

The copyright of this thesis vests in the author. No quotation from it or information derived from it is to be published without full acknowledgement of the source. The thesis is to be used for private study or non-commercial research purposes only.

Published by the University of Cape Town (UCT) in terms of the non-exclusive license granted to UCT by the author.

PARAMETERIZATION OF THE GAY-BERNE COARSE-GRAIN POTENTIAL FROM ATOMISTICALLY DETAILED ANISOTROPIC FREE ENERGY VOLUMES

Thesis presented to the
UNIVERSITY OF CAPE TOWN
in fulfilment of the requirements for the degree of
DOCTOR OF PHILOSOPHY

By
M. RIEDAA GAMIELDIEN

Supervisor: Professor Kevin J. Naidoo



Scientific Computing Research Unit
and Department of Chemistry
University of Cape Town
March 2012

University of Cape Town

PAGE LEFT BLANK INTENTIONALLY

ABSTRACT

Simulating a system of 300,000+ atoms, such as an explicitly solvated protein using all atom molecular dynamics on the microsecond time-scale, would require an enormous amount of computing power and specialized software, even with which would still require months of computing time. However, if the atomic degrees of freedom of the system can be reduced (or averaged) in some physically intuitive manner, while still retaining a connection with the underlying atomistic detail, microsecond simulations could be achieved within weeks or days. Coarse-graining, a sub-class of mesoscale modelling, is used to represent molecules in a reduced form as either regular spheroids (ellipsoids) or as continuum models, using specialized interaction potentials. The Gay-Berne is a one such coarse-grain potential, which has been particularly successful in that it has been used in modelling of liquid crystals, protein dynamics and lipid membrane and micelle formation. However, the challenge is obtaining a free energy dataset, from atomistically detailed all atom molecular dynamics simulations, against which the coarse-grain potential can be parameterized. In this study we present an unprecedented free energy hyper-volume for the benzene dimer association, which is in turn used to parameterize a coarse-grain benzene model.

The Free Energy from Adaptive Reaction Coordinate Forces (*FEARCF*) method is used to obtain the 4D anisotropic free energy hyper-volume for the benzene dimer association in vacuum and solution. The results in vacuum are found to agree with the structure and relative energy differences found in the literature, while the effect of the solvent is clearly evident from these hyper-volumes. This study is also extended to include biologically relevant aromatic amino acid pairs associating, namely phenylalanine and tyrosine, and compared to free energy hyper-volumes computed from ± 6000 protein structures. The latter shows that hyper-volumes obtained from the simulations are acceptable for describing aromatic association and, by extension, suggesting the role it plays in protein stability and folding.

By parameterizing against the benzene dimer vacuum free energy hyper-volume, a Gay-Berne interaction potential parameter set was obtained for a coarse-grain model of benzene. This parameter set was used to simulate liquid benzene in a coarse-grain rigid body dynamics simulation program governed by the Gay-Berne potential, developed during this thesis. An analysis of the results reveals the proof-of-concept: parameterization of the Gay-Berne potential against the benzene dimer free energy hyper-volume results in a coarse-grain dynamics simulation of liquid benzene showing good correlation (or connection) with the all atom model and experiment.

PREFACE

The mention and discussion of the Gay-Berne intermolecular interaction potential (Gay, J. G. and Berne, B. J. <http://dx.doi.org/10.1063/1.441483>) throughout this thesis brings to attention the central importance that this potential has in this work. Another important and central theme is the multi-dimensional anisotropic free energy hyper-volumes used to describe intermolecular interactions. This thesis project was presented, formulated, managed and executed in two parts, and the same rationale has been used in the design and layout of this thesis. The division also helps establish the boundary between the all atom and coarse-grain (CG) simulation techniques, separation of atomic isotropic and molecular anisotropic intermolecular interactions, and separate the application and method development components of this thesis. The two parts can be conceptualized as: (i) the application of the *FEARCF* method (Barnett, C. B. and Naidoo, K. J. <http://dx.doi.org/10.1080/00268970902852608> and Strumpfer, J. and Naidoo, K. J. <http://dx.doi.org/10.1002/jcc.21317>) for investigating aromatic association using free energy hyper-volumes; and (ii) the development of a rigid body CG dynamics program implementing the Gay-Berne CG interaction potential, parameterization of a CG potential, performance gain by using a CG model and comparison of the simulation results with all atom molecular dynamics (MD) simulations and experiment.

The Gay-Berne provides a means of describing molecular, rather than atomic, non-bonded interactions, thereby allowing for an all atom system to be described in a reduced representation. It is not only dependent on separation distance between the molecular, but their relative orientation as well. The Gay-Berne, $U_{GB}(r, \theta_i, \theta_j, \phi)$, is a linear site-site potential described by 4 molecular degrees of freedom: r the separation between the centre of masses; θ_i and θ_j the relative orientation of the i^{th} and j^{th} molecule with respect to the interconnecting **vector** (r); and ϕ the dihedral angle between molecule i and j . The form of the Gay-Berne potential allows for a group of atoms, or an entire molecule, to be conflated into a single, CG site. This reduction in the atomic degrees of freedom intrinsically reduces the number of non-bonded interactions and, concomitantly, reducing the computational cost of such a simulation significantly

The Gay-Berne potential, based on a Corner potential, was developed in the 1980s to overcome the computational expense incurred by calculating the atomic pairwise non-bonded interactions, by rather describing molecular pairwise non-bonded interactions between molecules. By doing so, the potential also had to include the effects the relative molecular orientation has on the interaction. A CG potential, such as the Gay-Berne, preferentially requires a free energy dataset to be parameterized against,

because this would ensure the all atom information is embedded within the CG model. The existence of such an anisotropic potential, the Gay-Berne potential, proved to be crucial for two reasons: (i) it already existed and is well represented in the literature; and (ii) for defining the free energy hyper-volume, $W(r, \theta_1, \theta_2, \phi)$, to which it would be parameterized against.

Chapter 1 introduces the Gay-Berne potential in a brief manner, showing that it is a generalized anisotropic form of the isotropic Lennard-Jones potential. Chapter 2 provides the theory for all atom MD simulations and statistical analysis, while the method used for determining the free energy hyper-volumes (*FEARCF*), as well as a thorough explanation of statistical mechanics ensembles and general free energy methodology, is provided in Chapter 3. The first part of the thesis provides the results for the application of the *FEARCF* method for describing the association free energy hyper-volumes for the benzene dimer (Chapter 4) and aromatic amino acid pairs (Chapter 5). The second part of this thesis begins with the introduction of rigid body dynamics, implemented with the Gay-Berne potential (Chapter 6); a description of the anisotropic Gay-Berne potential and the parameterization strategy employed (Chapter 7); and, finally, the last chapter presents the parameter set for the CG benzene model, obtained via parameterization from the free energy hyper-volume. The performance gain of the CG over the all atom model is reported and a comparison of the structure and dynamics results with all atom MD simulations and experiment is made.

Every chapter is designed to build up to the results obtained for the newly parameterized Gay-Berne CG benzene model presented in the final chapter. More importantly, does parameterizing a CG potential from a free energy hyper-volume produce a CG model that contains acceptable atomic information compared to other methods currently being used? Gregory A. Voth eloquently sums up the key aspect that brings this thesis together: "... coarse-graining should be cast within the context of statistical mechanics, and more specifically in terms of the Helmholtz free energy..." (Voth, G. A. *Coarse-Graining of Condensed Phase and Biomolecular Systems*, CRC Press: New York, 2009; p 1.)

$$\exp\left(-\frac{A}{k_B T}\right) = \text{const.} \int d\mathbf{x}_{AA} \exp\left[\frac{-V_{AA}(\mathbf{x}_{AA})}{k_B T}\right] \\ \approx \text{const.} \int d\mathbf{x}_{CG} \exp\left[\frac{-V_{CG}(\mathbf{x}_{CG})}{k_B T}\right]$$

where A , the atomistic Helmholtz free energy, obtained using an atomistically detailed potential energy function, $V(\mathbf{x}_{AA})$, is used to parameterize a CG potential, $V_{CG}(\mathbf{x}_{CG})$.

DECLARATION

I declare that the work in this thesis, "**Parameterization of the Gay-Berne Coarse-Grain Potential From Atomistically Detailed Anisotropic Free Energy Volumes**", is based on research carried out at the Scientific Computing Research Unit in the Department of Chemistry, University of Cape Town, South Africa. No part of this thesis has been submitted elsewhere for any other degree or qualification and is all my own work unless referenced to the contrary in the text.

M. Riedaa Gamieldien

March 2012

PUBLICATIONS AND CONFERENCES

Parts of this thesis has been presented at the following conferences:

- M. Riedaa Gamieldien and Kevin J. Naidoo. *Coarse-Grained Molecular Dynamics from Free Energy Adaptive Reaction Coordinate Forces (FEARCF)*. Centre for High Performance Computing National Conference, **2010**, 07-09 December 2010, The Westin Grand Cape Town Arabella Quays, South Africa.
- M. Riedaa Gamieldien and Kevin J. Naidoo. *Distance and Orientational Potential of Mean Force Elucidates Aromatic Amino Acid Association – With Implications to Coarse-Grained Molecular Dynamics*. The 2011 World Congress of WATOC, **2011**, 17 – 22 July 2011, Santiago de Compostela, Spain.

and Chapter 4 has been published as a journal article in the Journal of Physical Chemistry B:

- M. R. Gamieldien, J. Strümpfer, K. J. Naidoo. Hydration-Determined Orientational Preferences in Aromatic Association from Benzene Dimer Free Energy Volumes. *J. Phys. Chem. B* **2012**, 116 (1), 324-331.

ACKNOWLEDGEMENTS

I am grateful for the financial assistance received from the National Research Foundation (NRF), the South African Research Chair Initiative (SARChI) and the University of Cape Town for the pursuit of my PhD and for the conference travel to Spain for the Ninth Triennial WATOC conference held in Santiago de Compostela, Spain.

I would like to thank Christopher Barnett, Kyle Fernandes, Krishna Govender and Jestin Mandumpal for spending their precious time reading and providing invaluable corrections/additions for parts of this thesis. I am especially indebted to Krishna Govender for assuming my role as system administrator during the duration of writing up my thesis. Thanks to members of the Scientific Computing Research Unit (SCRU) for their support and encouragement during the last three years. Special thanks must be given to my other friends in the Chemistry department and outside for their encouragement and understanding during my plight of writing up my PhD thesis.

To my supervisor, Prof. Kevin J. Naidoo, many thanks for the guidance and patience you have shown during the course of my PhD thesis. Invaluable discussions and special attention to detail have shown me how to produce top-notch results. Thank for your understanding during the hard times I have had to endure during the course of my PhD. It is my belief that no other supervisor could have been more compassionate and supportive and I truly appreciate this.

Finally, to my familia, thank you very much for your love, support and constant nagging that never allowed me to rest my dairy air, butt always pushed me to provide an answer for “when are done, and when are you getting a job?” Well, now I am one step closer to the answer. And to the members of my family who could not be here (my father and grandmother), this thesis is dedicated to you.

ABBREVIATIONS

It should be noted here that the Helmholtz free energy, ΔA , is used exclusively in this thesis and will be hereto referred to as free energy.

Å	Angstrom (10^{-10} m)
CG	Coarse-Grain
CHARMM	Chemistry at Harvard Macromolecular Mechanics
FEARCF	Free Energy from Adaptive Reaction Coordinate Forces
GB	Gay-Berne
GR	Golubkov and Ren
LJ	Lennard-Jones
MD	Molecular Dynamics
MM	Molecular Mechanics
NVT	Canonical Ensemble
PDF	Pair Distribution Function
PME	Particle Mesh Ewald
PMF	Potential of Mean Force
QM	Quantum Mechanics
RDF	Radial Distribution Function
SDF	Spatial Distribution Function
WHAM	Weighted Histogram Analysis Method
ΔA	Helmholtz Free Energy
ΔH	Enthalpy
ΔE	Internal Energy
ΔS	Entropy
ps	Picosecond (10^{-12} s)
ns	Nanosecond (10^{-9} s)
μs	Microsecond (10^{-6} s)

NOTE:

Abbreviations not appearing here are provided in the text when they are used.

TABLE OF CONTENTS

ABSTRACT	III
PREFACE	IV
DECLARATION	III
PUBLICATIONS AND CONFERENCES	VII
ACKNOWLEDGEMENTS	VIII
ABBREVIATIONS	IX
TABLE OF CONTENTS	IX

CHAPTER 1: INTRODUCTION	1
1.1 INTERMOLECULAR FORCES	6
1.1.1 SHORT-RANGE INTERACTIONS	8
1.1.2 LONG-RANGE INTERACTIONS	8
1.2 THERMODYNAMICS OF INTERMOLECULAR FORCES	11
1.3 HYDROGEN BONDING, HYDROPHOBIC AND HYDROPHILIC INTERACTIONS	13
1.4 MOLECULAR ASSOCIATION	15
1.5 MODEL INTERMOLECULAR POTENTIALS	16
1.6 HOW TO REDUCE DEGREES OF FREEDOM?	19
1.7 COARSE-GRAIN DYNAMICS	21
1.8 OBJECTIVES AND OVERVIEW OF THESIS	23
1.9 REFERENCES	25

CHAPTER 2: MOLECULAR SIMULATIONS AND STATISTICAL ANALYSIS	33
2.1 MOLECULAR DYNAMICS	34
2.1.1 MOLECULAR MECHANICS – THE POTENTIAL ENERGY AND FORCES	34
2.1.2 LONG-RANGE INTERACTIONS	40
2.1.3 INTEGRATION SCHEMES I	41
2.1.4 BOUNDARY CONDITIONS AND MINIMUM IMAGE CONVENTION	43
2.1.5 POTENTIAL TRUNCATION	44
2.1.6 CONSTRAINT DYNAMICS	46
2.2 SAMPLING FROM ENSEMBLES	47
2.3 STATISTICAL ANALYSIS	49
2.3.1 PAIR DISTRIBUTION FUNCTIONS	49
2.3.2 SPATIAL DISTRIBUTION FUNCTION	50
2.3.3 CORRELATION FUNCTIONS	51
2.4 REFERENCES	53
 CHAPTER 3: STATISTICAL MECHANICS AND FREE ENERGY	 54
3.1 STATISTICAL MECHANICS: THE BEGINNINGS	54
3.2 STATISTICAL MECHANICS: THE FUNDAMENTALS	55
3.2.1 CONCEPTS	55
3.2.2 ENSEMBLE THEORY	56
3.3 THE FREE ENERGY METHOD	59
3.3.1 FREE ENERGY PERTURBATION METHODOLOGY	60
3.3.2 THE POTENTIAL OF MEAN FORCE METHODOLOGY	63
3.4 REFERENCES	72

PART 1: INFLUENCE OF ORIENTATION ON AROMATIC ASSOCIATION 75

CHAPTER 4: BENZENE DIMER – THE PROTOTYPE FOR AROMATIC ASSOCIATION – DYNAMICAL ORIENTATIONAL PREFERENCES 76

4.1 INTRODUCTION 76

4.2 METHODS 78

4.2.1 THE FEARCF METHOD APPLIED TO AROMATIC ASSOCIATION 80

4.2.2 REDUCING THE DIMENSIONALITY OF THE FREE ENERGY VOLUME 81

4.2.3 ENTROPY AND INTERNAL ENERGY CONTRIBUTIONS TO THE FREE ENERGY 81

4.3 RESULTS AND DISCUSSION 82

4.3.1 ANALYSIS OF THE FREE ENERGY VOLUMES IN VACUUM 82

4.3.2 ANALYSIS OF THE FREE ENERGY VOLUMES IN SOLUTION 85

4.3.3 THERMODYNAMIC ANALYSIS 88

4.4 CONCLUSION 90

4.5 REFERENCES 92

CHAPTER 5: AROMATIC AMINO ACID ASSOCIATION USING FREE ENERGY HYPER-VOLUMES 94

5.1 INTRODUCTION 94

5.2 METHODS 96

5.2.1 DEFINITION OF THE ASSOCIATION REACTION COORDINATES 96

5.2.2 PROTEIN DATABASE PREPARATION AND CALCULATIONS 96

5.3 RESULTS AND DISCUSSION 97

5.3.1 FREE ENERGY HYPER-VOLUME ANALYSIS IN VACUUM 97

5.3.2 FREE ENERGY HYPER-VOLUME ANALYSIS IN SOLUTION 103

5.3.3 AROMATIC AMINO ACID ASSOCIATION AS FOUND IN PROTEINS 109

5.3.4 THERMODYNAMIC ANALYSIS	114
5.4 CONCLUSIONS	117
5.5 APPENDIX	119
5.6 REFERENCES	123
 PART 2: REDUCTION OF REPRESENTATION – COARSE-GRAINED DYNAMICS	 126
 CHAPTER 6: COARSE-GRAIN DYNAMICS AND CODE DEVELOPMENT	 127
6.1 COARSE-GRAIN DYNAMICS	128
6.2 THE HAMILTONIAN	130
6.2.1 COARSE-GRAINED MECHANICS	130
6.3 ROTATIONAL AND RIGID BODY DYNAMICS	131
6.3.1 EULER EQUATIONS OF MOTION	131
6.3.2 ORIENTATION MATRIX	132
6.3.3 ROTATIONAL FRAMES OF REFERENCE	133
6.3.4 ANGULAR MOMENTUM AND VELOCITY	133
6.3.5 QUATERNIONS: A SUBSTITUTE FOR EULER ANGLES	135
6.4 ROTATIONAL KINEMATICS	136
6.4.1 INTEGRATION SCHEMES II	136
6.5 GAY-BERNE POTENTIAL AND FORCES	139
6.5.1 THE GAY-BERNE INTERACTION POTENTIAL	139
6.5.2 FORCES AND TORQUES	140
6.8 REFERENCES	141

CHAPTER 7: ANALYSIS OF THE GAY-BERNE INTERACTION POTENTIAL AND STRATEGY FOR PARAMETERIZATION	143
7.1 THE GAY-BERNE POTENTIAL: THE FULL STORY	144
7.1.1 GAY-BERNE POTENTIAL	144
7.1.2 VISUALIZING THE GAY-BERNE DATASETS	147
7.1.3 PARAMETRIC ANALYSIS	149
7.2 PARAMETERIZATION USING A GENETIC ALGORITHM (GA)	151
7.2.1 FITNESS FUNCTION	151
7.2.2 CROSSING BOUNDARIES	152
7.2.3 METHODS OF REPRESENTATION	153
7.2.4 METHODS OF SELECTION	154
7.2.5 METHODS OF CHANGE	155
7.2.6 STRENGTHS	156
7.2.7 LIMITATIONS	156
7.3 REFERENCES	158
 CHAPTER 8: COARSE-GRAIN DYNAMICS OF LIQUID BENZENE – A PROOF OF CONCEPT	 160
8.1 INTRODUCTION	160
8.2 METHODS	162
8.2.1 THE GAY-BERNE POTENTIAL, FORCES AND TORQUES	162
8.2.2 COARSE-GRAIN DYNAMICS SIMULATION DETAILS	163
8.2.3 CONVERSION FROM COARSE-GRAIN PARTICLES TO ALL ATOM BENZENE MOLECULES	163
8.2.4 ALL ATOM SIMULATION DETAILS	164
8.2.5 PARAMETERIZATION	165
8.2.6 CALCULATION OF THE DIFFUSION CONSTANT	166
8.3 RESULTS AND DISCUSSION	166

8.3.1 RADIAL PROBABILITY DISTRIBUTION FUNCTION	167
8.3.2 RADIAL-ANGULAR PROBABILITY DISTRIBUTION	169
8.3.3 THREE-DIMENSIONAL SOLVENT STRUCTURING	170
8.3.4 DIFFUSION CONSTANT	172
8.3.5 COMPARATIVE TIMINGS	173
8.4 CONCLUSIONS	174
8.5 REFERENCES	176

University of Cape Town

CHAPTER 1

INTRODUCTION

Chemistry and physics are two scientific disciplines that study the matter and forces that make up the universe, from the tiniest sub-atomic particles (e.g. electrons) to geckos climbing up trees. It is now well established that there are four fundamental forces in nature.¹ Two of these are the *strong* and *weak interactions* that act between neutrons, protons, electrons and other subatomic particles. They occur over very short distances, less than 10^{-5} nm. The other two forces are the *electromagnetic* and *gravitational interactions* that act between atoms and molecules, and operate over much larger distances from subatomic to practically infinity. For example, electromagnetic forces – the source of all intermolecular interactions – determine the properties of solid, liquids and gases, the behaviour of particles in solution and chemical reactions.¹

The Greeks were the first to view forces in a nonreligious way. They found that they only needed two fundamental forces to account for all natural phenomena: Love (attraction) and Hate (repulsion). Love brought things together, while Hate caused them to part. The idea was first proposed by Empedocles around 450 B.C, and was much “improved” by Aristotle, and formed the basis of chemical theory for 2000 years.¹ In the 4th century B.C., Aristotle observed the ability of a gecko to run up and down a tree in any way, even with its head pointing downwards.^{2, 3} Now, 2000 years later, we are uncovering the secrets of how geckos use millions of tiny foot-hairs to adhere to even smooth molecular surfaces (Figure 1.1a). The two competing hypotheses^{2, 4} for the gecko adhesion mechanism are (i) thin-film capillary forces (or other mechanisms reliant on hydrophilicity), and (ii) van der Waals forces. The mechanism for gecko adhesion has been attributed to van der Waals forces between the surfaces and tiny microscopic hairs (setae) found on their footpads.^{2, 5, 6} The thin-film capillary mechanism has been disputed and contradicted by a number of recent studies.⁷⁻⁹ In 2011, researchers found evidence that geckos leave footprints on the surfaces on which they walk, and these footprints were found to contain phospholipids with phosphocholine head groups.^{6, 10}

The presence of lipids has never been considered in current models of gecko adhesion. This finding further substantiates the van der Waals force adhesion mechanism. The analysis of gecko footprints and the toe pad–substrate interface has significant consequences for models of gecko adhesion and by extension, the design of synthetic mimics.⁶

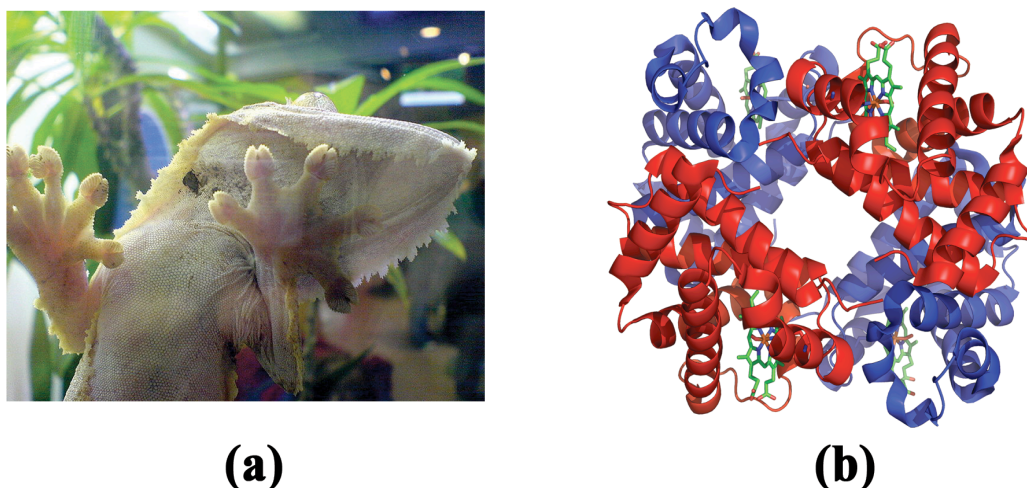


Figure 1.1: (a) Gecko adhering to a glass surface;¹¹ and (b) the three-dimensional structure of the oxygen transport protein, Hemoglobin.

There are other important intermolecular forces (or interactions) found in chemical and biochemical systems that include hydrogen bonding, electrostatic interactions (i.e., the multipole expansion, Section 1.1.2), solvent effects and the hydrophobic effect.^{10, 12} The importance of these intermolecular interactions in biology has been found to be the determining factor for the stability of a protein's three-dimensional (or native, Figure 1.1b) structure and protein folding.^{10, 12}

Protein folding, and understanding its mechanism, still remains a grand challenge problem in this day and age. In a recent article,¹³ a layman's term explanation of the protein-folding problem is given and the author explains the problem in terms of two important questions surrounding the field. First, what are the rules that govern the folding mechanism and kinetics? And secondly, can researchers predict the three-dimensional structure of unknown proteins?¹³ The answers to these two questions are the *Holy Grail* for theoretical chemists and biologists alike. To gain an understanding of such a complex problem, in retrospect, we do what we did when we were kids and learned, for ourselves, how an analog watch worked for first time. We opened the watch and observed how each piece operated in synchronicity to allow it to function. Therefore as with the watch example, the protein folding problem has to first be understood in terms of its constituent (smaller) parts, the knowledge of the working

parts collected and then we can begin to understand the problem of protein folding as a whole – with the concept in mind: the whole is equal to sum of the parts.¹⁴

It is undoubtedly clear that intermolecular forces (or interactions) play a critical part in not only science, but also in everyday life. John Rowlinson provides a history of the intermolecular forces in his book concerning the history of intermolecular forces.¹⁵ The work of Boyle and others lead to the equation of state for ideal gases, while the extension to include real gases showed the significance of intermolecular forces.

Implementing the famous work of Robert Boyle, who demonstrated that for an ideal gas $PV = \text{constant}$, and the later works of Charles, Dalton and Guy-Lusaac (who showed that $P \propto T$ at constant V), Clapeyron was able to derive the equation of state for ideal gases, while the extension to include real gases showed the significance of intermolecular forces.

Implementing the famous work of Robert Boyle, who demonstrated that for an ideal gas $PV = \text{constant}$, and the later works of Charles, Dalton and Guy-Lusaac (who showed that $P \propto T$ at constant V), Clapeyron was able to derive the equation of state for an ideal gas (i.e., the ideal gas law).¹

$$PV = nRT, \quad 1.1$$

where P is the pressure, V volume, n number of moles, R universal gas constant and T the temperature. However, this equation could not correctly predict the properties of real gases such as carbon dioxide and oxygen. In an attempt to show why real gases did not obey the ideal gas law, the Dutch physicist Johannes Diderik van der Waals considered the effects of attractive forces between molecules. In 1873, he arrived at his famous equation of state for liquids and gases:¹

$$\left(P + \frac{an^2}{V^2} \right) (V - nb) = nRT, \quad 1.2$$

in which he subtracted the term nb from the volume, to account for the finite size of molecules, and added the term an^2/V^2 to the external (measured) pressure, to account for attractive intermolecular forces now known as van der Waals forces.¹ James Clerk Maxwell remarked, "... that there can be no doubt the name of van der Waals will soon be amongst the foremost in molecular science".¹⁶

The work of van der Waals and many others have shown that intermolecular forces between atoms and molecules are very important for understanding their underlying

properties and dynamics. The strength of the intermolecular interactions leads to how strongly molecules associate, which is the critical difference between the phases of matter (solids, liquids and gases). Without intermolecular interactions we would not have molecular association, phases of matter, Archimedes observing geckos climbing up and down trees, proteins folding into functioning enzymes and life (and the universe). The aim is to study and understand these intermolecular interactions that are so crucial in the functioning of the universe.

There are numerous experimental and theoretical techniques for studying intermolecular interactions. In the last two decades computer simulation techniques have flourished and provided invaluable insight into intermolecular interactions. These simulation techniques make use of reduced intermolecular interaction potentials, which are usually pairwise in nature and computationally very expensive. The two most commonly used computational techniques are quantum mechanics (QM) and molecular dynamics (MD).^{17, 18} Quantum mechanics, due to its high computational expense, is used to study small molecular systems and molecular dynamics (implementing molecular mechanics)¹⁸ is a less computationally demanding classical technique, which can simulate large systems that explicitly includes the solvent.¹⁹ However, these techniques cannot reach the spatial (size) or temporal (time) scales required to study biologically relevant processes such as protein folding. Such systems have tens of thousands of atoms (by including the solvent explicitly) and require simulation times of microseconds or greater. The result of such a large system is that there are an astronomically large number of pairwise intermolecular interactions that need to be computed.²⁰ However, reducing the level of detail in the representation of the system (such as grouping atoms into a single interaction site), while retaining the underlying connections to atomistic models, will allow the study of larger systems on longer time scales in much less real time.²⁰ For a lipid bilayer simulation, Orsi et al. reported a speed up of 130x for the reduced system over the all atom MD simulation.²¹ This reduction in detail has been aptly referred to as coarse-graining, which is currently enjoying a revival and being developed by many researchers.²⁰ Coarse-grain (CG) dynamics has been implemented for studying a wide variety of systems from chemistry²² (e.g. benzene^{23, 24} and water²⁵) to biology (e.g. protein folding,²⁶⁻²⁸ nucleic acids²⁹ and lipids^{21, 30, 31}) to polymer science.^{32, 33}

CG dynamics has statistical mechanics at its foundation, and statistical mechanics merges the microscopic and macroscopic worlds. Thus, most CG methods are best cast

within the framework of the following formula for the Helmholtz free energy, A (Equation 1.3).³⁴

$$\exp\left(-\frac{A}{k_B T}\right) = \text{const.} \int d\mathbf{x} \exp\left[\frac{-V(\mathbf{x})}{k_B T}\right] \quad 1.3$$

$$\approx \text{const.} \int d\mathbf{x}_{CG} \exp\left[\frac{-V_{CG}(\mathbf{x}_{CG})}{k_B T}\right]$$

Where $V(\mathbf{x})$ and $\mathbf{x}$ refer to the potential energy and coordinates of all the atoms in system, T the temperature, k_B the Boltzmann constant and “const.” the normalization constant. The effective CG potential, $V_{CG}(\mathbf{x}_{CG})$, can be written in terms of $\mathbf{x}_{CG}$ variables that, by definition, are fewer in number the all atom variables.³⁴ Therefore, obtaining relevant information from all atom simulations for the parameterization of CG effective interaction potentials is of paramount importance, and should result in dynamics and thermodynamics similar to that obtained by all atom MD.

In a 2010 study, Ricci et al.³⁵ performed a massive CG-Monte-Carlo (MC) simulation of a sub-micrometric twisted nematic cell with nearly 10^6 particles using an off-lattice molecular model of a liquid crystal, which represents a single pixel of a LCD screen (Figure 1.2).

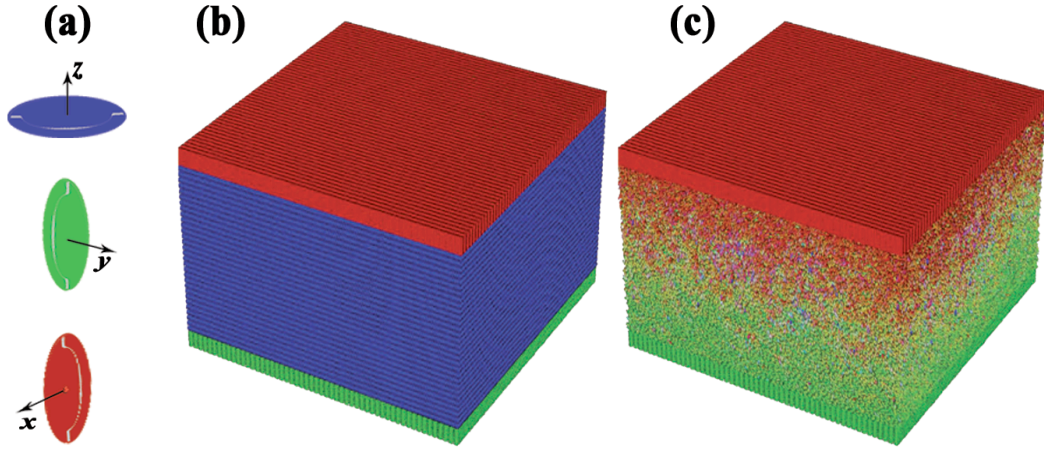


Figure 1.2: The (b) starting and (c) final configurations for a twisted nematic cell after relaxation. The ellipsoidally shaped particles are colour coded according to the orientation of molecular, $\mathbf{u}_i$, axis with respect to the laboratory frame shown in (a). Adapted from Ricci et al.³⁵

They used an effective anisotropic CG interaction potential, called the Gay-Berne potential (Section 1.5), to describe the interactions between the particles. The results demonstrated two important points in computer simulations of very large systems, using a reduced representation. First, was to show “that fairly massive parallel computing can now be applied to push the application of coarse-grained molecular

resolution simulations to the modeling of fairly realistic devices.”³⁵ And second, “to gain insight on the actual microscopic working of an important electro-optical device ...”³⁵ This study demonstrated that well parameterized CG models can *effectively* describe the anisotropic intermolecular interactions and can be used to investigate enormously large-sized (or real-world) systems.

1.1 Intermolecular Forces

Chemical systems do not consist of a single molecule but, rather, of a large number of molecules. To describe and understand many molecules we are required to enlarge the description to include intermolecular interactions. Intermolecular forces³⁶ of attraction are what give rise to liquid and solid phases, and are referred to as the weak electromagnetic interactions acting at a distance between molecules on a microscopic level.³⁷ They determine the chemical nature of molecular systems such as boiling points, viscosity of liquids, vapour pressure, the formation of surface monolayers, micelles and membranes, as well as transport of ions across biological membranes.^{1, 36}

To describe the intermolecular interactions we need to construct the partition function from two components. The first component is the molecular partition function, q , which is determined by the intramolecular interactions. The molecular partition function describes the different states that have a given energy, E .³⁷ The energy within a molecule can be distributed into different modes, where the total energy can be written as:

$$E = E_t + E_r + E_v + E_e + E_n, \quad 1.4$$

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

$$q = q_t \times q_r \times q_v \times q_e \times q_n. \quad 1.5$$

For systems of constant nuclear configuration, as is the case for the 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 translational and rotational energies, of molecules, are typically much less than $k_B T$, and can be described classically quite accurately. The vibrational energy can be very high depending on the atoms involved. Oxygen-hydrogen bonds, for example, have a very high vibrational frequency and should be described quantum

mechanically or have a constraint placed on them; while slower vibrational motions, such as the bond stretching between heavy atoms, can be described classically.^{37, 38}

The second component, Z , is the configuration partition function, which is determined by the intermolecular interactions.^{37, 38} It has been established that only the theory of quantum electrodynamics is required for calculations in typical chemical systems for which we are interested in. The theory was developed between the 1900s and 1950s and is very well tested,^{39, 40} but it remains too intricate to use on systems at the scale that we are interested (i.e., systems with tens of thousands of atoms). Thus, approximations must be introduced to perform computations on molecular systems of interest to chemists and biophysicists. The first approximation is to use reduced descriptions of quantum systems such as the Hartree-Fock⁴¹ and density functional theory methods.⁴² This description is accurate and can be done on small molecular systems, but the computations still remain too expensive for the spatial and temporal scales that are of interest in biophysics, chemical biology and materials science.

This high computational expense leads to the requirement that a system of molecules needs to be described effectively and classically. In this description, electrodynamics interactions can be split into two distinct groups: intramolecular (or short-range) and intermolecular (or long-range) interactions (Table 1.1).

Table 1.1: A summary of the short- and long-range interactions that is present in molecular systems.⁴³

Short-Range Contribution	Pairwise Additive	Sign
Exchange	No	Negative
Repulsion	No	Positive
Charge Transfer	No	Negative
Penetration	Yes	Negative
Damping	Approximately	Positive
Long-Range Contribution		
Electrostatic	Yes	Positive or Negative
Induction	No	Negative
Dispersion (van der Waals)	Approximately	Positive or Negative
Resonance	No	Positive or Negative
Magnetic	Yes	Positive or Negative

The intramolecular interactions consist of potentials that are parameterized to reproduce experimental measurements and/or quantum mechanical calculations.^{18, 19} This is done to give the correct behaviour of molecules in the simulation. It takes into account bond and angle vibrations and dihedral angle rotations, while ensuring that the atoms of the molecules remain together. The intermolecular interactions are, to date, typically described using atomic pair potentials. Intermolecular interactions are referred to as long-range interactions, since they describe the interactions of non-bonded atoms between different molecules. They are constructed to reproduce experimental measurements (e.g. heat capacities) and to be consistent with quantum chemical calculations of atomic interactions.

1.1.1 Short-range interactions

Short-range interactions arise when the electron densities of atoms, either open or closed shell, interact. Upon overlap, several interactions may occur that include exchange repulsion, charge transfer, electrostatic penetration and damping. Of these interactions, exchange repulsion is most prominent and is used to describe the repulsive force associated with the interaction of two closed shell systems. This short-range repulsion occurs when the wave function of two interacting atoms overlap, making the electrons, belonging to each respective atom, indistinguishable from each other.

The overlap between the orbitals of two atoms, A and B , is measured by the overlap integral, which is integrated over all space as

$$S = \int_{all\ space} \phi_A \phi_B d\tau. \quad 1.6$$

If A and B are separated by a significant distance, such that ϕ_A and ϕ_B do not overlap, then S will become zero. However if there is overlap, then the exchange energy involves the product of the two orbital overlap functions such that it is approximately proportional to S^2 .

1.1.2 Long-range interactions

The long-range interactions can be described by pairwise potentials, which are easy and quick to calculate by only including pair interactions. This greatly reduces the number of computations that are required for the system. Pair potentials are parameterized to incorporate not only the two-body interactions present in the system, but also to approximate many-body effects. Therefore, they are referred to as effective pair interaction potentials.^{18, 19} Long-range interactions can be classified into various

types: electrostatics, induction, dispersion and van der Waals and nuclear core repulsion (both described by the Lennard-Jones 12-6 potential⁴⁴). To calculate the interaction energy classically, the sum of the electrostatic and the Lennard-Jones interactions is required. The electrostatic component applies to charged species and only knowledge of the molecular charge distribution is required. Once determined, the induction and dispersion energies can be calculated based on the behaviour of the same chemical system's electrons under the influence of an external electric field. This implies that induction and dispersion interactions are polarization dependent. For the Lennard-Jones potential, which does not require knowledge of the electronic structure, only the size and interaction strength for each of the atoms is required.

The electrostatic interaction energy, between two molecules, is calculated as the sum of the Coulombic interactions between the particles. This can be obtained from the nuclear-nuclear and electron-electron repulsions and the nuclear-electron attractions, if the nuclei positions and electron distributions are known. However, in large molecular systems that are commonly encountered in MD simulations nowadays, this is not feasible.⁴⁵ In these situations a mathematical simplification of the problem has to be introduced, which is known as the multipole expansion. Using this approach, the attractive and repulsive terms are combined such that the electrostatic energy can be equated to the energy of interaction for the net charges and multipole moments (e.g. monopoles, dipoles, etc.), which arise from a non-uniform distribution of electric charge, $\rho(r)$, within a molecule.

