Computing Free Energy Hypersurfaces for Anisotropic Intermolecular Associations

Johan Strümpfer

Thesis presented for the degree of Master of Science

In the Department of Chemistry
University of Cape Town

June 2009

Abstract

Adaptive reaction coordinate force biasing methods have been previously used for calculating the free energy of conformation and chemical reactions amongst others. Here a generalized method is described that is able to produce free energies in multiple dimensions, descriptively named the free energies from adaptive reaction coordinate forces (FEARCF) method. To illustrate it a multidimensional intermolecular orientational free energy surface is calculated, and it is demonstrated how to investigate complex systems such as protein conformation and liquids. This multidimensional intermolecular free energy $W(r, \theta_1, \theta_2, \phi)$ provides a measure of orientationally dependent interactions that are appropriate for applications in systems that inherently have molecular anisotropic features. It is a highly informative free energy volume, which can be used to parameterize key terms such as the Gay-Berne intermolecular potential in coarse grain simulations.

To demonstrate the value of the information gained from the $W(r, \theta_1, \theta_2, \phi)$ hypersurfaces we calculated them for TIP3P, TIP4P and TIP5P dimer water models in vacuum. A comparison with a commonly used one dimensional distance free energy profile is made to illustrate the significant increase in configurational information. The $W(r)$ plots show little difference between the three models while the $W(r, \theta_1, \theta_2, \phi)$ hypersurfaces reveal the underlying energetic reasons why these potentials reproduce tetrahedrality in the condensed phase so differently from each.

Calculation of the free energy of association for the benzene dimer in aqueous solution and in vacuum is also perfomed. The dependence on the choice of water model to represent the environment is also illustrated. The TIP4P water model is shown to give more accurate results than TIP3P. The dimer association is also shown to be primarily enthalpic in nature and with the $W(r, \theta_1, \theta_2, \phi)$ hypersurface it is shown that the shifted stacked dimer is the unconstrained minimum free energy configuration.

Publications

The Journal of Computational Chemistry accepted chapter 4 of this thesis as an article in May 2009:

Strümpfer, J.; Naidoo. K. J., Computing free energy hypersurfaces for anisotropic intermolecular associations. *Journal of Computational Chemistry.* **2009**, *in press.*

The preliminary results of Chapter 5 and the results of Chapter 4 were also presented at the World Association of Theoretical and Computational Chemists (WATOC) conference in Sydney, Australia in 2008.

Acknowledgements

I am thankful for the financial aid I received from the University of Cape Town, the National Research Foundation, the South African Research Chair Initiative (SARChI) and my parents for funding my studies.

I would like to thank Riedaa Gamieldien and Chris Barnett for their effort and assistance in preparation of this thesis.

Many thanks go to Professor Kevin Naidoo for his guidance, supervision and advice with my research and thesis. Without his effort and assistance this would not have been possible for me.

I would also like to thank my family for their love and wonderful support and last but not least my girlfriend, Louise, who was a great help and support.

1 Introduction

1.1 The Forces Of Nature

All scientific endeavors to this day have provided us with a description of nature using four fundamental forces (in order of decreasing strength): the strong nuclear force, electromagnetic force, weak nuclear force and gravitational force[1]. Many people currently describe nature with three fundamental forces: strong, electroweak and gravitational since the electromagnetic force and the weak force have been unified into one theory. For this thesis, this unification is not relevant and we will continue to refer to them distinctly.

The strong nuclear force is described by the theory of quantum chromodynamics[2] and is responsible for the interactions between the quarks that make up the protons and neutrons in atomic nuclei. The force scales as a constant with distance but acts only on distances of a femtometer or less.

The electromagnetic force is two orders of magnitude weaker than the strong nuclear force and acts between charged species such as electrons and protons. It is described by the theory of quantum electrodynamics and scales with the inverse of distance squared but it can act over nearly infinite distances.

The weak force is thirteen orders of magnitude weaker than the strong nuclear force and is described by a theory, which encompasses quantum electrodynamics, called the electro-weak theory. It has a range much shorter than the strong nuclear force and scales as $\exp(-r)/r$. It acts on particles such as pions, electrons, neutrino and other subatomic particles.

The final of the four fundamental forces is the gravitational force. This force is the weakest of all at 38 orders of magnitude weaker than the strong force. It acts on all

particles over nearly infinite ranges and scales with inverse distance. The theory of general relativity describes the behaviour of objects interacting gravitationally.

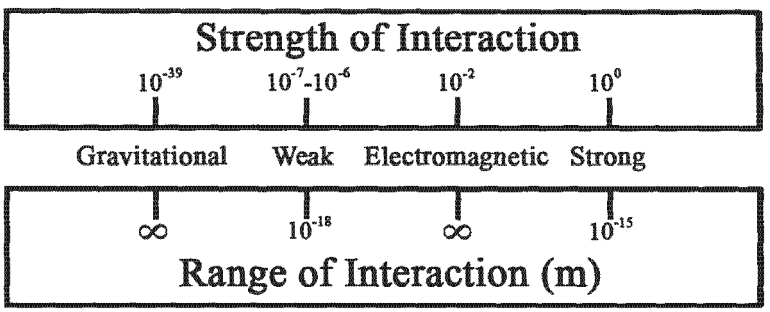

Figure 1-1: The strengths of interactions are taken from the structure constants and the range of interactions from the governing equations.[3]

The comparison of interaction strengths and interaction ranges are shown in Figure 1.1. Of the four fundamental forces, only one force strongly determines the nature of molecular systems: the electromagnetic force[4]. For much larger chemical systems than the ones studied in this thesis, the gravitational force may come into effect, but for all intents and purposes we may ignore all but the electromagnetic force. Atoms are made up of protons, neutrons and electrons. The protons and neutrons are held together by the strong nuclear force to form the atomic nucleus. The atomic nucleus is nearly unbreakable at the energy scales relevant to biological and chemical processes and thus the nucleus can be treated as a single positively charged object.

The atom is pictured as a positively charged nucleus surrounded by a number of negatively charged electrons that are spread out in space as and viewed as an electron density. This allows us to describe all relevant interactions in molecular systems using only the electromagnetic force.

A molecule is made up of a number of atoms that are bound together in a specific configuration. The force holding the molecule together is the electromagnetic force but this can be further categorised based on interaction strength.

The molecular partition function q describes the number of different states that have a given energy E [5]. The energy within a molecule can be distributed into different modes, from which our categorization is derived. The total energy can be written as:

$$E = E_t + E_r + E_v + E_e + E_n, \tag{1.1}$$

where the subscripts t, v, r, e and n stand for translational, vibrational, rotational, electronic and nuclear[4]. In this way the molecular partition function can also be separated according to:

$$q = q_t \times q_r \times q_v \times q_e \times q_n. \tag{1.2}$$

For systems of constant nuclear configuration, as is the case for systems of energies far from relativistic energies, E_n is a constant and for systems that are not electronically excited E_e is a constant. Thus the molecular partition function is primarily dependent on q_t, q_r and q_v.

The translation and rotational energies of molecules are typically much less that $k_B T$, thus these motions can be described classically to a high level of accuracy. The vibrational energies can be very high, depending on the atoms involved. Oxygen-hydrogen bonds, for example, have a very high vibrational frequency thus they have to be described quantum mechanically or constraints have to be placed on them and the energies of the system appropriately modified. Slower vibrational motions, such as the bond stretching between heavier atoms can be described classically.[4, 5]

1.2 Intermolecular Forces

Chemical systems consist not of a single molecule but of a large number of molecules. To describe many molecules we need to now enlarge the description to include intermolecular interactions. In doing so we can construct the partition

function from two components. The first component is the molecular partition function q (described above), which determined by the intramolecular interactions. The second component Z is the configurational partition function, which is determined by the intermolecular interactions.[4]

We have established that only the theory of quantum electrodynamics is required for calculations in typical chemical systems that we are interested in. The theory of quantum electrodynamics was developed between 1900 and 1950 and is very well tested[6, 7] but it remains too intricate to use on systems at the scale that we are interested in (tens of Ångstrom to nanometers and tens of thousands of atoms). Approximations must thus be introduced to perform computations on molecular systems of interest in chemistry and biophysics.

The first of these that can be made is to use reduced descriptions of the quantum systems such as the Hartree-Fock[8] based methods and density functional theory[9]. This description is accurate and can be done on small molecular systems but the computations remain too time consuming for the size scales and time scales common in biophysics, chemical biology and materials science.

This leads to a need to describe a system of molecules effectively and *classically*. In this description electrodynamics interactions are split into two distinct groups: intermolecular interactions and intramolecular interactions. Intermolecular interactions are to date typically described using atomic pair potentials. These potentials are so called long range since they describe the interactions of non-bonded atoms between different molecules. They are constructed to reproduce experimental measures (e.g. heat capacities, diffusion coefficients) and to be consistent with quantum chemistry calculations of atomic interaction.

The intramolecular interactions are made up of potentials that are parameterised to reproduce experimental measurements and quantum chemistry calculations [10, 11]. This is done so as to give the correct behaviour of the molecules in the simulations. These terms take into account bond vibrations, angle vibrations and dihedral angle vibrations while ensuring that the atoms of the molecules remain together.

The two main parts of the long-range interaction are the electrostatic Coulomb interaction, the van der Waals attraction and nuclear core repulsion, both described by the 6-12 Lennard-Jones potential. The electrostatic component applies to charged species and is the strongest interaction. In molecular dynamics simulations partial charges that arise from non-symmetric electron distributions around the atoms of a molecule are often assigned. This means that even for neutral molecules, there may be non-trivial electrostatic terms. The Lennard-Jones terms describes the weakly attractive dispersion interactions and the repulsive core interactions.

The long-range interactions are described by pairwise potentials. The pair potentials are easy and quick to calculate and by only including pair interactions, one greatly reduces the number of calculations that need to be done for a system. They are parameterised to incorporate not only the two body interactions present in the system, but also to approximate many-body effects. They are thus referred to as effective pair potentials.[10, 11]

For two molecules it is the sum of the effective pair potential terms between the atoms that make up the intermolecular interaction. Although the individual potential terms are isotropic, the sum over all the atoms of the two molecules is not. This is due to the fact that individual molecules are very rarely spherically symmetric. Any asymmetry in a molecule will introduce asymmetries in the interactions of that molecule with others and will hence not be able to be accurately described by isotropic potentials. For larger molecules this develops into a further complication since these molecules can adopt different conformations.

A description of intermolecular interactions must take into account the anisotropy in the interactions[12-22]. Methods to do this include using multipole expansions that can take into account higher order electrostatic terms, analytic two- and three-body functions and analytic functions of one or more variables that take into account relative orientations[13, 15-19, 22-26]. Most of these methods have as their primary aim to obtain a simpler but nonetheless accurate description of orientation dependent potential functions that can be used for larger scale and longer time simulation in the study of protein folding for protein structure prediction.

1.3 Thermodynamics of Intermolecular Forces

The interactions described above are the direct interactions between molecules arising from the quantum electrodynamics interactions between the atoms that make up the system. There are, however other effective forces that arise from multiple solvent interaction and other bulk effects and also averaged effects. To describe such effects we turn to the language of the physics of large numbers called thermodynamics.

The atoms of the molecules are not fixed in place but vibrate and move around. This means that the interaction between two molecules is also a function of the configurations of atomic positions of the two molecules. These fluctuations are controlled primarily by the intramolecular interaction potential and for smaller molecules are not very large. The total configurational space that the molecules may sample is large, but they do not sample all points uniformly. A reduced description of the intermolecular interaction could be given for only the most commonly sampled configurations. This would still require a complicated description of the interaction and thus a more common method is to average out the fluctuations with an appropriate bias to the areas of the configurational space sampled most often.

Another important effect on the interactions between molecules is the presence of other molecules (e.g. solute, solvent, ions, etc.) in the system. The molecules may also undergo similar 'internal' fluctuations and move about the molecules of interest. This greatly increases the configurational space that needs to be taken into account when considering intermolecular interactions. These forces can greatly affect the interactions between molecules. An example of this is the process of protein folding. The configurational space of a protein is very large and it would take immensely long for a protein to fold using only a conformational search. Solvent forces become a great boon here as they initiate the process and restrict the configurational and conformational space of the protein.

The effect of solvent interaction can be either direct/enthalpic by assisting the association of two molecules via solvent bridges or indirect/entropic. To maximise the number of available states of the solvent often implies greatly reducing the

configurational space of the associating molecules. This can drive molecules together in instances where they otherwise would not associate.

An example of this is the case of protein folding. The conformational space of a protein is extremely large. It has been shown that if a protein were to fold by an undirected search of its available conformational space it would take longer than the age of the universe to fold, this is known as Levinthal's paradox[27]. Since proteins fold much faster there has to be a directed search that takes place. It was demonstrated that by introducing only a small bias into the folding process, the folding time is significantly reduced to realistic folding times[28].

It is generally accepted now that proteins fold on a funnel-like energy landscape with the result that the folding process can be described as a two-state process. The initial state, however, corresponds to a large number of conformational states and the final state corresponds to a single or very few conformational states. Since proteins can be long chains of amino acids, the specific amino acid interactions are not sufficient to fold a protein. It has been shown that solvent effects are essential to establish amino acid associations that assist in the folding process[28-31].

To take into account both enthalpic and entropic effects we need to look at the free energy. The mean force acting on a molecule will take into account configurationally averaged effects of the molecule and its environment and the potential of mean force is exactly the free energy. By calculating the interaction free energy between molecules we can calculate the effective force that one molecule exerts on another.

Calculating free energy surfaces is not a trivial thing to do from molecular dynamics simulations. Various methods such as free energy perturbation, thermodynamic integration and umbrella sampling have been developed to overcome the difficulties involved[10, 32-35]. These techniques have been employed to study the interactions in many systems ranging from many different kinds of proteins[12, 16-18, 21, 33, 36-41] to simple organic molecules[42-46] to solvent molecules[46-50].

1.4 Hydrogen Bonds, Hydrophobic and Hydrophilic interactions

A system of particular interest in chemistry and biology is that of water. It is a pervasive substance in our bodies and on the earth. It is in a primarily water environment in which many of the processes that keep us alive takes place. It is also one of the most anomalous materials in having for example very high melting and boiling points. These properties are primarily due to the associations between water molecules.

A large factor in the association between water molecules is hydrogen bonding[51]. A hydrogen bond occurs when a hydrogen atom is strongly attracted to two other atoms. In water this is seen when the hydrogen forming a covalent bond with the oxygen atom of a water molecule is strongly attracted to the oxygen atom of another water molecule. The nature of this attraction is primarily electrostatic[52] thus it can be described by the intermolecular potentials used in molecular dynamic simulations. It does, however also contain a covalent part meaning that there is a directionality of the interaction, which if not taken explicitly into account, will mean that molecular dynamics simulations will not correctly describe hydrogen bonding. It is due to hydrogen bonding that we observe the tetrahedral structure of ice and near tetrahedral structures of liquid water.

To model water in simulations a number of models have been developed. These models include all-atoms approximations, three-, four- and five-site models, polarisable models and flexible models[10, 14, 20, 50, 53-58]. The models vary in their local and bulk properties but all produce fairly reasonable results for a wide range of systems[58, 59]. Although the importance of treating water explicitly and with high detail is well understood and is becoming more prominent, most molecular dynamics simulations use fixed charge and rigid three- or four-site models[58, 59].

As mentioned above, various effective forces are experienced by molecules that interact via electrostatic and van der Waals potentials due to the presence of other molecules. A particularly prominent water solvent effect, related to hydrogen bonding

(discussed above) is the hydrophobic effect. This effect is intuitively associated with the "demixing under standard conditions of oil-like materials from aqueous solutions"[60].

The current understanding of the hydrophobic effect it that a polar solvent, when surrounding a non-polar solute, will orient itself and arrange so as to minimise charged contact with the solute. The effect of this is that when there are two non-polar solutes near each other, the water molecules surrounding them will arrange to minimize the contact area and in doing so form an elastic hydrogen bond network surrounding the solute molecules[60, 61]. The effect is that the two solutes are pushed closer together and decrease in entropy, while the aqueous environment increases in entropy. The hydrophobic effect is considered to be the initial driving force behind protein folding and is responsible for micelle and lipid bilayer formation[37, 60-64].

As soon as the hydrophobic effect was discovered it was postulated that there might be a corresponding hydrophilic effect. The hydrophilic effect is considered to be the reverse hydrophobic effect. Instead of trying to minimize non-polar and polar contact, the system may be maximising polar-polar contact. This was proposed[37] and disputed[65] as being a major contributor to the protein folding process.

1.5 Molecular Association in Biology and Materials

The effect of solvation can greatly affect the intermolecular forces in any system. This is something that is clearly seen in all systems ranging from charged ion solutions to uncharged biopolymers.

Liquid crystals are a form of ordered fluids, such as lipid bilayers, micelles and those found in liquid crystal displays. They form by identical, repeated intermolecular interactions, which in theoretical and computations studies of these systems have typically been represented by a superposition of the isotropic atomic interactions[66]. It has, however, been shown that the intermolecular potentials derived from the isotropic interactions are insufficient to reproduce the crystal packing and thermal and vibrational properties of liquid crystal systems[67, 68].

Protein folding has been mentioned a number of times in this introduction and for good reason; protein folding and protein structure prediction is one of the most prominent directions of biomolecular research currently and in the past. It has been shown that anisotropy is of critical importance in describing amino acid associations for the purpose of assigning native and non-native protein conformations[12, 16-18, 26]. Aside from this it has also been shown that intermolecular potentials based solely on atomic electrostatic and van der Waals interactions are not sufficient to describe amino acid interactions for the purpose of protein folding[26, 31, 37, 62-65, 69].

1.6 Objectives and Extent of Thesis

The objective of this thesis is to design a multidimensional free energy method that can provide a solution to a number of chemical biological and physical problems. Here the fundamental anisotropic nature of molecular association has been shown to be the key to understanding many grand challenge problems such as protein folding and solvation forces.

In chapter 2 the groundwork for the methods that are employed is established. This includes a description of molecular dynamics simulations, how they are run and what protocols are involved to ensure relevant and efficient computations. Some of the basic statistical analysis tools used in molecular dynamics simulations are also introduced.

Chapter 3 continues the thesis with a general discussion of statistical mechanics. In this chapter the challenge of calculating free energies is elucidated as well as many of the methods that are used to overcome these challenges. The method that is developed and used in this thesis is the Free Energies from Adaptive Reaction Coordinate Forces (FEARCF) method. This is described in detail along with its implementation and details about its usage.

Chapter 4 presents the first set of results as a published journal article. In this chapter the methods developed in the previous two chapters are applied to the case of the water dimer for three similar water models. It is shown that even in the case of very similar models, the FEARCF method is sensitive enough to distinguish between the models where typical analysis methods cannot.

Chapter 5 presents the applications of the method to the case of two interacting benzene molecules in an aqueous environment. It is shown that the FEARCF method can successfully applied in solution and elucidates the minimum free energy conformation of the benzene dimer.

1.7 Chapter One References

1. Feynman, R. P., *The Character of Physical Law*. MIT Press: Boston, Mass., 1967.
2. Greiner, W.; Schäfer, A., *Quantum Chromodynamics*. Springer-Verlag: Berlin, 1994.
3. Rohlf, J. W., *Modern Physics from alpha to Z0*. John Wiley & Sons: New York, 1994.
4. Hinchliffe, A., *Chemical Modelling: From Atoms to Liquids*. John Wiley & Sons: New York, 1999.
5. McQuarrie, D. A., *Statistical Mechanics*. Harper & Row: New York, 1975.
6. Feynman, R. P., *QED: The strange theory of light and matter*. Princeton University Press: 1985.
7. Feynman, R. P., *Quantum electrodynamics*. Addison-Wesley: Reading, Mass., 1998.
8. Fischer, C. F., *The Hartree-Fock Method for Atoms: A Numerical Approach*. John Wiley & Sons: New York, 1977.
9. Parr, R. G.; Yang, W., *Density Functional Theory of Atoms and Molecules*. Oxford University Press: Oxford, 1989.
10. Allen, M. P.; Tildesley, D. J., *Computer Simulation of Liquids*. Oxford University Press: 1987.
11. Leach, A. R., *Molecular Modelling Principles and Applications*. Addison Wesley Longman Limited: Edingburgh Gate, Harlow, 1996.
12. Bartels, C.; Karplus, M., Multidimensional adaptive umbrella sampling: Applications to main chain and side chain peptide conformations. *J. Comput. Chem.* **1997,** 18, (12), 1450-1462.
13. Buchete, N. V.; Straub, J. E.; Thirumalai, D., Orientation-dependent coarse-grained potentials derived by statistical analysis of molecular structural databases. *Polymer* **2004,** 45, (2), 597-608.
14. Chowdhuri, S.; Tan, M.-L.; Ichiye, T., Dynamical properties of the soft sticky dipole-quadrupole-octupole water model: A molecular dynamics study. *J. Chem. Phys.* **2006,** 125, (14), 144513-8.
15. Laaksonen, A.; Stilbs, P., Molecular motion and solvation of benzene in water, carbon tetrachloride, carbon disulfide and. *J. Chem. Phys.* **1998,** 108, (2), 455-455.
16. Makowski, M.; Chmurzynski, L., Potentials of Mean Force of Two Hydrophobic Amino-Acid Side Chain Models Dependent on Orientation. *AIP Conf. Proc.* **2007,** 963, (2), 1282-1284.
17. Miyazawa, S.; Jernigan, R. L., How effective for fold recognition is a potential of mean force that includes relative orientations between contacting residues in proteins? *J. Chem. Phys.* **2005,** 122, (2), 024901-18.
18. Mukherjee, A.; Bhimalapuram, P.; Bagchi, B., Orientation-dependent potential of mean force for protein folding. *J. Chem. Phys.* **2005,** 123, (1), 014901-11.
19. Paramonov, L.; Yaliraki, S. N., The directional contact distance of two ellipsoids: Coarse-grained potentials for anisotropic interactions. *J. Chem. Phys.* **2005,** 123, (19), 194111-11.
20. Wallqvist, A.; Berne, B. J., Effective potentials for liquid water using polarizable and nonpolarizable models. *J. Phys. Chem.* **1993,** 97, (51), 13841-13851.

