thesis focuses on the evaluation of free energies using advanced potential energy functions to improve the accuracy of modelling intermolecular interactions in protein-ligand systems. Through computational simulations and calculations, we first implement this model on simple system by evaluating solvation free energies for small molecules in non-aqueous solvents as a starting point. Further investigation was then performed, by evaluating the protein-ligand binding free energies in cytochrome c peroxidase protein. In this chapter, a brief literature review of some of the basic concepts of protein-ligand interactions is addressed, followed by the motivation and aims of this work.

1.1 Protein ligand-binding

Proteins are the vital macromolecular players in many cellular processes. A protein is a chain of amino acids held together by peptide bonds. Inside our cells, these biomolecule units have structure and dynamics as varied as the functions they serve. Proteins are responsible for various tasks of cellular life – catalysing chemical reactions,[6] neuron signalling,[7] mediating cell responses,[8,9] immune protection[10] and growth control.[11,12] However, proteins are very complex

systems with distinct three-dimensional structure. They have to interact with other biomolecule components and change conformation under various conditions to form stoichiometric stable complexes and create functional modules and pathways in cells.

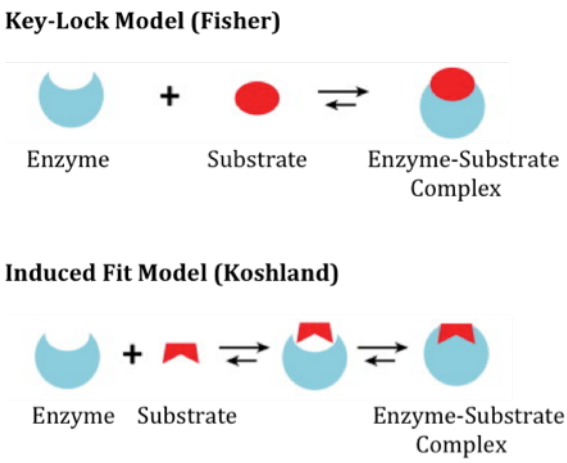


Figure 1.1: The basic concepts of molecular interaction described by two models of enzyme-substrate interactions: Key-lock model and induced fit model.

Emil Fischer's[13] classic 'key-lock' analogy, 1894 and Daniel E. Koshland's[14] 'induced fit theory', 1958 (Figure 1.1) introduced the fundamental concepts of molecular interactions. Fischer described the principles of the 'key-lock model', based on complementarity of the interacting surfaces associated with the appropriate shape and volume of the molecules coupled with non-covalent interactions - hydrogen-bonding, ionic interactions, van der Waals forces and hydrophobic effects - to stabilise the complex. Instead of only one particular shape able to fit each biomolecule, a modification called 'induced fit theory' proposed that dynamic rearrangements of the molecules to enable the fit of the macromolecule and the ligand occur in certain cases.

In protein-ligand binding systems, a ligand is complementary to a protein, acts as a signal-trigger and binds to a site on the target protein. A ligand is often considered to be a small molecule, however, anything that binds with specificity can be considered a ligand – ions[15] or proteins[16] that form a stable complex with other proteins to modulate biological activity, for example.

Protein binding interactions are the key aspect underlining protein function. Generally, the protein-ligand binding reaction is reversible - whether a stable or unstable state is formed depends on the structures and functions desired. Although covalent modification is possible, it will not be dealt with here, partly due to complexities of identifying reaction mechanism, thermodynamics and kinetics experimentally, and partly the inability to model covalent binding with classical molecular dynamics. Ligand-binding interactions are able to switch proteins

between states of different function. Several mechanisms significant to biological processes involve specific recognition of ligands by proteins; enzyme-substrate interaction and catalysis of key chemical reactions inside cells,[17] transporters routinely use recognition of specific molecules for their movement across membrane barriers,[18] receptors uniquely bind to hormones,[19] or other chemical messengers for inter and intracellular interaction, and antibodies specifically bind to chemical agents to mount vital defence mechanisms against infection and disease in immune systems.[10, 20]

Developing a detailed understanding of the structure-function relationships of protein-ligand interactions is essential to the molecular life sciences. Thermodynamic properties of binding between protein and ligand play an important role in structure-based drug design.[21-22] In general, the binding of a protein-ligand (PL) complex, i.e. the reversible reaction of protein (P) with a ligand (L) in an aqueous environment, is given by the equation:

$$\mathrm{P}_{(aq)} + \mathrm{L}_{(aq)} \rightleftharpoons \mathrm{PL}_{(aq)} \tag{1.1}$$

The equilibrium constant, K_{eq}, also known as association constant or affinity constant, K_A for the binding of a ligand to a protein is described by the following equation:

$$K_{eq} = \frac{[\mathrm{PL}]}{[\mathrm{P}][\mathrm{L}]} \tag{1.2}$$

Where [PL] is the concentration of the protein-ligand complex, [P] is the concentration of the protein, and [L] is the concentration of the free ligand. Note that the dissociation constant, K_D is just the inverse of K_{eq}:

$$K_D = \frac{[\mathrm{P}][\mathrm{L}]}{[\mathrm{PL}]} \tag{1.3}$$

K_D provides a qualitative measure of the binding affinity of ligand at the binding site, measured in molar units, M. Strictly speaking, equilibrium constants such as K_{eq} (K_A) should be calculated using activity of protein, ligand and complex, but this is commonly approximated to concentration in biochemical assays in dilute solution. The binding affinity indicates the strength of the interaction between two or molecules that bind reversibly. For a simple case of a ligand binding at a single site that is not affected by any other sites of binding on the target protein, the value of K_D is the concentration of the ligand that corresponds to half of the binding sites being occupied[23]. Hence the lower the K_D the tighter the binding.

At equilibrium, the binding constant K_D is related to the standard Gibbs free-energy change ($\Delta G°$) of the reaction through the equation below, where R is the gas constant and T is the absolute temperature (Kelvin).

$$\Delta G° = RT \ln K_D \quad (1.4)$$

This equation relates the affinity and specificity of protein-ligand binding to the change in binding free energy of the complex compared with other potential targets. In order to compare the specific ligand binding to multiple targets, the more negative the value of $\Delta G°$, the more favourable the reaction. By utilising the experimentally measurable quantity, K_D, in this equation, the binding free energy changes for the systems can be derived ranging from weak to strong binding. As reported, weak binding of coenzymes (e.g. nicotinamide and enzymes) is generally within 0.1 μM to 0.1 mM[23], while strong binding of complexes (e.g. avidin-biotin) exhibits K_D values of up to 0.1 fM (1fM = 10^{-15} M)[24]. In drug design, very high specificity and very high binding affinity, are desired to create potent drugs. Binding affinity refers to low K_D values. Specificity in the protein-binding event refers to possessing specific geometries of intermolecular interactions to satisfy specific counterparts at the binding interface.[25,26]

Change in binding free energy ($\Delta G°$) is influenced by two thermodynamics concepts: change in enthalpy ($\Delta H°$), "heat content" and change in entropy ($\Delta S°$). Each of these properties are in standard states[25] represented by the superscript '°'. The relationship between these quantities is written as:

$$\Delta G° = \Delta H° - T\Delta S° \quad (1.5)$$

The change in enthalpy in protein-ligand binding is primarily derived from van der Waals interactions, electrostatics, and hydrogen bonding. This change is result of the breaking and formation of non-covalent interactions such as loss of protein-solvent and ligand-solvent hydrogen bonds and the formation of protein-ligand hydrophobic contacts and hydrogen bonds.

Entropy upon binding is related to the changes in disorder, or degrees of freedom of the system. Change in entropy of the system (protein, ligand and solvent) is generally attributed to the solvation and desolvation energies, hydrophobic features, and structure changes of both ligand and receptor during complex formation. Energy is changes during the process in which the ligand is transferred from the hydrophilic environment of the solvent to the predominantly hydrophobic environment of the binding site. All these factors contribute to entropy and enthalpy summing to either a favourable or unfavourable change,[23,25] known as enthalpy-entropy compensation. This phenomenon confers a qualitative explanation for the strong link between the amount of mobility

($\Delta S°$) at a protein-ligand interface and the strength of the interaction ($\Delta H°$) between the protein-ligand at this interface.[27,28]

1.2 Protein-ligand binding drug discovery

Understanding of protein-ligand binding mechanisms and the ability to predict the binding affinities is extremely important in drug discovery studies. During the last two decades, experimental and computational techniques have been developed to address this factor. Nowadays, an appreciation of the three-dimensional (3D) structure of these protein targets is being exploited in every drug-discovery project. Indeed, this has brought new promise to drug design and development for therapeutic intervention, where the central interest is to affect the biological activity of a particular molecular target to provide a cure for diseases.

Generally, drug discovery is a high-cost and time-consuming process that mostly fails.[29,30] Based on retrospective analyses of the pharmaceutical industry during the 1990s, the process to develop each new drug in the market took an average of 14 years with an estimated cost of $ 800 million[29] and now costs up to $ 2.6 billion.[30] However, the probability of success is just one in nine.[31] Mostly, the failure of compounds was identifed in earlier stage of the drug development process due to the lack of safety and efficacy issues.[30] The current trend in modern pharmaceuticals is focussing at developing safe and innovative drugs that bind selectively to a single target to ensure a unique pharmacological response and reduce undesirable side effects. The necessity of such specificity addressing requires the detailed understanding of factors responsible for affecting affinity to the receptor binding site. Significant advances in biotechnology and huge amounts of new protein-ligand data available offer the possibility to accelerate the drug discovery process.

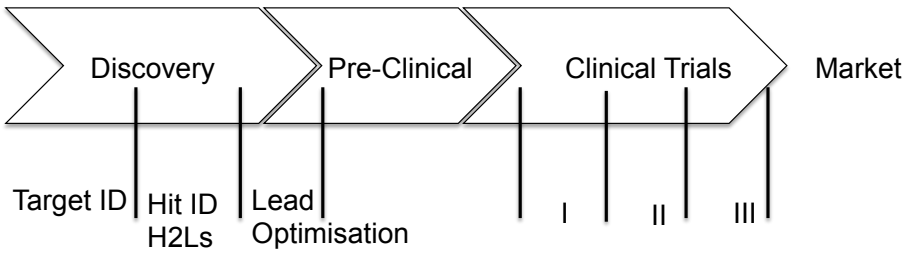


Figure 1.2: The stages of the drug discovery process. Structure-based methods play a significant role in the early stages. Clear features for the progress of every stage define the success of a drug making it to market. Figure taken from Hubbard, 2006.[32]

Figure 1.2 shows the stages involved in most modern drug discovery processes. The challenge in the drug discovery process is to verify desirable criteria for progression to the next stage as the stages become more expensive and consume more resources. The major concern in the drug discovery process is definitely on the initial stage. In this stage, protein-ligand binding activity is prerequisite to serve the success of the next phase. Identifying a relevant target from the established knowledge of protein-ligand binding mechanisms that lead to disease in a target-based study is the starting point. Next, a potent drug is identified from hits generated based on molecular compounds that bind to the target. The probability to produce higher numbers of hits depends on the structural library available for screening and computational methods applied. The high throughput screening (HTS) approach for example consumes larger quantities of assays and depends upon the availability of a large collection of target compounds.[33]

Recently, computational methods have been extensively utilised in drug discovery studies to assist in the design of novel ligands. Structure based drug design (SBDD)[32] is one of the structural methods engaged in the drug discovery process. Essentially SBDD involves the use of the three dimensional structure as the reference to design the drug-like ligands of interest. This approach is based on an understanding of the structural interaction of the protein-ligand during binding. The binding modes and binding affinity are suggested as part of this process. The best model must posses both structural and chemical features complementary to the protein at the receptor binding site.

The massive expansion and continuous development of computational methods has led to several successful applications. Various approaches have been established to win the drug discovery battle by exploiting the structure of the protein target as a route to discover novel hit compounds. Through these approaches, many new ideas of possible interactions for compounds templated on the active site have been generated through de novo design, virtual screening and fragment-based discovery techniques.

Amongst these methods, fragment-based lead discovery has emerged as one of the most successful approaches in drug discovery.[34] This has further inspired computational methods for docking and scoring small molecules in protein binding sites. Fragment-based discovery is a method for discovering potent lead compounds against the drug target based on the assumption that most ligands that bind strongly to a protein active site can be considered as a number of smaller fragments or functionalities. Choosing fragments with molecular weight < 250 Da allows higher hit rates through the screening of smaller libraries (400-20,000 molecules using by X-ray crystallography, NMR spectroscopy or functional assay) of compounds covering a larger chemical space compared to high throughput screening involving larger compounds.[35] The ligand efficiency

for fragments, designated by the amount of free energy change per atom upon binding, remains equally good as for larger hit molecules. Optimizing hits from fragment screening has been shown to be a potential alternative to HTS.[36,37] The structures of the fragments that have been identified to bind to the protein can be used to design new ligands by adding, merging and linking functionality to the fragment or grafting fragment features to existing ligands. These methods have been mentioned in several reviews and a number of studies have been published that perform various medicinal chemistry strategies to develop low-affinity fragments into high-affinity inhibitors against different targets.[34,38–40]

Essentially, the protein-ligand binding affinity has become the measurement of drug-like ligand selection for hit confirmation validated by experimental data. While affinity has been proven to be good measure to indicate the potential of drug-like compounds, binding activity inspired by kinetics may be equally crucial. The kinetic aspect is significant in understanding the effectiveness and biophysical process of reactions but is hard to assess for several cases.[41] However, qualitatively both affinity and specificity are primarily governed by thermodynamic properties. Using these to establish structure-activity relationships has become the key to success in the drug discovery process.

1.3 Ligand binding techniques

As discussed, the quantification of ligand binding interactions to specific receptors is essential in drug development projects. The main aspects of protein-ligand binding interactions are binding affinity and kinetics, including knowledge of the conformations of the target, binding thermodynamics and ligand efficiency. In the following, we discuss a few techniques performed to assess the ligand binding properties and protein-ligand binding activity.

1.3.1 Experimental methods

Several experimental methods for the determination of binding parameters in biomolecule systems involving protein-ligand have been developed over the past few years.[42,43] Continuous development is undergoing on the techniques to achieve accurate and consistent measurements of actual affinity. Measurements are frequently dependent on the sensitivity and the strength of the method used. As reported, more than 20 methods have been applied for determining biomolecular binding kinetics and thermodynamics.[44] Generally, the methods are categorised into two categories - direct and indirect methods.[45] Direct methods take to account the actual concentration of the bound and free protein separately. In contrast, indirect methods imply the concentrations from a signal observed as proportional to the concentration of product. Overall,

the most common methods to measure the binding affinity are isothermal titration calorimetry (ITC),[46] surface plasmon resonance (SPR)[47] fluorescence-based techniques[48] and Ultraviolet-Visible (UV/Vis) absorption spectroscopy.[49] Below the main features of the methods are briefly explained.

1.3.1.1 Isothermal Titration Calorimetry (ITC)

The ITC technique is classified as a direct method to elucidate the thermodynamic contributions to binding free energies of the system.[50,51] An ITC experiment measures directly the heat uptake or release associated with the molecular interaction between two or more molecules (Figure 1.3). The uniqueness of ITC is due to the fact that the quantities derived are not only the overall binding affinity but also the enthalpy, entropy and change in heat capacity of the system as well. For example, the Freire group observed that first-in-class inhibitors were entropy optimized, using ITC techniques to clarify the thermodynamic contributions of HIV-1 protease and HMG-CoA reductase inhibitors.[50,52,53] These calorimetric studies highlighted the importance of understanding independent entropic and enthalpic contributions for optimising binding affinity in molecular design through enhancing from first-in-class to best-in-class. Despite significant use in molecular design, this approach is limited to mostly high-affinity ligands since the weak binding ligands require an intractable protein concentration to perform the experiment. In addition, the experimental drawbacks are the fact that ITC is time-consuming and labour-intensive with low throughput compare to the SPR method. However, ITC remains a valuable experimental technique to determine thermodynamic parameters even with these limitations.

1.3.1.2 Surface Plasmon Resonance (SPR)

Surface plasmon resonance (SPR) is another technique in ligand binding that produces consistent affinity values with accurate indirect measurements.[54] SPR belongs to the indirect techniques class as it uses an optical method for determining the interaction between two different molecules in which one is mobile (binding partner) and one is fixed (ligand) on a thin gold film[55] (Figure 1.3). Since the development almost a decade ago[56] of the first biosensor based on SPR, the use of this technique has gradually increased. BIAcore[56] become the most widely establish biosensor, produced by BIAcore AB, which has developed into a range of instruments influenced by this technique. The measurement of the refractive index near a sensor surface (within ~ 300 nm) is the approach in this technique. In the BIAcore, this surface forms the floor of a small flow cell, 20-60 nL in volume.[57] The interaction between the molecules is detected via the change in refractive index as the binding partner binds to the ligand, the accumulation of protein on the surface results in an increase in the refractive index. Changes observed are measured in real time, and the result plotted as response or resonance units (RUs) versus time (a sensorgram). The

real-time binding data is very meaningful for the kinetic analysis of ligand-binding activity. SPR has been established as a method to determine K_D[48] for the macromolecular binding system. Although it is the preferred method for measuring binding kinetics, the weakness of this method is in assigning k_{on}. Other methods are suggested for k_{on}[58] due to the diffusion effects in SPR. In terms of equilibrium analysis this techniques is well suited to weak interactions. High affinity interactions (K_D < 10 nM) usually have very slow k_{off} values, and are therefore unsuitable for equilibrium analysis. In contrast, very weak interactions (K_D > 100 µM) are easily studied.[59] Less protein sample is required compared to calorimetry techniques. Equilibrium affinity measurements on the BIAcore are highly reproducible as precise temperature control makes it possible to estimate binding enthalpy by van't Hoff analysis.[60]

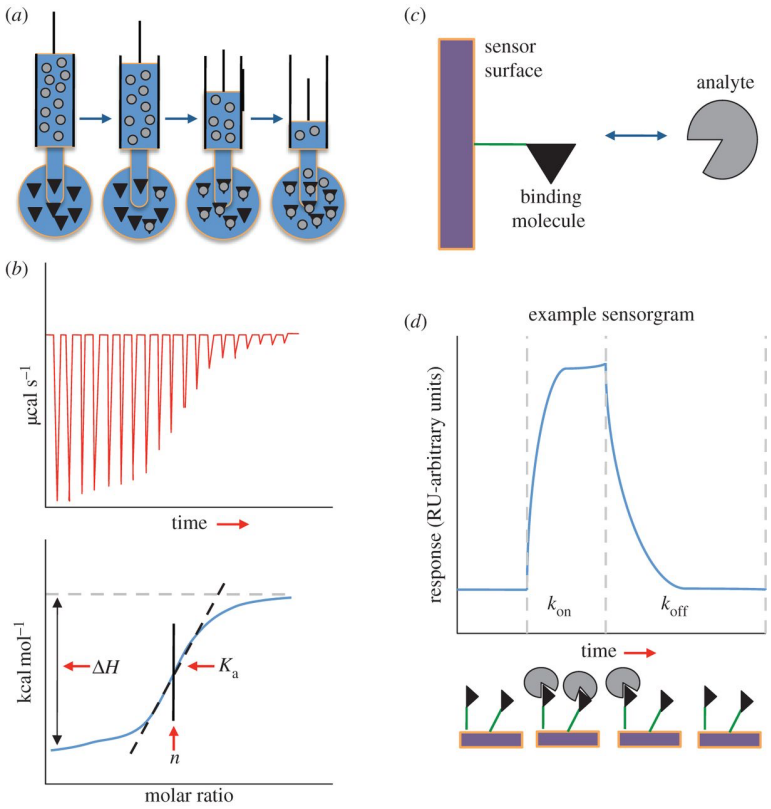


Figure 1.3: Isothermal titration calorimetry technique (ITC)- (a) Titrations used to measure heat capacity changes and (b) calculation of K_A. Surface plasmon resonance (SPR) methods- (c) SPR setup and (d) monitoring of the association/ dissociation process of the mobile agent. Figure taken from Kastritis *et al.*, 2013.[61]

1.3.1.3 Fluorescence-based Techniques

Fluorescence-based techniques have emerged as a popular way to understand protein or ligand features of biomolecular systems. Numerous applications developed on fluorescence-based techniques are highly sensitive in order to provide the dynamics of protein-ligand interaction. Questions regarding binding thermodynamics and kinetics governing the interaction may also be answered. Commonly, HTS will be the partner in this approach to identify the potential ligands to the protein target. Fluorescence-based techniques can be categorised into simple fluorescence, fluorescence polarisation and Förster resonance energy transfer approaches. In most of these methods, competitive binding assays are used in which a labelled ligand molecule is bound and subsequently displaced by a variety of competitive inhibitors.[48] Fluorescence-based techniques avoid the need to separate a bound and free fraction of ligand. Measurement of the concentration of the labelled ligand is essential in determining the absolute affinities. Such approaches are useful as in spectroscopic methods, on top of binding affinities, additional information can be derived including structural data, such as binding distances between the fluorophore and the protein. However, this is limited for more complicated equilibria as the response is not a direct measure of binding, but proportional to it.[45] These techniques are best applied mostly for high-affinity interactions. Diverse chemical environments and intermolecular interactions make it capable to investigate single and multiple molecules binding to the receptor. Bearing in mind that different complexes are in different environments, measurements are performed under diverse temperature, ionic strength and pH conditions. Consequently, differences in observation of the actual fluorescence levels measured in a binding experiment are strongly affected by the sensitivity of the detecting instrument. During measurement, these potential sources of errors must be treated carefully to obtain an accurate value.[62]

1.3.1.4 Ultraviolet-Visible (UV/Vis) absorption spectroscopy

UV/Vis absorption spectroscopy is recognised as a powerful tool for detecting the binding of ligand to protein in protein-ligand interactions studies. This method is sensitive to the supramolecular interaction of interest by measuring radiation absorbance, as a function of frequency (or wavelength), due to its interaction with a sample. The absorbance (or optical density) of a given sample is described by Beer's Law. UV/Vis absorption spectroscopy is a particularly useful technique in the case for the study of ligand binding associated to heme proteins, given the strongly absorbing heme prosthetic group that forms a coordination complex with the ligand using the central iron atom in the heme protein.[49,63,64] In UV/Vis absorption spectroscopy, the electronic absorption spectrum is highly sensitive to the surrounding

polypeptide environment, reflects the charge of the central heme iron, binding of ligands, and even protonation events and structural changes in its vicinity.[65]

1.3.2 Computational methods

Significant advances in high throughput experimental techniques and computer power have shed light on the enhanced development of the computational techniques in determining the molecular forces that control ligand-binding interactions. Today, computational methods are complementary to experimental methods in achieving higher rates of success and speeding up the drug discovery process. Numerous computational methods developed have opened up more opportunities to understand protein-ligand interaction in detail, often proving crucial to evaluating experiment. Previously, the need for computational approaches in drug design have been discussed in section 1.2. There are varieties of ways to predict ligand-binding interactions through computational estimation methods with different levels of accuracy, applicability and speed. These computational methods can be classified into three main approaches:[66] i) Free energy calculation based methods ii) MM-PBSA/GBSA methods iii) Docking and scoring. Further explanations of each method are discussed below.

1.3.2.1 Free Energy Calculation Based Methods

Free energy calculation based methods are rigorous accurate methods widely used to measure thermodynamic properties of systems, such as in ligand binding. The theory of free energy calculations is many decades old,[67,68] but applications have been restricted due to the lack of computational power. In early studies, triumphs in ligand-binding interaction were reported[69–71] but were limited to computing relative binding free energies and for a small number of compounds. In the past few years, tremendous development has shown free energy calculation methods to be useful when applied to molecular systems with sufficient computational resources.[72] There are two approaches in calculating binding free energies - absolute and relative[66] - which can be performed by free energy perturbation (FEP) and thermodynamic integration (TI) methods as described by Essex's group.[72] These free energy calculations are based on computational alchemy. In general, alchemical transformations perform molecular dynamics runs simulating unphysical intermediates and use rigorous statistical mechanics to calculate the free energy difference between two physically relevant states. The rigorous absolute binding free energy calculation methods are recognised as the most powerful approaches[73–75] in determining ligand-binding interaction compared to Molecular Mechanics-Poisson-Boltzmann Surface Area (MM-PBSA)[76,77] and docking.[78,79] Alchemical absolute free energy calculations show good correlation to experimental data with RMS error less than 3 kcal mol^{-1}. These may be classified as

fairly accurate,[78,80–84] and often much better,[85] depending on the particular systems studied and force field. Essentially, the system setup for this approach involves separate sets of simulation runs for solvated protein, ligand and the complex. In theory, no reference binding affinity of a reference complex is required for absolute free energy,[86] whereas relative binding free energy methods are based on calculation of the differences between the absolute free energy for related ligands of interest. Information regarding the structure and binding affinity of a related ligand for the complex is necessary as a reference in the calculation of relative binding free energies[83,85] Both methods are useful for comparing multiple ligands binding to the same complex in the drug discovery process or for experimental comparison purposes.[87] Generally, relative binding free energies refer to the difference in the binding of two compounds to the same protein receptor. This difference can be computed precisely using several different techniques[83,85] and sufficiently long molecular dynamics simulations for sampling,[66] which come with a high computational cost. However, the accuracy of calculations are highly sensitive to the choice of methodology utilised for the conformational sampling and the underlying atomic force field parameters assigned.[66,85,88] Further details of free energy calculations will be discussed in section 2.4.

1.3.2.2 MM-PBSA/GBSA Methods

The MM-PBSA or MM-GBSA methods are another computational approach that have been used in practice to determine the free energies of molecular systems including ligand-binding interactions.[84,89,90] MM stands for Molecular Mechanics, while PBSA refers to the Poisson-Boltzmann and Surface Area model or GBSA to the Generalised Born and Surface Area model. Reflected by their acronyms, MM-PBSA or GBSA methods combine molecular mechanics interaction energies with implicit solvation models for the calculation of the free energy. The GB method is based on an approximation to the PB equation.[91] These methods have been introduced in the late 1990s and continuously developed since then.[21,66,90,92,93] The principle of this method is fundamentally based on the separation of the contributions of the individual energy terms - intra-molecular, van der Waals, electrostatics and solvation - to the free energy of binding in the biomolecular interaction.[94] A particular difference between MM-PBSA/GBSA and TI/FEP is that MM-PBSA only calculates the end points of molecular interactions; it is therefore an approximation of the multiple small perturbations to alchemical intermediates as used in TI/FEP. It also has a reduced computational cost as the intensive calculation of individual contributions from each solvent molecule is ignored in this approach and replaced with an implicit PB or GB representation. This benefit improves large-scale data analysis involving HT datasets for virtual screening studies.[95,96] However, it remains a trade-off between efficiency and accuracy since ignoring the details of solvent effects may lead to incorrect contributions to binding free energies. In this approach, issues with the force field should be carefully treated as well. The parameters

used to represent the molecular systems must be good enough to generate the accurate result. Overall, MM-PBSA or GBSA methods are ranked higher than docking and scoring in terms of the accuracy of pose and affinity prediction[97] but still cannot beat rigorous free energy calculation based methods.[66]

1.3.2.3 Docking and Scoring

Docking and scoring methods are one of the most well-known computational tools for structure-based prediction of protein-ligand interactions since the invention of the DOCK technique in 1982.[98] Currently, these methods have become prominent in the modern drug discovery process and are often used in high throughput computational screening[99] due to the rapid computational speed offered. However, the accuracy of this method is less than the above-mentioned techniques in sections 1.3.2.1 and 1.3.2.2. Docking approaches concentrate more on generating structure poses and use an underlying empirical algorithm to determine the ligand-binding interactions. They only include endpoints in their calculations, similar to MM-PBSA/GBSA but different to alchemical free energy calculation based methods. Instead of running molecular dynamics to generate the conformational ensemble, docking generates a variety of potential binding poses. Each of these is scored using a scoring function in order to generate a rank order to determine the optimal binding pose. Two aspects that need to be considered in the generation of a ligand's best binding modes are ligand sampling and protein flexibility.[100] Both elements, ligand and protein - contribute to the binding affinity and specificity desired in drug design.[101] Approximations and scoring function algorithms also vary, which also contributes to the accuracy of the binding affinity predictions. Over past few years, significant progress in scoring functions has been made,[102–105] roughly categorised into three areas - force field based, empirical and knowledge-based approaches.[97] Although improvements have been made, the prediction of binding affinity with docking and scoring methods remains challenging in most cases.[105]

1.4 Potential energy function calculations

Various computational methods have been developed over the years to predict free energies in biomolecular systems, ranging from more to less rigorous methods as discussed previously.[66] However, which computational strategy to choose is the hardest part, as a clear answer depends on the particular goals of the study. Frequently, more accurate predictions with lower cost and time consumption are becoming the main aspects of consideration in choosing an appropriate method. To achieve high accuracy in predicting experimentas using any computational approach requires sufficient conformational sampling with an accurate underlying interatomic model to

represent the intermolecular interactions in the systems. With the intensive existing MD enhanced sampling methods[2,106,107] such as replica exchange or Monte Carlo techniques,[108] we will be able to deal with many of the sampling issues. However, the accuracy issues associated with the force field models still remain a major problem in free energy calculations.

1.4.1 Motivation

As pointed out before, the power of computational methods in broadening our knowledge about biomolecular forces, binding and mechanisms is undeniable. The major impact of computational methods on the modern drug discovery process is shown in the increased numbers of marketed drugs.[109] Vast advances in computational techniques have been introduced to accelerate the drug discovery process, aimed at fast and accurate methods for saving time and cost through reducing the number of compounds needed for synthesis and testing.[110,111]

As these methods currently stand they are not sufficiently accurate, as discussed previously in section 1.3.2. The ultimate challenge in most computational techniques is the accuracy of the underlying molecular mechanics force field model representing the interatomic interactions in complex systems. The extensive sampling with longer time alone will not lead to better results[97] without good quality parameters used to describe system features. Thus, an accurate potential energy function is a prerequisite for production of representative conformational ensembles for estimating free energies of molecules in systems.

Currently, the existing fixed-point-charge force fields,[112–115] may not be sufficiently accurate to ensure that calculated free energies are reliable. Classic fixed-point-charge models include the changes in electrostatic interaction by taking account of the polarisation implicitly, but limit the ability to fully adapt to the environment. In an effort to improve accuracy of interatomic potential for biomolecular interaction, the AMOEBA (Atomic Multipole Optimized Energetics for Biomolecular Application) force field has been introduced.[116] AMOEBA is an advanced potential energy function including a polarisation molecular mechanics model, designed to directly treat the polarisation effect by incorporating an explicit response to the environment. The ability of AMOEBA force field to capture this effect may be expected to give more accurate prediction of interaction energetics in systems. Greater details of AMOEBA force field will be discussed in chapter 2.

1.4.2 Aims

The ability to accurately represent the interatomic interactions for description of molecular systems remains a main challenge in force field development and molecular recognition applications. In this study we will tackle this problem head on by determining whether a better description of electrostatics term in potential energy function by the explicit inclusion of electronic polarisation is able to improve the accuracy of its free energy calculations over a traditional fixed, atom-centred, point charge potential energy function.

To investigate the performance of AMOEBA and where the success over the existing fixed-point-charge method may lies. We firstly, evaluated simple systems to see the advantages of the AMOEBA force fields in different non-aqueous solvent by calculating solvation free energies of a small molecule data set in a range of common organic solvents (chloroform, toluene, acetonitrile and dimethylsulfoxide) using solvents of different dielectric constants. Secondly, we investigated on more complex systems, by evaluating binding free energies on a clear case where fixed-point-charge force fields to fail due to the lack of explicit polarisation. Here, we tested on the cytochrome c peroxidase protein to investigate the effect of inclusion of a polarisable potential in binding free energy calculation. Initially we focus on binding free energy calculations using the AMBER force field to obtain a robust and reproducible free energy protocol required for the AMOEBA free energy calculations. To assess the capability of the AMOEBA force field in, identical evaluations will be carried out on the same systems with AMOEBA. This study will determine, whether AMOEBA is able to deliver more accurate estimates of binding free energies in the areas where fixed-point-charge model are unreliable.

Chapter 2: Theory

In this chapter, some theory underlying the computational approaches for the evaluation of potential energy functions by free energy calculations will be described. The AMOEBA advanced potential energy function implemented in this study will be compared against the classic fixed-point-charge model to distinguish the improvement arising from the explicit polarisation term. Then the simulation details to generate a molecular dynamics simulation for this study, as well as the theory and background of the free energy calculation techniques and corrections in free energy calculations will be discussed.

2.1 Molecular mechanics and force field

In a typical Molecular Mechanics (MM) potential energy function, atoms are represented as soft spheres with partial charges. These are built up to molecules by bonds modelled as springs. To calculate the potential energy of the system, a set of parameters needs to be assigned to each atom and atom pair. These parameters, illustrated in Figure 2.1, consist of five key terms- 1) bond stretching 2) angle bending 3) torsional angles 4) van der Waals (Lennard-Jones) and 5) electrostatics (Coulomb law), corresponding to the general functional form in the equation below:

$$U_{pot} = \sum_{bond} \frac{k_l}{2} \left(l - l_{eq}\right)^2 + \sum_{angles} \frac{k_\theta}{2} \left(\theta - \theta_{eq}\right)^2 + \sum_{dihedrals} \frac{V_n}{2} \left[1 + \cos(n\omega - \gamma)\right] + \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left(\underbrace{4\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6} \right]}_{Lennard-Jones\ 12-6\ potential} + \underbrace{\frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}}}_{Coulomb\ law} \right) \quad (2.1)$$

The first three terms specify bonded interactions between atoms adjacent to each other. The bonds and angles are modelled by harmonic potentials with the force constants k. Here, the energy is related to the deviation from ideal bond length and angle between two atoms. These are calculated with symbols l and θ denoting current length and angle and l_{eq} and θ_{eq} equilibrium structure parameters for bond length and angle. The third term refers to a torsional potential, calculated as a Fourier series, so there may be more than one set of parameters describing the same dihedral. Here ω is the observed angle, γ is the phase angle, V_n is the height of barrier and n periodicity. The fourth and the fifth contributions are for non-bonded interactions describing interactions between two atoms, incorporating both van der Waals and electrostatics. A Lennard-Jones 12-6 potential is commonly used to model the van der Waals (vdW) interactions, and Coloumb's law for the potential term of the electrostatic interactions. In

these equations r_{ij} defines the distance between atoms i and j, σ is an vdW radius, q is the point charge, ε_{ij} is well depth and ε_0 is the vacuum permittivity constant. Therefore, all of these sets of parameters must be summarised into the force field.

A force field is a core element that drives simulation of the Newtonian dynamics of the system (section 2.2) as it captures physicochemical properties of the interacting atoms in molecular interactions. Accurate MD simulations of inter- and intramolecular interactions requires precise representation of electrostatics[117] and van der Waals forces.[118] Various force fields exist, for example AMBER,[112] CFF,[119] MM3 / 4,[120,121] TRIPOS 5.2,[122] CHARMM,[113] OPLS-AA and GROMOS[114] to represent different molecules in different environments and they are mostly developed using both empirical knowledge and detailed calculations based on experimental data or quantum mechanical calculations.[123] Since all force fields are designed with different sets of parameters, and often with slightly different energy terms to calculate the total potential energy of a system, here we will describe one classical fixed-point-charge force field, the General AMBER force field (GAFF), and one next generation force field, the AMOEBA polarisable force field, both of which have been implemented for free energy calculations in this study.

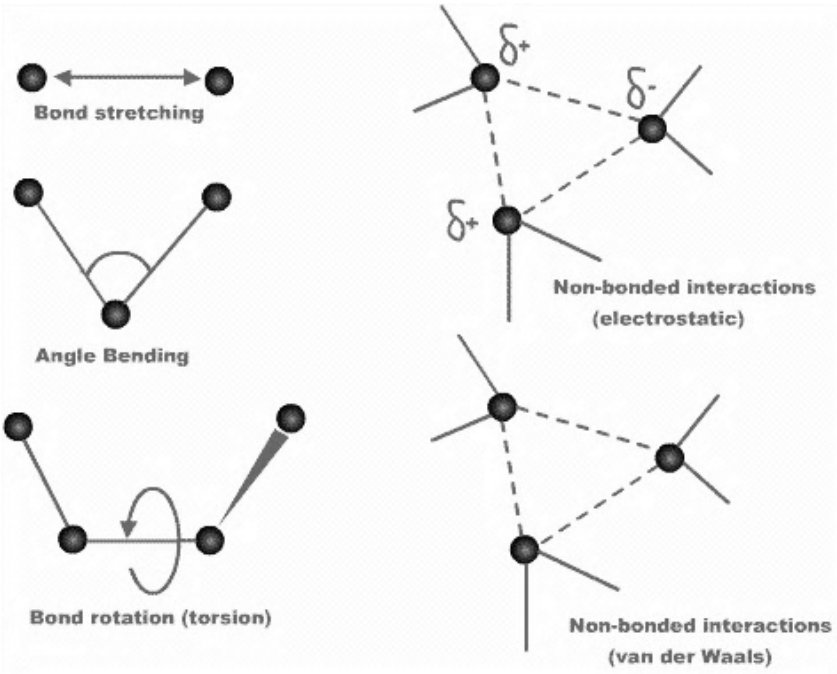


Figure 2.1: Schematic representation of the bonded (bond stretching, angle bending and bond rotation) and non-bonded (electrostatics and van der Waals) interactions contributing to a molecular mechanics force field's general functional form. (Taken from Leach, 2001)[124]

2.1.1 Classic fixed-point-charge force field- AMBER

Classic fixed-point-charge force fields, also known as pairwise additive force fields, have been the conventional model for inquiries into microscopic and macroscopic phenomena in chemistry and biology since the 1970s.[125] This class of force field model takes account of atomic polarisation in an implicit manner by using a fixed-point-charge electrostatic model, serving as an inexpensive description of polarisation. This approximation limits the capability of the model to adapt between various systems[115] as it cannot fully capture changes in many-body effects such as electronic polarisation.[126]

Several fixed-point-charge force fields have been developed such as AMBER,[112] CHARMM,[113] GROMOS[115] and OPLS.[114] Among them, the General Amber Force Field (GAFF) is one of the most widely used fixed-point-charge models utilised in drug design and other studies of ligand-protein or ligand-DNA interactions. GAFF is designed to be compatible with existing Amber force fields[127] and provides a parameter set particularly for small organic molecules to facilitate biomolecular simulation of small molecule ligands or even ligand binding interactions. Essentially, GAFF is composed of the same simple functional form for potential energy as in equation 2.1.[127]

The equilibrium bond parameters for length, angles and torsion are encoded based on atom type in GAFF. These parameters were assigned from experiment, ab initio calculations and the AMBER protein force fields. In order to precisely define the parameters, mean values are compared among the three resources.

In GAFF, non-bonded energies are contributed by electrostatic and van der Waals interactions. The electrostatic interactions between two interacting molecules are described by a static electric potential where each atom centre is assigned a partial point charge (monopole) and no higher electrostatic moments. Owing to this fact, a Coloumbic potential term is used to calculate the electrostatic potential energy for the interaction. Several charge models exist to designate the charges, however in GAFF the partial charges are commonly assigned using AM1-BCC (bond charge correction).[128] This is an alternative charge scheme designed to recreate charges from electrostatics modelled at RESP HF/6-31G* level and much cheaper to derive. For van der Waals parameters, the Lennard-Jones potential is implemented consistent with the Amber parm94 and parm99 force fields.[127] This potential consists of two components, a repulsion term and an attractive dispersion term. The repulsion decays as r^{-12}, and approximates the exchange repulsion between cores of atoms. Meanwhile the dipersion r^{-6} term covers the attractive van der Waals interactions between atoms.

2.1.2 An advanced polarisable force field- AMOEBA

In an effort to advance molecular mechanics force fields, numerous polarisable force fields have emerged to handle situations beyond the fixed-point-charge model. By including an explicit polarisation term in the functional form the transferability of parameters can be improved by accurately representing the chemical conformations and interaction energies in the systems. Therefore, a variety of approaches have been proposed for modelling the polarisation effect, such as induced dipoles,[112,116,129–140] fluctuating charges[141–147] and Drude oscillator models.[133,148–152] The AMOEBA force field is one of the most developed polarisable force fields and adopts the induced dipole approach as a way to address molecular polarisation. It was introduced by Ponder and co workers in 2002.[116]

The AMOEBA acronym stands for 'Atomic Multipole Optimised Energetics for Biomolecular Application'. As can be seen from the acronym, the AMOEBA polarisable force field is a model based on multipole electrostatics, 'optimised' with induced dipoles for polarisation effects, and applied for biomolecules, including parameters for small molecules, proteins, nucleic acids and ions. [153–155]

Fundamentally, the functional form of the AMOEBA force field is taken from the MM3 force field model.[120] AMOEBA was initially developed as an improved representation of electrostatics.[116] However, through continuous development, initiated with a water model, the AMOEBA parameters have been extended to other complex molecules. The general functional form for the AMOEBA force field is described as below.[116,140,156]

$$U_{pot} = U_{bond} + U_{angle} + U_{b\theta} + U_{oop} + U_{torsion} + U_{vdW} + U_{ele}^{perm} + U_{ele}^{ind} \quad (2.2)$$

As in the functional form of other force fields, the AMOEBA potential function is described by bonded and non-bonded interactions. The first five terms in equation 2.2 refer to valence interactions. The valence terms are similar to GAFF but with anharmonicity corrections. Instead of a description based simply on bonds, angles and torsions as in GAFF, full intramolecular flexibility with a bond-angle cross term (stretch-bend or Urey-Bradley) and out-of-plane bending for improper dihedral angles are included explicitly. Commonly, an up-to six-term traditional Fourier series expansion is used for estimation of torsional energies. The pi-torsion and torsion-torsion coupling terms utilise a grid-based[157] correction for torsions in rings and planes.

The last three terms in the equation above are the non-bonded interactions composed again by van der Waals and electrostatics. To accurately model short-range van der Waals interactions, the repulsion and dispersion are represented with a Halgren's buffered 14-7 potential form.[158] This function, produces a softer repulsive region compared to the Lennard-Jones 12-6 function and

better fits gas phase *ab initio* calculations and liquid properties of noble gases.[158] In AMOEBA, the van der Waals parameters are generated based on the best fits to experimental properties in gas and bulk phase.

In contrast to fixed-point-charge, the polarisable model treats the electrostatic interactions with higher order multipole moments up to quadrupoles. The electrostatic energies of interacting molecules are contributed from permanent atomic multipoles (monopole, dipole and quadrupoles) and induced dipoles at every atomic site. The atomic charge distributions (multipoles) are derived from high-level gas-phase QM calculations followed by distributed multipole analysis (DMA).[159,160]

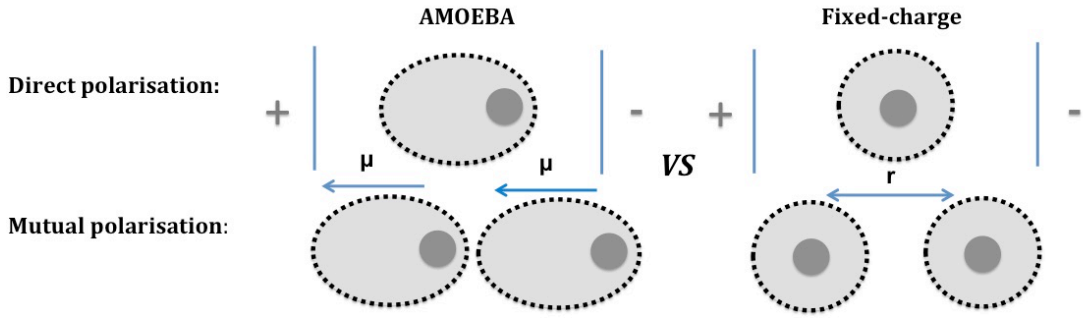


Figure 2.2: A schematic of the representation of polarisation effects in the AMOEBA polarisable force field and GAFF fixed-point-charge force field.

In the AMOEBA induced dipole model both direct polarisation from multipoles and mutual polarisation from induced dipoles are taken into account (Figure 2.2). In direct polarisation, the permanent multipoles create the field and point dipoles are assigned at each polarisable site.[161] This is followed by mutual polarisation in which the induced dipoles themselves will induced a dipole at each polarisable site according to the electric field felt by that site which in turn induces new dipoles and further polarises other atoms. This process continues until dipoles converge to a fixed precision. Finally, the same polarisation model is used for both intermolecular and intramolecular interactions. However, a group-based approach for modelling intramolecular polarisation is used. The permanent multipoles of atoms within a group do not polarise one another. These groups or fragments are then merged to represent the larger molecules for intermolecular interactions in the system. In AMOEBA, Thole's[162] damped model is utilised to avoid the polarisation catastrophe during the short range polarisation interactions. A Thole damping factor dampens polarisation interactions by replaces one of the point charges with a smeared-out charge distribution denoted by equation 2.3.

$$\rho_{Thole} = \frac{3a}{4\pi}\exp(-au^3), \; where\; u = \frac{R_{ij}}{(\alpha_i\,\alpha_j)^{1/6}} \qquad (2.3)$$

Where, a is a unitless damping coefficient , α_i is atomic polarisibility at site i, α_j is atomic polarisibility at site j, and R_{ij} is the linear separation between the site i, j.

Overall this significant improvement by incorporating a polarisation term in the AMOEBA force field model is expected to more accurately represent the structural and thermodynamic properties of atomic interactions in systems.

2.2 Molecular dynamics

Studying macromolecules at an atomic level by experimental techniques is very complex. It is both time consuming and costly, and besides provides limited insight into mechanistic details of the biomolecules. Ultimately, molecular dynamics (MD) simulations are exclusively useful in this respect, allowing atomistic insight into the dynamics of biomolecules in a solvated state, and providing a direct route from microscopic details of the system to macroscopic properties of experimental interest. The dynamics of a system composed of thousands of atoms can be explored by several approximations and techniques.