For an electric charge distribution, $\rho(r)$, the potential at point P , at a distance R from the centre of mass of $\rho(r)$, integrated over all space is given by:

$$V(P) = \frac{1}{4\pi\epsilon_o} \int_{all\ space} \frac{\rho(r)}{|R-r|} d\tau. \quad 1.7$$

By assuming that $R \gg r$, the expansion can be rewritten as:

$$V(P) = \frac{1}{4\pi\epsilon_o} \sum_{n=1}^{\infty} \frac{1}{R^{(a+b+1)}} \int_{all\ space} r^n P_n(\cos\theta) \rho(r) d\tau, \quad 1.8$$

where $P_n(x)$ is a Legendre polynomial of order n and the angular term $\cos\theta$ is given by the dot product of the two vectors $\hat{r}$ and $\hat{R}$. The multipole moments arise from the interaction of a site (e.g. monopole, dipole, etc.) of order a , with a site of order b . The terms a and b refer to the multipole expansion coefficients, where a and b equal to 0 for the zeroth charge moment (Coulombic interaction), 1 for the dipole moment, 2 for the

quadrupole moment, etc. If generalized, the multipole expansion giving the electrostatic interaction between two molecules is only valid when there is negligible overlap of the electron clouds of the two molecules. From the generalization, coefficients corresponding to the charge-charge, charge-dipole, dipole-dipole and higher order moments are obtained.³⁶

Inert gases, which have no permanent moment, would be assumed not to interact. However, inert gases are found to deviate from the ideal gas law because they possess instantaneous moments, arising from fluctuations in their electronic distributions. These fluctuations induce instantaneous moments in neighbouring molecules and there is a resultant energy of interaction. This resultant non-zero polarization energy is called the dispersion energy – first derived by F. London for the theory of dispersion. The Lennard-Jones potential describes the weakly attractive dispersion (r^{-6}) and repulsive core (r^{-12}) interactions given by

$$E_{LJ} = 4\varepsilon \left[\left(\frac{\sigma}{r} \right)^{12} - \left(\frac{\sigma}{r} \right)^6 \right], \quad 1.9$$

where ε is the strength of the interaction between the atoms and σ is the separation distance at which the maximum interaction occurs (see Section 1.5 for further details).

Therefore, for two molecules the sum of these long-range interactions, or effective pair interaction potential terms, makes up the intermolecular interaction. Although the individual atomic potential terms, discussed up to this point, are isotropic, the sum over all the atoms of two molecules is not. This is due to the fact that individual molecules are rarely spherically symmetric. Any asymmetry in a molecule will introduce asymmetries in the interactions of that molecule with others. A description of intermolecular interactions between molecules must take into account the anisotropy in the interaction.⁴⁵⁻⁵⁵ Methods to do this include using multipole expansions, described above, analytic two- and three-body functions and analytic functions of one or more variables that take into account relative orientations, such as the Gay-Berne potential (see Section 1.5).^{27, 45-47, 50-53, 56-58} Most of these methods have as their primary objective to obtain a simpler, but nonetheless, accurate anisotropic description of molecules, which can be used for large scale simulations. In molecular dynamics, where an isotropic potential energy function is used such as the Lennard-Jones, certain functions are used to ensure that the potential is zero beyond a certain distance, called the cutoff radius (see Chapter 2). However, this isotropic type of formulism cannot simply be employed when using an anisotropic potential, due to the different lengths of an ellipsoids semi-axes. While

stacked configurations do interact at 10 Å, end-to-end configurations are still able to interact. A function considering the relative orientation must therefore be considered.

1.2 Thermodynamics of Intermolecular Forces

The intermolecular forces described are as a result of the direct interactions between molecules arising from the quantum electrodynamical interactions between the atoms that make up the molecules. However, there are other effective forces that occur from multiple solvent interactions, bulk effects and other averaged effects. To describe such effects we use the language of statistical mechanics, which relates the macroscopic properties stemming from a very large number of microscopic particles.

The atoms of molecules are not fixed in place but can vibrate and translate. This transforms into the interaction occurring between two molecules as a function of the configurations of the atomic positions of the molecules. These fluctuations are controlled primarily by the intermolecular interactions. The total configurational space that the molecules may sample is enormous, but they do not sample all points uniformly. A reduced representation of the intermolecular interactions could be given for only the most frequently sampled configurations. However, this would still require a complicated description of intermolecular interactions and, thus, a more common method is to average out fluctuations with an appropriate bias to the areas of configurational space sampled most frequently.

Another important effect, which has to be considered, is the interactions between molecules in the presence of others (e.g. solvent, ions, etc). The molecules may also undergo similar 'internal' fluctuations and move about the molecule(s) of interest. This greatly increases the size of the configurational space that needs to be taken into account. These surrounding molecules can greatly affect the interactions, where a good example of this being the process of protein folding. The configurational space of a protein is very large and it would take an astronomically long time to reach the folded state based solely on a conformational search.^{59, 60} Therefore, nature makes use of the environment (or solvent) around the protein to drive the folding process in biologically realistic time scales.

The effect of solvent interactions can be either direct (enthalpic), by assisting the association of two molecules via solvent bridges, or indirect (entropic). To maximize the number of available states of the solvent often implies reducing the configurational space of the associating molecules. This can then drive molecules together where they

otherwise would not associate, and forms the basis for the hydrophobic effect. Using the protein-folding example again – proteins have a large number of atoms and, hence, its conformational space is extremely large. In a thought experiment done by Cyrus Levinthal, it was shown that if a protein were to fold in an undirected (or random) search of its available conformational space, it would take a time longer than the age of the universe to reach the final, folded state. However, proteins fold much faster and this paradox has become known as Levinthal's paradox.^{59, 60} Since, protein folding occurs so quickly there has to be some directed mechanism involved.⁶¹

It is generally accepted that proteins fold on a funnel-like energy landscape,⁶² where the folding process can be described as a two state process (Figure 1.3).⁶³ The initial state corresponds to a large number of conformational states, i.e. the unfolded state, and the final state corresponds to a single or very few conformational states, i.e. folded state.

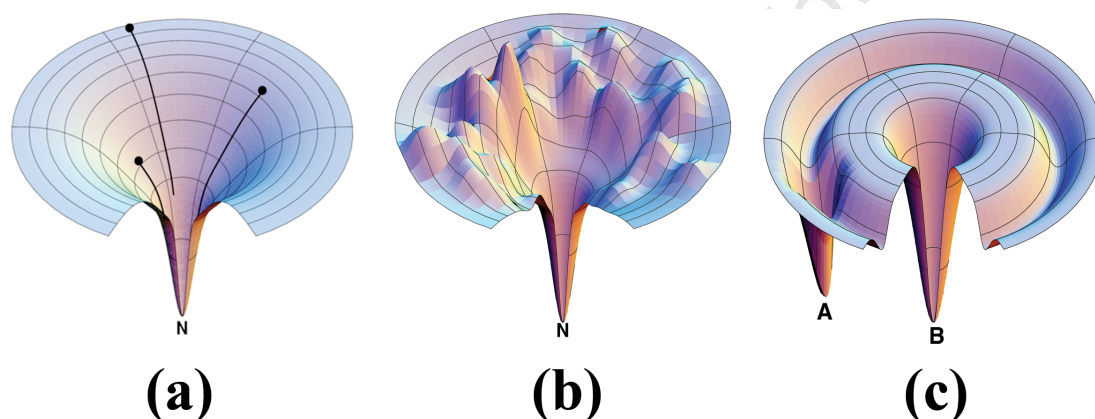


Figure 1.3: The folding funnel concept:^{62, 64} (a) in its simplest form, the straight collapse into native state; (b) the more rugged landscape with only a few pathways; and (c) a funnel landscape involving a kinetic trap A and the native state B.⁶³

Proteins consist of a long chain of amino acids and the specific interactions, between these amino acid residues, 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 protein folding process.^{61, 65-67} It is also widely accepted that the hydrophobic effect is one of the major driving forces for the folding of globular proteins.¹⁰

To take into account the interactions between molecules (the enthalpy) and the effects of entropy, 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; thus, it is called the potential of mean force (PMF). The PMF is exactly the free energy, and describes the effective force that one molecule exerts on another. Calculating the free energy surface from MD is not a trivial task. Various methods such

as free energy perturbation, thermodynamic integration and umbrella sampling have been developed to calculate the free energy.^{19, 68-71} These techniques have been employed to study the interaction in many systems from proteins^{12, 48, 50-52, 68, 70, 72-77} to simple organic molecules⁷⁸⁻⁸³ to solvent molecules.⁸⁴⁻⁸⁷

1.3 Hydrogen Bonding, Hydrophobic and Hydrophilic Interactions

A system of particular interest to chemists and biologists is water. It is the most abundant molecule on earth, in our bodies and it is particularly important for mediating and providing an environment for processes that are responsible for life. It is also one of the most anomalous molecules in having, for example, a very high melting and boiling point. These properties are a result of the particular association between water molecules called hydrogen bonding, Figure 1.4.⁸⁸ A hydrogen bond occurs when a hydrogen atom is strongly attracted to two other electronegative atoms. In water, this is seen when the hydrogen forming a covalent bond with the oxygen of a water molecule is strongly attracted to an oxygen atom of another water molecule. The nature of this attraction is primarily electrostatic.⁸⁹ It does, however, also contain a covalent part imparting directionality to the interaction, which if not explicitly taken into account will result in MD simulations that will incorrectly describe hydrogen bonding. It is due to hydrogen bonding that there is an observed tetrahedral structure of ice and near tetrahedral structure in liquid water. The prototypical hydrogen bond is described by the directionality of the interaction between water molecules (Figure 1.4).

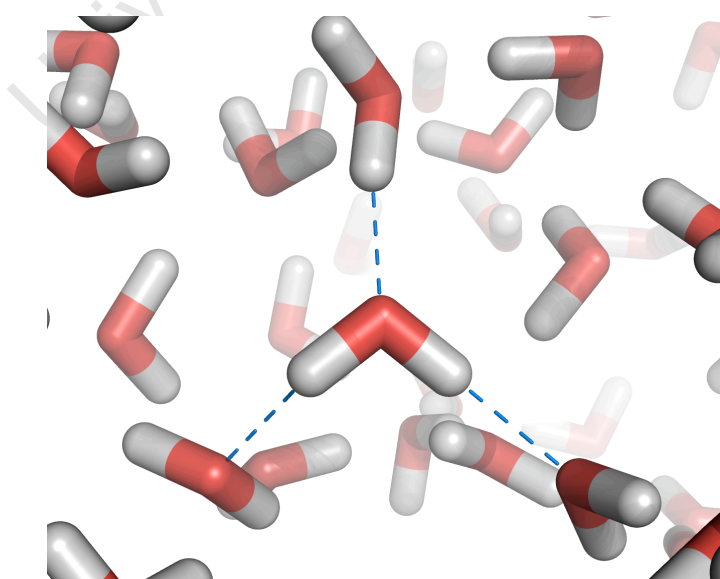


Figure 1.4: Hydrogen bonding between water molecules showing the tetrahedral structure of liquid water, which gives water its unique properties.

A number of water models for use in the simulations have been developed. These models include all atom approximations, three-, four- and five-site models, polarizable and flexible models.^{19, 49, 54, 86, 89-98} These models differ in their local and bulk properties, but all reproduce the results for a wide range of water properties reasonably well.^{90, 99} The importance of treating water molecules explicitly, and with high detail, is well understood and has become more prominent as a standard for running MD simulations. The consequence of this has been that most MD simulations use explicit, fixed charge and rigid three- or four-site water models.^{90, 99}

Molecules in the presence of other molecules, that interact via electrostatic and van der Waals interactions, experience various effective forces. A particularly prominent solvent effect (especially for water), related to hydrogen bonding, is the hydrophobic effect. This effect is intuitively associated with the notion of oil not mixing with water in the formation of oil droplets.¹⁰⁰ The current understanding of the hydrophobic effect is that a polar solvent, when surrounding a non-polar solute, will orient and rearrange itself so as to minimize charge contact with the non-polar solute. The effect of this is that when there are two non-polar solutes near each other, the surrounding water molecules will arrange to minimize the contact area and, in doing so, form an elastic hydrogen bond network surrounding the solutes.^{100, 101} The effect is that the two solutes are pushed closer together reducing the solutes configurational space resulting in a decrease in entropy, while the aqueous environment increases the entropy due to the release of bound water molecules. However, in a few recent studies, involving the association between a hydrophobic molecule and hydrophobic binding pocket, showed that the hydrophobic effect was being driven by enthalpy.¹⁰²⁻¹⁰⁴ They determined that shape of the interacting sites resulted in hydrophobic association being enthalpically dominated. This resulted from the difference in the thermodynamic signatures between water molecules near the hydrophobic surface and the bulk. Setny et al. showed that the favourable driving enthalpy change results from the release of water molecules from the hydrophobic environment to bulk water.¹⁰³ The actual solvent behaviour critically depends on the strength of the solute-solvent interaction and varies from the persistent hydration of strongly interacting objects.^{103, 105, 106} In a 2005 article, Chandler provided a detailed review on the varying role the hydrophobic effect plays on contrasting length scales.¹⁰⁷ For small hydrophobic solutes, water pays entropically as it makes a similar number of hydrogen bonds as in bulk. However, around large hydrophobic solutes or at interfaces water cannot form as many hydrogen bonds and, therefore, an interfacial enthalpic penalty is incurred enthalpic penalty is incurred.¹⁰⁷

The hydrophobic effect is considered to be a very important, initial driving force responsible for the mechanism for protein folding,¹⁰ and in the formation micelle and lipid bilayer.^{12, 100, 101, 108-110} 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 of the hydrophobic effect. Instead of trying to minimize the non-polar and polar contacts, the system attempts to maximize the polar-polar interactions. This was proposed¹² and disputed¹¹¹ as being a major contributor to protein folding.

1.4 Molecular Association

Molecular association plays a critical role in all facets of nature. The strength of molecular associations, arising from intermolecular interactions, has great significance in describing the microscopic structure and macroscopic observables of molecules. Examples include the recognition of specific nucleotide sequences by DNA binding proteins governed by amino acid interactions,¹¹² cellular communication and transport,^{113, 114} the role that $CH \cdots \pi$ interactions play in crystal packing and clathrate structures,¹¹⁵ the mechanism of crystallization,¹¹⁶ metal-ligand interactions and metallofoldamers upon ion coordination,¹¹⁷ the liquid crystal phase,¹¹⁸ the hydrophobic effect and the role it plays in protein folding^{100, 119-124} and the enhanced stabilization of secondary structures by non-covalent forces, i.e. hydrogen bonding.

Molecular association is a complex entity to describe in that it is not only dependent on the distance between non-bonded interacting molecules, but also on their relative orientation. For example, the sandwich and T configurations for benzene dimer are not found to be the global minima on the potential energy surface, but rather the parallel displaced and T-like (Bent-T and Y-shape) structures are the preferred arrangements.¹²⁵⁻¹³⁰ Furthermore, the competition between the solvent-solvent (e.g. solvation or hydration energy), solvent-solute (e.g. hydrogen bonding between water and glucose) and solute-solute (e.g. crystal lattice energy) interactions also play a crucial role in association,¹³¹ solubility,^{132, 133} ion pairing,¹³⁴ reaction rates¹³⁵⁻¹³⁸ and stability.¹³⁹

The strength of molecular interactions is also responsible for the phases of matter to which we have become accustomed. The interplay between molecular association (determined by the intermolecular forces) and molecular motion (determined by the kinetic energy and temperature) determines the phase in which matter exists (Figure 1.5a). Gases have the greatest degree of molecular motion with the least amount of association and appear chaotic, while solids exhibit the opposite behaviour, i.e. greater

molecular association and order that persists at larger distances.¹⁴⁰ Liquids, on the other hand, have strong forces of attraction due short intermolecular distances, while the amount of kinetic energy is sufficient to allow for tumbling of molecules. This leads to liquids having the attributes of both gases and solids – i.e. the randomness of gases (at long-range) and the structural stability of solids (at short-range). This order (or lack thereof) can be observed by using pair distribution functions (PDFs), from which it is possible to obtain an idea of the magnitude of molecular ordering (Figure 1.5b.)

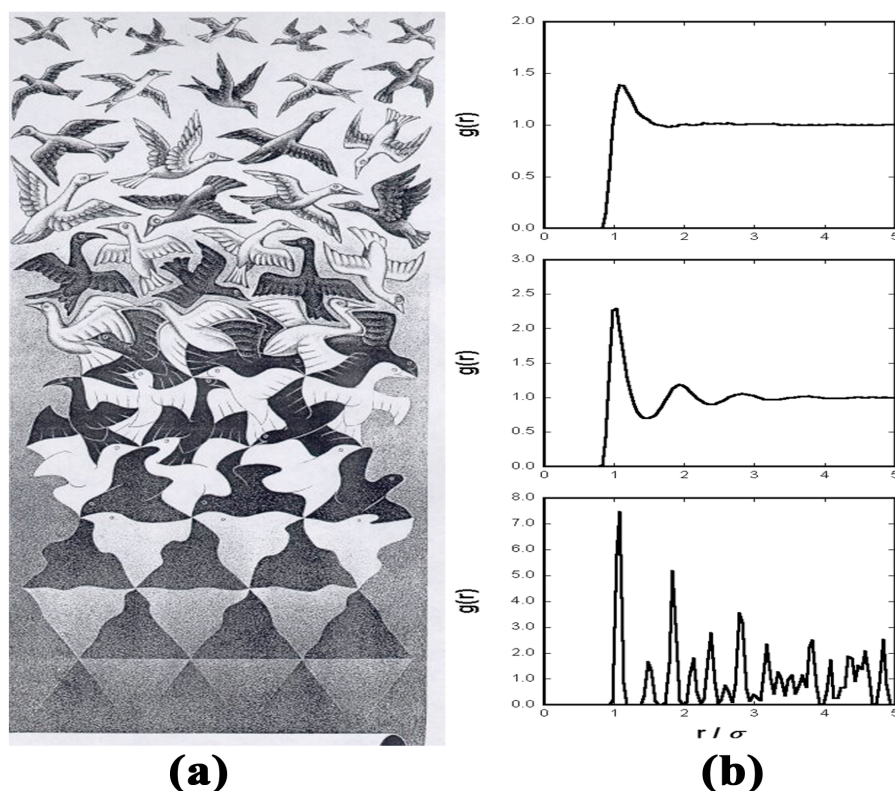


Figure 1.5: (a) Artwork entitled ‘Liberation’ shows all of the major points on the phases of matter.¹⁴¹ At the very bottom are rectilinear objects that fit perfectly, cover all the space and attain a packing density of unity – a solid, long-range order. In the middle layer the objects fail to fill the space and cavities appear, but only restricted motions is possible – a liquid, medium-range order. At the top the cavity space has been increased to the point where restrictions on motion have disappeared – a gas, weak short-range order. (b) The pair distribution functions (PDFs) for a solid, liquid and gas, corresponding to the layers in (a).¹⁴¹

1.5 Model Intermolecular Potentials

Analytical mathematical functions, constructed of terms defining contributions to the intermolecular potential derived from *ab initio* quantum mechanical calculations and experimental data, are routinely used to model the interactions for small to medium sized molecular systems. Several empirical potentials, shown in Figure 1.6, have been used to model the dispersion and repulsion interactions. Hard-sphere and square-well

(Figure 1.6a and b) potentials were used to study liquid properties; however, both types of potentials are discontinuous functions that lead to poor representation of the potential at large distances. There is one potential that has been the basis of most analytical investigations of atomic fluids mostly in part due to its low computational cost, accurate description of medium range interactions and, at the same time, its close representation to real potentials.¹⁴² This statistical potential, called the Lennard-Jones 12-6 potential (a special case of the Mie potential¹⁴³), first described by Sir John Edward Lennard-Jones in 1924, has the following form:¹⁴⁴

$$U_{LJ} = 4\varepsilon \left[\left(\frac{\sigma}{r} \right)^{12} - \left(\frac{\sigma}{r} \right)^6 \right], \quad 1.10$$

where ε is the interatomic interaction strength and σ is the interatomic size. The Lennard-Jones is a particularly impressive potential in that it describes the repulsion, by the r^{-12} term (accounting for the Pauli exclusion principle) and dispersion, by the r^{-6} term (to account for the attraction). U_{LJ} is zero when $r \rightarrow \infty$, while it has a global minimum at $r_{min} = 2^{1/6}\sigma$ where $U_{LJ} = -\varepsilon$ (Figure 1.6c and d).

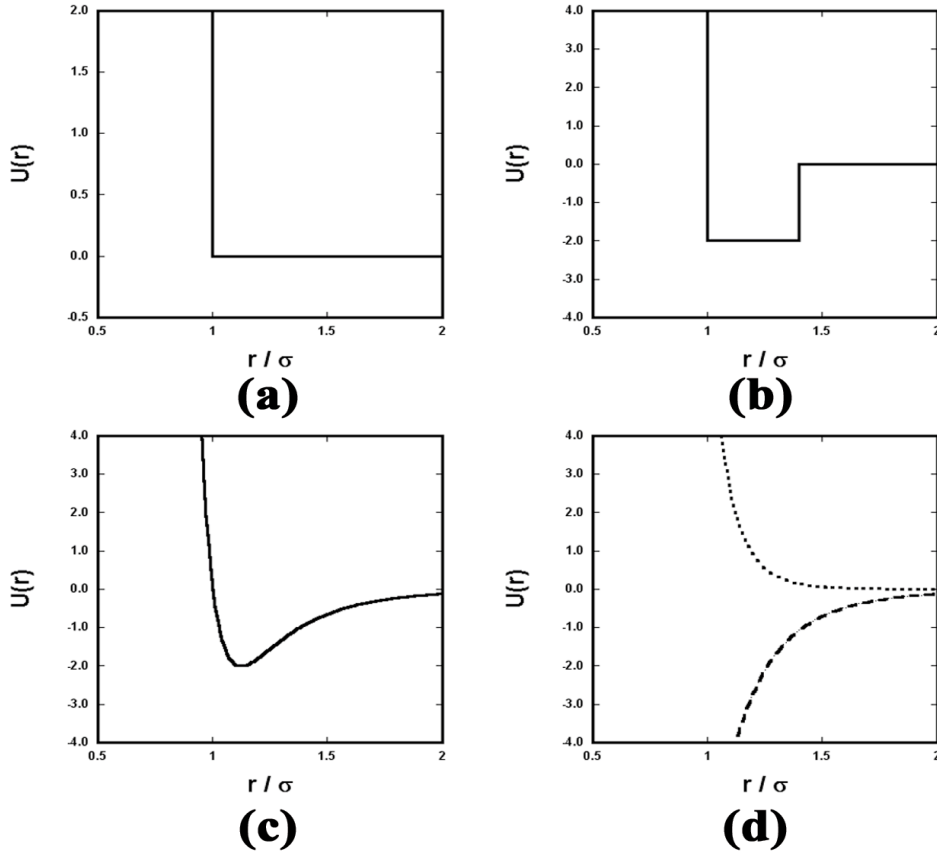


Figure 1.6: The idealized pair potentials: (a) hard-sphere potential; (b) square-well potential; (c) Lennard-Jones (LJ) potential, and its (d) the repulsive r^{-12} (dotted line) and attractive r^{-6} (dashed line) contributions.

These potentials only describe the repulsion and dispersion based on atomic size and propensity to associate; it does not take into account the charge that may reside on an atom or molecule. To accurately simulate molecules, with charge, requires the inclusion of an electrostatic term.¹⁴² Ideally one would like to model the electrostatics using the full multipole expansion (Section 1.1.2); however, the large computational cost of a distributed-multipole mode precludes this and the electrostatic energy, U_{Coul} , is more commonly described by atomic point charges and using Coulomb's law.

$$U_{Coul} = \sum_{j>i} \frac{q_i q_j}{4\pi\epsilon_o r_{ij}}, \quad 1.11$$

where q_i and q_j the atomic charges, r_{ij} is the interatomic distance and ϵ_o (exact value of $8.85419 \times 10^{-12} \text{ F.m}^{-1}$) is the permittivity of free space.

The above intermolecular potentials, which contribute to the overall potential energy, are based on that the interacting atoms that are spherical and only distance dependent. For large rigid molecules, a substantial number of sites would be required to model the pairwise interactions, resulting in a large amount of computation. For example, a crude model of the dimyristoylphosphatidylcholine (DMPC) lipid can be represented by 10 reduced sites (or macrounits)²¹ and would necessitate 100 site-site interaction between each pair of DMPC molecules. However, sites based on the all atom representation, which contains more than 100 atoms, would be more realistic but extremely computationally expensive. An alternative type of intermolecular potential, introduced by Corner in 1948,¹⁴⁵ involves a single site-site interaction between a pair of molecules, characterized by interaction strength (or energy) and size (or range) parameters that depend on the relative orientation of the molecules.¹⁴² This reduces the description of the system, resulting in fewer computations. A version of this family of molecular potentials, used in computer simulation studies, is the Gaussian overlap model generalized in a Lennard-Jones form, called the Gay-Berne potential.¹⁴⁶ The basic potential acting between two linear molecules is the Lennard-Jones interaction, Equation 1.10, where the angular relation between the two particles anisotropically describes ϵ and σ . This is achieved by considering the overlap of two ellipsoidal Gaussian functions, which represents the electron clouds of molecule.¹⁴² The anisotropic forms of ϵ and σ parameters have the following form:

$$\epsilon(u_i, u_j) = \epsilon_{sph} \left[1 - \chi^2 (u_i \cdot u_j)^2 \right]^{-1/2} \quad 1.12$$

$$\sigma(r_{ij}, u_i, u_j) = \sigma_{sph} \left[1 - \frac{1}{2} \chi \left(\frac{(r_{ij} \cdot u_i + r_{ij} \cdot u_j)^2}{r_{ij}^2 (1 + \chi(u_i \cdot u_j))} + \frac{(r_{ij} \cdot u_i + r_{ij} \cdot u_j)^2}{r_{ij}^2 (1 - \chi(u_i \cdot u_j))} \right) \right], \quad 1.13$$

where ϵ_{sph} and σ_{sph} are constants and χ is the anisotropy parameter relating the semi-major and semi-minor axes of the ellipsoid. This potential can describe the anisotropic interactions between molecules. Further discussion of this potential, its parameters and parameterization strategy is given in much more detail in Chapter 7.

1.6 How to Reduce Degrees of Freedom?

The modeling of polymeric materials is computationally challenging,^{22, 32} of high technological importance,^{22, 147} but also very fascinating from a general statistical mechanics point of view. All atom MD simulations would be ideal to study such systems, but the large computational expense prevents this. Therefore, to conduct simulations of polymeric materials, such as proteins, a reduced representation of the system has to be employed to reach the time scales at which biological events occur. There are several ways in which a molecule, or a portion of a molecule, can be reduced to decrease the overall computational cost. However, there are many decisions that must be made including the nature of the dynamics – conservative or dissipative – and the functional form of the interaction potentials. The way in which the reduction is performed has to be compliant with the all atom version of the molecule. Examples of system representation reduction include the aliphatic lipid tail moiety,^{21, 31, 148} water²⁵ and benzene (Figure 1.7).²³

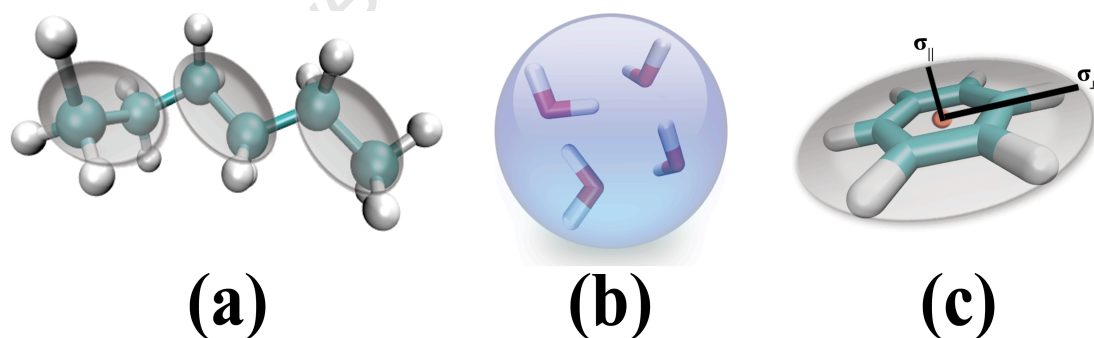


Figure 1.7: (a) Spheroidal coarse-grain (CG) mapping commonly used for the ethylene group of the tail moiety of a phospholipid (ellipsoid); (b) CG water site represented by 4 water molecules; and (c) CG benzene as an ellipsoid.

A common motif employed by many simulation methodologies is lumping groups of atoms together into a single interaction site. Examples of which are the CG model of Fukunaga,¹⁴⁸ the dissipative particle dynamics (DPD) model of Groot and Rabone¹⁴⁹ and the multiscale method of Goddard.¹⁵⁰ It is important not to overly distort the

geometrical shape of the molecules when applying this grouping. As an illustrative example, for amphiphilic molecules the head group size compared to the average tail size, normal to its length, determines whether micelles or inverse micelles are preferred. Cone shape molecules, such as phosphatidylethanolamine, have small head groups and tend to form inverse micelles; while inverted cone shaped molecules, such as lysophosphatidylcholine, tend to form micelles. Some models use anisotropic interaction sites to capture the underlying shape; one of the simplest is the Gay-Berne potential,⁵⁸ which is commonly used in liquid crystal simulations. It has also been used in lipid CG models³¹ and in united-residue protein folding models.^{27, 28}

The electrostatic partial charges may also be considered when choosing the CG grouping, if electrostatics is to be incorporated in the model. All atom MD force fields typically have atom-centred partial charges to capture electron-withdrawing capabilities of certain functional groups. Lumping atoms together into a single site imbues that site with a partial charge, which is the sum of all its constituent atomic charges. If this is not satisfactory, the interaction sites may have to be reduced in size or have their composition altered. One possibility is to associate several fictitious charges with an interaction site so as to reproduce the correct long-range electrostatics.

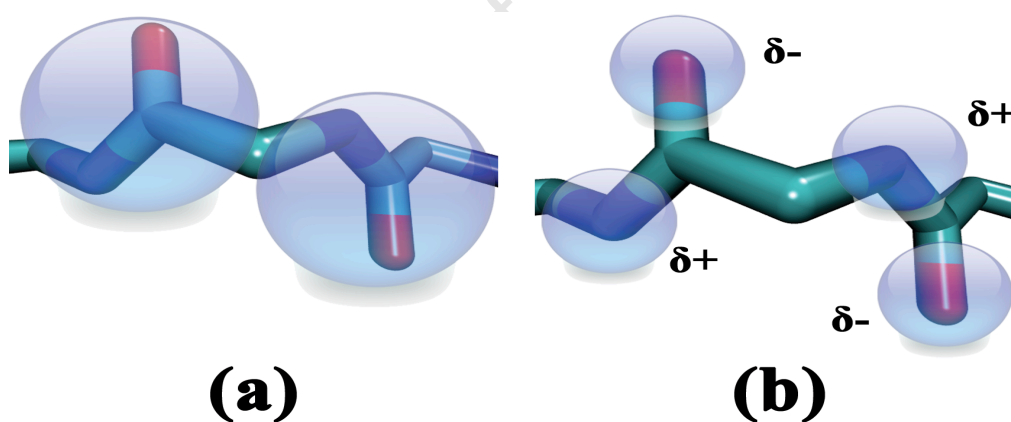


Figure 1.8: Modeling the electrostatics of the peptide backbone as either using (a) a grouped neutral entity ($-\text{C}_\alpha\text{H}-\text{CO}-\text{NH}-$) or using (b) separated partial positive ($-\text{NH}-\text{C}_\alpha\text{H}-$) and negative ($-\text{CO}-$) charges.

The backbone unit could be modeled as the electrically neutral $-\text{C}_\alpha\text{H}-\text{CO}-\text{NH}-$ site (with possibly two fictitious sites), Figure 1.8a, or as two sites, $-\text{CO}-$ and $-\text{NH}-\text{C}_\alpha\text{H}-$, with equal positive and negative partial charges, Figure 1.8b. The protein folding community often uses a single, uncharged site.^{27, 151} An alternative is to use a multipole expansion to capture the effect of an anisotropic charge distribution.²³ Clearly there are a plethora of options in designing reduced models. Many of these choices necessitate writing new simulation software to implement unusual or novel Hamiltonians.

1.7 Coarse-Grain Dynamics

Interesting chemical and biochemical processes, such as chemical reactions and protein folding, have been shown to occur over a broad range of spatial (size), from pico- to micrometres, and temporal (time) scales, from femtosecond dynamics to real-time macroscopic phenomena. The available simulation techniques for specific spatial scales and time scales are shown in Figure 1.9.²⁰

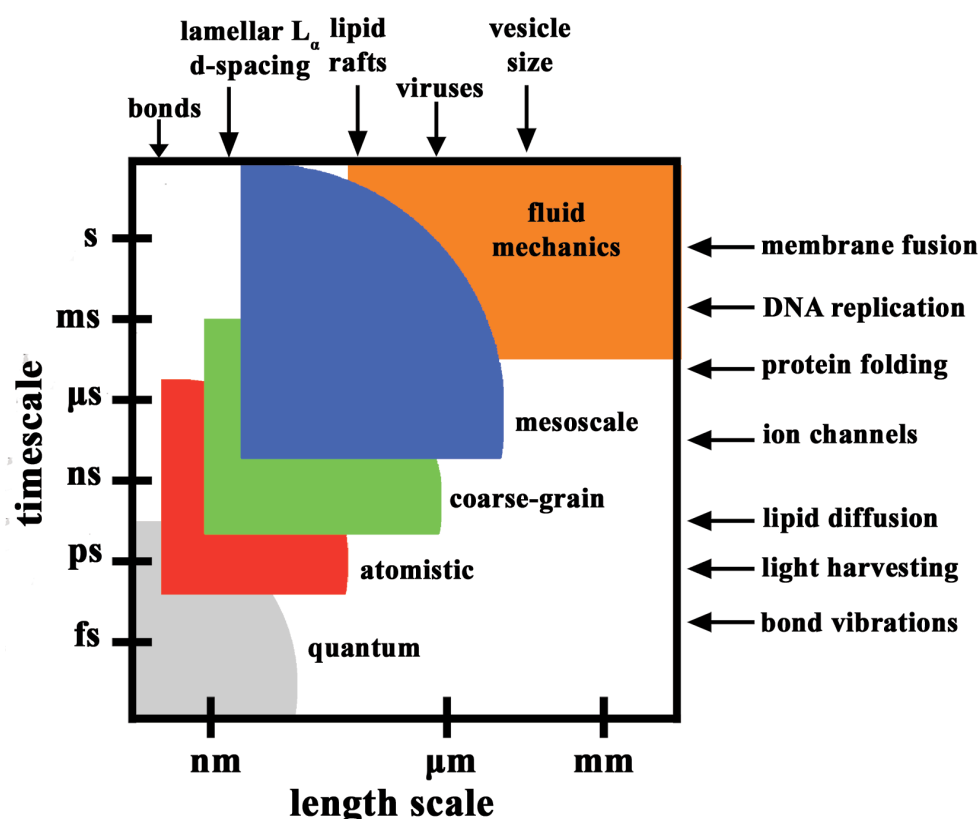


Figure 1.9: A schematic diagram of the accessible temporal and spatial scales by simulations, ranging from a very low-level description (quantum mechanics, which correlates electronic motion) to high-level description (as in fluid mechanics for studying air flow around supercars in a wind tunnel). Adapted from reference 21.²⁰

These techniques use a variety of approaches to reduce the level of detail in the representation of the system under consideration as the timescale and/or length scale grows. Bridging the gaps between the various levels is possible with multiscale modeling^{35, 152-155} in which the various levels are coupled and fed back into each other.²⁰ Reduced models, which preserve information from atomistic representations, allow the study of events that occur on timescales of hundreds of nanoseconds to milliseconds and spatial scales of microns.²⁰ Once a suitable means of reduction for the system has been decided upon, and the effective CG interaction potential chosen, the CG dynamics simulations can be performed.

There are many phenomena that lie within the mesoscopic spatio-temporal scale that may be explored with CG methods in biology and materials science.^{20, 148, 156} Examples of these applications have been in the fields of protein folding^{26-28, 157, 158}, dendrimers and polymers^{148, 150, 159}, various liquid crystals phases (e.g. smectic and nematic),^{35, 160-168} nucleic acid and nucleic acid-protein interactions,^{29, 169} simple organic^{23, 24} and water systems,²⁵ vesicle fusion,¹⁷⁰ lipid membranes and micelles^{21, 31, 149, 170} to name a few. The list of applications of CG methods to large-scale systems is already large and current trends indicate that it will continue to grow.²⁰ These systems and events can be studied due to the reduced representation of CG methods, which allows access to spatial scales and time scales much larger than that can be achieved by all atom representations. As mentioned before, Ricci et al. were able to use coarse-grain dynamics to investigate a twisted nematic cell with nearly 10^6 particles (representing a single pixel of a liquid crystal display) using an off-lattice molecular model of a liquid crystal.³⁵

There are several limitations, which occur when representing by coarse-graining a system.^{20, 171} It is well known that all models that use effective pair potentials, including condensed phase all-atom, united-atom and CG force fields, are thermodynamically inconsistent.^{20, 159, 172} Louis¹⁷² concludes that force fields that use effective pair potentials can be chosen to perform well for some particular physical properties of interest, which is clearly demonstrated for the various water models that replicate some physical properties better than others.^{86, 92, 94, 99, 173} CG force fields are usually parameterized from pair distribution functions (PDFs), all atom simulations¹⁷⁴ or experiments.¹⁷⁵ This has the advantage of incorporating the structural information into the CG model, but there is a loss of transferability. This is part of the trade-off in moving to a more efficient simulation method; generality is sacrificed.²⁰ However, the gain in speed, and the ability to reach much larger spatio-temporal scales, allows the study of very large systems that would otherwise not be accessible.

According to Nielsen et al., there are three fronts along which the current coarse-grain methods can be continued.²⁰ Firstly, systems can be studied whose constituent parts have already been parameterized. Secondly, existing parameters can be used as building blocks for new species. Thirdly, new species can be fully parameterized from a combination of atomistic simulations and experimental observables.²⁰ Undoubtedly, there are many possible futures for CG models.²⁰

1.8 Objectives and Overview of Thesis

This thesis is based primarily on statistical mechanics, its applications to calculating free energy and laying the foundation for CG dynamics simulations. This work can conceptually be divided into two parts: (i) the use of statistical mechanics for obtaining multi-dimensional anisotropic free energy hyper-volumes for aromatic association; and (ii) the parameterization of an anisotropic CG interaction potential, from the free energy hyper-volumes, and the subsequent CG dynamics and analysis.