21. Zhang, C.; Liu, S.; Zhou, H.; Zhou, Y., An accurate, residue-level, pair potential of mean force for folding and binding based on the distance-scaled, ideal-gas reference state. *Prot. Sci.* **2004,** 13, (2), 400-411.
22. Engkvist, O.; Åstrand, P.-O.; Karlstrom, G., Accurate Intermolecular Potentials Obtained from Molecular Wave Functions: Bridging the Gap between Quantum Chemistry and Molecular Simulations. *Chem. Rev.* **2000,** 100, (11), 4087-4108.
23. Gay, J. G.; Berne, B. J., Gay-Berne Potential. *J. Chem. Phys* **1981,** 74.
24. Skolnick, J., In quest of an empirical potential for protein structure prediction. *Curr. Opin. Struct. Biol.* **2006,** 16, (2), 166-171.
25. Buchete, N. V.; Straub, J. E.; Thirumalai, D., Anisotropic coarse-grained statistical potentials improve the ability to identify nativelike protein structures. *J. Chem. Phys* **2003,** 118, (16), 7658-7671.
26. Liwo, A.; lstrok, S. O.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Scheraga, H. A., A united-residue force field for off-lattice protein-structure simulations. I. Functional forms and parameters of long-range side-chain interaction potentials from protein crystal data. *J. Comput. Chem.* **1997,** 18, (7), 849-873.
27. Levinthal, C. In *Mossbauer Spectroscopy in Biological Systems*, Proceedings of a Meeting held at Allerton House, Monticello, IL, 1969; Debrunner, P.; Tsibris, J. C. M.; Münck, E., Eds. University of Illios Press: Monticello, IL, 1969; pp 22-24.
28. Zwanzig, R.; Szabo, A.; Bagchi, B., Levinthal's Paradox. *Proc. Natl. Acad. Sci.* **1992,** 89, 20-22.
29. Clark, P. L., Protein Folding in the Cell: Reshaping the Folding Funnel. *Trends Biochem. Sci.* **2004,** 29, (10), 527-534.
30. Dobson, C. M., Principles of protein folding, misfolding and aggregation *Semin. Cell Devel. Biol.* **2004,** 15, (1), 3-16.
31. Xu, D.; Lin, S. L.; Nussinov, R., Protein binding versus protein folding: the role of hydrophilic bridges in protein associations. *J. Mol. Biol* **1997,** 265, (1), 68-84.
32. Torrie, G. M.; Valleau, J. P., Nonphysical Sampling Distributions in Monte Carlo Free-Energy Estimation: Umbrella Sampling *J. Comp. Phys.* **1977,** 23, 187-199.
33. Trzesniak, D.; Kunz, A.-P. E.; van Gunsteren, A.-P. E., A Comparison of Methods to Compute the Potential of Mean Force. *ChemPhysChem* **2006,** 8, (1), 162-169.
34. Bash, P. A.; Singh, U. C.; Langridge, R.; Kollman, P. A., Free energy calculations by computer simulation. *Science* **1987,** 236, (4801), 564-568.
35. Mezei, M.; Beveridge, D. L., Free Energy Simulations. *Ann. N. Y. Acad. Sci.* **1986,** 482, (1 Computer Simulation of Chemical and Biomolecular Systems), 1-23.
36. Fixman, M., Classical Statistical Mechanics of Constraints: A Theorem and Application to Polymers. *Proc. Natl. Acad. Sci.* **1974,** 71, (8), 3050-3053
37. Ben-Naim, A., Solvent effects on protein association and protein folding. *Biopolymers* **1990,** 29, (3).
38. Kollman, P., Free energy calculations: Applications to chemical and biochemical phenomena. *Chem. Rev.* **1993,** 93, (7), 2395-2417.
39. Bartels, C.; Karplus, M., Probability Distributions for Complex Systems: Adaptive Umbrella Sampling of the Potential Energy. *J. Phys. Chem. B* **1998,** 102, (5), 865-880.
40. Kumar, S.; Payne, P. W.; Vasquez, M., Method for free-energy calculations using iterative techniques. *J. Comput. Chem.* **1996,** 17, (10), 1269-1275.

58. Wallqvist, A.; Mountain, R. D., Molecular models of water: Derivation and description. *Rev. Comp. Chem.* **1999,** 13, 183-247.
59. Jorgensen, W. L.; Tirado-Rives, J., Potential energy functions for atomic-level simulations of water and organic and biomolecular systems. *Proc. Natl. Acad. Sci.* **2005,** 102, 6665-6670.
60. Pratt, L. R., Molecular Theory of Hydrophobic Effects: "She is too mean to have her name repeated." *Annu. Rev. Chem. Phys.* **2006,** 53, 409-436.
61. Rao, B. G.; Singh, U. C., Hydrophobic Hydration: A Free Energy Perturbation Study. *J. Am. Chem. Soc.* **1989,** 111, 3125-3133.
62. Nicholls, A.; Sharp, K. A.; Honig, B., Protein folding and association: Insights from the interfacial and thermodynamic properties of hydrocarbon. *Proteins: Struct., Func., and Bioinf.* **1999,** 11, (4), 281-296.
63. Spolar, R. S.; Ha, J. H.; Record, M. T., Hydrophobic effect in protein folding and other noncovalent processes involving proteins. *Proc. Natl. Acad. Sci.* **1989,** 86, (21), 8382-8385.
64. Tsai, C. J.; Lin, S. L.; Wolfson, H. J.; Nussinov, R., Studies of protein-protein interfaces: a statistical analysis of the hydrophobic effect. *Prot. Sci.* **1997,** 6, (1), 53-53.
65. Sun, Y.; Kollman, P., Are There Water-Bridge-Induced Hydrophilic Interactions? *J. Phys. Chem* **1996,** 100, (16), 6760-6763.
66. Gavezzotti, A., *Theoretical Aspects and Computer Modelling of the Molecular Solid State.* John Wiley & Sons: Chichester, 1997.
67. Munowitz, M. G.; Wheeler, G. L., A critical evaluation of isotropic potential functions for chlorine. *Mol. Phys.* **1977,** 71.
68. Day, G. M.; Price, S. L., A Nonempirical Anisotropic Atom-Atom Model Potential for Chlorobenzene Crystals. *J. Am. Chem. Sci.* **2003,** 125, (52), 16434-16443.
69. Burley, S. K.; Petsko, G. A., Aromatic-aromatic interaction: a mechanism of protein structure stabilization. *Science* **1985,** 229, (4708).

2 Simulation Analysis and Protocols

2.1 Equilibrium Molecular Dynamics

Every probe of nature is done with some tool and accompanying methods. The tool used in this work is a computer and the methods are those of simulation of equilibrium molecular dynamics. Using the techniques described in this chapter and in chapter 3, the energies of interaction between molecules at equilibrium conditions can be computationally determined. To do this meaningfully, that is to obtain results that are useful and comparable with experiment, a high level of accuracy is needed.

The highest accuracy in molecular dynamics simulations comes from using the highest level of theory. This would mean employing the theories that have been found using reductionist approaches and working on the smallest scale possible to make the largest scale predictions. Chemistry is at its heart determined by building blocks of nature, the atoms. Atomic interactions, in particular, are responsible for all chemistry that we investigate today. Atomic interactions are made up of electron and nuclei interactions, which are very well described by quantum mechanics.

Simulations that are run using quantum mechanical interactions are called *ab initio* simulations. For systems incorporating at more than 100 atoms, it becomes too computationally expensive to use *ab initio* methods for long simulations.[1] There is thus an inverse relationship between high accuracy simulations and long sampling times.

We thus have to increase the smallest length scale used in the description of the system and make approximations for the quantum effects that are observed. These approximations will allow us sufficient accuracy at long enough simulation times to obtain the required sampling. This is done using empirical and semi-empirical force fields that determine the atomic interactions based on atom positions and not on electron or charge movement. The molecular force-field simulation package that was

used in this thesis was the Chemistry at HARvard Molecular Mechanics (CHARMM) version 33b2 package.[2]

In this chapter we present the most important details of the molecular dynamics simulations that were employed in this thesis. Included are descriptions of the assumptions of the methods, the potential energy function, the integration method and the methods employed to assist fast computations. These are all related to the technical concerns of molecular dynamics simulation. To ensure physical relevance and accurate descriptions that match experiments, the methods that ensure correct ensemble sampling and statistical mechanics are also discussed.

2.1.1 Molecular Dynamics

Molecular dynamics simulations based on the movement of atomic nuclei are employed, instead of quantum mechanical simulations incorporating the electron movement, to make it computationally viable to run multi nanosecond simulations with high accuracy. Due to the high ratio of nuclear and electronic masses, the Born-Oppenheimer approximation makes the valid assumption that the movements of electrons are much faster than that of the nuclei.[3] By employing this approximation, the electron's contribution to the momentum of the atom is averaged out. This allows us to describe the system classically with high precision. The molecular dynamics simulations describing the motion of the atoms are run using Hamiltonian dynamics. The Hamiltonian of a system can be written generally as:

$$H\left(\mathbf{q}^N,\mathbf{p}^N,t\right)=\sum_{i=1}^{N}\frac{\mathbf{p}_i^2}{2m_i}+V\left(\mathbf{q}^N,\mathbf{p}^N,t\right), \tag{2.1}$$

for N generalised coordinates and momenta, $\mathbf{q}^N$ and $\mathbf{p}^N$, and $V\left(\mathbf{q}^N,\mathbf{p}^N,t\right)$ is the potential energy of the system. For unmodified dynamics in a stationary reference frame the Hamiltonian describing the system is then also equal to the total energy of

the system $H(\mathbf{q}^N, \mathbf{p}^N, t) = E$. In molecular dynamics simulations the potential energy is a function only of the position $\mathbf{r}^N$ hence the Hamiltonian becomes:

$$H(\mathbf{r}^N, \mathbf{p}^N) = \sum_{i=1}^{N} \frac{\mathbf{p}_i^2}{2m_i} + V(\mathbf{r}^N), \tag{2.2}$$

where $\mathbf{r}_i$ is the 3D position of atom *i* and $\mathbf{p}_i$ is the corresponding 3D momentum. The equations of motion for Hamiltonian dynamics is given by the following two differential equations:

$$\dot{\mathbf{p}}_i = -\frac{\partial H}{\partial \mathbf{r}_i} = -\nabla_i V(\mathbf{r}^N) = \mathbf{F} \tag{2.3}$$

$$\dot{\mathbf{r}}_i = -\frac{\partial H}{\partial \mathbf{p}_i} = \frac{\mathbf{p}_i}{m_i} \tag{2.4}$$

By integrating these equations we can calculate the time evolution of the system and calculate its properties.

2.1.2 Interaction Potential

At the core of a molecular dynamics simulation is the force field, which determines all the dynamic interactions. The potential function V of the CHARMM force field determines the interactions between the atoms that make up the molecules in the simulation. The typical molecular dynamics potential consists of two parts: the non-bonded part $V_{Non-Bonded}$ and the bonded part V_{Bonded}.

The non-bonded part of the potential function (§2.1.2.1) is made up from the interactions between atoms that are not directly bonded to each other and it is present in all dynamics simulations of particles. The bonded part (§2.1.2.2), comprised of the interactions of atoms that are bonded to each other in a molecule, is specific to molecular dynamics simulations and ensures the correct behaviour of the atoms in a single molecule such that the molecule does not distort or break apart.

The interaction potentials also have to be constructed such that they reproduce the correct long-range interactions between molecules. This means taking into account many interactions between atoms in and also avoiding edge effects (§2.1.2.3). In computing the interaction potential, long-range interactions are taken into account in an approximate method (§2.1.2.4) to reduce the computational cost of the simulation.

From the potential function, the inter-atomic forces have to be calculated so that they can be used to calculate the trajectories of the atoms in the system. An example of this calculation is given in §2.1.2.5.

2.1.2.1 Inter-Molecular Potential

The non-bonded potential is a function of the atomic positions $\mathbf{r}$ of the atoms that are separated by 4 or more bonds within a molecule or are part of different molecules. The non-bonded potential can be expanded into terms consisting of single, pair, triplet and higher order interactions:

$$V_{Non-Bonded} = \sum_i V_1(\mathbf{r}_i) + \sum_i \sum_{j>i} V_2(\mathbf{r}_i, \mathbf{r}_j) + \sum_i \sum_{j>i} \sum_{k>j} V_3(\mathbf{r}_i, \mathbf{r}_j, \mathbf{r}_k) + \cdots \qquad (2.5)$$

The first of the non-bonded interaction terms accounts for the effect of an external field on the atoms in the simulation, which is typically not present in MD simulations. The second set of terms is known as the pair potential and is only a function of the separation $r_{ij} = |\mathbf{r}_i - \mathbf{r}_j|$ between atoms i and j. The triplet interaction can account for up to 10% of the non-bonded potential energy in liquids but it and higher order terms are usually excluded since their calculation are typically very time-consuming[4]. The non-bonded interaction may thus be written as:

$$V_{nb} = \sum_i \sum_{j>i} V_2(r_{ij}) \qquad (2.6)$$

The pair potential is further split into two components to take care of the quantum mechanical interaction of the electron clouds and the nuclei between the atoms as well as the electrostatic interaction from the net charges on the atoms.

The Lennard-Jones potential, with its strong short-range repulsion and weak long-range attraction, accounts for the quantum mechanical interaction between atoms. It has the form:

$$V_{LJ}\left(r_{ij}\right)=\varepsilon\left[\left(\frac{R_{\min,ij}}{r_{ij}}\right)^{12}-2\left(\frac{R_{\min,ij}}{r_{ij}}\right)^{6}\right]. \tag{2.7}$$

The parameters of the Lennard-Jones potential are ε, the well depth of the potential, and $R_{\min,ij}$, the separation distance at the minimum of the potential. To take care of the electrostatic interactions from the net atomic charges there is a Coulomb potential:

$$V_C\left(r_{ij}\right)=\frac{1}{4\pi\varepsilon}\frac{q_i q_j}{r_{ij}}. \tag{2.8}$$

The total non-bonded interaction is the sum of the Lennard-Jones terms and the Coulomb term $V_2 = V_{LJ} + V_C$.

2.1.2.2 Intra-Molecular Potential

To account for the interactions of atoms that are bonded to each other or near each other in the same molecule a different interaction potential is used. The bonded potential energy function that is used in CHARMM is:

$$\begin{aligned}V_b = &\sum_{\text{bonds}} k_b\left(b-b_0\right)^2+\sum_{\text{angles}} k_\theta\left(\theta-\theta_0\right)^2+\sum_{\text{dihedrals}} k_\phi\left[1+\cos\left(n\phi-\delta\right)\right]\\ &+\sum_{\text{improper dihedrals}} k_\omega\left(\omega-\omega_0\right)^2+\sum_{\text{Urey-Bradley}} k_u\left(u-u_0\right)^2\end{aligned}. \tag{2.9}$$

In the bonded potential function the five sets of terms account for the interaction energy of bonds, angles, dihedrals, improper dihedrals and Urey-Bradley interactions. An improper dihedral angle is defined in the same way as a dihedral angle, however the axis of rotation about which it is defined is not along a bond. As an example, an improper dihedral angle can be defined in a molecule of methane where no dihedral angle can.

The first term

$$\sum_{\text{bonds}} k_b (b - b_0)^2, \tag{2.10}$$

accounts for the potential energy of the bonds in the system. The two parameters, k_b and b_0, are the force constant and the equilibrium position for the vibrations of the bond between two bonded atoms in a molecule.

The second term

$$\sum_{\text{angles}} k_\theta (\theta - \theta_0)^2, \tag{2.11}$$

accounts for the potential energy of the bond angles in the molecule. The function behaves similarly to the first term: k_θ is the force constant for the vibrational motion of the angle and θ_0 is the equilibrium position.

The third term

$$\sum_{\text{dihedrals}} k_\phi [1 + \cos(n\phi - \delta)], \tag{2.12}$$

accounts for the potential energy of the dihedral angles in the molecules in the simulation, where k_ϕ is the force constant, $n = 1, 2, 3$ is the multiplicity, δ is the phase shift and ϕ is the dihedral angle. This takes care of the 1-4 bonded interactions in the molecule with rotations of the dihedral angle about the central bond.

The fourth term

$$\sum_{\text{improper dihedrals}} k_{\omega}(\omega-\omega_0)^2, \tag{2.13}$$

accounts for the potential energy of the bending of improper dihedral angles. In analogy to the second and first terms, k_{ω} is the force constant, ω_0 the equilibrium angle and ω the improper dihedral angle.

The fifth, Urey-Bradley, term

$$\sum_{\text{Urey-Bradley}} k_u(u-u_0)^2, \tag{2.14}$$

takes into account the potential energy of the 1-3 non-bonded interactions based on their distance apart. It is a harmonic term with k_u being the force constant, u_0 the equilibrium position apart and u the current distance between the atoms separated by two bonds.

The force fields are parameterised to give experimentally observed behaviours in well-known systems. The assumption in the force-field method is that the force fields, that are parameterised to give experimentally verified results for certain systems, can provide us with accurate results for other systems.

A common computation saving technique in many force fields is to remove some atomic details by grouping together functional groups as a single 'atom'. This can greatly increase the simulation time if there are many such united atom groups used, but there is also a substantial loss in detail. For the high accuracy simulations taking into account solvent effects and single atom motions in the long-range interaction, only all atom parameterisations of the CHARMM force field were used.

2.1.2.3 Boundary Conditions

To run a simulation in a cubic box would generate edge effects since there would be no interactions at the edge of a box. In a typical experimental container these effects are present but to simulate a typical experimental container, however, we would have to simulate on the order of 10^{26} atoms. This is unrealistic with current technology so as a solution periodic boundary conditions are used to emulate very large systems.

The periodic boundaries are created by surrounding the box with replicas of itself. Thus, when an atom or molecule leaves the one side of the box, it enters on the opposite face (shown in Figure 2-1). The interactions that are primarily affected by the boundary conditions are the long-range non-bonded interactions. It is thus important that the box size is large enough such that artificial interactions are not introduced. An example of this would be in simulating the interaction of two carbohydrates in a water box. If the box size is too small, then each carbohydrate will interact with multiple copies of the other carbohydrate and itself. If the aim of the simulation is only to study the single carbohydrate - carbohydrate interaction, then using a box size that is too small would produce erroneous results.

Using this method one can have a large number of pair interactions, though calculating them all would be too time consuming to be useful since beyond a certain distance the strength of these interactions would become negligible. To circumvent this problem, a minimum image convention is introduced.

Each replica image surrounding the box is used to calculate the minimum separation between atoms and only those less than a certain cut-off r_{cut} is considered when calculating the interactions. The cut-off distance r_{cut} is chosen to be half the box length thereby ensuring that atoms do not interact with their own images.

This reduces the number of computations performed in running the simulation, but creates a new problem that all the atoms need to be checked whether they are within the cutoff range. To counteract this Verlet proposed that a list is created that stores all the atoms that are near the cutoff range. This list is checked at every step to calculate

which atoms are within the cutoff distance. Every few steps this list must be updated to keep track of the neighbouring atoms.

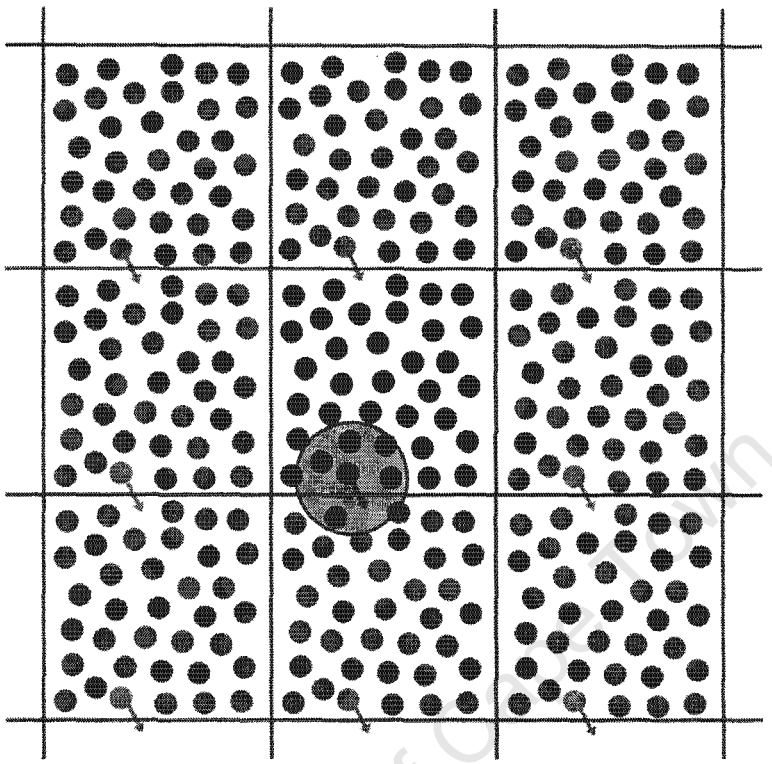

Figure 2-1: The effect of the periodic boundary conditions is depicted by surrounding the simulated cell (in the centre) by identical copies. As the central particle in the circle exits the central cell from the bottom, the particle re-enters it from the top. The minimum image convention is also shown where all particles in the circle are used to calculate non-bonded interactions.

2.1.2.4 Potential Cut-Offs

The minimum image convention reduces the number of interactions to calculate in the system but it is still $O(N^2)$. To further reduce the computational effort of the force calculation a truncation of the van der Waals interaction is performed. At a distance of $2.5R_{\min,ij}$, the Lennard-Jones potential drops to 1% of its value at $R_{\min,ij}$. Thus beyond a certain distance the effects of the potential become negligible. Simple truncation of the forces, however, is unphysical and introduces discontinuities in the force, which

destroys the energy conservation in the system and leads to unphysical behaviour. The solution is to use a switching function

The switching function is a smoothing polynomial, which smoothly brings the potential to zero. The form of the function is given in equation (2.15). The effect of the switching function is an effective truncation of long-range interactions at r_{off} and a smooth transition to zero between r_{on} and r_{off} without introduction discontinuities into the force.

$$V_{switched} = \begin{cases} V(r_{ij}) & r_{ij} < r_{on} \\ V(r_{ij})\left[\dfrac{\left(r_{off}^2 - r_{ij}^2\right)^2\left(r_{off}^2 + 2r_{ij}^2 - 3r_{on}^2\right)}{\left(r_{off}^2 - r_{on}^2\right)^3}\right] & r_{on} < r_{ij} \le r_{off} \\ 0 & r_{ij} > r_{off} \end{cases} \quad (2.15)$$

The simplest implementation of the switching function it to use it between all atoms in the neighbourhood list. To improve the energy conservation and dynamics, however, the switching function should rather be used with neutral groups of atoms. This means that atoms within each molecule should be grouped into neutral groups (when setting up the molecules) and the distance between the centres of masses of the groups is used to switch the potential to zero.

2.1.2.5 Force Calculation

To calculate the forces that arise from an interaction potential function we take the partial derivative of the potential with respect to the variables that parameterise the function. i.e.

$$F(\xi) = -\nabla_{\xi} V(\xi). \quad (2.16)$$

This, however, gives the forces in the ξ directions, which may not necessarily be in Cartesian coordinates, which the atomic positions and velocities are stored in. Thus

the forces from the potential functions have to be converted into forces on the Cartesian coordinates of the atoms.

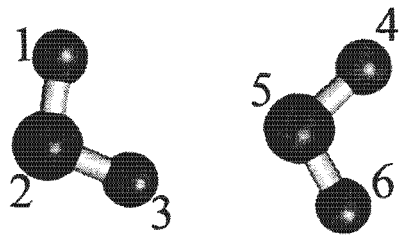

Figure 2-2: Two interacting water molecules with atoms labelled 1-6. The interaction terms that arise in this situation are listed in Table 1.

As an example, we calculate the forces on the atoms of two interacting water molecules shown in Figure 2-2. In Table 1 all the terms for this interaction are listed.