Molecular dynamics can be classified as a classical simulation technique applied for gases, liquids and solids. Undertaken by classical molecular mechanics, the electronic motions around atoms are neglected and only nuclear motion is treated, consistent with the Born-Oppenheimer approximation.[124] The instantaneous response of the electron cloud to the shift of the nuclei position means that we can describe the energetics of the systems used in MD with a classical interatomic potential.[163]

In MD, the system is simulated as a function of time by evaluating the equations of motion for a set of atoms. Therefore, the dynamics generated by an MD trajectory hold a detailed description of the position and momenta of each atom in the system over the time scale of the simulation. The trajectories are computed by approximating the equations of motion through numerical integration. By integrating Newton's laws of motion relating the net force $\boldsymbol{F}$ directly to the mass and acceleration of atoms, where m is the mass, $\boldsymbol{a}$ the acceleration and $\boldsymbol{v}$ the velocity of the atom i, the motion of atoms in the system can be solved. The force exerted on an atom causes an acceleration and the accelerations are used to update the velocities and positions $\boldsymbol{r}$. Therefore, both positions and velocities at time t can be related to atomic masses and a set of forces generated by a suitable potential energy function. By integrating over time t, the new velocities and position can be calculated for every atom, hence propagating dynamics over time as expressed in the following:

$$\boldsymbol{F}_i = m_i \boldsymbol{a}_i \tag{2.4}$$

$$\frac{d\boldsymbol{v}_i(t)}{dt} = \frac{\boldsymbol{F}_i}{m_i} \tag{2.5}$$

$$\frac{d\boldsymbol{r}_i(t)}{dt} = \boldsymbol{v}_i \tag{2.6}$$

The first MD simulation using a simple model was performed in 1957 by Alder and Wainwright[164] on 32 hard-spherical particles. In this model, no forces were added to the particles, hence positions by time after collisions were easy to represent analytically. It has become the cornerstone in the study of dynamics. Realistically, in intermolecular interactions, the force acting on each atom in the system is a function of all other atoms in the system, thus whenever the positions of the atoms change they will evolve together. This is a coupled process, which cannot be solved by simple differential equations. In 1964, molecular dynamics with continuous potentials were first applied in the simulation of argon performed by Rahman.[165] This potential did not allow for an analytical solution of the complex differential equation of molecular motion. Instead, a finite difference method was employed in solving the equations of motion to generate the molecular dynamics simulation. These approaches begin with approximating the accelerations, the velocities and positions at small time increments ($t + \delta t$) as Taylor series expansions, e.g.. [124]

$$\boldsymbol{r}(t + \delta t) = \boldsymbol{r}(t) + \delta t\, \boldsymbol{v}(t) + \frac{1}{2}\delta t^2\, \boldsymbol{a}(t) + \frac{1}{6}\delta t^3\, \boldsymbol{b}(t) + \frac{1}{24}\delta t^4\, \boldsymbol{c}(t) + \cdots \tag{2.7}$$

Respectively, $\boldsymbol{r}$, $\boldsymbol{v}$, t and δt are the position, velocity, time and time step. In MD simulations of biomolecules, the initial velocities are assigned randomly from a Maxwell-Boltzmann distribution and the time step is usually set to 1-2 fs to properly sample the fastest vibrations. The time step can be increased by e.g. the SHAKE constraint method that is applied to constrict hydrogen bond vibration, allowing a longer time step to be used, typically set to 2 fs [166]. A wide variety of numerical integration algorithms has been proposed, however the most commonly used are the Verlet algorithm (Verlet, 1967) and leapfrog algorithm.[167]

2.3 Molecular dynamics simulation

Molecular dynamics simulations are often used to calculate thermodynamic properties of a system. Potential energy functions describe the interaction energies. This is used to derive forces on each of the atoms. The simulation proceeds as described in equation 2.6 from an initial time point. As we discussed earlier in section 2.1, MD simulations involve the calculations of the force acting on all atoms followed by an integration step that updates the position, velocity and acceleration for each atom over small periods of time according to Newton's laws. In AMBER and TINKER a variety of underlying integration algorithms are available. Each starts from the same

laws of motion but varies in execution. Typically, the velocity Verlet[168] is used as default in AMBER and is employed in these simulations. In TINKER the same integrator, velocity Verlet, is used here for consistency. Both of these are for our liquid phase simulations. For the gas phase a stochastic integrator may be used to generate the dynamics trajectories. This is effective in order to sample properly, as the lack of explicit solvent molecules to cause changes in solute conformation may adversely affect sampling efficiency otherwise.

Time steps in simulations are restricted by sampling of the highest frequency vibrations in a system. Typically, these are vibrations of bonds involving hydrogen, and in general 1 fs time steps are sufficiently short to properly sample these motions. However there is no hard regulation for choosing an appropriate time step for MD simulation. Clearly, the choice must be appropriate for dealing with the sampling and the stability issues in a variety of integration algorithms.[169] Besides, the different types of motion present in various systems require different time steps.[124] To enable longer time steps to be used a constrained dynamics method can be applied, constraining the fastest motions in a system such that they can be sampled with longer time steps.

The value of the time step also determines the length of the MD simulation. The number of steps integrated with the equations of motion is multiplied by the time step between these steps. In the solvation free energy simulations herein, for example, consisting of 2 million ($2x10^6$) steps with a time step of 1 fs (10^{-15} s), simulations generate the dynamics of molecules over 2 ns ($2x10^{-9}$ s). Bear in mind that each step taken in MD simulation demands additional computer expenses. How long a simulation needs to be run is dependent on the systems and properties of interest,[170] computational capabilities[171–175] and sampling issues. For reliability, the simulation should be tested to see whether or not has converged to equilibrium before relying on the property averages calculated from it.

The chosen representative ensemble will also affect the behaviour of the system and further describes the key conditions of the simulation. The conditions are symbolised by N, V, P, T and E where N is a constant number of particles, V is constant volume, T constant temperature, P constant pressure and E is constant total energy. The choice of constant conditions fixes the ensemble employed in the system. Common ensembles are microcanonical (NVE), canonical (NVT) and isothermal-isobaric (NPT). The system was evolved according to the equations of the motions under constant pressure and temperature to achieve equilibrium. Under NPT, in addition to a thermostat controlling temperature, a barostat controls changes in the volume of the system to reach the target pressure. Therefore the time averages from simulations will equal the isothermal-isobaric ensemble averages. Free energies calculated under NPT correspond to the experimental Gibbs free energies (G) rather than the Helmholtz free energies (A) of NVT.

Thermostats and barostats[176–179] are the algorithms that control the temperature and the pressure in MD simulations. Thermostats work by coupling the atoms with a virtual heat bath, whereas barostats compute the pressure by adjusting the volume by scaling the box size in response to pressure fluctuations.[180] A small selection of thermostats that exist are Nosé-Hoover,[181] Berendsen, Andersen and Langevin, while Berendsen[182] and Martyna, Tuckerman and Klein (MTK) are commonly used for barostats.[183] However, among these constant temperature and pressure approaches, the Langevin[184–186] thermostat, Nosé-Hoover[181] thermostat and Berendsen barostat were implemented here.[182]

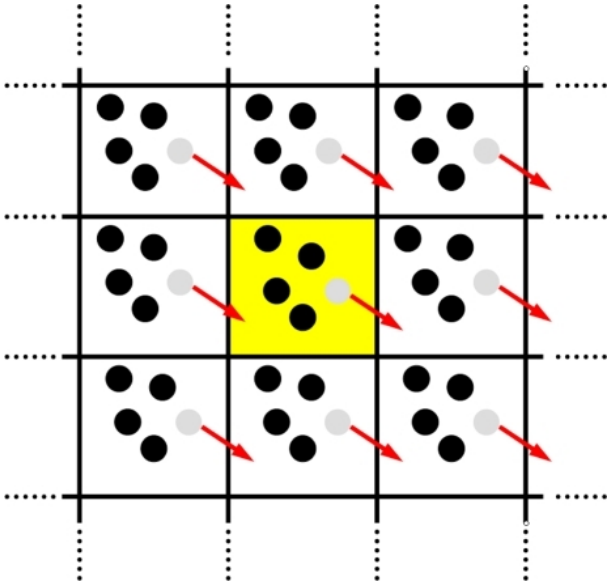


Figure 2.3: Schematic 2D representation of periodic boundary conditions. Each periodic image is an exact replica of the original box at the centre. Solvent and directions of motion are represented by red arrow. As a molecule exits one side of the periodic box, its image enters the opposite side. Figure taken from Steinhauser and Hiermaier, 2009.[187]

Periodic boundary conditions (PBC) were employed for simulations of molecules in the liquid phase to enable the macroscopic properties to be calculated using a relatively small number of particles. This technique represents the system as an infinitely tessellating central box surrounded by images of the same molecules in all directions, which replicate from the original box at the centre as shown in Figure 2.3. In the course of the simulation, the molecule that is close to the boundary is allowed to interact with those in neighbouring systems. As a molecule crosses the boundary of the box, the same molecule enters from the image box on the other side. Thus the number of molecules within the central box is conserved.

Even though PBC are often used in MD simulations, they have a limitation in simulations involving small box sizes, especially for solvation free energy calculations. The direct cut-off length for non-bonded interactions must be less than the half the size of the cell length. Otherwise

molecules will interact with their own images, causing incorrect interaction energies and possibly instability in a system. To handle long-range electrostatics interactions in these cases, Particle Mesh Ewald (PME) summation is commonly used.

An Ewald summation partitions the electrostatic interactions into short-range (real space) and long-range (reciprocal space) parts. Both real space and reciprocal space calculations converge quicker than explicitly calculating the pairwise Coulombic interactions. The Ewald sum is therefore a faster way of obtaining the total sum of electrostatic interactions than using the explicit Coulomb term. PME was developed by Darden in 1993[188] for long-range by using fifth order B-spline interpolation. Now, PME is predominantly implemented in MD simulations for fast and accurate treatment of electrostatics energies in a system.

2.4 Free energies calculations

Free energy calculations are often used to evaluate the macroscopic properties of molecular systems from microscopic simulation trajectories. The changes in free energies between two physically relevant states can be evaluated via statistical mechanics, for example, examining the thermodynamics of conformational changes in the molecule. Free energy is often expressed as the standard state Helmholtz free energy, A, corresponding to the NVT ensemble or Gibbs free energy, G, appropriate to the NPT ensemble. The partition function, 'a sum of states', can be used to connect the property to the macroscopic ensemble average from MD simulations. Generally, the free energy of our systems can be defined through the partition function by:

$$A_\lambda = -k_b\, T \ln Z_\lambda \tag{2.8}$$

Here k_b is Boltzmann's constant, T is the absolute temperature and Z is the partition function. Additionally the evaluation of the energy difference between two states is given as:

$$\Delta A_{0\to 1} = A_1 - A_0 \tag{2.9}$$

Since the free energy is a state function, the free energy difference can be calculated by scaling between states 0 and 1 through coupling to the variable λ, treating the partition function at each intermediate step as Z_λ.

The values above are ensemble averages from all states and simulations cannot access every single state. The ergodic hypothesis assumes that our ensemble average is equal to the time average from our simulations, thus an ensemble average can be evaluated by considering the relation of equation 2.8 above. There are many methods that have been proposed for calculating rigorous free energy differences: Free Energy Perturbation (FEP), Thermodynamic Integration (TI)

and Bennett's Acceptance Ratio (BAR) amongst them. These methods will be explained in the following section.

2.4.1 Free Energy Perturbation (FEP)

Free Energy Perturbation, also known as exponential averaging, is a method first attributed to the Zwanzig formula (1954).[68] It evaluates the potential energy differences between two states as follows:

$$A_{0\to1} = A_1 - A_O$$

$$= -k_b\,T \ln \frac{Z_1}{Z_0}$$

$$= -k_b\,T \ln \frac{\int_v e^{-\beta U_1(r)} dr}{Z_0}$$

$$= -k_b\,T \ln \frac{\int_v e^{-\beta U_1(r)} \times 1 dr}{Z_0}$$

$$= -k_b\,T \ln \frac{\int_v e^{-\beta U_1\,(r)}\, e^{\beta [U_0\,(r) - U_0\,(r)]} dr}{Z_0} \qquad , \; 1 = e^{\beta [U_0\,(r) - U_0\,(r)]}$$

$$= -k_b\,T \ln \frac{\int_v e^{-\beta U_0\,(r)}\, e^{-\beta [U_1\,(r) - U_0\,(r)]} dr}{Z_0}$$

$$= -k_b\,T \ln \langle e^{-\beta [U_1 - U_0]} \rangle_0$$

$$= -k_b\,T \ln \langle e^{-\beta \Delta U} \rangle_0 \qquad , \; \Delta U = U_1 - U_0 \qquad (2.10)$$

By considering two states as 0 and 1, the difference in energy between these two states is simply nothing more than an ensemble average taken over a simulation run for state 0 denoted by triangular brackets. In order to evaluate an ensemble average, we could run a simulation either state 0 or 1 and collect statistics. However, problems arise when state 0 and 1 do not overlap in phase space. In the case of one simulation visiting rare conformations, the exponential term can lead to poor convergence as the small overlap between states may result in a hugely different potential energy of one state to another state. The FEP calculation is carried out by evaluating conformation 1 using the parameters of state 0. Since the evaluation is only based on the sum of potential energy in the states, each trajectory can have a great effect on the free energy estimated. In order to get the free energy estimated right, efficient sampling is required to avoid

the large differences arising from the exponential average calculations. Therefore it is commonly necessary to divide the calculation into a series of 'windows' that require smaller perturbation steps.

2.4.2 Thermodynamics Integration (TI)

Thermodynamic Integration, also known as the integration method, is an alternative way to calculate the free energy differences given by:

$$A_{0\to1} = A_1 - A_O$$

$$= \int_0^1 \langle \frac{\partial U_\lambda}{\partial \lambda} \rangle_\lambda \; d\lambda \qquad (2.11)$$

The expression of TI is derived from an identical starting point to the exponential average estimator, but in a way less susceptible to issues of sampling. Rather than evaluate the energy differences at two states by an exponential average, the integration method calculates the gradient of the potential energy with respect to lambda, $\frac{\partial U_\lambda}{\partial \lambda}$ and integrates the total energy across all lambda by, for example, the Trapezium Rule.[189] The difficulty with TI is how to appropriately evaluate the gradient, $\frac{\partial U_\lambda}{\partial \lambda}$ compared to exponential averaging, which involves a straightforward calculation of potential energy at the next λ window. To represent the gradient as accurately as possible, one can perform a very small free energy perturbation with a very small λ window forward and backward to get an instantaneous gradient for each simulation. This is known as finite difference TI.

2.4.3 Bennett's Acceptance Ratio (BAR)

The Bennett acceptance ratio (BAR)[190] method was derived to address the problem of the exponential averaging approach where the rare states give large differences in free energy calculations. To minimise the statistical error an arbitrary weight function is adopted to weight the forward and backward free energy differences between the states sampled at each λ window. In BAR methods, two neighbouring simulations/ lambda windows are observed simultaneously, rather than one (forwards or backwards) as for FEP. A set of equations is solved for the weighting function iteratively until convergence, giving a single free energy for both neighbouring simulations. Overall, this method provides more accurate calculations of free energies in the systems compared to TI.[189] Alternatively, an extension of BAR called multistate BAR (MBAR),[191] which incorporates the information from all intermediate states in a single estimate, can be used.

2.4.4 Enhanced sampling methods

Fundamentally, free energy calculations can be categorised into two types: i) Alchemical free energy calculations ii) Conformational free energy calculations. Commonly, alchemical free energies were utilised in the study of free energy differences in the molecular interactions. The calculations involve a non-physical reaction coordinate of λ to connect the initial and final states by simulating unphysical intermediates. Thus, the sampling of the actual choice of coordinate is limited.

Sufficient sampling is crucial in order to obtain accurate and converged free energy differences in the free energy calculations. However, convergence in free energy difference is hard to achieve particularly in systems involving slow structural transitions or large environmental reorganization as λ changes.[192,193] Therefore, enhanced sampling methods are required to accelerate the conformational sampling to yield accurate and converged free energies. Replica exchange molecular dynamics (REMD),[194] is one of the enhanced sampling methods that is frequently applied to free energy calculations. The details of this method will be explained in the following section.

2.4.4.1 Replica exchange

Replica Exchange Molecular Dynamics (REMD), an enhanced sampling method was introduced by Sugita and Okamoto in 1999.[194] REMD method is one of particular interest because it does not require the potential energy surface but the weight of each state is a priori known (Boltzmann factor). In REMD simulations, N non-interacting copies (replicas) of a system are simulated at N different temperatures, wherein normal MD simulations are performed with an exchange of configurations between two adjacent temperatures periodically attempted.

A schematic of replica exchange simulation is shown in Figure 2.4. Each replica represents a normal MD simulation run for certain time period before stopping to attempt exchange with the adjacent replicas. Then, adjacent replicas are swapped if the Metropolis criterion is satisfied utilising a Monte Carlo test (configurations either swap or not based on their potential energies and temperatures). This process is repeated until swaps across temperature spaces are achieved and the full potential energy surface is explored. The potential energy will be increased and likely to swap to a higher temperature if the replica approaches an energy barrier, while if at this higher temperature, replica is able to overcome an energy barrier it is then likely to swap back to lower temperature.

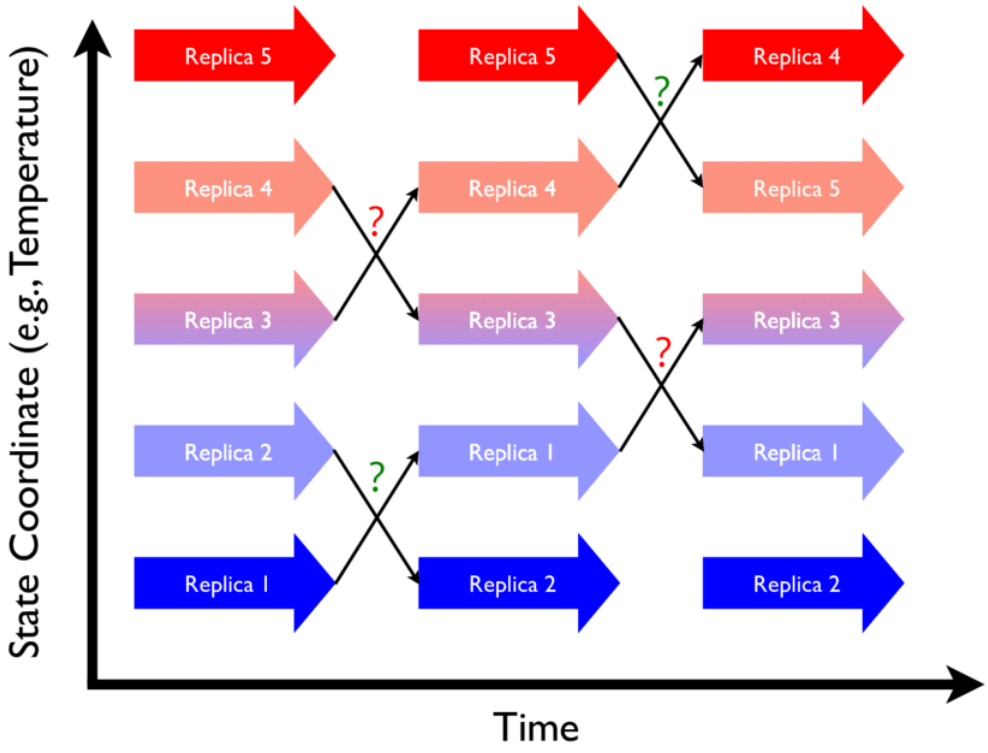


Figure 2.4: Ilustration of replica exchange for five replicas. MD trajectories represent by large arrows while attempted swaps between replicas are represent by small arrows. The question marks coloured in green indicates successful exchange while red indicated failed exchange. (Taken from AMBER16 manual)[195]

2.4.4.1.1 Hamiltonian replica exchange

Hamiltonian replica exchange (HREX)[196] is one of many variant of the original REMD method.[197–200] In HREX the replicas exchanges are between the 'Hamiltonian spaces' instead of 'temperature spaces' in REMD. The replicas differ in their Hamiltonian (potential energies) but have the same temperature. Given the total potential energy of the system consists of different terms there is a possibility of scaling parts of the potential function with the canonical distribution in each replica remains.

In HREX the exchanges between the adjacent replicas involve the exchanging of the coordinates and energies are evaluated from that configuration. This exchange attempt is based on a Monte Carlo-Metropolis criterion test [201] for evaluating whether the attempted swap of structures between two replicas should be accepted or not. The exchange probability for each molecular structure in replicas employed the detailed balance condition given in equation below:

$$P_{i\to j} = \min\{1, \exp\ (-\beta_1\ [H_1(x_2) - \ H_1(x_1)\ ] - \beta_2[H_2\ (x_1) - \ H_2(x_2)\ ])\ \} \qquad (2.12)$$

Where, states i is the replica with the combination of $\beta_1 H_1(x_1)$, $\beta_2 H_2(x_2)$,whlle states j refer to the replica with the combination of $\beta_1 H_1(x_2)$, $\beta_2 H_2(x_1)$. According to this equation (Equation 2.12) only the coordinate of each replica are exchanged however the temperature remains.

2.4.5 Corrections in free energy calculations

2.4.5.1 PME corrections

In molecular dynamics simulations for generating the binding free energies herein, the PME method was used for treating long-range electrostatics as explained in the previous section (Section 2.3). Several corrections need to be applied to the electrostatic free energies for non-neutral solute in PME simulations, involving the alchemical changes in the net charge (in solution and complex with protein) defined by Kastenholz and Hunenberger.[202] The changes in the net charge of the system in periodic simulations during the alchemical transformations creates artefact due to the finite size of the systems.[202,203] The finite size corrections for alchemical change in net charge to charging energies computed using PME calculations involves: i) Type B correction ii) Type C correction. Here, Type B and C corrections were applied to our calculations, the details of Type B and C correction will be discussed below:

2.4.5.1.1 Type B correction

Type B corrections are applied for correcting the artefact for charging an ion in a periodic box of pure solvent caused by: i) The solute-solute periodic copies interactions ii) The solute-neutralizing background charge interactions iii) The missing solute-solvent interactions beyond the length of the unit cell. The analytical correction for this artefact, ΔG_B (in kcal mol^{-1}) is denoted by the equation below: [203]

$$\Delta G_B = \frac{q^2}{4\pi\epsilon_0}\frac{1}{2L}\left\{\epsilon_S^{-1}\,\xi_{EW} - \left(\epsilon_i^{-1} - \epsilon_S^{-1}\right) \times \left[\frac{4\pi}{3}\left(\frac{R}{L}\right)^2 - \frac{16\pi^2}{45}\left(\frac{R}{L}\right)^5\right]\right\} \qquad (2.13)$$

Respectively, q is the charge of the ion, R is the radius of gyration (Rg) of the ligand, L is the edge length of the box (as a cube), ϵ_0 is vacuum permittivity, ϵ_S is the dielectric constant of solvent, ϵ_i is the dielectric constant inside the ion and $\xi_{EW} \approx 2.837279$.[203]

2.4.5.1.2 Type C correction

Type C correction accounts for the error for charging an ion in a periodic box of pure solvent caused by the constant offset to the electrostatic potential in the simulation cell due to the potential within the molecular envelope of the solvent molecule.[202] To obtain the corrected

electrostatic energies (ion charging in solution and protein), two analytical corrections artefacts: i) Type C1, ΔG_{C1} (equation 2.14) ii) Type C2, ΔG_{C2} (equation 2.15).

$$\Delta G_{C1} = -qf\ (6\epsilon_0)^{-1}\ \rho\gamma \tag{2.14}$$

$$\Delta G_{C2} = -q\ (f-1)\ (1.68 - 2.6R_I^{-1}) \tag{2.15}$$

Where q is the charge of the ion, R_I is the radius of the ion (in nanometer unit), f is the fraction of the box filled by solvent, ϵ_0 is vacuum permittivity, ρ is the number of density of water (number of water/ volume), γ is the quadrupole moment trace of the solvent molecule (TIP3P water). The C2 corrections, ΔG_{C2} was ignored due to the small energies associated (<0.04 kcal mol-1) for the systems larger than 128 water molecules, as suggested by Rocklin *et al.*[204]

2.4.5.2 Standard states dependence corrections in binding free energy

In our binding free energy calculations, the corrections to the binding free energies were applied by restraining the ligand to the protein to it position and orientation during the discharging and decoupling process. Restraints energy regulate to the position and orientation (a set of restraints) of the ligand relative to the protein are calculated analytically by having the standard states dependence correction instead of including the volume of the systems, we include the volume corresponding to one molar standard states. The details of these corrections are explained as follows.

Generally, the standard equation for calculating the binding free energies, ΔG_{bind} for the process follows:

$$(L)_{H_2O} + (P)_{H_2O} \xrightarrow{\Delta G_{bind}} (PL)_{H_2O} \tag{2.16}$$

Where, L is the ligand, P is the protein and PL is the complex of protein and ligand. The simulations for the PL initiates with the L bound to the P followed by the unbinding process to completely separate the ligand for the direct calculation according to equation 2.16. Owing to this complexity, such calculations are very difficult to compute computationally, however some work has been published using this approch.[205] Jorgensen *et al.* made it possible to overcome this problem by introducing the suitable thermodynamic cycle using 'double annihilation method' (DAM) since it involves two annihilation processes, through splitting the calculations from the equation 2.16 to 2.17 and 2.18. [206]

$$(L)_{H_2O} \xrightarrow{\Delta G_I} (L)_{gas} \tag{2.17}$$

$$(PL)_{H_2O} \xrightarrow{\Delta G_{II}} (L)_{H_2O} + (P)_{gas} \tag{2.18}$$

$$\Delta G_I - \Delta G_{II} = \Delta G_{bind} \quad (2.19)$$

Where, ΔG_I in equation 2.17 represents the ligand transferred from the solution to the gas phase and ΔG_{II} in equation 2.18, represent the ligand bound to protein formed complex in solution phase. Here, $\Delta G_{bind}\ is$ computed according to equation 2.19.

DAM is itself associated with several issues: i) The free energies of any reaction (Equation 2.16) depend on the standard states (ΔG^0_{bind}) but DAM does not take it into account ii) The calculations of ΔG_{II} (Equation 2.18) lead to issues of the wandering ligand at the end point of calculations (Figure 2.6).

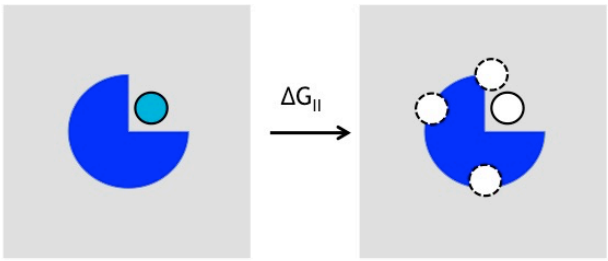


Figure 2.5: Portraying the wandering ligand, which occurs in DAM thermodynamic cycle of ΔG_{II} (Equation 2.18). The blue third quarter circle represents a protein, the cyan circles represent a fully charged ligand interacting with the environment, the no filled circles represent a discharged ligand, the no filled circle with the dotted line represents the ligand free to wander in the simulation systems, no vdW interactions, while the grey square box denotes a simulation run in solution phase. In this case, the simulation was absolutely fine at the initial states (electrostatics is on), but the problem occurs at the end states (electrostatics is off) and the ligand starts to freely move around at any point in the simulation, showed by the no filled circle with the dotted line, which leads to sampling problem in the calculations.

As solution to this problem, 'double decoupling method' (DDM) was proposed by Gilson *et al.*,[207] (similar to Roux *et al.*[75] and Wang and Hermans[74] approach), by expanding the equation 2.18 to equation 2.20 as followed:

$$(PL)_{H_2O} \xrightarrow{\Delta G^*_{II}} (P)_{H_2O} \cdots (L)_{gas} \xrightarrow{\Delta G_r} (P)_{H_2O} + (L)_{gas} \quad (2.20)$$

Here, equation 2.20 is separated to two parts: i) ΔG^*_{II}, from the PL in solution to the hypothetical states, where P in solution and L in gas with the protein and ligand held by the restraints during decoupling of the ligand (no interaction with the protein and solvent). i) ΔG_r computes for the cost of restraining the ligand to the position and orientation in the PL complex.

In this approach, the key to achieving the converged results is through introducing suitable auxiliary restraints in order to prevent the ligand from leaving the binding site (Figure 2.6). The

advantages of restraining the ligand to it positions and orientations are i) ΔG_{II}^{*} is free from sampling problem as ligand stays in place ii) ΔG_r, restraint energy to regulate the position and orientation (a set of restraints) of the ligand relative to the protein can be computed.

Setting up the restraints in MD simulations required the cross-link, which involves no other interaction between ligand and protein thats restraining the position and orientation of the ligand relative to the protein receptor. This is held by a set of restraints, commonly six harmonic restraints (one distance, two angles and three dihedrals) from six anchoring atoms (Figure 2.6).

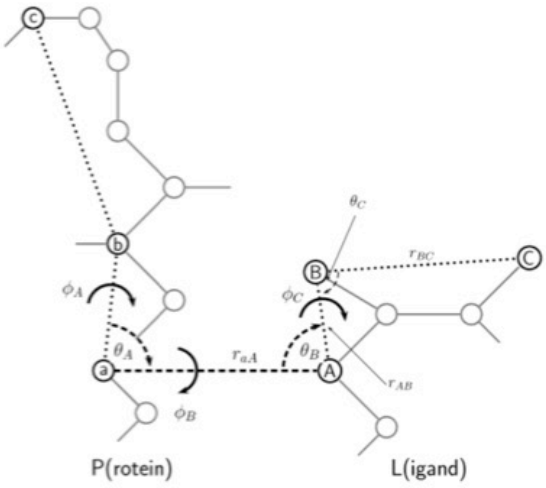


Figure 2.6: The illustration of the protein and ligand with the selected anchoring atom shows by circles labelled with a, b and c (protein anchor's atoms) while A, B and C (ligand anchor's atoms). The cross-link represents the six harmonic restraints consisting of one distance, two angles and three dihedrals denoted by r_{aA} (distance), ϑ_A *and* ϑ_B (angle) and $\emptyset_A$, $\emptyset_B$ *and* $\emptyset_C$ (dihedral). (Taken from Boresch *et al*.)[73]

Owing to the restraints added to the systems, the free energies cost, ΔG_r denoted by Equation 2.21, should be included in the binding free energy calculations. The detailed explanations of where are this contributions come from and how this derived explained in previously published paper by Boresch *et al*.[73] ΔG_r evaluated by having the standard states dependence correction instead of include the volume of the systems, volume corresponding to one molar standard states were included (Equation 2.22). Finally the ΔG_{bind}^{0} at the standard states were calculated employing the Equation 2.23.

$$\Delta G_r = -kT \ln \left[\frac{8\pi^2 V \left(K_r \; K_{\theta_A} K_{\theta_B} K_{\emptyset_A} K_{\emptyset_B} K_{\emptyset_C}\right)^{1/2}}{r_{a,A,0}^2 \sin\theta_{A,0} \sin\theta_{B,0} (2\pi kT)^3} \right] \tag{2.21}$$

$$\Delta G_r = -kT \ln \left[\frac{8\pi^2 V^0 (K_r \; K_{\theta_A} K_{\theta_B} K_{\phi_A} K_{\phi_B} K_{\phi_C})^{1/2}}{r_{a,A,0}^2 \; \sin\theta_{A,0} \; \sin\theta_{B,0} \; (2\pi kT)^3} \right] \tag{2.22}$$

$$\Delta G_{bind}^0 = \Delta G_I^0 - \Delta G_{II}^0 - \Delta G_r^0 \tag{2.23}$$

Chapter 3: Evaluation of solvation free energies for small molecules with the AMOEBA polarisable force field

3.1 Introduction

Electronic polarisation is one of the components that plays an important role in many biomolecular systems due to the changes in environment that might occur in a simulation, e.g. the difference between a protein interface and bulk solvent, or between a membrane surface and buried inside a bilayer. Classic fixed-point-charge models may not accurately represent changes in electrostatic interactions, as the charges are not able to fully adapt to the environment as discussed in Chapter 2 section 2.1.1 and 2.1.2. However, in theory, polarisable force fields should capture this effect. By incorporating an explicit response to the environment, such force fields may be expected to give more accurate predictions of the interactions in the systems.

Consequently, an evaluation of potential energy function accuracy is required to determine their performance, given the additional computational cost of incorporating the explicit polarisable potentials. Solvation free energy calculations are a common goal in computational solvation thermodynamics studies to assess force field properties.[156,208–213] They are also often recognised as a test of the ability of any force field to integrate well in many chemical environments. This is thanks to the availability of high accuracy experimental data, and the straightforward computational methodologies for free energy prediction. As such, evaluating the accuracy of solvation free energy prediction is often a crucial step for force field validation.

Solvation free energy calculations are sensitive to the accuracy of force field parameters, as interaction energies are reliant on the potential energy function portraying the interatomic interactions of the systems. In this chapter, the AMOEBA force field has been implemented in the solvation free energy calculations of small molecules in a variety of common organic solvents with different dielectric constants. The solvation free energies generated with the AMOEBA force field will be compared with experimental measurements and those of the fixed-point-charge force field GAFF, to evaluate the performance of the AMOEBA polarisable force field. The direct aim of this study is to understand the effects of AMOEBA in environments where polarisation might be

important. Ultimately, these will either identify applications to follow up with or target systems to compare other advanced methods. Further details of the tests carried out will be presented in this chapter.

3.2 Non-aqueous solvents

The local environment of a system plays in important role on electronic polarisation interaction in a system. To investigate this effect, non-aqueous solvents were used to represent different environments of the system, instead of using water as the solvent for assessing the accuracy of the force field in solvation free energy approaches, as in most studies.[210,213–218] Common organic solvents with a range of different dielectric constants, namely toluene (ε = 2.38), chloroform (ε = 4.81), DMSO (ε = 36.64) and acetonitrile (ε = 47.0) were chosen based on the availability of experimental data of solvation free energies for a variety of different molecules. Here, all the AMOEBA solvents were modelled utilising the parameters taken from amoeba09.prm[219] except for chloroform.[155] For fixed-charge simulations, the solvent's parameters were taken from Cieplak *et al.*,[126] (chloroform) Grabuleda *et al.*,[220] (acetonitrile) and Dupradeau *et al.*,[221] (DMSO and toluene).

3.3 Dataset

21 small molecules (Figure 3.1) were chosen in total, of which six had experimental solvation free energies for all four solvents, and 15 more had experimental solvation free energies for only toluene and chloroform. This choice of small molecules covered a variety of functional groups, and was obtained from Minnesota solvation database.[222] and Abraham *et al.*[223] Minnesota database consists of a large collection of solvation free energies (3000 data points) of unique solutes in different solvents. However our data set for this study was limited to molecules for which experimental solvation free energies were available in multiple organic solvents and availability of amoeba09 and chloroalkane AMOEBA force fields.[155,219] Solvent models for both force fields have previously undergone limited validation of liquid density and enthalpy of vaporisation to assess their suitability.[155,219]

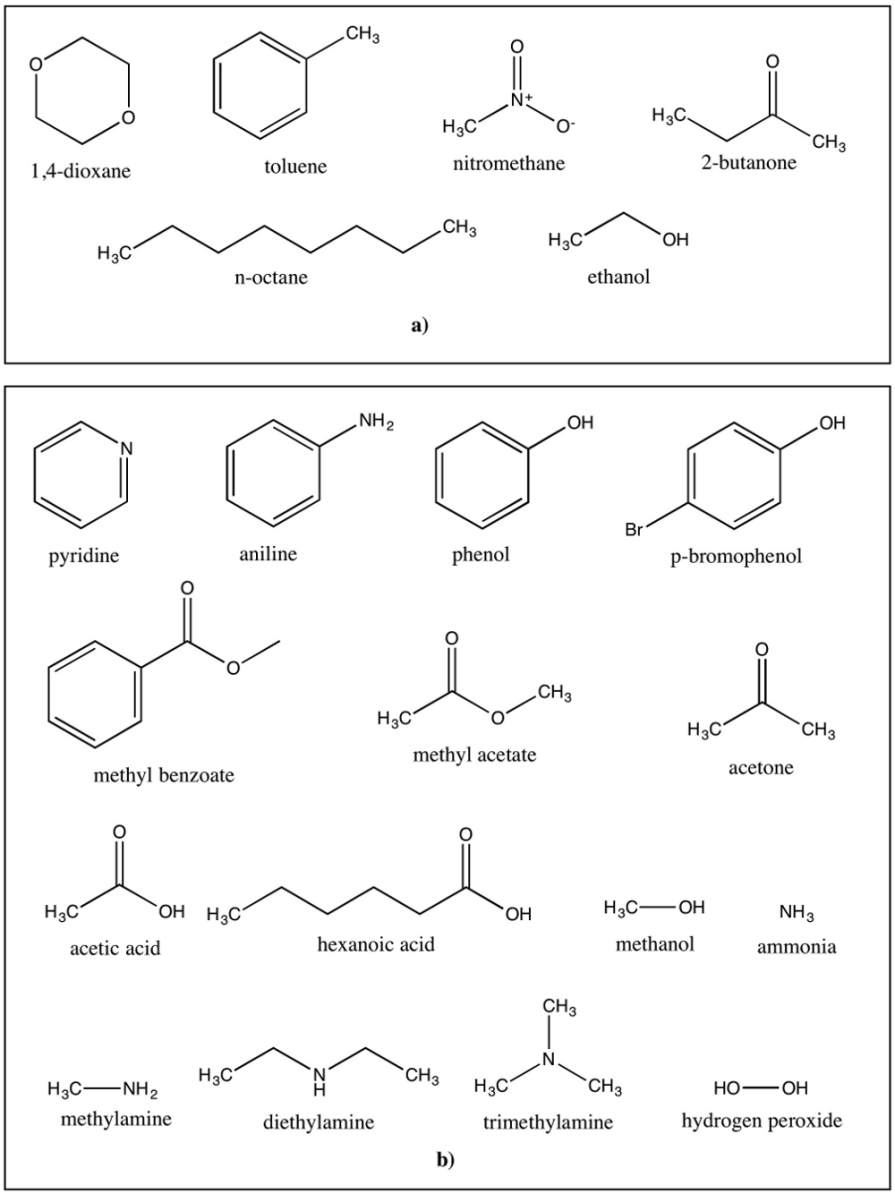


Figure 3.1: The selected small molecule data set employed in this study, taken from the Minnesota solvation database. a) Data set of small molecules for toluene, chloroform, acetonitrile and DMSO solvent. b) Data set of additional small molecules for toluene and chloroform solvent. (Figure taken from Mohamed *et al.*)[224]

3.4 Parameterisation

The standard AMOEBA parameterisation protocol for small molecules using automated parameter derivation by the POLTYPE software[225] with the underlying TINKER software[226] package and GAUSSIAN09 program[227] for QM calculation is depicted in Figure 3.2. The AMOEBA automated parameterisation procedure requires only the initial coordinates of a molecule in order to assign the entire AMOEBA parameter set for that molecule. The main steps involved are as follows.

The initial structure of each molecule was optimised quantum mechanically using GAUSSIAN09 at the HF/6-31G* level of theory. A single-point energy calculation was carried out subsequently at the MP2/6-311G(1d, 1p) level of theory followed by a Distributed Multipole Analysis facilitated by the Gaussian Distributed Multipole Analysis (GDMA) program[228] to compute an initial set of atomic multipoles. This was continued by further single point calculation of the molecular electrostatic potential using a larger basis set (MP2/6-311G++(2d, 2p)). Finally, the AMOEBA multipole parameters were optimised by fitting to the QM electrostatic potential. The parameters for van der Waals, bonds, angles, stretch-bends, torsions, and atomic polarisability values of the small molecule were assigned from a lookup database provided in the POLTYPE software. This internal set of lookup data was taken from chemical space covered by available parameters from the AMOEBA09 or MM3 potentials. In the absence of suitable specific parameters for individual chemical species, generic parameters were assigned by POLTYPE.

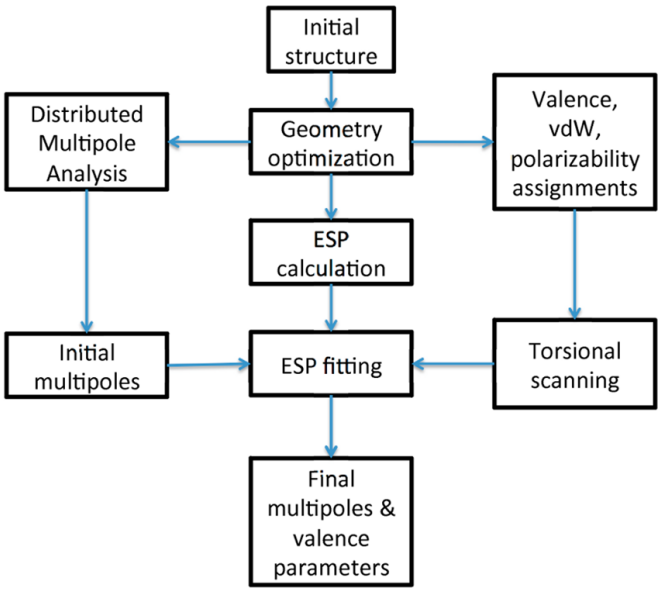


Figure 3.2: A summary of the standard AMOEBA parameterisations protocol employed TINKER software and GAUSSIAN program. (Figure taken from Bradshaw *et al.*)[229]

In an effort to improve the consistency and accuracy of small molecule parameters for AMOEBA, a manual parameterisation from scratch[153] was performed instead of automatic parameterisation using the POLTYPE software.[225] In manual parameterisation, the standard AMOEBA parameterisation protocol was adopted to generate all the parameters by hand using the TINKER 6.33 package[226] and GAUSSIAN09 program.[227] The issue that arose due to lack of human intervention was solved as we were able to manually make a choice of parameters to be assigned for the molecules of interest. In our case, all the valence parameters (bond, angle, stretch-bend, out-of-plane and torsion), van der Waals parameters and atomic polarisabilities for the small molecules and solvents were defined consistently based on the amoeba09.prm potential for small molecules available in TINKER6 except atomic multipoles. The atomic multipole parameters for these molecules were derived from the QM calculations obtained from the GAUSSIAN09 program[227] using the similar steps as a standard parameterisation protocol. Atomic charges, dipole and quadrupole value were obtained with Stone's distributed multipole analysis available through GDMA program.[228] However all the parameter assignment input was done manually to generate the final optimised multipole parameters.

For comparison, similar systems were set up for the GAFF fixed-point-charge force field. The standard GAFF fixed-point-charge parameterisation procedure was carried out for the small molecules. The ANTECHAMBER program[230] was used to generate the fixed-point-charge parameters for the MD simulations, using AM1-BCC charges for electrostatic interactions.[231,232] The parameters generated for all solute are available freely as online dataset.[233]

3.5 Free energy calculations

The protocol for solvation free energy calculations from Shi *et al.*, 2011[234] was adapted to calculate the solvation free energy of small molecules in four different solvents. The solvation free energy of each molecule was calculated based on the thermodynamic cycle as in Figure 3.3. The overall solvation free energy is given by:

$$\Delta G_{solv} = -\Delta G_{decoupling,sol} - \Delta G_{discharging,sol} + \Delta G_{discharging,vac} \qquad (3.1)$$

For evaluation of the solvation free energies for both AMOEBA and GAFF force fields in non-aqueous solvents, a similar system setup has been employed in molecular dynamics simulations. The free energy of each process was computed using Bennett's Acceptance Ratio (BAR) for AMOEBA and GAFF.

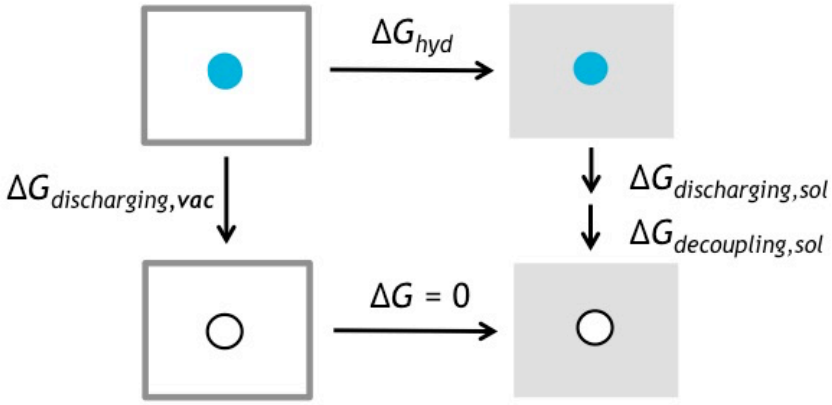


Figure 3.3: Thermodynamic cycle used to calculate the solvation free energies of small molecules in non-aqueous solvents. The simulations require three sets of calculations: i) solution phase: discharging of ligand in solution, ii) solution phase: decoupling of vdW interactions between the ligand and environment, iii) gas phase: discharging the ligand in vacuum. The cyan circles represent a fully charged ligand interacting with the environment, the unfilled circles represents a discharged ligand and completely decoupled with the environment, while the grey square denotes a simulation run in solution phase and the unfilled square box denotes a simulation run in gas phase.

3.6 Non-aqueous solvent box preparation

All solvents were first prepared in a cubic box with length of ~40 Å dimension on each side, containing ~400 to 800 molecules using TINKER utilities.[226] The number of solvent molecules varied depending on the size of a solvent molecule and the experimental density required (Table 3.1). The solvent box was minimised using the steepest descent algorithm for 2500 steps and then heated to 300 K at constant volume using NVT MD over a 50 ps time period, followed by 200 ps equilibration to 1 atm at constant pressure in the NPT ensemble. A Berendsen barostat was applied to control the pressure with the coupling time set to 2 ps.[235] This simulation was run with a 1 fs time step using the velocity Verlet integrator in TINKER. A Nosé-Hoover thermostat[236,237] was employed to restrain the temperature to 300 K with a coupling time parameter, from which the Nosé-Hoover chain masses are set in TINKER, of 0.2 ps. Final temperature and density equilibrated structures were used as solvent box inputs for the following series of solvation free energy calculations

Table 3.1: The non-aqueous solvents details for periodic systems setup in cubic cell with approximately the same size of the box with the consistent density match with the experiment after the equilibration.