Chapter 2 provides a concise description of the MD method, details of the Hamiltonian, calculation of long-range interactions, integration schemes, periodic boundary conditions and potential energy truncation. Also, some of the basic statistical analysis tools are introduced at the end of this chapter.

Chapter 3 suggests a few texts to read in order to review classical thermodynamics and its fundamental laws, which lays the groundwork for the following sections. A brief account of statistical mechanical ideas, fundamental postulates and ensemble theory, which is normal practice in any statistical mechanics text. Finally a concise history of free energy methodologies is given, along with more modern approaches. The chapter concludes with the free energy method employed throughout this thesis, the Free Energy from Adaptive Reaction Coordinate Forces (*FEARCF*) and Weighted Histogram Analysis Method (*WHAM*) both implemented in multi-dimensions.

The first part of this thesis begins with Chapter 4, which presents the first set of published results. In this chapter, the *FEARCF* and *WHAM* methods are used to determine the orientational preference of the benzene dimer in both vacuum and solution. We also investigate the thermodynamics of the benzene dimer in solution. The vacuum free energy hyper-volume, produced in this chapter, will serve as a foundation for the parameterization of the CG (Gay-Berne) interaction potential later.

Chapter 5, by virtue, is a natural extension of Chapter 4 to include the biologically relevant aromatic amino acid side chains. Here the *FEARCF* method is used to investigate the orientational preference and thermodynamics of aromatic amino acid side chains. The amino acid pairs focused on are phenylalanine dimer (PHE-PHE), tyrosine dimer (TYR-TYR) and PHE-TYR pairs. An analysis of these aromatic amino acid pairs are performed on approximately 5877 protein crystal structures to produce “experimental” free energy (referred to as PMF_{PDB}). This is compared to theoretical free

energy obtained from the simulations. An analysis of the association links the theoretical and “experimental” free energies.

The second part of this thesis is focused exploring a CG methodology that follows the third front of CG methods explained by Nielsen et al.²⁰ In other words, the parameterization of an effective CG interaction potential from a free energy.

Chapter 6 is at the methodological heart of this thesis, where an account of rigid body dynamics is provided. A description of the Hamiltonian, Euler equations of motion, another integration scheme and the Gay-Berne interaction potential are presented. This lays the theoretical framework for the development of a CG (or rigid body) dynamics code. References for deriving the derivation of forces and torques are also provided this chapter.

Chapter 7 is the parameterization methods chapter, which is concerned with obtaining parameters for the Gay-Berne potential from the benzene dimer vacuum free energy. An in depth description of the Gay-Berne potential, its variables and parameters are discussed. An explanation of the optimization scheme used, namely the Genetic Algorithm, is briefly provided (with further reading being cited in the chapter).

Finally, in Chapter 8, the thesis reaches its conclusion with the parameterization of a coarse-grained benzene particle, followed by the CG dynamics simulation of liquid benzene in a cubic box containing 500 CG benzene particles. The analysis of the liquid structure (using distribution functions) and dynamics (i.e., the diffusion constant) are investigated and compared with the all atom MD simulations and experiments.

1.9 References

1. Israelachvili, J. N., *Intermolecular and Surface Forces*. 3rd ed.; Elsevier: New York, 2011.
2. Autumn, K., *et al.*, Adhesive Force of a Single Gecko Foot-Hair. *Nature* **2000**, 405 (6787), 681.
3. Thompson, D. A. W., *Aristotle, Histor Ia Animalium, Trans.* Claredon: Oxford, 1918.
4. Stork, N. E., Experimental Analysis of Adhesion of Chrysolina Polita (Chrysomelidae: Coleoptera) on a Variety of Surfaces. *J. Exp. Biol.* **1980**, 88 (1), 91.
5. Autumn, K., *et al.*, Evidence for Van Der Waals Adhesion in Gecko Setae. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, 99 (19), 12252.
6. Hsu, P. Y., *et al.*, Direct Evidence of Phospholipids in Gecko Footprints and Spatula-Substrate Contact Interface Detected Using Surface-Sensitive Spectroscopy. *J. Roy. Soc. Interface* **2011**, 9 (69), 657.
7. Chen, B.; Gao, H., An Alternative Explanation of the Effect of Humidity in Gecko Adhesion: Stiffness Reduction Enhances Adhesion on a Rough Surface. *International Journal of Applied Mechanics (IJAM)* **2010**, 2 (1), 1.
8. Puthoff, J. B., *et al.*, Changes in Materials Properties Explain the Effects of Humidity on Gecko Adhesion. *J. Exp. Biol.* **2010**, 213 (21), 3699.
9. Prowse, M. S., *et al.*, Effects of Humidity on the Mechanical Properties of Gecko Setae. *Acta Biomaterialia* **2011**, 7 (2), 733.
10. Dill, K. A., Dominant Forces in Protein Folding. *Biochemistry* **1990**, 29 (31), 7133.
11. Montaron, L. P., Head and Front Legs of a Gecko. In http://en.wikipedia.org/wiki/File:Gecko_Leaftail_1.jpg, Leaftail1.jpg, G., Ed. Wikipedia: 2006.
12. Ben-Naim, A., Solvent Effects on Protein Association and Protein Folding. *Biopolymers* **1990**, 29 (3), 567.
13. Service, R. F., Problem Solved* (*Sort of). *Science* **2008**, 321, 784.
14. Kubovy, M.; Van Den Berg, M., The Whole Is Equal to the Sum of Its Parts: A Probabilistic Model of Grouping by Proximity and Similarity in Regular Patterns. *Psychological Review* **2008**, 115 (1), 131.
15. Rowlinson, J., *Cohesion: A Scientific History of Intermolecular Forces*. Cambridge University Press: Cambridge, 2002.
16. Nobelprize.Org J. D. Van Der Waals -- Biography. http://www.nobelprize.org/nobel_prizes/physics/laureates/1910/waals-bio.html (accessed 15 February 2012).
17. Lewars, E., *Computational Chemistry: Introduction to the Theory and Applications of Molecular and Quantum Mechanics*. Kluwer Academic Publishers: 2004; Vol. 2nd.
18. Leach, A. R., *Molecular Modeling*. 2nd ed.; Longman: 2001.
19. Allen, M. P.; Tildesly, D. J., Computer Simulations of Liquids. *Oxford University Press* **1987**, London, Chapter 2.
20. Nielsen, S. O., *et al.*, Coarse Grain Models and the Computer Simulation of Soft Materials. *J. Phys.: Condens. Matter* **2004**, 16 (15), R481.
21. Orsi, M., *et al.*, A Quantitative Coarse-Grain Model for Lipid Bilayers. *J. Phys. Chem. B* **2007**, 112 (3), 802.
22. Tries, V.; Paul, W., Modeling Polyethylene with the Bond Fluctuation Model. *J. Chem. Phys.* **1997**, 106 (2), 738.
23. Golubkov, P. A.; Ren, P. Y., Generalized Coarse-Grained Model Based on Point Multipole and Gay-Berne Potentials. *J. Chem. Phys.* **2006**, 125 (6), 064103.
24. Zacharopoulos, N., *et al.*, Coarse Graining Using Pretabulated Potentials: Liquid Benzene. *J. Chem. Phys.* **2005**, 122 (24), 11.

25. Darre, L., *et al.*, Another Coarse Grain Model for Aqueous Solvation: Wat Four? *J. Chem. Theory Comput.* **2010**, 6 (12), 3793.
26. Liwo, A., *et al.*, Simulation of Protein Structure and Dynamics with the Coarse-Grained Unres Force Field. In *Coarse-Graining of Condensed Phase and Biomolecular Systems*, Voth, G. A., Ed. CRC Press: New York, 2009; p 108.
27. Liwo, A., *et al.*, A United-Residue Force Field for Off-Lattice Protein-Structure Simulations .1. Functional Forms and Parameters of Long-Range Side-Chain Interaction Potentials from Protein Crystal Data. *J. Comput. Chem.* **1997**, 18 (7), 849.
28. Liwo, A., *et al.*, A United-Residue Force Field for Off-Lattice Protein-Structure Simulations .2. Parameterization of Short-Range Interactions and Determination of Weights of Energy Terms by Z-Score Optimization. *J. Comput. Chem.* **1997**, 18 (7), 874.
29. Schug, A., *et al.*, Coarse-Grained Structure-Based Simulations of Proteins and Rna. In *Coarse-Graining of Condensed Phase and Biomolecular Systems*, Voth, G. A., Ed. CRC Press: New York, 2009; p 123.
30. Fukunaga, H., *et al.*, A Coarse-Graining Procedure for Flexible Polymer Chains with Bonded and Nonbonded Interactions. *J. Chem. Phys.* **2002**, 116 (18), 8183.
31. Whitehead, L., *et al.*, Molecular Dynamics Simulation of the Hydrocarbon Region of a Biomembrane Using a Reduced Representation Model. *J. Comput. Chem.* **2001**, 22 (14), 1622.
32. Binder, K., *Monte Carlo and Molecular Dynamics Simulations in Polymer Science*. Oxford University Press: New York, 1995.
33. Paul, W., *et al.*, Structure-Property Correlation of Polymers, a Monte Carlo Approach. *Macromolecules* **1991**, 24 (23), 6332.
34. Voth, G. A., Introduction. In *Coarse-Graining of Condensed Phase and Biomolecular Systems*, Voth, G. A., Ed. CRC Press: New York, 2009; p 1.
35. Ricci, M., *et al.*, A Molecular Level Simulation of a Twisted Nematic Cell. *Faraday Discuss.* **2010**, 144, 171.
36. Stone, A. J., *The Theory of Intermolecular Forces*. Oxford University Press: New York, 1996.
37. Hinchliffe, A., *Chemical Modeling from Atoms to Liquids*. John Wiley & Sons, Ltd.: Chichester, West Sussex, 1999.
38. Mcquarrie, D. A., *Statistical Mechanics*. University Science Books: Sausalito, 1981.
39. Feynmann, R. P., *Q.E.D.: The Strange Theory of Light and Matter*. Princeton University Press: 1985.
40. Feynmann, R. P., *Quantum Electrodynamics*. Addison-Wesley: Reading, Mass., 1998.
41. Fischer, C. F., *The Hartree-Fock Method for Atoms: A Numerical Approach*. John Wiley & Sons: New York, 1977.
42. Parr, R. G.; Yang, W., *Density Functional Theory of Atoms and Molecules*. Oxford University Press: Oxford, 1989.
43. Matthews, R. P. Ion Pairing in Aqueous Metal Sulfates and Platinum Group Metal Ammonium Solutions. Thesis, University of Cape Town, Cape Town, 2010.
44. Lennard-Jones, J. E., *Proc. Roy. Soc. A* **1924**, 106, 463.
45. Engkvist, O., *et al.*, Accurate Intermolecular Potentials Obtained from Molecular Wave Functions: Bridging the Gap between Quantum Chemistry and Molecular Simulations. *Chem. Rev. (Washington, DC, U. S.)* **2000**, 100 (11), 4087.
46. Laaksonen, A.; Stilbs, P., Molecular Motion and Solvation of Benzene in Water, Carbon Tetrachloride, Carbon Disulfide and Benzene. A Combined Molecular Dynamics Simulation and Nuclear Magnetic Resonance Study. *J. Chem. Phys.* **1998**, 108 (2), 455.
47. Buchete, N. V., *et al.*, Orientation-Dependent Coarse-Grained Potentials Derived by Statistical Analysis of Molecular Structural Databases. *Polymer* **2004**, 45 (2), 144513.
48. Bartels, C.; Karplus, M., Multidimensional Adaptive Umbrella Sampling: Applications to Main Chain and Side Chain Peptide Conformations. *J. Comput. Chem.* **1997**, 18 (12), 1450.

49. Chowdhuri, S., *et al.*, Dynamical Properties of the Soft Sticky Dipole-Quadrupole-Octupole Water Model: A Molecular Dynamics Study. *J. Chem. Phys.* **2006**, 125 (14), 144513.
50. 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.
51. Miyazawa, S.; Jernigan, R. L., How Effective for Fold Recognition Is a Potential of Mean Force That Includes Relative Orientations between Interacting Residues in Proteins? *J. Chem. Phys.* **2005**, 122 (2), 024901.
52. Mukherjee, A., *et al.*, Orientation-Dependent Potential of Mean Force for Protein Folding. *J. Chem. Phys.* **2005**, 123 (1), 014901.
53. Paramonov, L.; Yaliraki, S. N., The Directional Contact Distance of Two Ellipsoids: Coarse-Grained Potentials for Anisotropic Interactions. *J. Chem. Phys.* **2005**, 123 (19).
54. Wallqvist, A.; Berne, B. J., Effective Potentials for Liquid Water Using Polarizable and Nonpolarizable Models. *J. Phys. Chem.* **1993**, 97 (51), 13841.
55. Zhang, C., *et al.*, An Accurate, Residue-Level, Pair Potential of Mean Force for Folding and Binding Based on the Distance-Scaled, Ideal-Gas Reference State. *Protein Sci.* **2004**, 13 (2), 400.
56. Buchete, N. V., *et al.*, Anisotropic Coarse-Grained Statistical Potentials Improve the Ability to Identify Nativelike Protein Structures. *J. Chem. Phys.* **2003**, 118 (16), 7658.
57. Skolnick, J., In Quest of an Empirical Potential for Protein Structure Prediction. *Curr. Opin. Struct. Biol.* **2006**, 16 (2), 166.
58. Gay, J. G.; Berne, B. J., Modification of the Overlap Potential to Mimic a Linear Site-Site Potential. *J. Chem. Phys.* **1981**, 74 (6), 3316.
59. Levinthal, C. In *How to Fold Graciously*, Mossbauer Spectroscopy in Biological Systems: Proceedings of a meeting held at Allerton House, Monticello, Illinois., Allerton House, Monticello, Illinois, 1969; Debrunner, J. T. P.; Munck, E., Eds. University of Illinois Press: Allerton House, Monticello, Illinois, 1969; pp 22.
60. Levinthal, C. In *How to Fold Graciously*, Mossbauer Spectroscopy in Biological Systems Allerton House, Monticello, Illinois, Allerton House, Monticello, Illinois, 1969; pp 22.
61. Zwanzig, R., *et al.*, Levinthals Paradox. *Proc. Natl. Acad. Sci. U. S. A.* **1992**, 89 (1), 20.
62. Dill, K. A.; Chan, H. S., From Levinthal to Pathways to Funnels. *Nat. Struct. Mol. Biol.* **1997**, 4 (1), 10.
63. Chan, H. S.; Dill, K. A., Protein Folding in the Landscape Perspective: Chevron Plots and Non-Arrhenius Kinetics. *Proteins: Struct., Funct., Bioinf.* **1998**, 30 (1), 2.
64. Dill, K. A. The Dill Research Group. <http://dillgroup.org/energy-landscapes.html> (accessed 13 February 2012).
65. Xu, D., *et al.*, Protein Binding Versus Protein Folding: The Role of Hydrophilic Bridges in Protein Association. *J. Mol. Biol.* **1997**, 265 (1), 68.
66. Clark, P. L., Protein Folding in the Cell: Reshaping the Folding Funnel. *Trends Biochem. Sci.* **2004**, 29 (10), 527.
67. Dobson, C. M., Principles of Protein Folding, Misfolding and Aggregation. *Semin. Cell Devel. Biol.* **2004**, 15 (1), 3.
68. Roux, B., The Calculation of the Potential of Mean Force Using Computer Simulations. *Comput. Phys. Commun.* **1995**, 91 (1-3), 275.
69. Bash, P. A., *et al.*, Free Energy Calculations by Computer Simulations. *Science* **1987**, 236 (4801), 564.
70. Trzesniak, D., *et al.*, A Comparison of Methods to Compute the Potential of Mean Force. *ChemPhysChem* **2007**, 8 (1), 162.
71. Torrie, G. M.; Valleau, J. P., Monte-Carlo Free-Energy Estimates Using Non-Boltzmann Sampling - Application to Subcritical Lennard-Jones Fluid. *Chem. Phys. Lett.* **1974**, 28 (4), 578.

72. Kumar, S., *et al.*, Method for Free-Energy Calculations Using Iterative Techniques. *J. Comput. Chem.* **1996**, *17* (10), 1269.
73. Bartels, C.; Karplus, M., Probability Distribution for Complex Systems: Adaptive Umbrella Sampling of the Potential Energy. *J. Phys. Chem. B* **1998**, *102* (5), 865.
74. Kollman, P. A., Free-Energy Calculations - Applications to Chemical and Biochemical Phenomena. *Chem. Rev.* **1993**, *93* (7), 2395.
75. Zhang, C., *et al.*, An Accurate, Residue-Level, Pair Potential of Mean Force for Folding and Binding on the Distance-Scaled, Ideal-Gas Reference State. *Prot. Sci.* **2004**, *13* (2), 400.
76. Fixman, M., Classical Statistical Mechanics of Constraints: A Theorem and Application to Polymers. *Proc. Natl. Ac. Sci.* **1974**, *71* (8), 3050.
77. Allen, T. W., *et al.*, Molecular Dynamics - Potential of Mean Force Calculations as a Tool for Understanding Ion Permeation and Selectivity in Narrow Channels. *Biophys. Chem.* **2006**, *124* (3), 251.
78. Jorgensen, W. L.; Severance, D. L., Aromatic Aromatic Interactions - Free-Energy Profiles for the Benzene Dimer in Water, Chloroform, and Liquid Benzene. *J. Am. Chem. Soc.* **1990**, *112* (12), 4768.
79. Podeszwa, R., *et al.*, Potential Energy Surface for the Benzene Dimer and Perturbational Analysis of Pi-Pi Interactions. *J. Phys. Chem. A* **2006**, *110* (34), 10345.
80. Chipot, C., *et al.*, Benzene Dimer: A Good Model for Pi-Pi Interactions in Proteins? A Comparison between the Benzene and the Toluene Dimers in the Gas Phase and in an Aqueous Solution. *J. Am. Chem. Soc.* **1996**, *118* (45), 11217.
81. Barnett, C. B.; Naidoo, K. J., Free Energies from Adaptive Reaction Coordinate Forces (F.E.A.R.C.F.): An Application to Ring Puckering. *Mol. Phys.* **2009**, *107* (8-12), 1243.
82. Smith, D. E.; Haymet, A. D. J., Free-Energy, Entropy, and Internal Energy of Hydrophobic Interactions - Computer-Simulations. *J. Chem. Phys.* **1993**, *98* (8), 6445.
83. Hooft, R. W. W., *et al.*, An Adaptive Umbrella Sampling Procedure in Conformational Analysis Using Molecular Dynamics and Its Applications to Glycol. *J. Chem. Phys.* **1992**, *97* (9), 6690.
84. 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* (2), 1359.
85. Tsunekawa, N., *et al.*, A Study of Water-Water Interactions in Hydrophobic Association by a Molecular Dynamics Simulation with an Optimized Umbrella Method. *J. Chem. Phys.* **2002**, *116* (15), 6725.
86. Liu, J., *et al.*, Comparison of Tip5p Water Model and Tip4p Water Model in Cluster Nucleation Kinetics Study through Umbrella Sampling and Free Energy Perturbation. *Chem. Phys.* **2006**, *323* (2-3), 579.
87. Ming, Y., *et al.*, Free Energy Perturbation Study of Water Dimer Dissociation Kinetics. *J. Chem. Phys.* **2004**, *121* (2), 773.
88. Pauling, L., *The Nature of the Chemical Bond*. 2 ed.; Cornell University Press: New York, 1948.
89. Isaacs, E. D., *et al.*, Compton Scattering Evidence for Covalency of the Hydrogen Bond in Ice. *J. Phys. Chem. Solids* **2000**, *61* (3), 403.
90. Wallqvist, A.; Mountain, R. D., Molecular Models of Water: Derivation and Description. In *Rev. Comput. Chem.*, John Wiley & Sons, Inc.: 2007; pp 183.
91. Aswani, R.; Li, J. C., A New Approach to Pairwise Potentials for Water-Water Interactions. *J. Mol. Liq.* **2007**, *134* (1-3), 120.
92. Amadei, A., *et al.*, A First Principles Polarizable Water Model for Molecular Simulations: Application to a Water Dimer. *J. Mol. Liq.* **2002**, *101* (1-3), 181.
93. Rick, S. W., *et al.*, Dynamical Fluctuating Charge Force Fields: Application to Liquid Water. *J. Chem. Phys.* **1994**, *101* (7), 6141.

94. Jorgensen, W. L., Quantum and Statistical Mechanical Studies of Liquids. 10. Transferable Intermolecular Potential Functions for Water, Alcohols, and Ethers. Application to Liquid Water. *J. Am. Chem. Soc.* **1981**, *103* (2), 335.
95. Jorgensen, W. L., *et al.*, Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79* (2), 926.
96. 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* (20), 8910.
97. Horn, H. W., *et al.*, Development of an Improved Four-Site Water Model for Biomolecular Simulations: Tip4p-Ew. *J. Chem. Phys.* **2004**, *120* (20), 9665.
98. Lamoureux, G., *et al.*, A Simple Polarizable Model of Water Based on Classical Drude Oscillators. *J. Chem. Phys.* **2003**, *119* (10), 5185.
99. Jorgensen, W. L.; Tirado-Rives, J., Potential Energy Functions for Atomic-Level Simulations of Water and Organic and Biomolecular Systems. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102* (19), 6665.
100. Pratt, L. R., Molecular Theory of Hydrophobic Effects: "She Is Too Mean to Have Her Name Repeated." *Annu. Rev. Phys. Chem.* **2002**, *53* (1), 409.
101. Rao, B. G.; Singh, U. C., Hydrophobic Hydration: A Free Energy Perturbation Study. *J. Am. Chem. Soc.* **1989**, *111* (9), 3125.
102. Dzubiella, J., How Interface Geometry Dictates Water's Thermodynamic Signature in Hydrophobic Association. *J. Stat. Phys.* **2011**, *145* (2), 227.
103. Setny, P., *et al.*, How Can Hydrophobic Association Be Enthalpy Driven? *J. Chem. Theory Comput.* **2010**, *6* (9), 2866.
104. Snyder, P. W., *et al.*, Mechanism of the Hydrophobic Effect in the Biomolecular Recognition of Arylsulfonamides by Carbonic Anhydrase. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108* (44), 17889.
105. Hummer, G., *et al.*, Water Conduction through the Hydrophobic Channel of a Carbon Nanotube. *Nature* **2001**, *414* (6860), 188.
106. Choudhury, N.; Pettitt, B. M., The Dewetting Transition and the Hydrophobic Effect. *J. Am. Chem. Soc.* **2007**, *129* (15), 4847.
107. Chandler, D., Interfaces and the Driving Force of Hydrophobic Assembly. *Nature* **2005**, *437* (7059), 640.
108. Spolar, R. S., *et al.*, Hydrophobic Effect in Protein Folding and Other Noncovalent Processes Involving Proteins. *Proc. Natl. Acad. Sci. U. S. A.* **1989**, *86* (21), 8382.
109. Nicholls, A., *et al.*, Protein Folding and Association: Insights from the Interfacial and Thermodynamic Properties of Hydrocarbons. *Proteins: Struct., Funct., Bioinf.* **1991**, *11* (4), 281.
110. Tsai, C.-J., *et al.*, Studies of Protein-Protein Interfaces: A Statistical Analysis of the Hydrophobic Effect. *Prot. Sci.* **1997**, *6* (1), 53.
111. Sun, Y.; Kollman, P., Are There Water-Bridge-Induced Hydrophilic Interactions? *J. Phys. Chem.* **1996**, *100* (16), 6760.
112. Luscombe, N. M., *et al.*, Amino Acid-Base Interactions: A Three-Dimensional Analysis of Protein-DNA Interactions at an Atomic Level. *Nucleic Acids Res.* **2001**, *29* (13), 2860.
113. Schliwa, M.; Woehlke, G., Molecular Motors. *Nature* **2003**, *422* (6933), 759.
114. Zhou, Y., *et al.*, Chemistry of Ion Coordination and Hydration Revealed by a K⁺ Channel-Fab Complex at 2.0[Thinsp][Angst] Resolution. *Nature* **2001**, *414* (6859), 43.
115. Suezawa, H., *et al.*, The Ch[Small Pi] Interaction as an Important Factor in the Crystal Packing and in Determining the Structure of Clathrates. *Journal of the Chemical Society, Perkin Transactions 2* **2001**, (11), 2053.
116. Strickland-Constable, R. F., *Kinetics and Mechanism of Crystallization from the Fluid Phase and of the Condensation and Evaporation of Liquids*. Academic Press London: 1968.

117. Maayan, G., Conformational Control in Metallofoldamers: Design, Synthesis and Structural Properties. *Eur. J. Org. Chem.* **2009**, 2009 (33), 5699.
118. Collings, P. J., *Nature's Delicate Phase of Matter*. Princetone University Press, Princeton: 1990.
119. Burley, S. K.; Petsko, G. A., Aromatic-Aromatic Interaction - a Mechanism of Protein-Structure Stabilization. *Science* **1985**, 229 (4708), 23.
120. Mcgaughey, G. B., *et al.*, Pi-Stacking Interactions - Alive and Well in Proteins. *J. Biol. Chem.* **1998**, 273 (25), 15458.
121. Samanta, U., *et al.*, Quantifying the Accessible Surface Area of Protein Residues in Their Local Environment. *Protein Eng.* **2002**, 15 (8), 659.
122. Collins, M. D., *et al.*, Cooperative Water Filling of a Nonpolar Protein Cavity Observed by High-Pressure Crystallography and Simulation. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, 102 (46), 16668.
123. Zhou, R., *et al.*, Hydrophobic Collapse in Multidomain Protein Folding. *Science* **2004**, 305 (5690), 1605.
124. Young, T., *et al.*, Motifs for Molecular Recognition Exploiting Hydrophobic Enclosure in Protein-Ligand Binding. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, 104 (3), 808.
125. Podeszwa, R., *et al.*, Potential Energy Surface for the Benzene Dimer and Perturbational Analysis of Pi-Pi Interactions. *J. Phys. Chem. A.* **2006**, 110, 10345.
126. Gräfenstein, J.; Cremer, D., An Efficient Algorithm for the Density-Functional Theory Treatment of Dispersion Interactions. *J. Chem. Phys.* **2009**, 130 (12), 124105.
127. Distasio, R. A., *et al.*, On the T-Shaped Structures of the Benzene Dimer. *Chem. Phys. Lett.* **2007**, 437 (4-6), 277.
128. Dinadayalane, T. C.; Leszczynski, J., Geometries and Stabilities of Various Configurations of Benzene Dimer: Details of Novel V-Shaped Structure Revealed. *Struct. Chem.* **2009**, 20 (1), 11.
129. Headen, T. F., *et al.*, Structure of Pi-Pi Interactions in Aromatic Liquids. *J. Am. Chem. Soc.* **2010**, 132 (16), 5735.
130. Pitonak, M., *et al.*, Benzene Dimer: High-Level Wave Function and Density Functional Theory Calculations. *J. Chem. Theory Comput.* **2008**, 4 (11), 1829.
131. Dzubiella, J.; Hansen, J. P., Competition of Hydrophobic and Coulombic Interactions between Nanosized Solutes. *J. Chem. Phys.* **2004**, 121 (11), 5514.
132. Anslyn, E. V.; Dougherty, D. A., *Modern Physical Organic Chemistry*. University Science Books: 2006.
133. Khossravi, D.; Connors, K. A., Solvent Effects on Chemical Processes, I: Solubility of Aromatic and Heterocyclic Compounds in Binary Aqueous—Organic Solvents. *J. Pharm. Sci.* **1992**, 81 (4), 371.
134. Benjamin, I., Structure and Dynamics of Hydrated Ions in a Water-Immiscible Organic Solvent. *J. Phys. Chem. B* **2008**, 112 (49), 15801.
135. Rhodes, T. A., *et al.*, Effect of Solvent-Solute and Solute-Solute Interactions on the Rate of a Michael Addition in Supercritical Fluoroform and Ethane. *J. Phys. Chem.* **1995**, 99 (24), 9903.
136. Kim, Y., *et al.*, Steric Effects and Solvent Effects on Sn2 Reactions. *J. Phys. Chem. A* **2009**, 113 (32), 9109.
137. Ananikov, V. P., *et al.*, Catalytic Triple Bond Activation and Vinyl-Vinyl Reductive Coupling by Pt(IV) Complexes. A Density Functional Study. *Organometallics* **2001**, 20 (8), 1652.
138. Litwinienko, G., *et al.*, The Frequently Overlooked Importance of Solvent in Free Radical Syntheses. *Chem. Soc. Rev.* **40** (5), 2157.
139. Kutt, A., *et al.*, Pentakis(Trifluoromethyl)Phenyl, a Sterically Crowded and Electron-Withdrawing Group: Synthesis and Acidity of Pentakis(Trifluoromethyl)Benzene, -Toluene, -Phenol, and -Aniline. *J. Org. Chem.* **2008**, 73 (7), 2607.

140. Richards, F. M.; Lim, W. A., An Analysis of Packing in Protein Folding Problem. *Q. Rev. Biophys.* **1994**, *26* (4), 423.
141. Escher, M. C., Liberation. Oldbourne Book Co. Ltd. 2 Partmen Street: London, 1961.
142. Allen, M. P.; Tildesly, D. J., *Computer Simulation of Liquids*. Oxford Univeristy Press: Oxford, 1987.
143. Mie, G., Zur Kinetischen Theorie Der Einatomigen Körper. *Annalen der Physik* **1903**, *316* (8), 657.
144. Jones, J. E., On the Determination of Molecular Fields. Ii. From the Equation of State of a Gas. *Proceedings of the Royal Society of London. Series A* **1924**, *106* (738), 463.
145. Corner, J., The Second Virial Coefficient of a Gas of Nonspherical Molecules. *Proc. R. Soc. Lond. A* **1948**, *A192*, 275.
146. Berne, B. J.; Pechukas, P., Gaussian Model Potentials for Molecular Interactions. *J. Chem. Phys.* **1972**, *56*, 4213.
147. Sommer, K., Correlation between Primary Chemical Structure and Property Phenomena in Polycondensates. *Adv. Mater. (Weinheim, Ger.)* **1991**, *3* (12), 590.
148. Fukunaga, H., *et al.*, A Coarse-Graining Procedure for Flexible Polymer Chains with Bonded and Nonbonded Interactions. *J. Chem. Phys.* **2002**, *116* (18), 8183.
149. Groot, R. D.; Rabone, K. L., Mesoscopic Simulation of Cell Membrane Damage, Morphology Change and Rupture by Nonionic Surfactants. *Biophys. J.* **2001**, *81* (2), 725.
150. Cagin, T., *et al.*, Multiscale Modeling and Simulation Methods with Applications to Dendritic Polymers. *Computational and Theoretical Polymer Science* **2001**, *11* (5), 345.
151. Kolinski, A.; Skolnick, J., Discretized Model of Proteins. I. Monte Carlo Study of Cooperativity in Homopolypeptides. *J. Chem. Phys.* **1992**, *97* (12), 9412.
152. Cagin, T., *et al.*, Multiscale Modeling and Simulation Methods with Applications to Dendritic Polymers. *Computational and Theoretical Polymer Science* **2001**, *11* (5), 345.
153. Goddard, W. A., *et al.*, Strategies for Multiscale Modeling and Simulation of Organic Materials: Polymers and Biopolymers. *Computational and Theoretical Polymer Science* **2001**, *11* (5), 329.
154. Muralidharan, K., *et al.*, A Concurrent Multiscale Finite Difference Time Domain/Molecular Dynamics Method for Bridging an Elastic Continuum to an Atomic System. *Modelling and Simulation in Materials Science and Engineering* **2003**, *11* (4), 487.
155. Rafii-Tabar, H.; Chirazi, A., Multi-Scale Computational Modelling of Solidification Phenomena. *Physics Reports-Review Section of Physics Letters* **2002**, *365* (3), 145.
156. Elliott, J. A., Novel Approaches to Multiscale Modelling in Materials Science. *Int. Mater. Rev.* *56* (4), 207.
157. Matysiak, S.; Clementi, C., Characterization of Protein-Folding Landscapes by Coarse-Grained Models Incorporating Experimental Data. In *Coarse-Graining of Condensed Phase and Biomolecular Systems*, Voth, G. A., Ed. CRC Press: New York, 2009; p 157.
158. Haliloglu, T.; Bahar, I., Coarse-Grained Simulations of Conformational Dynamics of Proteins: Applications to Apomyoglobin. *Proteins: Struct., Funct., Bioinf.* **1998**, *31*, 271.
159. Akkermans, R. L. C.; Briels, W. J., Coarse-Grained Interactions in Polymer Melts: A Variational Approach. *J. Chem. Phys.* **2001**, *115* (13), 6210.
160. Berardi, R., *et al.*, Monte-Carlo Investigations of a Gay-Berne Liquid-Crystal. *Journal of the Chemical Society-Faraday Transactions* **1993**, *89* (22), 4069.
161. Berardi, R., *et al.*, Phase Diagram of the Uniaxial and Biaxial Soft-Core Gay-Berne Model. *J. Chem. Phys.* **2011**, *135* (13).
162. Care, C. M.; Cleaver, D. J., Computer Simulation of Liquid Crystals. *Reports on Progress in Physics* **2005**, *68* (11), 2665.
163. Demiguel, E., *et al.*, Liquid-Crystal Phase-Diagram of the Gay-Berne Fluid. *Mol. Phys.* **1991**, *74* (2), 405.

164. Emsley, J. W., *et al.*, Computer-Simulation Studies of the Dependence on Density of the Orientational Order in Nematic Liquid-Crystals. *Liq. Cryst.* **1992**, *11* (4), 519.
165. Luckhurst, G. R., The Gay-Berne Mesogen: A Paradigm Shift? *Liq. Cryst.* **2006**, *33* (11-12), 1389.
166. Luckhurst, G. R., *et al.*, Computer-Simulation Studies of Anisotropic Systems .19. Mesophases Formed by the Gay-Berne Model Mesogen. *Liq. Cryst.* **1990**, *8* (4), 451.
167. Singh, R. C., Temperature-Dependent Study of Isotropic-Nematic Transition for a Gay-Berne Fluid Using Density-Functional Theory. *J. Phys.: Condens. Matter* **2007**, *19* (37).
168. Wilson, M. R., Progress in Computer Simulations of Liquid Crystals. *Int. Rev. Phys. Chem.* **2005**, *24* (3-4), 421.
169. Olson, W. K., *et al.*, Insights into the Sequence-Dependent Macromolecular Properties of DNA from Base-Pair Level Modeling. In *Coarse-Graining of Condensed Phase and Biomolecular Systems*, Voth, G. A., Ed. CRC Press: New York, 2009; p 205.
170. Marrink, S. J.; Mark, A. E., The Mechanism of Vesicle Fusion as Revealed by Molecular Dynamics Simulations. *J. Am. Chem. Soc.* **2003**, *125* (37), 11144.
171. Katsoulakis, M. A.; Vlachos, D. G., Coarse-Grained Stochastic Processes and Kinetic Monte Carlo Simulators for the Diffusion of Interacting Particles. *J. Chem. Phys.* **2003**, *119* (18), 9412.
172. Louis, A. A., Beware of Density Dependent Pair Potentials. *J. Phys.: Condens. Matter* **2002**, *14* (40), 9187.
173. Strumpfer, J.; Naidoo, K. J., Computing Free Energy Hypersurfaces for Anisotropic Intermolecular Associations. *J. Comput. Chem.* **2010**, *31* (2), 308.
174. Lyubartsev, A. P.; Laaksonen, A., Calculation of Effective Interaction Potentials from Radial-Distribution Functions - a Reverse Monte-Carlo Approach. *Phys. Rev. E* **1995**, *52* (4), 3730.
175. Toth, G.; Baranyai, A., Molecular Dynamics Analog of the Reverse Monte Carlo Method. *J. Chem. Phys.* **2001**, *114* (5), 2027.

CHAPTER 2

MOLECULAR SIMULATION AND STATISTICAL ANALYSIS

In this chapter, a description of the classical molecular mechanics Hamiltonian is given. A brief overview is provided of the governing equations and ideas that are used in standard molecular dynamics (MD) simulations. The Born-Oppenheimer approximation forms the basis of the validity of molecular mechanics, in that it allows the Hamiltonian to be expressed as a function of nuclear coordinates only.¹⁻³ This means that the system can be treated classically, using classical mechanics rather than quantum mechanics, with a high degree of accuracy.

Methods have also been developed to improve the efficiency and parallel scalability of MD simulations that include: periodicity, introduced to negate artificial edge/surface effects; various integration schemas, which try to reduce the truncation error of the Taylor expansion about $\mathbf{r}(t)$; potential truncation that serves to reduce the computational time of the simulation; and constraints that are used to maintain reasonable bond lengths and remove trivial high frequency motions to ensure stability of the simulation even when large timesteps are used. In this thesis, a modified version of the *CHARMM*⁴ program is used exclusively for running MD and free energy simulations and, therefore, only the *CHARMM* classical Hamiltonian (or empirical force-field) is discussed here.

Finally, an analysis of the trajectories produced by a MD simulation has to be performed. Since most of this work is done in the condensed liquid phase, we are mainly concerned about how the solvent is affecting the system. By making use of statistical analysis tools, we are able to investigate the solvent structure around molecules or liquid structure. The solvent structure is studied using distribution functions such as pair distribution functions (PDFs) and spatial or solvent distribution functions (SDFs). Transport properties such as diffusion can also be determined from MD trajectories by using correlation functions.

2.1 Molecular Dynamics

Classical MD simulations are based primarily on the movements of nuclei, whereas quantum mechanical calculations incorporate the motions of electrons as well. This reduction in the complexity of the calculation makes MD an attractive and computationally viable technique for simulating large biological systems into the nanosecond timescale.

A MD simulation describing the motions of the atoms are run using Hamiltonian dynamics, and can generally be written as:

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

for N generalized coordinates and momenta, $\mathbf{q}^N$ and $\mathbf{p}^N$, where $V(\mathbf{q}^N, \mathbf{p}^N, t)$ 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 mechanics, the potential energy is a function only of the coordinates, $\mathbf{r}^N$, and, hence, the Hamiltonian becomes

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

where $\mathbf{r}_i$ is the 3D coordinates of atom i , and $\mathbf{p}_i$ the conjugate 3D momenta. The equations of motion for Hamiltonian dynamics is given by the following two differential equations:

$$\dot{\mathbf{p}}_i = \frac{\partial H}{\partial \mathbf{r}_i} = -\nabla_{\mathbf{r}_i} V(\mathbf{r}_i) = \mathbf{F}_i \quad 2.3.1$$

$$\dot{\mathbf{r}} = \frac{\partial H}{\partial \mathbf{p}_i} = \frac{\mathbf{p}_i}{m_i} \quad 2.3.2$$

By integrating these equations we can obtain the time evolution, or dynamics, of the system and calculate various macroscopic properties.