Table 1: Force field terms from two interacting water molecules

Force Field Term	Atoms involved
Bond	1-2, 2-3, 4-5, 5-6
Angle	1-2-3, 4-5-6
Coloumb	1-4, 1-5, 1-6, 2-4, 2-5, 2-6, 3-4, 3-5, 3-6
Lennard-Jones	1-4, 1-5, 1-6, 2-4, 2-5, 2-6, 3-4, 3-5, 3-6

For a distance $\xi = r_{ij}$ between atom i and j, such as the case for the bond, Coloumb and Lennard-Jones terms and if the vector from i to j is $\mathbf{r}_{ij} = r_{ij}\hat{\mathbf{r}}_{ij}$, then the force on atom i is

$$\mathbf{F}_i = -\frac{\partial V}{\partial r_{ij}}\hat{\mathbf{r}}_{ij}, \tag{2.17}$$

and the force on atom j is $\mathbf{F}_j = -\mathbf{F}_i$. These forces are then in Cartesian coordinates and can directly be applied to the atoms i and j. If the force being calculated is due to a Lennard-Jones term, the potential cut-offs should also be taken into account to help save force computation time.

For an angle $\xi = \theta_{ijk}$ between three atoms i, j and k then the angle is calculated from

$$\cos\theta_{ijk} = \frac{\mathbf{r}_{ji} \cdot \mathbf{r}_{jk}}{|\mathbf{r}_{ji}||\mathbf{r}_{jk}|}. \tag{2.18}$$

The force on atom i is then

$$\mathbf{F}_i = -\frac{1}{\sin\theta_{ijk}} \frac{\partial V}{\partial \theta_{ijk}} \nabla_{\mathbf{r}_i} \cos\theta_{ijk}. \tag{2.19}$$

The force on atom k is calculated using the same equation (2.19) and inverting the sign of the angle θ_{ijk}. This calculates the forces in Cartesian coordinates that can be applied directly to the atoms i and k.

Once all the forces on all the atoms have been calculated in Cartesian coordinates from the current positions of the atoms in the simulation, the equations of motion must be integrated numerically to update the velocities and positions of the atoms.

2.1.3 Integration Methods

The Hamiltonian describing this many-body system does not have an analytic solution and thus requires the need for computational methods to numerically integrate the equations of motion giving the time evolution of the system. To integrate the dynamics equations, a modification of the velocity Verlet integration method was

employed. The method used was developed by Martyna *et al.*.[5] This section briefly develops and explains the integration method used.

The standard velocity Verlet integration methods [6-8] involve using the following to calculate the values of position and momentum at time $t + \Delta t$ from the position and momentum at time t

$$\mathbf{r}(t+\Delta t) = \mathbf{r}(t) + \dot{\mathbf{r}}(t)\Delta t + \frac{\ddot{\mathbf{r}}(t)}{2}(\Delta t)^2, \tag{2.20}$$

$$\dot{\mathbf{r}}(t+\Delta t) = \dot{\mathbf{r}}(t) + \frac{\ddot{\mathbf{r}}(t)\Delta t + \ddot{\mathbf{r}}(t+\Delta t)}{2}\Delta t. \tag{2.21}$$

The integration method that is employed uses the Liouville time evolution method. The Liouville operator iL is defined as:

$$iL = \sum_{i=1}^{N} \dot{\mathbf{r}}_i \cdot \frac{\partial}{\partial \mathbf{r}_i} + \dot{\mathbf{p}}_i \cdot \frac{\partial}{\partial \mathbf{p}_i} \tag{2.22}$$

The time evolution operator that propagates the system from $t=0$ to $t=t$ is then $\exp(iLt)$. The time evolution of any dynamics variable starting at $\zeta(0)$ is then calculated with:

$$\zeta(t) = \exp(iLt)\zeta(0) = \left[\exp(iL\Delta t)\right]^{N_t} \zeta(0) + O(N_t \Delta t^4), \tag{2.23}$$

where $t = N_t \Delta t$. A useful technique is to split the Liouville operator into two parts: $iL = iL_r + iL_p$ where:

$$iL_p = \sum_{i=1}^{N} \dot{\mathbf{p}}_i \cdot \frac{\partial}{\partial \mathbf{p}_i}, \tag{2.24}$$

$$iL_r = \sum_{i=1}^{N} \dot{\mathbf{r}}_i \cdot \frac{\partial}{\partial \mathbf{r}_i}. \tag{2.25}$$

The time evolution operator is then approximated to:

Certain atomic motions, such as the vibrational motion due to bonds between atoms, in the molecular dynamics may be too fast to model accurately for a chosen Δt (typically hydrogen bonds are treated in this way). In these cases, the motions of the bonded atoms need to be constrained. For a bond with a fixed length b between atoms at $\mathbf{r}_1$ and $\mathbf{r}_2$ we can write the constraint as a function χ:

$$\chi(\mathbf{r}_1, \mathbf{r}_2) = (\mathbf{r}_1 - \mathbf{r}_2)\cdot(\mathbf{r}_1 - \mathbf{r}_2) - b^2 = 0. \tag{2.29}$$

In equation (2.26) we see that the system first propagates the momenta half a step, then the coordinates gets propagated a full step using the newly calculated momenta, and then finally completes the propagation of the momenta for another half time step. The constraint thus has to be applied correctly at each update to the positions and momenta of the two atoms. If $(\bar{p}_1, \bar{p}_2)$ and $(\bar{r}_1, \bar{r}_2)$ are the unconstrained variables then we can define the updated constrained variables as:

$$\mathbf{p}_i\left(t + \frac{\Delta t}{2}\right) = \bar{\mathbf{p}}_i\left(t + \frac{\Delta t}{2}\right) + \lambda \mathbf{g}_i(t) \tag{2.30}$$

$$\mathbf{r}_i(t + \Delta t) = \bar{\mathbf{r}}_i(t + \Delta t) + \lambda \Delta t \mathbf{g}_i(t)/m \tag{2.31}$$

$$\mathbf{p}_i(t + \Delta t) = \bar{\mathbf{p}}_i(t + \Delta t) + \mu \mathbf{g}_i(t + \Delta t) \tag{2.32}$$

for $i=1,2$ and $\mathbf{g}_i(t) = -\partial\chi(t)/\partial\mathbf{r}_i$. The Lagrange multipliers λ and μ are determined by the SHAKE [9] and RATTLE [10] algorithms respectively.

2.2 Statistical Analysis Methods

In order to calculate observables and to determine macroscopic behaviours of a simulation of microscopic detail, it is necessary to ensure that the simulation has been run under the correct thermodynamic conditions and that the system has sampled a sufficient amount of the phase.

To accomplish sufficient sampling of microstates in molecular dynamics simulations the *ergodic hypothesis* is assumed. This states that a simulation that has run for a

sufficiently long time will sample sufficiently many independent microstates to calculate macroscopic observables from. This is equivalent to saying that a thermodynamic integral over all states is the same as a thermodynamic integral over all time. The probability of sampling a particular point in the phase space with an energy E and at a temperature T is given by the Boltzmann probability e^{-E/k_BT}. The energetically unfavourable areas of the phase space have a negligibly small probability of occurring and thus contribute negligibly to any macroscopic observable.

A further difficulty in sampling from simulation is that a system may fall into local energy minima or have very large energy barriers (>2kT) between minima. To overcome this problem, many simulations from different starting configurations are used along with non-Boltzmann molecular dynamics (discussed in Chapter 3) methods.

2.2.1 Ensemble Sampling from Simulations

To compare the results from analysis of simulations with experiments, the simulations have to mimic laboratory conditions as closely as possible. Thermodynamic properties such as total energy in the system, pressure, temperature, volume, number of particles and chemical potential are typical conditions that are controlled in experiment. The easiest conditions to control, and thus the most common, are pressure and temperature from the pressure-volume entropy-temperature thermodynamic conjugate pairs. The methods used in this thesis (described in Chapter 3) deal with the Helmholtz free energy, which is specified in the canonical ensemble where temperature and the volume are held constant.

To impose these restrictions (i.e. to run a simulation in the canonical ensemble) the dynamics has to be modified, which means modifying the Hamiltonian. This was done in the simulations by using Nosé-Hoover thermostats to generate canonically distributed positions and momenta.[11] The method of controlling the temperature proposed by Nosé was to have a heat reservoir that can interact with the system as an additional degree of freedom. This involved inserting additional terms (the

thermostats) into the Hamiltonian to control the temperature by allowing it to fluctuate around a certain value. The additional degree of freedom has a coordinate and corresponding momenta added to the system, s and p_s, with a mass Q. The Hamiltonian is modified to incorporate the thermostat is as follows:

$$H(\mathbf{r}^N,\mathbf{p}^N)=\frac{\mathbf{p}^2}{2m}+V(\mathbf{r})+\frac{p_s^2}{2Q}+3Nk_BT\ln s. \tag{2.33}$$

By defining alternative momenta $\mathbf{p}'=\mathbf{p}/s$ we can find that the microcanonical ensemble in $\mathbf{r},\mathbf{p},s,p_s$ can be written as a canonical ensemble in $\mathbf{r}$ and $\mathbf{p}'$. The additional coordinate s can thus be thought of as a scaling factor. By scaling s's associated momenta and the time step,

$$p_s'=p_s/s, \tag{2.34}$$

$$\Delta t'=\Delta t/s, \tag{2.35}$$

we can run a dynamics simulation that generates canonically distributed coordinates and momenta: the primed variables $\mathbf{r}'=\mathbf{r}$ and $\mathbf{p}'=\mathbf{p}/s$. The Hoover formulation adopts a new variable $\xi=sp_s'/Q$. This simplifies the Hamiltonian and the equations of motion in terms of the primed coordinates from those proposed by Nosé:

$$\dot{\mathbf{r}}_i'=\mathbf{p}_i'/m_i, \tag{2.36}$$

$$\dot{\mathbf{p}}_i'=-\nabla_i V\left(\mathbf{r}^N\right)-\xi\mathbf{p}_i', \tag{2.37}$$

$$\dot{\xi}=\frac{1}{Q}\left[\sum_{i=1}^{N}\frac{\mathbf{p}_i'^2}{m_i}-3Nk_BT\right], \tag{2.38}$$

$$\frac{\dot{s}}{s}=\xi=\frac{d\ln s}{dt}. \tag{2.39}$$

These are the dynamics equations that are used with the MTK equations of motion to calculate the time evolution of the system in the canonical ensemble.

2.2.2 Pair Distribution Functions

The atomic pair distribution function (PDF) describes the averaged environment surrounding an atom. The most commonly used pair distribution function is the radial distribution function, which is used to provide insight into the isotropic structure of a molecule around another. This describes the density of a type of atom or molecule from a point outwards as a function of distance r.

The radial distribution function (RDF) $g_{ab}(r)$ is calculated with

$$g_{ab}(r) = \frac{1}{N_a N_b} \sum_{i=1}^{N_a} \sum_{j=1}^{N_b} \delta(r_{ij} - r). \tag{2.40}$$

Here N_a is the number of molecules of type a, N_b is the number of molecules of type b and r_{ij} is the distance between molecule i and molecule j. This indication of density as a function of distance from a point or type of molecule provides an averaged view of the structure of the system. It is often used to investigate liquid and solid models where there are many molecules to average over. Typically a large number of molecules of types a and b are needed to calculate an accurate RDF.

2.2.3 Spatial Distribution Function

The spatial distribution function (SDF), similarly to the radial distribution function, provides structural information of the system. It is a simulation averaged, three-dimensional image of the anisotropic structure around a single molecule. The RDF averages out the angular coordinates and only looks at the distance dependence of the average density. The SDF, however, looks at the distance and angular dependence of the average density. This was used first by Svishchev and Kusalik to examine the structure of water for a given model[12] but remains to be used today to examine the 3D structural properties of liquids, in particular water.

To calculate the SDF of molecule 1 around molecule 2, each frame of the molecular dynamics trajectory is translated and rotated to align molecule 2 to a common orientation. The density of the atoms in molecule 1 about molecule 2 is then computed using a 3D histogram surrounding 2. The histogram is then normalized with respect to the bulk density of molecule 1.

This method is an improvement over the radial distribution function since it gives better insight into the average surrounding 3D environment of the molecule. To calculate smooth SDF's many simulation frames are needed. The large number of frames can be reduced if there are multiple molecules that can be used to calculate the 3D density. Due to this limitation, the SDF is usually calculated between a molecule of interest and the solvent molecules.

2.2.4 Orientational Order Parameter

A further probe of the local orientational structure is the orientational order parameter introduced by Tanaka[13, 14] and based originally on the angular order parameter of Chau and Hardwick[15].

The previous two measures of structure have either only a distance dependence (RDF) or a distance and angular dependence (SDF). This order parameter looks only at the orientational order in the structure and separates out the distance dependence. It has previously been used to elucidate and compare the structure of the nearest neighbour structure of liquid water.[16] The orientational order parameter is calculated by

$$q_i = 1 - \frac{3}{8}\sum_{j=1}^{3}\sum_{k=j+1}^{4}\left[\cos\theta_{ijk} + \frac{1}{3}\right]^2. \qquad (2.41)$$

Here θ_{ijk} is the angle between the central water molecule i and two of its neighbours j and k. The average orientational order is then calculated by taking an average of q_i over all the waters in the system. The average orientational order parameter q is thus

a measure of the averaged *short range* tetrahedral structure of water and is sensitive enough to probe the phase behaviour of water molecules.[16]

Using this measure the tetrahedrality of the first solvation shell can be quantified. An orientational order value of 1 means perfect tetrahedrality and a value of 0 means a completely uncorrelated system. This is especially informative when used in conjunction with other measures of local orientational order for water.

3 Free Energy from Molecular Simulation

The free energy of a system is the amount of thermodynamic energy that can be converted into work energy. As a result there are many reasons to calculate the free energy and the free energy changes within a system. Free energy calculations allow us to investigate solvation effects, ligand binding, relative stabilities of conformations and systems and as well as probabilities of different reaction pathways. In short, knowing the free energy gives us a great insight into many systems and their mechanisms of action. To numerically calculate the free energy of a system is not trivial and hence different methods have been developed. In this chapter we explore the principle methods that are used to calculate free energies from computer simulations (§3.2) and important considerations when calculating intermolecular orientational free energies (§3.3), starting with a brief review of some statistical mechanics (§3.1) and concluding with the free energy calculation method used in this thesis (§3.4).

3.1 Statistical Mechanics

When calculating thermodynamic quantities, we do so for a system in a particular state. For a given thermodynamic state (macrostate) there exist a large number of possible microstates (states of different microscopic configurations). For a given number of particles N, volume V, and internal energy E, we denote the number of these microstates as $\Omega(N,V,E)$. For a classical system the number of microstates is written as

$$\Omega(N,V,E)=\frac{1}{N!h^{3N}}\int\delta\left[H\left(\mathbf{p}^N,\mathbf{q}^N\right)-E\right]d^N\mathbf{p}\,d^N\mathbf{q}. \tag{3.1}$$

Consider two systems in thermal contact with energies E_1 and E_2, numbers of particles N_1 and N_2 and volumes V_1 and V_2. The total energy of the system is then $E_0 = E_1 + E_2$. The number of microstates of the total system is a product of the

number of microstates of system 1 and 2 since the two systems, while being in thermal contact, are statistically independent of each other. Thus we can write $\Omega_0(N_1,N_2,V_1,V_2,E_1,E_2)=\Omega_1(N_1,V_1,E_1)\Omega_2(N_2,V_2,E_2)$.

Consider further that the particles of the two systems cannot mix and that the exterior boundary of the system is fixed, thus we have that N_1 and N_2 are constant and the total volume $V_0=V_1+V_2$ is also constant. Natural systems evolve to their most probable macrostate. This means that the macrostate will be that for which the number of microstates is maximum. For a system of constant total energy and volume and numbers of particles, we find the energies E_1^* and E_2^* that Ω_0 is at a maximum:

$$\frac{\partial\left[\Omega_1(E_1,N_1,V_1)\Omega_2(E_2,N_2,V_2)\right]}{\partial E_1}=0$$
$$\left(\frac{\partial\Omega_1}{\partial E_1}\Omega_2+\Omega_1\frac{\partial E_2}{\partial E_1}\frac{\partial\Omega_1}{\partial E_2}\right)_{E_1=E_1^*,E_2=E_2^*}=0. \tag{3.2}$$

Since $\partial E_2/\partial E_1=-1$ this can be reduced to

$$\frac{1}{\Omega_1}\frac{\partial\Omega_1}{\partial E_1}(E_1^*)=\frac{1}{\Omega_2}\frac{\partial\Omega_1}{\partial E_2}(E_2^*), \tag{3.3}$$

which is the same as

$$\frac{\partial}{\partial E_1}\log\Omega_1(E_1^*)=\frac{\partial}{\partial E_2}\log\Omega_2(E_2^*)=const. \tag{3.4}$$

This equality can be generalised to any number of systems in thermal contact using $\partial\log\Omega/\partial E=const$ for each system. This constant can be calculated to get:

$$\frac{\partial\log\Omega}{\partial E}=\frac{1}{k_BT}. \tag{3.5}$$

From thermodynamics we know, however, that at constant volume and constant number of particles $\partial S / \partial E = 1/T$. This leads us to the important relation:

$$S = k_B \log \Omega . \tag{3.6}$$

Experiments measuring thermodynamic parameters involve knowledge of temperature and not internal energy. Calculations involving the same measurements have to then be done with the same knowledge of conditions. This leads us to the canonical ensemble, which fixes the number of particles N, volume V and temperature T. To calculate the distribution of states in for a given N,V and T, we couple our system of interest to a large heat reservoir. There are thus two systems, the bath with energy E_r and the system of interest in state i with energy E_i. As before the total energy of the two systems is constant, thus $E_r + E_i = E = const$. The probability of finding the bath in a particular state is then

$$p_i = \frac{\Omega(N,V,E-E_i)}{\sum_j \Omega(N,V,E-E_j)}. \tag{3.7}$$

We continue by expanding the logarithm of the microcanonical density of states:

$$\ln\Omega(N,V,E-E_i) = \ln\Omega(N,V,E) - E_i \frac{\partial \ln\Omega(N,V,E)}{\partial E} + O\left(\frac{1}{E}\right). \tag{3.8}$$

We can use the relation (3.5) to get

$$\ln\Omega(N,V,E-E_i) = \ln\Omega(N,V,E) - \frac{E_i}{k_B T} + O\left(\frac{1}{E}\right). \tag{3.9}$$

By inserting equation (3.9) into (3.7) we obtain the Boltzmann distribution:

$$p_i = \frac{\exp(-E_i / k_B T)}{\sum_j \exp(-E_j / k_B T)}, \tag{3.10}$$

from which we define the configurational partition function

$$Z(N,V,T)=\sum_j \exp\left(-E_j / k_B T\right). \tag{3.11}$$

3.2 Free Energy Methods

For constant volume and constant temperature systems, the Helmholtz free energy A is the amount of energy available for work. The Helmholtz free energy, as well as many other thermodynamics variables, can be calculated from the configurational partition function of the system.

For example, from equations (3.10) and (3.11) we can calculate the expectation value of the internal energy of a system:

$$U=\sum_i p_i E_i=-\frac{\partial \log Z}{\partial \beta}, \tag{3.12}$$

where $\beta=\left(k_B T\right)^{-1}$. If we look at small changes in the volume V and in β of $\log Z$ then we get:

$$d(\log Z)=\frac{\partial \log Z}{\partial \beta}d\beta+\frac{\partial \log Z}{\partial V}dV. \tag{3.13}$$

It is useful here to note that the pressure P is calculated by:

$$P=\sum_i p_i P_i=\frac{1}{\beta}\frac{\partial \log Z}{\partial V}. \tag{3.14}$$

This means that we have the following relation:

$$d(\log Z)=-Ud\beta+\beta P dV. \tag{3.15}$$

Rewriting $Ud\beta$ as $d(\beta U) - \beta dU$ allows us to rearrange equation (3.15) to write:

$$dU = k_B T d(\log Z + \beta U) - PdV. \tag{3.16}$$

Using the first law of thermodynamics, $dU = TdS - PdV$, we have that:

$$dS = d(k_B \log Z + \frac{U}{T}), \tag{3.17}$$

which implies that

$$S = k_B \log Z + \frac{U}{T} + C. \tag{3.18}$$

At the limit of $T \to 0$ we know that $S \to 0$, $U \to 0$ and $Z \to 1$, thus we have that the constant $C = 0$. Since the Helmholtz free energy is defined as $A = U - TS$ we obtain:

$$A = -k_B T \log Z \text{ or } Z = \exp\left(-\frac{A}{k_B T}\right). \tag{3.19}$$

The free energy and the partition function are not directly calculable from either experiment or simulation since they depend on the available configurational space of a system, which is infinite and hence not measurable. Various mathematical and computational techniques have thus been developed to investigate free energy differences in a system.

In the remaining sections we discuss some of the methods that have been previously developed (§3.2.1) and describe in detail the method that was developed and used (§3.4) for the research presented in this thesis.

3.2.1 Free Energy Perturbation

In many cases, we need not calculate the absolute free energy but rather the difference in free energy between two states. The two states can be for example: different conformations of the same molecule, functional group mutations on a molecule, whole molecular mutations or simply different points along a reaction coordinate.

The free energy perturbation method is well suited to calculate the free energy difference between states that can be considered perturbations of one another. Many free energy differences can, however, be calculated and combined on a series of perturbations between 2 different states to calculate their free energy difference.