	Solvent			
Details	Chloroform	Toluene	DMSO	Acetonitrile
Dielectric constant	4.81	2.38	47.0	36.64
No. of molecule	602	364	648	768
Box dimension (Å)	43.77	40.00	42.35	40.96
Comp. density (g/cc)	1.48	0.87	1.10	0.76
Exp. density (g/cc)	1.49	0.87	1.10	0.76

3.7 Production simulation details

AMOEBA MD simulations for solvation free energy calculations utilised either the AMBER 14[238] or TINKER 6.3.3 packages[226] depending on the solute/solvent system under investigation. All systems were initially prepared in TINKER[226] by soaking each molecule in a periodic box of pre-equilibrated solvent, generated as above, using the XYZEDIT utility of TINKER. Initial structures and parameters were then converted to AMBER format for subsequent minimisation, equilibration and simulation, using the tinker_to_amber utility of AMBER 14. However, solutes or solvents that included a 'Z-Bisector' multipole local frame (DMSO, Acetonitrile, Methylamine, Trimethylamine) could not be converted as the 'Z-Bisector' frame is not implemented in AMBER 14. Instead, these simulations were performed with an equivalent procedure in TINKER 6.3.3. Details of both protocols are provided below. All simulations were performed in triplicate, using the same starting structure but a different random number seed for the thermostat.

Solution phase simulations in AMBER used the pmemd.amoeba program and were performed as follows. Initially, the systems underwent minimisation for 2500 steps, of which the first 1000 steps were run with a steepest descent algorithm, and the next 1500 steps with a conjugate gradient algorithm. For each system, simulations were then performed in the NVT ensemble, heated slowly to 300 K over 50 ps, followed by another 100 ps of pressure equilibration using NPT at 300 K and 1 atm. A timestep of 1 fs and a velocity Verlet integrator was used to propagate dynamics. To maintain the temperature and pressure, the systems were treated using a Langevin thermostat and Berendsen barostat respectively.[235,239] A different random seed for the Langevin

thermostat was applied for each independent repeat. van der Waals interactions were evaluated explicitly up to a 9 Å cutoff with an analytical long-range correction. Long-range electrostatic interactions for all the systems were treated using a Particle Mesh Ewald (PME) summation,[240] with a real-space cutoff of 8 Å. The PME calculation used fifth order B-spline interpolation. At each step the atomic induced dipoles were converged until the root-mean square change was below 0.01 D/atom. Finally, the last configuration of the NPT simulation was used as the starting point for equilibration in all the intermediate λ states with AMOEBA.

A total of 11 intermediate state simulations with λ = 1.0, 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1 and 0.0 were applied to electrostatic interactions for discharging the solute in vacuum and in solvent.[241] λ = 1 refers to a fully interacting solute and λ = 0 to a noninteracting solute. However, for calculating the free energies of decoupling solute vdW interactions in the solvent, a different spacing of intermediate states was used with λ = 1.0, 0.9, 0.8, 0.75, 0.7, 0.65, 0.6, 0.5, 0.4, 0.2 and 0.0.[241] Furthermore, to allow the potential to disappear smoothly as the intermediate simulations progressed to zero, a soft-core Halgren buffered 14-7 van der Waals term[158] as previously described by Shi *et al.*[241] was applied. For each value of λ, 2 ns of constant pressure molecular dynamics were performed, using an identical protocol to the NPT pressure equilibration step. Atomic coordinates of the system were saved every 1 ps and the first 200 ps of each window were discarded as equilibration.

Solution phase simulations in TINKER[226] were performed identically to those in AMBER except for the following minor changes. Minimisation in TINKER was performed using the default minimisation algorithm, limited memory Broyden-Fletcher-Goldfarb-Shanno (BFGS) Quasi-Newton optimization[242] for 2500 steps. Additionally, the Nosé-Hoover thermostat[236,237] was employed during MD simulations instead of the Langevin thermostat[239] used with AMBER 14. All the other protocol options, including the λ windows applied, were identical.

All gas phase simulations were performed in TINKER.[226] In this simulation, a single solute molecule was simulated for 200 ps using a stochastic integrator with a time step of 0.1 fs and a temperature of 300 K. The induced dipoles were converged to 1×10^{-6} D/atom. Coordinates were saved every 0.1 ps. For free energy analysis, the first 20 ps were discarded. In each case, BAR was used to evaluate the free energy changes between the neighboring states (λ_i and λ_{i+1}).

For the GAFF simulations an identical protocol was implemented except that an 8 Å direct vdW cutoff was used rather than 9 Å. Importantly, the PMEMD and SANDER modules included in AMBER 14 were used for the GAFF simulations with identical λ windows employed throughout for both force fields. For free energy calculations, BAR was used as implemented in the PYMBAR

PYTHON package[191] for GAFF fixed-point-charge results, while an in-house script, BAR-amber[243] was used to analyze the results for the AMOEBA simulations.

3.8 Statistical error analysis

The error analysis and significance testing suggested by Mobley *et al.*[244] was employed to evaluate the calculated solvation free energies in four solvents simulated with both the AMOEBA and GAFF force fields. The agreement of estimated solvation free energies with experiment was evaluated using mean unsigned error (MUE), mean signed error (MSE), Pearson correlation coefficient (R), coefficient of determination (R^2) and Kendall's tau coefficient (τ) across three replicates. In addition, 1000 iterations of bootstrapping with replacement were performed to estimate the 95% confidence intervals on these values. Finally, a Student's paired t-test was applied to determine the significance of differences between MSE errors generated with AMOEBA and GAFF, assuming both are normally distributed. A Wilcoxon signed-rank test was used to similarly compare MUE since they are severely non-normally distributed. These tests will indicate whether the errors of our predictions are significantly different between different force fields.

3.9 Result and discussion

For evaluation of the solvation free energies for each solute in each solvent, the free energy of the small molecule was calculated as in equation 3.1. In order to ensure that we sample sufficiently with a long enough run time, the convergence of free energies was plotted. Here, each run was split into 200 ps periods and the total ΔG_{solv} evaluated every 200 ps over the course of trajectories. As a representative example, the convergence of ΔG_{solv} of 1,4-dioxane in chlorofom is shown in Figure 3.4. The solvation free energy does not change across the trajectory, supporting the fact that 2 ns length is a long enough run time for this simple molecule to converge. In addition, three independent replicates were run to observe the consistency of solvation free energies computed between each of the runs. Errors were calculated as the standard error in the mean between these three replicate results. Given the small standard error (SE), ~0.1 kcal mol^{-1} for AMOEBA and GAFF simulations in all data sets (Table 3.2 to 3.5) provides no evidence to indicate inadequate conformational sampling and hence we assessed the simulations to be of appropriate length.

A detailed analysis of results for each of the solvents will first be treated in turn in the following sections, followed by comparison of the results between them and analysis of the statistical error and performance for both AMOEBA and GAFF force fields.

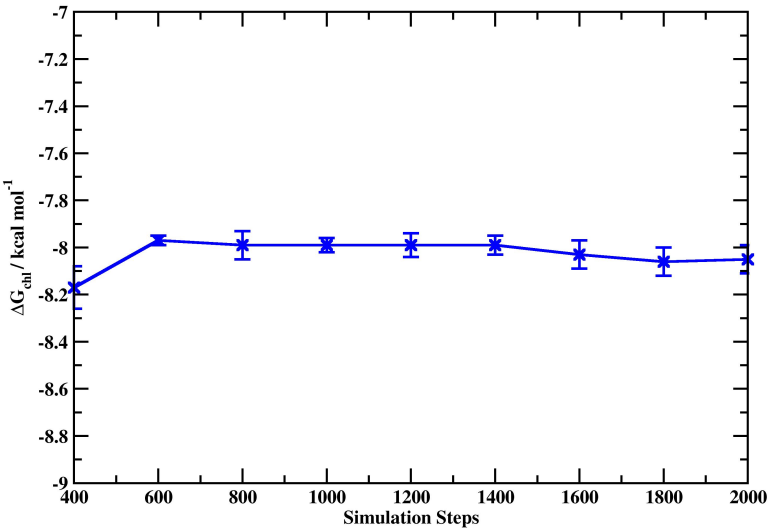


Figure 3.4: The energies convergence over the course of trajectories from 400ps to 2000ps of 1,4-dioxane for three independent repeats.

3.9.1 Toluene Solvent

Table 3.2 gives free energies of solvation in toluene (ε = 2.38) calculated with the AMOEBA and GAFF force fields for 21 small molecules compared against experimental data. The largest unsigned errors with respect to experiment of the calculated solvation free energies in these datasets comes from the ammonia calculation. Both AMOEBA and GAFF show poor results, but AMOEBA gives slightly worse estimates, with errors of 2.80 kcal mol^{-1} compared to 1.65 kcal mol^{-1}, respectively. Methanol however, demonstrated an excellent agreement with the experimental data with the smallest error of 0.08 kcal mol^{-1} and 0.02 kcal mol^{-1} with AMOEBA and GAFF, respectively. Ethanol also gives a good value but not quite as good as methanol with error 0.73 kcal mol^{-l} for AMOEBA and 0.39 kcal mol^{-1} for GAFF.

Table 3.2: AMOEBA calculated solvation free energies for small molecules in toluene (ε = 2.38) against experimental data and fixed-point charge (GAFF) data.

		ΔG_{tol} (kcal mol^{-1})		Unsigned Error to Experiment	
Molecule	Exp	AMOEBA	GAFF	AMOEBA	GAFF
1,4-dioxane	-4.91[a]	-5.43 ± 0.02	-5.82 ± 0.00	0.52	0.91
2-butanone	-4.27	-3.74 ± 0.01	-4.22 ± 0.02	0.53	0.05
Acetic acid	-4.00	-3.24 ± 0.04	-4.39 ± 0.34	0.76	0.39
Acetone	-3.59	-4.09 ± 0.01	-3.63 ± 0.03	0.50	0.04
Ammonia	-2.38	0.42 ± 0.03	-0.73 ± 0.03	2.80	1.65
Aniline	-6.69	-5.11 ± 0.04	-5.97 ± 0.03	1.58	0.72
Ethanol	-3.33	-2.60 ± 0.01	-2.94 ± 0.03	0.73	0.39
Methanol	-2.18	-2.10 ± 0.04	-2.16 ± 0.06	0.08	0.02
Methylamine	-2.65	-2.88 ± 0.03	-1.70 ± 0.04	0.23	0.95
n-octane	-5.38	-3.82 ± 0.04	-5.27 ± 0.01	1.56	0.11
Nitromethane	-4.31	-4.00 ± 0.03	-4.34 ± 0.01	0.31	0.03
Phenol	-6.93	-5.33 ± 0.08	-6.17 ± 0.06	1.60	0.76
Pyridine	-5.13	-4.58 ± 0.04	-4.81 ± 0.02	0.55	0.32
Toluene	-5.12	-4.04 ± 0.03	-4.45 ± 0.03	1.08	0.67
Diethylamine	-3.75	-2.58 ± 0.40	-4.10 ± 0.01	1.17	0.35
P-bromophenol	-8.70	-6.15 ± 0.02	-8.31 ± 0.05	2.55	0.39
Trimethylamine	-2.71	-3.33 ± 0.08	-3.34 ± 0.04	0.62	0.63
Hexanoic acid	-6.97	-5.83 ± 0.04	-7.69 ± 0.06	1.14	0.72
Methylacetate	-3.81	-3.72 ± 0.06	-4.56 ± 0.04	0.09	0.75
Methylbenzoate	-7.96	-7.18 ± 0.03	-8.13 ± 0.03	0.78	0.17
Hydrogen peroxide	-3.14	-3.34± 0.05	-3.18 ± 0.03	0.20	0.04

All the experimental solvation free energies are taken from Minnesota solvation database[245] except [a]Experimental solvation free taken from Abraham *et al.*[223] Errors report 1 SE over 3 repeats.

3.9.2 Chloroform Solvent

In the case of chloroform solvent, the same set of 21 small molecules was tested. Results are reported in Table 3.3. Clearly for Table 3.3, the solvation free energies estimated for ammonia and n-octane in chloroform are still far from the experimental data and have the largest errors to experiment, 3.17 kcal mol^{-1} for GAFF and 3.26 kcal mol^{-1} respectively for AMOEBA. In contrast, ethanol shows the best values with the smallest error to experiment, 0.06 kcal mol^{-1} for AMOEBA calculated solvation free energies in chloroform. Although AMOEBA values for ethanol fit well against experimental data, the opposite is true for the GAFF force field. GAFF resulted in a larger error of 1.20 kcal mol^{-1} for ethanol in chloroform. AMOEBA seems to give a better value as well for methanol compared to GAFF with 0.54 and 1.54 kcal mol^{-1}. However, phenol again demonstrated poor agreement with experimental data with the larger error calculated for AMOEBA, 2.62 kcal mol^{-l} compared to GAFF, 1.17 kcal mol^{-l}.

Table 3.3: AMOEBA calculated solvation free energies for small molecules in chloroform (ε = 4.81) against experimental data and fixed-point charge (GAFF) data.

		ΔG_{chl} (kcal mol^{-1})		Unsigned Error to Experiment	
Molecule	Exp.	AMOEBA	GAFF	AMOEBA	GAFF
1,4-dioxane	-6.21[a]	-8.06 ± 0.06	-6.47 ± 0.04	1.85	0.38
2-butanone	-5.43	-5.30 ± 0.05	-4.83 ± 0.07	0.13	0.60
Acetic acid	-4.74	-3.50 ± 0.01	-3.95 ± 0.11	1.24	0.79
Acetone	-4.42	-5.93 ± 0.02	-4.14 ± 0.01	1.51	0.28
Ammonia	-2.41	0.76 ± 0.01	-0.49 ± 0.01	3.17	1.92
Aniline	-7.34	-4.37 ± 0.01	-6.21 ± 0.03	2.97	1.13
Ethanol	-3.94	-4.00 ± 0.05	-2.74 ± 0.01	0.06	1.20
Methanol	-3.32	-3.86 ± 0.01	-1.78 ± 0.05	0.54	1.54
Methylamine	3.17	-5.14 ± 0.01	-1.82 ± 0.01	1.97	1.35
n-octane	-5.25	-1.99 ± 0.05	-6.52 ± 0.41	3.26	1.27
Nitromethane	-4.68	-3.92 + 0.03	-4.55 ± 0.11	0.76	0.13
Phenol	-7.14	-4.52 ± 0.02	-5.97 ± 0.02	2.62	1.17
Pyridine	-6.45	-5.46 ± 0.03	-5.31 ± 0.10	0.99	1.14
Toluene	-5.48	-3.08 ± 0.04	-5.03 ± 0.03	2.40	0.45
Diethylamine	-5.23	-2.65 ± 0.05	-4.79 ± 0.05	2.58	0.44
P-bromophenol	-8.59	-6.03 ± 0.04	-7.91 ± 0.10	2.56	0.68
Trimethylamine	-3.90	-6.17 ± 0.01	-4.18 ± 0.04	2.27	0.28
Hexanoic acid	-7.51	-5.23 ± 0.18	-7.97 ± 0.21	2.28	0.46
Methylacetate	-4.90	-4.54 ± 0.03	-5.18 ± 0.04	0.36	0.28
Methylbenzoate	-7.81	-7.57 ± 0.04	-9.00 ± 0.04	0.24	1.19
Hydrogen peroxide	-4.70	-3.20 ± 0.05	-2.00 ± 0.04	1.50	2.70

All the experimental solvation free energies are taken from Minnesota solvation database[245] except [a]Experimental solvation free energies taken from Abraham *et al.*[223] Errors report 1 SE over 3 repeats.

3.9.3 Acetonitrile Solvent

Owing to the limited data set, the solvation free energies in acetonitrile solvent were computed for only six molecules. The computed results directly compared with the experimental solvation free energies in acetonitrile are presented in Table 3.4. Although the data set evaluated was small, there was excellent agreement between simulation and experiment for the entire data set denoted by the small error computed (less than 1 kcal mol^{-1}). The same patterns were observed for other molecules for either AMOEBA or GAFF. Here, solvation free energies calculated for 2-butanone remain the best estimated with the lowest error to experiment of 0.09 kcal mol^{-1} by GAFF but poor for AMOEBA with 0.95 kcal mol^{-1}. Ethanol shows the same pattern, which is slightly better with GAFF, 0.36 kcal mol^{-1} rather than AMOEBA, 0.77 kcal mol^{-1}. However, 1, 4-dioxane molecule with AMOEBA consistently gave better results compared to GAFF as shown in Table 3.4. In addition, n-octane was still one of the solutes with higher error with 0.80 kcal mol^{-1} with AMOEBA and 0.29 kcal mol^{-1} by GAFF.

Table 3.4: AMOEBA calculated solvation free energies for small molecules in acetonitrile (ε = 36.64) against experimental data and fixed-point charge (GAFF) data.

	ΔG_{ace} (kcal mol^{-1})			Unsigned Error to Experiment	
Molecule	Exp	AMOEBA	GAFF	AMOEBA	GAFF
1,4-dioxane	-5.33[a]	-5.55 ± 0.01	-6.16 ± 0.04	0.22	0.83
2-butanone	-4.73	-3.78 ± 0.02	-4.82 ± 0.04	0.95	0.09
Ethanol	-4.43	-3.66 ± 0.02	-4.08 ± 0.01	0.77	0.36
n-octane	-3.57	-2.77 ± 0.02	-3.86 ± 0.05	0.80	0.29
Nitromethane	-5.62	-4.64 ± 0.02	-4.79 ± 0.02	0.98	0.83
Toluene	-4.68	-4.04 ± 0.04	-4.47 ± 0.03	0.64	0.21

All the experimental solvation free energies are taken from Minnesota solvation database[245] except [a]Experimental solvation free energies taken from Abraham *et* al.[223] Errors report 1 SE over 3 repeats.

3.9.4 DMSO Solvent

For perturbations performed in DMSO solvents, the identical data set employed for acetonitrile was tested for AMOEBA and GAFF. The solvation free energies in DMSO for these molecules are shown in Table 3.5. Again, the values estimated seem to produce consistent unsigned errors for all molecules, resulting in good agreement with the experimental data. Here, the largest error generated from 1,4-dioxane by 1.34 kcal mol^{-1} with GAFF giving the greater shift to the fitted line denoted by experiment value as shown in Figure 3.5. contrarily with AMOEBA, better result are obtained with only 0.37 kcal mol^{-1} error to experiment. While for 2-butanone, GAFF gave an excellent result with an error only 0.03 kcal mol^{-1}, compared to 1.36 kcal mol^{-1}, error against experiment with AMOEBA. In this analysis, n-octane remains as one of the larger contributors to the overall error to experiment, calculated as 1.80 kcal mol^{-1} for the AMOEBA force field and 0.75 kcal mol^{-1} with the GAFF force field. Meanwhile for ethanol, the value generated by AMOEBA and GAFF was quite good compared to rest of the molecules in the data set.

Table 3.5: AMOEBA calculated solvation free energies for small molecules in DMSO against experimental data and fixed-point charge, GAFF data.

	ΔG_{dmso} (kcal mol^{-1})			Unsigned Error to Experiment	
Molecule	Exp	AMOEBA	GAFF	AMOEBA	GAFF
1,4-dioxane	-4.90[a]	-5.27 ± 0.03	-6.24 ± 0.04	0.37	1.34
2-butanone	-4.23	-2.87 ± 0.07	-4.20 ± 0.03	1.36	0.03
Ethanol	-5.25	-4.48 ± 0.01	-5.12 ± 0.02	0.77	0.13
n-octane	-2.84	-1.04 ± 0.04	-2.09 ± 0.05	1.80	0.75
Nitromethane	-5.66	-4.56 ± 0.02	-4.81 ± 0.03	1.10	0.85
Toluene	-4.42	-3.09 ± 0.11	-3.86 ± 0.11	1.33	0.56

3.9.5 Solvent comparison

Figure 3.5 compares AMOEBA and GAFF solvation free energy results across all four solvents directly with those of experiment, while Table 3.6 provides summary metrics of the same results. The mean unsigned error to experiment for calculated solvation free energies across all solvents is approximately 1.22 kcal mol^{-1} for AMOEBA and 0.66 kcal mol^{-1} for GAFF (Tables 3.2 to 3.5). The

largest MUE is in chloroform solvent for both force fields as shown in Table 3.6. In terms of MSE both force fields underestimate solvation free energies (i.e. show positive MSE) particularly for ammonia simulated with AMOEBA in toluene and chloroform (Table 3.2 and Table 3.3). Predominantly, the AMOEBA MSE in all solvents is slightly larger than that of GAFF, as shown in Table 3.6.

Interestingly, the results of solvation free energies with GAFF often give better correlation to experimental data based on comparison of the four solvents in Figure 3.5 and Table 3.6. The best agreement was given in toluene with R^2 0.90 (Figure 3.5a) while the worst R^2 of 0.53 was observed in acetonitrile (Figure 3.5d). Similarly to the MUE metrics above, chloroform solvation free energies for small molecules using AMOEBA showed the worst correlation to experimental values with R^2 0.26 (Figure 3.5c). The best R^2 for AMOEBA was in DMSO (R^2 = 0.84), and may be due to a consistent underestimation of solvation free energy calculated across the whole data set, as suggested by the linear regression line observed in Figure 3.5d.

To allow performance comparison of AMOEBA and GAFF in different environments, the results of each solvent were also compared using their Kendall τ coefficients, which examined agreement in ranking of solvation free energies between theory and experiment. Kendall τ allowed the determination of a clear order of performance for all solvents in the two different force fields. With AMOEBA, toluene, DMSO and acetonitrile perform well (overall τ values of 0.74, 0.73 and 0.73 respectively), while chloroform again performs worst with only 0.23. For GAFF, the τ value for toluene indicates the best agreement in predicted rankings (τ = 0.87) followed by chloroform (0.77), acetonitrile (0.73) and DMSO (0.47). It should be noted that the small dataset sizes for acetonitrile and toluene (n = 6) may lead to large fluctuations in τ and R with small changes in results, as demonstrated by the broader confidence intervals for these measures.

Table 3.6: Summary of performance metrics for calculated solvation free energies with the AMOEBA polarizable force field and the GAFF fixed-point-charge force field in all four solvents. Upper and lower bounds are estimated as 95% confidence intervals in the mean using bootstrapping for 1000 iterations with replacement.

	AMOEBA Force Field			
	Solvent			
Metrics	**Toluene**	**Chloroform**	**Acetonitrile**	**DMSO**
MUE (kcal mol^{-1})	0.67 ≤ **0.92** ≤1.30	1.23 ≤ **1.68** ≤ 2.09	0.48 ≤ **0.73** ≤ 0.88	0.74 ≤ **1.12** ≤ 1.46
MSE (kcal mol^{-1})	0.37 ≤ **0.73** ≤ 1.14	0.12 ≤ **0.90** ≤ 1.57	0.10 ≤ **0.65** ≤ 0.88	0.20 ≤ **0.99** ≤ 1.4
R	0.74 ≤ **0.86** ≤ 0.92	0.18 ≤ **0.51** ≤ 0.79	-1.00 ≤ **0.89** ≤ 0.99	-0.63 ≤ **0.91** ≤ 1.00
***R*^{2}**	0.53 ≤ **0.74** ≤ 0.85	0.03 ≤ **0.26** ≤ 0.62	0.15 ≤ **0.79** ≤ 0.97	0.17 ≤ **0.84** ≤ 1.00
Kendall τ	0.53 ≤ **0.74** ≤ 0.88	-0.12 ≤ **0.23** ≤ 0.51	0.33 ≤ **0.73** ≤ 1.00	-0.09 ≤ **0.73** ≤ 1.00
	GAFF Force Field			
MUE (kcal mol^{-1})	0.32 ≤ **0.48** ≤ 0.68	0.68 ≤ **0.92** ≤ 1.23	0.21 ≤ **0.43** ≤ 0.67	0.27 ≤ **0.61** ≤ 0.98
MSE (kcal mol^{-1})	-0.14 ≤ **0.10** ≤ 0.40	0.18 ≤ **0.56** ≤ 1.01	-0.44 ≤ **0.03** ≤ 0.41	-0.68 ≤ **0.16** ≤ 0.58
R	0.89 ≤ **0.95** ≤ 0.98	0.78 ≤ **0.91** ≤ 0.96	-1.00 ≤ **0.73** ≤ 0.93	-0.05 ≤ **0.82** ≤ 0.99
***R*^{2}**	0.80 ≤ **0.90** ≤ 0.95	0.60 ≤ **0.83** ≤ 0.92	0.00 ≤ **0.53** ≤ 0.85	0.00 ≤ **0.68** ≤ 0.97
Kendall τ	0.72 ≤ **0.87** ≤ 0.96	0.59 ≤ **0.77** ≤ 0.88	-0.09 ≤ **0.73** ≤ 1.00	-0.23 ≤ **0.47** ≤ 1.00

Table 3.7: Calculated p-values of statistical tests between mean signed (Student's paired t-test) and unsigned error (Wilcoxon signed-ranked test) distributions for AMOEBA and GAFF. Significant differences ($p < 0.05$) denoted in bold. GAFF and AMOEBA perform identically in terms of MUE for acetonitrile and DMSO, and in terms of MSE in chloroform. For all other metrics GAFF performed better.

	Solvent			
p-value	**Toluene**	**Chloroform**	**Acetonitrile**	**DMSO**
Unsigned Error	**0.0071**	**0.0087**	0.2489	0.1730
Signed Error	**0.0015**	0.4363	**0.0098**	**0.0028**

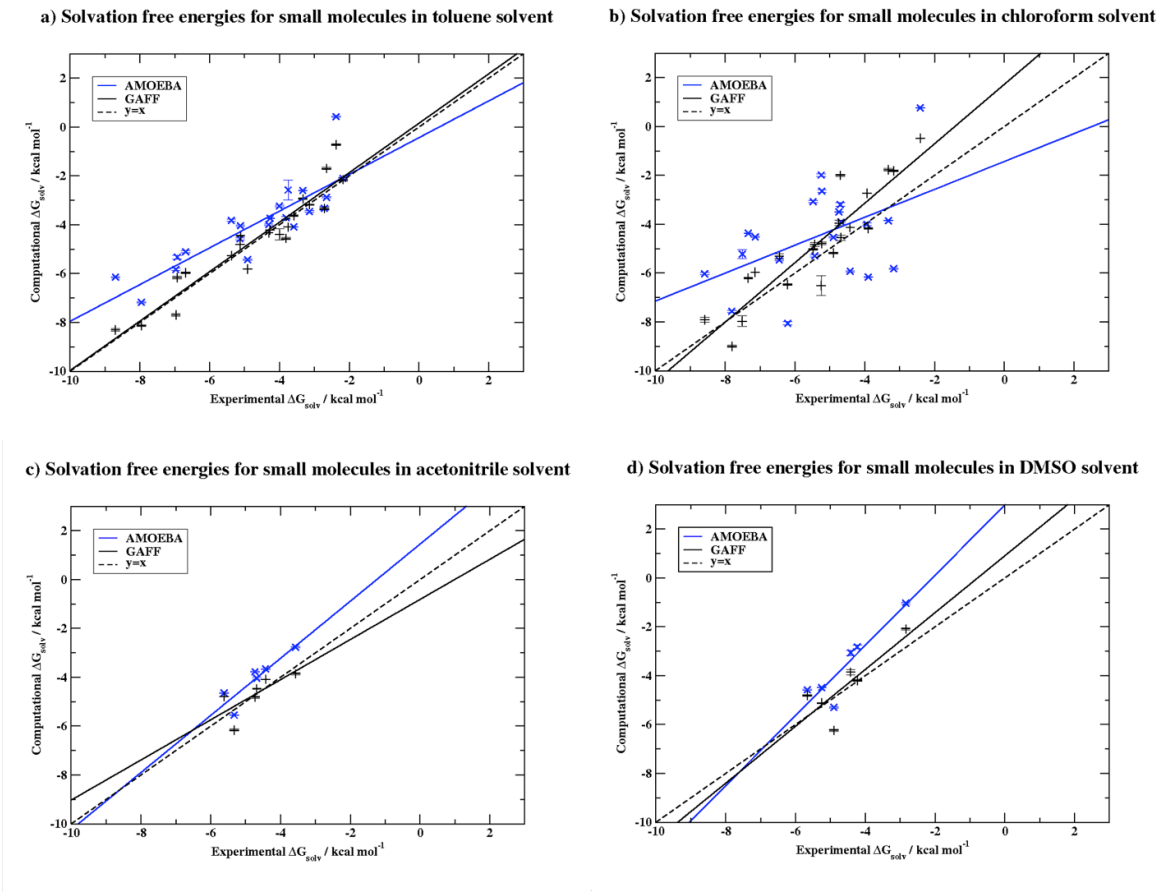


Figure 3.5: AMOEBA (blue) and GAFF (black) calculated ΔG_{solv} for small molecules in toluene, chloroform, acetonitrile and DMSO against experimental ΔG_{solv}. Line of perfect agreement, $y = x$, shown as dashed line. Linear regression in each solvent plot gives the following equations: a) AMOEBA ($y = 0.752\,x - 0.4375$), GAFF ($y = 1.012\,x + 0.153$) b) AMOEBA ($y = 0.571\,x - 1.435$) , GAFF ($y = 1.217\,x + 1.722$) c) AMOEBA ($y = 1.169\,x + 1.452$), GAFF ($y = 0.822\,x - 0.813$) and d) AMOEBA ($y = 1.436\,x + 2.986$), GAFF ($y = 1.164\,x + 0.907$). (Figure taken from Mohamed *et al.*)[224]

3.9.6 Statistical error analysis

Statistical error analysis were performed in this work, since the available experimental data set is very small, it is difficult to assess the performance of the force field based on the comparison of mean metrics. Here, statistical confidence intervals estimated *via* bootstrapping allow a more relevant comparison between metrics to be made. Bootstrapping with replacement was performed for 1000 iterations, and the 95% confidence intervals in all metrics were calculated

from the underlying distributions. Additionally, a Student's paired t-test and Wilcoxon signed-rank test were performed using the original signed and unsigned error distributions (respectively) for AMOEBA and GAFF, to assess whether differences between force fields were statistically significant. The ranges in these metrics computed in all solvents for both AMOEBA and GAFF represented in Table 3.6. Given the magnitude of the associated ranges remains similar between AMOEBA and GAFF, for the results of MUE and MSE for the solvation free energies provided in all solvents, suggesting that the performance across solvents is consistent in terms of error

Table 3.7 shows the t-test and Wilcoxon signed-rank test results, evaluating AMOEBA and GAFF differences in MSE and MUE, there is a significant difference between AMOEBA and GAFF MSE for all solvents except chloroform (with a significance threshold of $p = 0.05$). However, analysis of MUE distributions showed significant differences only in chloroform and toluene. DMSO and acetonitrile yield no significant difference between their very similar ranges of MUE.

3.9.7 Analysis of performance

Generally, the examination of results in Table 3.6 reveals that overall the AMOEBA polarisable force field performs well, but slightly worse than GAFF when compared to the experimental data. There are a number of molecules found to give the largest errors to experiment across all the solvents. Various aspects of the results associated to this performance are discussed as follows:

3.9.7.1 Experimental Error

Ammonia has consistently overestimated (too positive) solvation free energies with AMOEBA and GAFF force fields in both toluene and chloroform solvents. In this case, there may be a doubt in the veracity of experimental data. The experimental free energies of solvation for our solutes were calculated in one of two ways: i) using direct partition coefficients between gas phase and liquid phase, or ii) using partition coefficients between water and non-aqueous solvents, combined with hydration free energies. However, predominantly the latter approach was used - experimental measurements were determined by combining both experimental values for aqueous hydration free energies and partition coefficients measured between water and non-aqueous liquids.[245] The average uncertainty in experimental values of solvation free energies reported by the authors of the Minnesota solvation database is ~0.2 kcal mol^{-1} for the subset used in this study.[245–247] However, this uncertainty is likely to be non-normally distributed amongst the members of the database, such that individual molecules may have larger or smaller errors in

their experimental ΔG_{solv} estimates. The experimental errors for specific molecules are not provided by the Minnesota solvation database, but the consistently poor performance of a molecule across solvents and force fields studied, such as in the case of ammonia, may suggest a larger than average experimental error for that solute.

3.9.7.2 Parameterisation

The parameterisation aspect also has a great impact on the accuracy of solvation free energy calculations. For AMOEBA particularly, it has been shown elsewhere how small changes in parameteristion methodology can give significant differences in hydration free energies.[229] However, owing to the simplicity of the molecules constituting the data set used here, it is difficult to introduce further systematic modifications to the solute parametersation protocol without fundamental change to the underlying parameterisation philosophy (for example, by fitting to solvent-solute interaction energies). Here, the optimum AMOEBA parameterisation protocols has been followed closely for solutes. In particular, multipole coordinate frames and polarisation groups were manually defined, valence parameters were taken from the established amoeba09 parameter set, and atomic multipoles were fitted to molecular ESP calculated using the recommended large basis set (aug-cc-pVTZ). Thus, parameterisation on the whole was performed as per well-established guidelines.[229,241,248]

However, there may also be occasions where parameterisation remains challenging. Mostly, ammonia, n-octane and hexanoic acid molecules were observed to give the largest error in solvation free energies estimation to experiment with AMOEBA force field. The simplest of the molecules studied, such as ammonia, may be highly sensitive to small parameter changes. If the potential of each atom interacting with the solvent is even slightly overestimated, this may contribute to the significant overestimation of the solvation free energies for ammonia in chloroform and toluene. Additionally, generating parameters for n-octane or hexanoic acid was difficult as there are large and consist of extended chains. The parameters may also be affected by the conformation or conformations used in the multipole generation process. For these molecules with extended chains there are many conformations that are low in energy and visited during the MD simulation. It is challenging to select the correct low energy conformation for multipole assignment for those molecules. However here, we did not attempt to include multiple conformations in the ESP fitting process, as the majority of molecules studied had single, fairly rigid, well-defined low energy conformations.

Besides the solute parameters, the AMOEBA solvent models also need to be considered. Results for small molecules in chloroform are consistently the worst, even with solute parameterisation by hand. Liquid phase tests do exist in the paper, but they are fairly simple, describing the chloroform potential, including density and heat of vaporization.[155] These properties have also been evaluated for other solvent models used here. Nevertheless, it should be noted that these measures only validate solvent-solvent interactions and do not assess the accuracy of solute-solvent interactions, as would be necessary for accuracy in our solvation free energy calculations.[220,221,248,249] Since the other solvents models were taken straight from the amoeba09 parameters, perhaps these models have been better tested. Thus, better results were attained for molecules solvated in those solvents.

3.9.7.3 Sampling

Generally, sampling is a common issue when running molecular dynamics simulations. This study has attempted to deal with the sampling problem by performing three repeats for each solute/solvent combination, as demonstrated by the small uncertainties observed for the majority of molecules. In addition, the convergence of free energies was monitored every 200 ps from 400 ps to 2 ns of simulation time (Figure 3.4). The solvation free energies remained stable and consistently converged. Considering the molecules are fairly small, it is not surprising that they may quickly converge. Although it is assumed that sampling issues have been avoided for most of the small molecules, n-octane and hexanoic acid may be exceptions to this rule, as demonstrated by the higher than average standard errors observed in their estimates, particularly in chloroform (Table 3.6). The sampling of different conformations to reach equilibrium may have been problematic during the short timescales simulated here. However, variance in estimates due to differential sampling between repeats did not increase the error systematically between solvents. Moreover the increased uncertainty in predictions it caused was not the predominant driver of poor agreement in chloroform, where other solutes had equal or greater error to experiment.

3.9.7.4 Functional Group Trends

Typically, GAFF performed well for most functional groups with better accuracy to experimental solvation free energies compared to AMOEBA. This improvement spanned both polar and non-polar solvents, and solutes containing a multitude of functional groups. Here, the largest functional group subset tested is amines, consisting of five compounds (ammonia, aniline, methylamine, diethylamine and trimethylamine), for which experimental data was only available

in the non-polar solvents toluene and chloroform. However, there was no clear consistency in observed errors for particular solute functional groups, given its size, the current dataset is limited in its ability to discern trends in functional groups. Indeed, an extended study on a broader data set would be required to investigate functional group trends further.

3.9.7.5 Polarisation in different environments

The solvent models used in fixed-point-charge simulations with GAFF solute parameters had not been optimized for solvation free energy calculations during their respective parameterizations.[220,221,249] It is somewhat surprising, therefore, that all solvents showed consistently reasonable agreement with experiment. In general however, these results may suggest that electronic polarisation may not be crucial in this case. The solvents investigated correspond to common organic solvents with fairly low dielectric constants (all significantly less than water). In this sort of environment electrostatic screening will be lower, and molecular polarisation may have less of an effect on solvation free energies. This may be a reason why the GAFF fixed-point-charge model performs better, especially in toluene where the dielectric constant is very low. Again, in order to really draw firm conclusions, a non-aqueous solvent with the higher dielectric constant than water would be necessary, such as formamide (dielectric constant = 111).[250] However, computational non-aqueous solvation free energy studies are hampered by the scarcity of suitable experimental data for multiple solutes across multiple solvents. Therefore, considering the simplicity of solutes and solvents, simple systems perhaps are better represented by simple force fields rather than application of polarisation terms as incorporated in the AMOEBA force field.

3.10 Conclusion

Overall, both force fields estimated non-aqueous solvation free energies well, with only the AMOEBA chloroform and DMSO results exhibiting MUE above the 1.0 kcal mol^{-1} limit often considered as 'chemical accuracy' in free energy calculations. Our findings show that chloroform solvation free energies have the largest errors to experiment, despite reasonable correlation for GAFF, and are consistent with the recent results of Zhang *et al.*[251] GAFF showed statistically significant improvements in unsigned error over AMOEBA for the 21-solute datasets of toluene and chloroform, and in signed error for all but chloroform. Potential reasons for this discrepancy in AMOEBA performance is likely to be associated with i) parameters of solutes and solvents: AMOEBA parameters, both solvent and solute, have not been tested as extensively or empirically adjusted to recreate thermodynamic properties. This is particularly highlighted by the relatively

poor AMOEBA performance in chloroform. While AMOEBA parameters provide an excellent description of the electrostatic environment surrounding the chloroalkanes, including σ-hole effects, bulk phase thermodynamic properties were not included as targets in the parameter optimization process.[155] While the GAFF force field is more mature as it has been established since 2004,[127] is now a well developed and well understood small molecule force field, whose solute parameters (beyond the independently-derived point charges) have undergone multiple rounds of refinement and been used in multiple other free energy investigations and blind challenges.[214,218,244,252,253] ii) The solvents of low dielectric constant (and often simple solutes) tested here may not require the additional rigour of a polarisable force field for accurate free energy estimates. Evaluation of more challenging solutes and solvents is, however, extremely limited by a lack of suitable experimental data for comparison. Despite these challenges, our broad comparison of potential functions across a range of systems identifies clear opportunities for force field improvements, and we believe further work should ideally focus on more complex systems, where requirements for polarisation may be more apparent. Therefore, further investigations on the clear cases of failure in fixed-point-charge force fields towards the evaluation of the AMOEBA force field performance in protein-ligand interactions will be discussed in the next chapter.

Chapter 4: Evaluating parameter and methodology in binding free energy calculations of cytochrome c peroxidase

4.1 Introduction

This chapter begins to investigate extension of the AMOEBA force field from simple systems to more complex systems including the evaluation of protein-ligand interactions. Initially, clear cases of failure in fixed-point-charge force fields will be identified, followed by testing on a protein-ligand system to identify a protocol for equivalent later work with AMOEBA. This initial focus on binding free energy calculations using the AMBER force field instead of the AMOEBA force field was chosen owing to the limitation of AMOEBA in terms of the lengthy time required for testing.

To identify problems due to a force field model, a robust and reproducible free energy protocol is required for free energy calculations. Here, we explore the parameter and methodology sensitivity with the AMBER force field by changing: a) total protein charge, b) ligand parameters, c) restraints and d) simulation method. In this chapter, we assess the effects of these changes by evaluating the hydration free energies and absolute binding free energies of charged and neutral ligands in the model-binding site of the cytochrome c peroxidase (CCP) protein. We then compare calculated free energies generated against the previously published results from Rocklin *et al.*[204]

4.2 Dataset

In this study, cytochrome c peroxidase protein (Figure 4.1) was identified as a potential test system from the article published by Rocklin *et al.*, in which charged-ligand binding affinities are predicted within the model binding site of the cytochrome c peroxidase protein with the GAFF fixed-point-charge force field.[204] A systematic discrepancy between calculated and experimental binding affinities was observed in that work, and attributed to the absence of explicit electronic polarisation in the classical fixed-point-charge model.

This system is a mutant of yeast cytochrome c peroxidase that introduces a buried cavity binding site connected to the protein surface by a water channel. Upon binding, ligands displace water molecules to bind close to an ionised aspartate residue in the cavity, but remain accessible to solvent via the water channel. Here, electronic polarisation may play a substantial role in

electrostatic binding interactions, especially involving the charged ligand in the diverse environments of both the buried cavity and exposure to the solvent.[204]

Overall, 14 ligands were tested including 3 neutral ligands (Figure 4.2). The ligands were selected based on the availability of experimental binding data and quality of X-ray crystallographic structure (< 2 Å resolution). All bound ligand structures were taken from the Protein Databank (PDB), as shown in Figure 4.2. That high-resolution crystal structures are available for this dataset is a very good reason to choose it, as it eliminates another variable, that of uncertainty in binding pose when comparing force fields using binding free energies.

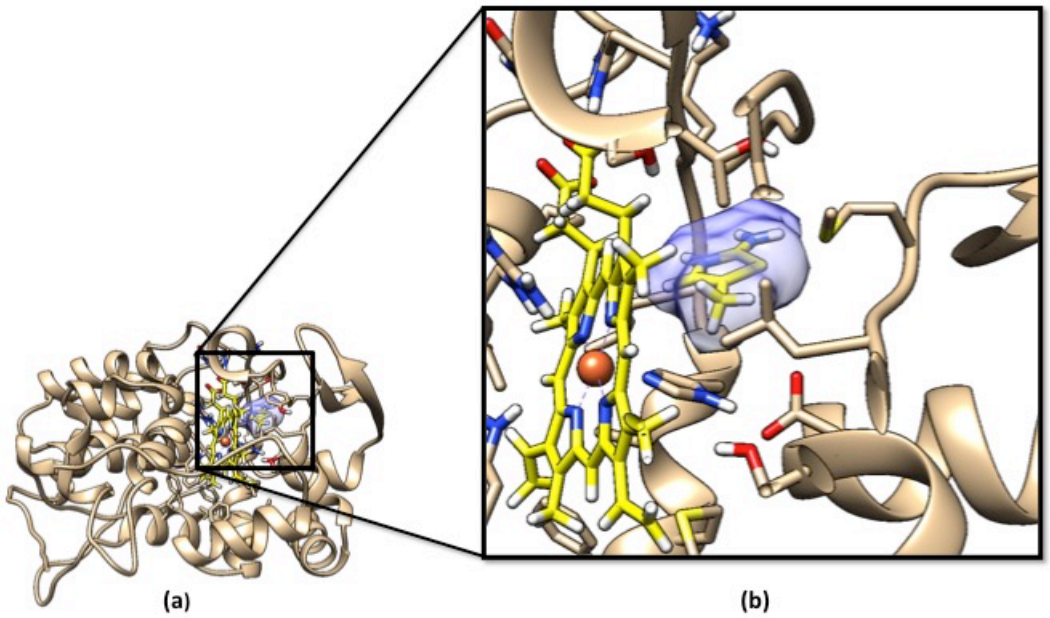
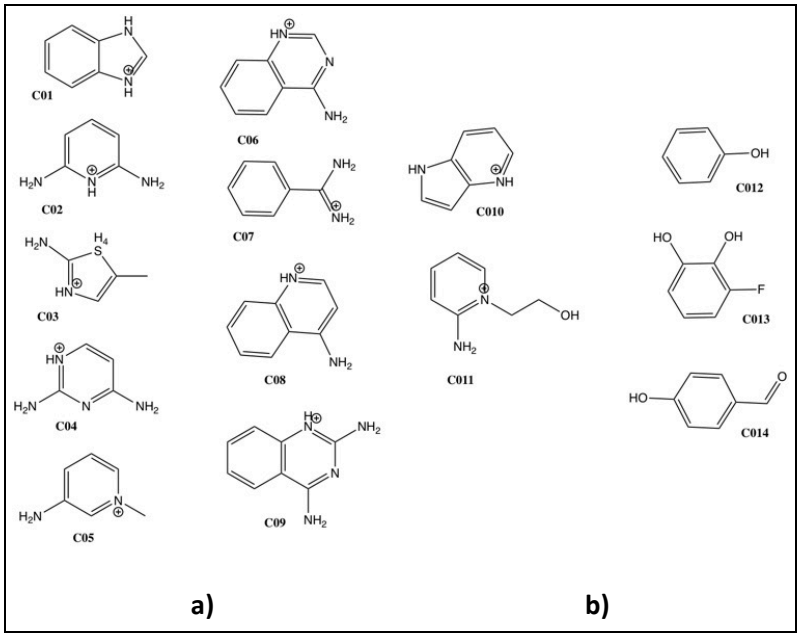


Figure 4.1: Cytochrome c peroxide protein open cavity binding site: a) The binding site in the context of the full CCP complex structure of PDB ID 4JM5. The protein is shown in brown with the ligands shown in yellow, while the illustration of the pocket surface around the C03 ligand is shown in blue b) The close-up view of the buried CCP protein-binding site bound to ligand C03 (blue surface) with the nearby metal ligand heme group on the left.



Ligand	PDB ID	Resolution (Å)
C01	1KXM	1.7
C02	4JM8	1.3
C03	4JM5	1.26
C04	4JM6	1.45
C05	4JM9	1.41
C06	4JMQ	1.41
C07	4JPU	1.41
C08	4JQJ	1.64
C09	4JPT	1.41
C010	4JPL	1.41
C011	4JQK	1.36
C012	4JMW	1.19
C013	4JMA	1.6
C014	4JQN	1.3

c)

Figure 4.2: The structures of 14 ligands chosen for this study: a) Charged ligands (at pH 7), b) Neutral ligands, c) The PDB accession codes and resolutions of the ligand-complex structures.