2.1.1 Molecular mechanics – The potential energy and forces

At the core of MD simulations is molecular mechanics (or empirical force field calculations as it is often referred to) and it is what governs all the interactions. Molecular mechanics is made possible by introducing several assumptions and approximations. Firstly, the potential energy can be written as a function of the nuclear

coordinates only (Born-Oppenheimer approximation), where atoms are modeled as balls and bonds as springs obeying Hooke's Law;² and, secondly, that the partial charge is representative of the electronic density on an atom.

The potential energy function, V , and the forces, $\mathbf{F} = -\nabla_{\mathbf{r}} V = -dV/d\mathbf{r}$, of any molecular mechanics force field determines the interactions between atoms that make up the molecules in the simulation. A molecular mechanics force field typically consists of intra- (bonded) and intermolecular (non-bonded) interactions. A typical empirical potential energy function, like that used in CHARMM,⁴ has the following form:

$$E = E_{bonds} + E_{angles} + E_{impropers} + E_{Urey-Bradleys} \quad (\text{Vibrational Modes}) \\ + E_{dihedrals} \quad (\text{Torsional Flexibility}) \quad , \\ + E_{Coulomb} + E_{Lennard-Jones} \quad (\text{Intermolecular Interactions})$$

and the subsequent forces for each of the terms are expressed as:

$$\mathbf{F} = - \left[\begin{array}{l} \nabla E_{bonds} + \nabla E_{angles} + \nabla E_{impropers} + \nabla E_{Urey-Bradleys} \\ + \nabla E_{dihedrals} \\ + \nabla E_{Coulomb} + \nabla E_{Lennard-Jones} \end{array} \right]$$

More specifically, the force field can conceptually be divided into three parts: vibrational modes, torsional flexibility and intermolecular interactions. The first two constitute the bonded interactions, while the last is the non-bonded interactions. A brief discussion of each term is provided in the following sections.

BONDED (INTRAMOLECULAR) INTERACTIONS

The bonded, or intramolecular, interactions are the interactions that the molecule experiences within itself, which includes the bonds, angles, dihedral angles, etc. Equation 2.4 is a typical intramolecular part of most molecular mechanics potential energy functions.

$$E_{bonded} = \sum_{bonds} k_b (b - b_o)^2 + \sum_{angles} k_\theta (\theta - \theta_o)^2 + \sum_{impropers} k_\varphi (\varphi - \varphi_o)^2 + \sum_{Urey-Bradleys} k_{UB} (b_{1,3} - b_{1,3,o})^2 \\ + \sum_{dihedrals} k_\phi [1 + \cos(n\phi - \delta)] \quad . \quad 2.4$$

Each of the terms in the above equation will be described in slightly more detail and the empirical formula for the forces are also given.

The vibrational modes, described by a harmonic function, of the intramolecular part of the potential energy is described using a ball and spring, where the stiffness of the spring is determined by a force constant. The torsional flexibility, described by a Fourier

expansion, describes the rotational motion around a central bond where the force constant governs the resistance to rotation.

- The first term, of Equation 2.4, accounts for the potential energy of the bonds in the system (Figure 2.1). The two parameters, k_b and b_o , are the force constant and equilibrium bond length, respectively.

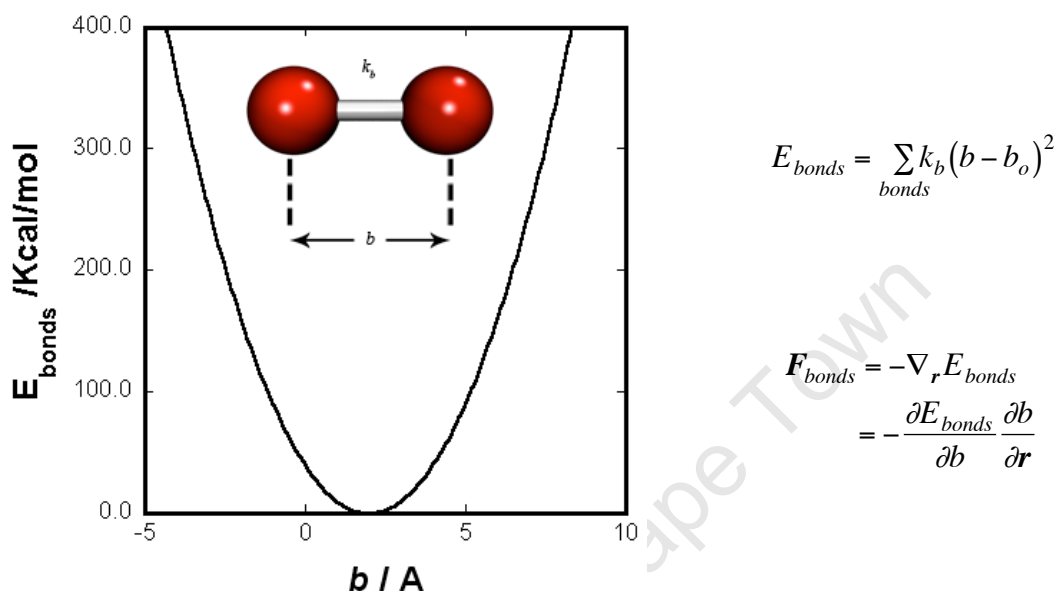


Figure 2.1: Graphical representation of the harmonic bond potential energy, forces are given in right panel.

- The second term accounts for the potential energy of the angles in the system (Figure 2.2). The function behaves similarly to the bonds term: k_θ is the force constant for the vibrational motion of the angle and θ_o is the equilibrium angle.

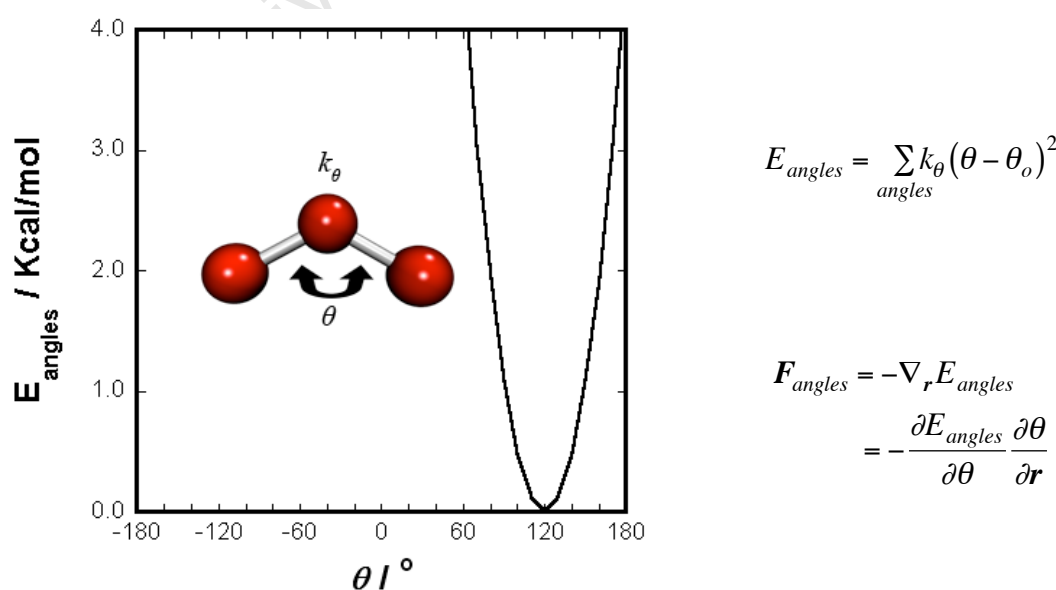
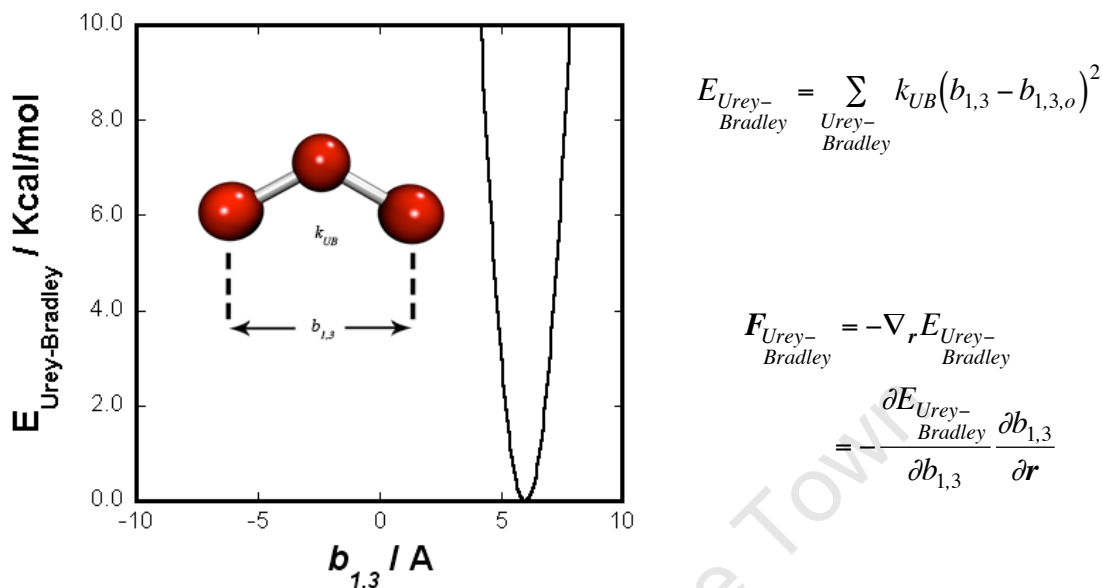


Figure 2.2: Graphical representation of the harmonic angle potential energy, forces are given in right panel.

- The third term accounts for potential of the improper dihedral angles in the system (Figure 2.3). In analogy with the first and second terms, k_φ is the force constant and φ_o is equilibrium improper dihedral angle.



- Figure 2.3:** Graphical representation of the harmonic Urey-Bradley potential energy, and forces.
- The fourth term takes into account the 1-3 non-bonded interactions based on their distance apart (Figure 2.4). It is a harmonic term with k_{UB} being the force constant and $b_{1,3,o}$ the equilibrium distance between atoms separated by two bonds.

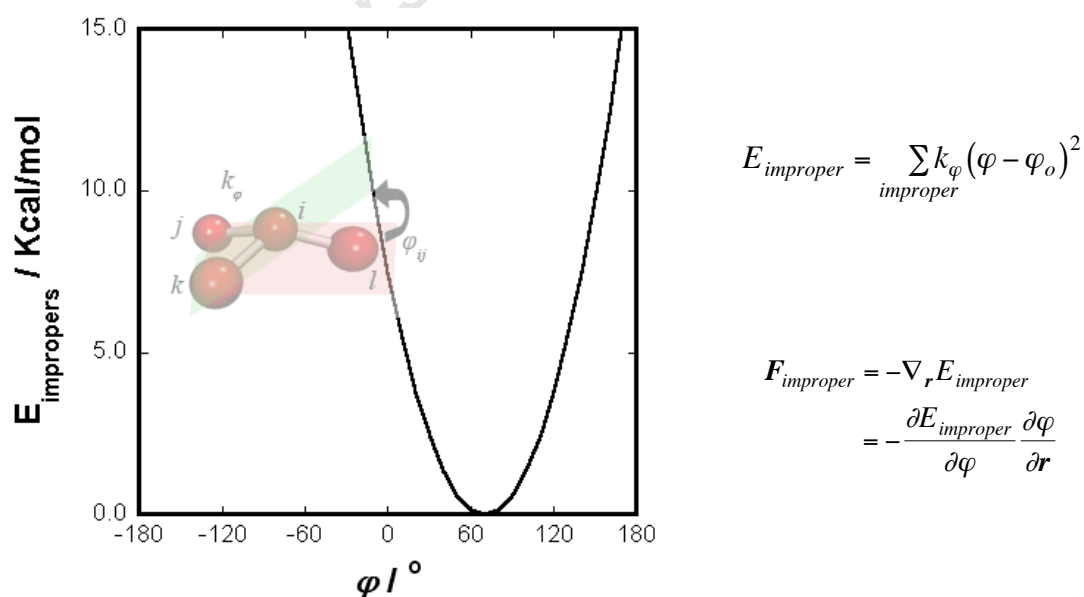
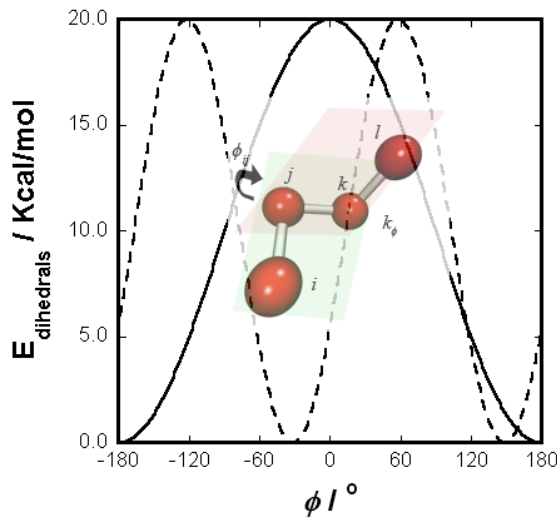


Figure 2.4: A graphical representation of the harmonic relationship used to describe the improper dihedral angles potential energy, forces given in right panel.

- Lastly, the force field also has to taken into account the degree of rotation that can occur around a central bond, $j-k$, made up of the four adjacent atoms (i, j, k, l) defining a dihedral angle (Figure 2.3). The dihedral of an eclipsed conformation is 0 and 360 (the bonds $i-j$ and $k-l$ are parallel), while the anti-parallel staggered dihedral is -180 and 180 . This sinusoidal behaviour allows the potential energy function and forces to be expressed as a Fourier expansion of the dihedral angle, ϕ , manipulated by the force constant (k_ϕ), multiplicity ($n=1,2,3,4$) and phase shift (δ) to correspond to empirically derived curves.



$$E_{dihedrals} = \sum_{dihedrals} k_\phi [1 + \cos(n\phi - \delta)]$$

$$\begin{aligned} \mathbf{F}_{dihedrals} &= -\nabla_{\mathbf{r}} E_{dihedrals} \\ &= -\left(\frac{\partial E_{dihedrals}}{\partial \cos \phi} \right) \frac{\partial \cos \phi}{\partial \mathbf{r}} \end{aligned}$$

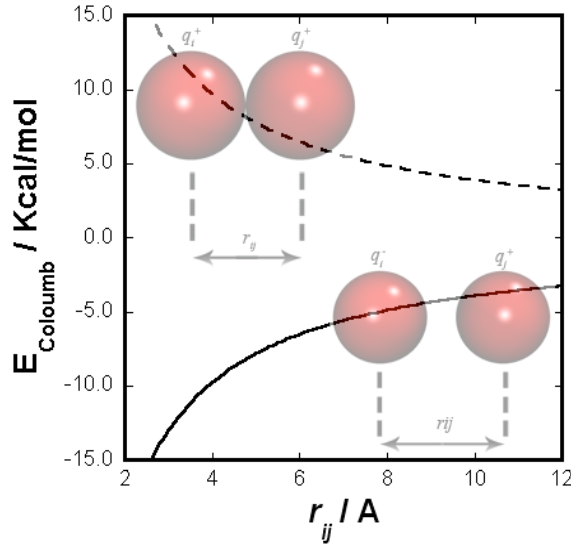
Figure 2.5: The schematic of the dihedral angle and the Fourier expansion of the dihedral angle to determine the potential energy function and forces.

NON-BONDED (INTERMOLECULAR) INTERACTIONS

The types of effective isotropic intermolecular potentials used by MD packages, such as *CHARMM*, are usually pairwise in nature and are a function of the separation distance, r_{ij} , between pairs of atoms i and j . The typical effective pairwise potential has to take into account, or mimic, the quantum mechanical interaction of the electron clouds and nuclei, between atoms, as well as the electrostatic interactions from the partial charges on the atoms.

$$E_{non-bonded} = \sum_i \sum_{j>i} \frac{q_i q_j}{4\pi\epsilon_o\epsilon_r r_{ij}} + \sum_i \sum_{j>i} 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^6 \right] \quad 2.5$$

The first term is the Coulomb potential and accounts for the electrostatic interactions between the partial charges on the atoms (Figure 2.6). The parameters are q_i and q_j are the partial charges on the atom i and j respectively, ϵ_o is the dielectric permittivity of the free space and ϵ_r is dielectric constant of the medium.



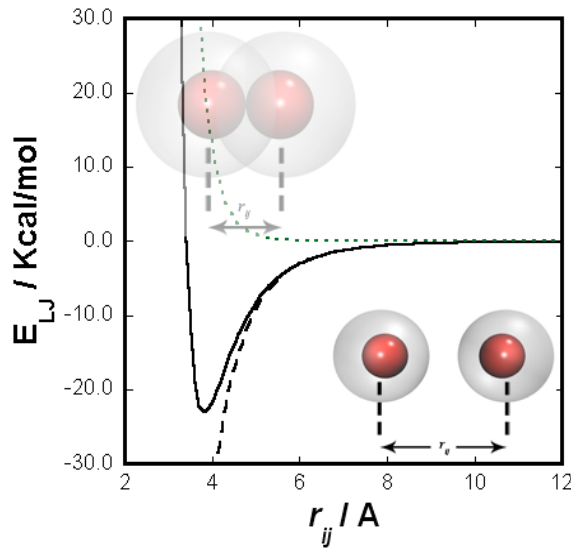
$$E_{Col} = \sum_i \sum_{j>i} \frac{q_i q_j}{4\pi\epsilon_o\epsilon_r r_{ij}}$$

$$\mathbf{F}_{Col} = -\nabla_{\mathbf{r}_{ij}} E_{Col} = -\frac{\partial E_{Col}}{\partial \mathbf{r}_{ij}};$$

by using Newton's third law of motion,
 $\mathbf{F}_j = -\mathbf{F}_i$

Figure 2.6: A schematic of the charge-charge (Coulomb) electrostatic interaction. The solid represents the attractive, while the dashed line the repulsive interactions. Here, $F_i = -F_j$, due to Newton's third law.

The second is a dispersion-like term called the Lennard-Jones potential (Figure 2.7). It has a strong short-range repulsion (r^{-12} term), which takes into account the Pauli Exclusion Principle, and weak long-range attraction (r^{-6} term) that accounts for the quantum mechanical attraction between electrons on one atom and the nucleus of another. The parameters are ϵ_{ij} , the interaction strength between atoms i and j , and σ_{ij} , (in the *CHARMM* force field it is given as $R_{min,ij} = \sigma_{ij}/2$) the range parameter, or the separation distance at the minimum of the potential. The potential and force resulting from the Lennard-Jones non-bonded interaction are described below in Figure 2.7.



$$E_{LJ} = \sum_i \sum_{j>i} 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^6 \right]$$

$$\mathbf{F}_{LJ} = -\nabla_{\mathbf{r}_{ij}} E_{LJ} = -\frac{\partial E_{LJ}}{\partial \mathbf{r}_{ij}},$$

and similar as above, $\mathbf{F}_j = -\mathbf{F}_i$.

Figure 2.7: A graphical representation of the Lennard-Jones (dispersion) interaction potential energy and forces. The solid line is the Lennard-Jones potential is composed of attractive (dashed line) and repulsive (dotted line) parts. The force on atom j due to i is given by $F_i = -F_j$ (Newton's third law).

2.1.2 Long-range interactions

The treatment of long-range interactions is of fundamental importance not only for charged systems, but also for neutral molecules with terms for the multipole expansion. The *Ewald summation*, first devised in 1921, is routinely used and is one of the best ways to incorporate all the effects of long-range interactions in a simulation.⁵ Use of the Ewald method has increased with the development of several efficient, $O(N\log N)$, algorithms that calculate the reciprocal space term. The *Smooth Particle Mesh Ewald (or PME)*^{6, 7} method has been implemented in CHARMM for its high speed, accuracy and application to constant pressure systems. The algorithm assigns point charges to grid points, which are interpolated using B-splines and calculates the reciprocal space using *3D fast Fourier transform (3-DFFT)*.⁷ Alternative methods are the *fast Fourier Poisson*⁸ and *Particle-Particle-Mesh Ewald* algorithms.⁹

A brief description of the *Ewald summation* and total charge-charge contribution to the potential energy and force routines are described below (in terms of the solute):¹⁰

1. Interactions within the central cell (cube).
2. Interactions of the central cell with all other images.
3. Interaction of image cells with the surrounding medium.

An expression, which incorporates all of these interactions, is given by

$$U = \frac{1}{2} \sum_{|n|=0}^{\infty} \sum_{i=1}^N \sum_{j=1}^N \frac{q_i q_j}{4\pi\epsilon_o |r_{ij} + n|}. \quad 2.6$$

The summation is conditionally convergent, as it is the sum of all the negative terms in the series. In the Ewald method, conditional convergence is overcome by converting the overall summation into two or more complicated but rapidly converging sums. The initial set of partial charges is surrounded by a Gaussian distribution of equal magnitude but opposite sign (real space calculation) to which a cancelling set of distributions must be added (reciprocal space calculation).

The real space summation, expressed by Equation 2.7, generally converges rapidly and the rate at which it converges is dependent on the width of the Gaussian distributions. The parameter α should be chosen in such a way that only terms in the series are those for which $|n|=0$. Altering α can modify the functional form of the Gaussian distributions.

$$U = \frac{1}{2} \sum_{|n|=0}^{\infty} \sum_{i=0}^N \sum_{j=0}^N \frac{q_i q_j \operatorname{erfc}(\alpha |r_{ij} + n|)}{4\pi\epsilon_o |r_{ij} + n|}, \quad 2.7$$

where

$$\operatorname{erfc}(x) = \frac{2}{\sqrt{\pi}} \int_x^{\infty} \exp(-t^2) dt \quad 2.8$$

is complementary error function associated with the Ewald summation in real space.

A second series of distribution is then added, cancelling the first series (the real space summation). This second summation is performed in the reciprocal space, where the vectors k , are defined as $k = 2\pi n/L^2$.

$$U = \frac{1}{2} \sum_{k \neq 0}^{\infty} \sum_{i=1}^N \sum_{j=1}^N \frac{q_i q_j}{4\pi\epsilon_o} \frac{1}{\pi L^3} \frac{4\pi^2}{k^2} \exp\left(-\frac{k^2}{4\alpha^2}\right) \cos(k \cdot r_{ij}). \quad 2.9$$

The real space summation includes the interaction of the Gaussian with itself and has to subsequently be removed by subtracting by

$$U = \frac{\alpha}{\sqrt{\pi}} \sum_{k=1}^N \frac{q_k^2}{4\pi\epsilon_o}. \quad 2.10$$

Finally, a correction term is required to account for the medium surrounding the image cells. This correction is only applicable if the surrounding medium has a relative permittivity of 1, i.e. a vacuum,

$$U_{\text{correction}} = \frac{2\pi}{3L^3} \left| \sum_{i=1}^N \frac{q_i}{4\pi\epsilon_o} r_i \right|^2. \quad 2.11$$

The overall Ewald summation energy is calculated using Equations 2.6 – 2.11.

2.1.3 Integration schemes I

Given that the generation of an MD trajectory involves integrating a large number of differential equations, it is of paramount importance that the numerical algorithm be efficient and accurate. The most important feature of any integration algorithm is the ability to support a long integration timestep with a small numerical error, rather than being fast itself. One of the most established and commonly used methods, for advancing the positions in the system, is the *Verlet* algorithm

$$\mathbf{r}(t + \Delta t) = 2\mathbf{r}(t) - \mathbf{r}(t - \Delta t) + \mathbf{a}(t)\Delta t^2. \quad 2.12$$

It is clear that the velocities do not feature in this integration scheme, as they have been eliminated by addition of the Taylor expansion about $\mathbf{r}(t)$.

$$\mathbf{r}(t + \Delta t) = \mathbf{r}(t) + \mathbf{v}(t)\Delta t + \frac{1}{2}\mathbf{a}(t)\Delta t^2 + \frac{1}{6}\mathbf{b}(t)\Delta t^3 + O(t)\Delta t^4 + \dots \quad 2.13.1$$

$$\mathbf{r}(t - \Delta t) = \mathbf{r}(t) - \mathbf{v}(t)\Delta t + \frac{1}{2}\mathbf{a}(t)\Delta t^2 - \frac{1}{6}\mathbf{b}(t)\Delta t^3 + O(t)\Delta t^4 - \dots, \quad 2.13.2$$

where $\mathbf{r}(t)$, $\mathbf{v}(t)$, $\mathbf{a}(t)$ and $\mathbf{b}(t)$ are the position, velocity, acceleration and jerk vectors, respectively, at time t .

Using this integration scheme the velocities are not required to compute the trajectories, but are useful for estimating the kinetic energy and, hence, the total energy ($H = K + V$). The velocities are estimated in the following way:

$$\mathbf{v}(t) = \frac{\mathbf{r}(t + \Delta t) - \mathbf{r}(t - \Delta t)}{2\Delta t}. \quad 2.14$$

Equation 2.14 is correct except for the local error of δt^4 and global error of δt^2 , whereas the velocity is subjected to large global and local errors of δt^2 .¹¹

An improvement upon the Verlet algorithm is the *leapfrog* algorithm given by Equations 2.15 and 2.16.

$$\mathbf{r}(t + \Delta t) = \mathbf{r}(t) + \mathbf{v}\left(t + \frac{1}{2}\Delta t\right)\Delta t, \quad 2.15$$

$$\mathbf{v}\left(t + \frac{1}{2}\Delta t\right) = \mathbf{v}\left(t - \frac{1}{2}\Delta t\right) + \mathbf{a}(t)\Delta t. \quad 2.16$$

The stored variables are the current coordinates $\mathbf{r}(t)$ and accelerations $\mathbf{a}(t)$ together with the mid-step velocities $\mathbf{v}(t - 1/2\delta t)$. The velocity in Equation 2.16 is implemented first, and then the velocities leap over the coordinates to give the next mid-step values $\mathbf{v}(1 + 1/2\delta t)$. During this step the current velocities are calculated as:

$$\mathbf{v}(t) = \frac{1}{2}\left(\mathbf{v}\left(t + \frac{1}{2}\Delta t\right) + \mathbf{v}\left(t - \frac{1}{2}\Delta t\right)\right). \quad 2.17$$

This is required because the total energy ($H = K + V$), at time t , must be evaluated, as well as any other variables that require both the coordinates and velocities at the same time. Following this, Equation 2.15 is then used to leap the coordinates over the velocities once again and the new accelerations are calculated ready for the next step, hence, the name *leapfrog*.¹¹ This method is equivalent to the Verlet algorithm but it avoids the need

to include the large errors for the coordinates and velocities. The velocity Verlet integration scheme, which will be discussed later in Chapter 6, calculates the velocity at each timestep.

2.1.4 Boundary conditions and the minimum image convention

Proper treatment of the boundary, as well as boundary effects, plays an important role in performing accurate MD simulations. If treated correctly, these methods enable the calculation of macroscopic properties by using a relatively small number of molecules.

One of the major concerns of MD simulations is that there are many molecules that lie near or on the surface of the simulation boundaries. These molecules produce edge and surface effects that experience energy fluctuations and forces very different from the molecules found in the bulk. Upon implementation of *periodic boundary conditions*, the molecules of the system are enclosed in a regularly shaped cell that is space filling. The commonly used regular shapes include cubes and truncated octahedra, which usually have a cell length significantly larger than the diameter of the solute molecule. This cell is then translated in all three-dimensions to produce a 3D periodic array of the system, analogous to a statistical ensemble. During an MD simulation, periodic boundary conditions are implemented as follows: the molecules coordinates, momenta, accelerations, etc are replicated in the other cells. When a molecule passes the boundary of the central cell its coordinates are translated (or modified) such that it reappears on the other side of the cell with the identical momentum, accelerations, etc (Figure 2.8a).

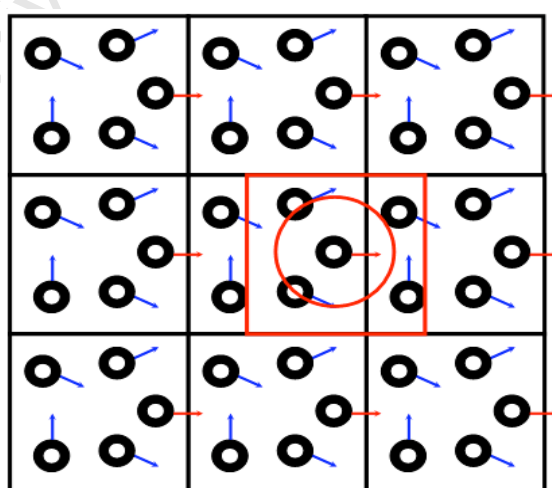


Figure 2.8: (a) 2D illustration of periodic boundary conditions, where the central box is repeated throughout space. (b) The red box is an illustration of the minimum image convention with the spherical potential cutoff. The central atom, in the red box (with the red arrow), is shown with its closest image of every other atom in the system.

The result of periodic boundary conditions is the overall elimination of edge and surface effects, allowing for the interaction of any molecule with all other molecules within the central cell as well as with the replicated images. It should be noted that no molecule or atom should interact with itself in an adjacent cell, an effect that would result in a crystalline system being simulated. Therefore, in principle each atom interacts with all of its infinite set of images as well as those of all other atoms. This is, however, most undesirable for both practical and theoretical reasons: firstly, the calculation of such forces in a periodic system is possible, but more computationally intensive; and secondly, long-range interactions in the potential will give rise to long range order, which is undesirable in a simulation of bulk liquid.

An elegant solution to this problem is to restrict the interaction of all atoms to its closest image of every other atom in the system, a condition known as the *minimum image convention* (Figure 2.8b). The idea of the minimum image convention is to prevent the system from behaving too much like a crystal. Setting a cutoff radius, for calculating the interaction potential and forces, to less than half the cell length can eliminate the possibility of this scenario. A satisfactory guess to the cutoff distance is to set it as the diameter of the solute plus the diameter of a water molecule (~ 2.8 Å). This cutoff distance should be less than half the cell length to prevent interaction of the molecule with itself in the adjacent cells.*

2.1.5 Potential truncation

The computationally most expensive part of any MD simulation is calculating the *non-bonded potential* and *forces*. This routine accounts for almost 90% of the time taken to run an MD simulation. The reason is due to the potential being calculated in a *pairwise* manner, which, by its nature, uses a *double loop* to calculate the potential energy and forces for all pairs of atoms within the system. Generally, with no forms of optimization, the number of non-bonded interaction terms increases with the square of the number of atoms and, hence, the computation time is of $O(N^2)$ order, rather than $O(N)$. In addition to this practical consideration, it is also undesirable to have long-range forces especially for a system that uses periodic boundary conditions. The present cutoff radii are generally between 8.0 and 12.0 Å, since in this range the pair potential is

* The solvated benzene dimer system was used to determine the type of water model, minimum box length and cutoffs to be used without introducing any undesired artifacts into the simulation. It was found that the system worked best with using the Tip4p-EW water model, a box length ≥ 33.1 Å, Ewald summation and cutoffs 10.0, 12.0 and 14.0 Å.

generally within 1% of its global minimum value. However, this introduces severe errors because the potential is no longer continuous. This results in a total energy that is no longer conserved, leading to an unstable trajectory. This sudden cutoff introduces a discontinuity in the potential function at the cutoff distance, which artificially increases the kinetic energy and temperature of the system. As a result it is necessary to either scale the velocities or strongly couple the system to a heat bath to keep the temperature constant. Both methods are unphysical and potentially detrimental, because they may obscure underlying defects in the simulation setup.^{10, 12} To avoid this problem, a method has been developed which uses various functions to zero the potential at the cutoff distance in a continuous fashion. The two methods commonly used are the “*shifting*” (Equation 2.18) and “*switching*” (Equation 2.19) functions.

The shifting function, as the name implies, shifts the entire potential energy surface by a constant in such a way that the new potential is zero at the cutoff distance (Figure 2.9, solid line). The potential energy, for separation distances less and greater than the cutoff distance, is modified as follows:

$$U_{ijSf}(r_{ij}) = \begin{cases} U_{ij} \left[1 - \frac{2r_{ij}^2}{r_{cut}^2} + \frac{r_{ij}^4}{r_{cut}^4} \right] & r_{ij} < r_{cut} \\ 0 & r_{ij} > r_{cut} \end{cases} \quad 2.18$$

However, this method may result in substantial changes in the calculated thermodynamic properties and is generally not preferred.

An alternative approach is by using the switching methodology, which applies a switching function (Figure 2.9, dashed line) to zero the potential smoothly and continuously at the cutoff distance, over a relatively short distance range (defined between r_{on} and r_{off})

$$U_{ijSw}(r_{ij}) = \begin{cases} U_{ij} & r_{ij} \leq r_{on} \\ U_{ij} \left[\frac{(r_{off} - r_{ij})^2 (r_{off} + 2r_{ij} - 3r_{on})}{(r_{off} - r_{on})^3} \right] & r_{on} < r_{ij} \leq r_{off} \\ 0 & r_{ij} > r_{off} \end{cases} \quad 2.19$$

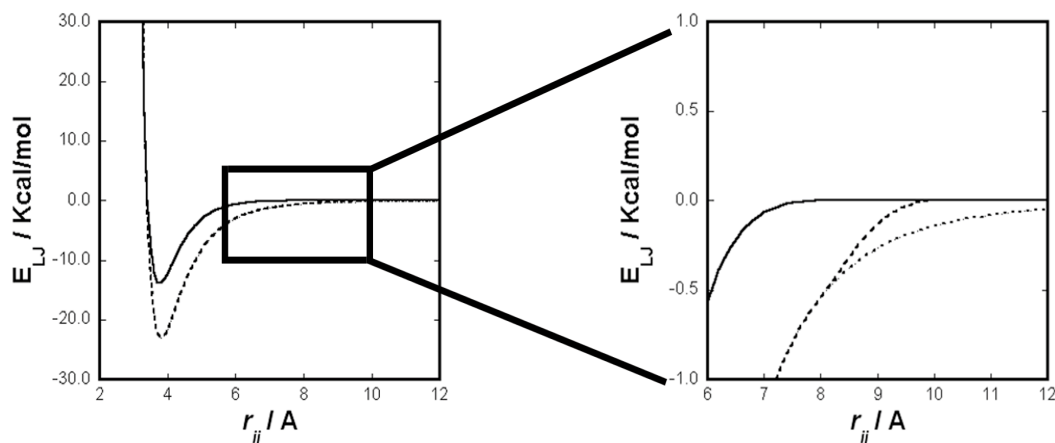


Figure 2.9: The effects of the shifting (solid line, cutoff 8.0 Å) and switching (dashed line, cutoff between 8.0 and 10.0 Å) potential truncation on the true Lennard-Jones potential (dotted line).

Smoothing functions can be applied on an atom-by-atom basis or to a predefined neutral group of atoms within the molecule. Switching functions applied on a group-by-group basis have shown to be the superior method, since it gives excellent energy conservation. However, there are many other methods of producing good cutoff behaviour. For example, applying the switching function directly to the forces such that the potential is given by the integral of the switched forces. These methods have been termed the “force-based” methods.¹³

2.1.6 Constraint Dynamics

The highest frequency motion present in the system dictates the timestep, Δt , used in an MD simulation. For example, bond vibrations involving light atoms (such as hydrogen) occur so quickly that an extremely short timestep would be required to accurately model the system. These high frequency motions are usually of little interest than the lower frequency modes, which often correspond to major conformational changes. Several algorithms have been developed to constrain these high frequency motions to allow for an increase in the MD timestep, without decreasing the accuracy and stability of the simulation. These algorithms constrain the bond length, b , between atoms at $\mathbf{r}_1$ and $\mathbf{r}_2$. The constraint function is written as:

$$\chi(\mathbf{r}_1, \mathbf{r}_2) = (\mathbf{r}_1 - \mathbf{r}_2) \cdot (\mathbf{r}_1 - \mathbf{r}_2) - b^2 = 0. \quad 2.20$$

This constraint has to be correctly applied at each update of the coordinates and momenta, for the two atoms involved, in the integration algorithm. If $(\mathbf{q}_{1uc}, \mathbf{q}_{2uc})$ and $(\mathbf{p}_{1uc}, \mathbf{p}_{2uc})$ are the unconstrained coordinates and momenta variables, respectively, then we can define the update constrained variables as

$$\mathbf{p}_i\left(t + \frac{\Delta t}{2}\right) = \mathbf{p}_{(UC)_i}\left(t + \frac{\Delta t}{2}\right) + \lambda \mathbf{g}_i(t) \quad 2.21$$

$$\mathbf{r}_i(t + \Delta t) = \mathbf{r}_{(UC)_i}(t + \Delta t) + \frac{\lambda \mathbf{g}_i(t) \Delta t}{m} \quad 2.22$$

$$\mathbf{p}_i(t + \Delta t) = \mathbf{p}_{(UC)_i}(t + \Delta t) + \mu \mathbf{g}_i(t + \Delta t) \quad 2.23$$

for $i = 1, 2$ and $\mathbf{g}_i(t) = -\partial\chi/\partial\mathbf{r}_i$. These equations form the basis for the two most popular constraint algorithms *SHAKE*¹⁴ and *RATTLE*,¹⁵ and the Lagrange multipliers, λ and μ , are determined by them. *SHAKE* is an algorithm usually employed when the Verlet and leapfrog integration algorithms are used, since the velocity is not directly calculated in these schemas. The *RATTLE* algorithm, on the other hand, is used with the velocity Verlet integration algorithm, which directly calculates the velocity at each half and full timestep. *RATTLE* serves to properly constrain the velocities of the atoms involved in the constraint.