The difference in free energy between two states 0 and 1 is calculated from (3.19) as

$$\Delta A = -k_B T \log \frac{Z_1\left(\mathbf{X}^N\right)}{Z_0\left(\mathbf{X}^N\right)}. \tag{3.20}$$

The fraction $Z_1\left(\mathbf{X}^N\right)/Z_0\left(\mathbf{X}^N\right)$ is given as

$$\frac{Z_1\left(\mathbf{X}^N\right)}{Z_0\left(\mathbf{X}^N\right)} = \frac{\int\ldots\int \exp\left(-\beta H_1\left(\mathbf{X}^N\right)\right) d\mathbf{X}^N}{\int\ldots\int \exp\left(-\beta H_0\left(\mathbf{X}^N\right)\right) d\mathbf{X}^N}. \tag{3.21}$$

By multiplying the integrand in the numerator by

$$1 = \exp\left(-\beta H_0\left(\mathbf{X}^N\right)\right)\exp\left(+\beta H_0\left(\mathbf{X}^N\right)\right), \tag{3.22}$$

we get

$$\frac{Z_1(\mathbf{X}^N)}{Z_0(\mathbf{X}^N)} = \frac{\int\cdots\int \exp\left(-\beta H_0(\mathbf{X}^N)\right)\exp\left(+\beta H_0(\mathbf{X}^N)\right)\exp\left(-\beta H_1(\mathbf{X}^N)\right)d\mathbf{X}^N}{\int\cdots\int \exp\left(-\beta H_0(\mathbf{X}^N)\right)d\mathbf{X}^N},$$
$$= \frac{\int\cdots\int \exp\left(-\beta H_0(\mathbf{X}^N)\right)\exp\left[-\beta\left(H_1(\mathbf{X}^N) - H_0(\mathbf{X}^N)\right)\right]d\mathbf{X}^N}{\int\cdots\int \exp\left(-\beta H_0(\mathbf{X}^N)\right)d\mathbf{X}^N}. \quad (3.23)$$

This, however, is the ensemble average of $\exp\left[-\beta\left(H_1(\mathbf{X}^N) - H_0(\mathbf{X}^N)\right)\right]$ over the configurations in state 0 of the system, i.e.:

$$\frac{Z_1(\mathbf{X}^N)}{Z_0(\mathbf{X}^N)} = \left\langle \exp\left[-\beta\left(H_1(\mathbf{X}^N) - H_0(\mathbf{X}^N)\right)\right]\right\rangle_0. \quad (3.24)$$

The initial and final states can be arbitrarily assigned and as such equation (3.24) could also be written as

$$\frac{Z_1(\mathbf{X}^N)}{Z_0(\mathbf{X}^N)} = \left\langle \exp\left[+\beta\left(H_1(\mathbf{X}^N) - H_0(\mathbf{X}^N)\right)\right]\right\rangle_1, \quad (3.25)$$

where the ensemble averaging is now done over a representative sample of configurations of the system in state 1. The free energy difference is thus calculated by

$$\Delta A = -k_B T \log\left\langle \exp\left[-\beta\left(H_1(\mathbf{X}^N) - H_0(\mathbf{X}^N)\right)\right]\right\rangle_0, \quad (3.26)$$

or by

$$\Delta A = -k_B T \log\left\langle \exp\left[+\beta\left(H_1(\mathbf{X}^N) - H_0(\mathbf{X}^N)\right)\right]\right\rangle_1. \quad (3.27)$$

Two simulations can be run and the free energy differences can be calculated using equation (3.26) or (3.27), with the average being the final result. This is called double-wide sampling. The difference in result of these two calculations is an

indication of the uncertainty in the result. When the states are very similar, however, this difference becomes very small. To stay within an acceptably small uncertainty, the energy difference between the two states should be $< 2k_BT$.

To investigate different states that have a free energy difference that is expected to be greater than $2k_BT$ the free energy perturbation equations (equations (3.26) and (3.27)) are applied on k intermediate states between state 0 and state 1 and summing up the free energy differences, i.e.

$$\Delta A = \sum_{i=0}^{k-1} \Delta A_{i \to i+1}. \tag{3.28}$$

The intermediate states need to be chosen such that for every i, $\Delta A_{i \to i+1} < 2k_BT$. The number of intermediate states is typically determined by trial where if a calculation of $\Delta A_{i \to i+1}$ is greater than $2k_BT$, an intermediate state is introduced.

For the most accurate calculation, each simulation (except for the first and last where $i = 0$ and $i = k$) is used twice: the first time using equation (3.26) to calculate $\Delta A_{i \to i+1}$ and then again with equation (3.27) to calculate $\Delta A_{i+1 \to i}$. The average of these two is then used as a more accurate value for the free energy difference between state i and state i+1. A simulation can be added on to each end such that double-wide sampling can be used for the each incremental step of the intermediate states. This is depicted in Figure 3-1.

The calculation of the free energy is then given by:

$$\Delta A = \frac{1}{2} \sum_{i=0}^{k-1} \left[\Delta A_{i \to i+1} + \Delta A_{i+1 \to i} \right]. \tag{3.29}$$

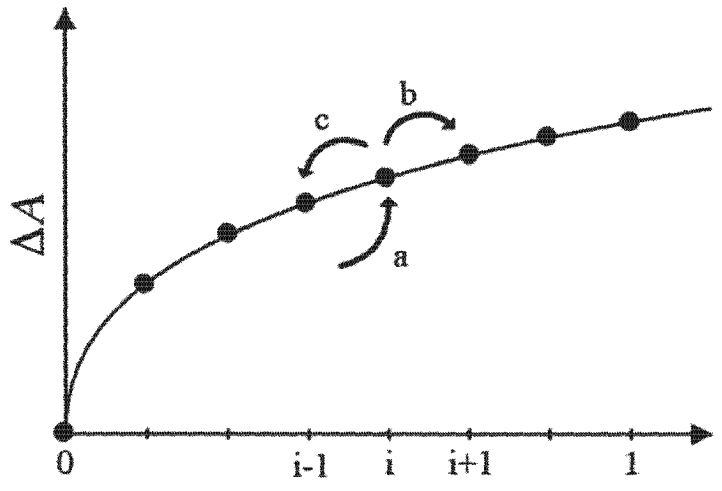

Figure 3-1: A schematic illustration of double-wide sampling where the free energy difference at (a) is calculated from the forward step (b) and the backward step (c).

3.2.2 Thermodynamic Integration

Thermodynamic integration involves the calculation of the free energy difference ΔA by introducing a coupling parameter between the two states. If we have two states, 0 and 1, each with Hamiltonian H_0 and H_1, we can construct a hybrid state than is part state 0 and part state 1 by using a hybrid Hamiltonian:

$$H\left(\mathbf{X}^N,\lambda\right)=\lambda H_1\left(\mathbf{X}^N\right)+\left(1-\lambda\right)H_0\left(\mathbf{X}^N\right). \tag{3.30}$$

Here the coupling parameter λ ranges between 0 and 1 and when $\lambda=0$ the system is in state 0 and when $\lambda=1$ the system is in state 1. The partition function is then dependent on the coupling parameter:

$$Z\left(\mathbf{X}^N,\lambda\right)=\int\ldots\int\exp\left(-\beta H\left(\mathbf{X}^N,\lambda\right)\right)d\mathbf{X}^N. \tag{3.31}$$

The summation over i in equation (3.31) represents the sum over all configurations in configurational space of the system. We can then calculate the free energy using equation (3.19):

$$A = -\frac{1}{\beta}\log\int\ldots\int\exp\left(-\beta H\left(\mathbf{X}^N,\lambda\right)\right)d\mathbf{X}^N. \tag{3.32}$$

To calculate the difference in free energy between the two states, we integrate $\partial A/\partial\lambda$ between $\lambda = 0$ and $\lambda = 1$. The partial derivative of (3.32) is:

$$\begin{aligned}\frac{\partial A}{\partial\lambda} &= -\frac{1}{\beta}\left(\frac{\partial\log Z}{\partial\lambda}\right)\\ &= -\frac{1}{\beta Z(\lambda)}\left(\frac{\partial Z(\lambda)}{\partial\lambda}\right)\\ &= \frac{\int\ldots\int\left(\frac{\partial H\left(\mathbf{X}^N,\lambda\right)}{\partial\lambda}\right)\exp\left(-\beta H\left(\mathbf{X}^N,\lambda\right)\right)d\mathbf{X}^N}{\int\ldots\int\exp\left(-\beta H\left(\mathbf{X}^N,\lambda\right)\right)d\mathbf{X}^N}.\end{aligned} \tag{3.33}$$

Equation (3.33) shows that the derivative of A with respect to λ is thus the ensemble average of the derivative of the Hamiltonian with respect to λ. Using the bracket notation we have then

$$\frac{\partial A}{\partial\lambda} = \left\langle\frac{\partial H\left(\mathbf{X}^N,\lambda\right)}{\partial\lambda}\right\rangle_\lambda. \tag{3.34}$$

The ensemble average in this case is weighted by the λ dependent probability function:

$$P\left(\mathbf{X}^N,\lambda\right) = \frac{\exp\left(-\beta H\left(\mathbf{X}^N,\lambda\right)\right)}{\int\ldots\int\exp\left(-\beta H\left(\mathbf{X}^N,\lambda\right)\right)d\mathbf{X}^N}. \tag{3.35}$$

If we then integrate equation (3.34) we get the difference in free energy as a function of λ:

$$\Delta A = \int_0^1 \left\langle \frac{\partial H(\mathbf{X}^N, \lambda)}{\partial \lambda} \right\rangle_\lambda d\lambda. \tag{3.36}$$

The derivative of the Hamiltonian can easily be evaluated using equation (3.30). Doing so gives us that

$$\Delta A = \int_0^1 \left\langle H_1(\mathbf{X}^N) - H_0(\mathbf{X}^N) \right\rangle_\lambda d\lambda. \tag{3.37}$$

This is the equation that is used to calculate the free energy differences between two states using thermodynamic integration. The λ in the integration comes from the λ dependent probability function in the ensemble average.

The numerical evaluation of the free energy difference is done using numerical integration of equation (3.37). A number of simulations are set up at discrete values of λ between 0 and 1. The ensemble average of $H_1(\mathbf{X}^N) - H_0(\mathbf{X}^N)$ is then calculated and integrated numerically to calculate the free energy difference between states 0 and 1.

3.2.2.1 Adaptive Biasing Force

The method of thermodynamic integration described above is particularly well suited to calculating the free energy difference as a single number between two states. To calculate, however, the free energy difference along a reaction coordinate one can still use the thermodynamic integration method. This, however, presents the common complication in free energy calculations that in an unconstrained molecular dynamics simulation there may be insurmountable free energy barriers along the reaction coordinates. The adaptive biasing force method proposed by Darve and Pohorille[1] presents us with an adaptive way to drive the simulation over free energy barriers.

The adaptive biasing force method calculates a running average of the force along a reaction coordinate. The negative of this force is then applied as an external force to the atoms involved in defining the reaction coordinate to assist the system in surmounting the free energy barriers. This allows the simulation to continue in a diffusion-like motion along the reaction coordinate allowing the system to sample high-energy regions of the configurational space.

For a free energy $A(\xi)$ parameterised by a reaction coordinate ξ, we obtain the derivative of the free energy

$$\frac{\partial A(\xi)}{\partial \xi} = \left\langle -F(\xi) + \frac{\partial \ln |J|}{\partial \xi} \middle| \xi \right\rangle, \tag{3.38}$$

where $F(\xi)$ is the mean force at ξ and J is the Jacobian $\partial \xi / \partial x_i$. From this we can see that we can calculate the free energy difference as a function of the reaction coordinate by integrating of the negative of the sum of the average force and the Jacobian term:

$$\Delta A(\xi) = -\int_{\xi_0}^{\xi} \left\langle F(\xi') + \frac{\partial \ln |J|}{\partial \xi'} \right\rangle d\xi'. \tag{3.39}$$

Calculating the average force in a molecular dynamics simulation can be done using the finite differences method and this can further made simpler using the prescription given in by Darve *et al.* in ref. 2 . A complication with this method, and indeed with any method that uses constraints[3-5], is that we have to compute the Jacobian correction term. While this may be simple to do for the case of only one reaction coordinate, this can become complicated for multiple reaction coordinates that are used when computing a multidimensional free energy surface as we wish to do in this work.

3.2.3 Umbrella Sampling

The free-energy difference between a state of interest compared to a reference state of a system can be expressed in terms of the configurational ensemble average of the energy difference ΔH between the two states:

$$\Delta A = -k_B T \ln\left\langle \exp\left(-\Delta H / k_B T\right)\right\rangle. \tag{3.40}$$

This can be expressed in integral form as:

$$\Delta \mathrm{A} = -k_B T \ln \int f_0(\Delta H)\exp\left(-\Delta H / k_B T\right) d\Delta H, \tag{3.41}$$

where $f_0(\Delta H)$ is the probability density of ΔH in the reference state. In order to calculate the free energy difference we would need to compute the values of $f_0(\Delta H)$ in the ranges where $f_0(\Delta H)\exp\left(-\Delta H / k_B T\right)$ is large. The region of configurational space that this corresponds to is the region that would be sampled in a standard simulation of the state of interest and not the reference state. We can relate $f_0(\Delta H)$ to the probability density $f(\Delta H)$ in a simulation of state of interest using

$$f(\Delta U) = f_0(\Delta U)\exp\left(-\Delta U / k_b T\right) Z_0 / Z, \tag{3.42}$$

where Z_0 and Z are the configurational partition functions of the reference state and the state of interest. This, however, does not help us calculate the free energy difference since we cannot calculate the configurational partition functions thus meaning that we can only calculate relative values of $f_0(\Delta H)$ where absolute values are required for the integral in (3.41). It is thus clear that simulations with standard Boltzmann sampling is not sufficient to explore the relevant parts of the configurational space required to calculate free energy differences.[6]

To adequately explore the configurational space requires that a bias be introduced into the system. Suppose a system is biased to sample a non-Boltzmann distribution, $p_w(\mathbf{q}^N) = w(\mathbf{q}^N)\exp(-\Delta H / k_B T) / \int w(\mathbf{q}^N)\exp(-\Delta H / k_B T) d\mathbf{q}^N$. Then the canonical average in equation (3.40) can be written in the form

$$\begin{aligned}\left\langle \exp\left(-\frac{\Delta H}{k_B T}\right)\right\rangle &= \frac{\int e^{-\frac{\Delta H}{k_B T}} e^{-\frac{H}{k_B T}} d\mathbf{q}^N}{\int e^{-\frac{H}{k_B T}} d\mathbf{q}^N} \\ &= \frac{\int\left[e^{-\frac{\Delta H}{k_B T}} / w(\mathbf{q}^N)\right] w(\mathbf{q}^N) e^{-\frac{H}{k_B T}} d\mathbf{q}^N}{\int\left[1/w(\mathbf{q}^N)\right] w(\mathbf{q}^N) e^{-\frac{H}{k_B T}} d\mathbf{q}^N} \times \frac{\int e^{-\frac{\Delta H}{k_B T}} d\mathbf{q}^N}{\int e^{-\frac{\Delta H}{k_B T}} d\mathbf{q}^N} \\ &= \frac{\int\left[e^{-\frac{\Delta H}{k_B T}} / w(\mathbf{q}^N)\right] p_w(\mathbf{q}^N) d\mathbf{q}^N}{\int\left[1/w(\mathbf{q}^N)\right] p_w(\mathbf{q}^N) d\mathbf{q}^N} \\ &= \frac{\left\langle e^{-\frac{\Delta H}{k_B T}} / w(\mathbf{q}^N)\right\rangle_w}{\left\langle 1/w(\mathbf{q}^N)\right\rangle_w},\end{aligned} \tag{3.43}$$

where $\langle \ldots \rangle_w$ denotes the average over the biased distribution $p_w(\mathbf{q}^N)$. This shows that from a biased simulation we can calculate the free energy difference between two states the have an energy difference of ΔH. This technique is known as umbrella sampling.

The simplest way to generate a biased simulation is to add an external potential into the system such that the Hamiltonian becomes

$$H = H_0 + U. \tag{3.44}$$

The biasing weight to the Boltzmann distribution $w(\mathbf{q}^N)$ is then given as $w = \exp(-U / k_B T)$.

Very often we wish to compute not only the free energy difference as a single value, but as a surface parameterised by reaction coordinates $\xi = f(\mathbf{q}^N)$. Thus $U \to U(\xi)$ and the Hamiltonian becomes

$$H(\xi) = H_0 + U(\xi). \tag{3.45}$$

This introduces a similar complication experienced by the Adaptive Biasing Force method that in changing configurational integral over Cartesian coordinates to an integral over reaction coordinates; a correction term in the form of the derivative of the logarithm of the Jacobian is introduced.[4, 5] Converting the forces that arise from the external potential from forces on the reaction coordinates to forces on Cartesian coordinates in the simulation can, however, circumvent need for this correction term.[5,7, 8] The equation to calculate the free energy then becomes:

$$\begin{aligned} \Delta A(\xi) &= -k_B T \ln \frac{\left\langle \exp\left(-\Delta H(\xi)/k_B T\right) \exp\left(U(\xi)/k_B T\right) \right\rangle_w}{\left\langle \exp\left(U(\xi)/k_B T\right) \right\rangle_w}, \\ &= -k_B T \ln P(\xi) \exp\left(U(\xi)/k_B T\right), \end{aligned} \tag{3.46}$$

where $P(\xi)$ is the weighted probability distribution of the reaction coordinate ξ.

3.2.3.1 Windowing Method

Biasing the simulations to run over the whole range of the reaction coordinates is often computationally impossible to do. What is done to avoid doing this is to bias different simulations to small overlapping ranges of the reaction coordinate space. Each range of the reaction coordinates being investigated is called a window; hence this method is referred to as the windowing method.

A typical implementation of the windowing method is to take the biasing potential as the form of a sum of harmonic terms:

$$U(\xi) = \sum_i k_i \left(\xi - \xi_{0,i}\right)^2 . \tag{3.47}$$

Multiple simulations are then run starting in each of the different harmonic wells (see Figure 3-2). The probability distributions of the reaction coordinates from the different simulations are combined and equation (3.46) is used to calculate the free energy surface.

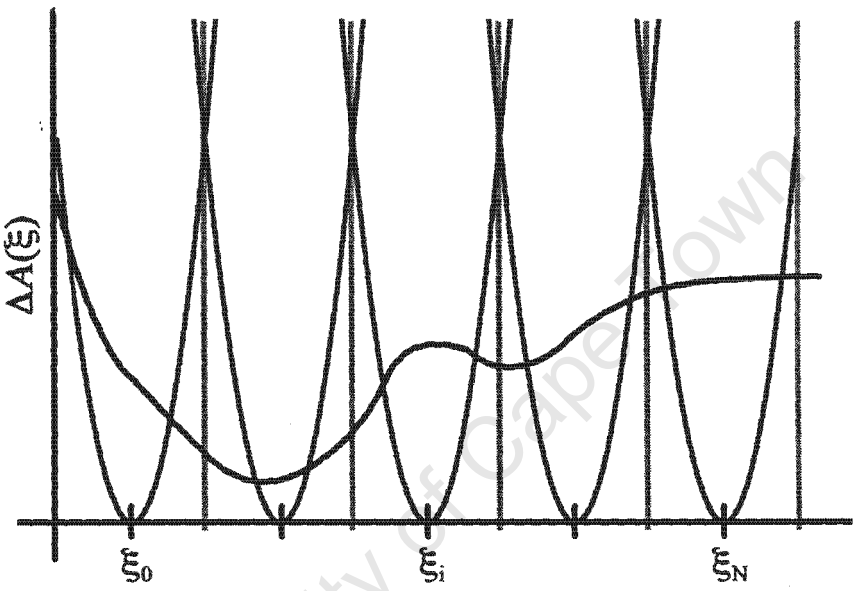

Figure 3-2: An illustration of the windowing method where harmonic potentials have been placed at $\xi_0, \cdots, \xi_N$ to create *N* windows for *N* simulations.

This method can also be used to construct *n*-dimensional free-energy surfaces with a potential

$$U(\xi_1, \cdots, \xi_n) = \sum_i k_i \left\| (\xi_1, \cdots, \xi_n) - (\xi_1, \cdots, \xi_n)_{0,i} \right\|^2 . \tag{3.48}$$

Care has to be taken in the multidimensional approach to cover the full *n*-dimensional volume of ξ-space. This makes the number of 'windows' very large as the volume of the space scales exponentially with *n*. As a result, this method is not typically used when we have $n > 2$.

3.2.3.2 Adaptive Umbrella Sampling

The second form of biasing potential that is often applied is one that is calculated in an adaptive way; hence the method is called Adaptive Umbrella Sampling first described by Mezei.[9] If one looks at a probability distribution that is calculated from the values of a reaction coordinate that is sampled during a simulation, one can observe that there are definite regions that are not sampled.

The calculated free energy surface from this simulation then describes where the free energy barriers in the system are located. This means that if one were to take the negative of the surface as the biasing potential in the following simulation, then the free energy barriers located in the previous simulation could be overcome (see Figure 3-3). This leads to improved sampling and provides an easy check of the calculation. Once a system has the correct free energy barriers calculated, the biasing potential should cancel them out exactly, allowing the system to sample freely along the reaction coordinate. Thus, when a simulation yields a near uniform sampling distribution the calculation has converged.

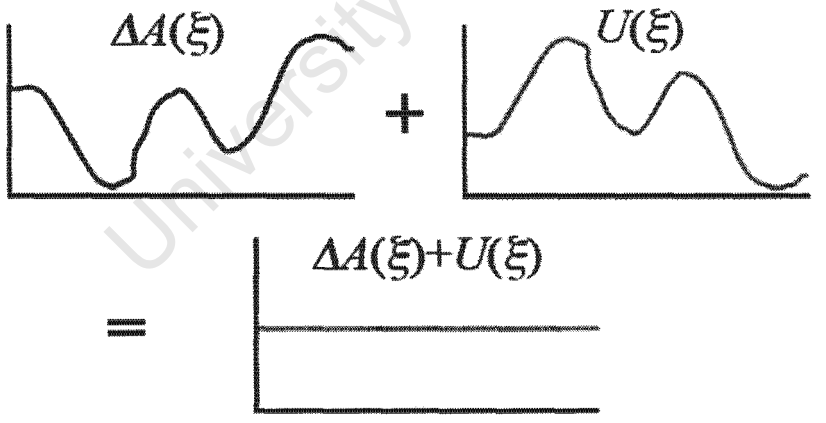

Figure 3-3: By taking the applied potential as the negative of the free energy surface the resulting potential in the simulation is completely flat, allowing uniform sampling along the reaction coordinate.

A simple description of the procedure is as follows. Set $U(\xi)$ equal to zero everywhere, then:

1. Run the simulation with applied potential $U(\xi)$.
2. From the histogram of the reaction coordinate values sampled in the simulation calculate the biased probability $P(\xi)$.
3. Use equation (3.46) to calculate the free energy $\Delta A(\xi)$.
4. If the biased probability distribution $P(\xi)$ is nearly uniform, then free energy surface calculated in step 4 is the converged potential of mean force.
5. Otherwise, set the biasing potential to the negative of the free energy surface $U(R) = -\Delta A(R)$ and start from step 1 again.

In the first few simulations there are often large regions of the reaction coordinate that are not sampled. To assign some value to these regions, we set them equal to the maximum value of the free energy in the regions that were sampled.

This means, however, that when using the procedure described above there are often sharp peaks and discontinuous jumps in the free energy surface and hence the applied potential. The forces that arise from the derivative of the applied potential can thus be unrealistically large for a molecular dynamics simulation and can cause it to crash. By employing a smoothing procedure we can remove the discontinuous jumps and reduce the size of the forces. For a reaction coordinate range that is discretised into $\{\xi_0, \xi_1, \cdots \xi_N\}$ we calculate the smoothed potential using:

$$U_{\text{new}}(\xi_i) = \frac{1}{3}\left[-0.3U(\xi_{i-2}) + 1.3U(\xi_{i-1}) + U(\xi_i) + 1.3U(\xi_{i+1}) - 0.3U(\xi_{i+2})\right]. \quad (3.49)$$

By applying this repeatedly we can increasingly smooth the biasing potential until the largest jumps between adjacent reaction coordinate values are small enough for a molecular dynamics simulation.

The Adaptive Umbrella Sampling method can also be easily extended into higher dimensions for multiple reaction coordinates.

3.2.3.3 Metadynamics

The metadynamics method was developed and introduced by Liao and Parrinello.[10-13] This method is similar to the Adaptive Umbrella Sampling and the Windowing methods, but can simply be implemented in multiple dimensions. The essence of the method lies in constructing the free energy surface from Gaussian terms that are evolved dynamically based on the sampling the reaction coordinates in the simulation.