4.3 Parameterisation

The parameterisation process for the systems was carried out by following the standard AMBER fixed-point-charge parameterisation procedure using the ANTECHAMBER program.[230] The AMBER14SB[238] force field was used to model the protein, while GAFF parameters were assigned for the ligand valence and vdW parameters, and electrostatic parameters assigned as AM1-BCC atomic charges. The water box with the counterions "CL-" and "NA+" was prepared using the TIP3P[254] water model and the ion parameters of Joung and Cheatham.[255] For heme, the parameters were taken from the hemoglobin model in the AMBER parameter database of Bryce and coworkers.[256] However, we modified the charge of iron, by adding +1 charge to create an Fe(III) parameter. Additional FE-NB bonded parameters (the bond between the Fe(III) and co-ordinated histidine) were also added to ensure the heme structure remains planar during the entire simulation. The new equilibrium bond length and angle for FE-NB and NB-FE-NP/NO (Figure 5.1 in Chapter 5, section 5.2) were assigned from the average of all crystal structures of CCP protein used in this work, using force constants of 100 kcal mol^{-1} $Å^{-2}$ and 50 kcal mol^{-1} rad^{-2} similarly to Banci *et al.*[257]

4.4 Protonation states

Protein protonation state is one of many components that may affect the calculation of charged binding affinities. The overall receptor protonation states are dependent on the pH of the system. Thus, the receptor protonation state was initially prepared at pH 4.5, consistent with the pH at which the crystallisation experimental work was performed. Each titratable residue protonation state was decided by running H++[258] and MCCE2.7[259] at pH 4.5 using the apo PDB structure 1KXN[260] without the heme cofactor, leading to a net neutral receptor, while a net charge -1 CCP model was created with the complete protein including the heme. The receptor was modelled using a combination of H++, MCCE prediction together with manual verifications depending on the surrounding environment (Table 4.1). The higher pK_a Asp and Glu were protonated according to the prediction pK_a. The protonation of His residues was determined by investigating the hydrogen-bond network and solvent accessibility. For comparison purposes, we also prepared receptor models with net charge +9 as proposed by Rocklin *et al.*[204] For charged ligands, the protonation states were assigned according to Rocklin *et al.*[204] as shown in Figure 4.2.

In Table 4.1, the comparison between our protonation state definitions and the protonation states defined by Rocklin *et al.*, MCCE and H++ prediction are listed. All the protonation state predictions were performed at pH 4.5 (Rocklin *et al.*, MCCE and our prediction) except for H++, where the prediction was performed at pH 4. The protein protonation residue numbers stated (Table 4.1) referred to the residue numbers assigned to all CCP protein models (Figure 4.2c). Initially, the protein PDB structures (Figure 4.2c) were aligned to ensure they were all the same length. Here, only one protein PDB ID 1KXM was trimmed by discarding three residues from beginning of the protein structure for this purpose. The residue names mean here, eg. ASH is protonated aspartic acid, GLH is protonated glutamic acid, HID is delta-protonated histidine on nitrogen and HIP is protonated histidine on both nitrogens (this is positively charged).

Table 4.1: Comparison of protonation states assigned for the protein–ligand systems in this study and in Rocklin *et al*., along with MCCE and H++ predictions for each residue at pH 4.5 and pH 4 respectively.

Net Charge of Systems					
System	Rocklin *et al.*[a]	Rocklin *et al.*[b]	MCCE[c]	H++[c]	Our prediction[c]
Ligand (L)	1	1	1	1	1
Heme (H)	-1	-1	-1	-1	-1
Protein (P)	-4	10	9	1	0
Receptor (H+P)	-5	9	8	0	-1
Complex (L+H+P)	-4	10	9	1	0

Protein ionisable residues						
Residue No.	Residue Name	Rocklin *et al.*[a]	Rocklin *et al.*[b]	MCCE[c]	H++[c]	Our prediction[c]
31	ASP		ASH	ASH	ASH	ASH
219	ASP			ASH	ASH	ASH
125	CYS					
8	GLU	GLH	GLH	GLH	GLH	GLH
14	GLU		GLH			
29	GLU		GLH	GLH	GLH	GLH
32	GLU		GLH	GLH	GLH	GLH
95	GLU		GLH	GLH		
115	GLU		GLH	GLH		
196	GLU		GLH	GLH		
204	GLU		GLH	GLH		
209	GLU		GLH	GLH	GLH	
216	GLU		GLH	GLH		
245	GLU		GLH			
262	GLU	GLH	GLH	GLH	GLH	GLH
266	GLU		GLH			
285	GLU		GLH	GLH	GLH	
286	GLU		GLH			
3	HIS	HIP	HIP	HIP	HID	HIP
49	HIS	HIP	HIP	HIP	HIP	HIP
57	HIS	HIP	HIP	HIP	HID	HIP
93	HIS	HIP	HIP	HIP	HIP	HIP
172	HIS	HID	HID	HIP	HID	HID
178	HIS	HIP	HIP	HIP	HIP	HIP

Rocklin *et al.*[204] assigned different protonation states for electrostatics and vdW simulations denoted by [a] electrostatics simulations [b] vdW simulations, while MCCE, [259] H++[258], and our prediction, assigned identical protonation states for all simulations including electrostatics and vdW free energy calculation steps. This is denoted by [c] all simulations.

4.5 Free energy calculations

The absolute binding free energies for all the cationic and neutral ligands bound to CCP protein were evaluated by performing alchemical free energy calculations using the existing protocol from Rocklin *et al.*[204] as the initial protocol to investigate the robustness of the method. The prediction of binding affinities of each compound was computed based on the thermodynamic cycle in Figure 4.3. The overall process is described by the equations below:

$$\Delta G_{bind} = G_{P+L \rightarrow PL} \quad (4.1)$$

$$\Delta G_{bind} = -\Delta G_{hyd} - \Delta G_{charging,vac} + \Delta G_{complex} \quad (4.2)$$

$$\Delta G_{complex,} = \Delta G_{charging,prot} + \Delta G_{coupling,prot} \quad (4.3)$$

Where ΔG_{bind} is the absolute binding free energy of ligand in complex, $\Delta G_{hyd,}$ is the hydration free energy of transferring the ligand from vacuum to solution, $\Delta G_{charging,vac}$ is the free energy of charging the ligand in vacuum (intramolecular interactions only) and $\Delta G_{complex}$ is the free energy change of complexation of the ligand and protein receptor (introduction of ligand into the binding site), divided into $\Delta G_{charging,prot}$ as the charging free energy of the ligand in complex with the protein receptor, and $\Delta G_{coupling,prot}$, the coupling free energy of the ligand in complex with the protein receptor.

Here, the binding free energies of the protein-ligand (PL) complex for the reactions of protein (P) and ligand (L) in solution were calculated in three sets of calculations (Figure 4.3): i) Ligand in solution: Free energies of transferring the ligand from solution to vacuum ($-\Delta G_{hyd}$) ii) Ligand in vacuum: Vacuum charging energy of ligand ($\Delta G_{charging,vac}$) iii) Ligand in complex : Free energy change of coupling the ligand with the protein ($\Delta G_{complex,}$), involve charging ($\Delta G_{charging,prot}$) and coupling ($\Delta G_{coupling,prot}$) interactions of the ligand in vacuum to the complex in protein in bulk solution.

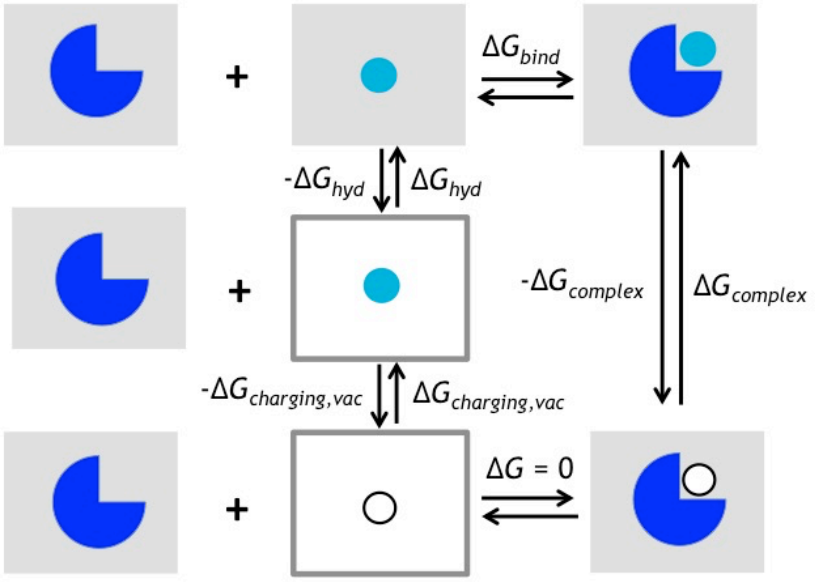


Figure 4.3: Thermodynamic cycle used to calculate the absolute binding free energy of ligand in complex, ΔG_{bind}. Three sets of calculations were required for evaluating the ΔG_{bind} : i) $-\Delta G_{hyd}$ simulations of ligand run in both solution and vacuum ii) $\Delta G_{charging,vac}$ simulations of ligand run in vacuum ii) $\Delta G_{complex}$ simulations of ligand in complex with the protein receptor run in solution. The blue third quarter circle represents a protein, the cyan circles represent a fully charged ligand interacting with the environment, the no filled circles represents a discharged ligand and completely decoupled from the environment, while the grey square box denotes a simulation run in solution phase and the no filled square box with denotes a simulation run in a gas phase.

For $-\Delta G_{hyd}$ and $\Delta G_{charging,vac}$ calculations, the simulations of ligand in solution and vacuum were performed by adopting the thermodynamic cycle of hydration free energies shown in the previous chapter, chapter 3, section 3.5 (Figure 3.3), where the $\Delta G_{charging,vac}$ is equivalent to $-\Delta G_{discharging,vac}$ calculation in Figure 3.3. However, in our case, corrections were applied to the electrostatic free energy calculations for charged ligands. There were two types of correction applied to the electrostatic calculations for Particle Mesh Ewald simulations that involved an alchemical change in net charge, defined by Kastenholz and Hunenberger[261] as: i) Type B artefact to correct for periodicity ii) Type C artefact to correct non-Coulombic errors. The details about this correction were discussed in chapter 2, section 2.4.5.1.

Here, the simulation of the ligand complexed with the protein receptor, $\Delta G_{complex}$ calculations were actually evaluated by performing simulations in the $-\Delta G_{complex}$ direction, shown in Figure 4.3. In this process (Figure 4.4), the systems were simulated starting with the complex with the ligand and protein fully interacting. To decouple the ligand from the system, the columbic interaction (ΔG_{ele}) needed to be removed first, followed by Lennard-Jones interactions (ΔG_{vdw}). Again, for

discharging the cationic ligands in the complex, identical PME corrections for charged ligands were applied to those used in the free energy calculations of ligands in solution. However, before proceeding with this step, a set of restraints needed to be added between the ligand and protein (ΔG_{rest_on}) to avoid the problem of 'wandering ligands' where the ligand leaves the binding pocket when the interactions are removed. Applying the restraints to the ligands is crucial, as this will restrain the ligand to its initial orientation and position in the protein. In this step, restraints on the protein receptor were also applied. By restraining defined protein dihedrals, sampling of conformations that have not been observed in the crystal structure can be prevented. Finally, the ligand harmonic restraints were then removed (ΔG_{rest_off}) using the analytical solution which includes a correction for the standard state described by Boresh et. al.[73] without the need to run any further simulations. The details about the set of restraints applied were discussed on the later section (section 4.6.2).

Figure 4.4: The protocol adopted for calculating the free energy change of forming the protein-ligand in complex, $\Delta G_{complex}$ (The sum of the free energies in the direction shown is $-\Delta G_{complex}$). Four sets of calculations run in solution (grey square box) are required: i) $\Delta G_{rest,on}$: confining the ligand with harmonic restraints ii) ΔG_{ele}: discharging the ligand inside the protein iii) ΔG_{vdw}: decoupling the ligand vdW interaction to the protein iv) $\Delta G_{rest,off}$: releasing the ligand harmonic restraints. The blue three-quarter circle represents a protein, the cyan circles represent a fully charged ligand interacting with the environment, the yellow circle represents a fully discharged ligand interacting with surrounding environment, the unfilled circle represents a discharged ligand completely decoupled with its environment, while the red dotted lines indicates the restraints applied to the ligand and protein

4.6 Simulation protocol

As mentioned previously, the simulation protocol performed here was adopted from the existing simulation protocol for calculating the absolute binding free energies developed by Rocklin *et al.*[204] These protocols were use as the initial protocol and were then modified across the investigation to explore the robustness of the simulation method to changes in simulation time and simulation method, e.g. replica exchange, with a fixed-point-charge force field.

4.6.1 System preparation

The systems were parameterised as per section 4.3. The ligand or protein-ligand complex was solvated in a cuboid box, filled with TIP3P[254] water molecules with a minimum distance 0.8 nm from the protein to the nearest box edge, using the LEAP tool in AMBER14.[195] An appropriate number of counterions Na^+ and Cl^- (~28 ions Na^+ and Cl^-) to reach a bulk concentration of 150 millimolar were added after neutralisation of the systems. Initially, the systems were minimised using the steepest descent algorithm for 2500 steps. The systems were then heated slowly to 300 K in the NVT ensemble for 50 ps, followed by 100 ps pressure equilibration to 1 atm using NPT at 300 K. A timestep of 2 fs and a Langevin integrator[186] was applied to the simulations to maintain temperature. A Berendsen barostat[235] was employed to maintain the pressure of the systems. Finally, these equilibrated structures were used as the initial structure in three independent repeat simulations for the following series of free energy calculations.

4.6.2 Unrestrained simulations for reference orientations selection

For comparison with experiment, binding free energies must be calculated to and from completely unrestrained endpoint simulations. However, here, a set of suitable restraints between the ligand and the protein receptor was introduced in the simulation protocol of the ligand in the complex. This step is essential to obtain converged results in the calculations by diminishing any sampling problems due to the ligand leaving the binding site when the ligand is fully decoupled from the protein. The binding free energy calculations therefore involve computing the energy of restraining the ligand to the binding site, and of releasing these restraints again at the alternate end of the coupling step. In the thermodynamic steps in Figure 4.4, the presence of restraints confines the ligand position and orientation relative to the receptor, confining the ligand to a certain volume. The addition of these restraints in the complex system can be evaluated by standard alchemical means, and there is an analytical solution for removing the restraints again and a correction for the standard state.[73]

To restrain the ligand's orientation and position in the binding site, six degrees of freedom – $\vartheta_A, \vartheta_B, \emptyset_A, \emptyset_B, \emptyset_C, r_{aA}$ (two angles, three dihedrals and one distance) between the ligand and protein were defined respectively according to Boresch *et al.*[73] Three reference atoms were selected from the largest domain in the protein as anchor atoms upon which to add harmonic restraints to the ligand. For our systems, a= Cβ, b= Cα and c= His172[204] on the protein receptor have been chosen as attachment point for two angles, three dihedrals and one distance for protein-ligand. Typically, reference atoms defined by A, B, C in the ligand was taken to be any three heavy atoms in the ligand.[73] Initially the same reference atoms were chosen for CCP systems

in our work for ligands C01 to C014. However the reference atoms chosen for each ligand were different depending on the ligand structure. The details of the atoms used as reference atoms between the ligand and protein for the 14 ligands in CCP are shown in Appendix A, Table S1.

To identify the reference orientation and position of the ligand, 2 ns unrestrained MD simulations of the fully coupled system in the bound state and unbound state were run to produce a preliminary orientation and position of residues in both states. Histograms of the distance, angle and dihedral distributions were created for each ligand-protein degree of freedom (Appendix A, Figure S1) to observe the range of the distribution sampled in both states of unrestrained simulations. From these probability distributions, the most favourable reference orientation was defined based on the centre of the histogram for each degree of freedom. The details of the reference position and orientation assigned for each ligand-protein in our study are listed in Table 4.2.

Table 4.2: The reference orientations of 14 ligands in CCP protein. Distance (r_{aA}) is in units of Ångstroms, remaining angles (ϑ_A, ϑ_B) and dihedrals ($\emptyset_A$, $\emptyset_B$, $\emptyset_C$) are in units of degrees.

	Degree of freedom					
Ligand	r_{aA}	ϑ_A	ϑ_B	$\emptyset_A$	$\emptyset_B$	$\emptyset_C$
C01	4.50	75	70	-140	-110	-50
C02	3.90	100	73	-138	-115	90
C03	4.10	95	75	-152	25	-100
C04	3.60	102	100	-156	-90	85
C05	3.75	90	80	-140	-130	100
C06	3.80	107	102	-163	-110	92
C07	4.10	103	69	-167	100	-88
C08	3.75	107	101	-160	-93	-92
C09	4.00	100	102	-160	-112	100
C010	4.10	90	80	-160	-160	110
C011	4.10	75	65	-163	77	-98
C012	4.62	99	60	-170	160	-62
C013	3.80	95	82	-137	-125	102
C014	4.20	115	95	-156	-62	52

4.7 Preliminary simulation protocol

The preliminary simulation protocol explained here was applied as the initial protocol and was then modified across the investigation to explore the robustness of the simulations method to changes in simulation time and simulation method, e.g. replica exchange, with the fixed-point-charge force field.

4.7.1 Production simulation details

Production simulations were performed using AMBER16[195] with a 2 fs time step using the Langevin integrator to propagate the dynamics at a temperature of 300 K. The MD simulations were run in the isothermal-isobaric ensemble (NPT) with pressure maintained using a Berendsen barostat over the simulations. The long range electrostatics interaction were treated using Particle mesh Ewald (PME) summation with the real space cutoff of 8 Å and van der Waals interaction cutoff were set to 8 Å with an analytical long-range correction. The details of each simulation performed are as followes.

4.7.1.1 In solution

The simulations of ligand in solution were performed to evaluate hydration free energies for transferring the ligand from solution to vacuum ($-\Delta G_{hyd}$) as in Figure 4.3., adopting the general protocol of hydration free energies for transferring the ligand from vacuum to solution explained in the previous chapter, Chapter 3 (Figure 3.3). However the simulations were run with slightly different methods e.g. simulation time. The simulations details for evaluating $-\Delta G_{hyd}$ were as follows:

Each calculation was performed using the equilibrated structure prepared in section 4.6.1 as the initial structure. 20 ns simulations were run for the electrostatic discharging steps in solvent and vacuum by linearly scaling down the electrostatics across 11 λ windows, λ = 0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0.[224] λ = 0.0 refers to the ligand fully charged and λ = 1.0 to the ligand fully uncharged. The same number of λ windows were also utilised for the van der Waals decoupling step with a softcore potential employed. To perform simulations and evaluate free energies, the PMEMD module was used for solution phase steps, while SANDER module was used for gas phase steps. The energy differences between each intermediate state were saved every 10 ps and the first 2 ns of simulations were discarded as equilibration. Free energies were analysed using BAR analysis. Finite size corrections for a) Type B artefacts and b) Type C1 artefacts were applied for the electrostatic free energy calculations (discharging free energies) computed using the equations described in chapter 2, section 2.4.5.1.

4.7.1.2 In complex

In the complex, the complexation free energies of the ligand ($\Delta G_{complex}$) were calculated using the protocol shown in Figure 4.4 and provided by equation 4.3. The $\Delta G_{complex}$ calculations were evaluated by performing -$\Delta G_{complex}$ simulations, as this will result in the same free energy difference just with the inverted sign. Here, the free energies of complexation were initially computed by restraining the ligand to the protein receptor only. No dihedral restraints on the receptor were applied.

Simulations in the complex were initiated from the unrestrained equilibrated structure used as input for determining ligand reference orientations. For the first calculation step, the ligands were restrained in the protein binding site, employing the potential form in equation 4.4.

$$U = \frac{k\lambda}{2}\,(\xi - \xi_0)^2 \tag{4.4}$$

Here, the ξ is the instantaneous value of the specific degree of freedom, ξ_0 is the reference value and k is the force constant (k = 10 kcal mol^{-1}Å^{-2} (distance restraints), k = 50 kcal mol^{-1} rad^2 (angle restraints), and k = 50 kcal mol^{-1} (torsional restraints). The restraints were linearly scaled up across 11 windows, with λ = 0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0. These intermediate steps are essential to prevent large forces while running these simulations. For each λ window, a 20 ns MD simulation using PMEMD was performed with atomic coordinates saved every 10 ps. The first 2 ns of simulation were discarded for equilibration. The free energies of confining the harmonic restraints on the six protein-ligand degrees of freedom were evaluated using BAR analysis.[190] To compute the free energies of releasing the ligand harmonic restraints, an analytical solution described by Boresch *et al.*[73] was used instead of running a simulation. The formula employed was as follows:

$$\Delta G_{rest,off} = -kT \ln \left[\frac{8\pi^2 V^0}{r_0^2 \sin\theta_{A,0} \sin\theta_{B,0}} \frac{\left(K_r K_{\theta,A} K_{\theta,B} K_{\emptyset,A} K_{\emptyset,B} K_{\emptyset,C}\right)^{\frac{1}{2}}}{(2\pi kT)^3} \right] \tag{4.5}$$

Respectively, k denotes the ideal gas constant, T the temperature (Kelvin), V^0 the volume corresponding to one molar standard state (1660 Å^3), r_0 the restraint's reference distance, ϑ_A and ϑ_B the restraints' reference angles, $\emptyset_A$, $\emptyset_B$ and $\emptyset_C$ the restraints' dihedral angles and K_x the assigned force constant for r_0 , ϑ_A, ϑ_B, $\emptyset_A$, $\emptyset_B$, $\emptyset_C$.[73]

The next calculation is discharging the ligand inside the protein. These simulations were run by scaling down the charges of the ligand in the protein. The same lambda values as those used for simulations of the ligand alone in solution were employed (λ = 0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0) to perform these simulations. Using PMEMD, simulations were run for 20 ns, with

the trajectories saved every 10ps. The first 2ns were discarded for equilibration. The energy differences between the states were analysed using BAR analysis.

The final calculations involved the decoupling of ligand in the protein. During the simulations, only the Lennard-Jones Interactions between the ligand and protein were turned off while the intramolecular Lennard-Jones interactions remained fully-on. Again, the same spaced λ windows and an identical protocol as above were utilised for simulating these calculations.

Ultimately, the free energies involving changes in net charge in the ligand were corrected for artefacts resulting from the artificial periodicity of the cell and non-Coulombic electrostatics, as explained in chapter2 section 2.4.5.1.

4.8 Optimised simulation protocol

The optimised simulation protocol explained here was applied to the further simulation protocol run for evaluating the sensitivity of components that affect the binding free energies of ligands in CCP. This protocol was then used as the optimised simulation protocol for calculating the absolute binding free energies of 14 ligands in cytochrome c peroxidase protein.

4.8.1 Production simulation details

For each system, the equilibrated structures prepared in section 4.6.1 again underwent minimization for 2500 steps of steepest descent, and another NVT equilibration over 150 ps at 300 K. The final structures from this equilibration were used as starting configurations for each series of free energy calculations. The following production simulations were performed with a 2 fs time step using the Langevin integrator in AMBER16[195] at the temperature 300 K. The production simulations were also run in canonical ensemble (NVT) at a fixed simulation volume.

4.8.1.1 In solution

The hydration free energy evaluation for transferring the ligand from solution to vacuum was performed as explained in Section 4.7.1.1 above with slight modifications as followed. All the simulations were performed in NVT ensemble for 6 ns, with the first 1ns discarded as equilibration. The electrostatic interactions (discharging in solvent and vacuum) were run with 5 λ windows, λ = 0, 0.25, 0.5,0.75,1.[204] Meanwhile, 16 λ windows were used for van der Waals decoupling[204], λ = 0, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95, 1.0[204] employing a soft core potential. BAR simulations were used to compute the free energies between adjacent windows. The energies differences between each intermediate states were saved every 2 ps. The identical method as mentioned previously, using BAR employed for analysis.

4.8.1.2 In complex

The binding free energies of the ligand in protein were calculated using the thermodynamic cycle shown in Figure 4.4 with both ligand and protein restraints applied. The same series of simulations as described in Section 4.7.1.2 were adopted with an additional contribution from dihedral restraint energies in the protein receptor.

Accompanying the alchemical simulation in the complex, the 'confine and release' protocol of Mobley and coworkers was applied for specific protein dihedrals in order to improve the convergence in the binding free energy calculations.[204] Dihedral restraints were applied to side chains that showed rare transitions between rotamers as ineffective sampling of this angle could contribute significantly and/or lead to poor convergence in the overall binding free energies. The following protein dihedral angles were restrained in all the alchemical simulations, using a force constant of 10 kcal mol^{-1} [204] i) Dihedral of Gly175; Leu174 Cα - LEU174 C - Gly175 N - Gly175 Cα (160°) ii) Dihedral of Met226; Met226 C- Met226 Cα- Met226 Cβ- Met226 Cγ (160°) iii) Dihedral of not contiguous atoms; Leu199 C - Asn200 Cα – Asn200 Cβ - Asn200 Cγ (-50°) and iv) Heme dihedral; Hem290 C2D - Hem290 C3D - Hem290 CAD – Hem290 CBD (90°) (Figure 4.7).

The free energies of applying the protein dihedral restraints in the binding site were calculated by performing 6 ns Hamiltonian Replica Exchange simulations in AMBER16.[195] The calculations involve the *apo* protein systems alone. 8 replicas were set up independently with defined dihedral force constants scaled as $k = \lambda, k_0$, and λ = 0.0 (zero force constant), 0.15, 0.35, 0.5, 0.65, 0.75, 0.85, 1.0 (with full force constant). Exchanges between adjacent replicas were trialled every 2 ps. The protein was fully unrestrained during the simulations. To evaluate the confine step protein dihedral restraint energies, post processing of the trajectory of the λ = 1.0 replica was carried out with the restraints applied. The free energies of confining the dihedral to the specific position were generated using the Zwanzig equation.

The free energies of releasing the protein dihedral restraints in the binding site were calculated identically to those of the confine step, except the *holo* protein-ligand complex was used in these calculations. The Zwanzig equation was again used evaluate the free energies of releasing protein dihedrals, except the definition of initial and final states was reversed compared to the confining step.

The free energies of applying restraints to the decoupled ligand were evaluated using the same analytical calculations as section 4.7.1.2. However, energies of confining the ligand harmonic restraints were calculated using MD simulations of 1 ns simulation in length. Here, 8 series of intermediate state simulations were run with, λ = 0.0 (zero restraint) 0.15, 0.35, 0.5, 0.65, 0.75,

0.85, 1.0 (full restraint). The energy of confining the ligand harmonics restraints was calculated using BAR analysis, on snapshots taken every 2 ps and with the first 200 ps discarded as equilibration.

The simulations of discharging the ligand inside the protein were performed using Hamiltonian Replica Exchange with AMBER16[195] with both ligand and protein restrained. 6 ns simulations with 8 replicas were simulated with exchanges between adjacent replicas trialled every 2 ps. Replicas were run with differently scaled ligand charges, from λ = 0.0 (ligand fully charged), 0.15, 0.35, 0.5, 0.65, 0.75, 0.85, 1.0 (ligand fully uncharged). The free energies of discharging the ligand in the protein were then evaluated using BAR analysis.

Free energies of decoupling protein-ligand Lennard-Jones interactions were also calculated from Hamiltonian Replica Exchange simulations, 6 ns in length, with the exchanges trialled every 2 ps. Here, 16 replicas were set up spaced at λ = 0 (fully interacting), 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95, 1.0 (non-interacting). The free energies between states were then analysed using BAR.

Finally, for charged ligands, identical free energy corrections to those mentioned previously (chapter2, section 2.4.5.1) were applied for correcting the error in electrostatic interactions.

4.9 Result and discussion

4.9.1 Electrostatics parameter sensitivity

To investigate the effect of parameters on the free energy calculations, free energies for transferring the ligand from solution to vacuum and absolute binding free energies were computed using the simulation method on section 4.7, with variations of either protein or ligand parameters. Binding free energies evaluated the effect of changing the protein overall charge (i.e. protonation state), while free energies for transferring the ligand from solution to vacuum evaluated ligand parameter effects.

4.9.1.1 Protein charge effect (protonation states)

The effect of protein charge was explored by preparing the system using different protein protonation states. Each system, (protein with ligand C01, C03 or C05) were set up using either a net charge +9 or a net charge -1 on the protein. The selection of these protonation states was discussed previously in section 4.4. Absolute binding free energies of ligand C01, C03 or C05 to cytochrome c peroxidase in different charge states are available in Table 4.3.

Table 4.3: Summary of calculated ΔG_{bind} for ligand C01, C03 and C05 with the net charge +9 and -1 on receptor compared to calculated absolute binding free energies by Rocklin *et al.*[a] [204] Uncertainties calculated as standard error over 3 repeats.

	ΔG_{bind} (kcal mol^{-1})		
Ligand	**Rocklin *et al.*[a]**	**Protein charge set 1[b]**	**Protein charge set 2[c]**
C01	-7.84 ± 0.00	-9.88 ± 0.23	-9.66 ± 0.59
C03	-6.41 ± 0.17	-3.17 ± 0.35	-4.30 ± 0.39
C05	-2.71 ± 0.24	-5.80 ± 0.21	-6.74 ± 0.32

The absolute binding free energies were evaluated with different net charge on receptor: [a] net charge (+9) for electrostatics and net charge (-5) for van der Waal simulation,[204] [b] net charge (+9) and [c] net charge (-1).

The different net charges of the protein give slightly different free energies, with a mean difference of < 1 kcal mol^{-1} in ΔG_{bind} between protein net charge +9 (Protein charge set 1) and net charge -1 (Protein charge set 2) across the three ligands. In fact, there were also differences between our calculated ΔG_{bind} (net charge +9) and the equivalent Rocklin *et al.* prediction (+9), with a mean $\Delta\Delta G$ of approximately ~ 2 kcal mol^{-1}. The differences to Rocklin *et al.* results were expected, however, as the simulations were run with slightly different setup. Their systems have been setup with net charge +9 for electrostatics calculations but net charge -5 for vdW calculations on the protein. For our simulations on protein net charge +9, a consistent system setup with the identical net charge for all simulations (including electrostatics and vdW) was employed, instead of using different net charge systems setup for electrostatics and vdW simulations. To better understand where the differences in the free energies came from, the effect of ligand parameters was also explored.

4.9.1.2 Ligand parameter effect

Initially, the free energies for transferring the ligand from solution to vacuum for all 14 ligands was compared to the equivalent free energies generated by Rocklin *et al.* (Figure 4.5). These free energies are equal to $-\Delta G_{hyd}$. Overall, the two datasets showed an excellent correlation, although some ligands, especially ligand C02, did not agree well. Upon examination of the generated GAFF parameters for ligand C02, we noted particular discrepancies between the point charges generated here and those quoted by Rocklin *et al*. (Figure 4.6).

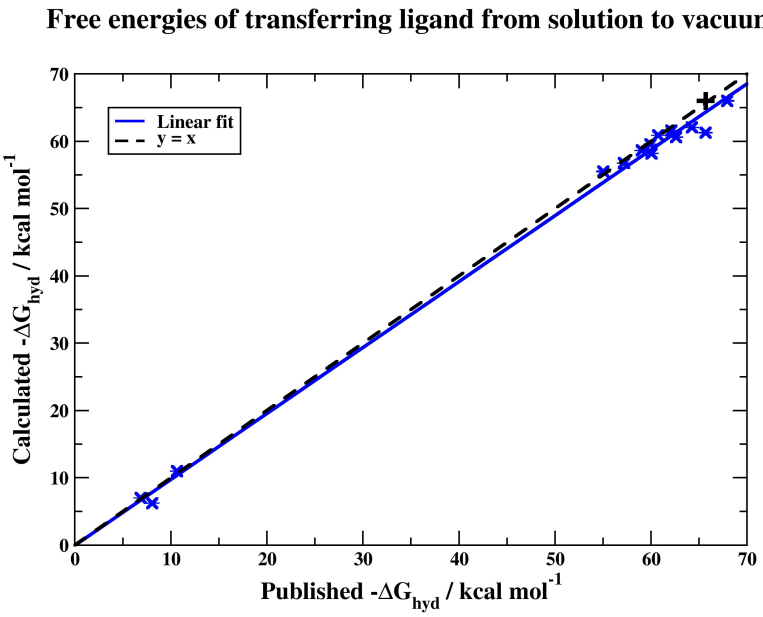


Figure 4.5: The hydration free energies of the ligands with our parameters (blue crosses, solid line) against those with Rocklin *et al.* parameters (black '+', dashed line).

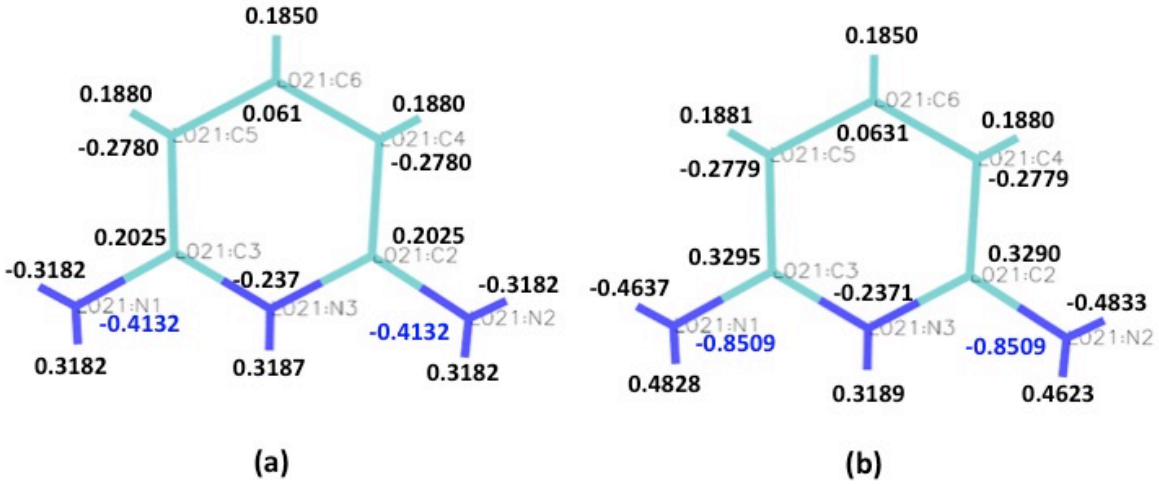


Figure 4.6: Comparison of parameters defined to each atom in ligand C02: a) Our generated parameters b) Rocklin *et al.* parameters. All the partial charges parameters assigned for ligand C02 atoms were labelled in black except for the Nitrogen, partial charges which were labelled in blue.

The most noticeable differences were shown on both symmetrical Nitrogen atoms in the compound (atoms N1 and N2) shown in Figure 4.6. To explore where the sensitivity in $-\Delta G_{hyd}$ came from we reran the simulations for ligand C02 using parameters taken from Rocklin *et al*. The hydration free energy computed with the original Rocklin *et al*. parameters was plotted with the '+' symbol in Figure 4.5 and a quantitative comparison between the free energies calculated using our parameters and Rocklin's parameter in Table 4.4.Using the identical parameters we observed

only a very small difference in -ΔG_{hyd} of 0.33 kcal mol^{-1} compared to the published results, while there is a substantial difference in free energies, 4.40 kcal mol^{-1}, when using our generated parameters. Thus, the difference to the published -ΔG_{hyd} for ligand C02 was owing to the slight differences in parameters generated using a different version of GAFF force field here, as described in the parameterisation process of Section 4.3.

Table 4.4: Summary of performance metrics for calculated hydration free energies with our generated parameters. Uncertainties calculated as standard error over 3 repeats

Metrics	**Calculated -ΔG_{hyd}, Parameter set 1[a]**
MUE (kcal mol^{-1})	1.22 ± 0.63
MSE (kcal mol-^{1})	-1.06 ± 0.71
***R*^{2}**	0.99
Ligand C02	**Difference -ΔG_{hyd}, to Rocklin *et al.* (kcal mol^{-1})**
Parameter set 1[a]	4.40
Parameter set 2[b]	0.33

[a] Our generated parameters [b] Rocklin *et al.* parameters.[204]

The different partial charges assigned to the atoms, especially to Nitrogen (atoms N1 and N2) resulted in a huge impact to the -ΔG_{hyd} calculated for ligand C02. Dynamical differences between the two parameter sets were therefore also probed (Appendix A, Figure S2).

For a basic conformational analysis of the ligand C02 atoms, all-atom RMSD (Root Mean Square Deviation) and RMSF (Root Mean Square Fluctuation) were calculated using trajectories from the electrostatics interaction simulations. Both parameter sets showed similar RMSD over the course of the simulation. RMSF over all the atoms indicated the fluctuations involved the NH_2 atoms - 3,2, 9 and 10 in compound C02 using our parameter set, and equivalent to atoms 13, 14, 15 and 16 in the Rocklin *et al.* parameter set. Thus, the difference in the ΔΔ*G* calculated between the parameter sets is not due to the hydrogen atom dynamics, but just the nitrogen charges.

4.9.2 Methodology sensitivity

We explored the effects of both protein and ligand restraints (to avoid asymmetric conformational sampling) and Hamiltonian replica exchange (to enhance ligand sampling across lambda windows) applied during the simulations.

4.9.2.1 Restraint effects (Confine and release method)

In this section, the ligand and protein restraints effects to the binding free energies were assessed.

4.9.2.1.1 Ligand restraints

At the endpoint of the ligand decoupling simulations, when the protein-ligand intermolecular interactions are fully switched off alchemically, the decoupled ligand is free to explore the entire volume of the periodic box. This may cause problems in the calculation of Δ*G* for the final decoupling step, as Δ*G* will be highly dependent on the configurations sampled by the ligand and any potential 'clashes' with protein or solvent atoms. As such, a set of ligand restraints was introduced.[204] Six suitable harmonic restraints (Table 4.2) were applied to ligand C01 in the binding site as the solution for better sampling during this simulation. To determine an appropriate strength of force constant for this set of restraints, the effect of two sets of force constants on the ligand sampling and calculated absolute binding free energies was evaluated.

The two sets of restraint force constants were employed as follows: i) k = 10 kcal mol^{-1} Å^{-2} for distance restraints, k = 10 kcal mol^{-1} rad^{-2} for angle restraints and k = 10 kcal mol^{-1} for torsional restraints; or ii) k = 10 kcal mol^{-1}Å^{-2} for distance restraints, k = 50 kcal mol^{-1} rad^{-2} for angle restraints, and k = 50 kcal mol^{-1} for torsional restraints. The applied restraints involved three reference atoms on the protein receptor C01 (172@CB, 172@CA and 172@N) and three reference atoms on ligand C01 (291@C3A, 291@C2 and 291@C2) as listed in Table 4.2 (section 4.6.2). The ligand harmonic restraints were then removed using the analytical solution includes a corrections for the standard state (Equation 4.5) describe by Boresh *et al.*[262] Computed free energies for these simulations are given in Table 4.5.

Table 4.5: The comparison of each component of Δ*G* (kcal mol^{-1}) generated in the simulation with different sets of force constants on the ligand harmonic restraints applied to free energy calculations of C01 ligand. Uncertainties calculated as standard error over three repeats.

Component	**Δ*G* (kcal mol^{-1})**		
	Rocklin *et al.*	**Ligand restraint set 1[a]**	**Ligand restraint set 2[b]**
Binding	-7.84 ± 0.00	-9.66 ± 0.77	-9.46± 0.52
Confine harmonic restraints	9.07± 0.00	7.04 ± 0.00	9.44 ± 0.00
Release harmonic restraints	-1.10 ± 0.00	-1.45 ± 0.02	-2.93 ± 0.02

Two set of ligand restraints with difference force constant (*k*) on harmonics restraints applied: [a] k = 10 kcal mol^{-1} Å^{-2} (distance restraints), k = 10 kcal mol^{-1} rad^{-2} (angle restraints) and k = 10 kcal mol^{-1} (dihedral restraints) and [b] k = 10 kcal mol^{-1} (distance restraints) and k = 50 kcal mol^{-1} rad^{-2} (angle restraints) and k = 50 kcal mol^{-1} (dihedral restraints).

More positive ΔG_{bind} with smaller uncertainty, ± 0.52 kcal mol^{-1} were obtained with ligand restraint set 2 using the greater angle/torsional restraint force constant, compared to ligand restraint set 1 with uncertainty of ± 0.77 kcal mol^{-1}. As expected, tighter angle distributions were observed by having greater force constants (Appendix A, Figure S3 and S4). Although ligand restraint set 1 (the smaller force constants) was able to restrain the ligand in the electrostatic interactions calculations, ligand orientation fluctuated noticeably more in van der Waals calculations, as the ligand was slowly decoupled completely. Thus, the force constants of ligand restraint set 2 is a suitable strength of force constant for restraining the ligand in order to avoid larger errors in overall ΔG_{bind} estimations.

4.9.2.1.2 Protein restraints

The effects of protein restraints were explored by evaluating the Δ*G* of this additional step and the variance in the binding free energy calculations. In this assessment, the protein dihedrals (Figure 4.7) suggested by Rocklin *et al.* [204] were restrained over the simulation using a protocol explained in Section 4.7.1.2.

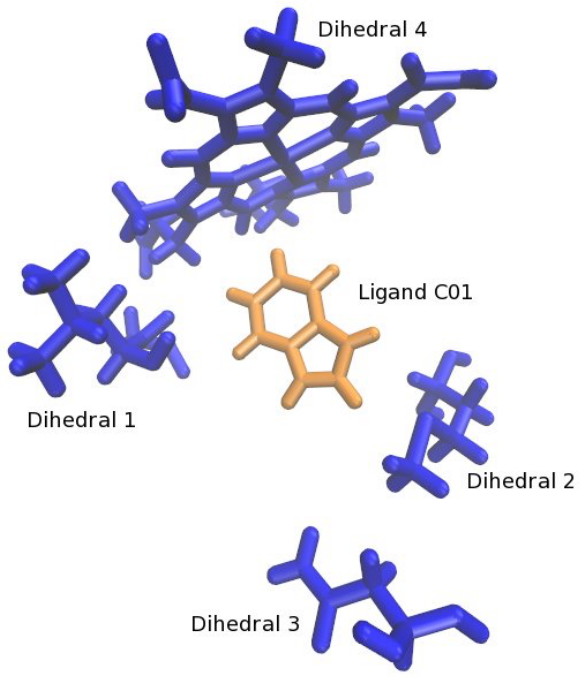


Figure 4.7: Four restrained protein dihedrals positioned at the ligand binding site for complex of C01 ligand: a) Dihedral 1 (Dihedral of Gly175; Leu174 Cα - LEU174C - Gly175 N - Gly175 Cα (160°)) b) Dihedral 2 (Dihedral of Met226; Met226 C- Met226 Cα- Met226 Cβ- Met226 Cγ (160°)) c) Dihedral 3 (Dihedral of not contiguous atom; Leu199 C - Asn200 Cα – Asn200 Cβ - Asn200 Cγ (-50°)) and d) Dihedral 4 (Heme dihedral; Hem290 C2D - Hem290 C3D - Hem290 CAD – Hem290 CBD (90°)).

The positions of each dihedral restraint on atoms in the ligand-binding site are illustrated in Figure 4.7. These dihedral restraints were chosen as the given protein receptor residues had been identified to frequently adopt conformations that had not been observed in the CCP crystal structure.[204] In order to examine appropriate protein dihedral restraints and reference orientations to be restrained in the systems, 200 ns normal MD simulations were performed of both apo and holo protein, starting from the CCP crystal structure, to examine the unrestrained populations of these dihedrals. Figure S5 (Appendix A) shows the dihedral angle populations of each potential dihedral restraint in the protein.

As discussed previously, we performed the confine and release method in our calculations to improve the convergence in conformational sampling for these protein dihedrals.[204] However, there tone of issues corresponding to these protein dihedrals arose for simulating these restraints. In particular, the protein dihedral defined using four non-contiguous atoms; Leu199 C - Asn200 Cα – Asn200 Cβ - Asn200 Cγ (Dihedral 3 in Figure 4.7) has no defined torsional potential. As an alternative to specifying an arbitrary restraint, restraints were instead applied to other dihedrals with motions correlated to this torsion. There were three potential dihedrals corresponding to this torsion (Appendix A, Figure S6a,b and c); i) Dihedral Asn_200C: 200@C- 200@CA- 200@CB- 200@CG, ii) Asn_200N: 200@N- 200@CA- 200@CB- 200@CG and iii) Asn_200CN: 199@C- 200@N- 200@CA- 200@CB. However, dihedral Asn_200C: 200@C- 200@CA- 200@CB- 200@CG was chosen among the other potential dihedral owing to the close correlation between the movement of this dihedral which trigger the movement of the not contiguous atom dihedral examine proposed by Rocklin *et al.* here (corresponding to the histogram distributions in Appendix A, Figure S6d and Figure S6e). A similar issue was observed for the fourth dihedral (Hem290 C2D - Hem290 C3D - Hem290 CAD – Hem290 CBD), which has no force constant value defined in the heme parameters. Therefore this dihedral could not have its force constant scaled down during the replica exchange simulations. Thus, no enhanced sampling was executed on this particular dihedral to evaluate the free energy contribution of this restraint however, the restraints were still applied during the alchemical steps of ligand discharging/decoupling.

The effect of including these protein restraints on absolute binding free energies is reported in Table 4.6. Restraint set 2, which includes ligand and protein restraints, shows a smaller uncertainty in ΔG_{bind} estimations (0.45 kcal mol^{-1}) compared to restraint set 1, which corresponds to the results previously simulated with no restraints on the protein (i.e. Ligand restraint set 2 in Table 4.6). There are also slight differences (less than 0.5 kcal mol^{-1}) in the confine and release protein restraints energies between our estimations (Restraint set 2) and the Rocklin *et al.* predictions. It is probable that this difference arises from a different set of dihedral reference angles assigned although similar dihedrals were restrained.

Table 4.6: The comparison of each component of Δ*G* (kcal mol^{-1}) generated in the simulation with and without protein restraints applied to C01 ligand free energy calculations. Uncertainties calculated as one standard error over three repeats.