2.2 Sampling From Ensembles

To compare the results between simulations and experiments, we have to ensure that we run the simulations under the same conditions as the experiments were performed^{10, 16} (i.e., comparing apples with apples). Thermodynamic properties such as the pressure, volume, temperature and the number of particles are all thermodynamic internal constraints that can easily be controlled in an experiment. This thesis deals with the Helmholtz free energy, which is derived from the canonical (NVT) ensemble (Chapter 3). It requires that the number of particles, volume and temperature be held constant. To impose this restriction the Hamiltonian has to be modified. The temperature is held constant by using the Nosé-Hoover thermostat, where a heat bath is coupled to the system. This generates canonically distributed positions and momenta much in the same spirit that the canonical ensemble configurational partition function, Z , was derived (Chapter 3, section 3.2.2). The thermostat adds another degree of freedom to the Hamiltonian and allows the temperature to fluctuate about a certain value. The temperature is an intensive property and, therefore, the additional degree of freedom is to serve as the heat reservoir (or bath). The extensive counterpart to temperature is the kinetic energy, and is related to it in the following manner:

$$T = \frac{1}{3Nk_B} \sum_{i=1}^N \frac{\mathbf{p}_i^2}{m_i}. \quad 2.24$$

The additional degree of freedom has a coordinate and conjugate momentum, s and p_s , added to the system with a thermal mass, Q_m

$$H(\mathbf{r}^N, \mathbf{p}^N) = \frac{\mathbf{p}^2}{2m} + V(\mathbf{r}) + \frac{\mathbf{p}_s^2}{2Q_m} + 3Nk_B T \ln s. \quad 2.25$$

By defining the alternative momenta as $\mathbf{p}' = \mathbf{p}/s$ we can find that the microcanonical ensemble in $\mathbf{r}$, $\mathbf{p}$, s and 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. The scaling factor's (s) associated momenta and the timestep is

$$\mathbf{p}'_s = \frac{\mathbf{p}_s}{s}, \quad 2.26$$

$$\Delta t = \frac{\Delta t}{s}, \quad 2.27$$

By scaling these, we can run a dynamics simulation that generates canonically distributed coordinates and momenta. The Hoover formulation adopts a new variable,

$$\xi = \frac{s\mathbf{p}'_s}{Q_m} \quad 2.28$$

This simplifies the Hamiltonian and equations of motion in terms of the primed coordinates from those proposed by Nosé:

$$\dot{\mathbf{r}}'_i = \frac{\mathbf{p}'_i}{m_i}, \quad 2.29$$

$$\dot{\mathbf{p}}'_i = -\nabla_i V(\mathbf{r}^N) - \xi \mathbf{p}'_i, \quad 2.30$$

$$\dot{\xi} = \frac{1}{Q_m} \left[\sum_{i=1}^N \frac{\mathbf{p}_i'^2}{m_i} - 3Nk_B T \right], \quad 2.31$$

$$\frac{\dot{s}}{s} = \xi = \frac{d \ln s}{dt}, \quad 2.32$$

Equations 2.29 – 2.32 provide modified equations of motion that maintain a constant temperature, which corresponds to a canonical (NVT) ensemble.

2.3 Statistical Analysis

2.3.1 Pair Distribution Functions

The atomic *pair distribution function (PDF)* is a method used to describe the radially averaged environment surrounding an atom. The most commonly used pair distribution function is the *radial distribution function (RDF)*, which is used to give insight into the isotropic intermolecular structuring around a molecule. This function gives the probability of finding pairs of atoms a distance r apart, relative to the probability expected for a completely random distribution of atoms at the same density.¹¹ The RDF can easily be extended to molecular liquids, such as water, and solids by calculating the site-site distributions, for example O–O or O–H. Typically, a large number of molecules (or sites) are needed to calculate an accurate RDF.¹⁷ The RDF is expressed in the following way

$$g(r) = \rho^{-2} \left\langle \sum_i \sum_{j \neq i} \delta(\mathbf{r}_i) \delta(\mathbf{r}_j - \mathbf{r}) \right\rangle = \frac{V}{N^2} \left\langle \sum_i \sum_{j \neq i} \delta(\mathbf{r} - \mathbf{r}_{ij}) \right\rangle. \quad 2.33$$

where, r is the interatomic distance from the selected atoms in the solute molecule and the solvent molecule, ρ is the number of molecules (N) in the volume (V), the number density and the factor ρ^{-2} is used to normalization factor to ensure the RDF approaches unity for $r \rightarrow \infty$ (i.e., completely random distribution of configurations).¹⁶⁻¹⁸

In the case of gases, the RDF does not display significant structuring beyond the first peak, which indicates only local ordering in the system. In solids, definite sharp peaks are observed and appear to repeat almost infinitely even at large intermolecular separations; indicative of crystal/periodic long-range ordering. The liquid RDF is the intermediate scenario; it has strong local structuring (similar to gases) and slightly weaker long-range order (similar to solids) shown in Figure 2.10.

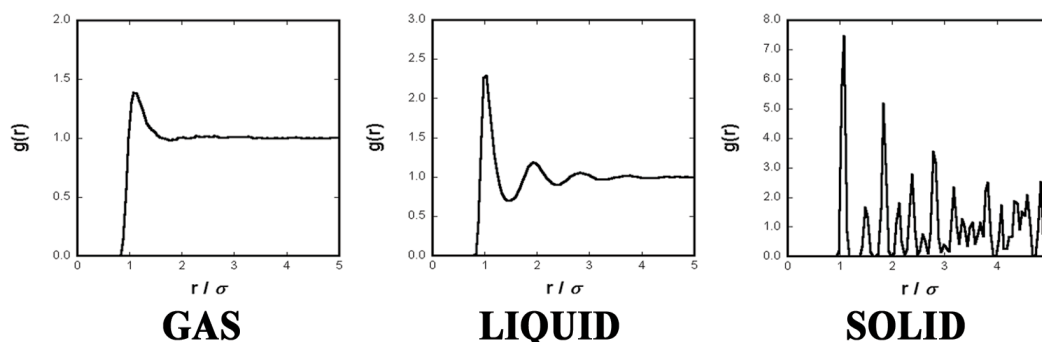


Figure 2.10: The radial distribution function (RDF) for an arbitrary molecule in the gas, liquid and solid phase. Long-range order (or structure) decreases when the molecules have more freedom to move.

2.3.2 Spatial Distribution Function

The spatial distribution function (SDF), also known as the solvent probability distribution function, is a method commonly used to ascertain the anisotropic solvent structuring (probability distribution) around a solute molecule. Whereas the site-site RDF provides a radially averaged isotropic description, the SDF presents a complete spatial, or 3D, view of the solvent distribution around the solute allowing for any anisotropy to be observed.¹⁹⁻²² To calculate an SDF, all of the translational and rotational motions have to be removed by a best least squares fitting procedure to some user defined structure, or by using the first coordinate frame of the trajectory (Figure 2.11).

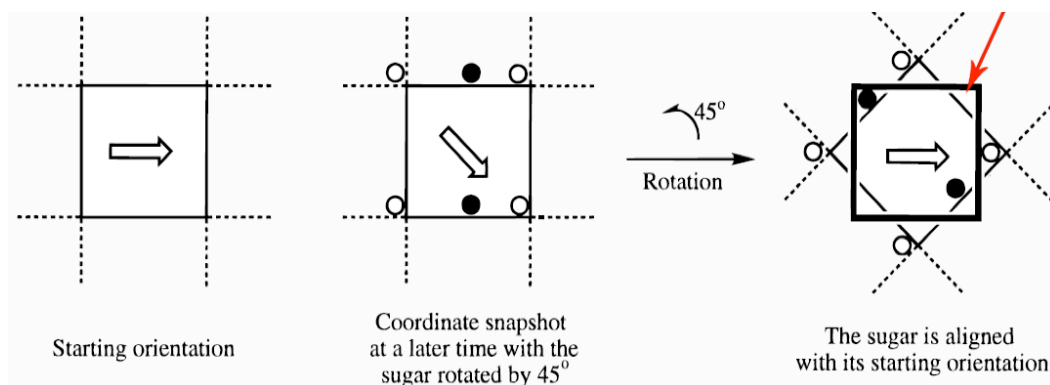


Figure 2.11: The procedure for removing the translation and rotation, around the solute, from a MD trajectory. The box (indicated by the red arrow, far right) is where the averaging and calculation of the density is performed.

Care should be taken in the fitting procedure, as too much movement in either the solute or solvent can result in a “smearing” of the 3D probability distribution. There are two ways to display the 3D probability: (1) by using the nuclear coordinates; or (2) by assigning the atom (or the site of interest) of the solvent a Gaussian function to smoothly represent its size, Equation 2.34. The latter is used in this thesis, because it displays the SDF as a smooth 3D probability.

$$G(r) = \left(\frac{a}{\pi}\right)^{\frac{3}{2}} \exp(-ar^2) \quad 2.34$$

After the fitting procedure a 3D histogram (or 3D array of bins) is created around the solute and the 3D density (Equation 2.30) of the solvent is calculated around the solute.

The histograms are then normalized with respect to the bulk density of the surrounding solvent molecules¹²

$$\bar{\rho}(x_i, y_i, z_i) = \rho(x_i, y_i, z_i) \frac{n_{xbin} n_{ybin} n_{zbin}}{n_{molecular_site} n_{frames}}. \quad 2.35$$

To obtain a smooth SDF many frames of a trajectory are required. However, this large number frames can effectively be reduced if there are multiple molecules (of similar kind, e.g. solvent molecules) that can be used to calculate the 3D density.¹⁶

2.3.3 Correlation Functions

When the need for calculating transport properties or time-dependent quantities arises, we need to link dynamical motions and fluctuations with some function that can correctly describe it in a statistical manner.¹¹ For instance, investigating how the rotation about a protein's dihedral changes with time and determine if there is any correlation in the motion at any given time. This can be achieved by using correlation functions, which can provide numerical insight into the strength of correlations of motion (or any dynamical property) at any time. These can be linked to experimentally obtained observables via statistical mechanics. For correlation functions of the same variable, known as an autocorrelation function, we define it as follows:

$$c(t) = \int_0^T A(t + \tau) A(\tau) d\tau, \quad 2.36$$

or it can be expressed in the more general ensemble average form

$$c(t) = \langle \delta A(0) \delta A(t) \rangle. \quad 2.37$$

$\delta A(t)$ represents the instantaneous fluctuation in $A(t)$ from its time-independent equilibrium average, $\langle A \rangle$, at time t

$$\delta A(t) = A(t) - \langle A \rangle. \quad 2.38$$

The correlation function is therefore a measure of the correlation of motion in the system. It determines how quickly the motions or quantities, under investigation, become uncorrelated. At the limits of a correlation function, where $t = 0$, the $c(t)$ is equal to 1, which indicates that the quantities are exactly the same. At some later time, $t = t'$, the $c(t')$ tends toward 0 as the quantities become more uncorrelated.¹¹

Many different properties can be determined from correlation functions including diffusion,¹¹ shear viscosity,¹¹ recover the Fourier phase information from the Patterson function from X-ray diffraction experiments, to name a few. For example, the diffusion constant is found by calculating the autocorrelation function of the velocity vector

$$D = \frac{1}{3} \int_0^\infty \langle \mathbf{v}(0) \cdot \mathbf{v}(t) \rangle dt. \quad 2.39$$

The Einstein relation makes the approximation, that at very long times, the diffusion becomes proportional to mean square displacement in the following way

$$2tD = \frac{1}{3} \langle |\mathbf{r}(0) - \mathbf{r}(t)|^2 \rangle, \quad 2.40$$

where t is the time, $\mathbf{r}(0)$ and $\mathbf{r}(t)$ are the initial position (at $t = 0$) and position at some time t , respectively, and $\langle \rangle$ denotes the ensemble average.

Clearly statistics plays a central role in all atom molecular dynamics (MD) simulations. It not only lays the foundations for the statistical mechanics, but it is also central in analyzing the microscopic trajectory for comparison with macroscopic observables. Hence, the following chapter discusses the theory of statistical mechanics and the calculation of the Helmholtz free energy (A) from molecular dynamics simulations.

2.4 References

1. Born, M.; Oppenheimer, R., Zur Quantentheorie Der Molekule. *Ann. Phys.* **1927**, *84*, 457.
2. Leach, A. R., *Molecular Modeling*. 2nd ed.; Longman: 2001.
3. Lewars, E., *Computational Chemistry: Introduction to the Theory and Applications of Molecular and Quantum Mechanics*. Kluwer Academic Publishers: 2004; Vol. 2nd.
4. Brooks, B. R., *et al.*, Charmm: The Biomolecular Simulation Program. *J. Comput. Chem.* **2009**, *30* (10), 1545.
5. Ewald, P. P., Berechnung Optischer Und Elektrostatischer Güiterpotentiale. *Annalen der Physik* **1921**, *369* (3), 253.
6. Darden, T., *et al.*, Particle Mesh Ewald - an Nlog(N) Method for Ewald Sums in Large Systems. *J. Chem. Phys.* **1993**, *98* (12), 10089.
7. Essman, U., *et al.*, A Smooth Particle Mesh Ewald Method. *J. Chem. Phys.* **1995**, *103* (19), 8577.
8. Hockney, R. W.; Eastwood, J. W., *Computer Simulations Using Particles*. McGraw-Hill: New York, 1981.
9. York, D.; Yang, W. T., The Fast Fourier-Poisson Method for Calculating Ewald Sums. *J. Chem. Phys.* **1994**, *101* (4), 3298.
10. Matthews, R. P. Ion Pairing in Aqueous Metal Sulfates and Platinum Group Metal Ammonium Solutions. Thesis, University of Cape Town, Cape Town, 2010.
11. Allen, M. P.; Tildesly, D. J., *Computer Simulation of Liquids*. Oxford Univeristy Press: Oxford, 1987.
12. Kuttel, M. M., Developing Analytical Tools for Saccharides in Condensed Phases. *MSc. Thesis, Univeristy of Cape Town*. **1999**.
13. Molteni, C.; Parrinello, M., Glucose in Aqueous Solution by First Principles Molecular Dynamics. *J. Am. Chem. Soc.* **1998**, *120* (9), 2168.
14. Ryckaert, J.-P., *et al.*, Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of N-Alkanes. *J. Comp. Phys.* **1977**, *23* (3), 327.
15. Andersen, H. C., A "Velocity" Version of the Shake Algorithm for Molecular Dynamics Simulations. *J. Comp. Phys.* **1983**, *52* (1), 24.
16. Strumpfer, J. Computing Free Energy Hypersurfaces for Anisotropic Intermolecular Associations. Thesis, University of Cape Town, Cape Town, 2009.
17. Gamielien, M. R., Investigations of the Solution Properties of Both Unmodified and Modified Cyclodetrins. *MSc. Thesis, University of Cape Town* **2008**, Chapter 3.
18. Rossky, P. J.; Karplus, M., A Molecular Dynamics Study of a Dipeptide in Water. *J. Am. Chem. Soc.* **1979**, *110* (8), 1913.
19. Lui, Q.; Brady, J. W., Anisotropic Solvent Structuring in Aqueous Sugar Solutions. *J. Am. Chem. Soc.* **1996**, *118*, 12276.
20. Schmidt, R. K., *et al.*, The Anomeric Equilibrium in D-Xylose Free Energy and the Role of Solvent Structuring. *J. Am. Chem. Soc.* **1996**, *118*, 541.
21. Lui, Q., *et al.*, Molecular Dynamics Studies of the Hydration Alpha, Alpha-Trehalose. *J. Am. Chem. Soc.* **1997**, *119* (33), 7862.
22. Lui, Q.; Brady, J. W., Model Dependence of the Anisotropic Structuring of Solvent Water around Sugars in Molecular Dynamics Simulations. *J. Phys. Chem. B* **1997**, *101* (8), 1817.

CHAPTER 3

STATISTICAL MECHANICS AND FREE ENERGY

Free energy is a term used to refer to the amount of energy there is available for the system to do work. How do we calculate a macroscopic property, such as free energy, from a microscopically detailed experiment such as MD? Molecular dynamics renders a great deal of microscopic information (e.g. positions, velocities, etc) and the conversion of this detailed information into macroscopic, measurable quantities (e.g. pressure, volume, etc.) is the province of statistical mechanics.¹ Statistical mechanics is the link between the microscopic dynamics or fluctuations and the macroscopic measurable observables. In other words, statistical mechanics deals with large numbers of particles/molecules, treating them as matter in bulk.¹⁻⁴ In statistical mechanics there is the idea of an ensemble that holds a few thermodynamic variables (or internal constraints) constant, while letting the others fluctuate. Each ensemble has its conjugate partition function, which contains all the statistical information to calculate the various ensemble free energies. The microcanonical, canonical and isothermal-isobaric ensembles each yield the entropy, Helmholtz and Gibbs free energies, respectively.^{1, 2, 4, 5} In this thesis, the Helmholtz free energy is used exclusively and all the free energy methods described have, thus, been derived for the canonical ensemble. This chapter introduces the basic concepts of thermodynamics, statistical mechanics and the Helmholtz free energy. A brief overview of free energy methodology is then provided. The chapter concludes with the newly developed multi-dimensional Free Energy from Adaptive Reaction Coordinate Forces (*FEARCF*) and Weighted Histogram Analysis Method (*WHAM*) methods used in this thesis.

3.1 Statistical Mechanics: The Beginnings

The history of thermodynamics stems as far back as the late 1700s with the advent of thermochemistry, it was only in the early 1800s that classical thermodynamics fully developed pioneered by Nicolas Leonard Sadi Carnot, the “father of thermodynamics”, in his work on the efficiency of an idealized engine.⁶ This was followed by the formulation of statistical mechanics in the mid 1800s mostly by the efforts of Ludwig Boltzmann and

James Clerk Maxwell. Josiah W. Gibbs finally put it into a more general context around 1900, to what we know as the laws of thermodynamics. The reader is encouraged to briefly review the fundamental laws of classical thermodynamics.^{2, 3, 7-9} Later, we see classical thermodynamics relates to the microscopic world by employing statistical laws (mechanics) that govern large systems.^{1, 5}

3.2 Statistical Mechanics: The Fundamentals

The first key concept of statistical mechanics is the ergodic hypothesis, which relates measurable observables with ensemble averages. Secondly, Liouville's theorem, which is key to Hamiltonian dynamics, asserts that we can calculate the density of microstates from any statistical ensemble. Finally, the fundamental postulate of statistical mechanics states that at equilibrium the probability of all the microstates of a system are equally probable, which leads to the most probable macrostate. The idea of an ensemble is essential for determining the configurational partition function (Z). From the Z various thermodynamic functions, such as free energy, can be calculated.

3.2.1 Concepts

The most seemingly presumptuous concept is the ergodic hypothesis, which postulates that the expectation value of an observable for a system is equal to the measurable macroscopic value given a sufficient amount of time, based on some internal constraints (or thermodynamic variables).^{1, 4, 5}

$$A_{obs} = \langle A \rangle_{ens} = \frac{1}{t} \sum_{i=1}^N A_i d\tau_i = \frac{1}{t} \int A dt, \quad 3.1$$

where A is some measurable, $\langle A \rangle_{ens}$ is that measurable's ensemble average, t is the time and τ is the number of steps in the simulation.

Another crucial theorem, which plays a central role in sampling, is Liouville's theorem. This theorem is probably the hardest to come to grips with as a beginner to statistical mechanics and, more specifically, hyper-dimensional phase space (or Hilbert space). Phase space can be thought of as a continuous fluid that remains continuous when it is allowed to flow. It begins as a definite shape evolving, in time, to represent a shape with many tentacle-like arms. Liouville's theorem is a fundamental theorem to statistical mechanics and states that the systems, in an ensemble, are neither created nor destroyed. This forms the mathematical facet for the conservation of energy and mass.¹⁰

This theorem deals with how the phase space distribution evolves with time, and the mathematical form Liouville's theorem is given as the Liouville equation¹¹

$$\frac{d\rho(\Omega, t)}{dt} = -\sum_{i=1}^N \left(\frac{\partial \rho(\Omega, t)}{\partial \mathbf{q}_i} \dot{\mathbf{q}}_i + \frac{\partial \rho(\Omega, t)}{\partial \mathbf{p}_i} \dot{\mathbf{p}}_i \right), \quad 3.2$$

If we have an equilibrium ensemble then the phase space distribution, ρ , is only a function of the Hamiltonian, H . It can be shown, using the canonical equations of the Hamiltonian, that

$$\dot{\mathbf{q}} = \frac{\partial \mathbf{q}}{\partial t} = \frac{\partial H}{\partial \mathbf{p}} \quad 3.3.1$$

$$\dot{\mathbf{p}} = -\frac{\partial \mathbf{p}}{\partial t} = -\frac{\partial H}{\partial \mathbf{q}}, \quad 3.3.2$$

and the phase space distribution remains constant over time¹¹

$$\frac{d\rho(\Omega)}{dt} = \frac{\partial \rho(H)}{\partial H} \left[-\sum_{i=1}^N \left(\frac{\partial H}{\partial \mathbf{q}_i} \frac{\partial H}{\partial \mathbf{p}_i} - \frac{\partial H}{\partial \mathbf{p}_i} \frac{\partial H}{\partial \mathbf{q}_i} \right) \right] = 0. \quad 3.4$$

Finally, the fundamental postulate of statistical mechanics states:⁵

For an isolated system with a fixed total energy (E) and size (perhaps specified by the volume and total number of particles) all the microstates are equally likely at thermodynamic equilibrium.

In other words, at equilibrium the macrostate of the system corresponds to the most random microstate – the distribution of the microstates at the same energy and size is entirely uniform. This leads to the definition of the Boltzmann entropy

$$S = k_B \ln \Omega(N, V, E), \quad 3.5$$

where k_B is the Boltzmann constant and $\Omega(N, V, E)$ corresponds to the number of microstates with the same number of particles (N), volume (V) and energy (E).

3.2.2 Ensemble theory

Josiah W. Gibbs introduced the idea of an ensemble around 1900. The idea is as follows: create N (where N is large) identical macroscopic copies of the system but with one prerequisite, that all the copies of the system should correspond to some real

macroscopic state (or macrostate) composed of many different microscopic states (microstates).^{2, 12}

THE MICROCANONICAL ENSEMBLE

When calculating thermodynamic quantities we do so for a system in a particular macroscopic state. For a given macroscopic state (macrostate) there exist a large number of possible microstates (different microscopic configurations). For the internal constraints that a given number of particles, N , volume, V , and internal energy, E , remain constant, we denote the number of these microstates as $\Omega(N, V, E)$. This is referred to as the *microcanonical ensemble*. For a classical system, of indistinguishable particles, the number of microstates is written as:

$$\Omega(N, V, E) = \frac{1}{N! h^{3N}} \int d^N \mathbf{q} d^N \mathbf{p} \delta[H(\mathbf{q}^N, \mathbf{p}^N) - E] \quad 3.6.1$$

$$\Omega(N, V, E) = \frac{1}{N! h^{3N}} \sum_{i=1}^N d\mathbf{q}_i d\mathbf{p}_i \delta[H(\mathbf{q}_i, \mathbf{p}_i) - E]. \quad 3.6.2$$

Here $H(\mathbf{q}^n, \mathbf{p}^n)$ is the Hamiltonian for a continuous distribution, while $H(\mathbf{q}_i, \mathbf{p}_i)$ is the Hamiltonian of a discrete one. The delta function $\delta(H - E)$ selects states of an N -particle system in a container of volume V that have the desired energy E . When the states are continuous, the δ is the Dirac Delta function (Equation 3.6.1);^{1, 13} when the set of states is discrete, δ is just the Kronecker delta (Equation 3.6.2), taking on values of 0 or 1.¹

Consider two systems in thermal contact with energies E_1 and E_2 , number of particles N_1 and N_2 and volumes of V_1 and V_2 .⁵ The total energy of the system is simply the sum of the two systems in contact, i.e. $E_0 = E_1 + E_2$. The number of microstates of the total system is the 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.⁵

$$\Omega_o(N_1, N_2, E_1, E_2, V_1, V_2) = \Omega_1(N_1, V_1, E_1) \Omega_2(N_2, V_2, E_2). \quad 3.7$$

Consider further that the particles of the two systems cannot mix and that the exterior boundary of the system is fixed such that we have N_1 and N_2 are constant and the total volume, $V_0 = V_1 + V_2$, is also constant. Naturally, systems evolve to their most probable macrostate, which means that the most probable macrostate is one in which maximizes its number of microstates.

For a system of constant total energy, volume and number of particles we find the energies E_1^* and E_2^* that maximizes Ω_o as

$$\frac{\partial [\Omega_1(N_1, V_1, E_1) \Omega_2(N_2, V_2, E_2)]}{\partial E_1} = 0 \quad 3.8$$

$$\left(\frac{\partial \Omega_1}{\partial E_1} \Omega_2 + \Omega_1 \frac{\partial E_2}{\partial E_1} \frac{\partial \Omega_1}{\partial E_2} \right)_{E_1=E_1^*, E_2=E_2^*} = 0. \quad 3.9$$

Where $\Omega(N, V, E)$ has been condensed as Ω . Since, $\frac{\partial E_2}{\partial E_1} = -1$, Equation 3.9 can be reduced to

$$\frac{1}{\Omega_1} \frac{\partial \Omega_1}{\partial E_1}(E_1^*) = \frac{1}{\Omega_2} \frac{\partial \Omega_1}{\partial E_2}(E_2^*), \quad 3.10$$

and we find that it is the same as

$$\frac{1}{\Omega_1} \frac{\partial \Omega_1}{\partial E_1}(E_1^*) = \frac{1}{\Omega_2} \frac{\partial \Omega_1}{\partial E_2}(E_2^*) = \text{const}. \quad 3.11$$

This equality can be generalized to any number of systems in thermal contact using $\frac{\partial \ln \Omega}{\partial E} = \text{const}$ for each system.^{1, 4, 5} This constant can be calculated to yield:

$$\frac{\partial \ln \Omega}{\partial E} = \frac{1}{k_B T}. \quad 3.12$$

From the Second Law of thermodynamics we know that at constant volume and number of particles $\frac{\partial S}{\partial E} = \frac{1}{T}$, which leads to the very important Boltzmann entropy^{1, 4, 5}

$$S = k_B \ln \Omega. \quad 3.13$$

THE CANONICAL ENSEMBLE

We often cannot accurately control, or keep constant, the internal energy. Instead, it is easier to hold the temperature fixed, which also makes the interpretation of the experiments easier. Theoretical calculations involving the same thermodynamic variables as the experiments have to be performed. Experiments, measuring thermodynamic parameters, usually involve knowledge of the temperature and not internal energy.^{14, 15} This leads us to the *canonical ensemble*, which has fixed the number of particles, N , volume, V , and temperature, T . To calculate the microstate distribution for a given set of N , V and T , we have to couple our system of interest to a large heat reservoir. There are essentially two systems: the reservoir heat bath with energy E_r ; and

the system of interest in state i with energy E_i , in other words the configurational energy. As in the preceding section, the total energy of the two systems is constant, $E_r + E_i = E = \text{const}$. The probability of finding the bath in a particular state is then^{1, 2, 4, 5, 10, 12}

$$p_i = \frac{\Omega(N, V, E - E_i)}{\sum_{i=1}^N \Omega(N, V, E - E_i)}. \quad 3.14$$

We can continue to expand the natural logarithm of the microcanonical density of states:

$$\ln \Omega(N, V, E - E_i) = \ln \Omega(N, V, E) - E_i \frac{\partial \ln \Omega(N, V, E)}{\partial E} + O\left(\frac{1}{E}\right). \quad 3.15$$

We can use the relation from Equation 3.12 to obtain

$$\ln \Omega(N, V, E - E_i) = \ln \Omega(N, V, E) - \frac{E_i}{k_B T} + O\left(\frac{1}{E}\right). \quad 3.16$$

By inserting Equation 3.16 into 3.14 we obtain the Boltzmann distribution

$$p_i = \frac{\exp(-E_i \beta)}{\sum_{i=1}^N \exp(-E_i \beta)}, \quad 3.17$$

where $\beta = \frac{1}{k_B T}$ and the $\sum_i^N p_i = 1$. From Equation 3.17 the *configurational partition function*, Z , can be defined from the *configurational energy*, E_i , as:

$$Z(N, V, T) = \sum_{i=1}^N \exp(-E_i \beta). \quad 3.18$$

3.3 The Free Energy Method

In the canonical ensemble, the Helmholtz free energy is the amount of energy available for the system to do work, while the number of particles, temperature and volume is held constant. The Helmholtz free energy, as well as many other thermodynamic quantities, can be obtained from the configurational partition function, Z . The expectation (average) value of any microscopic value is determined using Z :^{1, 5}

$$\langle X \rangle = X = \sum_i p_i X_i = \sum_i \frac{X_i \exp(-\beta E_i)}{Z}. \quad 3.19$$

The total internal energy is then the expectation value of all the energies of all microstates

$$\langle E \rangle = U = \sum_i p_i E_i = \sum_i \frac{E_i \exp(-\beta E_i)}{Z}, \quad 3.20$$

Taking the derivative with respect to the inverse temperature (β), we obtain the expression for the internal energy, U ,

$$\langle E \rangle = U = -\frac{1}{Z} \frac{\partial Z}{\partial \beta} = -\frac{\partial \ln Z}{\partial \beta}. \quad 3.21$$

The entropy defined by the Gibbs entropy (inferred from the fundamental postulate of statistical mechanics), a discretized version of the Shannon entropy, is given by

$$S = -k_B \sum_i p_i \ln p_i. \quad 3.22$$

By inserting Equation 3.17 and expanding the logarithm, we find the entropy in terms of the U and Z :

$$\frac{S}{k_B} = -\left[-\beta \left(\sum_i \frac{E_i \exp(-\beta E_i)}{Z} \right) - \ln Z \right] = \beta U + \ln Z. \quad 3.23$$

Finally, the Helmholtz free energy ($A = U - TS$) is obtained, with further rearrangement of Equation 3.24,

$$A = -k_B T \ln Z, \quad 3.24.1$$

$$Z = \exp\left(-\frac{A}{k_B T}\right). \quad 3.24.2$$

3.3.1 Free energy perturbation methodology

Free energy perturbation methodologies deal with calculating the difference in free energy between two states, by mutating from one state to another. This can be achieved because free energy differences between closely related states can be calculated by using both experimental and computational methodologies. Examples of free energy perturbation methods are the *free energy perturbation (FEP)* and *thermodynamic integration (TI)* methods. These relate the free energy difference between two states, A and B , to a thermodynamic ensemble average dependant on the potential energy properties of the individual states.^{16, 17} *FEP* and *TI* methods employ the coupling

parameter approach,¹⁸ in which the Hamiltonian of each state, A and B , is written as a function of the coupling parameter, λ

$$H(\mathbf{p}, \mathbf{q}, \lambda_A) = H_A(\mathbf{p}, \mathbf{q}) \quad 3.25$$

$$H(\mathbf{p}, \mathbf{q}, \lambda_B) = H_B(\mathbf{p}, \mathbf{q}), \quad 3.26$$

where the Hamiltonians $H_A(\mathbf{p}, \mathbf{q})$ and $H_B(\mathbf{p}, \mathbf{q})$ correspond to states A and B , and where $\lambda_A + \lambda_B = 1$. This is possible because the free energy is a state function, implying that we can follow any, but realistic, path and only require the knowledge about the initial and final states.

It follows that the configurational partition function, Z , can be written as a function of λ , given by:

$$Z(\lambda) = \frac{1}{N! h^{3N}} \int d\mathbf{q}^N d\mathbf{p}^N \exp[-\beta H(\mathbf{q}^N, \mathbf{p}^N, \lambda)] \quad 3.27$$

and the Helmholtz free energy, A , which depends on Z , can be written in terms of λ :

$$A(\lambda) = -k_B T \ln Z(\lambda). \quad 3.28$$

Therefore, the calculable change in the Helmholtz free energy, ΔA , can be written as a function of λ .

FREE ENERGY PERTURBATION

The free energy perturbation method treats the change between states as a set of discrete steps.^{19, 20} However, representation of the free energy by an analytical function is only valid for simple model systems. To overcome this drawback, the Helmholtz free energy differences is calculated between the related systems, A and B (with Hamiltonians, H_A and H_B). The change in the free energy can be written as follows:

$$\Delta A = A_B - A_A = -k_B T \ln \langle \exp[-\beta(H_B - H_A)] \rangle_{\lambda_A} \quad 3.29$$

where $\langle \rangle_{\lambda_A}$ refers to an ensemble average over a system represented by H_A . This can be generalized by writing H_A and H_B as functions of λ , such that λ varies from 0 ($H = H_A$) to 1 ($H = H_B$). The free energy difference between the molecular systems is calculated using a thermodynamic cycle, as shown in Figure 3.1, and a method used to carry out a smooth conversion from an initial state to a final state. The most commonly used conversion methods are the single- and dual-topology methods. Only the dual-topology method is

discussed here.^{18, 21} In the dual-topology method the topologies of both the initial and final states co-exist throughout the perturbation, and the Hamiltonian is given by:

$$H(q, p, \lambda) = \lambda H_B(q, p) + (1 - \lambda) H_A(q, p). \quad 3.30$$

In this method, the atom types and internal parameters are never changed. The two sets of atoms interact with the rest of the system but never with each other, and their relative contribution to the Hamiltonian, $H(q, p, \lambda)$, changes as λ changes.

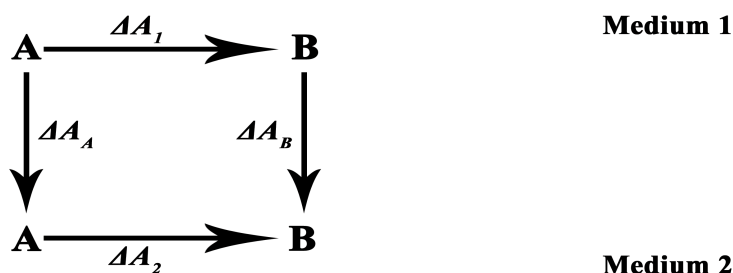


Figure 3.1: The free energy perturbation method based on a generalized thermodynamic cycle. A and B are the initial and final chemical species molecules, respectively, in different media 1 and 2.

The final point to consider, in running *FEP* calculations, is the data collection method. When a molecular system is perturbed from one state to another, data collection is required at each λ point. The λ points are chosen such the energy differences between the points are less than $2.0 k_B T$ (or $\sim 1.2 \text{ kcal/mol}$ at 298.15 K). Various sampling methods exist such as the double-wide,²² overlap and double-ended sampling methods;^{18, 23, 24} only the double-wide sampling method is detailed here. In the double-wide sampling method (Figure 3.2), q is an intermediate point ($\lambda = n$) at which a simulation is performed. q is then used to calculate the average free energy differences for reactants (r) and products (p) as $\Delta A_n \rightarrow \Delta A_{n-1}$ for $\lambda_n \rightarrow \lambda_{n-1}$ and $\Delta A_n \rightarrow \Delta A_{n+1}$ for $\lambda_n \rightarrow \lambda_{n+1}$.

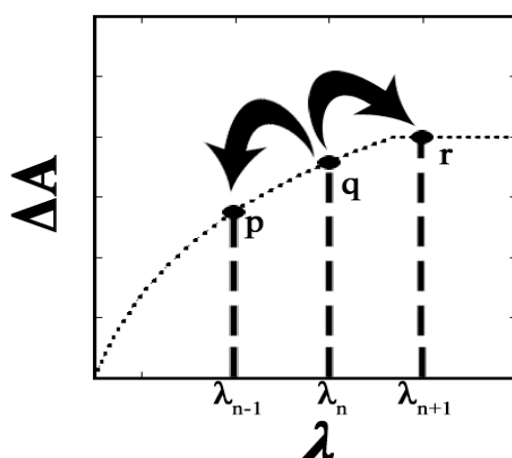


Figure 3.2: Pictorial representation of the double-wide sampling method. q is the intermediate, r and p are the reactant and products, respectively.

This provides an efficient manner to calculate the overall free energy difference

$$\Delta A = \frac{1}{2} \sum_{n=0}^{n-1} [\Delta A_{n \rightarrow n-1} + \Delta A_{n \rightarrow n+1}], \quad 3.31$$

and allow for, if required, the addition of several extra increments in λ . However, increasing the number of λ points does not necessarily improve the calculated free energy difference.

THERMODYNAMIC INTEGRATION

An alternative to *FEP* is thermodynamic integration (*TI*), where the difference is calculated by the integration of an ensemble average over λ , giving the free energy as

$$\Delta A = \int_{\lambda=0}^{\lambda=1} \left\langle \frac{\partial H(\mathbf{q}, \mathbf{p}, \lambda)}{\partial \lambda} \right\rangle_{\lambda} d\lambda. \quad 3.32$$

The Hamiltonian of the initial state is given by $H_{(\lambda=0)}$, while $H_{(\lambda=1)}$ represents the Hamiltonian of the final state.¹⁸ The derivative of the Hamiltonian can be evaluated using Equation 3.32, which yields

$$\Delta A = \int_0^1 \langle H_B(\mathbf{q}, \mathbf{p}) - H_A(\mathbf{q}, \mathbf{p}) \rangle_{\lambda} d\lambda. \quad 3.33$$

This is the equation used to calculate the free energy differences between states in computational simulations. The evaluation of the change in the free energy is carried out using numerical integration of Equation 3.33 by setting up several simulations at discrete values of λ (between 0 and 1), and then calculating the free energy differences.¹⁴

3.3.2 The potential of mean force methodology

Free energy perturbation and thermodynamic integration methods obtain free energies as a function of a chemical mutation from state *A* to *B*. The *potential of mean force* (*PMF*) is the potential required to produce the average force over all configurations of all the $n+1 \dots N$ molecules acting on a particle at any fixed configuration, keeping a set of $1 \dots n$ molecules fixed.¹⁹ The PMF was first introduced by Kirkwood in 1935,¹⁹ and it allows for a reasonably chosen parameter to link state *A* to *B*. The parameter is represented as a reaction coordinate, ξ , where

$$\xi(\mathbf{q}) = \xi(\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_N). \quad 3.34$$

The reaction coordinate is a function of positions of the particles in a system, rather than mutations from state *A* to *B*. The reaction coordinate could be any meaningful metric

such as a distance, angle, dihedral angle or a combination of them. This allows for an N -dimensional free energy hyper-volume being described by N -dimensional reaction coordinates.²⁵

As an example, we take a generalized case such as the distance reaction coordinate between two solute molecules (A and B), in a fixed system containing N solvent molecules. The reaction coordinate probability, $p(r')$, integrated over phase space, is given as:

$$p(r') = \frac{\int \dots \int \exp[-\beta H(\mathbf{q}, \mathbf{p})] \delta(r' - r(\mathbf{q})) d\mathbf{q} d\mathbf{p}}{\int \dots \int \exp[-\beta H(\mathbf{q}, \mathbf{p})] d\mathbf{q} d\mathbf{p}}. \quad 3.35$$

From this it follows, that in terms of the configurational partition function, the reaction coordinate probability can be written as:

$$p(r') = \frac{Z_R(r')}{Z}, \quad 3.36$$

where $Z_R(r')$ is the reaction coordinate configurational partition function. By substituting Equation 3.36 into Equation 3.27 the Helmholtz free energy, defined as the potential of mean force, $W(r')$, is given as

$$W(r') = -k_B T \ln p(r') - kT \ln Z. \quad 3.37$$

The potential of mean force may be obtained using several methods. The simplest of these is the calculation of $W(r')$ from unbiased MD simulations. In this method the $W(r')$, between the two solutes (A and B) due to solvent molecules, is obtained from the radial or pair distribution function, $g(r')$, between the solute molecules

$$W(r') = -k_B T \ln g(r') + C, \quad 3.38$$

where

$$g(r') = \frac{p(r')}{4\pi(r')^2}, \quad 3.39$$

and C is a constant that allows the freedom to choose the energy scale.²⁶⁻²⁸

The formulation as given by Equations 3.38 and 3.39 is the classical way of computing the PMF via the inversion of the radial distribution function from MD simulations.^{4, 5, 29} However, this method only works for systems with low energy barriers ($< 3.0 k_B T$, for

simulations performed at 298.15 K), else the method fails to sample all of the reaction coordinates phase space.³⁰

In order to overcome large energy barriers, an alteration to the potential energy function is required. The PMF can be calculated using constraint-biased real space methods, which apply holonomic constraints or restraints on the reaction coordinates, and employ either forces or finite differences of the physical Hamiltonian to bias the system.³⁰⁻³² However, the use of constraint-biased methods requires the introduction of metric tensor corrections in order to deal with the quenching of momenta on the constrained atoms.³³⁻³⁵ Unlike methods employing Cartesian coordinates, constraint-biased methods require the calculation and application of logarithmic Jacobian corrections.^{31-33, 35-38} Alternatively, umbrella-biased simulations can be conducted with a harmonic biasing potential and multiple window sampling methodologies, along with window matching weighted histogram analysis. Adaptive umbrella sampling methods, using non-harmonic and non-local biasing potentials, can also be employed.