The metadynamics simulation restrains the simulation to a particular value of the reaction coordinate using inverted Gaussian potentials. The Hamiltonian for a multidimensional parameterisation of the free energy surface using $(\xi_1,\cdots,\xi_n)$ is modified to become:

$$H = H_0 + \sum_i \frac{p_{\xi,i}^2}{m_{\xi,i}} - \sum_i \sum_{k=1}^{N_i} W_{i,k} \exp\left(-\frac{\left(\xi_i - \zeta_{i,k}\right)^2}{2\sigma_{i,k}^2}\right). \qquad (3.50)$$

Here along each reaction coordinate ξ_i there are N_i Gaussians, with heights $W_{i,k}$, widths $\sigma_{i,k}$ and each centred at $\zeta_{i,k}$, that are added to the unbiased Hamiltonian. The heights and widths and centres are chosen to reach a compromise between accuracy and efficiency. Typically for a single parameter free energy surface 200 or more Gaussian terms are used.[11] This number increases exponentially for higher dimensions. This makes the computations using metadynamics slow.

The method is made adaptive by using the histories of the reaction coordinates in a simulation to update the positions of the Gaussians. This updates $\zeta_{i,k}$ so that the Gaussians are optimally places in the centres of the wells of the potential energy surface, enhancing the sampling of the reaction coordinates in a subsequent simulation. The method can also be modified to continuously update the Gaussians.[10]

As with other methods, there arises a Jacobian correction term due to the reaction coordinate dependent potential. This correction term is often neglected in the calculations of 1D and 2D free energy surfaces[11] but for higher dimensional surfaces this correction may not be as trivial and consideration thereof cannot be neglected.

3.2.3.4 Weighted Histogram Analysis Method

The umbrella sampling methods require the calculation of the sampling distribution of the reaction coordinates. In many instances, multiple histograms are calculated from different simulations with different potential biases. These biases can be used to reach a stage where the reaction coordinates are sampled uniformly in one long (or many shorter) simulations and then the final histogram and bias can be used to calculate the sampling distribution and in turn the free energy surface.

The calculation can, however, be sped up if one takes into account all previous simulations with their different biasing potentials and resulting histograms of reaction coordinate sampling. The Weighted Histogram Analysis Method by Kumar *et al.*[14, 15] is a method to combine multiple histograms into a single one in this context. This is done so that all previous histograms that resulted from the simulations, each with different biases, can be used and that no information is discarded.

To combine the histograms from different simulations, each has to be converted into an unbiased histogram. These are combined with the correct weights to obtain the total unbiased probability distribution $P(R)$. For N simulations with biasing potentials $U_i(R)$, resulting histograms $n_i(R)$ and $N_i = \sum_R n_i(R)$, for each $i = 1, \cdots, N$, the equation for the total unbiased probability distribution is

$$P(R) = \frac{\sum_{i=1}^{N} n_i(R)}{\sum_{i=1}^{N} N_i \exp\left(\left[F_i - U_i(R)\right] / k_B T\right)}. \tag{3.51}$$

Here F_i is given by

$$F_i = -k_B T \log\left\{\sum_R P(R)\exp\left[-U_i(R)/k_B T\right]\right\}. \tag{3.52}$$

Initially all F_i are set to zero. Equations (3.51) and (3.52) are applied iteratively until the maximum change in F_i between iterations is less than a tolerance ε. Then the WHAM equations, (3.51) and (3.52), are said to have converged. The WHAM equations are applied in place of step 3 in the algorithm described in §3.2.3.2.

3.3 Intermolecular Orientational Free Energies

The implementations of free energy methods described previously involve calculation of the free energies along reaction coordinates that are chosen such that they elucidate the free energy difference of the reaction, conformational change or other chemical conversion under investigation. This work, however, aims at computing the interaction free energies between two chosen molecules as a function of their relative positions and orientations. This produces considerations that may not be encountered in typical free energy calculations from molecular dynamics simulations.

One such consideration is how to choose relevant reaction coordinates with which to parameterise the free energy surfaces that would describe relative interactions in sufficient detail. The most immediate choice to parameterise such an interaction would be the relative Cartesian coordinates of the molecular centres and relative orientation described by the Euler angles. This set thus comprised of six independent variables with which to parameterise the free energy surface and under the assumption of the molecules being mostly rigid, totally describes all possible relative positions and orientations. Three of these consist of relative Cartesian coordinates and three are angular coordinates.

The use of the umbrella sampling method requires that the forces arising from the derivative of the applied potential with respect to the different reaction coordinate

variables need to be applied to the atoms of the molecules under investigation in the molecular dynamics simulation. The second consideration to take into account is that the forces arising from the derivatives of the applied potential are relevant to the molecule as a whole, but in the molecular dynamics that are run, the forces are all applied to the atoms that make up the molecule. Furthermore, these atoms are described by their Cartesian coordinates, and thus the forces that are integrated to calculate the trajectories are all in Cartesian coordinates. In the umbrella sampling method that we employ, a conversion of the forces that arise from the reaction coordinate variables to ones that can be applied to the atoms of the molecules must be done.

One of the main implications of the description that we have chosen is that we assume that the molecules are rigid. The forces that arise must be applied in such a manner to the comprising atoms of the molecules such that the molecules do not deform. More explicitly, the conversion from derivatives of the applied potential, which apply as forces on molecules, to forces that are to be applied on atoms, must be done while considering the molecules as rigid bodies.

The force conversion procedure is thus specific to the description of the free energy surface that is chosen and must also ensure that the molecules in the simulation are not deformed. In the proceeding chapter, the description of the free energy surface is defined and the force conversion procedure is detailed.

3.4 Free Energy from Adaptive Reaction Coordinate Forces

To compute multidimensional free energy surfaces using the Free Energy from Adaptive Reaction Coordinate Forces (FEARCF) method we make use of histograms. Histograms are used to construct probability distributions from which we can calculate free energy surfaces. The histograms are multidimensional arrays, which represent the sampling distributions of the chosen parameters, where each array index corresponds to a dimension in the parameter space.

Each bin in the histogram is a volume in the chosen parameter space, thus the bin size directly influences the accuracy of the free energy surface. To achieve high accuracy many bins are thus needed, but this means that the method would suffer a similar limitation as the multidimensional methods described in §3.2.2.1 and §3.2.3.3; the number of required bins scales exponentially with the number of dimensions. Since these free energy surfaces are smooth, there is some redundancy in the stored information and high detail (many bins) is only needed for sharply detailed surfaces. This allows us to use cubic splines to describe the free energy surface. By employing cubic splines we also enforce differentiability of the surface, which is required for the method, and are also able to use fewer bins in a dimension.

From equation (3.19) we see that the free energy is calculated from the logarithm of the partition function. The partition function is a sum over macroscopic states (equation (3.11)) and each individual term of the sum is proportional to the probability that these states will be visited in the associated ensemble. A histogram of the sampled states from a simulation is then a measure of the same probability and thus can be used to calculate the free energy.

Since the conformations are canonically distributed we can write their distribution as

$$\rho(U,N,V,T)=\frac{\exp(-\beta U)\Omega(N,V,U)}{Z(N,V,T)}. \tag{3.53}$$

The probability of finding the system at temperature T with N number of particles in a macrostate with energy $U \pm dU/2$ is then

$$\rho(U,N,V,T)dU\,. \tag{3.54}$$

In a simulation in the canonical ensemble we set $\Delta U \neq 0$ and compute the sampling distribution $f(U)$, which is the number of times the energy in the range $[U-\Delta U/2, U+\Delta U/2]$ is sampled. The sampling distribution can be converted into a normalised energy distribution using:

$$\bar{\rho}(U) = \frac{f(U)}{\Delta U \sum_{U'} f(U')}. \tag{3.55}$$

As the bin width $\Delta U \to 0$ and $\sum f(U) \to \infty$ we have that $\bar{\rho} \to \rho$. With judiciously chosen bin widths and sufficiently long simulations we can then obtain an estimate of the density of states with:

$$\bar{\Omega}(U) = \bar{\rho}(U)\exp(\beta U)Z(T) \tag{3.56}$$

Using this description we may include a biasing potential energy function U into the simulation and calculate the density of states (equation (3.56)) or equivalently the free energy surface (equation (3.46)) from the sampling histogram. By parameterising the biasing potential function with the reaction coordinates of interest, we can similarly calculate the free energy surface as parameterised by the reaction coordinates. The forces arising from the biasing potential are applied to the relevant atoms in the simulation, taking into consideration the discussion in §3.3.

To improve the sampling efficiency of the simulations different biases may be used in each simulation to overcome free energy barriers in the system. The optimum choice of the bias is, as discussed before, the negative of the free energy surface. Since we do not have this to start with, we can iteratively update our estimate of the surface from the previous simulations and apply the negative of that as the biasing potential in the proceeding simulation. The resulting histograms from the simulations can be combined using the WHAM technique (§3.2.3.4) to further improve the efficiency of the calculation.

The implementation of the FEARCF method that we used is provided in depth in Appendix 3-A.

3.5 Chapter Three References

1. Darve, E.; Pohorille, A., Calculating free energies using average force. *J. Chem. Phys.* **2001**, (115).
2. Darve, E.; Rodriguez-Gomez, D.; Pohorille, A., Adaptive biasing force method for scalar and vector free energy calculations. *J. Chem. Phys.* **2008**, (128), 144120.
3. Fixman, M., Classical Statistical Mechanics of Constraints: A Theorem and Application to Polymers. *Proc. Natl. Acad. Sci.* **1974**, 71, (8), 3050-3053
4. Boresch, S.; Karplus, M., The Jacobian factor in free energy simulations. *J. Chem. Phys.* **1996**, 105, (12), 5145-5154.
5. Trzesniak, D.; Kunz, A.-P. E.; van Gunsteren, A.-P. E., A Comparison of Methods to Compute the Potential of Mean Force. *ChemPhysChem* **2006**, 8, (1), 162-169.
6. Torrie, G. M.; Valleau, J. P., Nonphysical Sampling Distributions in Monte Carlo Free-Energy Estimation: Umbrella Sampling *J. Comp. Phys.* **1977**, 23, 187-199.
7. Berkowitz, M.; Karim, O. A.; McCammon J. A.; J., R. P., Sodium chloride ion pair interaction in water: computer simulation. *Chem. Phys. Lett.* **1984**, 105, (6), 577-580.
8. Belch, A. C.; Berkowitz, M.; McCammon, J. A., Solvation structure of a sodium chloride ion pair in water. *J. Am. Chem. Soc.* **1986**, 108, (8), 1755-1761.
9. Mezei, M., Adaptive Umbrella Sampling: Self-consistent Determination of the Non-Boltzmann Bias. *J. Comp. Phys.* **1987**, 68, (1).
10. Ianuzzi, M.; Laio, A.; Parrinello, M., Efficient Exploration of Reactive Potential Enenergy Surfaces Using Car-Parinello Molecular Dynamics. *Phys. Rev. Lett.* **2003**, 90, (23), 238302.
11. Laio, A.; Parrinello, M., Escaping free-energy minima. *Proc. Natl. Acad. Sci.* **2002**, 99, (20).
12. Martonak, R.; Laio, A.; Parrinello, M., Predicting Crystal Structures: The Parrinello- Rahman Method Revisited *Phys. Rev. Lett.* **2003**, 90, (7).
13. Micheletti, C.; Laio, A.; Parrinello, M., Reconstructing the Density of States by History-Dependent Metadynamics. *Phys. Rev. Lett.* **2004**, 92, (17).
14. Kumar, S.; Payne, P. W.; Vasquez, M., Method for free-energy calculations using iterative techniques. *J. Comput. Chem.* **1996**, 17, (10), 1269-1275.
15. Kumar, S.; Rosenberg, J. M. D.; Bouzida, J.; Swendsen, R. H.; Kollman, P., The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *J. Comput. Chem.* **1992**, 13, (8), 1011-1021.

Appendix 3-A: Multidimensional Cubic Splines and Implementation

Cubic Splines

For any numerical description of a function we store values of the function at discrete points and use interpolation techniques to calculate the value of the function in between the discrete points. To calculate the forces arising from the biasing potential we need to be able to calculate smooth first derivatives of the potential. The natural cubic spline interpolation method is thus chosen since it has smooth first derivatives (by requiring continuity of the second derivatives).

A function $f(x)$ is known on discrete values of $x = \{x_i : i = 1, \cdots, N\}$. In other words we know

$$y_i = f(x_i) \quad \forall i = 1, \cdots, N. \tag{3.57}$$

To interpolate the function, a piecewise smooth cubic spline is fitted to the discrete data points. In other words, we approximate $f(x)$ by $S(x)$ where:

$$S(x) = \left\{ \begin{array}{lcl} S_1(x) & : & x_1 \le x \le x_2 \\ S_2(x) & : & x_2 \le x \le x_3 \\ \cdot & & \\ \cdot & & \\ \cdot & & \\ S_{N-1}(x) & : & x_{N-1} \le x \le x_N \end{array} \right\}. \tag{3.58}$$

There are three conditions that are satisfied for adjacent cubic polynomials $S_i(x)$ and $S_{i+1}(x)$:

1. $S_i(x_{i+1}) = S_{i+1}(x_{i+1}) \leftrightarrow$ Continuity of the function.
2. $S_i'(x_{i+1}) = S_{i+1}'(x_{i+1}) \leftrightarrow$ Continuity of the first derivative.
3. $S_i''(x_{i+1}) = S_{i+1}''(x_{i+1}) \leftrightarrow$ Continuity of the second derivative.

Using these conditions we can derive the equation for $S_i(x)$:

$$S_i(x) = y''_{i+1}\frac{(x-x_i)^3}{6h_i} + y''_i\frac{(x_{i+1}-x)^3}{6h_i} + \left[\frac{y_{i+1}}{h_i} - \frac{h_i y''_{i+1}}{6}\right](x-x_i) + \left[\frac{y_i}{h_i} - \frac{h_i y''_i}{6}\right](x_{i+1}-x), \tag{3.59}$$

with $x \in [x_i, x_{i+1}]$ and $h_i = x_{i+1} - x_i$. The continuity condition of the first and second derivatives fixes the values of y''_i for $i = 2, \cdots, N-1$ with

$$\frac{h_{i-1}}{6}y''_{i-1} + \frac{h_i + h_{i-1}}{3}y''_i + \frac{h_i}{6}y''_{i+1} = \frac{y_{i+1}-y_i}{h_i} - \frac{y_i - y_{i-1}}{h_{i-1}}. \tag{3.60}$$

The only remaining free parameters are y''_1 and y''_N. For natural splines, these are chosen to be zero. To calculate the value of $f(x)$, first y''_i need to be calculated using the natural spline conditions $y''_1 = y''_N = 0$ and equation (3.60). Then the value of $f(x)$ can be calculated using y_i, x_i, y''_i and equation (3.59).

The first derivative of $f(x)$ can easily be calculated from:

$$f'(x) = S'_i(x) = \frac{y''_{i+1}(x-x_i)^2}{2h_i} + \frac{y''_i(x_{i+1}-x)^2}{2h_i} + \frac{y_{i+1}+y_i}{h_i} - \frac{\left(y''_{i+1}+y''_i\right)h_i}{6}, \tag{3.61}$$

N-Dimensional Cubic Splines

In N-dimensions, the interpolation is done recursively by *splining out* each dimension to get the value of the function. To calculate $f(x_1, x_2, \cdots, x_N)$ from the stored values of $f\left(x_{1,i_1}, x_{2,i_2}, \cdots, x_{N,i_3}\right)$ (where $[i_1 = 1, \cdots, n_1]$, $[i_2 = 1, \cdots, n_2]$, $\cdots$, $[i_N = 1, \cdots, n_N]$) we use

the one dimensional spline method described above. Suppose that we have splined out dimensions 1 to $k-1$:

1. Calculate the values of the 2nd derivative in the direction of the k th dimension. In other words, solve the $n_k - 1$ equations arising from equation (3.60) for all values of x_{j,i_j}, where $j = k,\cdots,N$ and $i = 1,\cdots,n_j$ to get $f''(x_1, x_2, \cdots, x_{k-1}, x_{k,i_k}, \cdots, x_{N,i_N})$.
2. The k th dimension is then splined out using $f''\left(x_1, x_2, \cdots, x_{k-1}, x_{k,i_k}, \cdots, x_{N,i_N}\right)$ and equation (3.59) to get $f\left(x_1, x_2, \cdots, x_k, x_{k+1,i_{k+1}}, \cdots, x_{N,i_N}\right)$.
3. Set $k = k+1$ and repeat from step 1 till $k = N$.

To calculate the forces from the biasing potential to be applied in the simulations the first derivative in each dimension also needs to be calculated. To do this we repeat the above method N times, using equation (3.61) in step 2 instead of equation (3.59) and permuting the order in which the dimensions are splined out so that we can calculate $\partial f / \partial x_i$ for all $i = 1,\cdots,N$.

Implementation in CHARMM

The CHemistry At HARvard Molecular Mechanics (CHARMM) program implements a one dimensional umbrella sampling technique using the windowing method for generally defined reaction coordinates in its RXNCOR code. We modified modified this from a single dimension umbrella sampling method to a multidimensional adaptive umbrella sampling method up to and including six dimensions. To do this numerous modifications had to be made to the code, this included:

- 1 to 6 dimensional cubic spline routines (*rxnene.src*)
- The reaction coordinate calculations and application of the forces that arise from the biasing potential (*rxnene.src*)
- The implementation of the modified reaction coordinate definitions (*rxndef.src*)

To run a molecular dynamics simulation in CHARMM we need to:

1. Specify the force field parameters.
2. Define the atoms and molecules in the simulation.
3. Specify the coordinates of the atoms that make up the molecules.
4. Specify the simulation conditions and integration method
5. Start the simulation.

The running simulation calculates at each dynamics step the updates to the momentum and positions according to the forces on the atoms. At each dynamics step, the energy must also be calculated, which is done in CHARMM by calling the ENERGY routine.

In calculating the free energy surface from CHARMM simulations using the umbrella sampling method some modifications are made to this procedure. Between steps 3 and 4, described above, there is a further step inserted that defines the reaction coordinates from the set of atoms defined in step 2. The further change in the procedure is in the running simulation. In addition to calculating the energy, momentum and position of each atom, the value of the reaction coordinates need to be calculated and the energy and forces from the biasing potential need to be calculated. This is implemented in CHARMM by calling the RXNENE subroutine from the ENERGY routine. The RXNENE routine uses the ASCEND routine to calculate the values of the defined reaction coordinates. The value of and forces from the biasing potential are then calculated using the multi-dimensional cubic spline method described in §3.4. The energy is added to the total energy in the system and the forces are applied to the atoms involved in the reaction coordinate definitions using the DECEND0 routine.

describe important phenomena such as ion pairing.[14] However, since the molecular rotation is folded into the overall configurational sampling important information describing relative molecular orientation is lost in this free energy profile.

Regardless of the many successes, computer simulations employing atomistic force fields are unable to fully explore conformational and configurational space adequately to model complex mechanisms such a protein folding. It may be possible to use course grain simulations where orientation dependent intermolecular pair potentials provide access to experimentally observable mechanisms on microsecond and longer timescales. However, using this approach the atomic accuracy embedded in molecular mechanics force fields is lost at the expense of significantly improving configurational sampling. It may be possible to gain reasonable anisotropic intermolecular accuracy from a four-dimensional potential energy functional form that incorporates molecular orientation dependence. The fundamental variables of this functional form are illustrated with two water molecules in Figure 4-1. The function is shown as two interacting vectors having as fundamental variables, the internuclear distance (r), the molecular orientations for the first and second water molecules (θ_1 and θ_2) and their relative molecular rotation (ϕ). This intermolecular potential is used for interacting ellipsoids in a generalized form of the Gay-Berne potential[15, 16] and in molecular terms represent amino acid side-chain interactions in coarse grain protein simulations.[17] Here a generalised force based PMF method is introduced from which orientationally dependent molecular association free energies may be investigated.

4.2 The Orientational PMF

Dipole interactions are typically described by two degrees of freedom i.e., the separation of the centres and the angle between them. This is a sparse description and contains not much more information than W(r). A complete depiction of intermolecular interactions requires six degrees of freedom, i.e., the relative Cartesian coordinates and the relative orientation specified by the Euler angles. While it is possible to calculate this six dimensional PMF using the methods presented here, it

demands computational resources out of reach of present day midrange computational technologies and is therefore impractical.

A reduced, but useful description is the 4-dimensional $W(r,\theta_1, \theta_2, \phi)$. Here the parameters are distance between centres of masses (r), the molecular vector angles (θ_1 and θ_2) and their relative orientation (ϕ) as described in Figure 4-1 and mentioned above. However, there is a loss of configurational information since the rotations about the semi major axes of the ellipsoids (see for example the broken lines shown in Figure 4-1) are to be averaged into the sampling. Subsequently there have been several attempts to derive accurate potentials of mean force (PMFs) for side chain interactions, which include their relative orientation. Here so called knowledge based potentials, using the frequencies of amino acid pairing from the Protein Data Bank (PDB)[18] was proposed by Tanaka and Scheraga[19]. These have been calculated from distribution functions that have been constructed from PDB structures.[8, 20, 21] These potentials are invaluable as they include atomic level information such as environmental entropic effects due to protein conformational and condensed phase configurational sampling. However, they are heavily biased as they are compiled from only solved protein structures that are an incredibly small fraction of all natural proteins. Furthermore it has been formally shown that PDB derived PMFs are neither the potentials nor the PMFs for interactions between pairs of amino-acid residues.[22]

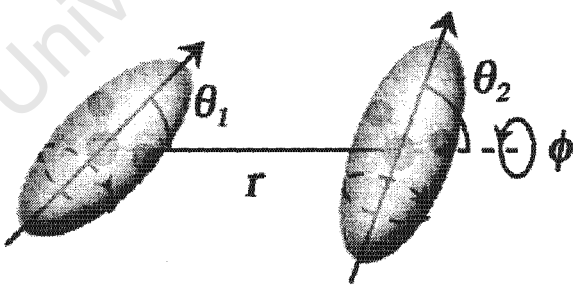

Figure 4-1: Vector interaction parameterized by r, θ_1, θ_2 and ϕ. The broken lines indicate configurations that cannot be identified from the surfaces as they folded into W(r, θ_1, θ_2 and ϕ).

The four dimensional intermolecular PMF for molecular pairs, $W(r, \theta_1, \theta_2, \phi)$, is calculated as an inter vector orientational potential using an adaptive reaction coordinate force based method. From here on the general method is referred to as the

Free Energies from Adaptive Reaction Coordinate Forces (FEARCF) method. These orientationally dependent PMFs can be calculated for any molecular pair in any environment by choosing a vector to that will yield the most physical and chemical information when the four reaction coordinates are deconvoluted into the atoms making up the molecules. Here the PMF surfaces for the rigid three-, four- and five-site TIP3P[23], TIP4P and TIP5P[24] models are calculated. To do this a vector is chosen such that it lies along the water dipole (Figure 4-1), i.e. it lies along the plane made from the atomic centres of the Oxygen and two Hydrogens. This is to test if the multidimensional free energy hypersurfaces are sensitive to relatively minor differences in potential functions and if they are able to reveal important anisotropic details of molecular association.