Component	**Δ*G* (kcal mol^{-1})**		
	Rocklin *et al.*	**Restraint Set 1[a]**	**Restraint Set 2[b]**
Binding	-7.84 ± 0.00	-9.46 ± 0.52	-9.90 ± 0.45
Confine protein restraints	1.08± 0.00	0.00 ± 0.00	0.77 ± 0.02
Release protein restraints	-0.80 ± 0.00	0.00 ± 0.00	-0.72 ± 0.01

Two set of absolute binding free energy calculations: [a] no protein restraints applied (ligand restraints only) and [b] protein restraints applied (both ligand and protein restraints).

4.9.2.2 Simulation method effect (enhanced sampling method)

Simulation methods also plays an important role in obtaining accurate and precise free energies right. To generate a converged and free conformation sampling in the ΔG_{bind} calculations, an enhanced sampling method was employed in our runs using the optimised simulation protocol (Section 4.7). Effects of the enhanced sampling method are discussed in following section.

4.9.2.2.1 Hamiltonian replica exchange simulation

A Hamiltonian replica exchange (HREX) method, exchanging system configurations between adjacent lambda windows, was introduced in order to sufficiently sample conformational space in the free energy calculations. For CCP systems, a lengthy time was required to generate the converged binding free energies as reported in Rocklin *et al.*[204] To evaluate whether converged binding free energies could be calculated with shorter simulation timescales, we initially performed our simulations using windows of 20 ns length to extensively evaluate all the interactions in our systems but without replica exchange (section 4.7). However, with the Hamiltonian replica exchange applied, our simulations were run for only 6 ns per window, with coordinate exchange trials between neighbouring replicas every 2 ps. An exchange path for each starting replica during the simulation is given in Figure 4.8. Based on this analysis, an appropriate number of replicas were applied to the calculation indicated by good overlaps on the exchange probability between each replica over the whole simulation performed in Table 4.7.

Table 4.7: The exchange probability between each replica over the whole electrostatics and vdW simulations performed for ligand C01.

Simulation	No. of replica	Replica from	Replica to	Exchange probability
Electrostatics	8	1	2	0.18
		2	3	0.09
		3	4	0.20
		4	5	0.19
		5	6	0.37
		6	7	0.35
		7	8	0.16
		8	1	0.00
vdW	16	1	2	0.85
		2	3	0.81
		3	4	0.34
		4	5	0.12
		5	6	0.24
		6	7	0.38
		7	8	0.46
		8	9	0.74
		9	10	0.73
		10	11	0.77
		11	12	0.77
		12	13	0.78
		13	14	0.81
		14	15	0.79
		15	16	0.81
		16	1	0.00

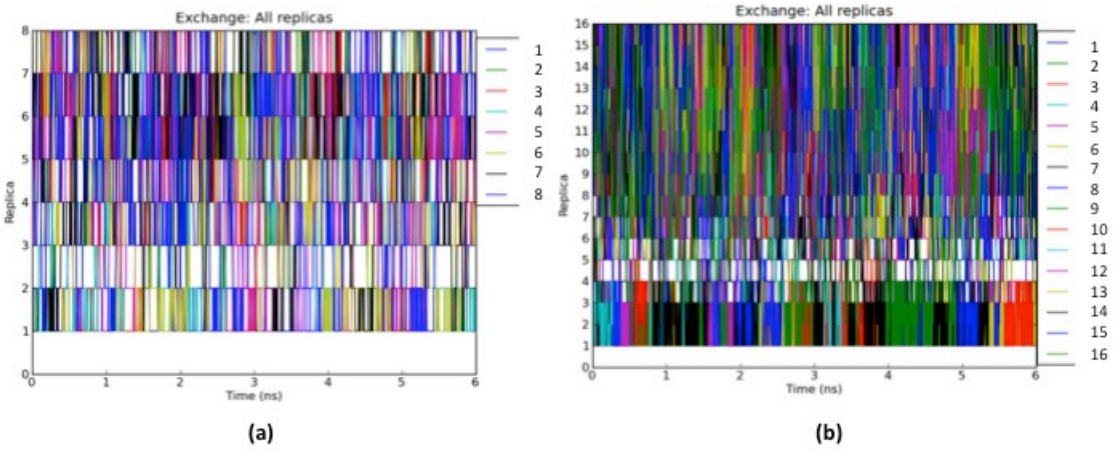


Figure 4.8: The exchange paths for all replicas during Hamiltonian replica exchange simulations for ligand C01. a) Exchange paths of 8 replicas applied during an electrostatics interactions simulation b) Exchange probability with 16 replicas applied for the vdW interactions simulations.

The computed absolute binding free energies, along with each component of the free energies, evaluated for ligand C01 with the HREX method are presented in Table 4.8. The simulation method performance was evaluated by looking at the variance computed in the total ΔG_{bind} calculations and each component of the free energies compared to the simulations without HREX. In the total ΔG_{bind}, a smaller variance was observed (Method Set 2), compared to the previous method (Method Set 1), with only 0.33 kcal mol^{-1} computed with the application of enhanced sampling method, compared to 0.45 kcal mol^{-1} without. Comparing each component of the free energies calculated, there is a slight improvement in estimation of vdW contributions, with a drop in variance from 0.15 kcal mol^{-1} to 0.06 kcal mol^{-1}. However there is no improvement in the overall value of $-\Delta G_{hyd}$ nor its individual components using this enhanced sampling method. Thus, an enhanced sampling method is not crucial for the $-\Delta G_{hyd}$ calculations nor its individual components, however, the HREX calculations method would be more efficient simulation method for total ΔG_{bind} calculations.

Table 4.8: Components of the absolute binding free energy calculations on charged compound C01. Uncertainties calculated as one standard error over three repeats.

Component		**Δ*G* (kcal mol^{-1})**	
	Rocklin *et al.*	**Method Set 1[a]**	**Method Set 2[b]**
In solution			
Uncharged Ligand	43.93 ± 0.00	43.66 ± 0.02	43.71 ± 0.03
PME non-neutral ligand corrections	17.10 ± 0.00	17.33 ± 0.00	17.33 ± 0.00
Decouple ligand VdW	-0.29 ± 0.00	-0.10 ± 0.01	-0.13 ± 0.04
In Complex			
Ligand restraints	7.97 ± 0.00	6.51 ± 0.02	6.60 ± 0.03
Protein restraints	0.28 ± 0.00	-0.03 ± 0.03	-0.03 ± 0.03
Couple ligand vdW	-6.90 ± 0.00	-7.81 ± 0.15	-5.48 ± 0.06
Charge ligand	-53.39 ± 0.00	-53.58 ± 0.45	-55.57 ± 0.46
PME non-neutral ligand corrections	-16.13 ± 0.00	-15.07 ± 0.00	-15.07 ± 0.00
Include ligand symmetry	-0.41 ± 0.00	-0.41 ± 0.00	-0.41 ± 0.00
ΔG_{bind}	-7.84 ± 0.00	-9.90 ± 0.45	-9.43 ± 0.33

Two set of absolute binding free energy calculations: [a] no enhanced sampling method applied and [b] an enhanced sampling method applied (Hamiltonian Replica Exchange).

4.9.3 Overall result

4.9.3.1 Free energies of transferring the ligand from the vacuum to solution

The free energies of transferring the ligand in solution to vacuum, -ΔG_{hyd} calculated for GAFF force field using the optimised protocol (section 4.8.1.1) against the published -ΔG_{hyd},[204] are provided in Table 4.9 with the regression plotted in Figure 4.9.

Our calculated GAFF -ΔG_{hyd} using the optimised protocol shown (Figure 4.9) an excellent agreement to the previously published GAFF -ΔG_{hyd} published by Rocklin *et al.*[204] with R^2 = 0.998. Our calculated GAFF -ΔG_{hyd} match up well for most of the ligands tested here, except for the ligand C02 given by a huge differences of 4.44 kcal mol^{-1} to -ΔG_{hyd} published by Rocklin *et al.*[204] Parameters investigation has demonstrated that differences in the -ΔG_{hyd} estimated for this particular ligand is due to the electrostatics parameter assigned between both parameters (The old and updated GAFF force field).

Table 4.9: Comparison of the the free energies of transferring the ligand in solution to vacuum, $-\Delta G_{hyd}$ using the optimised protocol with the GAFF force field to the published free energies of transferring the ligand in solution to vacuum by Rocklin *et al.*[204]

	$-\Delta G_{hyd}$ (kcal mol^{-1})		
Ligand	**Published**[a]	**Calculated**[b]	**Difference calculated to Published**[a]
C01	60.74	60.95 ± 0.09	0.21
C02	65.67	61.23 ± 0.01	4.44
C03	62.60	60.59 ± 0.04	2.01
C04	67.92	66.10 ± 0.04	1.82
C05	57.18	56.76 ± 0.03	0.42
C06	62.12	61.61 ± 0.02	0.51
C07	59.94	59.50 ± 0.07	0.44
C08	59.03	58.63 ± 0.07	0.40
C09	64.26	61.96 ± 0.02	2.30
C010	60.04	58.24 ± 0.04	1.80
C011	55.03	55.45 ± 0.02	0.42
C012	5.44	5.59 ± 0.02	0.15
C013	6.97	5.22 ± 0.05	1.75
C014	9.76	9.98 ± 0.02	0.22

All the published hydration free energies are taken from [a]Rocklin *et al.*[204] All the [b]calculated hydration free energies report one standard error over three repeats.

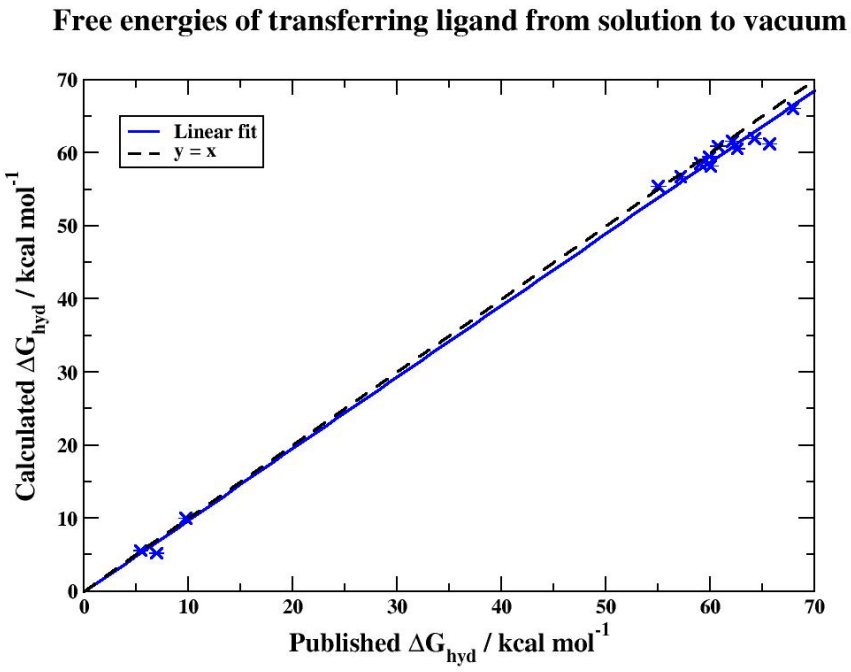


Figure 4.9: Calculated (blue) free energies for transferring ligand from solution to vacuum($-\Delta G_{hyd}$) against published by Rocklin *et al.*[204] free energies for transferring ligand from solution to vacuum($-\Delta G_{hyd}$) . Line of linear fit (blue) and y=x (dashed line). Linear regression plots gives the following equation for the calculated results (y = -0.105 + 0.981 x), R^2 =0.998.

4.9.3.2 Absolute binding free energy

Absolute binding free energies, ΔG_{bind} for the CPP protein using our calculated GAFF parameter with the optimised protocol and previously published[204] GAFF results against to the experimental ΔG_{bind} ere presented in Table 4.10 and regression plotted in Figure 4.10.

A better gradient of the ΔG_{bind} given by our calculated GAFF force field to the experiment with R^2 = 0.562, compared to the published GAFF results R^2 = 0.315 (Figure 4.10) is observed, although no charge corrections have been applied to our calculations as used by Rocklin *et al.* for their estimations. In table 4.10, our calculated ΔG_{bind} data reported an overestimation in ΔG_{bind} compared to the experimental data for most of the ligands, with our predictions showing, a too negative ΔG_{bind} while published data show less negative ΔG_{bind} but only after using scaled charges.

Presumably, the differences in the estimation between our calculated and the published data are due to the difference in protocol implemented for the overall ΔG_{bind} calculations. Here, we only implemented the absolute binding free energy protocol calculations, while the Rocklin protocol are evaluated by combined relative binding free energy calculations and absolute binding free energy calculations to compute the overall ΔG_{bind} . However, their findings suggest that the

discrepancy in their prediction were due to the lack of the explicit polarisation in the binding site during the binding event for the charged ligand resulting in overpolarisation in the electrostatics calculations interactions. Charged scaling, then was applied in their calculation to correct for this artefect. As noted, in our method no charge correction have been applied to the calculations resulting in the larger systematic error observed in our estimations (Table 4.10).

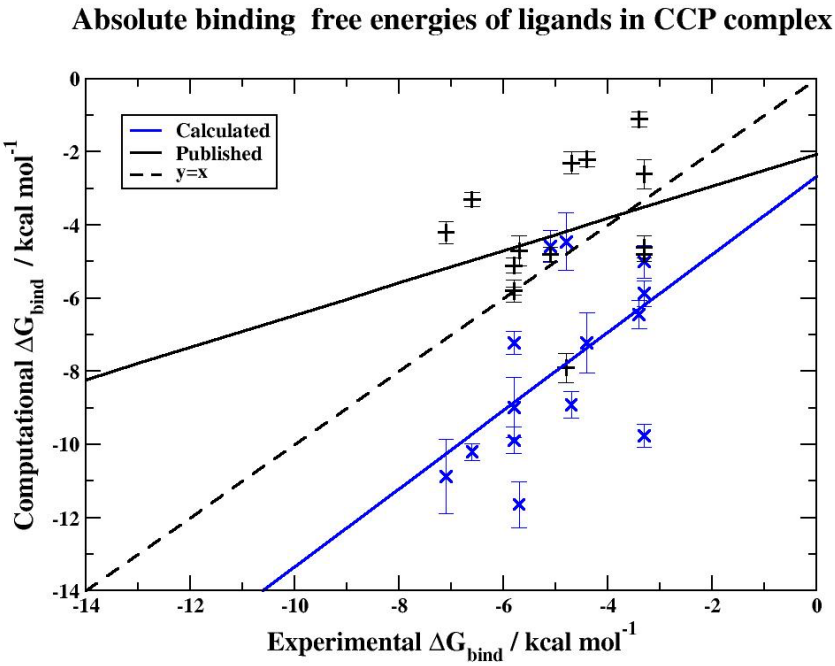


Figure 4.10: Calculated (blue) and previously published Rocklin *et al.* [204](black) computational binding free energies for ligands to the CCP complex against experimental binding free energies. Line of perfect agreement, y=x (dashed line). Linear regression plots give the following equation: a) Calculated (y = -2.657 + 1.068x), R^2 =0.562 b) Published (y = -2.052 + 0.441x), R^2=0.315.

Table 4.10: Absolute binding free energies using optimised protocol with GAFF force field.

	ΔG_{bind} (kcal mol^{-1})			Unsigned error ΔG_{bind} (kcal mol^{-1}) to experiment	
Ligand	**Experimental[a]**	**Published[a]**	**Calculated[b]**	**Published[a]**	**Calculated[b]**
C01	-5.80[c]	-5.80 ± 0.10[d]	-8.98 ± 0.82	0.00	3.18
C02	-5.80 ± 0.20	-5.10 ± 0.20[d]	-7.22 ± 0.32	0.70	1.42
C03	-5.10 ± 0.30	-4.80 ± 0.20[d]	-4.58 ± 0.43	0.30	0.52
C04	-4.40 ± 0.20	-2.20 ± 0.20[d]	-7.22 ± 0.82	2.20	2.82
C05	-3.40 ± 0.40	-1.10 ± 0.20[d]	-6.45 ± 0.39	2.30	3.05
C06	-7.10 ± 0.20	-4.20 ± 0.30	-10.88 ± 1.02	2.90	3.78
C07	-6.60 ± 0.20	-3.30 ± 0.20	-10.20 ± 0.23	3.30	3.60
C08	-5.80 ± 0.20	-5.80 ± 0.30	-9.89 ± 0.36	0.00	4.09
C09	-5.70 ± 0.20	-4.70 ± 0.40	-11.64 ± 0.63	1.00	5.94
C010	-4.80 ± 0.20	-7.90 ± 0.40	-4.45 ± 0.79	3.10	0.35
C011	-4.70 ± 0.20	-2.30 ± 0.30	-8.91 ± 0.36	2.40	4.21
C012	>-3.3	-4.60 ± 0.30	-5.87 ± 0.35	1.30	2.57
C013	>-3.3	-4.80 ± 0.20	-9.76 ± 0.32	1.50	6.46
C014	>-3.3	-2.60 ± 0.40	-5.00 ± 0.44	0.70	1.70

All the experimental binding free energies are taken from [a]Rocklin *et al.*[204] except [c]Rosenfeld *et al.*[260] All the published computational binding free energies are taken from [a]Rocklin *et al.* with the scaled charge of 0.986 except [d]published binding free energies with scaled charge of 0.981. [b]All the calculated binding free energies are without charge scaling. Calculated results report one standard error over 3 repeats.

4.10 Conclusion

Evaluation of parameter and methodology sensitivity is crucial in order to obtain a robust and reproducible protocol for free energy calculations. In these tests, a detailed investigation has been performed to generate both sensible parameters and optimised protocols with a fixed-charge force field, to inform the binding free energy calculations with the Amoeba force field in Chapter 6. For the electrostatics part of the calculation: i) Protein protonation states with a net charge -1 on the receptor would be the ideal charge state for this system. ii) For the ligand parameters, having small differences in parameter may result in substantial effects in the overall calculations. A good parameter is essential in generating high accuracy results in free energy calculation: i) Additional steps in the thermodynamic cycle to confine and then release suitable ligand and protein restraints are crucial for obtained converged results in binding free energy calculations ii) The alchemical binding free energy steps in the complex were improved by using a HREX method for sufficient conformational sampling in the simulations. Overall, the evaluation of the free energy of transferring the ligand from the solution to vacuum with the optimised protocol gives very consistent and reproducible result, but slightly overestimated the binding free energy evaluation. This suggested that, this might be due to the lack of representation of explicit polarisation for these systems, that perhaps can be corrected by using the polarisable force field such as AMOEBA. Ultimately, with all the issues addressed here and appropriate parameters and methodology suggested, we now have a validate protocol to use with the AMOEBA force field. Although, this chapter has resulted in an optimised method proposed for the AMOEBA calculations, the heme group parameterisation would next need to be considered, since there are no AMOEBA parameters available for the heme group. The detailed parameterisation of heme will be discussed in the next chapter.

Chapter 5: Assessment of AMOEBA polarisable force field heme parameters

5.1 Introduction

For the calculations of the AMOEBA ligand-binding interactions, suitable parameters for the full protein complex are required. While there are defined AMOEBA parameterisation protocols for protein, solvent, ions (Na^+ and Cl^-) and small molecules (ligands),[140,154,219] generating the heme group parameters is more challenging. Heme is not only a large ligand but it is also directly coordinated to a part of the protein through a histidine residue. Additionally, the heme group in cytochrome C peroxidase also has a transition metal (Fe (III)) at the centre, which is potentially highly polarisable and certainly highly polarising as a large charge centre. Thus, having correct heme group parameters and interactions with the surroundings is an essential but complex and extremely challenging part of setting up the AMOEBA system. In this chapter we will discuss the parameterisation and the assessment of AMOEBA heme group parameter sets for the cytochrome C peroxidase protein complex.

5.2 Parameterisation

Initially, a literature search for available and relevant AMOEBA heme group parameters was performed. In April 2018, de la Lande and coworkers published AMOEBA parameters of the heme cofactor in both its ferric (Fe(III)) and ferrous (Fe(II)) forms.[263] However, at the time of starting this study, suitable heme group parameters for our system did not exist. Thus, we developed an in-house set of AMOEBA heme group parameters, including both bonded and non-bonded interactions, consistent with the parameterisation methodology of the rest of the AMOEBA force field (chapter 3, section 3.4). Bonded interactions refer to the valence parameters including bond, angle, stretch-bend, out-of-plane and torsion terms, and was parameterised using reference data (equilibrium values and force constants). The non-bonded interactions comprise the electrostatic parameters (atomic multipoles, polarisabilities and damping coefficients) and vdW parameters (radii and well depths).

Mainly, the parameterisation for the heme group was divided into two parts: i) the porphyrin ring parameters, and ii) the central iron atom parameters (Figure 5.1). The detailed parameter derivation for both parts of the heme group will be discussed in the sections below.

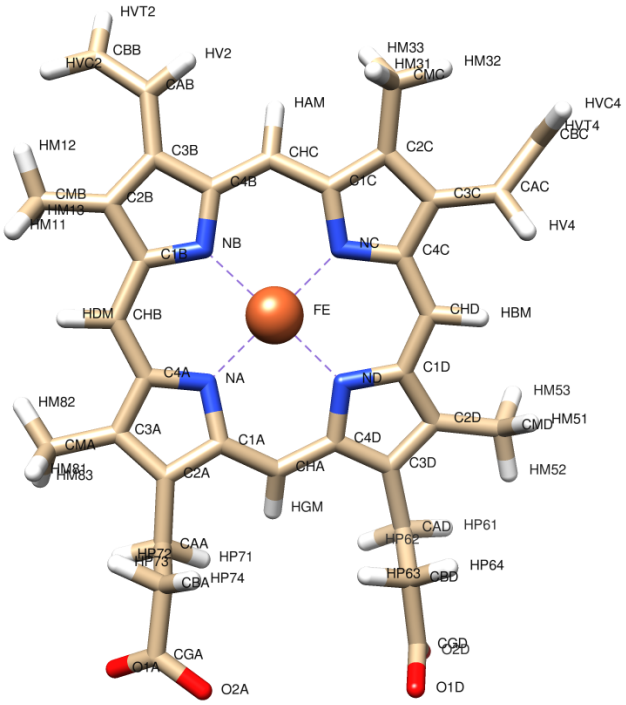


Figure 5.1: The structure of the cytochrome c peroxidase heme group comprised of porphyrin ring and ferric iron in the centre, labelled with the atom names.

5.2.1 Porphyrin ring parameters

The parameterisation of the porphyrin ring for the heme group was performed using a manual parameterisation (chapter 3, section 3.4),[224] with slight modifications to the standard AMOEBA parameterisation protocol, using the TINKER7.1 package[226] and GAUSSIAN09[264] for generating the bonded and non-bonded parameters. Modifications to the standard parameterisation protocol are due to the problem encountered and explained in section 5.3.1.1, owing to the size (large ligand) and complexity (metal ligand) of the heme structure.

Atomic multipole parameters of the porphyrin ring were derived from QM calculations performed with GAUSSIAN09[264] with the multipoles obtained by the Stone's GDMA Distributed Multipole Analysis followed by electrostatic potential refinement using the *potential* utility in the TINKER7.1 package.[226] The heme with a coordinating histidine sidechain (modelled as imidazole) and a water molecule was employed as the initial structure (Figure 5.2) for the QM calculations. The initial structure was first optimised in gas phase at the HF/6-31G* level with a net charge -1 and high-spin multiplicity of 6. During this optimisation step, two rotable dihedrals (totalling 8 carbon atoms) of the heme group were frozen (Figure 5.3). This was followed by another optimisation in implicit solvent using the continuum solvent model PCM (Polarisable Continuum Model) at the HF/6-311G(1d, 1p) level, with both dihedrals now freely rotatable (unfrozen). A single-point

energy calculation at the MP2/6-311G(1d,1p) level in gas phase was then performed, for the DMA analysis and generation of initial point multipoles. Finally, a single point calculation of electrostatic potential outside the molecular volume (ESP) in gas phase was carried out with a larger basis set, MP2/6-311++G(2d, 2p).

The multipole coordinate frames, polarisation groups and polarisibilities were manually defined with the *poledit* program in TINKER7.1,[226] while the valence parameters were assigned using the *valence* program in the same package. The valence parameters generated were then manually refined, according to the suggested parameters available in the TINKER amoeba09.prm[226] and amoebapro13.prm[226] force fields for similar atom types.

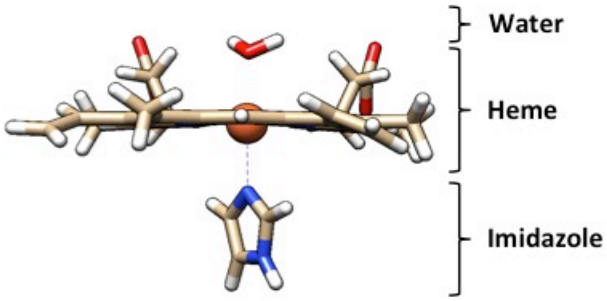


Figure 5.2: The initial structure (heme with coordinating histidine, modelled as imidazole, and a water molecule) employed for QM calculation using GAUSSIAN09.[264]

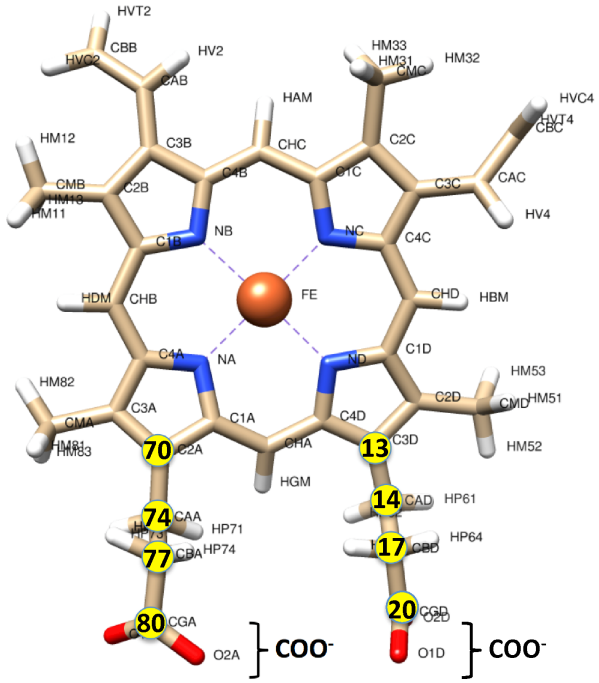


Figure 5.3: The structure of the cytochrome c peroxidase heme group, labelled with the atom names. Two rotable dihedrals (carboxylic acid substituents of heme group) with 8 carbon atoms were frozen during optimisation step and are denoted by yellow circles numbered with atom number.

5.2.2 Iron parameter

Although the catalytically active ground state of CCP is high spin Fe(III),[265] Fe(II) heme parameters however are also of interest to test their stability in this system, and because they may be transferable to other heme complexes. Here, the ferric ion parameters in the centre of heme group were adopted from the closely available non-bonded ferric parameters[266,267] There were two sets of iron parameters that have been published relevant to our work i) Fe(II) parameters[267] and ii) Fe (III) parameters[266] (Table 5.1). Both iron states were parameterised and tested in slightly different systems. To determine the most sensible and stable iron parameters for our systems: a) physical heme group geometries and b) simulation stabilities, for both competing parameters were explored.

Table 5.1: Two sets of AMOEBA non-bonded parameters for Fe(II)[267] and Fe(III) iron.[266] R^0 is vdW radius, ε^0 is well depth, α is polarisibility and a is the Thole damping factor.

Iron	Spin multiplicity	R^0 (Å)	ε^0 (kcal mol^{-1})	α (Å^3)	a
Fe (II)	Quintet	2.798	0.390	0.550	0.113
Fe (III)	Sextet	2.066	0.833	0.258	0.052

5.2.2.1 Fe (II) parameters

The Fe (II) non-bonded parameters, (electrostatic parameters (polarisabilities and damping coefficients) and vdW parameters (radii and well depths)) in the quintet state were taken from the paper published by David *et al.*[267] (Table 5.1). Although in this paper,[267] Fe (II) ion parameters were parameterised in three spin states, singlet, triplet and quintet, the high-spin quintet was chosen as the most suitable state for our system. The quintet spin state of Fe^{2+} alone was parameterised with six coordinated water molecules in the octahedral $[Fe(H_2O)_6]^{2+}$ complex.

Polarisability for all Fe^{2+} spin states was defined as α = 0.550 Å^3. However, an optimised damping factor of, a = 0.113[267] was proposed for Fe^{2+} rather than the general damping factor a = 0.390 used elsewhere in AMOEBA for atoms in molecules and singly charged ions. In fact, this proposed damping factor, a = 0.113[267] is consistent to the damping factor for other cations Zn^{2+} and Cu^{2+} (a = 0.16), recently published.[268] The terms governing vdW parameters (R^0 , ε^0) were defined as R^0 = 2.789 Å^2, ε^0 = 0.390 in the quintet state and optimised geometry.

All the parameters of Fe^{2+} extracted here, have been validated by examining the geometry relaxation of hexaaquo structures in gas phase.[267] Besides this, normal MD simulations in water also have been performed to assess the structural and the energetics properties against the available experimental data.[267]

5.2.2.2 Fe (III) parameters

The non-bonded parameters of Fe(III) were obtained from a more recent published paper Xia *et al.*[266] Again, three sets of parameters at different spin states (doublet, quartet and sextet) were generated. Again, the high-spin sextet spin state parameters were selected to be consistent with the multiplicity defined for our systems, were applied to generate the Fe (III) iron parameters. Here, a similar geometric structure to that employed for Fe^{2+} parameterisation was used, with six coordinated water molecules in an $[Fe(H_2O)_6]^{3+}$ complex. This geometry is the most stable and sensible structure to explore the nature for generating the optimised AMOEBA Fe(III) iron parameters.

The optimised non-bonded parameters of Fe (III) in the sextet spin state are given in Table 5.1. Different polarisibility values were calculated for each spin state for Fe^{3+} with $\alpha = 0.258$ Å^3 assigned for the sextet spin state. There are slightly smaller values for the damping factor of Fe^3 compared to the Fe^{2+} damping factor parameter $^+$, with damping defined by a = 0.052. This smaller damping factor is more likely to lead to overpolarisation (potential for suffering a polarisation catastrophe). vdW parameters of the repulsion- dispersion are given by $R^0 = 2.066$ Å^2, $\varepsilon^0 = 0.833$.

These optimal AMOEBA force field parameters of Fe^{3+}, have been validated by evaluating the structural and energetic properties of the optimised geometry of $[Fe(H_2O)_6]^{3+}$.[266] However here, the Fe^{3+} parameters were assessed in solution phase by evaluating the hydration of both Ferric ion, and in the integrated Fe^{3+}-porphine complex.[266]

5.3 Results and discussion

5.3.1 Parameterisation

So far, two sets of AMOEBA heme group parameters generated with iron parameters, for Fe(II) and Fe(III) respectively, had been extracted from literature sources for all non-bonded interaction parameters, except the atomics multipoles (permanent electrostatics parameter). Here, the bonded interactions refer to the valence parameters including bond, angle, stretch-bend, out-of-plane and torsion terms, and were parameterised by using reference data (equilibrium values and force constants). In this section the development of porphyrin ring parameters is discussed, along with the assessment of the Fe(II) and Fe(III) heme cofactors as a whole.

5.3.1.1 Porphyrin ring parameters

Generating the AMOEBA porphyrin ring parameters, particularly the electrostatic parameters, was not as straightforward as generating the parameters for ligands or small molecules owing to the size of the heme group. However, the complexity of the porphyrin ring itself, particularly the two interacting negatively-charged carboxylic acid chains, was the biggest challenge in the parameterisation process during the computationally expensive QM calculations. To obtain appropriate parameters for the porphyrin ring that represent the correct heme group geometry in a protein complex, electrostatic parameters were derived from QM calculations incorporating models of all Fe-coordinating ligands in the protein complex. That is, an axial water molecule and axial imidazole (representing the histidine sidechain in the CCP crystal structure) were included in all calculations. The multiplicity of 6 was used for the geometry optimisation in GAUSSIAN09.[269] The multiplicity of 6, which is the highest spin states, was chosen based on the surrounding environment. In theory, this would be the most reliable spin state, which should give the most

stable states with the lowest energies.[270,271] Smaller basis sets were used for the first optimisation (HF/6-31G*) in the gas phase (Figure 5.4), followed by the second optimisation with the larger basis set (HF/6-311G(1d, 1p) in solution phase, to achieve the optimised geometry for DMA (Figure 5.5).

A problem was encountered during the first gas-phase geometry optimisation, as the carboxylic acid substituents of the heme group both rotated such that their negatively-charged COO^- groups interacted with the water molecule coordinating the central iron (Figure 5.4). This occurred for both sets of iron parameters and is likely thanks to the gas phase environment. As a result, dihedral angles in the carboxylic acid chains were constrained and the optimisations re-run (Figure 5.5). However, for the second optimisation, both restrained dihedrals were released and the optimisation performed in polarisable implicit solvent instead of gas phase to avoid this favourable interaction (Figure 5.6). Following this, a single point energy calculation with larger basis set and at the MP2 level was performed in gas phase for the DMA, and subsequently calculation of the ESP outside the molecule was also performed in gas phase. The size of the fully coordinated heme group made MP2/aug-cc-pVTZ (used in the standard AMOEBA protocol described in chapter 3, section 3.4) intractable for the ESP calculations. Hence these calculations were carried out with a slightly smaller basis set (MP2/6-311++G(2d, 2p)) but larger than that for the DMA single point calculation. All QM calculations were performed in GAUSSIAN09[264] (Figure 5.7).

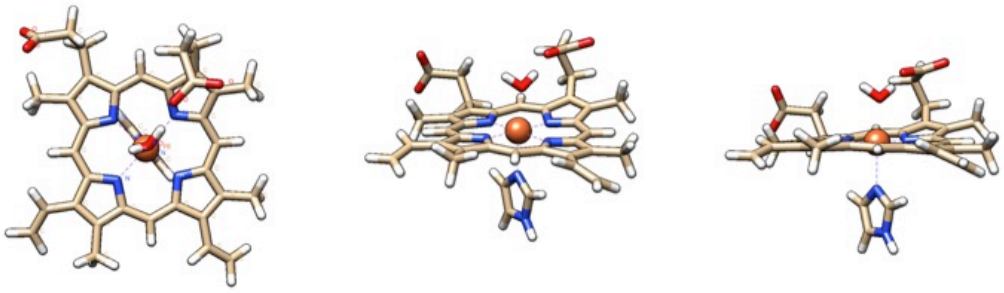


Figure 5.4: The structures of the heme group after undergoing the HF/6-311G (1d,1p) in gas phase with all atoms unrestrained. The geometry of the ring is clearly distorted, caused by the negative charge of a carboxylate group flipping to interact with the water molecule that is coordinated to the ferric ion.

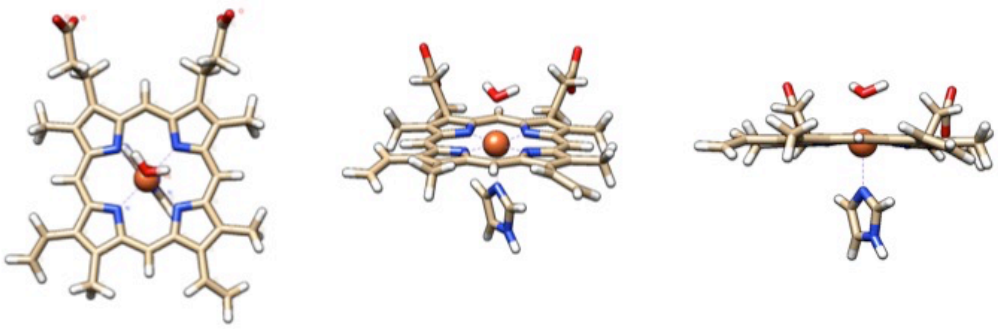


Figure 5.5: The structures of the heme group after undergoing the optimisation at HF/6-31G* in gas phase with the both rotated dihedral restrained. The heme group showed the correct flat geometry structure with the water and imidazole molecules remaining coordinated to the central iron.

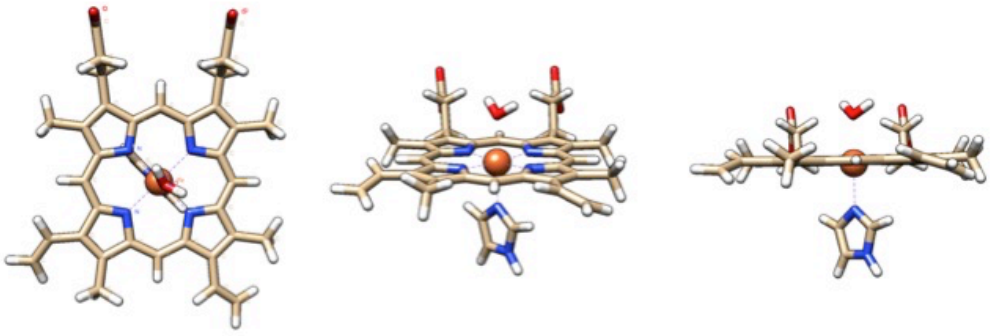


Figure 5.6: The structure of the heme group after undergoing optimisation with HF/6-311G (1d,1p) in implict solvent using the continuum solvent model PCM with both rotated dihedral (labelled with O) freely rotatable. A flat geometry of the ring is still observed with the correct coordination of a water and imidazole molecule.

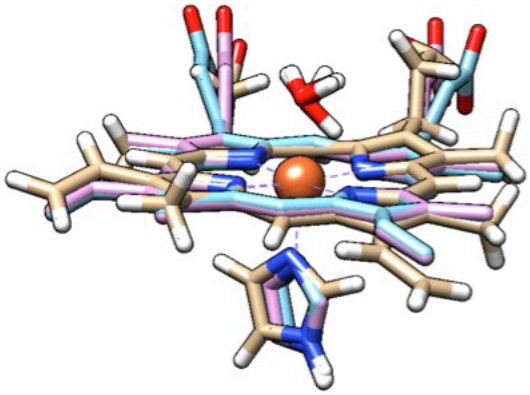


Figure 5.7: The superimposed geometries of the heme group at different stages of optimisation. The initial structure (brown), an optimised structure after gas phase (blue) and the final structure after implicit solvent optimisation and DMA calculation (pink).

5.3.1.2 Iron parameter

As stated above, there are two alternative parameter sets for Fe (II) and Fe (III) respectively that we wished to investigate in this study. Each had advantages and disadvantages to its proposed use. Here, heme groups with Fe (II) parameters taken from Semrouni *et al.*,[267] appears to have been well parameterised for a solution-phase ion and for recreating an octahedral geometry. However, these Fe (II) parameters are not the relevant charge state for the biological system and are not tested in the protein environment. On the other hand, we also generated heme group parameter with the Fe(III) parameters, obtained from the recent published parameter of Fe(III).[266] This set of parameters are the initial charge state for the electron transport cycle, parameterised in a porphine ring environment. However, their performance with additional ligands and/or in a protein environment is unknown. Thus, to explore the most sensible parameters in the protein systems, both sets of parameters were tested to examine how they performed in our protein systems.

5.3.2 Validation of heme group parameters

Both sets of Fe parameters were combined with those of the porphyrin ring and first assessed by examining the geometry of the heme group during minimisation in gas phase, solution and in protein. Additionally, MD simulations of the heme group in the CCP protein were performed to compare the structural and dynamical properties for both parameter sets.

5.3.2.1 Heme geometry in gas phase, solution and protein

For each set of iron parameters, the initial structure of the heme group with coordinated imidazole and water molecule was first minimised in the gas phase using the full set of AMOEBA parameters. Both parameter sets ended up with the planar geometry of the heme group shown in Figure 5.8. Fe and coordinating water oxygen (Fe-O), Fe and coordinating nitrogen in the porphrin ring (Fe-N (porphrin ring)) and Fe and coordinating nitrogen in imidazole (Fe-N (imidazole)) did not appreciably change from that observed in the QM geometry (Table 5.2).

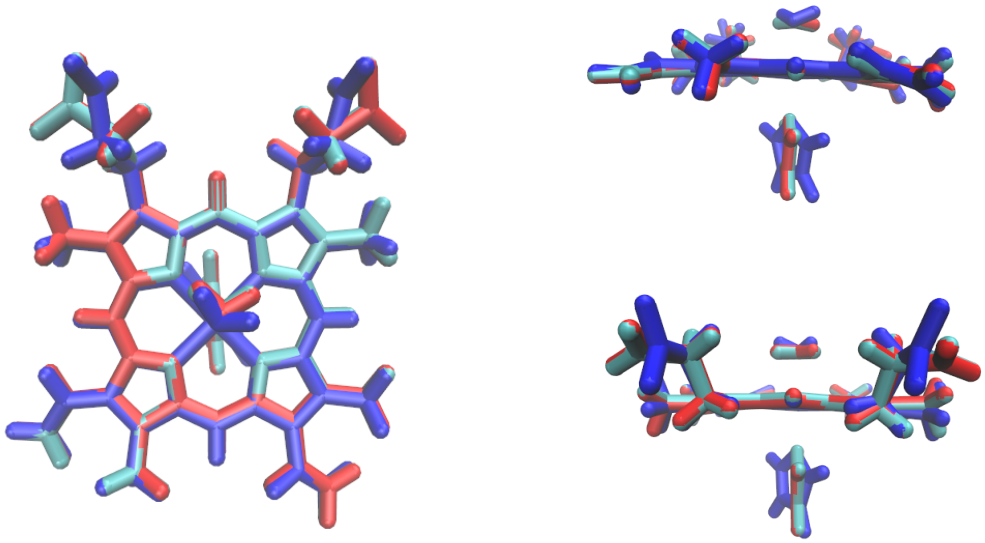


Figure 5.8: AMOEBA Fe(II) (cyan) and Fe(III) (red) heme structure after minimisation in gas overlaid with the reference structure generated from QM calculation during parameterisation (blue). The AMOEBA Fe(II) (RMSD = 0.632) and Fe(III) (RMSD = 0.606) structures overlay very well with that of QM after minimisation in gas phase, with only minor differences in orientation of the Fe (III) heme structure.

Table 5.2: Comparison of distances between the iron and coordinating water oxygen (Fe-O), coordinating nitrogen in the porphyrin ring (Fe-N (porphyrin ring)) and coordinating nitrogen in imidazole (Fe-N (imidazole)). AMOEBA final structures after minimisation in gas using Fe (II) and Fe (III) parameters are compared with the QM optimised structure in gas.

	Distance (Å)		
Parameter	**Fe-O**	**Fe-N (porphrin ring)**	**Fe-N (imidazole)**
QM	2.27	2.06	2.20
Fe (II)	2.09	2.07	2.22
Fe (III)	2.12	2.08	2.22

The assessment of the effects of parameters on geometry continued with the minimisation of the heme group in solution phase to observe the response of both parameter sets in the presence of a large number of AMOEBA water molecules. Here, the initial heme structure (identical to that employed in the gas phase) was soaked in a cubic box of water of side length 24.662 Å, containing 500 molecules of AMOEBA water14,[272]applying the heme group parameter of Fe (II) and Fe(III), respectively. Again, a planar structure with a flat geometry of the heme group (Figure 5.9) was observed, with only slightly modified Fe-O distances at the end of the minimisation (Table 5.3).

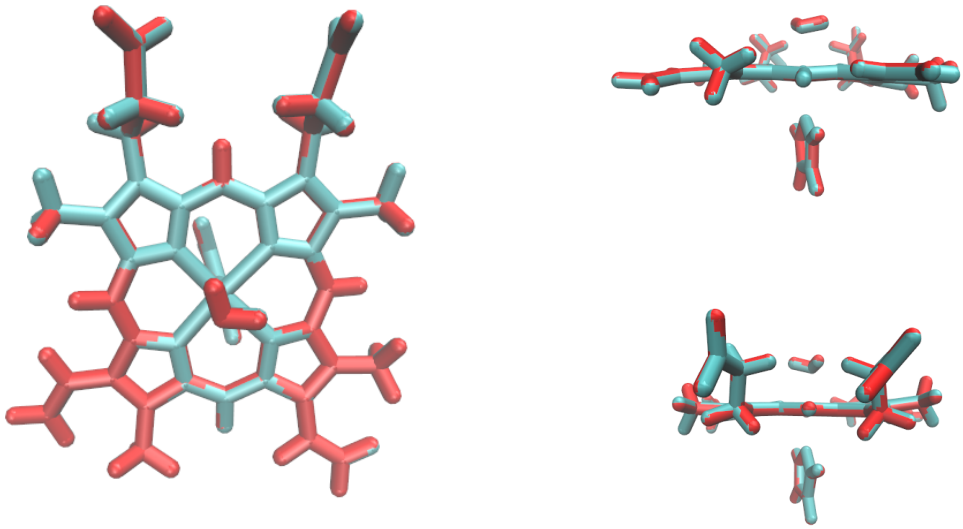


Figure 5.9: The geometry of AMOEBA heme group structures after the minimisation in solution (RMSD = 0.021). Structure using Fe(II) parameters shown in cyan and Fe (III) parameters shown in red.

Finally, minimisations of the heme groups in the fully solvated CCP protein complex crystal structure (PDB ID: 4JM8) were performed with both sets of parameters. Instead of using the heme group with coordinating imidazole and water, the imidazole was replaced with the sidechain of the coordinating histidine (His175 singly protonated at the δ-nitrogen) in the full protein environment. The standard AMOEBA histidine and water14 parameters were used for the coordinating residues. Both Fe(II) and Fe(III) parameter sets resulted in suitable flat structures, shown in Figure 5.10, after the minimisation in the protein. The tests were extended by running short MD simulations. The systems were slowly heated to 300 K in the NVT ensemble for 50 ps, followed by 100 ps pressure equilibration to 1 atm using NPT at 300 K. A timestep of 2 fs and a Langevin integrator[184] was applied to the simulations to maintain temperature. A Berendsen barostat[235] was employed to maintain the pressure of the systems. Again, during the simulation, both parameters shown a correct geometry as demonstrated in Figure 5.10.