The following sections provide a brief overview and description of the popular non-Boltzmann sampling simulation techniques. These include *adaptive biasing force method*, *umbrella* and *adaptive umbrella sampling methods* and the *Free Energy from Adaptive Reaction Coordinate Forces (FEARCF)* and *Weighted Histogram Analysis Method (WHAM)*. The last two methods are used for the studies in this thesis.

ADAPTIVE BIASING FORCE

The adaptive biasing force method, proposed by Darve and Pohorille,³⁷ presents an adaptive way to drive a simulation over free energy barriers, using a thermodynamic integration like method. This is accomplished by calculating a running average of the force along a reaction coordinate and then applies the negative of this force, as an external force, to only the atoms involved in the reaction coordinate. This allows the simulation to continue in a continuous motion along the prescribed reaction coordinate.

The derivative of the free energy, defined by a reaction coordinate (e.g. distance, r), is given by

$$\frac{\partial A(\mathbf{r})}{\partial r} = \left\langle -\mathbf{F}(\mathbf{r}) + \frac{\partial \ln|J|}{\partial \mathbf{r}} \middle| \mathbf{r} \right\rangle, \quad 3.40$$

where $\mathbf{F}(\mathbf{r})$ is the mean force at $\mathbf{r}$ and J is the Jacobian correction. The Jacobian correction is required because the force has to be converted from reaction coordinate

forces to atomic (Cartesian) coordinate forces. The free energy can be calculated by integrating the negative of Equation 3.40

$$\Delta A(\mathbf{r}) = -\int_{r_o}^r \left\langle \mathbf{F}(\mathbf{r}') + \frac{\partial \ln|J|}{\partial \mathbf{r}} \right\rangle d\mathbf{r}'. \quad 3.41$$

The average force, $\mathbf{F}(\mathbf{r}')$, is calculable from MD simulations using the finite differences method. However, the calculation and application of the Jacobian correction term is also required.³⁷⁻³⁹

UMBRELLA SAMPLING

Methods employing harmonic restraints or umbrella potentials are free from momentum quenching problems and do not forbid the use of constraints involving restrained atoms. Umbrella sampling, developed by Torrie and Valeau,⁴⁰ generates non-Boltzmann sampled simulations by adding an external biasing potential, focusing the sample on a range between r_i-r_o centred on r_o . The biasing potential is given by:

$$\omega(r_i; r_o) = \frac{1}{2} k (r_i - r_o)^2. \quad 3.42$$

This is done in order to enhance the sampling along the reaction coordinate, r , and is restricted to the range r_i-r_o to help improve the configurational sampling. The new harmonically modified Hamiltonian is written as

$$H(\mathbf{r}) = H(\mathbf{r}, \mathbf{p}) + \omega(r_i; r_o). \quad 3.43$$

Computational limitations generally make biasing simulations over the entire range of a reaction coordinate implausible. This can be overcome by using the windowing method, which biases different simulations with small overlapping reaction coordinate ranges. This is usually accomplished by taking the biasing potential to be the sum of the harmonic terms in the reaction coordinate space. Several simulations are run, each starting in different harmonic wells. Ultimately the harmonic biases are removed to obtain the unbiased probability distribution, $p(\xi)$, and potential, $U(\xi)$, and the results are recombined to obtain the Helmholtz free energy (or PMF)¹⁴

$$\Delta A(\xi) = -k_B T \ln \left[p(\xi) \exp \left(-\frac{U(\xi)}{k_B T} \right) \right]. \quad 3.44$$

ADAPTIVE UMBRELLA SAMPLING

Another form of a biasing potential method, which is regularly employed, is the Adaptive Umbrella Sampling method, first described by Mezei.⁴¹ In this method, the biasing potential is iteratively adapted based on the probability distributions of the reaction coordinates sampled during the previous simulations. Unlike conventional umbrella sampling, which only applies a bias to a small region along the reaction coordinate, adaptive umbrella sampling applies a bias along the entire range of the reaction coordinate (e.g., the distance, r). This biasing potential, $U(r)$, is calculated from the probability distributions of the previous simulations and is set to zero for the initial simulation. The calculation converges when uniform sampling, along the reaction coordinate, is observed. At this point the biasing potential is the exact inverse of the potential of mean force, Figure 3.3.

Now, using a generalized reaction coordinate (ξ), a simple description of the procedure is as follows.

For the initial simulation(s), set biasing potential, $U(\xi)$, equal to zero everywhere, then¹⁵

1. Run the simulation with applied potential, $U(\xi)$.
2. From the histograms of the reaction coordinate values sampled in the simulation, calculate the biased probability, $p(\xi)$.
3. Use Equation 3.44 to calculate the free energy, $\Delta A(\xi)$.
4. If the biased probability distribution, $p(\xi)$, is nearly uniform, then the free energy surface calculated in step 3 is the converged potential of mean force, PMF.
5. Otherwise, set the biasing potential as the negative of the free energy surface ($U(\xi) = -\Delta A(\xi)$) and return to step 1.

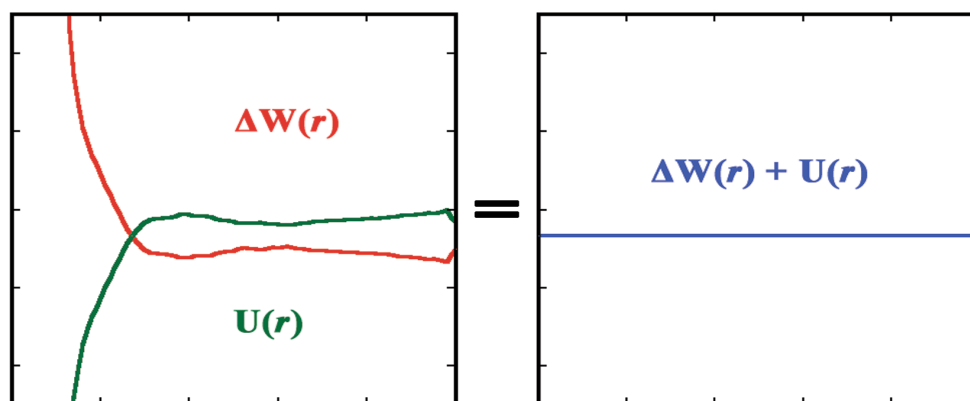


Figure 3.3: $U(r)$ is the negative of $\Delta A(r)$, and is applied to bias the simulation to regions that are high-energy and not sampled well. When $\Delta A(r) + U(r) = 0$, then $-U(r)$ represents the final $A(r)$.

In the first few simulations, there are often large regions of the reaction coordinate that are not sampled, resulting an infinite value for the free energy. To overcome this, they are set to equal the maximum value of the free energy in the regions that have already been sampled. This means, however, that when using the procedure described above sharp peaks and discontinuous jumps exist in the free energy surface and, hence, the applied potential. The forces that arise from the derivative of such an applied potential can, thus, be unrealistically large and varying for an MD simulation. This would cause instability in the trajectory and crash the simulation. By employing a smoothing procedure, discontinuous jumps can be removed and reduce the unrealistic size of the forces to more realistic, MD friendly forces.

METADYNAMICS

The metadynamics method was developed and introduced by Laio and Parrinello.⁴²⁻
⁴⁵ This method is similar to the Adaptive Umbrella Sampling, and can be implemented in multi-dimensions. The essence of the method lies in constructing the free energy surface from Gaussian terms that evolve dynamically based on the amount of sampling, for the reaction coordinates, during the simulation.

The metadynamics simulation restrains the simulation to a particular region of the reaction coordinate using inverted Gaussian potentials. The Hamiltonian for a multi-dimensional parameterization of the free energy surface using $(\xi_1, \dots, \xi_n)$ is modified to become:

$$H = H_o + \sum_i \frac{p_{\xi,i}^2}{m_{\xi,i}} - \sum_i \sum_{k=1}^{N_i} W_{i,k} \exp\left(-\frac{(\xi_i - \xi_{i,k})^2}{2\sigma_{i,k}^2}\right). \quad 3.45$$

Here along the reaction coordinate, ξ_i , there are N_i Gaussians, with heights, $W_{i,k}$, widths, $\sigma_{i,k}$, and each centred at $\xi_{i,k}$, which are added to the unbiased Hamiltonian. The heights, widths and centres are chosen to reach a compromise between accuracy and efficiency. Typically, for a singly parameterized free energy surface 200 or more Gaussians terms are required.⁴³ This number increases exponentially for higher dimensions, which makes the computations using metadynamics very 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 $\xi_{i,k}$ so that the Gaussians are optimally placed at 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.⁴² As with other

methods, there arises a Jacobian correction term due to the reaction coordinate dependent potential. This correction is often neglected in the calculation of 1D and 2D free energy surfaces,⁴³ but for higher dimensional surfaces this correction may not be as trivial and its consideration can no longer be neglected.

FREE ENERGY FROM ADAPTIVE REACTION COORDINATE FORCES

The *Free Energy from Adaptive Reaction Coordinate Forces* (or *FEARCF*), developed by Strumpfer and Naidoo,²⁵ is a flat histogram method⁴⁶ that calculates the biasing force from the biasing potential. It applies the biasing forces, resulting from the reaction coordinate definition, to only the atoms involved in the reaction coordinate. The atoms involved in the reaction coordinate are treated as a rigid body, which prevents chemically unmeaningful deformation of the molecules. Since the forces are applied directly to the atoms, there is no need for a Jacobian correction when using the FEARCF method.^{25, 47} In describing the deceptively simple association of aromatic molecules, such as benzene, we are required not only to describe the distance between the molecules but also the directionality of this interaction. The benzene dimer is a prototype for aromatic association and there is currently much debate in the literature about which is the most favoured orientation for an associated benzene dimer.⁴⁸⁻⁵¹ In this thesis, the FEARCF^{25, 47, 52} method is used to investigate the anisotropic association of the benzene dimer,⁴⁷ as well as of some aromatic amino acids pairs (i.e. PHE-PHE homodimer or PHE-TYR heterodimer, etc). First the 4D reaction coordinate frame is defined, Figure 3.4, the simulation(s) is then biased using FEARCF and, finally, the free energy is calculated using the *Weighted Histogram Analysis Method* (WHAM).

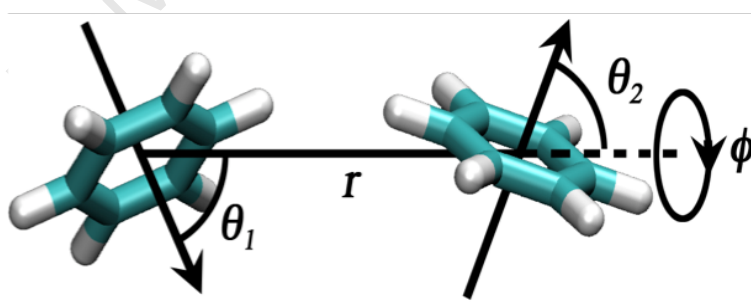


Figure 3.4: A multi-dimensional reaction coordinate chosen based on distance and angular relations that would yield a multi-dimensional free energy hyper-volume.

FEARCF is based on the adaptive umbrella sampling method of Mezei⁴¹ and makes use of multi-dimensional histograms to construct multi-dimensional probability distributions, from which the multi-dimensional free energy (hyper-volume) can be calculated. To increase the sampling in the high-energy regions of the reaction coordinates (ξ) an

adaptive force, $F(\xi)$, is applied along each reaction coordinate. The force, in the simple 1D case where $\xi = r$, is calculated from the biased umbrella potential

$$\mathbf{F}(r) = -\nabla_r U(r), \quad 3.46$$

where $U(r)$ is the inverse of the latest guess of the free energy, i.e. $W(r) = -U(r)$. This approach forces the molecules into poorly sampled regions of configurational (or conformational) space over a series of simulations, until near-uniform sampling over the reaction coordinate(s) is achieved. To improve the sampling efficiency many simulations should be performed, with each simulation using a different starting configurations and momenta including high-energy regions of the reaction coordinate space. The procedure initially begins with a zero biasing potential, and the system is allowed to evolve in the usual, unbiased manner. All subsequent simulations (or iterations) are biased using the biasing potential calculated from the histograms of previous simulations, which then produce new biased histograms. The resulting histograms, from all the simulations, can be combined using the *WHAM* method to calculate the free energy and further improve the efficiency of the calculation.

WEIGHTED HISTOGRAM ANALYSIS METHOD

The *Weighted Histogram Analysis Method (WHAM)*, developed by Kumar et al.,^{53, 54} is a technique whereby biased probability distributions, from several different simulations, are combined using computed weighting factors for each of the probability distribution histograms. This produces the best estimate for the actual unbiased probability distribution histograms. These biases can be used to reach a stage where the reaction coordinates are uniformly sampled during a simulation.

The final histogram and biasing potential can be used to calculate the sampling distribution and in turn the free energy surface. To combine the histograms from different simulation requires the unbiased histograms from each simulation. Combined by iteratively optimizing the their weights to achieve the unbiased probability distribution $p(r)$.

For N simulations with biasing potential, $U_i(r)$, the resulting histograms $n_i(r)$ and $N_i = \sum_{i=1}^N n_i(r)$, for each simulation ($i=1,...,N$), gives the equation for the total unbiased probability

$$p(r) = \frac{\sum_{i=1}^N n_i(r)}{\sum_{i=1}^N N_i \exp(\beta[F_i - U_i(r)])}, \quad 3.47$$

where $\beta = \frac{1}{k_B T}$. The F_i (the free energy) is calculated in the following way

$$F_i = -k_B T \ln \left\{ \sum_R p(r) \exp[-\beta U_i(r)] \right\}. \quad 3.48$$

Initially all F_i are set to zero. Equations 3.47 and 3.48 are iteratively updated until the maximum change in F_i between iterations is less than a certain tolerance, ε (i.e. $|F_i - F_{i-1}| \leq 0.001$). Once the WHAM algorithm has converged, the potential is smoothed for the initial simulations to avoid discontinuity in the potential energy surface, which would result in extremely large forces. The WHAM method allows for the combination of a large amount of biased probability distributions from different simulations. Thus, increasing the amount of phase space that can be sampled (see the preceding FEARCF section) in a single iteration.

3.4 References

1. Allen, M. P.; Tildesley, D. J., *Computer Simulation of Liquids*. Oxford University Press: Oxford, 1987.
2. Atkins, P. W.; De Paula, J., *Physical Chemistry*. 7th ed.; Oxford University Press: 2002.
3. Chandler, D., Thermodynamics, Fundamentals. In *Introduction to Modern Statistical Mechanics*, Oxford University Press: New York, 1987; pp 3.
4. Mcquarrie, D. A., *Statistical Mechanics*. University Science Books: Sausalito, 1981.
5. Chandler, D., Statistical Mechanics. In *Introduction to Modern Statistical Mechanics*, Oxford University Press: New York, 1987; pp 54.
6. Perrot, P., *A to Z of Thermodynamics*. Oxford University Press: Oxford; New York, 1998.
7. Kittel, C.; Kroemer, H., *Thermal Physics*. 2nd ed.; W. H. Freeman: 1980.
8. Hsu, W.; Barakat, R., Statistics and Thermodynamics of Luminescent Radiation. *Phys. Rev. B* **1992**, 46 (11), 6760.
9. Purcell, E. M.; Pound, R. V., A Nuclear Spin System at Negative Temperature. *Physical Review* **1951**, 81 (2), 279.
10. Toda, M., *et al.*, *Statistical Physics I*. Springer: Berlin Heidelberg, 1983.
11. Helrich, C. S., *Modern Thermodynamics with Statistical Mechanics*. Springer-Verlag: Berlin Heidelberg, 2009.
12. Wilde, R. E.; Singh, S., *Statistical Mechanics: Fundamentals and Modern Applications*. John Wiley & Sons, Inc.: Toronto, 1998.
13. Dirac, P., *Principles of Quantum Mechanics*. 4th ed.; Clarendon Press: Oxford, 1958.
14. Matthews, R. P. Ion Pairing in Aqueous Metal Sulfates and Platinum Group Metal Ammonium Solutions. Thesis, University of Cape Town, Cape Town, 2010.
15. Strumpfer, J. Computing Free Energy Hypersurfaces for Anisotropic Intermolecular Associations. Thesis, University of Cape Town, Cape Town, 2009.
16. Beveridge, D. L.; Dicapua, F. M., Free-Energy Via Molecular Simulation - Applications to Chemical and Biomolecular Systems. *Annual Reviews of Biophysics and Biophysical Chemistry* **1989**, 18, 431.
17. Straatsma, T. P., *et al.*, Holonomic Constraint Contributions to the Free-Energy Differences from Thermodynamic Integration from Molecular Dynamics Simulations. *Chem. Phys. Lett.* **1992**, 196 (3-4), 297.
18. Kollman, P. A., Free-Energy Calculations - Applications to Chemical and Biochemical Phenomena. *Chem. Rev.* **1993**, 93 (7), 2395.
19. Kirkwood, J. G., Statistical Mechanics of Fluid Mixtures. *J. Chem. Phys.* **1935**, 3 (5), 300.
20. Zwanzig, R. W., High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. *J. Chem. Phys.* **1954**, 22 (8), 1420.
21. Axelsen, P. H.; Li, D. H., Improved Convergence in Dual-Topology Free-Energy Calculations through Use of Harmonic Restraints. *J. Comput. Chem.* **1998**, 19 (11), 1278.
22. Jorgensen, W. L.; Thomas, L. L., Monte-Carlo Simulation of Differences in Free-Energies of Hydration. *J. Chem. Phys.* **1985**, 83 (6), 3050.
23. Jorgensen, W. L.; Thomas, L. L., Perspective on Free-Energy Perturbations Calculations for Chemical Equilibria. *J. Chem. Theory Comput.* **2008**, 4 (6), 869.
24. Lu, N. D., *et al.*, Improving the Efficiency and Reliability of Free Energy Perturbation Calculations Using Overlap Sampling Methods. *J. Comput. Chem.* **2004**, 25 (1), 28.
25. Strumpfer, J.; Naidoo, K. J., Computing Free Energy Hypersurfaces for Anisotropic Intermolecular Associations. *J. Comput. Chem.* **2010**, 31 (2), 308.
26. Ben-Naim, A., *Water and Aqueous Solutions*. Plenum Press: New York, 1974.

27. 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* (2), 1359.
28. Mezei, M.; Beveridge, D. L., Free Energy Simulations. *Ann. N. Y. Acad. Sci.* **1986**, *482* (Comput. Simul. Chem. Biomol. Syst.), 1.
29. Leach, A. R., *Molecular Modeling*. 2nd ed.; Longman: 2001.
30. Trzesniak, D., *et al.*, A Comparison of Methods to Compute the Potential of Mean Force. *ChemPhysChem* **2007**, *8* (1), 162.
31. Berkowitz, M., *et al.*, Sodium-Chloride Ion-Pair Interaction in Water - Computer Simulation. *Chem. Phys. Lett.* **1984**, *105* (6), 577.
32. Ghoufi, A.; Malfreyt, P., Calculations of the Potential of Mean Force from Molecular Dynamics Simulations Using Different Methodologies: An Application to the Determination of the Binding Thermodynamics Properties of an Ion Pair. *Mol. Phys.* **2006**, *104* (22-24), 3787.
33. Coluzza, I., *et al.*, Constrained Reaction Coordinate Dynamics for Systems with Constraints. *Mol. Phys.* **2003**, *101* (18), 2885.
34. Fixman, M., Simulation of Polymer Dynamics. I. General Theory. *J. Chem. Phys.* **1978**, *69* (4), 1527.
35. Go, N.; Scheraga, H. A., Use of Classical Statistical Mechanics in Treatment of Polymer-Chain Conformation. *Macromolecules* **1976**, *9* (4), 535.
36. Carter, E. A., *et al.*, Constrained Reaction Coordinate Dynamics for the Simulation of Rare Events. *Chem. Phys. Lett.* **1989**, *156* (5), 472.
37. Darve, E.; Pohorille, A., Calculating Free Energies Using Average Force. *J. Chem. Phys.* **2001**, *115* (20), 9169.
38. Den Otter, W. K., Thermodynamic Integration of the Free Energy Along a Reaction Coordinate in Cartesian Coordinates. *J. Chem. Phys.* **2000**, *112* (17), 7283.
39. Den Otter, W. K.; Briels, W. J., Free Energy from Molecular Dynamics with Multiple Constraints. *Mol. Phys.* **2000**, *98* (12), 773.
40. Torrie, G. M.; Valleau, J. P., Monte-Carlo Free-Energy Estimates Using Non-Boltzmann Sampling - Application to Subcritical Lennard-Jones Fluid. *Chem. Phys. Lett.* **1974**, *28* (4), 578.
41. Mezei, M., Self-Consistent Determination of the Non-Boltzmann Bias. *J. Comput. Chem.* **1987**, *68* (1), 237.
42. Ianuzzi, M., *et al.*, Efficient Exploration of Reactive Potential Energy Surfaces Using Car-Parinello Molecular Dynamics. *Phys. Rev. Lett.* **2003**, *90* (23), 238302.
43. Laio, A.; Parrinello, M., Escaping Free-Energy Minima. *Proc. Natl. Acc. Sci.* **2002**, *99* (20).
44. Martonak, R., *et al.*, Predicting Crystal Structures: The Parinello-Rahman Method Revisited. *Phys. Rev. Lett.* **2003**, *90* (7).
45. Micheletti, C., *et al.*, Reconstructing the Density of States by History-Dependent Metadynamics. *Phys. Rev. Lett.* **2004**, *92* (17).
46. Naidoo, K. J., Fearc a Multidimensional Free Energy Method for Investigating Conformational Landscapes and Chemical Reaction Mechanisms. *Chem. Sci. Chin.* **2011**, *54*, 1962.
47. Gamielien, M. R., *et al.*, Hydration Determined Orientational Preferences in Aromatic Association from Benzene Dimer Free Energy Volumes. *J. Phys. Chem. B* **2011**.
48. Sinnokrot, M. O.; Sherrill, C. D., High-Accuracy Quantum Mechanical Studies of Pi-Pi Interactions in Benzene Dimers. *J. Phys. Chem. A* **2006**, *110* (37), 10656.
49. Linse, P., Stacked or T-Shaped Benzene Dimer in Aqueous-Solution - a Molecular-Dynamics Study. *J. Am. Chem. Soc.* **1992**, *114* (11), 4366.
50. Linse, P., Orientation-Averaged Benzene Benzene Potential of Mean Force in Aqueous-Solution. *J. Am. Chem. Soc.* **1993**, *115* (19), 8793.

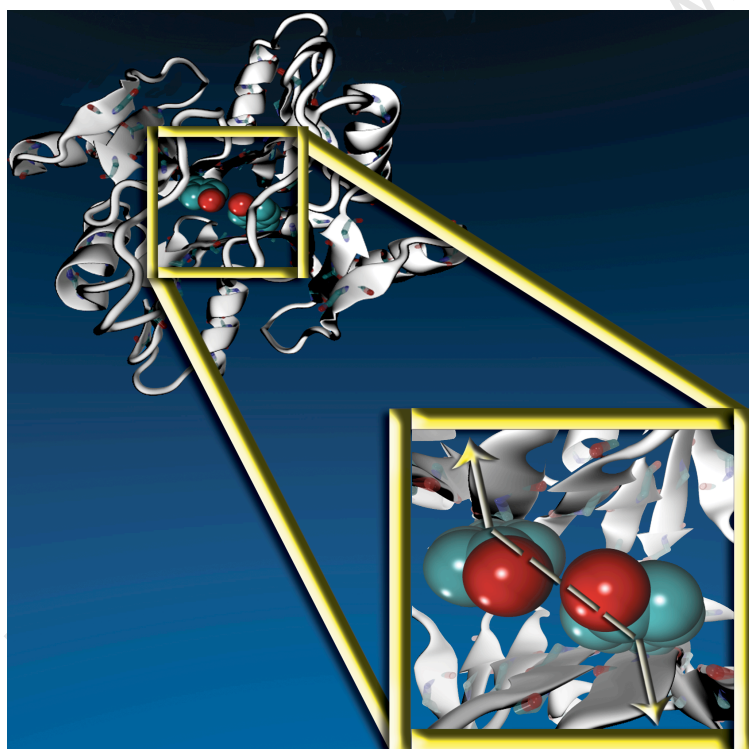
51. Jorgensen, W. L.; Severance, D. L., Aromatic-Aromatic Interactions: Free Energy Profiles for the Benzene Dimer in Water, Chloroform, and Liquid Benzene. *J. Am. Chem. Soc.* **1990**, *112* (12), 4768.
52. Barnett, C. B.; Naidoo, K. J., Free Energies from Adaptive Reaction Coordinate Forces (Fearcf): An Application to Ring Puckering. *Mol. Phys.* **2009**, *107* (8-12), 1243.
53. Kumar, S., *et al.*, The Weighted Histogram Analysis Method for Free-Energy Calculations on Biomolecules. I. The Method. *J. Comput. Chem.* **1992**, *13* (8), 1011.
54. Kumar, S., *et al.*, Method for Free-Energy Calculations Using Iterative Techniques. *J. Comput. Chem.* **1996**, *17* (10), 1269.

University of Cape Town

PART 1

INFLUENCE OF ORIENTATION ON AROMATIC ASSOCIATION

“... there can be no doubt the name of van der Waals will soon be amongst the foremost in molecular science.” – James Clerk Maxwell



Tyrosine association between the two subunits of the protein *FOSFOMYCIN RESISTANCE PROTEIN A*.

van der Waals interactions encompasses all the intermolecular forces, including dispersion (London) and electrostatic forces. Aromatic association exhibits a strong orientational bias due to these intermolecular forces. The orientational preference for aromatic association has been investigated using a variety of theoretical techniques (e.g. high-level quantum mechanical and free energy calculations), and experimental analyses (protein crystal structure analysis and various spectroscopic methods). In the first part of this thesis, aromatic association (in vacuum and solution) is explored using free energy hyper-volumes.

CHAPTER 4

THE BENZENE DIMER – PROTOTYPE FOR AROMATIC ASSOCIATION – SOLVATED

4.1 Introduction

Molecular association between aromatic molecules is at the heart of many important biochemical processes such as cellular communication and transport,¹ protein folding,²⁻⁴ the mechanism of crystallization,⁵ and liquid crystal phases.⁶ The particular case of two benzene molecules associating is often used as a central narrative for interior hydrophobic interactions in globular proteins that are responsible for their overall fold, π - π interactions that help to stabilize base pair stacking in DNA,⁷ and that inform the packing of organic crystals and liquid crystalline materials.⁶ In all of these important phenomena it is striking that the aromatic interactions exhibit a strong orientational bias. Arguably the benzene dimer is the simplest model that is most useful to chemists, chemical biologist and chemical physicists when attempting to understand the intermolecular interactions emanating from a range of fundamental dispersion effects.

The benzene dimer is weakly associating⁸⁻¹⁰ and has been the subject of considerable study because of its simplicity and its effectiveness as a model for complex aromatic interactions.⁸⁻¹¹ Various studies have shown that it adopts different stable configurations in different phases and environments. The geometry of an aromatic pair of molecules may be reduced to 4 variables that describe the distance and relative orientation of each molecule. In the case of the benzene dimer, vectors normal to each benzene plane and the centre of mass of each benzene molecule are used. The 4 variables are then defined (Figure 4.1) as the distance of the interconnecting vector between the two benzene centers (r), the angle between each benzene normal vector and interconnecting vector (θ_1 , θ_2) and the relative angle between these vectors (ϕ). Using these parameters we are able to distinguish between previously identified

minimum energy configurations that are Sandwich (S), T-shaped, (T), V-shaped (V), Bent T-shaped (BT), Side-by-Side (SbS) and Parallel Displaced (PD) (Figure 4.1).

Gräfenstein and Cremer comprehensively mapped out the potential energy surface (PES) of the benzene dimer employing a quasi-self-consistent-field dispersion-corrected density-functional theory formalism (QSCF-DC-DFT).¹² They distinguished between minimum PD and T configurations by specifying when the H-C $\cdots$ C-H fragment of the upper benzene was located exactly over the H-C $\cdots$ C-H fragment of the lower benzene (AoA), or whether it had an atom-over-bond (AoB) or bond-over-bond (BoB) arrangement. In addition, they specified when the benzene rings were edge-over-plane (EoP) and edge-over-edge (EoE). This led to the identification of 16 stationary points on the surface that, when ignoring the relative orientation of the benzene C6 symmetry axis, can be conflated to correspond with the S(4), PD (4), B(2), T(5) and SbS(1) geometries shown in Figure 4.1.

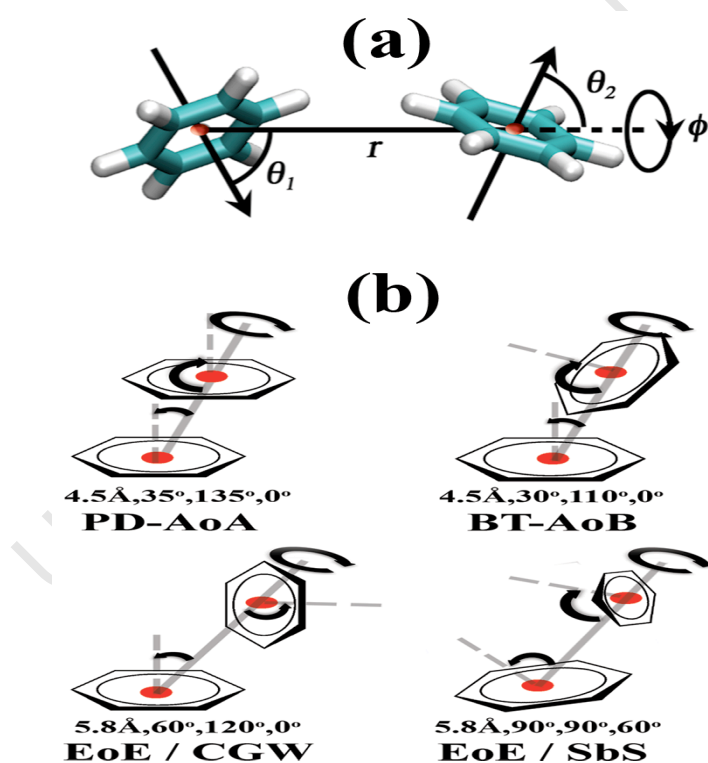


Figure 4.1: (a) The reaction coordinate definitions in the centre, and the various benzene dimer configurations using Gräfenstein and Cremer nomenclature,¹² and (b) the various configurations found from our free energy 2D surfaces.

The search for the definitive benzene dimer global minima geometry is important in the context of understanding dispersion interactions and the electronic arrangements within aromatic systems. However, our interests here are focused on macromolecular conformations that are driven by the thermodynamics of the hydrophobic association of

aromatic groups. It is for this reason that we seek to understand the effect of water on the relative orientation of benzene molecular pairs. To this end, we present a converged multi-dimensional free-energy volume, $W(r, \theta_1, \theta_2, \phi)$, for two benzene molecules in vacuum and solution. These free energy volumes are highly informative as they produce not only relative stationary points, that correspond with configurations obtained with *ab initio* methods, but the configurational pathways linking these favoured geometries. From this we gain insight into the delicate interplay between solvent-solvent, solvent-solute and solute-solute interactions that lead to the configurational preferences seen in the association of aromatic molecules in solution.

4.2 Methods

All simulations were performed using our modified version of *CHARMM* v33b2¹³ that incorporates the previously developed Free Energy from Adaptive Reaction Coordinate Forces (*FEARCF*), which is a flat histogram method.¹⁴⁻¹⁶ The reaction coordinate, ξ , space is a discretized 4D grid where there are 4 independent reaction coordinates $\xi_1 = r$, $\xi_2 = \theta_1$, $\xi_3 = \theta_2$, $\xi_4 = \phi$. The sampling frequency for a bin site is recorded for each simulation and 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 multi-dimensional 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 iteration's trajectory away from previously sampled volumes. The entire reaction coordinate space is equally sampled when the biasing forces are derived from the true potential of mean force (*PMF*). The forces are applied to the Cartesian coordinates and, therefore, the *FEARCF* method does not need logarithmic Jacobian corrections.¹⁷⁻²⁰

The system Hamiltonian is modified by adding the driving potential, $U(\xi)$, derived from the population density of the sampled volume, to the unbiased Hamiltonian, H^o , to give

$$H(\xi) = H^o + U(\xi). \quad 4.1$$

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

$$p(\xi) = Cp'(\xi) \exp\left(\frac{U(\xi)}{k_B T}\right), \quad 4.2$$

where C is the normalization constant, k_B is the Boltzmann constant, and T is the temperature at which the simulation was carried out. The *PMF*, $W(\xi)$, is then calculated from the probability distribution of the reaction coordinate of all trajectories up to that point and is given by

$$W(\xi) = -k_B T \ln p(\xi). \quad 4.3$$

At each step of the simulation, the driving forces for each ξ_i are applied to the atoms involved in the definition of the molecular vectors (for example as shown in Figure 4.1). For the 4D hyper-volumes, the forces are calculated as follows:

$$F(\xi) = -\sum_{i=1}^4 \frac{\partial U(\xi)}{\partial \xi_i}. \quad 4.4$$

The force, $F(\xi)$, arising from the driving potential is calculated by treating the molecules, which were used in the reaction coordinate definitions, as rigid bodies. This ensures that during the simulation the atoms within each of the benzenes do not move apart and deform due to differing accelerations that arise from the driving potential. Consequently, when the resultant acceleration for a benzene molecule is calculated as a whole, the forces on the atoms are determined such that they experience the same resultant acceleration.

For the first *FEARCF* iteration $U(\xi) = 0$ that then evolves progressively along subsequent iterations to eventually mirror $W(\xi)$. To achieve this each of the iterations were comprised of several sets of relatively short simulations, which were initiated from configurations of dimers that occupy significantly different points in reaction coordinate and momentum space. This was followed by an increase in the simulation times to reach convergence (sampling ratio for configurations of at least 1:10 for the most probable to least probable).

The 1D free energy can be deconvoluted into an entropy and internal energy (enthalpy) of association using the “finite difference” method.^{21, 22} The calculation of the 2D ($W(r, \theta_1)$, $W(r, \phi)$, $W(\theta_1, \phi)$ and $W(\theta_1, \theta_2)$) and 3D ($W(r, \theta_1, \phi)$ and $W(\theta_1, \theta_2, \phi)$) Boltzmann averaged plots were produced to identify configurations from the 4D $W(r, \theta_1, \theta_2, \phi)$ hyper-volume, taking into account that there can be significant loss of information when reducing the dimensionality of the free energy hyper-volumes.²³

4.2.1 The *FEARCF* method applied to aromatic association

The vacuum *FEARCF* simulations were carried out using the *Velocity-Verlet2* (vv2) integrator with a time step of 1 fs,²⁴ with the temperature maintained at 298.15 K using the *TPCONTROL* method (based on the Nose-Hoover thermostat) with *TAU* set at a value of 0.1.²⁴ Non-bonded cutoffs of 10, 12 and 14 Å with potential and force switching functions were used. A harmonic restraining potential (a linear force) was applied to the edges of the distance reaction coordinate, ensuring the benzenes molecules would not drift off resulting in chemically meaningless sampling of vacuum configurational space that is far outside of the dimer electrostatic interaction field (i.e. $r > 16$ Å). Each of the iterations was comprised of 8 simulations with different initial configurations and velocities. A total of 50 iterations were run, where the first 5 iterations comprised sets of 100 ps simulations, the following 15 iterations comprised sets of 300 ps simulations, and last 30 iterations comprised sets of 500 ps simulations. The simulation time for the entire *FEARCF* experiment was 160 ns (i.e. 20ns per simulation). We used a multi-dimensional implementation of *WHAM*²⁵ to combine the data from the all simulations in the set at iteration n and preceding sets from iterations $i = 1, 2, \dots n-1$ to obtain the free energy hypersurfaces. The data was weighted using free energy weighting factors (f_i) described previously¹⁴ and in Chapter 3, section 3.3.2. The *WHAM*²⁵ equations were applied iteratively until the maximum difference between the previous and current iteration weighting coefficients ($\max|f_i^i - f_i^{i-1}|$) was less than 0.001. The accuracy of the free energy surfaces is a factor 100 less accurate than the tolerance chosen.²⁶ We are therefore able to distinguish between free energies that are 0.1 kcal/mol apart.

The solution *FEARCF* simulations were performed using the TIP4P-Ew²⁷ water model in a periodic cubic box of length 34.1 Å. Each simulation was run using an *NVT* ensemble with a constant temperature of 298.15 K, maintained by the Nose-Hoover thermostat,^{28, 29} and leapfrog integrator algorithm.^{30, 31} Here we use double the number of simulations per iteration compared with the vacuum case (i.e. 16 simulations). Unlike the vacuum experiment, no harmonic potential was used to bind the distance component of the reaction coordinate as this lead to an artificial periodic sampling. The Particle Mesh Ewald (*PME*) method^{32, 33} was used to estimate the long-range interactions, with a grid size of 36 x 36 x 36 and a kappa value of 0.34 with a 6th order B-Spline interpolation. Each of the simulations was initiated with a unique dimer configuration and compliment of molecular velocities in the simulation box. A total of 30 iterations were required to reach convergence. The first 3 iterations comprised sets of simulations of 100 ps in length, the next 17 iterations comprised sets of 300 ps in length, and the last 10

iterations comprised sets of 500 ps in length. This translated into a total of 160 ns (10 ns per simulation). The *WHAM*²⁵ procedure was used to combine the data from the various simulations as described above.