4.3 The FEARCF Method

This adaptive reaction coordinate force biasing method is a generalization of the method used to produce two dimensional conformational[25, 26] and reaction[27] free energy surfaces. The concept of adjusting the biasing potential iteratively to evolve the free energy was first presented by Mezei as adaptive umbrella sampling.[28] Other methods have subsequently employed this concept and extended it using either forces to directly bias the system or potential biasing. Of particular relevance here is Darve and Pohorille's adaptive biasing force routine for thermodynamic integration that continuously updates the biasing forces at every step during the simulation.[29, 30] A popular adaptive umbrella potential method developed by Laio and Parrinello uses a combination of Gaussians as a biasing potential from which course grained forces are calculated to direct the system to regions that have not been previously sampled.[31]

The FEARCF method is based on probability distributions and histograms. The reaction coordinate space is a discretized n-dimensional grid were the sampling frequency for a bin site is recorded for each simulation. In the case of molecular pair association a four dimensional grid is required. The population of this grid represents a running tally of the reaction coordinate probability density, derived from the history of simulations to that point. It is used as input for a multidimensional cubic-spline interpolation from which the reaction coordinate biasing forces are calculated. These

forces are applied to all atoms used in the reaction coordinate definition to bias the next simulation's reaction coordinate trajectory away from previously sampled areas. The entire reaction coordinate space is equally sampled when the biasing forces are derived from the true PMF. The forces are applied on Cartesian coordinates and therefore the PMF does not need logarithmic Jacobian corrections.[32-35] The method requires no intervention from the user other than to make a judicious choice of reaction coordinate, and simulation length at each update of the biasing force. In general a combination of relatively short simulation times, initiated from well separated positions on the reaction coordinate surface at the start of FEARCF, followed by increasing these to longer (x3) times as convergence is reached, is recommended.

The macromolecular program CHARMM[36] was modified to calculate the effect of the perturbing forces generated from any multidimensional reaction biasing potential $U(\xi)$. To make this more instructive, it is described here for the four dimensional orientational PMF representing the Free energy of association between two molecules. In this case there are four independent reaction coordinates $\xi = (\xi_1 = r, \xi_2 = \theta_1, \xi_3 = \theta_2, \xi_4 = \phi)$. The system Hamiltonian is modified by adding $U(\xi)$ to the unbiased Hamiltonian H_0,

$$H(\xi) = H_0 + U(\xi). \tag{4.1}$$

The biased Hamiltonian $H(\xi)$ is then used in the simulation instead of H_0 and so generates a biased probability distribution of sampled coordinates $P'(\xi)$. This $P'(\xi)$ can then be converted to an unbiased probability distribution by accounting for the biasing potential as follows:

$$P'(\xi) = CP(\xi)\exp\left(-\frac{U(\xi)}{k_B T}\right), \tag{4.2}$$

where C is the normalization constant, k_B is the Boltzmann constant and T is the temperature of the system. The PMF for a calculated unbiased probability distribution is then

$$W(\xi) = -k_B T \log P(\xi). \tag{4.3}$$

At each step of the simulation the biasing forces for ξ_i are applied to the atoms involved in the definition of the molecular vectors (for example as shown in Figure 4-1). For 4-dimensional surfaces these are

$$F(\xi) = \sum_i -\frac{\partial U(\xi)}{\partial \xi_i}. \tag{4.4}$$

The PMF, $W(\xi)$, is then calculated from the probability distribution of the reaction coordinate assembled from the sampling of all trajectories to that point. If the negative of the PMF is iteratively added after each simulation to the unbiased Hamiltonian, the estimate of the PMF is improved. The umbrella potential is used to calculate the biasing forces, which expands the sampling to all parts of the reaction coordinate ξ. Convergence is reached when we observe uniform sampling in a simulation throughout ξ space. The best biasing potential is therefore when $U(\xi) = -W(\xi)$ which will result in the system sampling the entire coordinate space. This provides us with an accurate check that the PMF is correct.

The force arising from the biasing potential is calculated by treating the molecules, which were used in the reaction coordinates definition, as rigid bodies. This ensures that during the simulation, the atoms within each molecule do not move apart and deform the molecules due to differing accelerations that arise from the biasing potential. Consequently when the resultant acceleration for these molecules is calculated as a whole, the forces on the atoms were determined such that the atoms experience the same resultant acceleration.

The force arising from the biasing potential is calculated by treating the molecules, which were used in the reaction coordinates definition, as rigid bodies. This ensures that during the simulation, the atoms within each molecule do not move apart and deform the molecules due to differing accelerations that arise from the biasing potential. Consequently when the resultant acceleration for these molecules is calculated as a whole, the forces on the atoms were determined such that the atoms experience the same resultant acceleration.

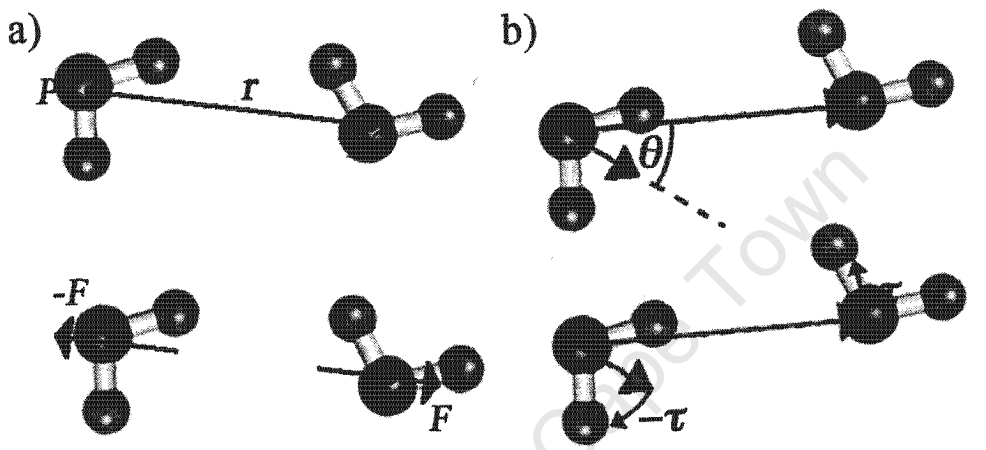

Figure 4-2: (a) Translational force applied to each water molecule as calculated from the partial derivative of the adaptive biasing potential. (b) Illustration of the torque applied to each water molecule as calculated from the partial derivative of the adaptive biasing potential

A translational force arises from a reaction coordinate, which is defined as a distance between two points P_1 and P_2 (Figure 4-2a). A positive force in this case should increase the distance, and a negative deceases it by applying equal but opposite forces to the points that define the molecules. The force arising from the biasing potential for the reaction coordinate ξ_1 (i.e., the distance between P_1 and P_2) is $F(r=d)=-\nabla U|_{r=d}$ which results in accelerations $a_{P1}=-F/M_{P1}$ and $a_{P2}=F/M_{P2}$ on each molecule. Here M_{P1} and M_{P2} are the total masses of the molecules at P_1 and P_2. The accelerations, a_{P1} and a_{P2}, are then spread equally amongst the atoms of the respective molecules, such that the k'th atom of the molecule at P_1 experiences a force

$$f_k(\xi_i) = \frac{F(\xi_i)}{M} m_k \; ; i = 1, \tag{4.5}$$

where m_k is the mass of atom k.

In the case where the reaction coordinate is an angle between the two molecular vectors, an axial vector is defined about which the angle of rotation is calculated. The derivative of the potential with respect to the angular coordinate is a torque that is to be applied about the axial vector. In general for rotational motion, ($\xi_2=\theta_1$, $\xi_3=\theta_2$, $\xi_4=\phi$) the resultant angular acceleration about the pivoting axes is calculated so that the torque τ on a molecule with a moment of inertia I, can be used to calculate the force on an atom k

$$f_k(\xi_i) = \frac{\tau(\xi_i)}{I} r_k m_k \; ; \mathrm{i} = 2,3,4, \tag{4.6}$$

where r_k is the distance from the axis of rotation to atom k.

For the reaction coordinates $\xi_2=\theta_1$ and $\xi_3=\theta_2$ the torque is applied to each molecule about the axial vector passing through the molecule's centre of mass. The magnitude of the torque is then calculated by $\tau(\theta) = -\frac{\partial U}{\partial \theta}$. The points of rotation about which to apply the torque are the beginning points of each vector. In the $\xi_2=\theta_1$ and $\xi_3=\theta_2$ cases respectively the points are located on the centres of mass of the water molecules. The axis, about which the rotation is applied, is perpendicular to both vectors defining the angle and projects out from the page (Figure 4-2b).

For $\xi_4=\phi$ the torque is applied to each molecule about the axial vector connecting the two molecular vectors and also rotate about the respective molecules' centres of mass.

The total force on atom k is in the molecule a that comprises N atoms is

$$F_k = -\frac{\partial U}{\partial r}\frac{m_k}{\sum_{i=1}^{N} m_i} - \frac{\partial U}{\partial \theta_1}\frac{m_k r_k}{\sum_{i=1}^{N} m_i r_i^2} - \frac{\partial U}{\partial \theta_2}\frac{m_k r_k}{\sum_{i=1}^{N} m_i r_i^2} - \frac{\partial U}{\partial \phi}\frac{m_k r_k}{\sum_{i=1}^{N} m_i r_i^2}. \quad (4.7)$$

The following section defines the exact atomic forces on all N_1 atoms of molecule 1 and N_2 atoms of molecule 2. To do this, the following vectors are defined: $\mathbf{r}$ – from molecule 1 to molecule 2, $\hat{\mathbf{d}}_1$ – the dipole unit-vector along molecule 1 and $\hat{\mathbf{d}}_2$ the dipole unit-vector along molecule 2. The four reaction coordinates that are used are then defined in terms of these vectors as:

$$\begin{aligned}
\xi_1 &= r = |\mathbf{r}| \\
\xi_2 &= \theta_1 = \arccos\left(\hat{\mathbf{d}}_1 \cdot \hat{\mathbf{r}}\right) \\
\xi_3 &= \theta_2 = \arccos\left(\hat{\mathbf{d}}_2 \cdot \hat{\mathbf{r}}\right) \\
\xi_4 &= \phi = \arccos\left(\frac{\left(\hat{\mathbf{d}}_1 \times \mathbf{r}\right)\cdot\left(\hat{\mathbf{d}}_2 \times \mathbf{r}\right)}{\left|\left(\hat{\mathbf{d}}_1 \times \mathbf{r}\right)\right|\left|\left(\hat{\mathbf{d}}_2 \times \mathbf{r}\right)\right|}\right)
\end{aligned} \quad (4.8)$$

The Cartesian force arising from the derivative of the potential with respect to r is then applied to the atoms of the molecules as:

$$\mathbf{f}_{1,k}^{r} = -\frac{\partial U}{\partial r}\frac{m_k}{\sum_{i=1}^{N_1} m_i}(-\hat{\mathbf{r}}), \quad \mathbf{f}_{2,k}^{r} = -\frac{\partial U}{\partial r}\frac{m_k}{\sum_{i=1}^{N_2} m_i}(\hat{\mathbf{r}}) \quad . \quad (4.9)$$

with $\mathbf{f}_{1,k}^{r}$ being the force on the atom k of molecule 1 and $\mathbf{f}_{2,k}^{r}$ being the forces on atom k of molecule 2.

The Cartesian forces that arise from θ_1, θ_2 and ϕ are calculated using the equation that describes relation between the torque arising from a Cartesian force at a point in the body: $\boldsymbol{\tau} = \mathbf{q} \times \mathbf{F}$ (where q is the vector from the torque axis to the point). Since $\boldsymbol{\tau}$ and $\mathbf{q}$ are construct (shown below) perpendicular to each other we can write $\hat{\mathbf{F}} = \hat{\boldsymbol{\tau}} \times \hat{\mathbf{q}}$.

The torque unit vector for θ_1, θ_2 is calculated with $\hat{\boldsymbol{\tau}}_i^{\theta} = \hat{\mathbf{d}}_i \times \mathbf{r} / \left|\hat{\mathbf{d}}_i \times \mathbf{r}\right|$ for $i = 1,2$.

We then use $\mathbf{q} = \mathbf{r}_{i,k}$ for $i = 1,2$, the perpendicular distance from the axis along $\hat{\boldsymbol{\tau}}_i^{\theta}$

positioned at the centre of mass of molecule 1 (for θ_1) or molecule 2 (for θ_2), to all the atoms of molecules 1 and 2. The force due to the derivatives of the biasing potential with respect to θ_1 and θ_2 for the atoms on molecule 1 is:

$$\mathbf{f}_{1,k}^{\theta} = -\frac{\partial U}{\partial \theta_1}\frac{m_k r_{1,k}}{\sum_{i=1}^{N_1} m_i r_{1,i}^2}(-\hat{\boldsymbol{\tau}}_1^{\theta} \times \hat{\mathbf{r}}_{1,k}) - \frac{\partial U}{\partial \theta_2}\frac{m_k r_{2,k}}{\sum_{i=1}^{N_1} m_i r_{2,i}^2}(\hat{\boldsymbol{\tau}}_2^{\theta} \times \hat{\mathbf{r}}_{2,k}), \tag{4.10}$$

and for molecule 2 is:

$$\mathbf{f}_{2,k}^{\theta} = -\frac{\partial U}{\partial \theta_1}\frac{m_k r_{1,k}}{\sum_{i=1}^{N_2} m_i r_{1,i}^2}(\hat{\boldsymbol{\tau}}_1^{\theta} \times \hat{\mathbf{r}}_{1,k}) - \frac{\partial U}{\partial \theta_2}\frac{m_k r_{2,k}}{\sum_{i=1}^{N_2} m_i r_{2,i}^2}(-\hat{\boldsymbol{\tau}}_2^{\theta} \times \hat{\mathbf{r}}_{2,k}). \tag{4.11}$$

For ϕ we use $\hat{\boldsymbol{\tau}}^{\phi} = \hat{\mathbf{r}}$ and $\mathbf{q} = \mathbf{r}_{i,k}^{\phi}$ for i =1,2 with $\mathbf{r}_{1,k}^{\phi}$ being the perpendicular distance from the axis along $\hat{\boldsymbol{\tau}}^{\phi}$ positioned at the centre of mass of molecule 1 and $\mathbf{r}_{2,k}^{\phi}$ being the perpendicular distance from the axis along $\hat{\boldsymbol{\tau}}^{\phi}$ positioned at the centre of mass of molecule 2. The forces on the atom k of molecule 1 arising from the derivative of the biasing potential with respect to ϕ is

$$\mathbf{f}_{1,k}^{\phi} = -\frac{\partial U}{\partial \phi}\frac{m_k r_{1,k}^{\phi}}{\sum_{i=1}^{N_1} m_i \left(r_{1,i}^{\phi}\right)^2}(-\hat{\boldsymbol{\tau}}^{\phi} \times \hat{\mathbf{r}}_{1,k}^{\phi}), \tag{4.12}$$

and for molecule 2 it is:

$$\mathbf{f}_{2,k}^{\phi} = -\frac{\partial U}{\partial \phi}\frac{m_k r_{1,k}^{\phi}}{\sum_{i=1}^{N_2} m_i \left(r_{1,i}^{\phi}\right)^2}(\hat{\boldsymbol{\tau}}^{\phi} \times \hat{\mathbf{r}}_{2,k}^{\phi}). \tag{4.13}$$

The total force in Cartesian coordinates on atom *k* of molecule *i* is then

$$\mathbf{F}_{i,k} = \mathbf{f}_{i,k}^{r} + \mathbf{f}_{i,k}^{\theta} + \mathbf{f}_{i,k}^{\phi}. \tag{4.14}$$

Combining histograms of previous simulations and weighting them appropriately can significantly increase the rate of achieving convergence in the adaptive umbrella sampling method. Kumar et al. devised a weighted histogram analysis method (WHAM)[37-41] that allowed optimal weighting of previous simulation data by recasting the Ferrenberg-Swendsen multiple histogram equations.[42] The histograms for differently biased simulations were combined into an unbiased distribution by applying the correct weights calculated from the WHAM equations. Adequate sampling is generally accepted to be achieved when the ratio of most sampled to least sampled states is 1:50 or better.[27] Using the adaptive reaction coordinate forces method this minimum criterion is exceeded and have reported ratios of 1:5 resulting in a greater accuracy of the calculated free energy surfaces.[25]

4.4 Results and Discussion

A key measure of the accuracy of the water model potential function is the extent to which it can produce local tetrahedral order in the first hydration shell in the condensed phase as observed from neutron diffraction partial structure factors.[1] Here we compare the 1 dimensional and 4 dimensional PMFs of rigid three-, four- and five-site TIP3P, TIP4P[23] and TIP5P[24] models with results taken from 5 ns MD liquid simulations of the same models. The simulations were done in NVT ensembles of 512 water molecules in 24.8 Å cubic boxes where periodic boundary conditions were applied.

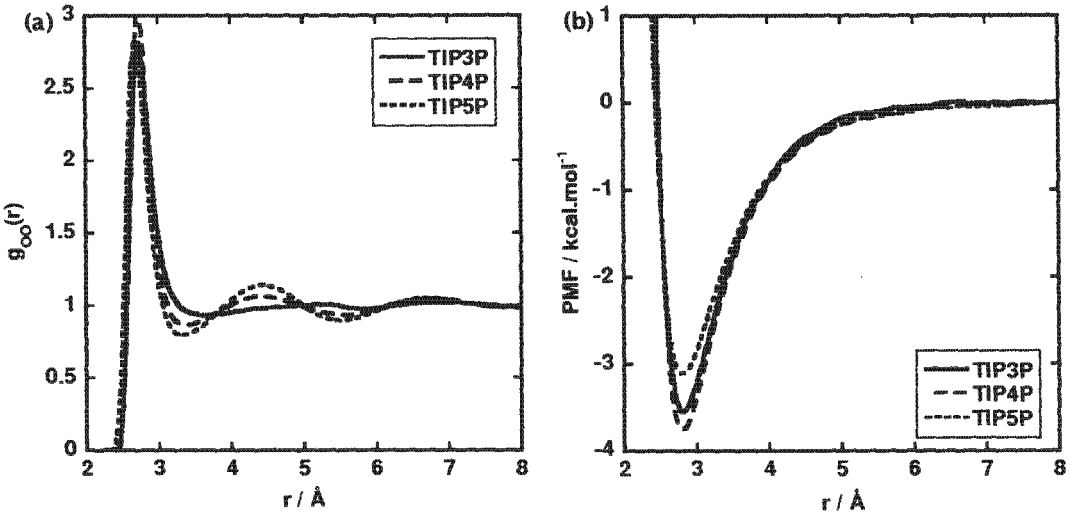

Figure 4-3: (a) $g_{O\text{-}O}(r)$ RDF for TIP3P, TIP4P and TIP5P liquid bulk water; (b) W(r) for two individual water molecules in vacuum.

The structures that the various models form in bulk are typically measured with radial distribution functions (RDFs) shown in Figure 4-3a. This can be compared with the radially averaged PMFs, W(r), for single water dimers of these models shown in Figure 4-3b to understand the relationship between the intermolecular free energy and the condensed phase structure. The radially averaged structures of the first solvation shell are very similar, as inferred from the first peak in the gO-O(r) RDFs appearing at 2.7 Å, for all three models (Table 4-1). This corresponds to the intermolecular attractive free energies observed in the W(r) profiles of 3.50 kcal/mol, 3.69 kcal/mol and 3.08 kcal/mol for TIP3P, TIP4P and TIP5P respectively each at an equilibrium distance of 2.80 Å. The second RDF peak provides information about the most probable distances between the neighbors of a central water molecule. While this is absent for TIP3P both TIP4P and TIP5P gO-O(r) RDFs have broad peaks centred at ~4.5 Å.

Table 4-1: Geometries of the first hydration shell taken from distribution functions.

Tetrahedrality Metrics	TIP3P	TIP4P	TIP5P
r (Å)	2.7	2.7	2.7
(w_D-w_C-w_D) (°)	99.3	105.3	101.3
(w_D-w_C-w_A) (°)	117.5	106.5	115.9
(w_A-w_C-w_A) (°)	88.0	121.5	95.0
q	0.78	0.78	0.73

w_D, water donating a hydrogen bond; w_A, water accepting a hydrogen bond; w_C, central water

The distance dependent description of the first solvation shell and the free energy calculated for a pair of water molecules provides ambiguous information about the nature of the first hydration shell. The structure of the first neighbour shell of liquid water could be understood by separating the translational and orientational order.[43] This can be done using two order parameters introduced by Tanaka[44, 45] where the orientational order parameter

$$q_i = 1 - \frac{3}{8}\sum_{j=1}^{3}\sum_{k=j+1}^{4}\left[\cos\theta_{ijk} + \frac{1}{3}\right]^2, \qquad (4.15)$$

is a rescaled version of the angular parameter of Chau and Hardwick[46, 47] used to quantify the tetrahedrality of the first shell. Here θ_{ijk} is the angle between the central water molecule i and two of its neighbors j and k. The average orientational order for each model is 0.777, 0.780 and 0.733 for TIP3P, TIP4P and TIP5P respectively where a value of q = 1 indicates perfect tetrahedrality. Based on this measure there is no difference in the local symmetry and order exhibited by the three models. The local orientational order that the water models exhibit in bulk can be illustrated from observations of the spatial distribution functions (SDFs). These were calculated for TIP3P, TIP4P and TIP5P and are shown in Figure 4-4 where the white and red contours represent hydrogen and oxygen probability densities. While there are four

hydrogen bond sites there are two symmetric areas *viz,* sites where the central water contributes a hydrogen and one where it accepts a hydrogen via its electron lone pairs. In all three models the hydrogen donating sites are quantitatively equivalent. The hydrogen probability densities in the TIP3P and TIP4P models are diffuse while in the case of the TIP5P model the hydrogen probability densities reveal a directional preference for the lone pair sites of the central water. This is despite the W(r) minima of TIP5P (Figure 4-3b) being indicative of a lower overall attraction between pairs of water molecules. Measuring the angles formed from the maxima of the oxygen contours to the central oxygen we derive a quantitative measure of the water hydrogen bond donor (w_D) and water hydrogen bond acceptor (w_A) orientation about the central molecule. From this a comparison of the extent of the local order in the three models (Table 4-1) can be made.