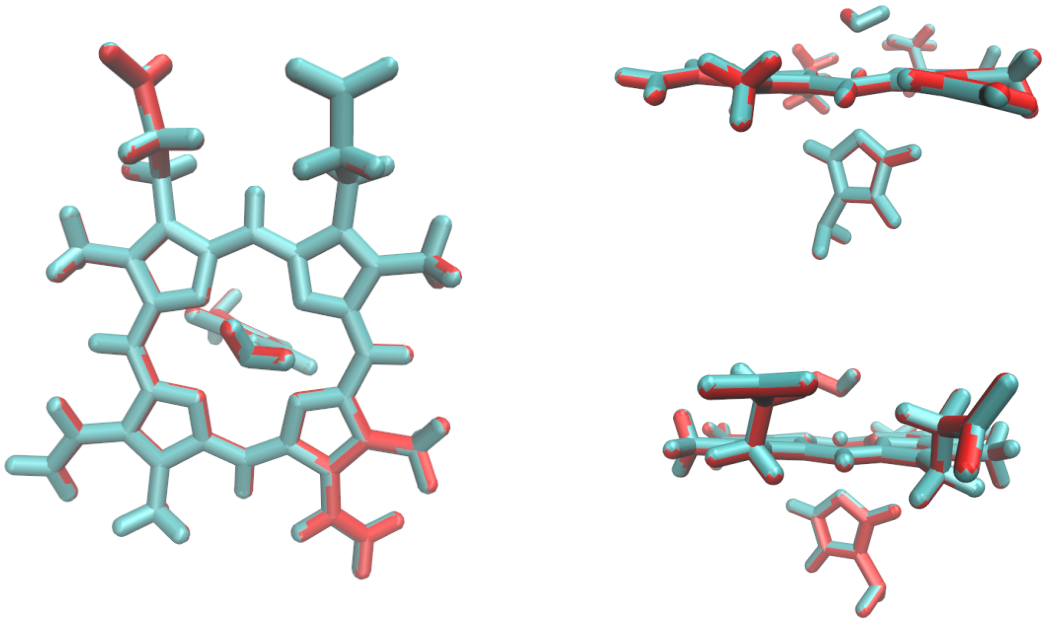


Figure 5.10: The geometry of AMOEBA heme group structures after the minimisation in protein complex (RMSD = 0.014). Structure using Fe (II) parameters shown in cyan and Fe (III) parameters shown in red.

Further NVT simulations were carried out for 150 ps for both parameters. However the simulation with Fe (III) was unsuccessful to complete. Simulation of the heme group with Fe (III) parameters encountered a 'polarisation catastrophe'-style error, in which the water molecule coordinating the Fe (III) rotated such that its H atoms also appeared to interact closely with the Fe, resulting in runaway dipole convergence. This situation caused instability in our system and clearly was not the correct geometry for this system (Figure 5.11).

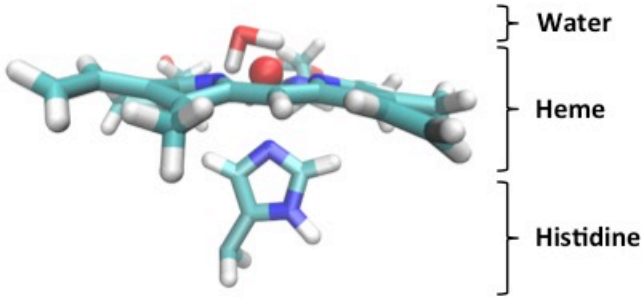


Figure 5.11: Representative geometry of the heme group during MD simulation of the full protein complex using Fe(III) parameters. The coordinating water molecule reorients itself towards the iron in the centre of the heme group. This unphysical geometry caused instability in the protein complex simulation, but was not observed in the simulations with Fe(II) parameters.

Fixing the water distance by directly bonding the water to Fe (III) in heme group would be one of the solutions to this problem, as the Fe, O and H atoms would then have their electrostatic interactions excluded from one another (as they share either 1-2 or 1-3 interactions). Although direct bonding of Fe to water might be helpful, this is not ideal because the water should be free to exchange with other solvent molecules during the simulation. This also precludes specific parameterisation of the coordinated water as a separate ligand because it may exchange with bulk solvent throughout the simulation.

Table 5.3: Distances between the iron and oxygen of the axial coordinating water in the final structures after minimisation (in gas, water and protein) or simulation (in protein) using Fe (II) and Fe (III) parameters.

	Parameter	
Simulation	**Fe(II)-O distance/ Å**	**Fe(III)-O distance/ Å**
Minimisation in gas	2.09	2.12
Minimisation in solution	2.02	2.09
Minimisation in protein	1.90	1.90
MD simulation in protein (50 ps NVT, 100 ps NVT)	1.79	1.66
Further MD simulation in protein (150 ps NVT)	1.81	-

5.4 Conclusion

Consequently, the only iron parameters for heme group that worked in our system use the Fe(II) parameters of Semrouni *et al.*[267] Based on the tests performed, running the simulations with the Fe(III) parameters of Xia *et al.*[266] is impossible as MD simulations in cytochrome C peroxidase resulted in unphysical heme group geometries and simulation instabilities, despite a parameterisation process that included optimisation in a porphine ring environment.

Ideally, the heme group itself should have been parameterised inside the protein environment and with the coordinated water molecule and histidine attached. However, the size of the basis sets traditionally used for the ESP fit in the last stage of AMOEBA electrostatic parameter refinement is large, to incorporate diffuse terms and pick up the effects of intramolecular

polarisation. This means that the traditional AMOEBA parameter generation method cannot incorporate the protein environment beyond the directly coordinated ligands.

Despite the lack of biological relevance of Fe(II) as a stage in the CCP electron transfer cycle, the tests performed in the protein complex have shown that sensible heme geometry, coordination distances and dynamic stability can result from simulations with Fe(II) parameters. Given the parameters of vdW and polarisibility have not been reoptimized from those of Fe(II) but with +3 charge, this inspires confidence that the overall Fe(II) parameter set is robust to its environment.

In the CCP crystal structures complexed (e.g. PDB ID: 4JM8) with small molecules there is no direct short range interaction between the ligand and iron of the heme group. Therefore the predominant effect of the heme group on binding free energies of the small molecule ligands is likely to result from effects through the surrounding protein and long range electrostatic effects. Thus, getting a stable geometry of the heme group is likely more important for the AMOEBA free energy calculations than exact refinement of the iron vdW and polarisibility parameters. Ultimately, the heme group parameters with: a) porphyrin ring parameters developed here, and b) Fe(II) iron parameters of Semrouni *et al.* will be employed in the protein-ligand binding free energy calculations with the AMOEBA force field explained in the next chapter.

Chapter 6: Evaluation of Protein-Ligand Binding Free Energies of cytochrome c peroxidase with AMOEBA polarisable force field

6.1 Introduction

In this chapter, evaluation of the AMOEBA polarisable force field performance will be extended to more complex systems by calculating the binding free energies of charged (cationic) and neutral ligands to cytochrome c peroxidase protein. Ideally, this protein-ligand binding interaction system was chosen to represent a high-field environment, where it is more feasible polarisation may be required, especially for the binding interactions involving charged ligands. As pointed out in Chapter 4, Rocklin et al.[204] have reported a systematic discrepancy of the fixed-pointcharged- force field in estimating accurate binding free energies to experiment. This was blamed on the absence of an explicit electronic polarisation response in the classical fixed-pointcharge- potential. Thus, by incorporating an explicit response to the environment polarisable force fields may be expected to give more accurate predictions of the interactions in this system.

To understand the extent to which the AMOEBA polarisable force field is able to capture this effect, the binding free energies of cytochrome c peroxidase protein with charged and neutral ligands utilising the AMOEBA force field were compared to the binding free energies generated by the GAFF force field employing the optimised methodology previously discussed in Chapter 4. In this chapter, the performance of the potential energy functions were assessed by validating the calculated binding free energies with the AMOEBA force field against the experimental results. Discussion of the improvements arising from the application of polarisation terms in AMOEBA will also be explained in further detail.

6.2 Data set

Overall, 14 ligands were selected including three neutral ligands (Figure 6.1). All 14 ligands had their hydration free energies calculated in solution and seven (six charged and one neutral ligand) ligands had their binding free energies calculated in the protein complex, with both AMOEBA and GAFF force fields. Although a smaller data set was tested for AMOEBA in the complex, this choice of ligands however covered a good range of chemistry available in this study. As noted in Chapter 4, section 4.2, this selection of ligands was chosen based on their availability of experimental

binding data and the quality of X-ray crystallographic structures of cytochrome c peroxidase with bound ligand. All had high-resolution crystal structures (< 2 Å resolution) obtained from the Protein Databank (PDB), as shown in Figure 4.2. (Chapter 4, Section 4.2).

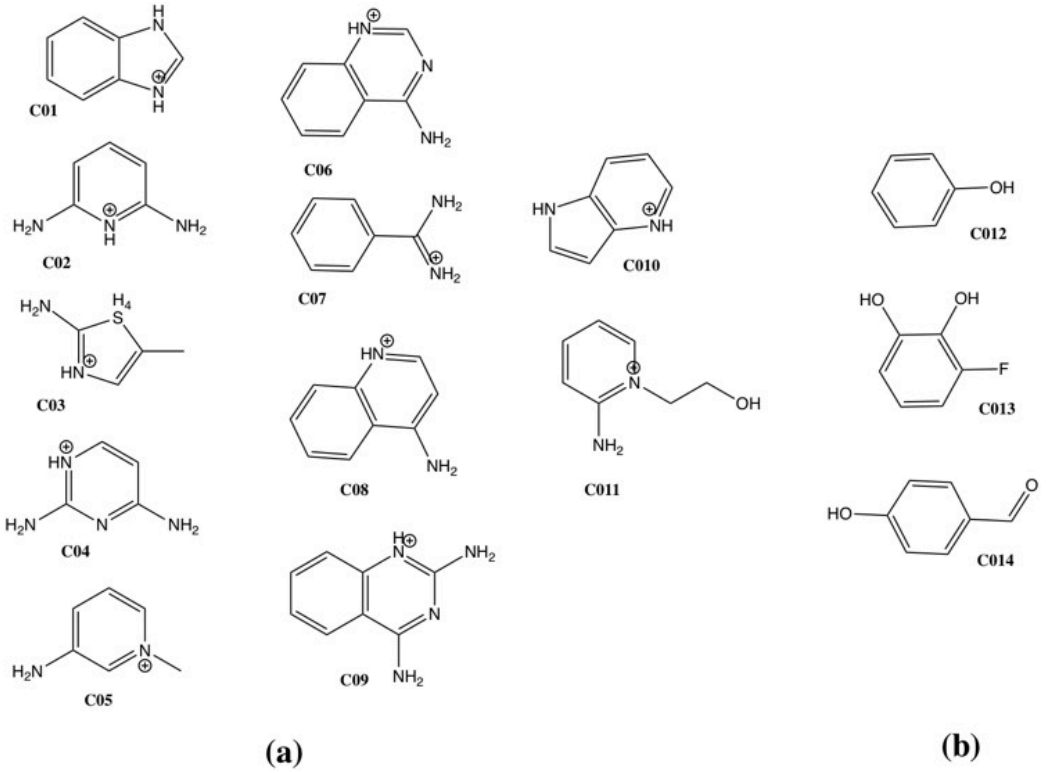


Figure 6.1: The structures of selected ligands in this study. a) Data set of charged ligands, b) Data set of neutral ligands. All 14 ligands from both data sets selected for free energy calculations in solution and only seven ligands (C01, C02, C03, C04, C06, C07 and C012) selected for the free energy calculations in protein complex

6.3 Parameterisation

The standard AMOEBA parameterisation procedure for small molecules, as described in Chapter 3, Section 3.3, was carried out for all ligands using TINKER 7.0[226] and GAUSSIAN09 programs.[227] The 2013 set of AMOEBA parameters was used to model the protein, while vdW, bonds, angles, stretchbends, torsions, and atomic polarisibility- parameters of the ligands and counterions (Cl^- and Na^+) were taken from the 2009 AMOEBA parameter set .[219] For heme, in-house parameters developed in the previous chapter (Chapter 5) was employed. Here the 2014 water model parameters were used to model the water molecules in the simulations[272]. For fixed-point-charge simulations the parameters described in Chapter 4 were utilised. For consistency, the protein

protonation states with net charge -1 were implemented for AMOEBA similarly to the GAFF system setup previously discussed in Chapter 4, section 4.4.

6.4 Free energies calculations

The absolute binding free energies for the ligands in cytochrome c peroxidase were calculated by adopting an identical absolute binding free energy calculation protocol to that described in Chapter 4, section 4.5 for both AMOEBA and GAFF force field. However in this study, all the corrections in binding free energy calculation will not be evaluated and corrected in the ΔG_{bind}, as they are not directly transferable to non-additive potentials. According to Chapter 4, section 4.5, to compute the binding free energies, a series of MD simulations is required to perturb the ligand in bulk water and in the complex according to the thermodynamic cycle shown in Figure 4.3. Therefore two sets of free energy calculations were involved, to calculate: i) In solution, $-\Delta G_{hyd}$, and ii) In complex, $\Delta G_{complex}$. Finally, the total binding free energy for the ligands in CCP protein were calculated as in equation 4.3.

The simulations for the GAFF fixed-point-charged force field were run using the optimised protocol with the simulation details as explained in Chapter 4, section 4.8. AMOEBA simulations were run consistently to GAFF simulations except with a shorter simulation time (2ns for simulations in solution and 1ns for simulations in complex), owing to the added computational cost of the polarisable potential in AMOEBA. The computational details for AMOEBA were as follows.

6.5 Simulation Details

AMOEBA simulations for binding free energies calculations were run using AMBER16 modules.[238] Initially, all the system were parameterised as per section 4.3 and prepared using the XYZEDIT utility in TINKER 7.1[226] to solvate each ligand in a cuboid box, filled with amoeba14[161] water model with an appropriate number of counterions Na+ and Cl- added to reach 150 millimolar ion concentration (approximately 28 each of ion Na+ and Cl-, depending on the exact box size). The initial structures and AMOEBA parameters in TINKER format were then converted to AMBER format for the following minimisation, equilibration and production simulations.

The systems were first minimised using the steepest descent algorithm for 2500 steps. The systems were then heated slowly to 300 K in the NVT ensemble for 50 ps, followed by 100 ps pressure equilibration to 1 atm using NPT at 300 K. A timestep of 2 fs and a Langevin

integrator/thermostat[184] was applied to the simulations to maintain temperature. A Berendsen barostat[235] was employed to maintain the pressure of the system. Finally, these equilibrated structures were used as the initial structure in three independent repeat simulations for the following series of free energy calculations.

The van der Waals cutoff was set to 9 Å with an isotropic LRC (Long Range Correction). Long-range electrostatics for all the systems was treated using Particle Mesh Ewald (PME) summation, with a real-space cutoff of 8 Å. The PME calculation used fifth order B-spline interpolation. The induced dipoles were iterated until the root-mean square change was below 0.001 D per atom. This lower convergence criterion was used to speed up the simulations. However a tighter convergence criterion of 0.00001 D per atom was then used to post-process each MD trajectory and recalculate system potential energies during the calculation of free energies. The energy differences between each intermediate states were saved every 2 ps and the first 200 ps were discarded as equilibration.

Simulations were performed for 2 ns for all ligand perturbations in solution while simulations of 1 ns length were performed for the ligand perturbations in complex. Each AMOEBA free energy calculation in solution and complex were simulated using the same lambda window spacing as the GAFF simulations (Chapter 4, section 4.8).

6.6 Result and discussion

The case of CCP binding free energies had been reported as a failure of fixed-point-charge potentials. In order to obtain a measure of how accurately AMOEBA could estimate the free energies by incorporating explicit polarisation, the performance of AMOEBA and GAFF force fields in calculating free energies of transferring the ligand from solution to vacuum, $-\Delta G_{hyd}$ and binding free energies of the ligands in CCP complex will be presented and discussed. Three independent replicates were run to check for the consistency of free energies computed between each of the runs. Errors were calculated as the standard error (SE) in the mean between these three replicate results.

6.6.1 Free energies of transferring the ligand from the vacuum to solution

Free energies of transferring the ligand from solution to vacuum, $-\Delta G_{hyd}$, are calculated for AMOEBA and GAFF force fields, compared to the GAFF published $-\Delta G_{hyd}$,[204] and shown in Table 6.1. Figure 6.2 shows the linear regression of both datasets to the published GAFF results. AMOEBA hydration free energies for the same set of ligands had been previously calculated by the Ren group,[273] using a different free energy methodology and different parameters. Therefore

calculated AMOEBA -ΔG_{hyd} were also compared to published AMOEBA -ΔG_{hyd},[273] provided in Table 6.2 and with regression plotted in Figure 6.3.

Both AMOEBA and GAFF force field shown an excellent correlation to the previously published -ΔG_{hyd}[204] with the GAFF fixed-point-charged force field, with R^2 = 0.98 (AMOEBA) and R^2 = 0.99 (GAFF) respectively (Figure 6.2). Neutral ligand -ΔG_{hyd} were particularly well-recreated, however a less positive free energy was obtained for all the charged ligands (Table 6.1), with the calculated -ΔG_{hyd} for AMOEBA significantly overestimated by approximately 7 kcal mol^{-1}. To understand why the -ΔG_{hyd} were overestimated to this significant amount, we then compared our AMOEBA -ΔG_{hyd} results by evaluating the difference of our calculated -ΔG_{hyd} results to the previously published -ΔG_{hyd} results performed by Abella *et al.* These results are shown in Table 6.2 and Figure 6.3. Again, we observed that both calculated ΔG_{hyd} with the AMOEBA force fields give an excellent correlation with R^2 = 0.99. In this case, our calculated -ΔG_{hyd} was still overestimated compared to the published AMOEBA results.[273] Predominantly, this is due to the differences of water parameters employed to model the solvated systems (water box) for -ΔG_{hyd} calculations. Abella *et al.* in their works were using Amoeba 2003, water03.prm[274] to model their solvated systems while, AMOEBA 2014, water14.prm[272] was used in our solvated systems. Although the implementation of AMOEBA 2014, water14.prm, has been previously tested in the calculations of ΔG_{hyd} for neutral ligands, and provided an overall more accurate estimation of ΔG,[275] this is the first case of AMOEBA 2014, water14.prm parameters being used in the calculations of -ΔG_{hyd} for charged ligands.

As noted above, the -ΔG_{hyd} of neutral ligands, were estimated to be very similar to the -ΔG_{hyd} published with either the AMOEBA or GAFF force field. However, ligand C013 (Table 6.1) showed the largest difference in -ΔG_{hyd} between our AMOEBA and GAFF results calculated across all neutral ligands (6.16 kcal mol^{-1}). We hypothesised this was due to the fluorine parameters designated for this particular ligand. Supporting this idea, the AMOEBA ΔG_{hyd} published by Ren *et al.* for this ligand (C013) was also overestimated compared to GAFF, in fact with a larger energy difference (7.28 kcal mol^{-1}). Low AMOEBA ΔG_{hyd} energies were also observed in SAMPL4 ligands (L28 to L33) with similar structures but chloro substituents, suggesting that the anomalous result here is specifically due to the fluorine functional group associated with the C013 ligand.[275]

Table 6.1: The free energies of transferring the ligand form solution to vacuum,-ΔG_{hyd} with AMOEBA and GAFF force field against the published -ΔG_{hyd} taken from Rocklin *et al.*[204]

	-ΔG_{hyd}(kcal mol^{-1})				
Ligand	**Published**[a]	**Calculated**		**Difference -ΔG_{hyd} calculated to Published**[a]	
	GAFF	**AMOEBA**	**GAFF**	**AMOEBA**	**GAFF**
C01	43.64	36.52 ± 0.01	43.63 ± 0.09	7.12	0.01
C02	48.67	36.19 ± 0.09	43.90 ± 0.01	12.48	4.77
C03	45.47	36.91 ± 0.01	43.26 ± 0.04	8.56	2.21
C04	50.90	39.37 ± 0.08	48.76 ± 0.04	11.53	2.14
C05	40.02	31.34 ± 0.08	39.41 ± 0.03	8.68	0.61
C06	45.10	35.34 ± 0.11	44.29 ± 0.02	9.76	0.81
C07	42.99	38.70 ± 0.10	42.15 ± 0.07	4.29	0.84
C08	42.14	31.72 ± 0.14	41.32 ± 0.07	10.42	0.82
C09	47.30	36.91 ± 0.05	44.66 ± 0.02	10.39	2.64
C010	43.00	34.19 ± 0.21	40.87 ± 0.04	8.81	2.14
C011	38.03	34.34 ± 0.06	38.12 ± 0.02	3.69	0.09
C012	5.44	5.37 ± 0.06	5.59 ± 0.05	0.07	0.15
C013	6.97	13.13 ± 0.05	5.22 ± 0.10	6.16	1.75
C014	9.76	9.58 ± 0.01	9.98 ± 0.02	0.18	0.22

All the published free energies of transferring the ligand from solution to vacuum,-ΔG_{hyd} taken from [a]Rocklin *et al*.,[204] with no charge corrections applied to the charging energies for the charged ligand. All the calculated free energies of transferring the ligand from solution to vacuum,-ΔG_{hyd}, provided as 1 standard error over 3 repeats.

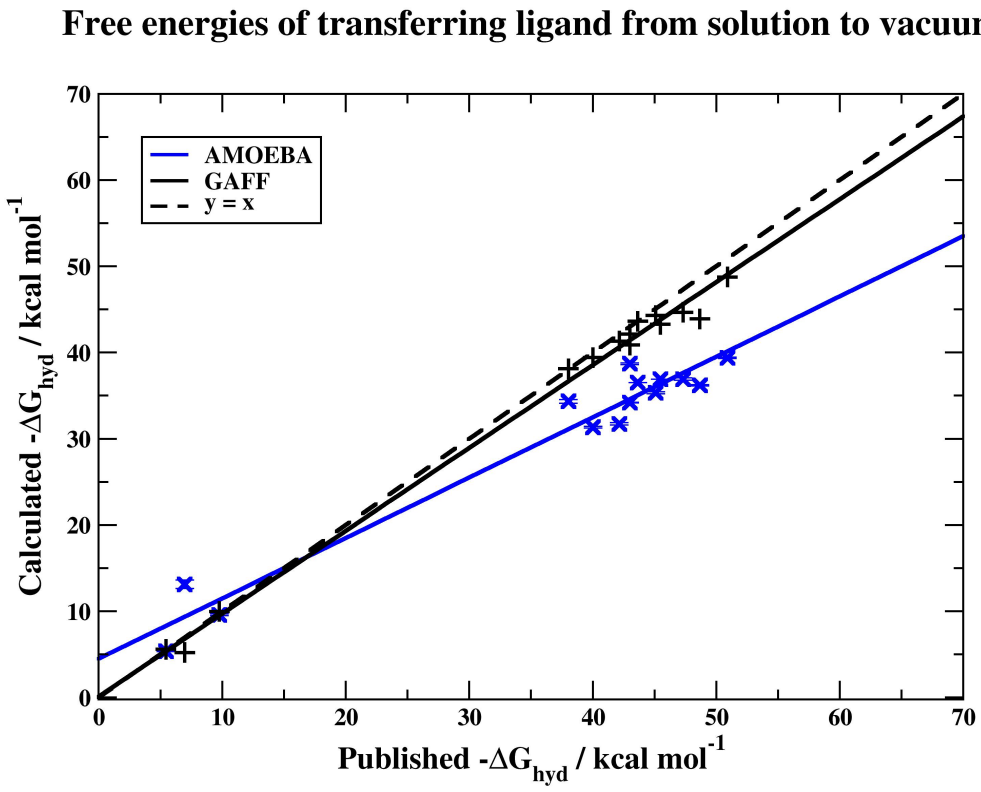


Figure 6.2: The AMOEBA (blue) and GAFF (black) free energies of transferring the ligand from solution to vacuum,-ΔG_{hyd} against the published -ΔG_{hyd} taken from Rocklin *et al.*[204] without corrections applied to the charging free energies for charged ligands. Line of perfect agreement, y = x, denoted by dashed line. Linear regression for each force field plot gives the following equation: a) AMOEBA (y = 0.700 x + 4.487), R^2 = 0.98 b) GAFF (y = 0.961 x + 0.115), R^2 = 0.99.

Table 6.2: Comparison of the free energies of transferring the ligand form solution to vacuum,-ΔG_{hyd} with AMOEBA force field between calculated and the published -ΔG_{hyd} taken from Abella *et al.*[273]

	-ΔG_{hyd}(kcal mol^{-1})		
Ligand	Published[a] AMOEBA	Calculated AMOEBA	Difference -ΔG_{hyd} calculated to Published[a]
C01	46.61 ± 0.82	36.52 ± 0.01	10.09
C02	44.59 ± 0.52	36.19 ± 0.09	8.40
C03	45.76 ± 0.77	36.91 ± 0.01	8.85
C04	49.70 ± 0.83	39.37 ± 0.08	10.33
C05	41.68 ± 0.79	31.34 ± 0.08	10.34
C06	46.36 ± 0.57	35.34 ± 0.11	11.02
C07	48.70 ± 0.57	38.70 ± 0.10	10.00
C08	41.66 ± 0.56	31.72 ± 0.14	9.94
C09	46.22 ± 0.62	36.91 ± 0.05	9.31
C010	44.33 ± 0.51	34.19 ± 0.21	10.14
C011	49.25 ± 0.15	34.34 ± 0.06	14.91
C012	5.38 ± 0.39	5.37 ± 0.06	0.01
C013	14.25 ± 0.32	13.13 ± 0.05	1.12
C014	9.08 ± 0.47	9.58 ± 0.01	0.50

All the published free energies of transferring the ligand from solution to vacuum,-ΔG_{hyd} taken from [a]Abella et al.,[273] All the calculated free energies of transferring the ligand from solution to vacuum,-ΔG_{hyd}, have error bars as 1 standard error over 3 repeats.

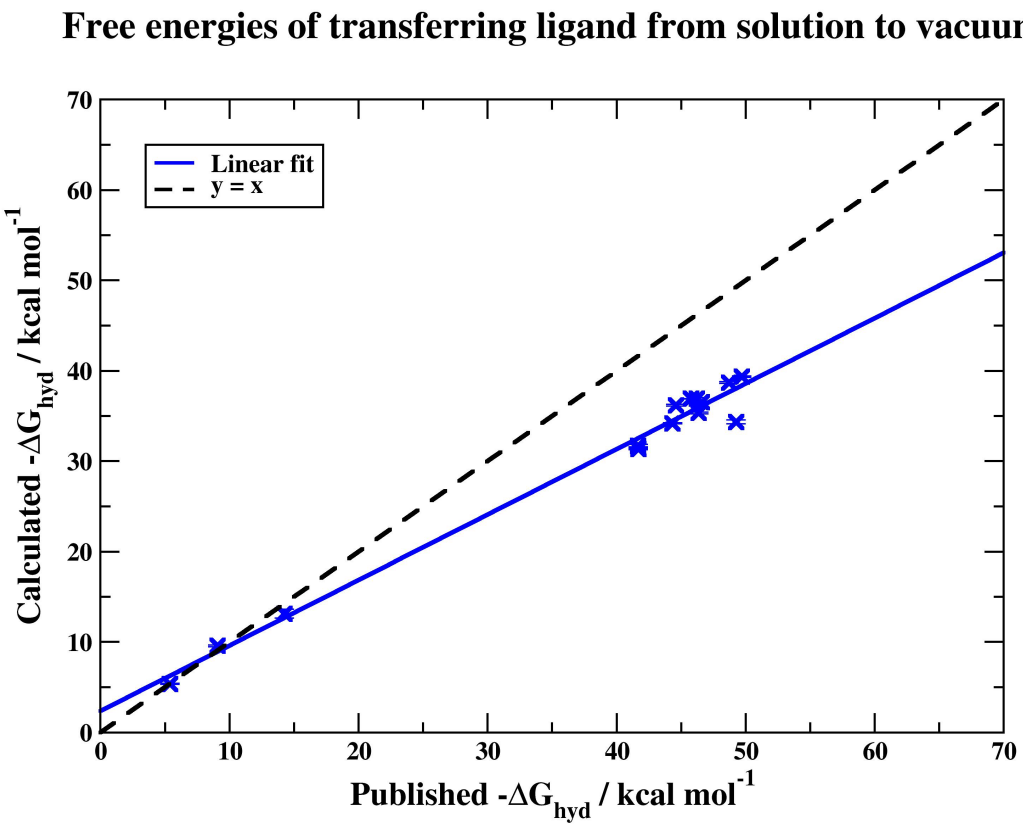


Figure 6.3: The AMOEBA calculated (blue) free energies of transferring the ligand form solution to vacuum,-ΔG_{hyd} against the published -ΔG_{hyd} taken from Abella *et al.*[273] line of perfect agreement, y = x (dashed line). Linear regression for AMOEBA calculated plot gives the following equation: (y = 0.724 x + 2.362), R^2 = 0.99.

6.6.2 Absolute binding free energy

Absolute binding free energies, ΔG_{bind} for seven ligands (6 charged and one neutral) In CCP protein are presented for the AMOEBA and GAFF force field in Table 6.3 and regression to the experimental data[204] plotted in Figure 6.4.

The GAFF force field gives better agreement of the ΔG_{bind} to the experiment with R^2 = 0.90, compared to the AMOEBA force field R^2 = 0.36 (Figure 6.4). However this correlation is limited in meaning considering that only seven ligands were tested here. AMOEBA ΔG_{bind} free energies agree well with the experimental ones for the neutral ligand with only 0.01 kcal mol^{-1} difference (although this may be fortuitous), but were generally overestimated (too negative ΔG) for charged ligands. As previously discussed in section 6.6.1, this is presumably affected by the choice of the water model employed in the system. Given that, this is the first attempt of implementing the AMOEBA 2014, water14 model[161] in a protein system for calculating ΔG_{bind} of ligands (charged and neutral).

Nevertheless, there is no clear systematic difference, positive or negative, between the AMOEBA and GAFF predictions in Table 6.3. Some AMOEBA predictions also have large error bars, and hence large differences in predictions between repeats. We hypothesised the source of this uncertainty could be the shorter simulations with the AMOEBA force field. Interestingly, looking at the free energy contributions of each component evaluated here, the largest variation was provided by large standard errors mostly in the vdW calculations (Figure 6.5). Figure 6.5 represents the convergence in electrostatics and vdW calculations across the length of the simulations. The comparison of the convergence between both electrostatics and vdW calculations show that as the simulation time increased, the simulations tended not to be converged particularly in the vdW coupling steps. Apart from this, we noticed that the exchange probability rates between replicas in the AMOEBA force field were not equivalent to the GAFF force field (the exchange probability rates have previously been evaluated in Chapter 4, section 4.9.2.2.1). Extremely poor exchange probability rates were observed for certain replicas in vdW calculations (Figure 6.6). This is likely to be a substantial contribution to the larger variation to the free energies calculated.

Table 6.3: Comparison of absolute binding free energies, ΔG_{bind} for the charged and neutral ligand in CCP protein with both AMOEBA and GAFF force field against to experimental

ΔG_{bind} (kcal mol^{-1})					
Ligand	**Experiment**[a]	**Calculated**		**Unsigned Error calculated to Experiment**	
		GAFF	**AMOEBA**	**GAFF**	**AMOEBA**
C01	-5.80[c]	-11.22 ± 0.82	-6.86 ± 1.66	5.42	**1.06**
C02	-5.80 ± 0.20	-9.44 ± 0.32	-10.98 ± 1.64	**3.64**	5.18
C03	-5.10 ± 0.20	-6.70 ± 0.43	-11.59 ± 1.13	**1.60**	6.49
C04	-4.40 ± 0.20	-8.83 ± 0.89	-11.32 ± 0.91	**4.43**	6.92
C06	-7.10 ± 0.20	-13.03 ± 1.02	-10.69 ± 2.46	5.93	**3.59**
C07	-6.60 ± 0.20	-12.33 ± 0.23	-6.62 ± 0.50	5.73	**0.02**
C012	> -3.30	-5.87± 0.35	-3.29 ± 0.90	2.57	**0.01**

All the experimental binding free energies are taken from [a]Rocklin *et al.*[204] except [c]Rosenfeld *et al.*[260] All the calculated binding free energies report 1 standard error over 3 repeats.

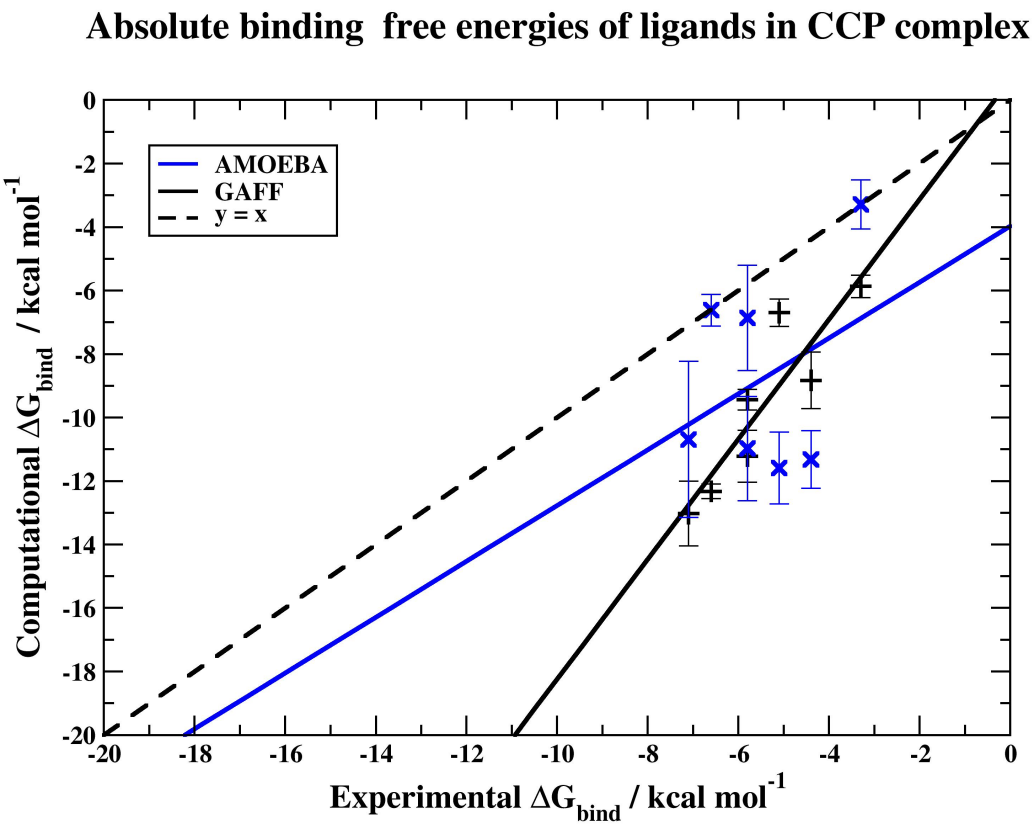


Figure 6.4: The AMOEBA (blue) and GAFF (black) binding free energies for the ligands in CPP protein, ΔG_{bind} against the experimental ΔG_{bind} taken from [a]Rocklin *et al.*[204] and [c]Rosenfeld et al.[260] and line of perfect agreement, y = x (dashed line). Linear regression for each force field plot gives the following equation: a) AMOEBA (y = 0.879 x - 3.976), R^2 = 0.36 b) GAFF (y = 1.890 x + 0.653), R^2 = 0.90.

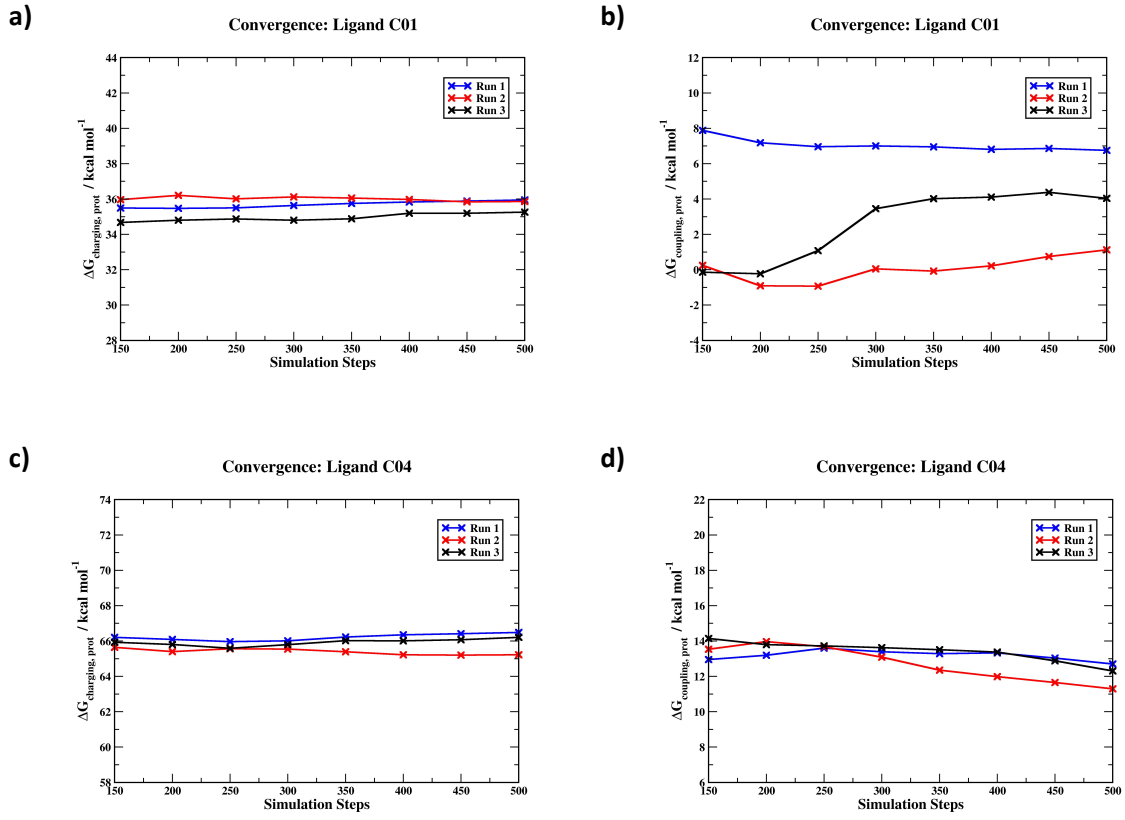


Figure 6.5: The energies convergence of charging and coupling of ligand in CCP protein over the course of the trajectories from 150 to 500 simulation steps where: a and b represented the energies convergence for ligand C01, while c and d represented the energies convergence for ligand C04.

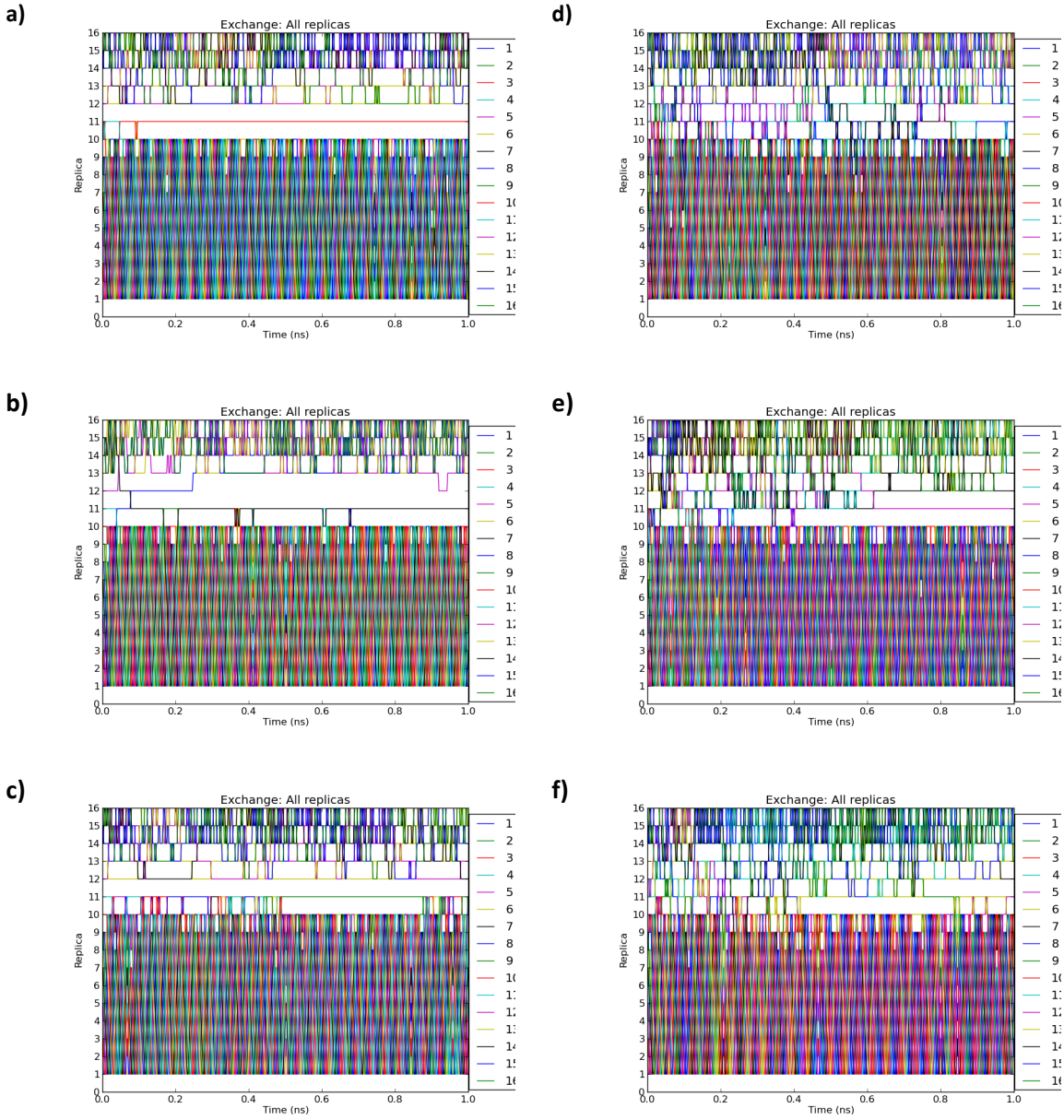


Figure 6.6:The exchange paths during Hamiltonian replica exchange simulations with 16 replicas applied for the vdW interactions simulations with AMOEBA force field where: a, b and c represented the exchange paths for ligand C01, while d,e and f represented the exchange paths for ligand C04.

6.7 Conclusion

Overall, AMOEBA force fields estimated the -ΔG_{hyd} well, consistent to the published Rocklin *et al.*[204] results for neutral ligands, however -ΔG_{hyd} for charged ligands was systematically lower than that predicted with GAFF. Additionally, the ligand with fluorine was an outlier in the AMOEBA -ΔG_{hyd} calculations in their comparison with GAFF. When comparing AMOEBA results here to charged ligands in calculations in Abella *et al.* we note a similar underestimation of -ΔG_{hyd}.[273] In this case, it is therefore predominantly the water model that is having an effect for the -ΔG_{hyd} calculations for the charged ligands. Both AMOEBA -ΔG_{hyd} predictions for the fluoridated ligand match up well, suggesting that the AMOEBA parameters themselves might be responsible here. For the evaluation of $\Delta G_{bind,}$ calculations, AMOEBA force fields accurately estimated a few ligands with the very small unsigned error to the experimental results, although no systematic improvement over GAFF could be determined. Additionally, issues in convergence and exchange probabilities between adjacent replicas have been recognised and associated with the overall performance in AMOEBA ΔG_{bind} estimations. Presumably, by having an optimised AMOEBA ΔG_{bind} calculation protocol including optimum spacing of vdW lambda windows, a consistent and accurate experimental ΔG_{bind} would be recreated. However, such a detailed investigation and derivation of protocols lies beyond the scope and timeline of the initial investigation presented here.

.

Chapter 7: Conclusion

An improvement of the existing force fields to accurately represent the interatomic interactions for description of molecular systems and better correlate dynamics with experimental observations remains a main challenge in force field development and molecular recognition applications. Instead of describing electrostatics as the interactions of fixed, atom-centred, point charges, a better representation of electrostatics by the inclusion of electronic polarisation is one such improvement. The AMOEBA polarisable force field is one of many possible force fields that included a polarisable molecular mechanics model, designed to directly capture the polarisation effect by incorporating an explicit response to the environment. Consequently, an evaluation of potential energy function accuracy is required to determine their performance, given the additional computational cost of incorporating the explicit polarisable potentials.

The aim of this thesis was to determine whether the explicit inclusion of polarisation in the AMOEBA potential energy function was able to give greater accuracy and precision compared to the much simpler and cheaper GAFF potential energy function in free energy calculations. Investigation of two main applications of AMOEBA was performed in order to assess where its successes over existing fixed-charge methods might lie. We conclude with a discussion of various aspects of the results in this study.

7.1 Evaluation of solvation free energies for small molecules with the AMOEBA polarisable force field

Initially, the investigation of the effect of AMOEBA in environments where polarisation might be important was carried out by the evaluation of the solvation free energy calculations of small molecules in a variety of common organic solvents with different dielectric constants. It was anticipated that non-aqueous solvents would have been the environment in which fixed-point-charged force fields may not perform as well as polarisable force fields. However, this did not seem to be the case in the low dielectric environment (low dielectrics constant solvents) and fairly simple solutes tested here. The simplicity of solutes and solvents suggested a simple system would perhaps be better represented by a simple force field rather than application of polarisation. However to completely draw this conclusion, a broader data sets with more challenging solutes and solvents are required. Ideally, further study on more complex systems, focusing particularly on the higher dielectric environments may be more promising as they may require more complex potential forms to be accurate, particularly as additive potentials have not traditionally been parameterised against higher dielectric environments than water.

7.2 Evaluating parameters and methodology in binding free energy calculations of cytochrome c peroxidase

Evaluation of AMOEBA performance was extended to a more challenging problem (protein-ligand binding interactions) by first identifying a clear case of failure in fixed-point-charged potentials. In this case, the real challenge in free energy calculations is the binding free energy calculations for the charged ligands. Here, cytochrome c peroxidase protein was selected as a test system owing to the systematics discrepancy observed in the binding free energies with fixed-point-charged force field. Evaluations on a range of protein-ligand systems were performed to explore the sensitivity of results to parameters and protocols for equivalent later work with AMOEBA. In this study, we addressed a few aspects that are crucial and highly sensitive in the free energy calculations. Molecule parameters are clearly one of the most important aspects that matter a lot in any free energy calculation. Small changes in the parameters can give rise to huge effects in the free energies evaluated. In addition, an optimised protocol with restraints applied to ligand is required for the convergence of free energies in the simulations in a reasonable amount of time. The application of an efficient sampling methodology, such as HREX, is highly important not only for treating the convergence and sampling issues but also very useful to speed up the simulations. However, we observed that although the optimised parameters and protocols have been used for the free energy calculations, the overall binding free energies with the GAFF fixed-point-charge force field were still underestimated (too negative). This suggests that the inclusion of the polarisation description for the interactions of this system might be important here in order to get the free energies right.