4.2.2 Reducing the dimensionality of the free energy volume

Boltzmann averaged 2D surface plots were calculated for $W(r, \theta_1)$ and $W(r, \phi)$ (where $r \leq 10.0$ Å), and $W(\theta_1, \phi)$ and $W(\theta_1, \theta_2)$ (where $r \leq 6.5$ Å) were taken over the respective remaining parameters (θ_2, ϕ) , (θ_1, θ_2) , (r, θ_2) and (r, ϕ) from the 4D $W(r, \theta_1, \theta_2, \phi)$ volumes. It must be noted that in some instances, there is significant loss of information when reducing the dimensionality of the free energy surfaces. This has been recently illustrated for the reduction of the 3-, to 2- and then to 1D ring pucker free energy surfaces.²³ Unambiguous geometry assignments for molecular pairs are best made from the complete $W(r, \theta_1, \theta_2, \phi)$ free energy hyper-volume. However, in the case of dimers, a well-chosen 3D free energy volume can adequately provide this information. We calculated $W(r, \theta_1, \phi)$ and $W(\theta_1, \theta_2, \phi)$ where the benzene molecules were in each other's attractive field (i.e. ≤ 6.5 Å) by taking Boltzmann averages from the 4D anisotropic free energy hyper-volumes over the remaining orientational parameters θ_2 and r , respectively.

4.2.3 Entropy and internal energy contributions to the free energy

The variation in the change in the Helmholtz free energy with temperature (at constant number of particles and volume) yields the following relation to the entropy:

$$\Delta S = - \left(\frac{\partial \Delta A}{\partial T} \right)_{N,V}, \quad 4.5$$

and, therefore, we calculate the entropy of association using a “finite-difference” method. This is accomplished by determining the 1D *PMFs* at 3 well-separated temperatures.^{21, 22} The 1D solution *PMFs* were run using the same conditions as described for the solution 4D free energy hyper-volume calculation. The 3 temperature simulations were carried out at 278.15, 298.15 and 318.15 K. The entropy was then calculated as

$$\Delta S = - \frac{\Delta A(T + \Delta T) - \Delta A(T - \Delta T)}{2\Delta T}. \quad 4.5.1$$

The uncertainty in ΔS using the finite difference method is inversely proportional to ΔT , implying a larger temperature difference will yield a more precise estimate of ΔS .³⁴ Since

we wish to know the change in entropy of association of benzene dimers in solution (i.e. in water) at 298.15 K, we take the maximum temperature difference of 20 K, which satisfies a lower temperature bound that is above the maximum density temperature of water (277.15 K).

The change in the internal energy, which can be related to the enthalpy of the system (using $\Delta H = \Delta E + p\Delta V$), was calculated using the Helmholtz free energy equation

$$\Delta A = \Delta E - T\Delta S. \quad 4.6$$

The Helmholtz and Gibbs free energies are uniquely related by the work function (where $p\Delta V \sim 0$, for NVT simulations).

4.3 Results And Discussion

The complete 4D free energy volumes $W(r, \theta_1, \theta_2, \phi)$ were computed for the benzene dimer in vacuum and solution. Boltzmann averaged 2D surface plots were then calculated for $W(r, \theta_1)$, $W(r, \phi)$, $W(\theta_1, \phi)$ and $W(\theta_1, \theta_2)$. The 2D free energy surfaces for the dimer in vacuum (Figure 4.2a-d) are convenient to visualize and they contain essential information about the transformation between configurational minima. Unambiguous geometry assignments for molecular pairs are best made from the complete $W(r, \theta_1, \theta_2, \phi)$ free energy volume. However, in the case of dimers, a well-chosen 3D free energy volume can adequately provide this information. We calculated $W(r, \theta_1, \phi)$ and $W(\theta_1, \theta_2, \phi)$ where the benzene molecules were in each other's attractive field (i.e. $r \leq 6.5$ Å), Figure 4.2e and f. When these are viewed together with the 2D volumes provide an accurate account of the dimer configurational free energy landscape emerges.

4.3.1 Free energy hyper-volume analysis in vacuum

This investigation for the benzene dimer in vacuum reveals a collage of BT, tilted BT and PD configurations, within 0.05 kcal/mol from each other as can be seen from the large inner contour volumes, that contribute to the large minimum energy well that stretches from 3.8 Å through to 5.8 Å (Figure 4.2a and b and Table 4.1). This is in agreement with high level *ab initio* studies, which found BT and PD to be the lowest energy geometries.³⁵ 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 we are able to calculate the 6D *PMF* using the

methods we present here, it demands computational resources out of reach of present day midrange computational technologies and is therefore impractical.

The reduced, however, useful 4D $W(r, \theta_1, \theta_2, \phi)$ description does suffer from some loss of configurational information due to rotations about the vectors through the C6 symmetry axes. Therefore, we cannot definitively assign configurations derived from the free energy volumes to be AoA, AoB or BoB. This is not a major disadvantage as Gräfenstein and Cremer showed that these bond and atom alignments are close in energy for specific (PD, T and EoE) orientational configurations.¹² The global minima takes on an approximate tilted PD (AoA)/BT (AoA) form ($r = 4.5$, $\theta_1 = 54/126$, $\theta_2 = 27/153$, $\phi = \pm 180$) that is similar to the low energy PD-til-AoB and global minimum T-til-AoB configurations found by Gräfenstein and Cremer.

The global minimum found from the free energy surface is only 0.04 and 0.08 kcal/mol lower in energy than the PD ($r = 4.5$, $\theta_1 = 36/144$, $\theta_2 = 145/35$, $\phi = 0$) and V-Shape (V)/Cogwheel (CW)/Edge-over-Edge (EoE) ($r = 4.8$, $\theta_1 = 54/126$, $\theta_2 = 153/27$, $\phi = \pm 180$) configurations, respectively. The vacuum free energy relationship between the lowest energy benzene dimer configurations calculated here, from a classical force field, correspond well with the findings of Gräfenstein and Cremer. They showed that all T and PD stationary points are within 0.12 kcal/mol from each other on the QSCF-DC-DFT potential energy surface (PES).¹²

Table 4.1: The benzene dimer configurations found in vacuum, arranged from lowest to highest free energy.

Configuration Name	Reaction Coordinates ($r; \theta_1; \theta_2; \phi$)	ΔA (kcal/mol)
Tilted-PD(AoA)/BT (AoA)	4.0 Å; 150/30°; 30/150°; 0/180°	0.00
PD	4.0 Å; 150/30°; 30/150°; 0/180°	0.04
V/CW/EoE	4.0 Å; 150/30°; 30/150°; 0/180°	0.08

We performed a molecular mechanics optimization starting from the PD structure and found a minimum at 3.8 Å, which corresponds to the Gräfenstein and Cremer PD-AoA configuration. The same calculation was initiated from a T configuration, and this time the T-Til-AoB was found at 4.8 Å and is the lowest energy configuration, therefore agreeing with Gräfenstein and Cremer's *ab initio* PES results. Moreover, the free energy derived centre-of-mass-to-centre-of-mass distances are in strong agreement with those of vacuum experiments of 4.96 Å.³⁶ The vacuum free energy results using classical

CHARMM force field parameters show a strong correlation with the minima configurations, and their relative energetic preference, as found from high-level QM calculations.^{11, 35, 37} The accuracy of the molecular mechanics potential energy function is therefore validated as it allows for discrimination between low energy conformers as observed from detailed quantum mechanical simulations.

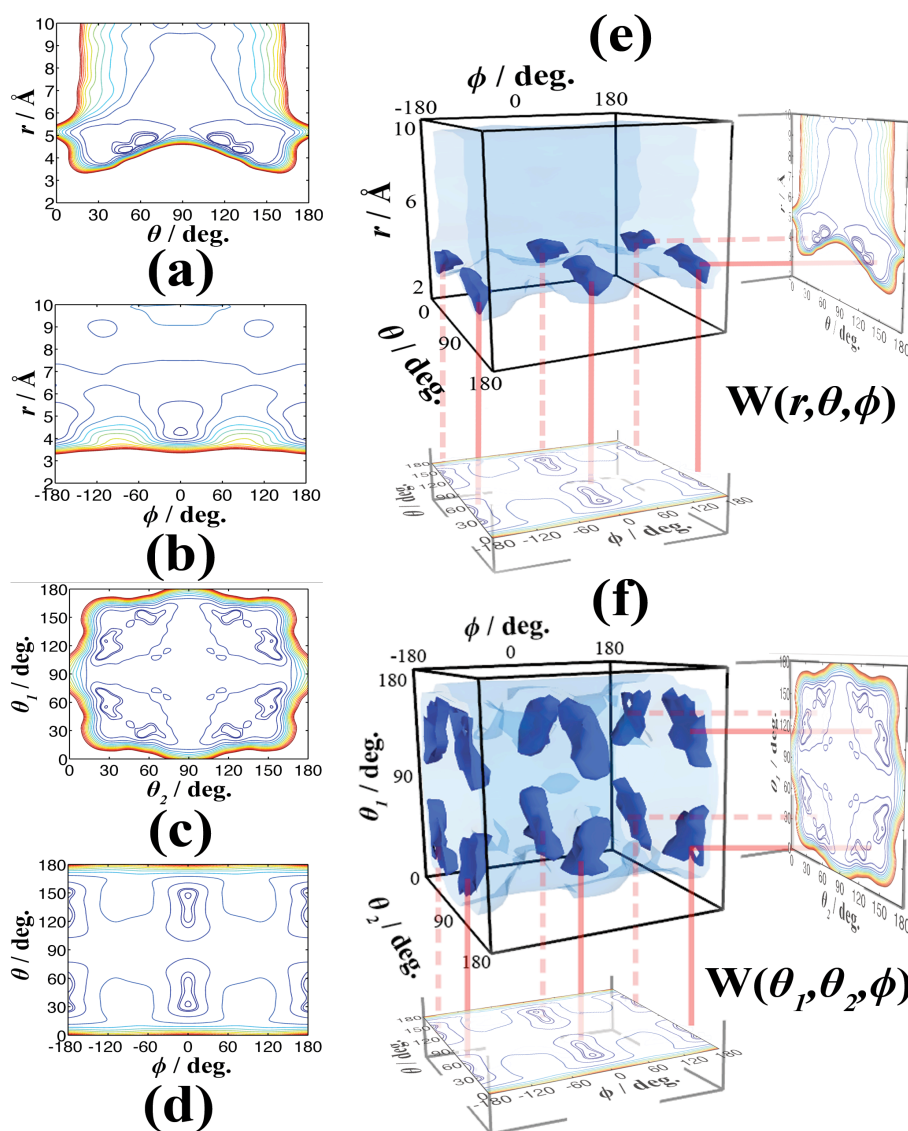


Figure 4.2: 2D contour (a-d) and 3D (e-f) volume plots for the benzene dimer in vacuum. The 2D contour plots are contoured at 0.5 kcal/mol from a minimum of 0.0 kcal/mol up to a maximum of 5.0 kcal/mol. The 3D volume plots have an inner contour of 1.0 k_BT and outer contour of 3.0 k_BT.

The broadness of the low energy (≤ 0.5 kcal/mol) configuration well (Figure 4.2a-d) suggests that the mix of the PD, BT and V (or EoE) are in dynamic equilibrium. From the Boltzmann averaged $W(r, \theta_1)$, $W(r, \phi)$, $W(\theta_1, \phi)$ and $W(\theta_1, \theta_2)$ surfaces it can be seen that the minima are symmetrical about $\theta = 90$ and $\phi = 0$, and that they display no significant

energy barriers between canonical configurations. Therefore, the benzene dimer has access to a variety of configurations in vacuum.

The basin of the free energy minima (Figure 4.2a-d) is well within the thermal energy envelope, which leads to a random ensemble of low energy configurations in vacuum. This result opposes the notion that there are specific energetically favourable benzene dimer orientations that bias the association of these molecules, since the geometries can easily vary between the various minimum energy configurations. The orientationally biased aromatic associations observed in protein structures are, therefore, not explained by the nature of the potential energy function.^{2, 38} In an analysis of interaction energies performed on optimized coupled cluster structures Sherrill et al. concluded that two closely associating benzene molecules either maximizes individual contributions from the van der Waals and electrostatic interactions or optimizes the balance between these two sources of attraction.³⁷ Using symmetry adapted perturbation theory combined with density functional theory (SAPT-DFT), Podesszwa et al. showed that dispersion is the dominant contribution to the PD configuration's stability. In the BT case electrostatics, in particular quadrupole-quadrupole, have been proposed to stabilize that configuration.³⁵ Vacuum experiments such as Raman-vibronic double resonance spectroscopy³⁹ and Fourier transform microwave spectroscopy³⁶ attempt to deconvolute very complex spectra for the dimer and have concluded that the dimer is characterized by two inequivalent benzene sites, suggestive of T-shape type structures. The free energy volumes for the vacuum benzene dimer shown here reveal more T and T-like structures than parallel or stacked structures in the low energy basins. The experimental observation that there is much uncertainty in the dimer configurations along with the very minor differences in the dispersion and electrostatic terms derived from computed couple cluster methods correlates⁴⁰ well with the ensemble of PD, BT and V/CGW (or EoE) configurations observed from the free energy vacuum simulations. Our results suggest a 2:1 stoichiometric ratio between T (and T-like) and PD configurations. This may possibly explain the discrepancy between experiment and high-level static quantum mechanical calculations.

4.3.2 Free energy hyper-volume analysis in solution

In solution, unlike in vacuum, the thermodynamically distinct low energy geometries are separated by at least $1.0 k_B T$ from each other; as can be seen from the ~ 0.59 kcal/mol contour volumes coloured in dark blue, Figure 4.3e and f. These are well-defined low energy regions that are far fewer and smaller in number compared

with the vacuum case contoured at the same energy levels, Figure 4.2e and f. The global minimum observed from the $W(r, \theta_1)$, $W(r, \phi)$, $W(\theta_1, \phi)$ and $W(\theta_1, \theta_2)$ 2D free energy surfaces is what Gräfenstein and Cremer referred to as EoE¹² ($r = 5.8$, $\theta_1 = 63/117$, $\theta_2 = 117/63$, $\phi = -53$) and ($r = 5.8$, $\theta_1 = 63/117$, $\theta_2 = 63/117$, $\phi = 53$).

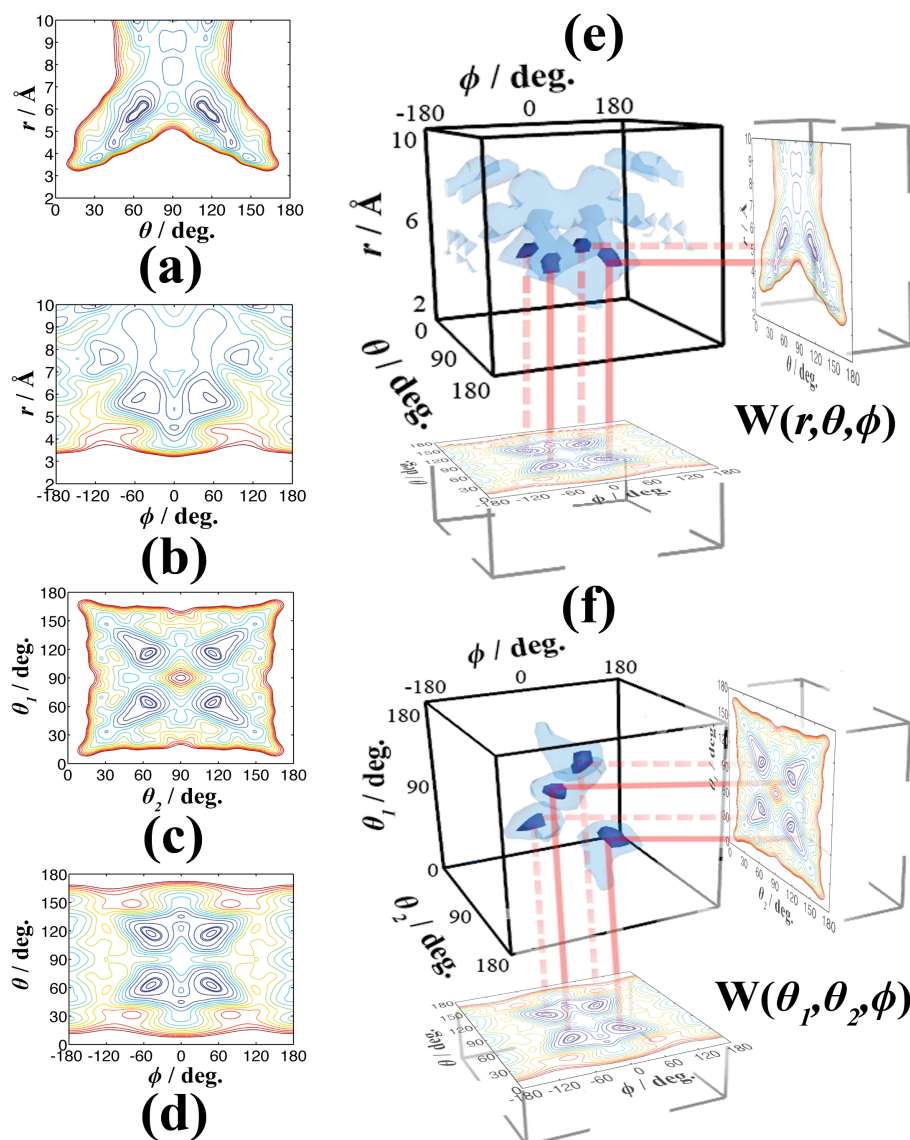


Figure 4.3: 2D contour (a-d) and 3D (e-f) volume plots for the benzene dimer in solution (TIP4P-Ew water).

The 2D contour plots are contoured at 0.5 kcal/mol from a minimum of 0.0 kcal/mol up to a maximum of 5.0 kcal/mol. The 3D volume plots have an inner contour of 1.0 $k_B T$ and outer contour of 3.0 $k_B T$.

They found this configuration to be nearly 1.0 kcal/mol higher in energy on their PES than the global minimum. The EoE configuration is similar in form to the cogwheel (CW) structures observed by Thornton et al., which is the predominantly observed configuration for phenylalanine dimers in proteins.³⁸ The PD-tilted-AoB ($r = 5.8$, $\theta_1 = 45/135$, $\theta_2 = 45/135$, $\phi = 0$) configurations are found to be 0.45 kcal/mol higher in energy than the global minimum on the solution free energy surface (Table 4.2).

Table 4.2: The benzene dimer configurations found in solution, arranged from lowest to highest free energy.

Configuration Name	Reaction Coordinates (r ; θ_1 ; θ_2 ; ϕ)	ΔA (kcal/mol)
EoE	5.8; 63/117; 117/63; ± 53	0.00
PD-tilted-AoB	5.8; 45/135; 45/135; 0	0.45

Nonetheless, the PD-tilted-AoB and EoE configurations are separated from each other by less than $1.0 k_B T$ allowing for regular interconversion between them. As the benzene molecules move out of each other's potential energy field, where specific configurations PD-tilted-AoB and EoE are favoured, a solvent shared minimum appears at $r = 7.5 \text{ \AA}$. This minimum found at larger benzene separations has been observed by others as a very weak minima in 1D $W(r)$ PMFs.^{10, 11} The orientational preference for the two benzene molecules at this distance is an SbS arrangement, which could be interpreted as a distant EoE configuration. The competing attractive forces between water, water and benzene and the benzene molecules result in the formation of a solvent shared minimum.

This secondary minimum observed at $7.5 \text{ \AA}$, which is not present in the vacuum free energy curves, confirms that the occurrence of this configuration is due to the solvent effect. The free energy for benzene dimer in vacuum at $7.5 \text{ \AA}$ is found to be > -0.2 kcal/mol, implying that there is very weak interaction occurring at this separation distance. However, in solution there are two additional intermolecular interactions that occur, viz. benzene-water and water-water interactions. We find that the benzene-water interaction is favourable and would account for the strong first solvation shell (indicated by the high probability of water) around each isolated benzene molecule (i.e. for benzene-benzene separation distances $> 10 \text{ \AA}$). The benzene-benzene and benzene-water interactions are found to be nearly isoenergetic. This not only allows the water interact with an individual benzene molecule, but also destabilize the benzene contact pair such that the global free energy minimum has been distorted toward an EoE configuration. As the separation increases, the benzene-benzene interaction diminishes allowing for the water to interact with benzene, thereby increasing the water probability in the first solvation shell. At a separation distance of $6.0 \text{ \AA}$, each benzene molecule begins to develop its own first solvation shell; and, while, the benzene dimer interaction has significantly decreased to < -0.2 kcal/mol, there are now more favourable benzene-water interactions. At $7.5 \text{ \AA}$, the interactions are balanced such that the water bound to benzene is strong enough to remain bound, while also interacting

with the adjacent benzene's first solvation shell and the bulk water. This solvent effect is also the reason why the global free energy minimum configuration is EoE and not PD or T-like. This phenomenon is predominantly found in charged systems,⁴¹ but has been found by others for the benzene dimer.^{8, 10, 11} We find that the SbS minima is 1.29 kcal/mol higher in energy than the close contact global minimum configuration found at $r = 5.8 \text{ \AA}$.

While these configurations are energetically distinct from each other (Figure 4.3a and b), the energy barriers separating them are less than $3.0 k_B T$ and should result in rapid and frequent interconversion between EoE, PD-tilted-AoB, BT-AoA, PD-AoA, and SbS configurations at room temperature. The BT, V/CW and SbS arrangements that appear have been conjectured from 6D deconvolution calculations performed on neutron diffraction benzene solvent structure factor data.⁴² Although the energy barriers separating dimer configurations are more than 3-fold greater in solution ($< \sim 3.0 k_B T$) than they are in vacuum ($< \sim 1.0 k_B T$), they remain very low and would make the observation of specific dimer configurations experimentally very difficult.

4.3.3 Thermodynamics analysis

The interaction in vacuum is mainly driven by a large enthalpic contribution to the free energy, as found by Podeszwa et al from their PES of the benzene dimer obtained using SAPT-DFT.³⁵ The thermodynamic data for the solution case, i.e. the change in free energy ΔA , internal energy ΔE and entropy ΔS , is shown in Figure 4.4a.

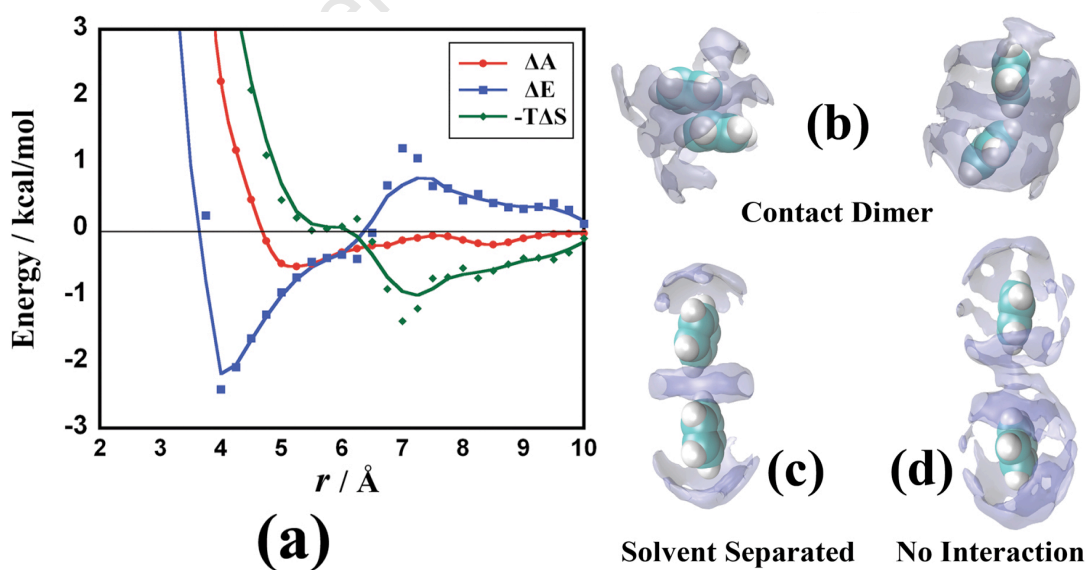


Figure 4.4: The thermodynamic curves in solution (a) and water probability density (b-d) profiles for the benzene dimer.

To obtain the thermodynamic data, three *PMFs* were calculated (as a function of the distance between the benzenes' centres of mass) at 278.15 K, 298.15 K and 318.15 K. The 1D distance *PMF* reveals a global minimum at a separation of $r = 5.8 \text{ \AA}$, which is consistent with previous studies^{10, 11, 43} and with the multi-dimensional anisotropic free energy of association *PMFs* shown in Figure 4.3. However, the thermodynamic analysis shows that while the contact ($r = 5.5 \text{ \AA}$) EoE dimer configuration (along with the PD and BT configurations) are enthalpically favoured (Figure 4.4a, blue curve), the location of the global free energy minima is not at the lowest attractive energy point at $4.0 \text{ \AA}$ but closer to the *local* entropy minima at $5.5 \text{ \AA}$ (Figure 4.4a, green curve). The benzenes are driven toward each other by the attractive interactions of their potential fields and the hydrophobic effect. While, the separation distance of $5.5 \text{ \AA}$ (compared with the potential energy minimum at $4.0 \text{ \AA}$) of the close contact association is determined by the *local* entropy of the benzene molecules. Loosely interpreted, this means that while there are external forces and attractive internal forces driving the association of the benzene in solution, the orientation of the benzenes, in relation to each other, is determined by the extent of the rotational freedom that the individual benzenes may have at $5.5 \text{ \AA}$.

The broad appearance of the $W(r)$ global minimum well indicates that several configurations are contributing to it. This is verified by the 2D surfaces and 3D volumes (Figure 4.3), where the nature of the contributing structures have been described above. Minor but important points of inflection are observed for both the entropy and internal energy curves at a distance of approximately $6.0 \text{ \AA}$. These points of inflection suggest that, for the benzene dimer, the enthalpically dominated contact dimer association character diminishes and that the underlying drive in free energy of the solvent-shared configuration is entropically favoured. It may therefore seem that this is the critical distance at which a close contact benzene dimer in solution may be defined (at $6.0 \text{ \AA}$). This suggests that the entropy (of associating benzenes in solution) is driving the initial association of the benzene dimer in solution. The free energy of association of benzene-water is weaker than that of the benzene dimer at $6.0 \text{ \AA}$, Figure 4.5. However, the former is dominant at distances $< 4.0 \text{ \AA}$. As the benzene molecules approach each other, taking on the characteristics of the solvent-shared minimum configuration, there is a decrease in water probability around the dimer (Figure 4.4b-d), i.e. a decrease in water density, as shown by a minimum for entropy curve, Figure 4.4a.

The decrease in water probability (Figure 4.4b-d) suggests that the waters are released causing an increase in entropy, which results in the observed gain in the entropic contribution to the free energy. A similar observation was made for methane molecules

where the association of 3 methane molecules was driven entropically at the solvent-shared minimum and enthalpically favoured at the contact minimum.

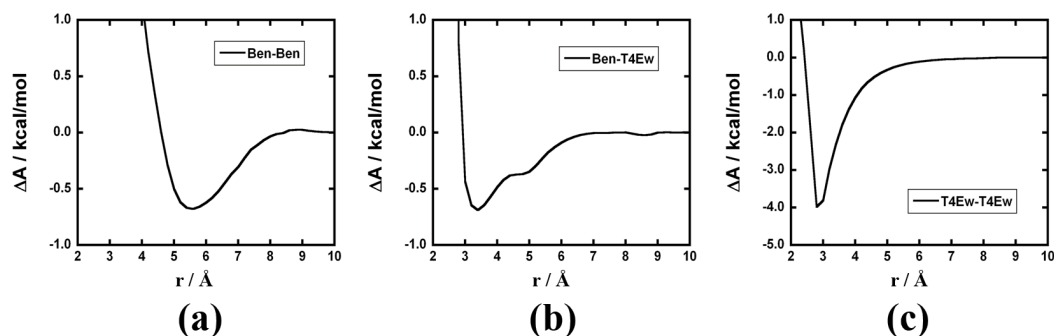


Figure 4.5: The various interactions between (a) benzene-benzene, (b) benzene-water (TIP4P-Ew) and (c) water (TIP4P-Ew)-water (TIP4P-Ew) molecules, in vacuum.

4.5 Conclusion

In this study, we have shown that the thermodynamics of benzene association in solution is central to their relative orientation. Our results agree with high-level quantum mechanical studies that the potential energy and enthalpy play an important role in the association of benzene dimers. However, we find that these factors do not determine the orientation of these molecules on close contact in solution. Specifically, we find that there is large variation in configurations that contribute to the vacuum free energy wells of the benzene contact pair, while the close contact configurations of benzene dimers are fewer and more energetically distinct from each other in solution. The EoE configuration is most likely in solution and appears to evolve from the entropically favoured SbS solvent-shared configuration. Benzene association in water therefore balances three effects: (1) a reduction in the entropic loss of waters adjacent to the benzene molecules; (2) the freedom of the benzenes to rotate in solution; and (3) the energetic gain from the interaction between higher order terms in the electrostatic multipole expansion and dispersion attraction between the benzene molecules. Although the association of the benzene dimer is accepted to be the result of the hydrophobic effect, the relative orientation of the benzene molecules is a result of maximizing the benzene-benzene entropic contribution to the free energy.

In the next chapter, an analysis of pairs of biologically relevant aromatic amino acid residues is investigated namely the PHE-PHE, TYR-TYR and PHE-TYR pairs. These pair interactions are also investigated for 5877 protein crystal structures and compared to the solution free energy hyper-volumes obtained from the simulation. The free energy, entropy and internal energy analysis, obtained from the simulations, provides a

thermodynamic description of the PHE-PHE, TYR-TYR and PHE-TYR association in solution and proteins.

5.5 References

1. Schliwa, M.; Woehlke, G., Molecular Motors. *Nature* **2003**, 422 (6933), 759.
2. Burley, S. K.; Petsko, G. A., Aromatic-Aromatic Interaction - a Mechanism of Protein-Structure Stabilization. *Science* **1985**, 229 (4708), 23.
3. Bhattacharyya, R., *et al.*, Aromatic-Aromatic Interactions in and around Alpha-Helices. *Protein Eng.* **2002**, 15 (2), 91.
4. Mcgaughey, G. B., *et al.*, Pi-Stacking Interactions - Alive and Well in Proteins. *J. Biol. Chem.* **1998**, 273 (25), 15458.
5. Strickland-Constable, R. F., *Kinetics and Mechanism of Crystallization from the Fluid Phase and of the Condensation and Evaporation of Liquids*. Academic Press London: 1968.
6. Collings, P. J., *Nature's Delicate Phase of Matter*. Princetone University Press, Princeton: 1990.
7. Cockroft, S. L., *et al.*, Electrostatic Control of Aromatic Stacking Interactions. *J. Am. Chem. Soc.* **2005**, 127 (24), 8594.
8. Linse, P., Stacked or T-Shaped Benzene Dimer in Aqueous-Solution - a Molecular-Dynamics Study. *J. Am. Chem. Soc.* **1992**, 114 (11), 4366.
9. Sinnokrot, M. O.; Sherrill, C. D., High-Accuracy Quantum Mechanical Studies of Pi-Pi Interactions in Benzene Dimers. *J. Phys. Chem. A* **2006**, 110 (37), 10656.
10. Jorgensen, W. L.; Severance, D. L., Aromatic-Aromatic Interactions: Free Energy Profiles for the Benzene Dimer in Water, Chloroform, and Liquid Benzene. *J. Am. Chem. Soc.* **1990**, 112 (12), 4768.
11. Chipot, C., *et al.*, Benzene Dimer: A Good Model for Pi-Pi Interactions in Proteins? A Comparison between the Benzene and the Toluene Dimers in the Gas Phase and in an Aqueous Solution. *J. Am. Chem. Soc.* **1996**, 118 (45), 11217.
12. Gräfenstein, J.; Cremer, D., An Efficient Algorithm for the Density-Functional Theory Treatment of Dispersion Interactions. *J. Chem. Phys.* **2009**, 130 (12), 124105.
13. Brooks, B. R., *et al.*, Charmm: The Biomolecular Simulation Program. *J. Comput. Chem.* **2009**, 30 (10), 1545.
14. Strumpfer, J.; Naidoo, K. J., Computing Free Energy Hypersurfaces for Anisotropic Intermolecular Associations. *J. Comput. Chem.* **2010**, 31 (2), 308.
15. Naidoo, K. J., Fearcf a Multidimensional Free Energy Method for Investigating Conformational Landscapes and Chemical Reaction Mechanisms. *Chem. Sci. Chin.* **2011**, 54, 1962.
16. Barnett, C. B.; Naidoo, K. J., Free Energies from Adaptive Reaction Coordinate Forces (Fearcf): An Application to Ring Puckering. *Mol. Phys.* **2009**, 107 (8-12), 1243.
17. Belch, A. C., *et al.*, Solvation Structure of a Sodium Chloride Ion Pair in Water. *J. Am. Chem. Soc.* **1986**, 108 (8), 1755.
18. Berkowitz, M., *et al.*, Sodium-Chloride Ion-Pair Interaction in Water - Computer Simulation. *Chem. Phys. Lett.* **1984**, 105 (6), 577.
19. Trzesniak, D., *et al.*, A Comparison of Methods to Compute the Potential of Mean Force. *ChemPhysChem* **2007**, 8 (1), 162.
20. Khavrutskii, I. V., *et al.*, Computing Accurate Potentials of Mean Force in Electrolyte Solutions with the Generalized Gradient-Augmented Harmonic Forier Beads Method. *J. Chem. Phys.* **2008**, 128 (4), 13.
21. Smith, D. E., *et al.*, Entropy of Association of Methane in Water: A New Molecular Dynamics Computer Simulation. *J. Am. Chem. Soc.* **1992**, 114 (14), 5875.
22. Smith, P. E.; Van Gunsteren, W. F., Predictions of Free Energy Differences from a Single Simulation of the Initial State. *J. Chem. Phys.* **1994**, 100 (1), 577.
23. Barnett, C. B.; Naidoo, K. J., Ring Puckering: A Metric for Evaluating the Accuracy of Am1, Pm3, Pm3carb-1, and Scc-Dftb Carbohydrate Qm/Mm Simulations. *J. Phys. Chem. B* **2010**, 114 (51), 17142.

24. Lamoureux, G., *et al.*, A Simple Polarizable Model of Water Based on Classical Drude Oscillators. *J. Chem. Phys.* **2003**, *119* (10), 5185.
25. Kumar, S., *et al.*, The Weighted Histogram Analysis Method for Free-Energy Calculations on Biomolecules. I. The Method. *J. Comput. Chem.* **1992**, *13* (8), 1011.
26. Ghosh, I.; Mccammon, J. A., Sidechain Rotational Isomerization in Proteins. Dynamic Simulation with Solvent Surroundings. *Biophys. J.* **1987**, *51* (4), 637.
27. Horn, H. W., *et al.*, Development of an Improved Four-Site Water Model for Biomolecular Simulations: Tip4p-Ew. *J. Chem. Phys.* **2004**, *120* (20), 9665.
28. Nose, S., A Molecular Dynamics Method for Simulations in the Canonical Ensemble. *Mol. Phys.* **1984**, *52* (2), 255.
29. Hoover, W. G., Canonical Dynamics: Equilibrium Phase-Space Distributions. *Phys. Rev. A* **1985**, *31* (3), 1695.
30. Hockney, R. W.; Eastwood, J. W., *Computer Simulations Using Particles*. McGraw-Hill: New York, 1981.
31. Amini, M., *et al.*, Time Integration in Particle Models. *Comput. Phys. Commun.* **1987**, *44* (1-2), 83.
32. Essman, U., *et al.*, A Smooth Particle Mesh Ewald Method. *J. Chem. Phys.* **1995**, *103* (19), 8577.
33. York, D. M., *et al.*, The Effect of Long-Range Electrostatic Interactions in Simulations of Macromolecular Crystals: A Comparison of the Ewald and Truncated List Methods. *J. Chem. Phys.* **1993**, *99* (10), 8345.
34. Smith, D. E.; Haymet, A. D. J., Free-Energy, Entropy, and Internal Energy of Hydrophobic Interactions - Computer-Simulations. *J. Chem. Phys.* **1993**, *98* (8), 6445.
35. Podeszwa, R., *et al.*, Potential Energy Surface for the Benzene Dimer and Perturbational Analysis of Pi-Pi Interactions. *J. Phys. Chem. A* **2006**, *110*, 10345.
36. Arunan, E.; Gutowsky, H. S., The Rotational Spectrum, Structure and Dynamics of a Benzene Dimer. *J. Chem. Phys.* **1993**, *98* (5), 4294.
37. Sherrill, C. D., *et al.*, An Assessment of Theoretical Methods for Nonbonded Interactions: Comparison to Complete Basis Set Limit Coupled-Cluster Potential Energy Curves for the Benzene Dimer, the Methane Dimer, Benzene-Methane, and Benzene-H₂s. *J. Phys. Chem. A* **2009**, *113* (38), 10146.
38. Singh, J.; Thornton, J. M., The Interaction between Phenylalanine Rings in Proteins. *FEBS Lett.* **1985**, *191* (1), 1.
39. Henson, B. F., *et al.*, Raman-Vibronic Double-Resonance Spectroscopy of Benzene Dimer Isotopomers. *J. Chem. Phys.* **1992**, *97* (4), 2189.
40. Tsuzuki, S., *et al.*, Origin of Attraction and Directionality of the Pi/Pi Interaction: Model Chemistry Calculations of Benzene Dimer Interaction. *J. Am. Chem. Soc.* **2002**, *124* (1), 104.
41. Matthews, R. P.; Naidoo, K. J., Experimentally Consistent Ion Association Predicted for Metal Solutions from Free Energy Simulations. *J. Phys. Chem. B* **2010**, *114* (21), 7286.
42. Headen, T. F., *et al.*, Structure of Pi-Pi Interactions in Aromatic Liquids. *J. Am. Chem. Soc.* **2010**, *132* (16), 5735.
43. Gao, J., Supercritical Hydration of Organic Compounds. The Potential of Mean Force for Benzene Dimer in Supercritical Water. *J. Am. Chem. Soc.* **1993**, *115* (15), 6893.