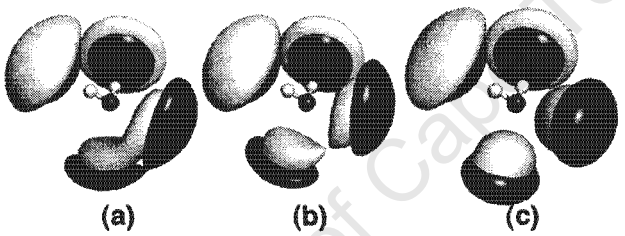

Figure 4-4: Spatial Distribution Functions for (a) TIP3P, (b) TIP4P and (c) TIP5P. White represents hydrogen probability and dark grey oxygen probability

While the SDF plots indicate that the TIP5P model best reproduces tetrahedrality the measurement of the angles made between the oxygen contour maxima (Table 4-1) imply that TIP4P best displays local water ordering as observe in experiments. Simultaneously the orientational order parameter is unable to distinguish between the three models.

To understand the effect of the pair potential on the local ordering of the bulk liquid we calculate four dimensional orientational PMFs $W(r, \theta_1, \theta_2, \phi)$ for water dimers by placing vectors along each water dipole (i.e., along the oxygen and bisecting the HOH angle). The parameters are as defined in Figure 4-2 resulting in orientation dependent PMFs for the TIP3P, TIP4P and TIP5P water models. Since it is not possible to visualize the PMF in five dimensions we extract 2 dimensional subsets of the free

energy surface (Figure 4-5) by averaging over the remaining ensemble parameters. This is done for combinations of reaction coordinates where $\xi_i = r, \theta_1, \theta_2, \phi$ and $\xi_1 \neq \xi_2 \neq \xi_3 \neq \xi_4$ such that

$$W(\xi_1, \xi_2) = -k_B T \log\left[\sum_{\xi_3, \xi_4} P(\xi_1, \xi_2, \xi_3, \xi_4)\right]. \tag{4.16}$$

Since $W(r, \theta_1, \theta_2, \phi)$ is calculated between identical molecules we do not show $W(r, \theta_2)$ and $W(\theta_2, \phi)$. Furthermore the interchange of water molecules 1 and 2 is equivalent to setting $\theta_1 = 180 - \theta_2$.

The values of r, θ_1, θ_2 and ϕ for the free energy favoured orientations are listed in Table 2 along with structural data from single point MP2 optimization calculations[48], Radio frequency and microwave experimental spectra[49] for water dimers. The energetically favored geometries extracted from the PMFs are similar to those reported for experimental water dimers[49] and correspond well with the probability distributions from bulk TIP3P and TIP4P simulations.[2]

Table 4-2: Favored Water Dimer Configurations

Classical Model	r / Å	θ_1 / °	θ_2 / °	$\bar{\phi}$°
TIP3P	2.79	27	54	180
TIP4P	2.80	45	59	180
TIP5P	2.78	49	53	180
QM / Experiment				
MP2 / (O:13s,8p,4d,2f H:8s,4p,2d)[48]	2.93	58.2	57.7	180
Experiment [49]	2.98	51±10	57±10	180

To achieve an ideal water tetrahedral geometry (i.e., an O···O···O bond angle of 109.47°) values of θ_1=52.25°, θ_2 =52.25°, and ϕ = 180° are expected. The variations

in the inter vector rotational angle, ϕ, as seen in the orientational PMF's reveal that TIP3P and TIP4P allows a significant movement away from the minimum ϕ-orientation compared with TIP5P. This configurational flexibility is well within the thermal $3K_BT$ envelope. Examination of the (θ_1, θ_2) minima for the three models shows that TIP3P is the least tetrahedral (θ_1 = (27°, 54°) while TIP4P shows an improvement in its tendency toward tetrahedrality (45°, 59°) and TIP5P being near ideal (49°, 53°)). These minima correspond to eight equivalent forms of the equilibrium dimer configurations with Cs symmetry. The equilibrium configurations and the transitions between them have been studied using Ab initio methods and reported in terms of the molecular symmetry (MS) group.[50, 51]

The differences between the TIP models observed in the SDFs which were taken from bulk phase water (Figure 4-4) can be understood in the context of the intermolecular orientational free energy by examining the barrier heights separating the water pairs as they exchange roles between being a hydrogen bond donor (1) and a hydrogen bond acceptor (2). We plot a minimum free energy contour path (Figure 4-6) between the minima of the $W(\theta_1, \theta_2)$ PMF's and represent this as a broken line joining the two minima in Figure 4-5g, h, and i. A local minima is present for all three models at the height of the reaction path barrier (θ_1 ,θ_2) ≈ (90,90) which corresponds to no hydrogen bonding but the most favorable (anti parallel) dipole–dipole orientation. The free energy barrier between the hydrogen bond favoured configurations (i.e., water 1 (donor) to water 2 (acceptor) and water 2′ (acceptor) to water 1′ (donor)) is highest (4.9 kcal/mol) for TIP5P. The TIP3P model has the next highest barrier (2.89 kcal/mol) with TIP4P being the lowest (1.97 kcal/mol). An analysis of infrared vibration-rotation-tunneling (VRT) spectroscopy[52] revealed the donor acceptor interchange rearrangement barrier of the water dimer to be 0.59 kcal/mol while calculations at the MP2/aug-cc-pVXZ (X = D, T, Q) level of theory gave a value of 0.8 kcal/mol.[51]

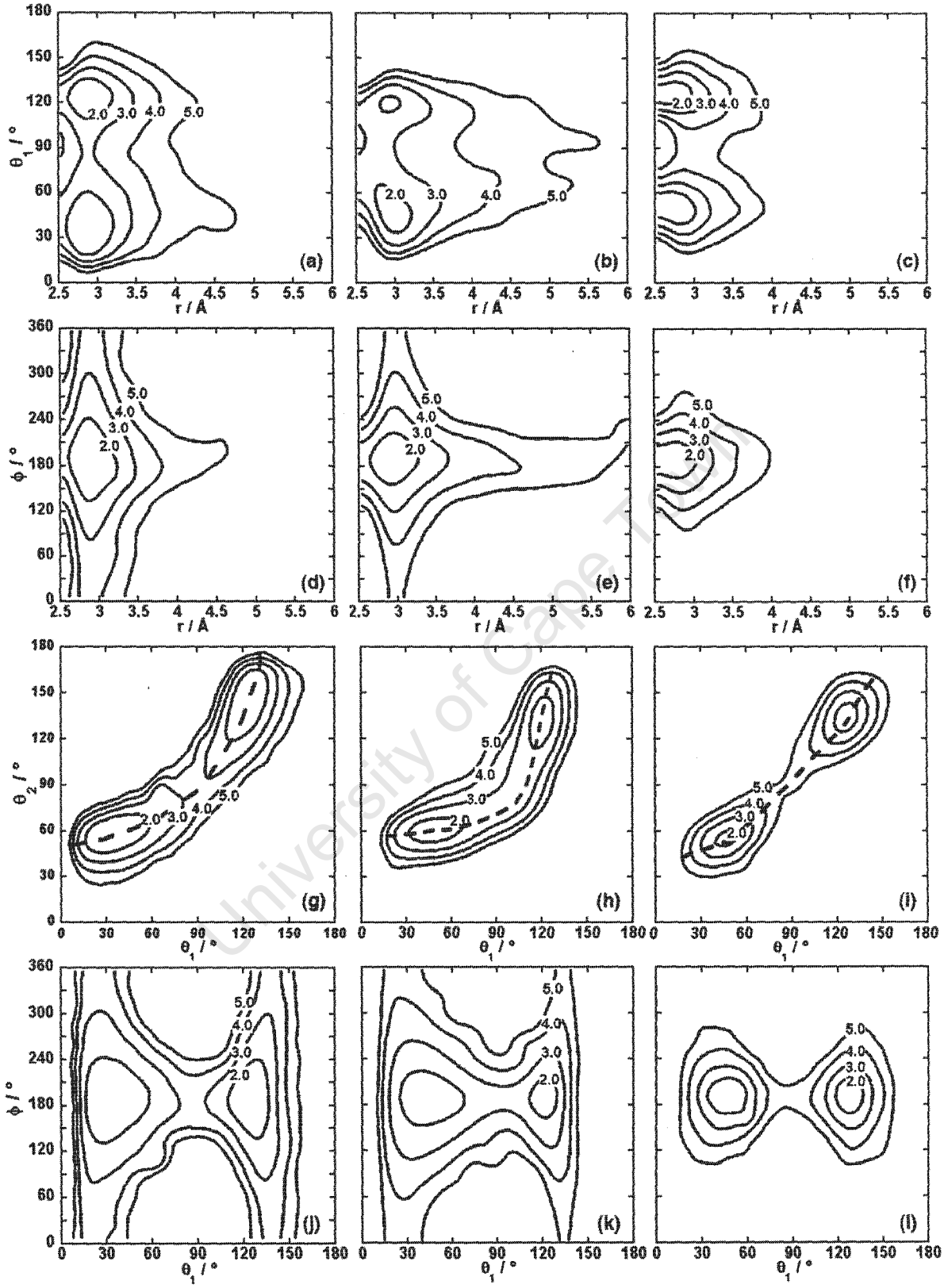

Figure 4-5: The 2D extracted PMF's for TIP3P (a),(d),(g),(j), TIP4P: (b),(e),(h), (k), TIP5P: and (c),(f),(i),(l). The contours are plotted at 1 kcal/mol intervals starting from 2 kcal/mol.

This activation barrier is for a transition between single hydrogen bonded (with C_s symmetry) equilibrium configurations going via a double hydrogen bonded transition state of C_i symmetry. Recently scanning tunneling microscopy experiments of single water dimers suggested that this mechanism involves quantum tunneling.[53] The activation barriers we extracted (Figure 4-6) from the multidimensional free energy of association surfaces (Figure 4-5) are not the same as those gained from the *ab initio* calculations.[50, 51] The TIP models negotiate the interchange rearrangement via an anti-aligned dipole minimum, which in the TIP5P case is near perfect ($\theta_1 = 89°$, $\theta_2 = 91°$). This corresponds to the doubly bifurcated (C_{2h}) structure which is a minimum on an AM1 potential energy surface but has been shown[54], using MP4 / 6-31+G(2d, 2p), to be a saddle point that is 3.75 kcal/mol above the energies of the equilibrium configurations. [50]

The reasons for this discrepancy between the $W(r, \theta_1, \theta_2, \phi)$ derived transition state and the transition state calculated from static ab initio optimizations are twofold. Firstly as we pointed out above our use of four instead of six parameters to describe intermolecular association do not include all possible molecular configurations. This is because the contributing configurations that arise from rotations about the molecular vectors C_2 symmetry axes (shown in Figure 4-2 as broken arrows) are averaged into the four-dimensional free energy surface. Specifically the C_i symmetry cyclic transition structures observed in a full molecular description will be folded into the equilibrium configuration free energy wells when using the inter vector definition shown in Figure 4-2. Secondly the TIP potentials are classical models and so are unable to accurately reproduce quantum effects present in these intermolecular associations. Therefore the expected ordering of the stationary points observed from a six dimensional free energy surface is not likely to coincide with the results of the detailed quantum studies[50, 51] as has been previously noted.[55]

It is not the objective of the present work to investigate via classical models the details of the water dimer potential energy surface. We intend here only to show the sensitivity of these novel multidimensional PMFs that include orientational information in distinguishing between three relatively similar classical water models. With this in mind we observe that the TIP5P model energetically favors stronger

intermolecular hydrogen bonding which should lead to less frequent rotational motion compared to the other TIP models. The accurate representation of tetrahedrality (Figure 4-4) of the TIP5P model and its ability to reproduce a density maximum of 4°C[4, 56] may be due to the higher free energy barrier it exhibits preventing frequent interchange from a hydrogen bond donor to that of an acceptor with neighbouring waters.

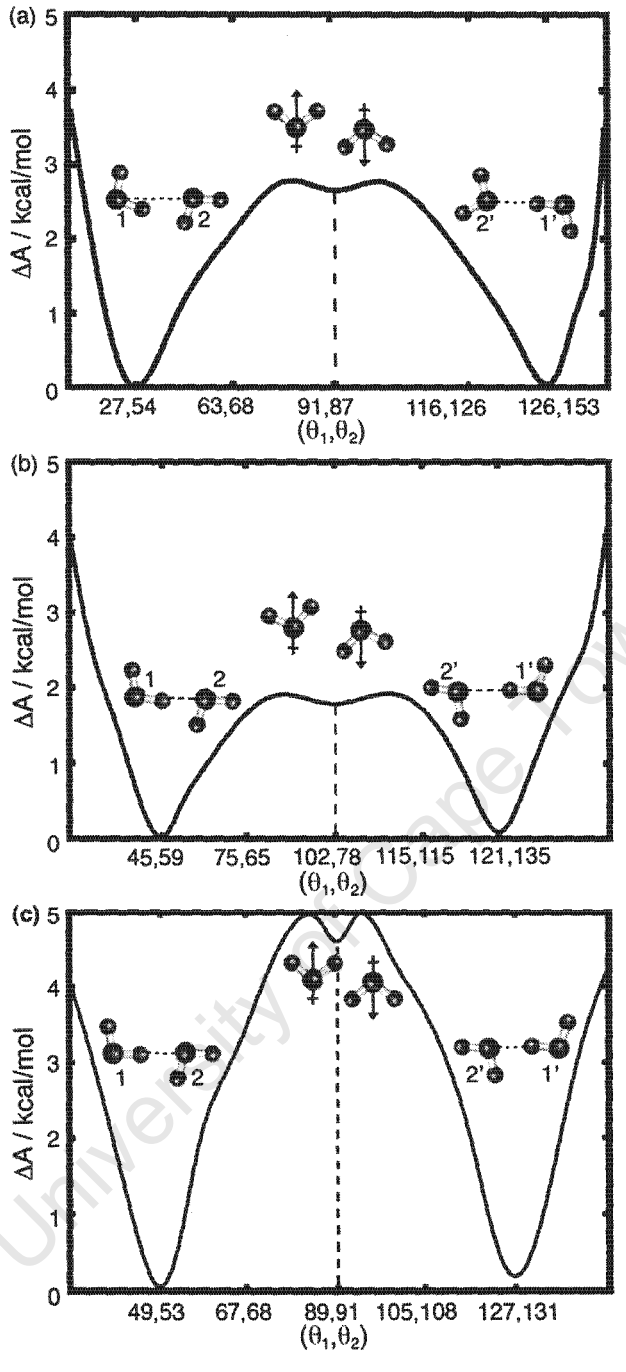

Figure 4-6: The minimum transition paths for (a) TIP3P along the broken line in Figure 4-5g (b): TIP4P along the broken line in Figure 4-5h and (c): TIP5P along the broken line in Figure 4-5i. The energy barriers separating the switch from a hydrogen bond donor to an acceptor are due to anti aligned dipole interactions.

4.5 Conclusions

The adaptive reaction coordinate force method, which we have previously developed and used to derive specific conformational and configurational reaction free energy surfaces, have been generalised here. The FEARCF approach gives access to a fully converged multidimensional free energy volume. While any PMF can be calculated using this method its use is particularly significant for the calculation of free energy volumes from atomic based forces expressed as derivatives of the reaction coordinate free energy surface. To illustrate the accuracy and versatility of the method we apply it to a demanding problem water dimer association and rotational dynamics. Although it was not our intention to match the molecular accuracy of previous high level *ab initio* investigations into the water dimer hydrogen bond interchange mechanism, we were able to show the origin of the discrepancies between the classical and quantum models. Excellent sampling of the conformational space is achieved demonstrating how high energy configurations are accessible using this adaptive reaction coordinate force to bias the trajectory about the reaction coordinate multidimensional volume.

The calculation of the multidimensional intermolecular orientational free energy calculated here illustrates the sensitivity and accuracy of a four-dimensional free energy surface in discriminating between relatively similar intermolecular water (TIP) potentials that display significantly different bulk configurations. The PMFs provide energetic reasons for the extent to which each model is able to reproduce local ordering in water. The calculation of the four dimensional orientational PMF can therefore provide an atomistic accuracy check for course grain potential functions such as Gay-Berne that are used in the simulation of materials such as liquid crystals and macromolecules such as proteins. The $W(r, \theta_1, \theta_2, \phi)$ bridges the gap between atomistic simulations with embedded molecular details and course grained molecular potentials that allow access to events on the microsecond-second timescale.

4.6 Chapter Four References

1. Soper, A. K., Orientational correlation function for molecular liquids: The case of liquid water. *J. Chem. Phys.* **1994,** 101, 6888-6888.
2. Mason, P. E.; Brady, J. W., "Tetrahedrality" and the Relationship between Collective Structure and Radial Distribution Functions in Liquid Water. *J. Phys. Chem. B* **2007,** 111, (20), 5669-5679.
3. Svishchev, I. M.; Kusalik, P. G., Structure in liquid water: A study of spatial distribution functions. *J. Chem. Phys.* **1993,** 99, 3049-3049.
4. Miyazawa, S.; Jernigan, R. L., How effective for fold recognition is a potential of mean force that includes relative orientations between contacting residues in proteins? *J. Chem. Phys.* **2005,** 122, 024901-024901.
5. Buchete, N. V.; Straub, J. E.; Thirumalai, D., Orientational potentials extracted from protein structures improve native fold recognition. *Polymer* **2004,** 13, (4), 862-862.
6. Liwo, A.; Oldziej, S.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Scheraga, H. A., A united-residue force field for off-lattice protein-structure simulations. I. Functional forms and parameters of long-range side-chain interaction potentials from protein crystal data. *J. Comput. Chem.* **1997,** 18, (7).
7. Makowski, M.; Liwo, A.; Scheraga, H. A., Simple physics-based analytical formulas for the potentials of mean force for the interaction of amino acid side chains in water. 1. Approximate expression for the free energy of hydrophobic association based on a Gaussian-overlap model. *J. Phys. Chem. B* **2007,** 111, (11), 2910-2916.
8. Mukherjee, A.; Bhimalapuram, P.; Bagchi, B., Orientation-dependent potential of mean force for protein folding. *J. Chem. Phys.* **2005,** 123, 014901-014901.
9. Skolnick, J., In quest of an empirical potential for protein structure prediction. *J. Curr. Opin. Struct. Biol.* **2006,** 16, (2), 166-171.
10. Paramonov, L.; Yaliraki, S. N., The directional contact distance of two ellipsoids: Coarse-grained potentials for anisotropic interactions. *J. Chem. Phys.* **2005,** 123, 194111-194111.
11. Aswani, R.; Li, J. C., A new approach to pairwise potentials for water–water interactions. *J. Mol. Liq.* **2007,** 134, (1-3), 120-128.
12. Chowdhuri, S.; Tan, M. L.; Ichiye, T., Dynamical properties of the soft sticky dipole-quadrupole-octupole water model: a molecular dynamics study. *J. Chem. Phys.* **2006,** 125, 144513-144513.
13. Mezei, M.; Ben Naim, A., Calculation of the solvent contribution to the potential of mean force between water molecules in fixed relative orientation in liquid water. *J. Chem. Phys.* **1990,** 92, 1359-1359.
14. Naidoo, K. J.; Lopis, A. S.; Westra, A. N.; Robinson, D. J.; Koch, K. R., Contact Ion Pair between Na+ and PtCl62-Favored in Methanol. *J. Am. Chem. Soc* **2003,** 125, (44), 13330-13331.
15. Gay, J. G.; Berne, B. J., Modification of the overlap potential to mimic a linear site–site potential. *J. Chem. Phys.* **1981,** 74, 3316-3316.
16. Vorobjev, Y. N., Block-units method for conformational calculations of large nucleic acid chains. I. Block-units approximation of atomic structure and conformational energy of polynucleotides. *Biopolymers* **1990,** 29.

17. Vorobjev, Y. U. N., Block-units method for conformational calculations of large nucleic acid chains. II. The two-hierarchical approach and its application to conformational arrangement of the unusual T C loop of rabbit tRNAval. *Biopolymers* **1990,** 29, (12-13), 1519-1529.
18. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N., PE Bourne The Protein Data Bank. *Nucl. Acids Res.* **2000,** 28, 235-242.
19. Tanaka, S.; Scheraga, H. A., Medium-and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. *Macromolecules* **1976,** 9, (6), 945-950.
20. Makowski, M.; Chmurzy ski, L. In *Potentials of Mean Force of Two Hydrophobic Amino Acid Side Chain Models Dependent on Orientation*, AIP Conf. Proc., 2007; 2007; pp 1282-1282.
21. Buchete, N. V.; Straub, J. E.; Thirumalai, D., Anisotropic coarse-grained statistical potentials improve the ability to identify nativelike protein structures. *J. Chem. Phys.* **2003,** 118, 7658-7658.
22. Ben-Naim, A., Statistical potentials extracted from protein structures: Are these meaningful potentials? *J. Chem. Phys.* **1997,** 107, 3698-3698.
23. 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,** 79, 926-926.
24. Mahoney, M. W.; Jorgensen, W. L., A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions. *J. Chem. Phys.* **2000,** 112, 8910-8910.
25. Kuttel, M. M.; Naidoo, K. J., Free Energy Surfaces for the [alpha](1 4)-Glycosidic Linkage: Implications for Polysaccharide Solution Structure and Dynamics. *J. Phys. Chem. B* **2005,** 109, (15), 7468-7474.
26. Naidoo, K. J.; Brady, J. W., Calculation of the Ramachandran potential of mean force for a disaccharide in aqueous solution. *J. Am. Chem. Soc* **1999,** 121, (10), 2244-2252.
27. Rajamani, R.; Naidoo, K. J.; Gao, J., Implementation of an adaptive umbrella sampling method for the calculation of multidimensional potential of mean force of chemical reactions in solution. *J. Comput. Chem.* **2003,** 24, (14).
28. Mezei, M., Adaptive umbrella sampling: self-consistent determination of the non-Boltzmann bias. *J. Comp. Phys* **1987,** 68, 237-237.
29. Darve, E.; Pohorille, A., Calculating free energies using average force. *J. Chem. Phys.* **2001,** 115, 9169-9169.
30. Darve, E.; Rodríguez-Gómez, D.; Pohorille, A., Adaptive biasing force method for scalar and vector free energy calculations. *J. Chem. Phys.* **2008,** 128, 144120-144120.
31. Laio, A.; Parrinello, M., Escaping free-energy minima. *Proc. Natl. Acad. Sci.* **2002,** 99, (20), 12562-12566.
32. Belch, A. C.; Berkowitz, M.; McCammon, J. A., Solvation structure of a sodium chloride ion pair in water. *J. Am. Chem. Soc* **1986,** 108, (8), 1755-1761.
33. Berkowitz, M.; Karim, O. A.; McCammon, J. A.; Rossky, P. J., Sodium chloride ion pair interaction in water: computer simulation. *Chem. Phys. Lett.* **1984,** 105, (6), 577-580.
34. Khavrutskii, I. V.; Dzubiella, J.; McCammon, J. A., Computing accurate potentials of mean force in electrolyte solutions with the generalized gradient-

53. Kumagai, T.; Kaizu, M.; Hatta, S.; Okuyama, H.; Aruga, T.; Hamada, I.; Morikawa, Y., Direct Observation of Hydrogen-Bond Exchange within a Single Water Dimer. *Phys. Rev. Lett.* **2008,** 100, (16), 166101-166101.
54. Herndon, W. C.; Radhakrishnan, T. P., An evaluation of water cluster geometries derived from semi-empirical AM 1 calculations. *Chem. Phys. Lett.* **1988,** 148, (6), 492-496.
55. Millot, C.; Soetens, J. C.; Costa, M.; Hodges, M. P.; Stone, A. J., Revised anisotropic site potentials for the water dimer and calculated properties. *J. Phys. Chem. A* **1998,** 102, (4), 754-770.
56. Mahoney, M. W.; Jorgensen, W. L., Diffusion constant of the TIP5P model of liquid water. *J. Chem. Phys.* **2001,** 114, 363-363.