7.3 Evaluation of Protein-Ligand Binding Free Energies of cytochrome c peroxidase with the AMOEBA polarisable force field

Further investigations were then carried out to understand to which extent AMOEBA can give more accurate predictions of ΔG_{bind} polarisable force field by incorporating an explicit response to the environment in the system. A performance comparison of AMOEBA and GAFF force field free energies of transferring the ligand from the vacuum to solution, and binding free energies of the ligands in the CCP complex were evaluated and validated against the available published (free energies of transferring the ligand from the vacuum to solution) and experimental data (binding free energies).

In general, AMOEBA force fields perform well in estimation of the free energies of transferring the ligand from the vacuum to solution as shown by the consistency of predictions to the published

Rocklin *et al*.[204] for neutral ligands, however the free energies of transferring the ligand from the vacuum to solution were underestimated (less negative) for charged ligands. Here, the comparison of our AMOEBA calculated free energies of transferring the ligand from the vacuum to solution against the AMOEBA published results performed by Abella *et al*.[273] also showed less negative estimation of charged ligand free energies in our calculations. We hypothesised this was caused by the differences between water models implemented in the studies. We also observed a consistent discrepancy between AMOEBA and GAFF predictions for the neutral ligand with a fluorine functional group in both calculations with the AMOEBA force field. Hence we proposed that AMOEBA parameters might be responsible here.

For the evaluation of binding free energies, the AMOEBA force field accurately estimated a few ligands with very small signed errors to experiment compared to the GAFF force field. However, a few issues must be addressed associated with the overall performance of AMOEBA binding free energy estimations, especially in steps involving the decoupling of vdW interactions. Firstly, many lambda window simulations were not converged as the simulations were run only for 1 ns length. We recognised that the free energies in the simulations were not converged across the simulation steps in vdW decoupling simulation thanks to the huge variation observed for three independent runs. A longer length of simulation time will be required to provide converged free energies in the simulations. Secondly, and relatedly, there were issues regarding the exchange probability rate of replicas across the simulations. The poor exchange rates between the replicas observed in vdW simulations again contributed to the huge variation in the estimated free energies. In this case, the optimised methodology for the GAFF force field calculations does not seem to be directly applicable to the AMOEBA polarisable force field. Therefore further investigation and an optimised protocol particularly for AMOEBA binding free energy calculations are required. It is possible that by having an optimised AMOEBA binding free energy calculation protocol, a consistent and more accurate agreement with experimental binding free energies would be recreated. Ultimately, this will lead to better understanding and improvement of AMOEBA to be fully implemented in biological applications.

Appendix A

Table S1: The restraint atom involved in protein-ligand complex for restraining the position and orientation of ligand in protein for 14 ligands in the CCP systems. The reference atoms in protein the receptor assigned by a, b and c, while the reference atoms in ligand assigned by A, B and C.

	Reference Atom					
Ligand	a	b	c	A	B	C
C01	172@CB	172@CA	172@N	291@C3A	291@C2	291@C2
C02	172@CB	172@CA	172@N	291@C6	291@C3	291@C2
C03	172@CB	172@CA	172@N	291@C5	291@C4	291@C2
C04	172@CB	172@CA	172@N	291@C5	291@C4	291@C2
C05	172@CB	172@CA	172@N	291@C1	291@C5	291@N3
C06	172@CB	172@CA	172@N	291@C4	291@C2	291@C6
C07	172@CB	172@CA	172@N	291@C1	291@C5	291@C3
C08	172@CB	172@CA	172@N	291@C04	291@C09	291@C07
C09	172@CB	172@CA	172@N	291@C4A	291@C4	291@C2
C010	172@CB	172@CA	172@N	291@C03	291@C04	291@C06
C011	172@CB	172@CA	172@N	291@C05	291@C01	291@C03
C012	172@CB	172@CA	172@N	291@C2	291@C1	291@C5
C013	172@CB	172@CA	172@N	291@C6	291@C2	291@C1
C014	172@CB	172@CA	172@N	291@C4	291@C2	291@C1

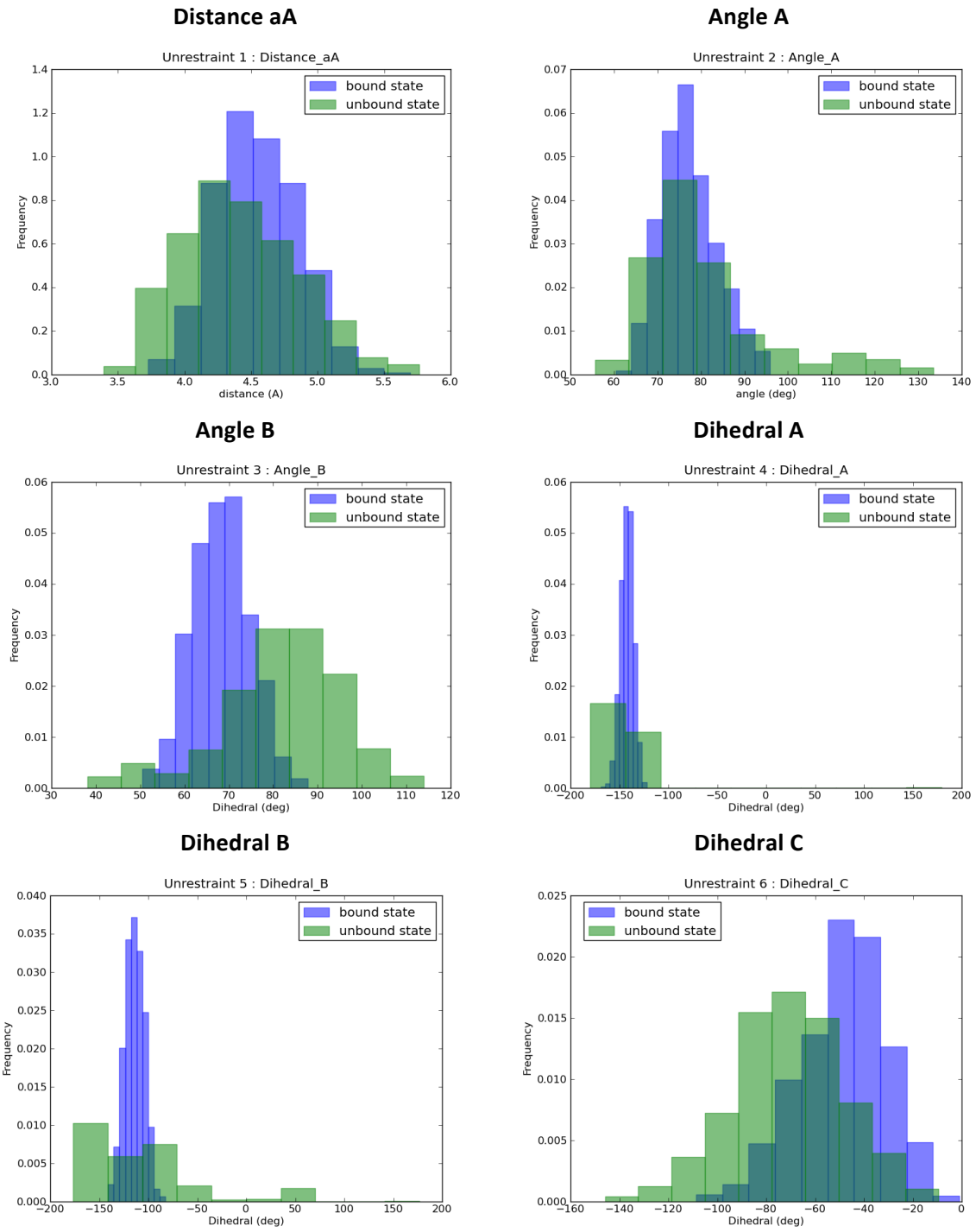


Figure S1: The histograms generated independently from six degree of freedom (one distance, two angles and three dihedrals) from unrestrained simulation trajectories in both bound (blue histogram) and unbound (green histogram) states of C01 structure for reference orientation identification.

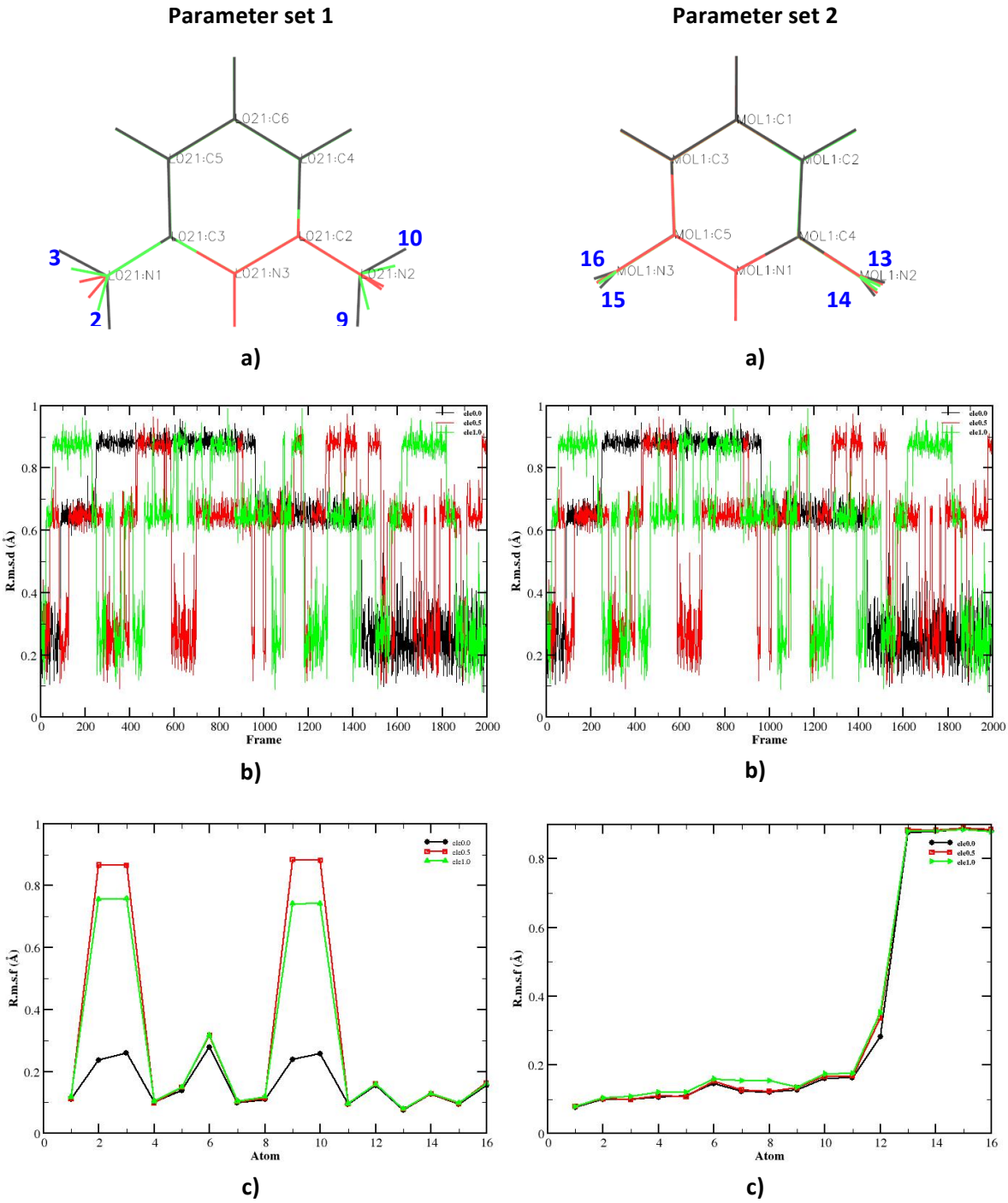
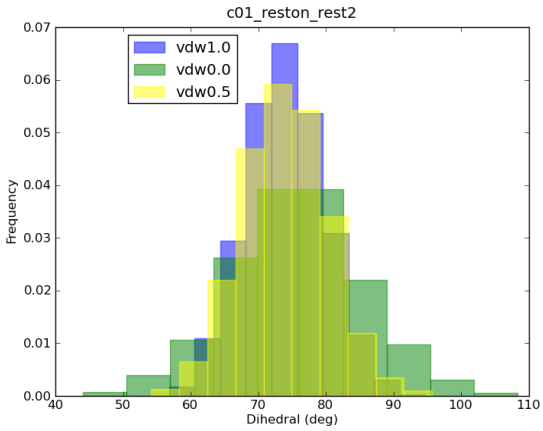
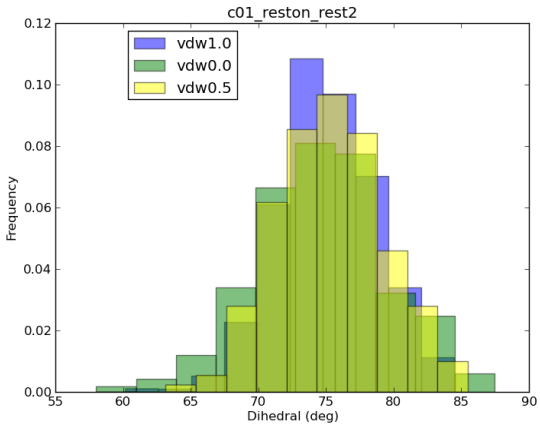


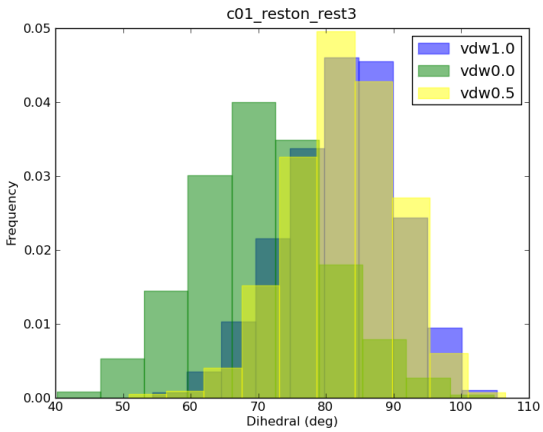
Figure S2: Dynamical differences between the C02 ligand simulated with our parameters, parameter set 1 (left) or those of Rocklin *et al.*,parameter set 2 (right) during the electrostatics interaction simulation. a) Average ligand structure observed during the simulations. b) All-atom ligand RMSD across the simulations. c) Atomic RMSF averaged across the simulations. Results from independent repeats are shown respectively in black, red and green.

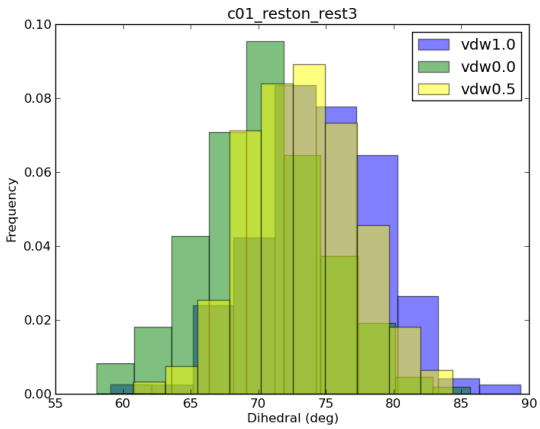
Ligand restraint set 1[a] **Ligand restraint set 2**[b]





Restraint 2: Angle A **Restraint 2: Angle A**

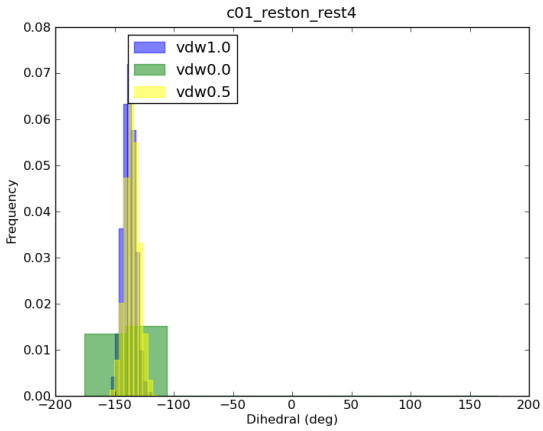


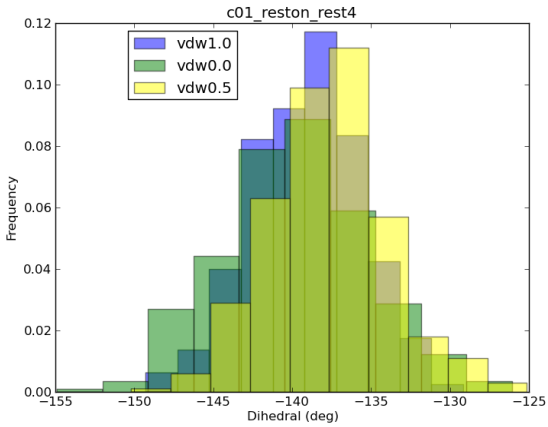


Restraint 3: Angle B **Restraint 3: Angle B**

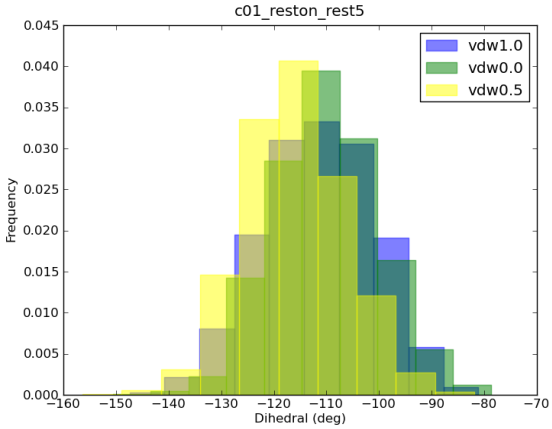
Figure S3: Histogram of angle distributions for angle restraints applied to the ligand C01 in vdW free energy simulations with different force constants: [a]k = 10 kcal mol^{-1} rad^{-2} [b]k 50 kcal mol^{-1} rad^{2}.

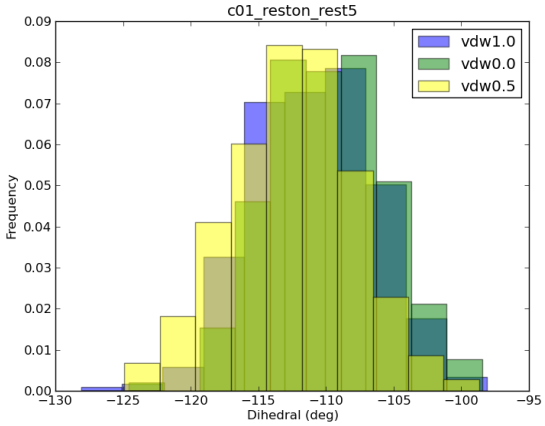
Ligand restraints set1[a] **Ligand restraint set 2**[b]



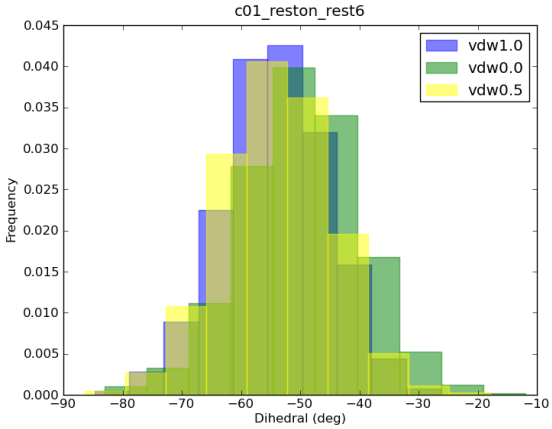


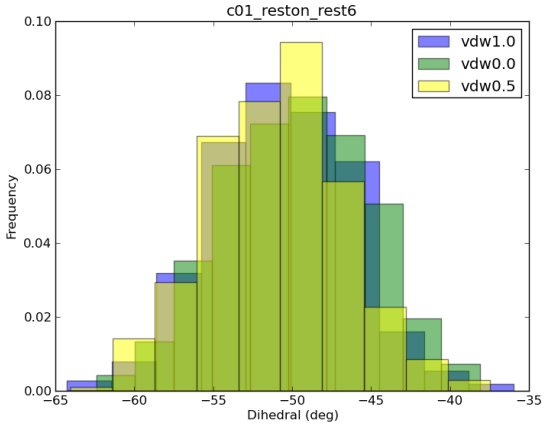
Restraint 4: Dihedral A **Restraint 4: Dihedral A**





Restraint 5: Dihedral B **Restraint 5: Dihedral B**





Restraint 6: Dihedral C **Restraint 5: Dihedral C**

Figure S4: Histogram of angle distributions for dihedral restraints applied to the ligand C01 in vdW free energy simulations with different force constants: [a] k = 10 kcal mol^{-1} and [b] k 50 kcal mol^{-1}.

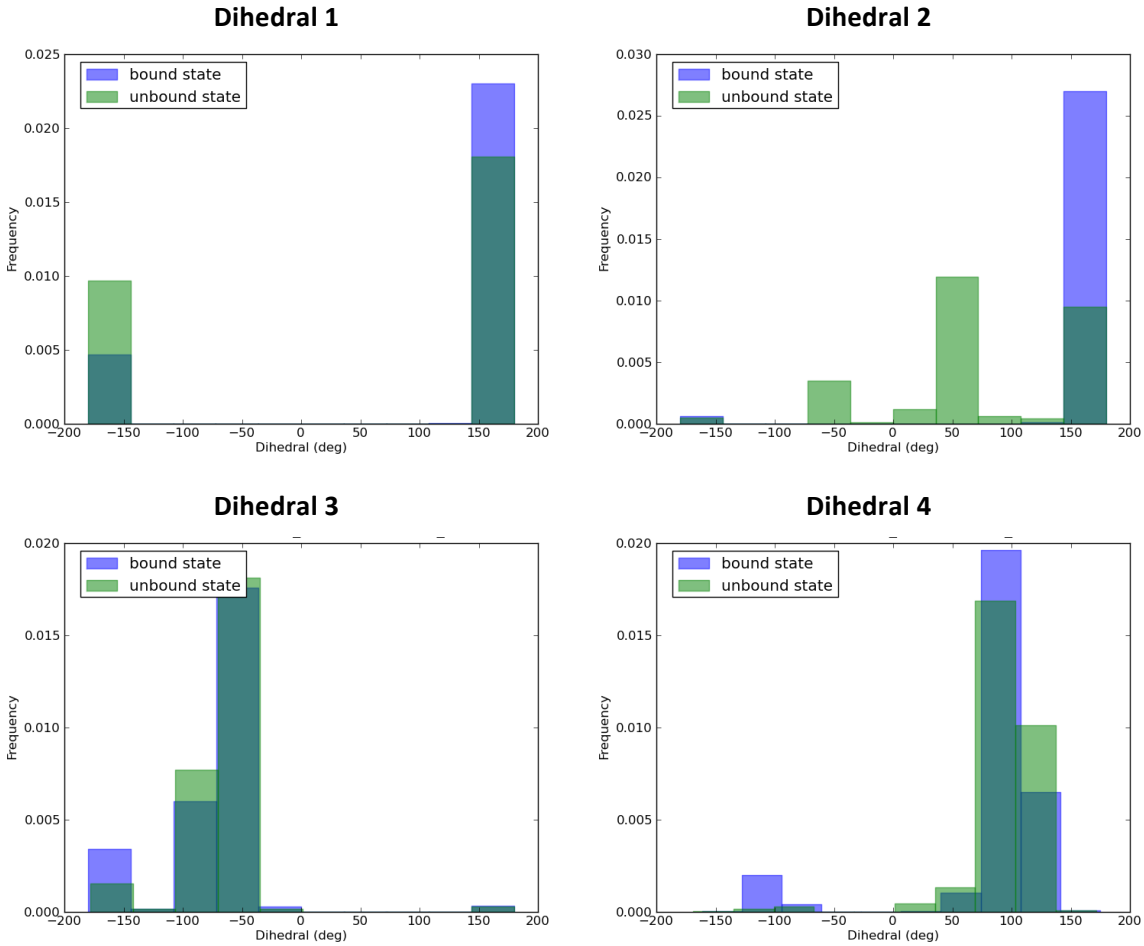


Figure S5: The histograms of dihedral angle distributions for four protein dihedral generated by 200 ns normal MD in both bound (blue histogram) and unbound (green histogram) states of C01 for protein dihedral restraints reference orientations identification.

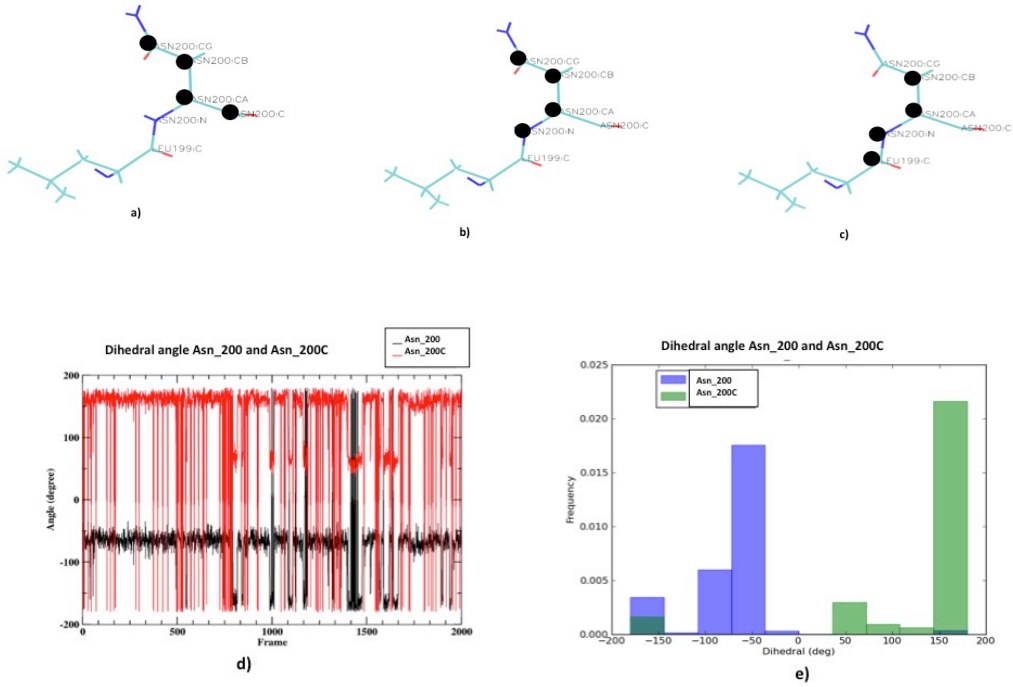


Figure S6: Analysis of dihedral of not contiguous atom (Dihedral Asn_200) on the protein receptor. Top panels: Illustrations of three potential dihedral corresponding to dihedral of the non contiguous atom: a) Dihedral Asn_200C: 200@C- 200@CA- 200@CB- 200@CG b) Asn_200N: 200@N- 200@CA- 200@CB- 200@CG and c) Asn_200CN: 199@C- 200@N- 200@CA- 200@CB. Atoms involvedED represent by black circle. Bottom panels: Analysis of dihedral Asn_200 and Asn_200C projected from trajectories of 200 ns unrestrained normal MD simulation in bound states: a) Fluctuation between dihedral Asn200 (black lines) and Asn_200C (red lines). b) Histogram of dihedral distributions between dihedral Asn200 (blue histogram) and Asn_200C (green histogram

List of References

(1) Lei, H.; Duan, Y. Improved Sampling Methods for Molecular Simulation. *Current Opinion in Structural Biology*. 2007, 17(2), 187–191.

(2) Roe, D. R.; Bergonzo, C.; Cheatham, T. E. Evaluation of Enhanced Sampling Provided by Accelerated Molecular Dynamics with Hamiltonian Replica Exchange Methods. *J. Phys. Chem. B* **2014**, *118* (13), 3543–3552.

(3) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. Comparison of Multiple Amber Force Fields and Development of Improved Protein Backbone Parameters. *Proteins Struct. Funct. Bioinforma.* **2006**, *65* (3), 712–725.

(4) Beauchamp, K. A.; Lin, Y.-S.; Das, R.; Pande, V. S. Are Protein Force Fields Getting Better? A Systematic Benchmark on 524 Diverse NMR Measurements. *J. Chem. Theory Comput.* **2012**, *8* (4), 1409–1414.

(5) Dickson, C. J.; Madej, B. D.; Skjevik, Å. A.; Betz, R. M.; Teigen, K.; Gould, I. R.; Walker, R. C. Lipid14: The Amber Lipid Force Field. *J. Chem. Theory Comput.* **2014**, *10* (2), 865–879.

(6) Vuong, T. V.; Wilson, D. B. Glycoside Hydrolases: Catalytic Base/Nucleophile Diversity. *Biotechnology and Bioengineering*. 2010, 107(2) 195–205.

(7) Neves, S. R.; Ram, P. T.; Iyengar, R. G Protein Pathways. *Science* **2002**, *296* (5573), 1636–1639.

(8) Voskoboinik, I.; Dunstone, M. A.; Baran, K.; Whisstock, J. C.; Trapani, J. A. Perforin: Structure, Function, and Role in Human Immunopathology. *Immunological Reviews*. 2010, 235(1), 35–54.

(9) Cao, X.; Cai, S. F.; Fehniger, T. A.; Song, J.; Collins, L. I.; Piwnica-Worms, D. R.; Ley, T. J. Granzyme B and Perforin Are Important for Regulatory T Cell-Mediated Suppression of Tumor Clearance. *Immunity* **2007**, *27* (4), 635–646.

(10) Overbaugh, J.; Morris, L. The Antibody Response against HIV-1. *Cold Spring Harb. Perspect. Med.* **2012**, *2* (1).

(11) Thomas, M. J. The Molecular Basis of Growth Hormone Action. *Growth Horm. IGF Res.* **1998**, *8* (1), 3–11.

(12) Quik, E. H.; van Dam, P. S.; Kenemans, J. L. Growth Hormone and Selective Attention: A

Review. *Neuroscience and Biobehavioral Reviews*. 2010, 34(8) 1137–1143.

(13) Fischer, E. Einfluss Der Configuration Auf Die Wirkung Der Enzyme. *Ber. Dtsch. Chem. Ges.* **1894**, *27*, 2985–2993.

(14) Koshland, D. E. The Key–Lock Theory and the Induced Fit Theory. *Angew. Chemie Int. Ed. English* **1995**, *33*, 2375–2378.

(15) Teif, V. B. Ligand-Induced DNA Condensation: Choosing the Model. *Biophys. J.* **2005**, *89*, 2574–2587.

(16) Teif, V. B.; Rippe, K. Statistical-Mechanical Lattice Models for Protein-DNA Binding in Chromatin. *J. Phys. Condens. Matter* **2010**, *22* (41), 414105.

(17) Li, J.; Fitzpatrick, P. F. Regulation of Phenylalanine Hydroxylase: Conformational Changes upon Phosphorylation Detected by H/D Exchange and Mass Spectrometry. *Arch. Biochem. Biophys.* **2013**, *535* (2), 115–119.

(18) Fisher, J.; Devraj, K.; Ingram, J.; Slagle-Webb, B.; Madhankumar, A. B.; Liu, X.; Klinger, M.; Simpson, I. A.; Connor, J. R. Ferritin: A Novel Mechanism for Delivery of Iron to the Brain and Other Organs. *Am. J. Physiol. Cell Physiol.* **2007**, *293* (2), C641–C649.

(19) Brooks, A. J.; Waters, M. J. The Growth Hormone Receptor: Mechanism of Activation and Clinical Implications. *Nat. Rev. Endocrinol.* **2010**, *6* (9), 515–525.

(20) Diebolder, C. a; Beurskens, F. J.; de Jong, R. N.; Koning, R. I.; Strumane, K.; Lindorfer, M. a; Voorhorst, M.; Ugurlar, D.; Rosati, S.; Heck, A. J. R.; van de Winkel, J. G. J.; Wilson, I. a; Koster, A. J.; Taylor, R. P.; Saphire, E. O.; Burton, D. R.; Schuurman, J.; Gros, P.; Parren, P. W. H. I. Complement Is Activated by IgG Hexamers Assembled at the Cell Surface. *Science* **2014**, *343* (6176), 1260–1263.

(21) Gohlke, H.; Klebe, G. Approaches to the Description and Prediction of the Binding Affinity of Small-Molecule Ligands to Macromolecular Receptors. *Angewandte Chemie - International Edition*. 2002, 41(15), 2644–2676.

(22) Whitesides, G. M.; Krishnamurthy, V. M. Designing Ligands to Bind Proteins. *Q. Rev. Biophys.* **2005**, *38* (4), 385–395.

(23) Dunn, M. F. Protein – Ligand Interactions : General Description. **2010**, 1–12.

(24) Wilchek, M.; Bayer, E. A. The Avidin-Biotin Complex in Bioanalytical Applications. *Anal. Biochem.* **1988**, *171* (1), 1–32.

(25) Fisher, H. F. Protein–Ligand Interactions: Thermodynamic Basis and Mechanistic Consequences. In *eLS* ; John Wiley & Sons, Ltd, **2001.**

(26) Gohlke, H.; Hendlich, M.; Klebe, G. Predicting Binding Modes, Binding Affinities and "Hot Spots" for Protein-Ligand Complexes Using a Knowledge-Based Scoring Function. *Perspect. Drug Discov. Des* **2000**, *20*, 115–144.

(27) Williams, D. H.; Stephens, E.; O'Brien, D. P.; Zhou, M. Understanding Noncovalent Interactions: Ligand Binding Energy and Catalytic Efficiency from Ligand-Induced Reductions in Motion within Receptors and Enzymes. *Angewandte Chemie - International Edition*. 2004, 43(48), 6596–6616.

(28) Searle, M. S.; Westwell, M. S.; Williams, D. H. Application of a Generalised Enthalpy-Entropy Relationship to Binding Co-Operativity and Weak Associations in Solution. *J. Chem. Soc. Perkin Trans. 2* **1995**, No. 1, 141–151.

(29) Kola, I.; Landis, J. Can the Pharmaceutical Industry Reduce Attrition Rates? *Nat. Rev. Drug Discov.* **2004**, *3* (8), 711–715.

(30) Avorn, J. The $2.6 Billion Pill - Methodological and Policy Considerations. *Perspective* **2015**.

(31) Dickson, M.; Gagnon, J. P. Key Factors in the Rising Cost of New Drug Discovery and Development. *Nat. Rev. Drug Discov.* **2004**, *3* (5), 417–429.

(32) Hubbard, R. E. *Structure-Based Drug Discovery: An Overview*; Royal Society of Chemistry, **2006**; Vol. 3.

(33) Davis, A. M.; Keeling, D. J.; Steele, J.; Tomkinson, N. P.; Tinker, A. C. Components of Successful Lead Generation. *Curr. Top. Med. Chem.* **2005**, *5* (4), 421–439.

(34) Congreve, M.; Chessari, G.; Tisi, D.; Woodhead, A. J. Recent Developments in Fragment-Based Drug Discovery. *Journal of Medicinal Chemistry*. 2008, 60(13), 3661–3680.

(35) Fischer, M.; Hubbard, R. E. Fragment-Based Ligand Discovery. *Mol. Interv.* **2009**, *9* (1), 22–30.

(36) Hajduk, P. J.; Greer, J. A Decade of Fragment-Based Drug Design: Strategic Advances and Lessons Learned. *Nat. Rev. Drug Discov.* **2007**, *6* (3), 211–219.

(37) Hajduk, P. J. SAR by NMR: Putting the Pieces Together. *Mol. Interv.* **2006**, *6* (5), 266–272.

(38) Erlanson, D. A.; McDowell, R. S.; O'Brien, T. Fragment-Based Drug Discovery. *J. Med. Chem.*

2004, *47* (14), 3463–3482.

(39) Zartler, E. R.; Shapiro, M. J. Fragonomics: Fragment-Based Drug Discovery. *Curr. Opin. Chem. Biol.* **2005**, *9* (4), 366–370.

(40) de Kloe, G. E.; Bailey, D.; Leurs, R.; de Esch, I. J. P. Transforming Fragments into Candidates: Small Becomes Big in Medicinal Chemistry. *Drug Discovery Today*. 2009, 14(13) 630–646.

(41) Tummino, P. J.; Copeland, R. A. Residence Time of Receptor-Ligand Complexes and Its Effect on Biological Function. *Biochemistry* **2008**, *47* (20), 5481–5492.

(42) Piehler, J. New Methodologies for Measuring Protein Interactions in Vivo and in Vitro. *Current Opinion in Structural Biology*. 2005, 15(1), 4–14.

(43) Shoemaker, B. A.; Panchenko, A. R. Deciphering Protein-Protein Interactions. Part I. Experimental Techniques and Databases. *PLoS Computational Biology*. **2007**,3, 0337–0344.

(44) Vuignier, K.; Schappler, J.; Veuthey, J.-L.; Carrupt, P.-A.; Martel, S. Drug-Protein Binding: A Critical Review of Analytical Tools. *Anal. Bioanal. Chem.* **2010**, *398* (1), 53–66.

(45) Wilkinson, K. D. Quantitative Analysis of Protein-Protein Interactions. *Methods Mol. Biol.* **2004**, *261*, 15–32.

(46) Ladbury, J. E.; Chowdhry, B. Z. Sensing the Heat: The Application of Isothermal Titration Calorimetry to Thermodynamic Studies of Biomolecular Interactions. *Chem. Biol.* **1996**, *3* (10), 791–801.

(47) Willander, M.; Al-Hilli, S. Analysis of Biomolecules Using Surface Plasmons. *Methods in molecular biology (Clifton, N.J.)*. 2009, 544, 201–229.

(48) Masi, A.; Cicchi, R.; Carloni, A.; Pavone, F. S.; Arcangeli, A. Optical Methods in the Study of Protein-Protein Interactions. *Adv. Exp. Med. Biol.* **2010**, *674*, 33–42.

(49) Nienhaus, K.; Nienhaus, G. U. Probing Heme Protein-Ligand Interactions by UV/Visible Absorption Spectroscopy. In *Protein-Ligand Interactions: Methods and Applications*; Ulrich Nienhaus, G., Ed.; Humana Press: Totowa, NJ, **2005**, 215–241.

(50) Ladbury, J. E.; Klebe, G.; Freire, E. Adding Calorimetric Data to Decision Making in Lead Discovery: A Hot Tip. *Nat. Rev. Drug Discov.* **2010**, *9* (1), 23–27.

(51) Velázquez-Campoy, A.; Ohtaka, H.; Nezami, A.; Muzammil, S.; Freire, E. Isothermal Titration Calorimetry. In *Current Protocols in Cell Biology*; John Wiley & Sons, Inc., **2004**.

(52) Freire, E. Do Enthalpy and Entropy Distinguish First in Class from Best in Class? *Drug Discovery Today*. **2008**, 13(19-20), 869–874.

(53) Freire, E. A Thermodynamic Approach to the Affinity Optimization of Drug Candidates. *Chem. Biol. Drug Des.* **2009**, *74* (5), 468–472.

(54) Rich, R. L.; Myszka, D. G. Advances in Surface Plasmon Resonance Biosensor Analysis. *Curr. Opin. Biotechnol.* **2000**, *11* (1), 54–61.

(55) Schuck, P. Use of Surface Plasmon Resonance to Probe the Equilibrium and Dynamic Aspects of Interactions between Biological Macromolecules. *Annu. Rev. Biophys. Biomol. Struct.* **1997**, *26*, 541–566.

(56) Jonsson, U.; Fagerstam, L.; Ivarsson, B.; Johnsson, B.; Karlsson, R.; Lundh, K.; Lofas, S.; Persson, B.; Roos, H.; Ronnberg, I.; Sjolander, S.; Stenberg, E.; Stahlberg, R.; Urbaniczky, C.; Ostlin, H.; Malmqvist, M. Real-Time Biospecific Interaction Analysis Using Surface Plasmon Resonance and a Sensor Chip Technology. *Biotechniques* **1991**, *11* (5).

(57) Merwe, P. A. Van Der. Surface Plasmon Resonance GENERAL PRINCIPLES OF BIACORE EXPERIMENTS. *Physics (College. Park. Md).* **2010**, *627*, 1–50.

(58) Qin, S.; Pang, X.; Zhou, H. X. Automated Prediction of Protein Association Rate Constants. *Structure* **2011**, *19* (12), 1744–1751.

(59) Wyer, J. R.; Willcox, B. E.; Gao, G. F.; Gerth, U. C.; Davis, S. J.; Bell, J. I.; van der Merwe, P. A.; Jakobsen, B. K. T Cell Receptor and Coreceptor CD8αα Bind Peptide-MHC Independently and with Distinct Kinetics. *Immunity* **1999**, *10* (2), 219–225.

(60) Willcox, B. E.; Gao, G. F.; Wyer, J. R.; Ladbury, J. E.; Bell, J. I.; Jakobsen, B. K.; van der Merwe, P. A. TCR Binding to Peptide-MHC Stabilizes a Flexible Recognition Interface. *Immunity* **1999**, *10* (3), 357–365.

(61) Kastritis, P. L.; Bonvin, A. M. J. J. On the Binding Affinity of Macromolecular Interactions: Daring to Ask Why Proteins Interact. *J. R. Soc. Interface* **2013**, *10*, 20120835.

(62) Klotz, I. M. Ligand–Receptor Interactions: Facts and Fantasies. *Q. Rev. Biophys.* **1985**, *18* (03), 227–259.

(63) Brenk, R.; Vetter, S. W.; Boyce, S. E.; Goodin, D. B.; Shoichet, B. K. Probing Molecular Docking in a Charged Model Binding Site. *J. Mol. Biol.* **2006**.375(5), 1449-1470

(64) Musah, R. A.; Jensen, G. M.; Bunte, S. W.; Rosenfeld, R. J.; Goodin, D. B. Artificial Protein

Cavities as Specific Ligand-Binding Templates: Characterization of an Engineered Heterocyclic Cation-Binding Site That Preserves the Evolved Specificity of the Parent Protein. *J. Mol. Biol.* **2002**. 315(4), 845-857.

(65) Nienhaus, K.; Lamb, D. C.; Deng, P.; Nienhaus, G. U. The Effect of Ligand Dynamics on Heme Electronic Transition Band III in Myoglobin. *Biophys. J.* **2002**. 82(2), 1059-1067.

(66) Mobley, D. L.; Dill, K. A. Binding of Small-Molecule Ligands to Proteins: "What You See" Is Not Always "What You Get." *Structure* **2009**, *17*, 489–498.

(67) Kirkwood, J. G. Statistical Mechanics of Fluid Mixtures. *J. Chem. Phys.* **1935**, *3* (5), 300–313.

(68) Zwanzig, R. High - Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. *J. Chem. Phys.* **1954**, *22* (8), 1420–1426.

(69) Tembre, B. L.; Mc Cammon, J. A. Ligand-Receptor Interactions. *Comput. Chem.* **1984**, *8* (4), 281–283.

(70) Jorgensen, W. L.; Ravimohan, C. Monte Carlo Simulation of Differences in Free Energies of Hydration. *J. Chem. Phys.* **1985**, *83* (6), 3050.

(71) Bash, P. A.; Singh, U. C.; Langridge, R.; Kollman, P. A. Free Energy Calculations by Computer Simulation. *Science* **1987**, *236* (4801), 564–568.

(72) Michel, J.; Essex, J. W. Prediction of Protein-Ligand Binding Affinity by Free Energy Simulations: Assumptions, Pitfalls and Expectations. *Journal of Computer-Aided Molecular Design*. **2010**, 224(8), 639–658.

(73) Boresch, S.; Tettinger, F.; Leitgeb, M.; Karplus, M. Absolute Binding Free Energies: A Quantitative Approach for Their Calculation. *J. Phys. Chem. B* **2003**, *107* (35), 9535–9551.

(74) Hermans, J.; Wang, L. Inclusion of Loss of Translational and Rotational Freedom in Theoretical Estimates of Free Energies of Binding. Application to a Complex of Benzene and Mutant T4 Lysozyme. *J. Am. Chem. Soc.* **1997**, *119* (11), 2707–2714.

(75) Roux, B.; Nina, M.; Pomès, R.; Smith, J. C. Thermodynamic Stability of Water Molecules in the Bacteriorhodopsin Proton Channel: A Molecular Dynamics Free Energy Perturbation Study. *Biophys. J.* **1996**, *71* (2), 670.

(76) Pearlman, D. A. Evaluating the Molecular Mechanics Poisson-Boltzmann Surface Area Free Energy Method Using a Congeneric Series of Ligands to P38 MAP Kinase. *J. Med. Chem.* **2005**, *48* (24), 7796–7807.

(77) Steinbrecher, T.; Case, D. A.; Labahn, A. A Multistep Approach to Structure-Based Drug Design: Studying Ligand Binding at the Human Neutrophil Elastase. *J. Med. Chem.* **2006**, *49* (6), 1837–1844.

(78) Mobley, D. L.; Graves, A. P.; Chodera, J. D.; McReynolds, A. C.; Shoichet, B. K.; Dill, K. A. Predicting Absolute Ligand Binding Free Energies to a Simple Model Site. *J. Mol. Biol.* **2007**, *371* (4), 1118–1134.

(79) Pearlman, D. A.; Charifson, P. S. Are Free Energy Calculations Useful in Practice? A Comparison with Rapid Scoring Functions for the P38 MAP Kinase Protein System. *J. Med. Chem.* **2001**, *44* (21), 3417–3423.

(80) Deng, Y.; Roux, B. Calculation of Standard Binding Free Energies: Aromatic Molecules in the T4 Lysozyme L99A Mutant. *J. Chem. Theory Comput.* **2006**, *2* (5), 1255–1273.