5 Benzene Dimer

In chapter 4 we demonstrated the sensitivity of the orientational PMF in distinguishing very subtle differences in water models. Here we use the free energy surface $W(r,\theta_1,\theta_2,\phi)$ to investigate molecular association on a very simple but important test case that is the benzene dimer in an aqueous solvent. The benzene dimer is a model system to study the very important π-π interactions that are present in DNA and many proteins.

5.1 Introduction

The strength of molecular association in solution is at the heart of many important macroscopic observations such as cellular communication and transport[1], protein folding[2-4], the mechanism of crystallization[5], and liquid crystal phases[6] to name but a few. In a solution the competition of the solute-solute, solute-solvent and solvent-solvent associations determine properties such as solubility. In the case of a polymer molecule in dilute solution the chain expansion depends on the balance between intramolecular segment-segment (monomer-monomer) and intermolecular segment-solvent and solvent-solvent interactions. If the polymer is dissolved in a *good solvent* it will expand to increase the number of segment-solvent contacts. In a poor solvent, however, the segment-solvent interactions are weak and their free energy of interaction may be positive and so unfavourable. In this case the polymer chain contracts so as to reduce the number of segment-solvent contacts and so become compact. The chain is therefore subjected to two opposing influences being i) expansion due to unfavourable segment-segment interactions and ii) contraction due to unfavourable segment-solvent interactions.[7]

Protein folding is the process by which polypeptides obtain their secondary and tertiary structures. The amino acids in the polypeptide are chemically connected along the c-α backbone to form a long polymer chain. The individual amino acids interact through space via intermolecular interactions due to the length of the chain. As can be

seen from free energy and solubility data given in Table 5-1, there is an affinity difference between polar and non-polar amino acids for solvents of different polarities.[8] This difference is further shown in protein structure where non-polar residues tend to the interior of the protein and polar residues to the exterior.[9] The stability of the protein structure is clearly affected by the solute-solute and solute-solvent association, making this competition decisive in the steering of the folding process.

Table 5-1: Classification of amino acid properties

Residue	Type	Zwitterion solubility, 25°C (mol/kg)	Side chain transfer ΔG_t, EtOH $\rightarrow H_2O$ (kcal/mol)
Trp	Nonpolar	0.07	3.00
Ile	Nonpolar	0.26	2.95
Pro	Polar	14.1	2.60
Ser	Polar	4.02	No data
Thr	Polar	No data	0.45
Arg	Charged	4.06	0.75

Table adapted from ref. 8

Aromatic interactions particularly benzene dimer interactions are often used as a model for amino acids such as phenylanlanine, DNA base stacking and in template directed and asymmetric synthesis.[10] Furthermore a principle component of liquid crystalline materials is aromatic in nature.[6] What is striking in all of these important phenomena is the strong orientational dependence of the aromatic interactions.

behaviours of a system. We chose to use TIP3P and TIP4P water environments as well as vacuum to match previous studies. The choice of using both water models was made to elucidate that TIP3P water, although having been used to parameterise the force field, can often lead to incorrect behaviour and TIP4P water can give more accurate and experimentally comparable results.

5.3 The FEARCF Method

As in the previous chapter the Free Energies from Adaptive Reaction Coordinate Forces (FEARCF) method was used to calculate the free energy surfaces. The same 4D parameterisation was used with the vectors describing the relative molecular orientations chosen perpendicular to the planes of the benzene molecules and is illustrated in Figure 5-2.

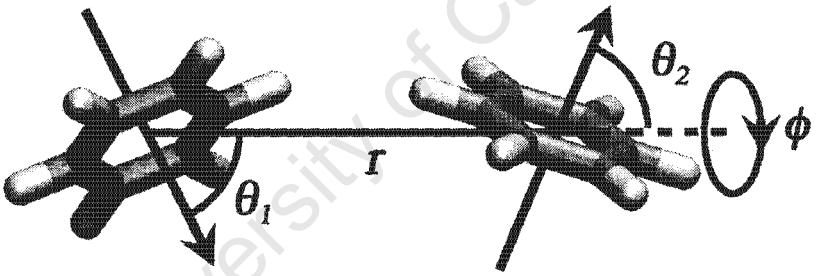

Figure 5-2: The 4D parameterisation of the free energy surface $W(r,\theta_1,\theta_2,\phi)$ for the benzene dimer.

The FEARCF method calculates free energy surfaces in n-dimensions based on adaptive umbrella sampling[17, 18] and is described in detail in Chapter 3. A numerical biasing potential is applied to the simulation using multidimensional cubic-spline interpolation.

$$H(\xi) = H_0 + U(\xi) \tag{5.1}$$

The probability distribution $P(\xi)$ of sampled reaction coordinates is then calculated and the resulting free energy surface is calculated using:

$$W(\xi) = -k_B T \log\left[P(\xi) \exp\left(\frac{U(\xi)}{k_B T} \right) \right] \tag{5.2}$$

For the next iteration of the method the biasing potential is set to the negative of the free energy $U(\xi) = -W(\xi)$. This iterative procedure is followed until there is no more significant change in the free energy surface and the sampling from the previous simulation is near uniform.

The forces arising from the cubic spline interpolation are then forces on the chosen reaction coordinates, which cannot be directly applied to the atoms of the molecules within the simulation. These forces must thus be converted into Cartesian forces and applied correctly so as not to deform the molecules. For a distance reaction coordinate this is done using

$$f_k(\xi) = -\frac{\partial U(\xi)}{\partial \xi} \frac{m_k}{\sum_{j=1}^{N_i} m_j}, \tag{5.3}$$

where m_k is the mass of and f_k is the force on atom k and N_i is the number of atoms in molecule i = 1,2. For an angular reaction coordinate this is done using

$$f_k(\xi) = -\frac{\partial U(\xi)}{\partial \xi} \frac{m_k r_k}{\sum_{j=1}^{N_i} m_j r_j^2}, \tag{5.4}$$

where r_k is the perpendicular distance from the axis of rotation to the position of atom k. For the 4D parameterisation given above and illustrated in Figure 5-2, the total force on an atom k is then:

$$F_k = -\frac{\partial U}{\partial r} \frac{m_k}{\sum_{i=1}^{N} m_i} - \frac{\partial U}{\partial \theta_1} \frac{m_k r_k}{\sum_{i=1}^{N} m_i r_i^2} - \frac{\partial U}{\partial \theta_2} \frac{m_k r_k}{\sum_{i=1}^{N} m_i r_i^2} - \frac{\partial U}{\partial \phi} \frac{m_k r_k}{\sum_{i=1}^{N} m_i r_i^2}. \tag{5.5}$$

We further used the Weighted Histogram Analysis Method[19-21] to combine all the histograms of previous simulations to increase the convergence of the free energy calculations.

5.4 Results

In this section the results from the benzene dimer calculations are presented. The simulations come from all-atom simulations of two benzene rings in either TIP3P or TIP4P water in a cubic box with side length 34.0 Å. The simulations were run in the NVT ensemble at a temperature of 298.15 K using a Nosé-Hoover thermostat and CHARMM's `vv2` integrator. Simulations were also run to calculate the 1D pmf at 278.15 K and 318.15 K to calculate the entropy contribution to the potential of mean force using $-\Delta S = \left[\Delta G(T_1) - \Delta G(T_2)\right]/(T_1 - T_2)$.

5.4.1 1D PMF

The one dimensional PMF's were constructed from 99 simulations each of 0.5 ns in TIP3P water and 23 simulations of 0.5 ns for TIP4P water. The PMF's were calculated as a function of the benzene – benzene centre of mass distance.

In Figure 5-3 the 1D free energy curves for TIP3P and TIP4P clearly show the difference between the water models. Both have minima in similar locations: the contact association at ~ 5.7 Å and the solvent separated association at ~ 8.6 Å. The prominent difference in the PMF's is the well depths of the minima. For TIP3P, the benzene dimer prefers the solvent separated position while for TIP4P the benzene dimer has its global minumum at the contact position.

This contact minima is located close to the same separation distance of 5.5 Å presented in previous results with similarly shallow well depths[13, 14]. This distance seems indicative of a T-shaped configuration as the minimum energy configuration, since a stacked configuration would allow a closer approach of the molecules.

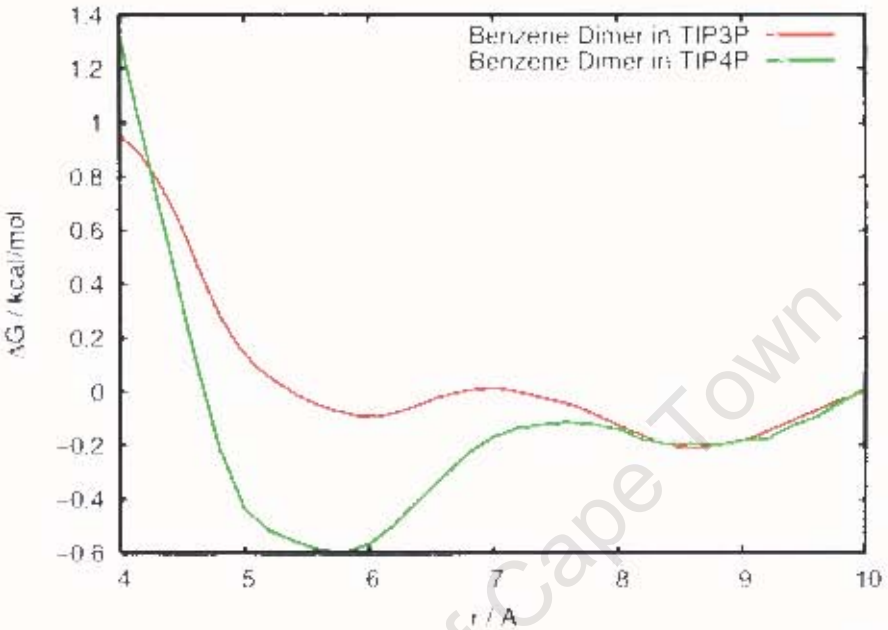

Figure 5-3: The distance potential of mean force for two benzene dimers in TIP3P (red) and TIP4P (green) water.

The next part of the investigation was to determine whether the largest contribution to the minima of the PMF was enthalpic or entropic. Since the stacked configuration is the entropically favoured position, the expectation is that the enthalpic contribution should be the greatest for the minimum position that we see in the PMF.

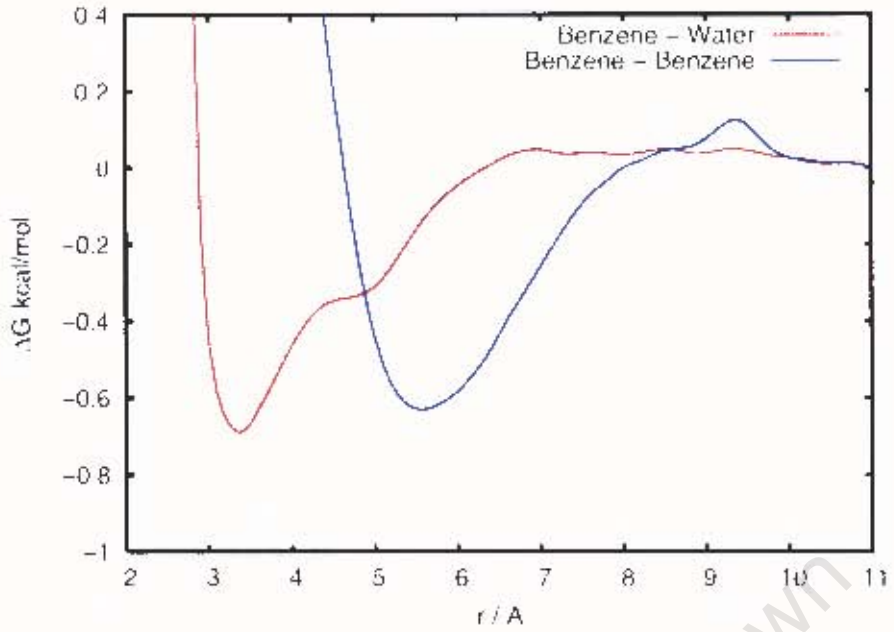

Figure 5-5: Vacuum free energies between benzene – benzene (blue) and benzene – water (red)

A further consideration is the water – water association, which is also in competition with benzene – water. The association strength between two TIP4P waters was calculated in Chapter 4 to be -3.6 kcal/mol. This is clearly much greater than either of the other interactions and is likely the cause for the benzene – benzene association in aqueous solution.

5.4.2 4D PMF

The full four dimensional free energy surfaces were calculated for the benzene dimer in TIP4P water and in vacuum. In Figure 5-7 the most illuminating slices are shown. From this figure we can see that the minimum energy occurs at the following angles: $\theta_1, \theta_2 = 60, 120$ and $\phi = 0, 180$. The subfigures are calculated by taking a Boltzmann average over the angle coordinate not shown in each case and over the distance coordinate between 5Å and 6 Å. Figure 5-7a and Figure 5-7c were extracted using the following equation:

$$W(\xi_2,\xi_3) = -k_B T \log\left[\sum_{\xi_1=5}^{6}\sum_{\xi_4=0}^{360} P(\xi_1,\xi_2,\xi_3,\xi_4)\right]. \quad (5.6)$$

Figure 5-7b and Figure 5-7d were extracted from the 4D surface using:

$$W(\xi_2,\xi_4) = -k_B T \log\left[\sum_{\xi_1=5}^{6}\sum_{\xi_3=0}^{180} P(\xi_1,\xi_2,\xi_3,\xi_4)\right]. \quad (5.7)$$

These figures however do not clearly show the minimum energy configuration since they are averaged slices of the full 4D surface. The absolute minimum, for both water and vacuum, occurs at four distinct but equivalent orientations:
$(\theta_1,\theta_2,\phi) = (60,60,0),(120,120,0),(120,60,180),(60,120,180)$.

The minimum orientation is depicted in Figure 5-6. Obtaining the same minima for both water and vacuum cases is unsurprising since the minimum position in the 1D PMF is shown to be enthalpically driven (Figure 5-4).

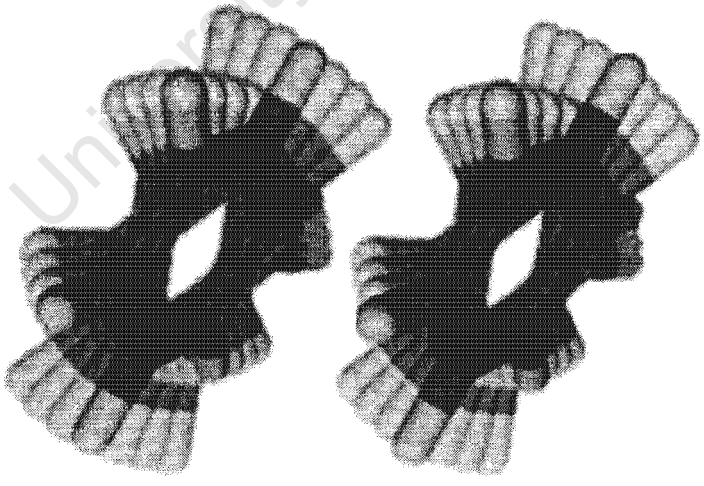

Figure 5-6: "Shifted stacked" minimum energy configuration for the benzene dimer in vacuum and solution. The faded benzene rings in the background represent the movement freedom due to the shallow minima and the circular arrows represent the direction in which the rotational symmetry is assumed.

The minimum free energy orientation corresponds well to the minimum energy orientation computed from high level quantum mechanics calculations by Podeszwa *et al.*[15]. It is, as they showed, neither the directly stacked nor T orientation that were typically thought to be the minimum energy but rather the shifted-stacked configuration.

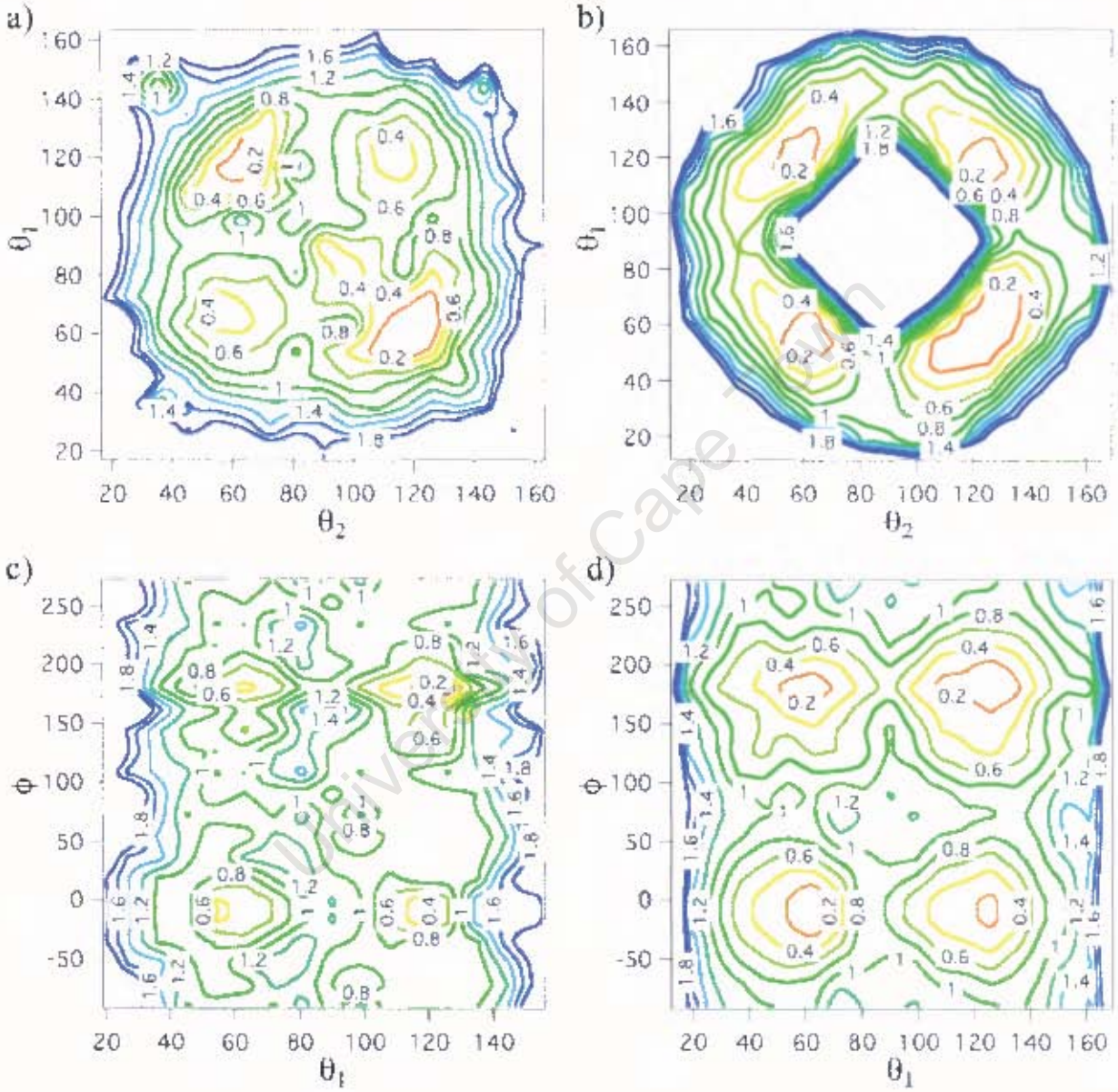

Figure 5-7: Selected slices of the 4D PMF for the benzene dimer in TIP4P water (a & c) and in vacuum (b & d). The slices are taken by taking a Boltzmann average between 5 and 6 Å and over the angular coordinate not shown in each case. The contour energies are in kcal/mol with the minima being set to 0 kcal/mol.

5.5 Conclusion

We calculated the free energy surfaces for the benzene – benzene interaction in aqueous solution and in vacuum. In solution, there was a clear difference in the association of benzene in TIP3P and TIP4P. It is interesting to note that although the force field was parameterised to the TIP3P model, it did not give the correct contact association, while TIP4P did.

The interaction energies in vacuum compared well with previous results. Further calculations of relative entropic and enthalpic constributions to the free energy revealed that the minimum positions were enthalpic and not entropic. This is contrary to what was expected since the hydrophobic association is thought to be the primary driving force of non-polar molecules. The results could be an indication that either the molecular volume decrease linked with hydrophobic association is too small for the benzene dimer for it to be a driving force. This is something that could be investigated in future studies.

The calculated anisotropic free energy surface from the molecular dynamics simulation was able to clearly show the minimum free energy configuration for the benzene dimer. This result compared well to previous high level quantum calculations and reiterated (along with Chapter 4) the sensitivity of the method. Further studies using the FEARCF method could be done with benzene and more complex aromatic molecules, possibly being treated quantum mechanically through semi-empirical methods. These results could be used to provide greater insight into the interactions of aromatic molecules in liquid crystals or biomolecules such as proteins.

18. Bartels, C.; Karplus, M., Probability Distributions for Complex Systems: Adaptive Umbrella Sampling of the Potential Energy. *J. Phys. Chem. B* **1998,** 102, (5), 865-880.
19. Bouzida, D.; Kumar, S.; Swendsen, R. H., Efficient Monte Carlo methods for the computer simulation of biological molecules. *Phys. Rev. A* **1992,** 45, (12), 8894.
20. Kumar, S.; Payne, P. W.; Vasquez, M., Method for free-energy calculations using iterative techniques. *J. Comput. Chem.* **1996,** 17, (10), 1269-1275.
21. Kumar, S.; Rosenberg, J. M. D.; Bouzida, J.; Swendsen, R. H.; Kollman, P., The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *J. Comput. Chem.* **1992,** 13, (8), 1011-1021.

6 Conclusion

This work presented the free energy from adaptive reaction coordinate forces method that extends previous 2-D adaptive umbrella sampling methods to higher dimensions and employs a previously suggested parameterisation for intermolecular association. It was successfully shown to have the sensitivity to distinguish between the TIP3P, TIP4P and TIP5P water models and elucidate which of their association properties may lead to their different bulk behaviours. It was also shown to have the required detail to study the solvent effect on benzene-benzene association. Here it showed that the shifted-stacked configuration is the solvent induced minimum and that the typical choice of water model, TIP3P for which it was parameterized, cannot reproduce this minimum.

The sensitivity and detail of the method lends itself to the study of more complicated systems. The free energy surfaces of more complex molecules, particularly those with higher anisotropy should be investigated to determine the best choice of orientation vectors to use. This information can then lead to a systematic study of amino-acids, carbohydrates, nucleic acids and similar molecules that can lead to insights into systems important in biology and technology.