(81) Fujitani, H.; Tanida, Y.; Ito, M.; Jayachandran, G.; Snow, C. D.; Shirts, M. R.; Sorin, E. J.; Pande, V. S. Direct Calculation of the Binding Free Energies of FKBP Ligands. *J. Chem. Phys.* **2005**, *123* (8), 4108.

(82) Jayachandran, G.; Shirts, M. R.; Park, S.; Pande, V. S. Parallelized-over-Parts Computation of Absolute Binding Free Energy with Docking and Molecular Dynamics. *J. Chem. Phys.* **2006**, *125* (8), 84901–84912.

(83) Shirts, M. R.; Mobley, D. L.; Chodera, J. D. Alchemical Free Energy Calculations: Ready for Prime Time? *Annu. Rep. Comput. Chem.* **2007**.

(84) Wang, J.; Hou, T.; Xu, X. Recent Advances in Free Energy Calculations with a Combination of Molecular Mechanics and Continuum Models. *Curr. Comput. - Aided Drug Des.* **2006**, *2*, 287–306.

(85) Wang, L.; Wu, Y.; Deng, Y.; Kim, B.; Pierce, L.; Krilov, G.; Lupyan, D.; Robinson, S.; Dahlgren, M. K.; Greenwood, J. Accurate and Reliable Prediction of Relative Ligand Binding Potency in Prospective Drug Discovery by Way of a Modern Free-Energy Calculation Protocol and Force Field. *J. Am. Chem. Soc.* **2015**, *137* (7), 2695–2703.

(86) Woo, H.-J.; Roux, B. Calculation of Absolute Protein–Ligand Binding Free Energy from Computer Simulations. *Proc. Natl. Acad. Sci. United States Am.* **2005**, *102* (19), 6825–6830.

(87) Selzer, T.; Albeck, S.; Schreiber, G. Rational Design of Faster Associating and Tighter Binding Protein Complexes. *Nat. Struct. Mol. Biol.* **2000**, *7* (7), 537–541.

(88) H.-J. Böhm, G. S. *Introduction to Molecular Recognition Models*; **2003**.

(89) Wang, W.; Donini, O.; Reyes, C. M.; Kollman, P. A. Biomolecular Simulations: Recent Developments in Force Fields, Simulations of Enzyme Catalysis, Protein-Ligand, Protein-Protein, and Protein-Nucleic Acid Noncovalent Interactions. *Annu. Rev. Biophys. Biomol. Struct.* **2001**, *30*, 211–243.

(90) Kollman, P. a.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Donini, O.; Cieplak, P.; Srinivasan, J.; Case, D. a.; Cheatham, T. E. Calculating Structures and Free Energies of Complex Molecules: Combining Molecular Mechanics and Continuum Models. *Acc. Chem. Res.* **2000**, *33* (12), 889–897.

(91) Feig, M.; Brooks, C. L. Recent Advances in the Development and Application of Implicit Solvent Models in Biomolecule Simulations. *Curr. Opin. Struct. Biol.* **2004**, *14* (2), 217–224.

(92) Kumari, R.; Kumar, R.; Lynn, A. G-Mmpbsa -A GROMACS Tool for High-Throughput MM-PBSA Calculations. *J. Chem. Inf. Model.* **2014**, *54*, 1951–1962.

(93) Srinivasan, J.; Cheatham, T. E.; Cieplak, P.; Kollman, P. A.; Case, D. A. Continuum Solvent Studies of the Stability of DNA, RNA, and Phosphoramidate - DNA Helices. *J. Am. Chem. Soc.* **1998**, *120*, 9401–9409.

(94) Homeyer, N.; Gohlke, H. Free Energy Calculations by the Molecular Mechanics Poisson-Boltzmann Surface Area Method. *Mol. Inform.* **2012**, *31*, 114–122.

(95) Rastelli, G.; Del Rio, A.; Degliesposti, G.; Sgobba, M. Fast and Accurate Predictions of Binding Free Energies Using MM-PBSA and MM-GBSA. *J. Comput. Chem.* **2010**, *31*, 797–810.

(96) Okimoto, N.; Futatsugi, N.; Fuji, H.; Suenaga, A.; Morimoto, G.; Yanai, R.; Ohno, Y.; Narumi, T.; Taiji, M. High-Performance Drug Discovery: Computational Screening by Combining Docking and Molecular Dynamics Simulations. *PLoS Comput. Biol.* **2009**, *5*.

(97) Houa, T.; Wangb, J.; Lia, Y.; Wang, W. Assessing the Performance of the MM/PBSA and MM/GBSA Methods: I. The Accuracy of Binding Free Energy Calculations Based on Molecular Dynamics Simulations. *J. Chem. Inf. Comput. Sci.* **2011**, *51* (1), 69–82.

(98) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A Geometric Approach to Macromolecule-Ligand Interactions. *J. Mol. Biol.* **1982**, *161*, 269–288.

(99) Ferreira, R. S.; Simeonov, A.; Jadhav, A.; Eidam, O.; Mott, B. T.; Keiser, M. J.; McKerrow, J.

H.; Maloney, D. J.; Irwin, J. J.; Shoichet, B. K. Complementarity between a Docking and a High-Throughput Screen in Discovering New Cruzain Inhibitors. *J. Med. Chem.* **2010**, *53*, 4891–4905.

(100) Huang, S. Y.; Zou, X. Advances and Challenges in Protein-Ligand Docking. *International Journal of Molecular Sciences*. **2010**, 11(8) 3016–3034.

(101) Verlinde, C. L.; Hol, W. G. Structure-Based Drug Design: Progress, Results and Challenges. *Structure* **1994**, *2* (7), 577–587.

(102) Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V; Mee, R. P. Empirical Scoring Functions: I. The Development of a Fast Empirical Scoring Function to Estimate the Binding Affinity of Ligands in Receptor Complexes. *J. Comput. Aided. Mol. Des.* **1997**, *11* (5), 425–445.

(103) Mitchell, J. B. O.; Laskowski, R. A.; Alex, A.; Thornton, J. M. BLEEP—Potential of Mean Force Describing Protein–Ligand Interactions: I. Generating Potential. *J. Comput. Chem.* **1999**, *20* (11), 1165–1176.

(104) Muegge, I.; Martin, Y. C. A General and Fast Scoring Function for Protein-Ligand Interactions: A Simplified Potential Approach. *J. Med. Chem.* **1999**, *42* (5), 791–804.

(105) Ferrara, P.; Gohlke, H.; Price, D. J.; Klebe, G.; Brooks, C. L. Assessing Scoring Functions for Protein-Ligand Interactions. *J. Med. Chem.* **2004**, *47*, 3032–3047.

(106) Paluch, A. S.; Mobley, D. L.; Maginn, E. J. Small Molecule Solvation Free Energy: Enhanced Conformational Sampling Using Expanded Ensemble Molecular Dynamics Simulation. *J. Chem. Theory Comput.* **2011**, *7* (9), 2910–2918.

(107) Abrams, C.; Bussi, G. Enhanced Sampling in Molecular Dynamics Using Metadynamics, Replica-Exchange, and Temperature-Acceleration. *Entropy* **2014**, *16* (1), 163–199.

(108) Paluch, A. S.; Mobley, D. L.; Maginn, E. J. Small Molecule Solvation Free Energy: Enhanced Conformational Sampling Using Expanded Ensemble Molecular Dynamics Simulation. *J. Chem. Theory Comput.* **2011**, *7* (9), 2910–2918.

(109) Overington, J. P.; Overington, J. P.; Al-Lazikani, B.; Al-Lazikani, B.; Hopkins, A. L.; Hopkins, A. L. How Many Drug Targets Are There? *Nat. Rev. Drug Discov.* **2006**, *5* (12), 993–996.

(110) Baldi, A. Computational Approaches for Drug Design and Discovery: An Overview. *Syst. Rev. Pharm.* **2010**, *1* (1), 99.

(111) Sliwoski, G.; Kothiwale, S.; Meiler, J.; Lowe, E. W. Computational Methods in Drug Discovery. *Pharmacol. Rev.* **2014**, *66* (1), 334–395.

(112) Cornell, W.; Cieplak, P.; Bayly, C.; Gould, I.; Merz …, K. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *Journal of the American Chemical Society*. **1995**. 117(19), 5179-5197.

(113) MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B* **1998**, *102* (18), 3586–3616.

(114) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.* **1996**, *118* (45), 11225–11236.

(115) Oostenbrink, C.; Villa, A.; Mark, A. E.; Van Gunsteren, W. F. A Biomolecular Force Field Based on the Free Enthalpy of Hydration and Solvation: The GROMOS Force-field Parameter Sets 53A5 and 53A6. *J. Comput. Chem.* **2004**, *25* (13), 1656–1676.

(116) Ren, P.; Ponder, J. W. Consistent Treatment of Inter-and Intramolecular Polarization in Molecular Mechanics Calculations. *J. Comput. Chem.* **2002**, *23* (16), 1497–1506.

(117) Dong, F.; Olsen, B.; Baker, N. A. Computational Methods for Biomolecular Electrostatics. *Methods Cell Biol.* **2008**, *84*, 843–870.

(118) Israelachvili, J. N. *Intermolecular and Surface Forces*; **2011**.

(119) Maple, J. R.; Hwang, M.-J.; Stockfisch, T. P.; Dinur, U.; Waldman, M.; Ewig, C. S.; Hagler, A. T. Derivation of Class II ForceF Fields. I. Methodology and Quantum Force Field for the Alkyl Functional Group and Alkane Molecules. *J. Comput. Chem.* **1994**, *15* (2), 162–182.

(120) Allinger, N. L.; Yuh, Y. H.; Lii, J. H. Molecular Mechanics. The MM3 Force Field for Hydrocarbons. 1. *J. Am. Chem. Soc.* **1989**, *111* (23), 8551–8566.

(121) Allinger, N. L.; Chen, K.; Lii, J. An Improved Force Field (MM4) for Saturated Hydrocarbons. *J. Comput. Chem.* **1996**, *17* (5-6), 642–668.

(122) Clark, M. Validation of the General Purpose Tripos 5.2 Force Field. *J. Comput. Chem.* **1989**, *10*, 982–1012.

(123) Mackerell, A. D. Empirical Force Fields for Biological Macromolecules: Overview and Issues. *J. Comput. Chem.* **2004**, *25* (13), 1584–1604.

(124) Leach, A. R. *Molecular Modelling: Principles and Applications*; 2001; Vol. 2nd.

(125) McCamm0n, J. A. Dynamics of Folded Proteins. *Nature* **1977**, *267*, 16.

(126) Cieplak, P.; Dupradeau, F.-Y.; Duan, Y.; Wang, J. Polarization Effects in Molecular Mechanical Force Fields. *Journal of Physics: Condensed Matter*. **2009**, 21(33) 333102.

(127) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and Testing of a General Amber Force Field. *J. Comput. Chem.* **2004**, *25*, 1157–1174.

(128) Jakalian, A.; Jack, D. B.; Bayly, C. I. Fast, Efficient Generation of High-quality Atomic Charges. AM1-BCC Model: II. Parameterization and Validation. *J. Comput. Chem.* **2002**, *23* (16), 1623–1641.

(129) Kaminski, G. A.; Stern, H. A.; Berne, B. J.; Friesner, R. A. Development of an Accurate and Robust Polarizable Molecular Mechanics Force Field from Ab Initio Quantum Chemistry. *J. Phys. Chem. A* **2004**, *108* (4), 621–627.

(130) Ren, P.; Ponder, J. W. Temperature and Pressure Dependence of the AMOEBA Water Model. *J. Phys. Chem. B* **2004**, *108* (35), 13427–13437.

(131) Jiao, D.; King, C.; Grossfield, A.; Darden, T. A.; Ren, P. Simulation of Ca2+ and Mg2+ Solvation Using Polarizable Atomic Multipole Potential. *J. Phys. Chem. B* **2006**, *110* (37), 18553–18559.

(132) Xie, W.; Pu, J.; MacKerell, A. D.; Gao, J. Development of a Polarizable Intermolecular Potential Function (PIPF) for Liquid Amides and Alkanes. *J. Chem. Theory Comput.* **2007**, *3* (6), 1878–1889.

(133) Sprik, M. Computer Simulation of the Dynamics of Induced Polarization Fluctuations in Water. *J. Phys. Chem.* **1991**, *95* (6), 2283–2291.

(134) Dang, L. X.; Chang, T.-M. Molecular Dynamics Study of Water Clusters, Liquid, and Liquid–Vapor Interface of Water with Many-Body Potentials. *J. Chem. Phys.* **1997**, *106* (19), 8149–8159.

(135) Brdarski, S.; Åstrand, P.-O.; Karlström, G. The Inclusion of Electron Correlation in Intermolecular Potentials: Applications to the Formamide Dimer and Liquid Formamide. *Theor. Chem. Acc.* **2000**, *105* (1), 7–14.

(136) Stern, H. A.; Kaminski, G. A.; Banks, J. L.; Zhou, R.; Berne, B. J.; Friesner, R. A. Fluctuating Charge, Polarizable Dipole, and Combined Models: Parameterization from Ab Initio

Quantum Chemistry. *J. Phys. Chem. B* **1999**, *103* (22), 4730–4737.

(137) Burnham, C. J.; Li, J.; Xantheas, S. S.; Leslie, M. The Parametrization of a Thole-Type All-Atom Polarizable Water Model from First Principles and Its Application to the Study of Water Clusters (N= 2–21) and the Phonon Spectrum of Ice Ih. *J. Chem. Phys.* **1999**, *110* (9), 4566–4581.

(138) Kaminski, G. A.; Stern, H. A.; Berne, B. J.; Friesner, R. A.; Cao, Y. X.; Murphy, R. B.; Zhou, R.; Halgren, T. A. Development of a Polarizable Force Field for Proteins via Ab Initio Quantum Chemistry: First Generation Model and Gas Phase Tests. *J. Comput. Chem.* **2002**, *23* (16), 1515–1531.

(139) Burnham, C. J.; Xantheas, S. S. Development of Transferable Interaction Models for Water. I. Prominent Features of the Water Dimer Potential Energy Surface. *J. Chem. Phys.* **2002**, *116* (4), 1479–1492.

(140) Ren, P.; Ponder, J. W. Polarizable Atomic Multipole Water Model for Molecular Mechanics Simulation. *J. Phys. Chem. B* **2003**, *107* (24), 5933–5947.

(141) Rappe, A. K.; Goddard III, W. A. Charge Equilibration for Molecular Dynamics Simulations. *J. Phys. Chem.* **1991**, *95* (8), 3358–3363.

(142) Rick, S. W.; Stuart, S. J.; Berne, B. J. Dynamical Fluctuating Charge Force Fields: Application to Liquid Water. *J. Chem. Phys.* **1994**, *101* (7), 6141–6156.

(143) Banks, J. L.; Kaminski, G. A.; Zhou, R.; Mainz, D. T.; Berne, B. J.; Friesner, R. A. Parametrizing a Polarizable Force Field from Ab Initio Data. I. The Fluctuating Point Charge Model. *J. Chem. Phys.* **1999**, *110*, 741–754.

(144) Ando, K. A Stable Fluctuating-Charge Polarizable Model for Molecular Dynamics Simulations: Application to Aqueous Electron Transfers. *J. Chem. Phys.* **2001**, *115* (11), 5228–5237.

(145) Yoshii, N.; Miyauchi, R.; Miura, S.; Okazaki, S. A Molecular-Dynamics Study of the Equation of State of Water Using a Fluctuating-Charge Model. *Chem. Phys. Lett.* **2000**, *317* (3), 414–420.

(146) Patel, S.; Brooks, C. L. CHARMM Fluctuating Charge Force Field for Proteins: I Parameterization and Application to Bulk Organic Liquid Simulations. *J. Comput. Chem.* **2004**, *25* (1), 1–16.

(147) Patel, S.; Mackerell, A. D.; Brooks, C. L. CHARMM Fluctuating Charge Force Field for Proteins: II Protein/Solvent Properties from Molecular Dynamics Simulations Using a Nonadditive Electrostatic Model. *J. Comput. Chem.* **2004**, *25* (12), 1504–1514.

(148) van Maaren, P. J.; van der Spoel, D. Molecular Dynamics Simulations of Water with Novel Shell-Model Potentials. *J. Phys. Chem. B* **2001**, *105* (13), 2618–2626.

(149) Yu, H.; Hansson, T.; van Gunsteren, W. F. Development of a Simple, Self-Consistent Polarizable Model for Liquid Water. *J. Chem. Phys.* **2003**, *118* (1), 221–234.

(150) Lamoureux, G.; MacKerell Jr, A. D.; Roux, B. A Simple Polarizable Model of Water Based on Classical Drude Oscillators. *J. Chem. Phys.* **2003**, *119* (10), 5185–5197.

(151) Harder, E.; Anisimov, V. M.; Whitfield, T.; MacKerell, A. D.; Roux, B. Understanding the Dielectric Properties of Liquid Amides from a Polarizable Force Field. *J. Phys. Chem. B* **2008**, *112* (11), 3509–3521.

(152) Lopes, P. E. M.; Lamoureux, G.; Roux, B.; MacKerell, A. D. Polarizable Empirical Force Field for Aromatic Compounds Based on the Classical Drude Oscillator. *J. Phys. Chem. B* **2007**, *111* (11), 2873–2885.

(153) Ren, P.; Wu, C.; Ponder, J. W. Polarizable Atomic Multipole-Based Molecular Mechanics for Organic Molecules. *J. Chem. Theory Comput.* **2011**, *7* (10), 3143–3161.

(154) Shi, Y.; Xia, Z.; Zhang, J.; Best, R.; Wu, C.; Ponder, J. W.; Ren, P. Polarizable Atomic Multipole-Based AMOEBA Force Field for Proteins. *J. Chem. Theory Comput.* **2013**, *9* (9), 4046–4063.

(155) Mu, X.; Wang, Q.; Wang, L.-P.; Fried, S. D.; Piquemal, J.-P.; Dalby, K. N.; Ren, P. Modeling Organochlorine Compounds and the σ-Hole Effect Using a Polarizable Multipole Force Field. *J. Phys. Chem. B* **2014**, *118* (24), 6456–6465.

(156) Ponder, J. W.; Wu, C.; Ren, P.; Pande, V. S.; Chodera, J. D.; Schnieders, M. J.; Haque, I.; Mobley, D. L.; Lambrecht, D. S.; DiStasio Jr, R. A. Current Status of the AMOEBA Polarizable Force Field. *J. Phys. Chem. B* **2010**, *114* (8), 2549–2564.

(157) MacKerell, A. D.; Feig, M.; Brooks, C. L. Improved Treatment of the Protein Backbone in Empirical Force Fields. *J. Am. Chem. Soc.* **2004**, *126* (3), 698–699.

(158) Halgren, T. A. The Representation of van Der Waals (VdW) Interactions in Molecular Mechanics Force Fields: Potential Form, Combination Rules, and VdW Parameters. *J. Am.*

Chem. Soc. **1992**, *114* (20), 7827–7843.

(159) Stone, A. J. Distributed Multipole Analysis, Or How To Describe A Molecular Charge-Distribution. *Chem. Phys. Lett.* **1981**, *83* (2), 233–239.

(160) Stone, A. J. Distributed Multipole Analysis: Stability for Large Basis Sets. *J. Chem. Theory Comput.* **2005**, *1* (6), 1128–1132.

(161) Wang, L.-P.; Head-Gordon, T.; Ponder, J. W.; Ren, P.; Chodera, J. D.; Eastman, P. K.; Martinez, T. J.; Pande, V. S. Systematic Improvement of a Classical Molecular Model of Water. *J. Phys. Chem. B* **2013**, *117* (34), 9956–9972.

(162) Thole, B. T. Molecular Polarizabilities Calculated with a Modified Dipole Interaction. *Chem. Phys.* **1981**. 59, 341-350.

(163) Mura, C.; McAnany, C. E. An Introduction to Biomolecular Simulations and Docking. *Mol. Simul.* **2014**, *40* (10–11), 732–764.

(164) Alder, B.; WainWright, T. Phase Transition for a Hard Sphere System. *J. Chem. Phys.* **2012**, *27* (May 2015), 1208–1211.

(165) Rahman, A. Correlations in the Motion of Atoms in Liquid Argon. *Phys. Rev.* **1964**, *136* (2A), A405.

(166) Miyamoto, S.; Kollman, P. A. Settle: An Analytical Version of the SHAKE and RATTLE Algorithm for Rigid Water Models. *J. Comput. Chem.* **1992**, *13* (8), 952–962.

(167) Hockney, R. W. The Potential Calculation and Some Applications(Potential Calculation from given Source Distribution, Including Direct and Iterative Methods, Error Analysis. *Meth. Compt. Phys* **1970**, *9*, 136–211.

(168) Swope, W. C.; Andersen, H. C.; Berens, P. H.; Wilson, K. R. 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.* **1982**, *76* (1), 637–649.

(169) Fincham, D.; Heyes, D. M. Integration Algorithms in Molecular Dynamics. *CCP5 Q.* **1982**, *6*, 4–10.

(170) Henzler-Wildman, K.; Kern, D. Dynamic Personalities of Proteins. *Nature* **2007**, *450* (7172), 964–972.

(171) Xu, D.; Williamson, M. J.; Walker, R. C. Advancements in Molecular Dynamics Simulations

of Biomolecules on Graphical Processing Units. *Annu. Rep. Comput. Chem.* **2010**, *6*, 2–19.

(172) Friedrichs, M. S.; Eastman, P.; Vaidyanathan, V.; Houston, M.; Legrand, S.; Beberg, A. L.; Ensign, D. L.; Bruns, C. M.; Pande, V. S. Accelerating Molecular Dynamic Simulation on Graphics Processing Units. *J. Comput. Chem.* **2009**, *30* (6), 864–872.

(173) Götz, A. W.; Williamson, M. J.; Xu, D.; Poole, D.; Le Grand, S.; Walker, R. C. Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 1. Generalized Born. *J. Chem. Theory Comput.* **2012**, *8* (5), 1542–1555.

(174) Dror, R. O.; Pan, A. C.; Arlow, D. H.; Borhani, D. W.; Maragakis, P.; Shan, Y.; Xu, H.; Shaw, D. E. Pathway and Mechanism of Drug Binding to G-Protein-Coupled Receptors. *Proc. Natl. Acad. Sci.* **2011**, *108* (32), 13118–13123.

(175) Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y. Atomic-Level Characterization of the Structural Dynamics of Proteins. *Science (80).* **2010**, *330* (6002), 341–346.

(176) Parrinello, M.; Rahman, A. Polymorphic Transitions in Single Crystals: A New Molecular Dynamics Method. *J. Appl. Phys.* **1981**, *52* (12), 7182–7190.

(177) Tuckerman, M. E.; Alejandre, J.; López-Rendón, R.; Jochim, A. L.; Martyna, G. J. A Liouville-Operator Derived Measure-Preserving Integrator for Molecular Dynamics Simulations in the Isothermal–Isobaric Ensemble. *J. Phys. A. Math. Gen.* **2006**, *39* (19), 5629–5651.

(178) Shinoda, W.; Shiga, M.; Mikami, M. Rapid Estimation of Elastic Constants by Molecular Dynamics Simulation under Constant Stress. *Phys. Rev. B* **2004**, *69* (13), 134103.

(179) Martyna, G. J.; Tobias, D. J.; Klein, M. L. Constant Pressure Molecular Dynamics Algorithms. *J. Chem. Phys.* **1994**, *101* (5), 4177–4189.

(180) Scarpazza, D. P.; Ierardi, D. J.; Lerer, A. K.; Mackenzie, K. M.; Pan, A. C.; Bank, J. A.; Chow, E.; Dror, R. O.; Grossman, J. P.; Killebrew, D. Extending the Generality of Molecular Dynamics Simulations on a Special-Purpose Machine. In *Parallel & Distributed Processing (IPDPS), 2013 IEEE 27th International Symposium on*; IEEE, **2013**; 933–945.

(181) Evans, D. J.; Holian, B. L. The Nose–Hoover Thermostat. *J. Chem. Phys.* **1985**, *83* (8), 4069–4074.

(182) Hünenberger, P. H. Thermostat Algorithms for Molecular Dynamics Simulations. In *Advanced computer simulation*; Springer, 2005;173, 105–149.

(183) Martyna, G. J.; Klein, M. L.; Tuckerman, M. Nosé–Hoover Chains: The Canonical Ensemble via Continuous Dynamics. *J. Chem. Phys.* **1992**, *97* (4), 2635–2643.

(184) Pastor, R. W.; Brooks, B. R.; Szabo, A. An Analysis of the Accuracy of Langevin and Molecular Dynamics Algorithms. *Molecular Physics*. **1988**, 65 1409–1419.

(185) Lzaguirre, J. A.; Catarello, D. P.; Wozniak, J. M.; Skeel, R. D. Langevin Stabilization of Molecular Dynamics. *J. Chem. Phys.* **2001**, *114* (5), 2090–2098.

(186) Bussi, G.; Parrinello, M. Accurate Sampling Using Langevin Dynamics. *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.* **2007**, *75* (5).

(187) Steinhauser, M. O.; Hiermaier, S. A Review of Computational Methods in Materials Science: Examples from Shock-Wave and Polymer Physics. *International Journal of Molecular Sciences*. 2009, 10(12) 5135–5216.

(188) Darden, T.; York, D.; Pedersen, L. In Large Systems. **1993**, No. June, 10089–10092.

(189) de Ruiter, A.; Boresch, S.; Oostenbrink, C. Comparison of Thermodynamic Integration and Bennett Acceptance Ratio for Calculating Relative Protein-ligand Binding Free Energies. *J. Comput. Chem.* **2013**, *34* (12), 1024–1034.

(190) Bennett, C. H. Efficient Estimation of Free Energy Differences from Monte Carlo Data. *J. Comput. Phys.* **1976**, *22* (2), 245–268.

(191) Shirts, M. R.; Chodera, J. D. Statistically Optimal Analysis of Samples from Multiple Equilibrium States. *J. Chem. Phys.* **2008**, *129* (12), 124105.

(192) Pohorille, A.; Jarzynski, C.; Chipot, C. Good Practices in Free-Energy Calculations. *J. Phys. Chem. B* **2010**.114(32), 10235-10253.

(193) Ross, G. A.; Bruce Macdonald, H. E.; Cave-Ayland, C.; Cabedo Martinez, A. I.; Essex, J. W. Replica-Exchange and Standard State Binding Free Energies with Grand Canonical Monte Carlo. *J. Chem. Theory Comput.* **2017**. 13(12), 6373-6381

(194) Sugita, Y.; Okamoto, Y. Replica-Exchange Molecular Dynamics Method for Protein Folding. *Chem. Phys. Lett.* **1999**. 314(1-2), 141-151.

(195) Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. The Amber Biomolecular Simulation Programs. *Journal of Computational Chemistry*. **2005**, 26(16) 1668–1688.

(196) Meng, Y.; Sabri Dashti, D.; Roitberg, A. E. Computing Alchemical Free Energy Differences with Hamiltonian Replica Exchange Molecular Dynamics (H-REMD) Simulations. *J. Chem. Theory Comput.* **2011**. 7(9), 2721-2727

(197) Fukunishi, H.; Watanabe, O.; Takada, S. On the Hamiltonian Replica Exchange Method for Efficient Sampling of Biomolecular Systems: Application to Protein Structure Prediction. *J. Chem. Phys.* **2002**. 16(20), 9058-9067

(198) Liu, P.; Kim, B.; Friesner, R. A.; Berne, B. J. Replica Exchange with Solute Tempering: A Method for Sampling Biological Systems in Explicit Water. *Proc. Natl. Acad. Sci.* **2005**. 102(39), 13749-13754.

(199) Liu, P.; Voth, G. A. Smart Resolution Replica Exchange: An Efficient Algorithm for Exploring Complex Energy Landscapes. *J. Chem. Phys.* **2007**. 126(4), 045106

(200) Lyman, E.; Zuckerman, D. M. Resolution Exchange Simulation with Incremental Coarsening. *J. Chem. Theory Comput.* **2006**. 2(3), 656-666

(201) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys.* **1953**. 21(6), 1087-1092

(202) Kastenholz, M. A.; Hünenberger, P. H. Computation of Methodology-Independent Ionic Solvation Free Energies from Molecular Simulations. II. the Hydration Free Energy of the Sodium Cation. *J. Chem. Phys.* **2006**. 124(22), 224501

(203) Hünenberger, P. H.; McCammon, J. A. Ewald Artifacts in Computer Simulations of Ionic Solvation and Ion-Ion Interaction: A Continuum Electrostatics Study. *J. Chem. Phys.* **1999**.110(4), 1856-1872

(204) Rocklin, G. J.; Boyce, S. E.; Fischer, M.; Fish, I.; Mobley, D. L.; Shoichet, B. K.; Dill, K. A. Blind Prediction of Charged Ligand Binding Affinities in a Model Binding Site. *J. Mol. Biol.* **2013**, *425* (22), 4569–4583.

(205) Pranata, J.; Jorgensen, W. L. Monte Carlo Simulations Yield Absolute Free Energies of Binding for Guanine-Cytosine and Adenine-Uracil Base Pairs in Chloroform. *Tetrahedron* **1991**. 1, 357-369

(206) Jorgensen, W. L.; Buckner, J. K.; Boudon, S.; Tirado-Rives, J. Efficient Computation of Absolute Free Energies of Binding by Computer Simulations. Application to the Methane Dimer in Water. *J. Chem. Phys.* **1988**. 89, 3741-3746.

(207) Gilson, M. K.; Given, J. A.; Bush, B. L.; McCammon, J. A. The Statistical-Thermodynamic Basis for Computation of Binding Affinities: A Critical Review. *Biophysical Journal*. **1997**. 7(23), 1047-1069.

(208) Mobley, D. L.; Wymer, K. L.; Lim, N. M.; Guthrie, J. P. Blind Prediction of Solvation Free Energies from the SAMPL4 Challenge. *Journal of Computer-Aided Molecular Design*. **2014**, 28(3), 135–150.

(209) Manzoni, F.; Söderhjelm, P. Prediction of Hydration Free Energies for the SAMPL4 Data Set with the AMOEBA Polarizable Force Field. *J. Comput. Aided. Mol. Des.* **2014**, *28* (3), 235–244.

(210) Mobley, D. L.; Dumont, E.; Chodera, J. D.; Dill, K. A. Comparison of Charge Models for Fixed-Charge Force Fields: Small-Molecule Hydration Free Energies in Explicit Solvent. *J. Phys. Chem. B* **2007**, *111* (9), 2242–2254.

(211) Baker, C. M.; Lopes, P. E. M.; Zhu, X.; Roux, B.; MacKerell, A. D. Accurate Calculation of Hydration Free Energies Using Pair-Specific Lennard-Jones Parameters in the CHARMM Drude Polarizable Force Field. *J. Chem. Theory Comput.* **2010**, *6* (4), 1181–1198.

(212) Lundborg, M.; Lindahl, E. Automatic GROMACS Topology Generation and Comparisons of Force Fields for Solvation Free Energy Calculations. *J. Phys. Chem. B* **2015**, *119* (3), 810–823.

(213) Shivakumar, D.; Williams, J.; Wu, Y.; Damm, W.; Shelley, J.; Sherman, W. Prediction of Absolute Solvation Free Energies Using Molecular Dynamics Free Energy Perturbation and the OPLS Force Field. *J. Chem. Theory Comput.* **2010**, *6* (5), 1509–1519.

(214) Mobley, D. L.; Bayly, C. I.; Cooper, M. D.; Shirts, M. R.; Dill, K. A. Small Molecule Hydration Free Energies in Explicit Solvent: An Extensive Test of Fixed-Charge Atomistic Simulations. *J. Chem. Theory Comput.* **2009**, *5* (2), 350–358.

(215) Martins, S. A.; Sousa, S. F.; Ramos, M. J.; Fernandes, P. A. Prediction of Solvation Free Energies with Thermodynamic Integration Using the General Amber Force Field. *J. Chem. Theory Comput.* **2014**, *10* (8), 3570–3577.

(216) Kaminski, G.; Duffy, E. M.; Matsui, T.; Jorgensen, W. L. Free Energies of Hydration and Pure Liquid Properties of Hydrocarbons from the OPLS All-Atom Model. *J. Phys. Chem.* **1994**, *98* (4), 13077–13082.

(217) Udier-Blagović, M.; Morales De Tirado, P.; Pearlman, S. A.; Jorgensen, W. L. Accuracy of

Free Energies of Hydration Using CM1 and CM3 Atomic Charges. *J. Comput. Chem.* **2004**, *25* (11), 1322–1332.

(218) Shivakumar, D.; Harder, E.; Damm, W.; Friesner, R. A.; Sherman, W. Improving the Prediction of Absolute Solvation Free Energies Using the Next Generation OPLS Force Field. *J. Chem. Theory Comput.* **2012**, *8*, 2553–2558.

(219) Ren, P.; Wu, C.; Ponder, J. W. Polarizable Atomic Multipole-Based Molecular Mechanics for Organic Molecules. *J. Chem. Theory Comput.* **2011**, *7* (10), 3143–3161.

(220) Grabuleda, X.; Jaime, C.; Kollman, P. A. Molecular Dynamics Simulation Studies of Liquid Acetonitrile: New Six-site Model. *J. Comput. Chem.* **2000**, *21* (10), 901–908.

(221) Dupradeau, F.-Y.; Pigache, A.; Zaffran, T.; Savineau, C.; Lelong, R.; Grivel, N.; Lelong, D.; Rosanski, W.; Cieplak, P. The R.E.D. Tools: Advances in RESP and ESP Charge Derivation and Force Field Library Building. *Phys. Chem. Chem. Phys.* **2010**, *12* (28), 7821–7839.

(222) Marenich, A. V; Kelly, C. P.; Thompson, J. D.; Hawkins, G. D.; Chambers, C. C.; Giesen, D. J.; Winget, P.; Cramer, C. J.; Truhlar, D. G. Minnesota Solvation Database.

(223) Abraham, M. H.; Platts, J. A.; Hersey, A.; Leo, A. J.; Taft, R. W. Correlation and Estimation of Gas–Chloroform and Water–Chloroform Partition Coefficients by a Linear Free Energy Relationship Method. *J. Pharm. Sci.* **1999**, *88* (7), 670–679.

(224) Mohamed, N. A.; Bradshaw, R. T.; Essex, J. W. Evaluation of Solvation Free Energies for Small Molecules with the AMOEBA Polarizable Force Field. *J. Comput. Chem.* **2016**.

(225) Wu, J. C.; Chattree, G.; Ren, P. Automation of AMOEBA Polarizable Force Field Parameterization for Small Molecules. *Theor. Chem. Acc.* **2012**, *131* (3), 1–11.

(226) Ponder, J. TINKER: Software Tools for Molecular Design. *Washington University School of Medicine, St. Louis, MO*. **2001.**

(227) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A. Gaussian 09, Revision A. 02; Gaussian, Inc. *Wallingford, CT* **2009**, *19*, 227–238.

(228) Stone, A. J. GDMA: A Program for Performing Distributed Multipole Analysis of Wave Functions Calculated Using the Gaussian Program System, Version 1.0. *Univ. Cambridge Cambridge, UK* **1999**.

(229) Bradshaw, R. T.; Essex, J. W. Evaluating Parameterization Protocols for Hydration Free

Energy Calculations with the AMOEBA Polarizable Force Field. *J. Chem. Theory Comput.* **2016**, *12* (8), 3871–3883.

(230) Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. Automatic Atom Type and Bond Type Perception in Molecular Mechanical Calculations. *J. Mol. Graph. Model.* **2006**, *25* (2), 247–260.

(231) Jakalian, A.; Bush, B. L.; Jack, D. B.; Bayly, C. I. Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: I. Method. *J. Comput. Chem.* **2000**, *21* (2), 132–146.

(232) Jakalian, A.; Jack, D. B.; Bayly, C. I. Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: II. Parameterization and Validation. *J. Comput. Chem.* **2002**, *23* (16), 1623–1641.

(233) Mohamed, N. A.; Essex, J. W.; Bradshaw, R. T. Underlying data for "Evaluation of solvation free energies for small molecules with the AMOEBA polarizable force field" http://dx.doi.org/10.5281/zenodo.59203.

(234) Shi, Y.; Wu, C.; Ponder, J. W.; Ren, P. Multipole Electrostatics in Hydration Free Energy Calculations. *J. Comput. Chem.* **2011**, *32* (5), 967–977.

(235) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular Dynamics with Coupling to an External Bath. *J. Chem. Phys.* **1984**, *81* (8), 3684–3690.

(236) Nosé, S. A Unified Formulation of the Constant Temperature Molecular Dynamics Methods. *J. Chem. Phys.* **1984**, *81* (1), 511–519.

(237) Hoover, W. G. Canonical Dynamics: Equilibrium Phase-Space Distributions. *Phys. Rev. A* **1985**, *31* (3), 1695–1697.

(238) Case, D. A.; Babin, V.; Berryman, J. T.; Betz, R. M.; Q. Cai; Cerutti, D. S.; Cheatham, T. E.; Darden, T. A.; Duke, R. E.; Gohlke, H.; Goetz, A. W.; Gusarov, S.; Homeyer, N.; Janowski, P.; Kaus, J.; Kolossváry, I.; Kovalenko, A.; Lee, T. S.; LeGrand, S.; Luchko, T.; Luo, R.; Madej, B.; Merz, K. M.; Paesani, F.; Roe, D. R.; Roitberg, A.; Sagui, C.; Salomon-Ferrer, R.; Seabra, G.; Simmerling, C. L.; Smith, W.; Swails, J.; Walker, R. C.; Wang, J.; Wolf, R. M.; Wu, X.; Kollman, P. A. AMBER 16. *AMBER 16*. University of California, San Francisco **2016**.

(239) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Oxford University Press, New York, **1987**.

(240) Darden, T.; York, D.; Pedersen, L. Particle Mesh Ewald: An N Log(N) Method for Ewald

Sums in Large Systems. *J. Chem. Phys.* **1993**, *98* (12), 10089.

(241) Shi, Y.; Wu, C. J.; Ponder, J. W.; Ren, P. Y. Multipole Electrostatics in Hydration Free Energy Calculations. *J. Comput. Chem.* **2011**, *32* (5), 967–977.

(242) Byrd, R. H.; Nocedal, J.; Schnabel, R. B. Representations of Quasi-Newton Matrices and Their Use in Limited Memory Methods. *Math. Program.* **1994**, *63* (1–3), 129–156.

(243) Ren, P. BAR-amber http://biomol.bme.utexas.edu/wiki/index.php/Research_ex:Amber (accessed Jul 29, 2016).

(244) Mobley, D. L.; Wymer, K. L.; Lim, N. M.; Guthrie, J. P. Blind Prediction of Solvation Free Energies from the SAMPL4 Challenge. *J. Comput. Aided. Mol. Des.* **2014**, *28* (3), 135–150.

(245) Marenich, A. V.; Kelly, C. P.; Thompson, J. D.; Hawkins, G. D. .; Chambers, C. C.; Giesen, D. J.; Winget, P.; Cramer, C. J.; Truhlar, D. G. Minnesota Solvation Database-version **2012** http://comp.chem.umn.edu/mnsol/ (accessed Jul 29, 2016).

(246) Nicholls, A.; Mobley, D. L.; Guthrie, J. P.; Chodera, J. D.; Bayly, C. I.; Cooper, M. D.; Pande, V. S. Predicting Small-Molecule Solvation Free Energies: An Informal Blind Test for Computational Chemistry. *J. Med. Chem.* **2008**, *51* (4), 769–779.

(247) Geballe, M. T.; Skillman, A. G.; Nicholls, A.; Guthrie, J. P.; Taylor, P. J. The SAMPL2 Blind Prediction Challenge: Introduction and Overview. *J. Comput. Aided. Mol. Des.* **2010**, *24* (4), 259–279.

(248) Ren, P. Y.; Wu, C.; Ponder, J. W. Polarizable Atomic Multipole-Based Molecular Mechanics for Organic Molecules. *J. Chem. Theory Comput.* **2011**, *7* (10), 3143–3161.

(249) Cieplak, P.; Caldwell, J.; Kollman, P. 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/. *J. Comput. Chem.* **2001**, *22* (10), 1048–1057.

(250) Lide, D. R. *CRC Handbook of Chemistry and Physics, 84th Edition, 2003-2004*; CRC Press, **2003**, Vol. 53.

(251) Zhang, J.; Tuguldur, B.; Van Der Spoel, D. Force Field Benchmark of Organic Liquids. 2. Gibbs Energy of Solvation. *J. Chem. Inf. Model.* **2015**, *55* (6), 1192–1201.

(252) Jämbeck, J. P. M.; Lyubartsev, A. P. Update to the General Amber Force Field for Small

Solutes with an Emphasis on Free Energies of Hydration. *J. Phys. Chem. B* **2014**, *118* (14), 3793–3804.

(253) Fennell, C. J.; Wymer, K. L.; Mobley, D. L. A Fixed-Charge Model for Alcohol Polarization in the Condensed Phase, and Its Role in Small Molecule Hydration. *J. Phys. Chem. B* **2014**, *118* (24), 6438–6446.

(254) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**.

(255) Joung, I. S.; Cheatham, T. E. Determination of Alkali and Halide Monovalent Ion Parameters for Use in Explicitly Solvated Biomolecular Simulations. *J. Phys. Chem. B* **2008**.

(256) Giammona D; D, C.; C., B. Force field modifications for all-atom heme http://research.bmh.manchester.ac.uk/bryce/amber. (accessed Jul 29, 2017).

(257) Banci, L.; Carloni, P.; Savellini, G. G. Molecular Dynamics Studies on Peroxidases: A Structural Model for Horse Radish Peroxidase and a Substrate Adduct. *Biochemistry* **1994**.

(258) Gordon, J. C.; Myers, J. B.; Folta, T.; Shoja, V.; Heath, L. S.; Onufriev, A. H++: A Server for Estimating p K as and Adding Missing Hydrogens to Macromolecules. *Nucleic Acids Res.* **2005**, *33* (suppl_2), W368–W371.

(259) Gunner, M. R.; Zhu, X.; Klein, M. C. MCCE Analysis of the PKas of Introduced Buried Acids and Bases in Staphylococcal Nuclease. *Proteins Struct. Funct. Bioinforma.* **2011**, *79* (12), 3306–3319.

(260) Rosenfeld, R.; Hays, A.; Musah, R. A.; Goodin, D. B. Excision of a Proposed Electron Transfer Pathway in Cytochrome c Peroxidase and Its Replacement by a Ligand-binding Channel. *Protein Sci.* **2002**.

(261) Kastenholz, M. A.; Hünenberger, P. H. Computation of Methodology\hyphen Independent Ionic Solvation Free Energies from Molecular Simulations: I. The Electrostatic Potential in Molecular Liquids. *J. Chem. Phys.* **2006**.

(262) Brooks, B. R.; Brooks, C. L.; Mackerell, A. D.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caflisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoscek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M. CHARMM: The Biomolecular Simulation Program. *J. Comput. Chem.* **2009**, *30*, 1545–1614.

(263) Wu, X.; Clavaguera, C.; Lagardère, L.; Piquemal, J. P.; De La Lande, A. AMOEBA Polarizable Force Field Parameters of the Heme Cofactor in Its Ferrous and Ferric Forms. *J. Chem. Theory Comput.* **2018**. 14(5), 2705-2720.

(264) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A. J.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V; Cioslowski, J.; Fox, D. J. Gaussian 09, Revision D.01. *Gaussian 09, Revision D.01*. Gaussian, Inc.: Wallingford CT 2009.

(265) Coulson, A. F.; Erman, J. E.; Yonetani, T. Studies on Cytochrome c Peroxidase. XVII. Stoichiometry and Mechanism of the Reaction of Compound ES with Donors. *J. Biol. Chem.* **1971**. 246(4), 917-924

(266) Xia, M.; Chai, Z.; Wang, D. Polarizable and Non-Polarizable Force Field Representations of Ferric Cation and Validations. *J. Phys. Chem. B* **2017**. 121(23), 5718-5729.

(267) Semrouni, D.; Isley, W. C.; Clavaguéra, C.; Dognon, J. P.; Cramer, C. J.; Gagliardi, L. Ab Initio Extension of the AMOEBA Polarizable Force Field to Fe 2+. *J. Chem. Theory Comput.* **2013**. 9(7), 3062-3071.

(268) Xiang, J. Y.; Ponder, J. W. A Valence Bond Model for Aqueous Cu(II) and Zn(II) Ions in the AMOEBA Polarizable Force Field. *J. Comput. Chem.* **2013**. 34(9), 739-749.

(269) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A. Gaussian 09, Rev. B. 01. *Gaussian Inc., Wallingford CT*. **2010**.

(270) Oda, A.; Yamaotsu, N.; Hirono, S. New AMBER Force Field Parameters of Heme Iron for Cytochrome P450s Determined by Quantum Chemical Calculations of Simplified Models. *J. Comput. Chem.* **2005**.

(271) Shahrokh, K.; Orendt, A.; Yost, G. S.; Cheatham, T. E. Quantum Mechanically Derived AMBER-Compatible Heme Parameters for Various States of the Cytochrome P450 Catalytic Cycle. *J. Comput. Chem.* **2012**. 33(2), 119-133.

(272) Laury, M. L.; Wang, L. P.; Pande, V. S.; Head-Gordon, T.; Ponder, J. W. Revised Parameters for the AMOEBA Polarizable Atomic Multipole Water Model. *J. Phys. Chem. B* **2015**. 119(29), 9423-9437

(273) Abella, J. R.; Cheng, S. Y.; Wang, Q.; Yang, W.; Ren, P. Hydration Free Energy from Orthogonal Space Random Walk and Polarizable Force Field. *J. Chem. Theory Comput.* **2014**. 10(7), 2792-2801.

(274) Ren, P.; Ponder, J. W. Polarizable Atomic Multipole Water Model for Molecular Mechanics Simulation. *J. Phys. Chem. B* **2003**, *107* (24), 5933–5947.

(275) Bradshaw, R. T.; Essex, J. W. Evaluating Parametrization Protocols for Hydration Free Energy Calculations with the AMOEBA Polarizable Force Field. *J. Chem. Theory Comput.* **2016**. 12(8), 3871-3883.