CHAPTER 5

INVESTIGATING AROMATIC AMINO ACID ASSOCIATION USING FREE-ENERGY HYPER-VOLUMES

5.1 Introduction

Aromatic interactions play an important role in wide range of chemical and biochemical problems, including stereochemistry of organic reactions,¹ binding affinities in host-guest chemistry,² molecular design,³ protein folding and structure,⁴⁻⁶ DNA and RNA base stacking⁷⁻⁹ and recognition of DNA by drugs and regulatory proteins.¹⁰⁻¹² To understand the complex nature of protein tertiary structure it is important to first understand the interactions that occur between the amino acid residues,^{4, 6, 12-14} as well as the interactions between the protein and its surrounding environment.

Indeed, a systematic study on the specific interactions between aromatic amino acid residues is extremely difficult from the experimental standpoint. From a theoretical standpoint,¹⁰ however, the use of well-established computational techniques on simple molecular systems can give invaluable insight on this matter. Several studies have been undertaken in this regard. For example, computer simulations using force fields were used to calculate the potential of mean force (PMF) of pairs of aromatic molecules in pure liquid phases¹⁵ and in various solvents.¹⁵⁻¹⁸ Efforts were also devoted to investigate the potential energy surface (PES) using semi-empirical and *ab initio* methods.^{16, 19-24}

Nonetheless, important questions regarding the nature and directionality of aromatic-aromatic association still remain a commonly studied topic.^{15, 16, 18, 20, 22, 25-34} All the non-bonded interactions between aromatic amino acid residues involve several different effects that can be classified as (1) van der Waals (dispersion) interactions, (2) electrostatic interactions, and (3) polarization and charge transfer effects. There is clear experimental and theoretical evidence that van der Waals and electrostatic interactions play an important role in the stability in aromatic association.^{10, 19} On the other hand, there is experimental evidence showing that charge transfer and electron donor-

acceptor effects in interacting ground state aromatic molecules are negligible compared to the electrostatics.³⁵

From a structural point of view, aromatic association can be characterized as the competition between the stacked and T-shape configurations.³⁶ The stacked configurations are stabilized by dispersion interactions, while T-shape arrangements are strongly affected by the electrostatic interactions and possibly hydrogen bonding.³¹ Furthermore, solvation must be taken into account to predict the correct intermolecular configurations in proteins and solution.¹⁶⁻¹⁸

In this study, the anisotropic free energy hyper-volumes for the aromatic association of the PHE-PHE (toluene), TYR-TYR (p-cresol) and PHE-TYR pairs in vacuum and solution are discussed. The geometry for the PHE-PHE, TYR-TYR and PHE-TYR pair configurations, found in vacuum and solution, are collectively shown in Figure 5.1. The study in solution shows the importance of the solvent in determining the configurations found for these aromatic pairs. The solution free energy hyper-volumes are compared to the PMFs derived from these aromatic interactions found in 5877 protein crystal structures. Finally, the thermodynamics of the PHE-PHE, TYR-TYR and PHE-TYR pairs in solution is discussed and used to explain the aromatic amino acid association.

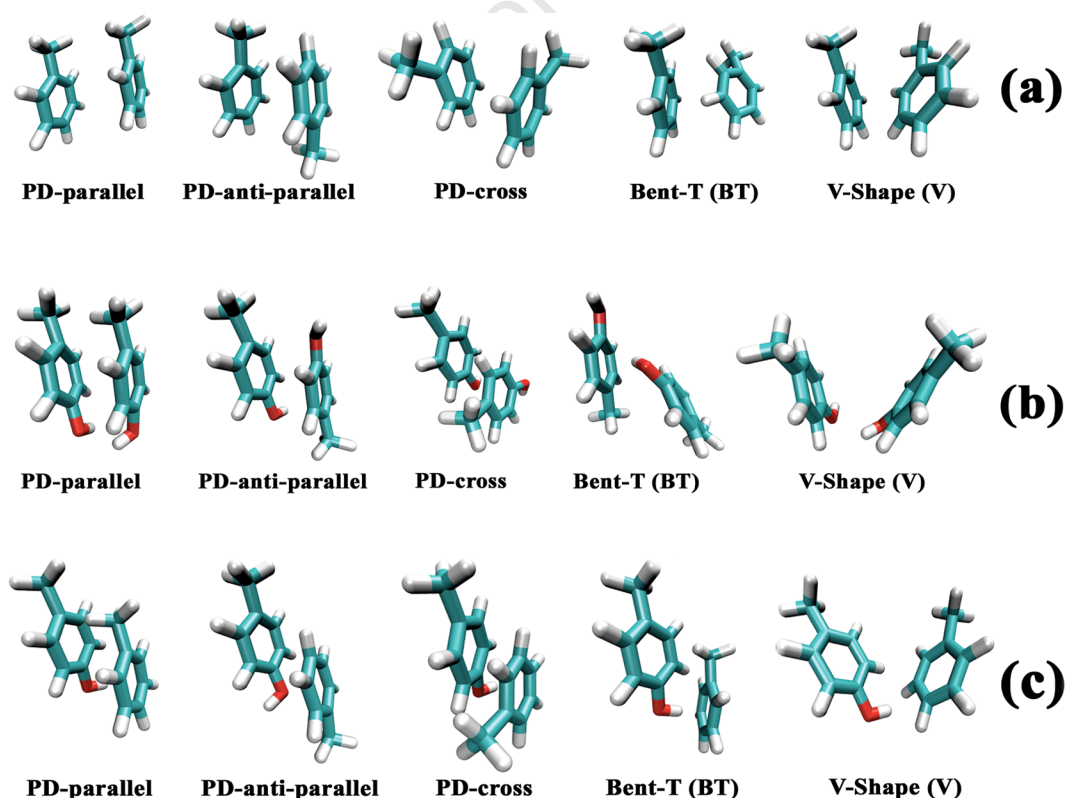


Figure 5.1: The collection of configurations found in vacuum and/or solution for the (a) PHE-PHE, (b) TYR-TYR and (c) PHE-TYR pairs.

5.2 Methods

The theoretical methodology, used in this study, has already been described in the previous chapter. Since the way in which we performed the simulations for the benzene dimer was successful, it was decided to stick to such an effective protocol. Therefore, the simulation details for vacuum and solution are exactly the same as for the benzene dimer. The previously described Boltzmann averaging methodology has also been used to reduce the dimensionality of the data to the corresponding 2-D or 3-D datasets.

5.2.1 Definition of the association reaction coordinates

In order to calculate the 4-D reaction coordinates, the centre of mass and vector normal to the plane of the aromatic rings have to be defined. In all cases, only the heavy ring atoms are used in the definition of the centre of mass, and two adjacent ring carbons (same ones used for all normal vector definitions) are used to calculate the normal vectors, exactly the same as for the benzene dimer (Figure 4.1).

5.2.2 Protein database preparation and calculations

There are two major problems that prohibit the direct use of Protein Databank (PDB) structures for statistical analysis. First, there are many closely related or identical proteins in the PDB, with, for instance, more than 200 T4 lysozyme mutants and more than 100 closely related hemoglobin structures. The second problem is that the structure quality varies greatly among structural entries, from unrefined, initial models with serious errors in connectivity, packing, and even global topology, to high-quality final structures. The obvious solution is to create a PDB subset that would be as large as possible, but would contain only unrelated protein structures of relatively high quality.³⁷ The selection of protein crystal structures used in this study was based on only three, but important, criteria: resolution, R-factor and sequence identity. The criteria are as follows:

- **Resolution < 2.2Å**
- **R-Factor ≤ 0.3**
- **Sequence Identity < 30%**

A list of 5877 PDB protein structures was the resultant of the culling procedure. The protein structures were obtained directly from the Royal Society of Chemistry Protein Databank website (RSC PDB at <http://www.pdb.org/pdb/home/home.do>).

The 4-D reaction coordinates (r , θ_1 , θ_2 and ϕ) between the PHE-PHE, TYR-TYR and PHE-TYR aromatic pairs in proteins were calculated as described in Figure 5.2. This is similar to the 4-D reaction coordinates calculated in the simulation 4-D PMF calculations. Using the WHAM methodology (described in previous chapter), the probability distributions and free energy hyper-volumes (PMF_{PDB}) for these pairs were obtained.

5.3 Results and Discussion

The configurations of the PHE-PHE, TYR-TYR and PHE-TYR pairs are collected in Figure 5.1, which were assigned from the 2-D and 3-D plots in vacuum and solution. The 2-D free energy surfaces $W(r, \theta)$, $W(r, \phi)$, $W(\theta_1, \theta_2)$ and $W(\theta, \phi)$ provides a convenient way of visualizing the various the configurations and the transformation pathways that may exist between them. Along with the 3-D volume plots $W(r, \theta, \phi)$ and $W(\theta_1, \theta_2, \phi)$, they present a clear picture of the homodimers configurational free energy landscape and topology. The contact minimum is defined for a centre-of-mass-to-centre-of-mass separation distance of ≤ 6.0 Å, as previously postulated.¹⁸ The orientation of the configurations of the contact pair can be classified into 2 categories: stacked (sandwich and PD), and T-shaped (BT and EoE/CW/V) structures.

5.3.1 Free energy hyper-volume analysis in vacuum

The free energy basin for the PHE-PHE pair is found to be relatively flat (Figure 5.3a) in the contact pair regime between 3.0 and 5.0 Å. The global minimum configuration observed from the $W(r, \theta_1)$, $W(r, \phi)$, $W(\theta_1, \phi)$ and $W(\theta_1, \theta_2)$ 2-D free energy surfaces takes on a PD form (Figure 5.3a-d and Table 5.1). In the case of the PHE pair, the association is dependent on the in-plane orientation of the PHE-PHE (i.e., the orientation about the axes passing perpendicularly through the centre of mass).

This was not found to be the case for the benzene dimer.¹⁸ This is due to methyl group lowering the symmetry of PHE molecule compared to benzene. An analysis of the relative methyl (or in-plane) orientation for each PHE (Figure 5.4a and b) shows that the PD-cross ($\psi = \pm 85^\circ$) is the preferred configuration (global minimum), while the PD-parallel ($\psi = 0^\circ$) and PD-anti-parallel ($\psi = \pm 180^\circ$) arrangements are found to be equally probable. Tsuzuki et al. showed using ARS-F (a model chemistry for the evaluation of intermolecular interaction energy between Aromatic Systems using Feller's method)³⁸ that the PD-cross, PD-parallel and PD-anti-parallel configurations are favoured, where the latter two configurations were found to be isoenergetic.³⁸

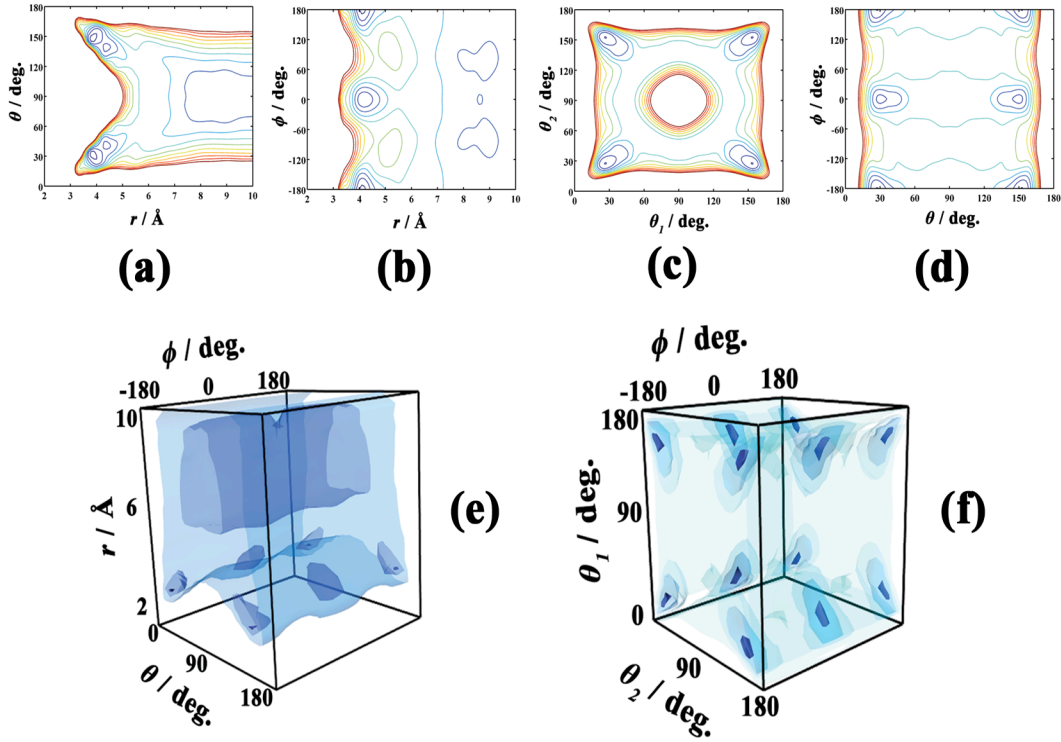


Figure 5.3: The vacuum PHE-PHE 2-D surfaces for (a) $W(r, \theta)$; (b) $W(r, \phi)$; (c) $W(\theta_1, \theta_2)$; and (d) $W(\theta, \phi)$. The 3-D volumes for (e) $W(r, \theta, \phi)$; and (f) $W(\theta_1, \theta_2, \phi)$. The 2-D plots are contoured at 0.5 kcal/mol from 0.0 to 5.0 kcal/mol, and volume levels of $1.0 k_B T$ (dark blue); $3.0 k_B T$ (lighter blue); and $8.3 k_B T$ (5.0 kcal/mol). The same contour and level volumes are applied throughout this chapter, unless otherwise stated.

From the $W(r, \theta_1)$ and $W(\theta_1, \theta_2)$ 2-D free energy surfaces, the PD-cross is found to be 0.5 and 1.0 kcal/mol lower in energy than the PD-anti-parallel and PD-parallel configurations, respectively. This does not oppose the result obtained by Tsuzuki et al., since the energy barrier separating the PD-anti-parallel and PD-parallel is $\leq 1.0 k_B T$ allowing them to exist in dynamic equilibrium. It is observed that the PD-cross and PD-anti-parallel have a shorter separation distance (4.0 Å) than the PD-parallel (4.5 Å) configuration.

This infers that the PD-cross and PD-anti-parallel can have stronger dispersion interactions than the PD-parallel configurations. The PHE pair arranged in the PD-parallel configuration cannot get closer to each other because of the bulky methyl groups. Therefore, it would appear that the steric effect imposed by the symmetry breaking methyl groups is also a key factor for determining the association of the PHE contact pair in vacuum. A simple example of this would be shaking hands, where the cross configuration is preferred over the more sterically hindered parallel and anti-parallel arrangements.

Table 5.1: The 4-D reaction coordinates for the PHE-PHE pair configurations in vacuum. The PD-cross, PD-parallel and PD-anti-parallel arrangements were assigned using Figure 5.4a.

Configuration Type	Reaction Coordinates (r ; θ_1 ; θ_2 ; ϕ)	ΔA (kcal/mol)
PD-cross	4.0 Å; 160/20°; 20/160°; 0/±180°	0.00
PD-anti-	4.0 Å; 145/35°; 35/145°; 0/±180°	0.50
PD-	4.5 Å; 135/45°; 45/135°; 0/±180°	1.00

A molecular mechanics optimization was performed for the PD and BT configurations at distances between 4.0 – 6.0 Å, and it was found that only PD configurations were obtained. The main source of attraction between the PHE contact pair in vacuum is found to be dispersion. Even though PHE-PHE has a non-zero dipole moment, the electrostatic contribution was found to be repulsive, which is consistent with results obtained by others.^{13, 16, 29, 31, 38} This is not unexpected for the PD-cross configuration where the angle between the dipoles, which corresponds to the ψ angle ($\psi = \pm 85^\circ$), results in a near zero dipole-dipole interaction. In the case of the PD-anti-parallel, the electrostatics interactions are still repulsive but less so than the PD-parallel configuration.

The free energy plots show that the PHE-PHE contact pair is more structured than the benzene dimer in vacuum, in that the PHE contact pair is only composed of one type of dimer configuration (i.e., PD configurations). This suggests that “stacked” configurations are preferred over “T” configurations. This is a similar observation to that made by Chipot et al.¹⁶ Chelli et al. using quantum mechanical and 2-D potential of mean force (PMF) calculations found that the PD structures were most favourable in vacuum and stabilized by dispersion.¹³ Stacked configurations allow more overlap between the PHE-PHE pair¹³ and take up less space overall compared to T-shape configurations.

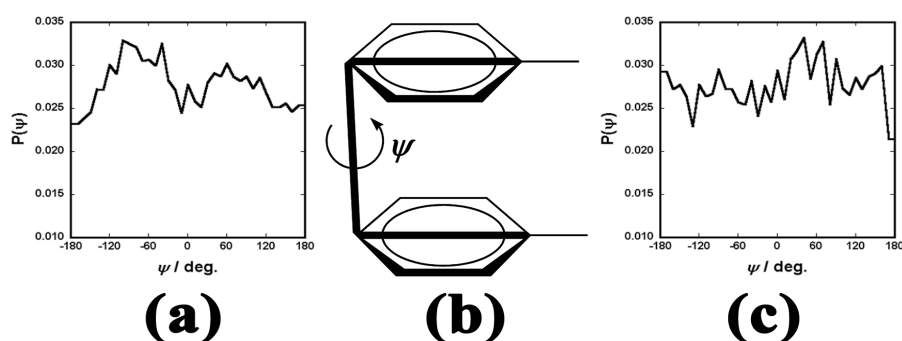


Figure 5.4: The PHE-PHE contact pair in-plane orientation, ψ , (b) definition; and the probability distributions, $P(\psi)$, (a) in vacuum; and (c) in solution.

The TYR-TYR free energy presents as a flat basin for separation distances ≤ 6.0 Å as shown by Figure 5.5a. Most noticeable is that there is a strong secondary minimum observed to begin at 7.5 Å, indicating a “solvent-shared” like minimum (Figure 5.5a). This can be attributed to edge-to-edge arrangement with the OHs hydrogen bonding.

The global minimum free energy configuration is found to be of the PD form (Figure 5.5 and Table 5.2). A statistical analysis of the trajectories was performed to determine the alignment of the PD configurations (Figure 5.6b). The preferred geometry of the TYR-TYR PD contact pair is found to be such that the OH groups are staggered in the PD configurations (Figure 5.6a), which promotes the possibility of hydrogen bonding between the OHs and OH-to-ring-centre (RC). This corresponds to a PD-cross arrangement, while the PD-parallel and PD-anti-parallel configurations are found to be equally probable.

For TYR-TYR contact pair Chelli et al obtained the PD-(cross, parallel and anti-parallel), T-shape, BT and EoE/V/CW (OH-OH hydrogen bonding) configurations from their potential energy surface (PES).¹³ From our surfaces (Figure 5.5a-f), the contact pair configurations are composed of the PD-cross, the equally probable isoenergetic PD-parallel and PD-anti-parallel, EoE/V/CW and BT configurations (Table 5.2). The PD-cross is found to be 0.5, 1.0 and 1.5 kcal/mol lower in energy than in energy than the PD-(parallel and PD-anti-parallel), EoE/V/CW and BT structures, respectively.

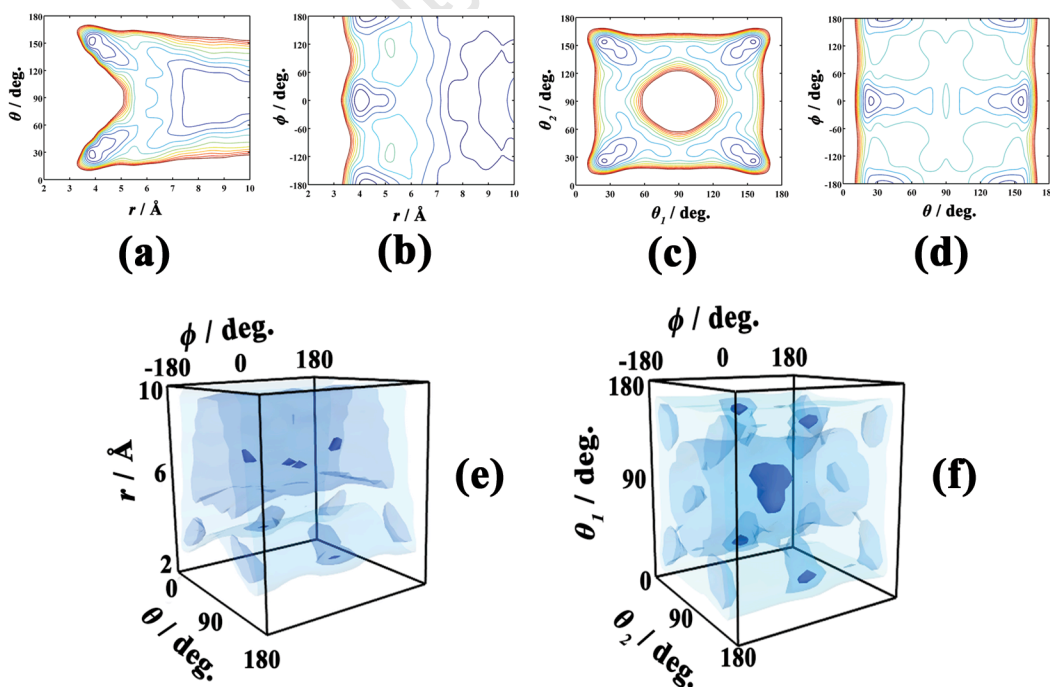


Figure 5.5: The vacuum TYR-TYR pair 2-D surfaces for (a) $W(r, \theta)$; (b) $W(r, \phi)$; (c) $W(\theta_1, \theta_2)$; and (d) $W(\theta, \phi)$. The 3-D volumes for (e) $W(r, \theta, \phi)$; and (f) $W(\theta_1, \theta_2, \phi)$.

Molecular mechanics optimizations were performed for a number of configurations with separation distances ranging from 3.0 – 7.0 Å. The BT (OH-RC hydrogen bonding at a separation distance of 4.7 Å), PD-cross (at 3.9 Å) and EoE/V/CW (at 5.9 Å) configurations were obtained. The interactions are all dominated by electrostatic interactions through hydrogen bonding.

Table 5.2: The 4-D reaction coordinates for the TYR pair in vacuum. The PD-cross, PD-parallel and PD-anti-parallel arrangements were assigned using Figure 5.6a.

Configuration Name	Reaction Coordinates (r ; θ_1 ; θ_2 ; ϕ)	ΔA (kcal/mol)
PD-cross	3.8 Å; 160/20°; 20/160°; 0°	0.00
PD-(and anti-)	4.0 Å; 135/45°; 45/135°; 0°	0.50
EoE	5.2 Å; 30/150°; 150/30°; $\pm 180^\circ$	1.00
BT	4.5 Å; 115/40°; 40/115°; 0/ $\pm 180^\circ$	1.50

The molecular mechanics optimizations indicate that the CHARMM force field (for the TYR-TYR pair) shows strong correlation with the minimum energy configurations found using *ab initio* quantum mechanics.¹³

Sherrill et al. have done a large dimer quantum mechanical study on a variety of substituted benzene molecules.^{31, 39-41} They found that the TYR analogue (phenol), as well as other substituted benzenes, binds more strongly in sandwich configuration than the benzene dimer.^{40, 41} This is found to agree with the 1-D free energy profiles for TYR-TYR and TYR-water (see Appendix Interactions and Solvent-Structuring section).

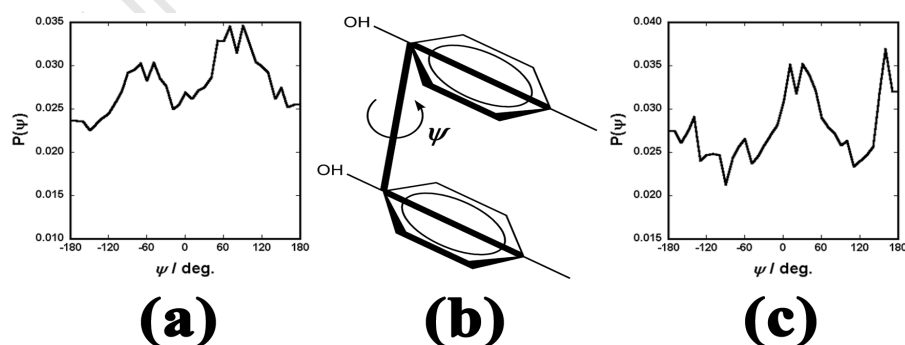


Figure 5.6: The TYR-TYR contact pair in-plane orientation (ψ) definition (b); and the probability distributions, $P(\psi)$, (a) in vacuum and (c) in solution.

The “solvent-shared” like minimum is indicative of configurations where the molecules are laying in a plane slightly tilted, maximizing hydrogen bonding occurring between the

OHs. This results in the observed centre-of-mass-to-centre-of-mass separation distance being ≥ 7.5 Å. The hydrogen bonding orientation is further substantiated by the dihedral angle being between -120° and 120° (Figure 5.5b), the correct alignment for hydrogen bonding. It can easily be seen that this configuration will form the basis of the solvent-shared minimum in solution.

The extracted 2-D contour and 3-D volume PMFs for the PHE-TYR pair (Figure 5.7) shows that the free energy basin, within the critical distance for a contact pair (i.e., ≤ 6.0 Å), is found to be fairly flat, similar to all the aromatic pairs investigated thus far. This implies that an ensemble of configurations, with energy barriers ≤ 3.0 $k_B T$, that exist in dynamic equilibrium, populates the area in and around the global minimum.

The global free energy minimum configuration was found to be of a PD form (Figure 5.7 and Table 5.3). An analysis of the in-plane orientation revealed that the most probable PD configuration is PD-cross ($\psi = \pm 75^\circ$), while the PD-parallel ($\psi = 0^\circ$) and PD-anti-parallel ($\psi = \pm 180^\circ$) configurations are equally likely (Figure 5.8a and b). It should be noted that the dihedral analysis was only performed for configurations that had a separation distance ≤ 5.0 Å and specifically for PD configurations. The BT configuration is found to be approximately 2.0 $k_B T$ (~ 1.2 kcal/mol) higher in energy than the PD-cross global minimum, while the EoE is 1.5 kcal/mol higher in energy (Table 5.3) – in agreement with others.^{13, 42-44} However this falls within the 3.0 $k_B T$ range, making them thermally accessible at room temperature.

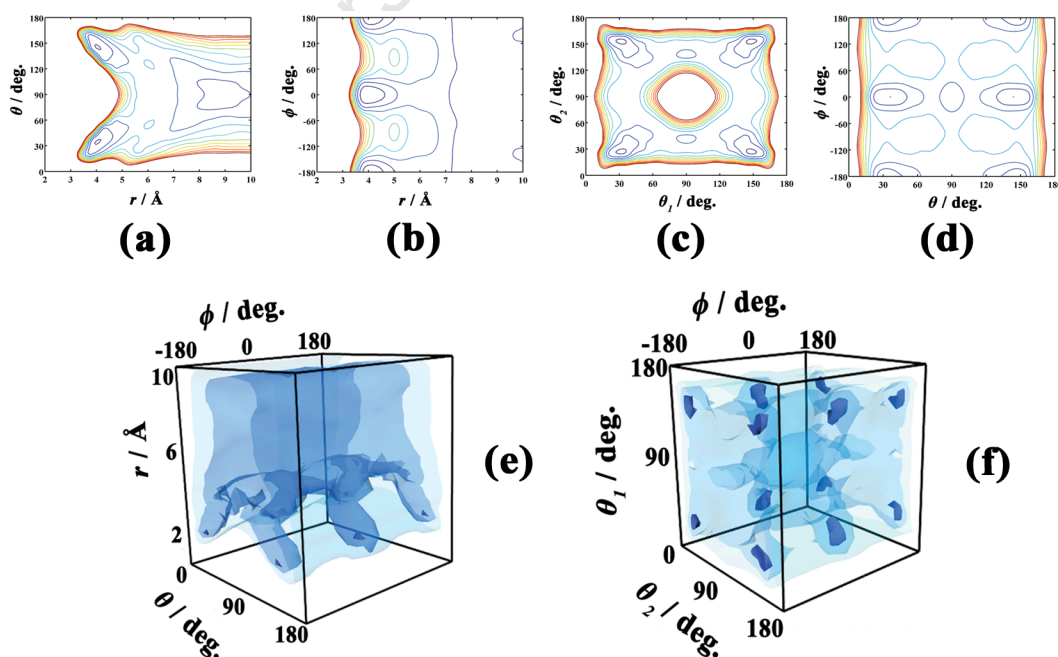


Figure 5.7: The vacuum PHE-TYR pair 2-D surfaces for (a) $W(r, \theta)$; (b) $W(r, \phi)$; (c) $W(\theta_1, \theta_2)$; and (d) $W(\theta, \phi)$; contoured at 0.5 kcal/mol from 0.0 to 5.0 kcal/mol.

Table 5.3: The 4-D reaction coordinates for the PHE-TYR pair in vacuum. The PD-cross, -parallel and -anti-parallel arrangements were assigned using Figure 5.8a.

Configuration Name	Reaction Coordinates (r ; θ_1 ; θ_2 ; ϕ)	ΔA (kcal/mol)
PD-cross	4.0 Å; 160/20°; 20/160°; 0/±180°	0.00
PD-(and anti-)	4.0 Å; 135/45°; 45/135°; 0°	0.50
BT	5.0 Å; 90/135°; 135/90°; 0/±180°	1.20
EoE	5.0 Å; 90/25°; 25/90°; 0/±180°	1.50

All the structures are stabilized mostly by large attractive electrostatic interactions. This is consistent with the result obtained by Chelli et al.,¹³ using quantum mechanical and 2-D PMF calculations, and Gervasio et al.,³⁶ using molecular mechanics and *ab initio* calculations. The electrostatic attraction is due to the OH-PHE ring-centre (RC) interaction, where aromatic rings have been shown to be good hydrogen bond acceptors.⁴⁵

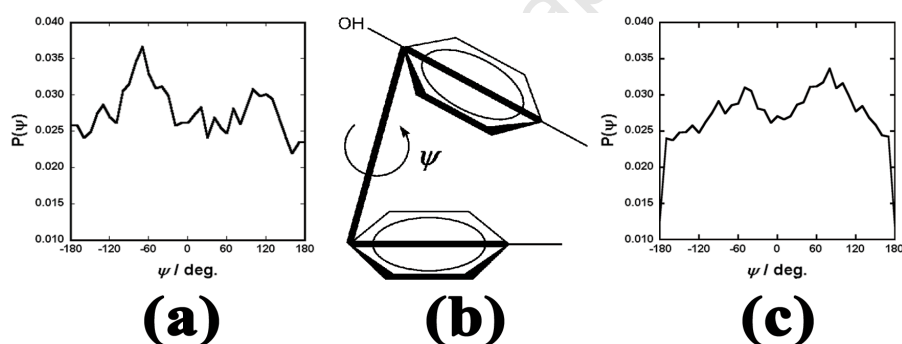


Figure 8: The PHE-TYR contact pair in-plane orientation (ψ) definition (b); and the probability distributions, $P(\psi)$, in (a) vacuum and (c) solution.

5.3.2 Free energy hyper-volume analysis in solution

Upon solvation of the PHE dimer, the overall free energy 2-D surfaces and 3-D volumes change dramatically (Figure 5.9a-d). Firstly, the global minimum is shifted from 4.0 to 6.0 Å; and secondly, the free energy basin has become very flat and much broader extending to separation distances between 5.5 to 6.8 Å (Figure 5.9a). Chipot et al. showed that the contact pair distance of approximately 5.5 Å for PHE in solution.¹⁶ It appears that, in solution, the local structure of the PHE-PHE dimer becomes more unstructured (Figure 5.9a-f) than compared to vacuum (Figure 5.3a-f). The opposite situation is observed for the benzene dimer, where the free energy configurational space is reduced in solution compared to vacuum.¹⁸ The energy barriers in the free

energy basin is found to be $\ll 3.0 k_B T$, indicating that an ensemble of contact pair configurations contribute to the global free energy well and the low energy barrier allows for regular interconversion between them (Figure 5.9a). The global minimum energy configuration is found to be an approximate PD, which has a slight twist in the dihedral angle (Figure 5.9 and Table 5.3). The BT and EoE/V/CW are found to be only 0.5 kcal/mol higher in energy than the PD configuration.

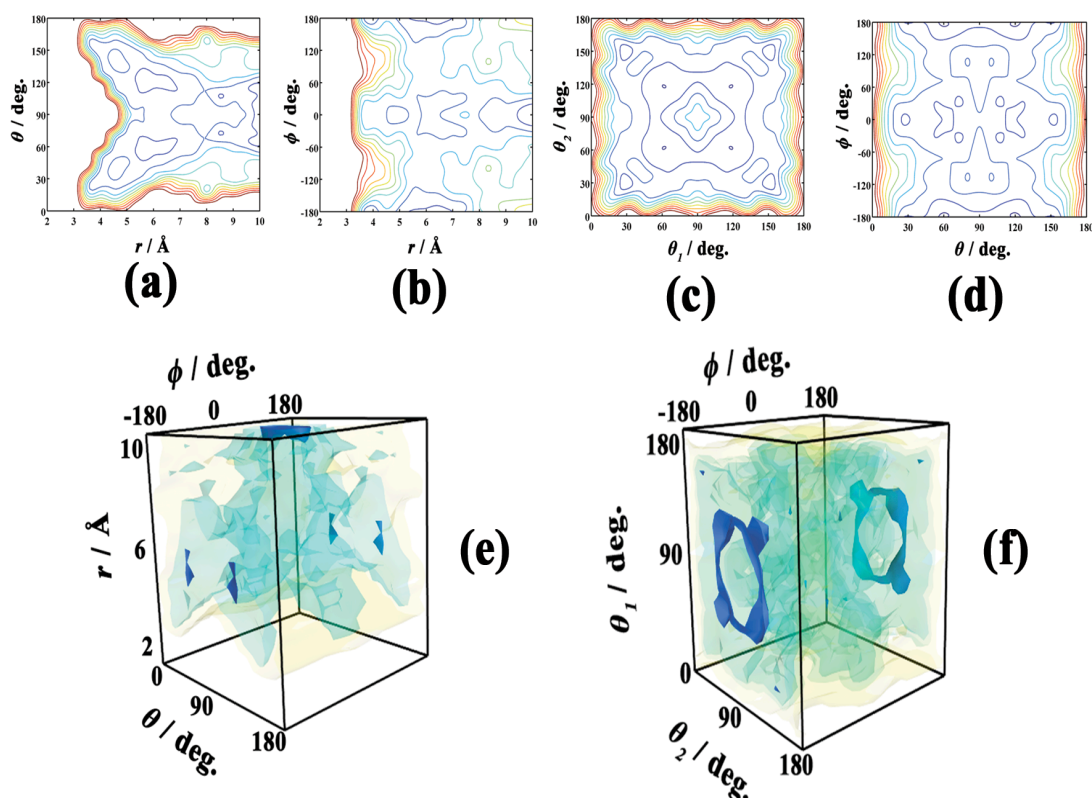


Figure 5.9: The solution PHE-PHE pair 2-D surfaces for (a) $W(r, \theta)$; (b) $W(r, \phi)$; (c) $W(\theta_1, \theta_2)$; and (d) $W(\theta, \phi)$. The 3-D volumes for (e) $W(r, \theta, \phi)$; and (f) $W(\theta_1, \theta_2, \phi)$.

The water destabilizes the contact pair formation allowing for PHE-PHE pair to assume a larger ensemble of configurations, shown by the larger $3.0 k_B T$ volume in Figure 5.9e-f. The solute-solvent interactions plays an important role here; the PHE-water interaction is found to be slightly stronger than that found for the benzene-water interaction¹⁸ (see Appendix Interactions and Solvent-Structuring section for the PHE-PHE pair).

The relative orientation of the plane of PHE-PHE ring does not prefer PD-cross, PD-parallel or PD-anti-parallel (Figure 5.4c), emphasizing the unstructured nature of this pair in solution. The EoE is found to be similar to the cogwheel (CW) structure observed by Singh and Thornton in their analysis of phenylalanine pairs in proteins.⁶ As with the case of the benzene dimer, a solvent-shared local minimum is found to begin at 8.5 Å, in

agreement with Chipot et al.¹⁶ It shows a preference for a distant-EoE/SbS configuration and is only $1.0 k_B T$ higher in energy than the global minimum.

Table 5.4: The 4-D reaction coordinates for the PHE-PHE pair in solution. The PD-cross, -parallel and -anti-parallel arrangements were assigned using Figure 5.4c.

Configuration Type	Reaction Coordinates (r ; θ_1 ; θ_2 ; ϕ)	ΔA (kcal/mol)
Contact Pair		
PD-(no preference)	5.5 Å; 60/120°, 120/60°, 0-±30/±150-180°	0.00
BT	5.5 Å; 90/60°, 60/90°, 0/180°	0.30
EoE	5.5 Å; 70/110°, 110/70°, 0-±30°	0.60
Solvent-Shared		
Distant-EoE/SbS	8.0 Å; 60/120°, 120/60°, 0°	0.60

In solution, the TYR-TYR contact pair minima range between $3.5 \leq r \leq 5.0$ Å. The 2-D surface has fractioned into two legs and they are observed to be narrower, with respect to θ (Figure 5.10a), when compared to the vacuum case, PHE and benzene dimers.¹⁸ The free energy basins, in each leg, appear to be flat with energy barriers $\leq 3.0 k_B T$. Secondary local minima at 7.5 Å are observed, which corresponds to the solvent-shared minima as anticipated from the vacuum analysis (Figure 5.10a). However, the solvent-shared minimum for the TYR dimer appears to be more favourable in the case of TYR than either PHE or benzene.¹⁸ This is due to the greater hydrogen bonding capability of TYR. The solvent-shared minima have a free energy of association that is only $1.0 k_B T$ higher in energy than the global minimum contact pair free energy.

The contact pair configurations are made up of an ensemble of PD, EoE/V/CW and BT configurations (Table 5.5 and Figure 5.10a-f). The energy barriers separating these configurations are low ($\leq 3.0 k_B T$), which allows them to exist in dynamic equilibrium. The global free energy minimum configuration is found to be of PD form. The in-plane orientation analysis shows that the PD-parallel and PD-anti-parallel configurations are the preferred configurations, while the PD-cross arrangement becomes less probable compared with the vacuum case (Figure 5.6c).

The PD-parallel has increased electrostatic interactions due to the hydrogen bonding between the OHs groups and with water, while PD-anti-parallel geometry allows for hydrogen bonding with the ring centre and water. The favourable TYR-water interaction is shown the Appendix – Interactions and Solvent-Structuring section for the PHE-PHE

pair. The PD configurations are found to have the shortest separation distance of the all the contact pair configurations. This enhances the PD configurations stability brought by stronger dispersion interactions.

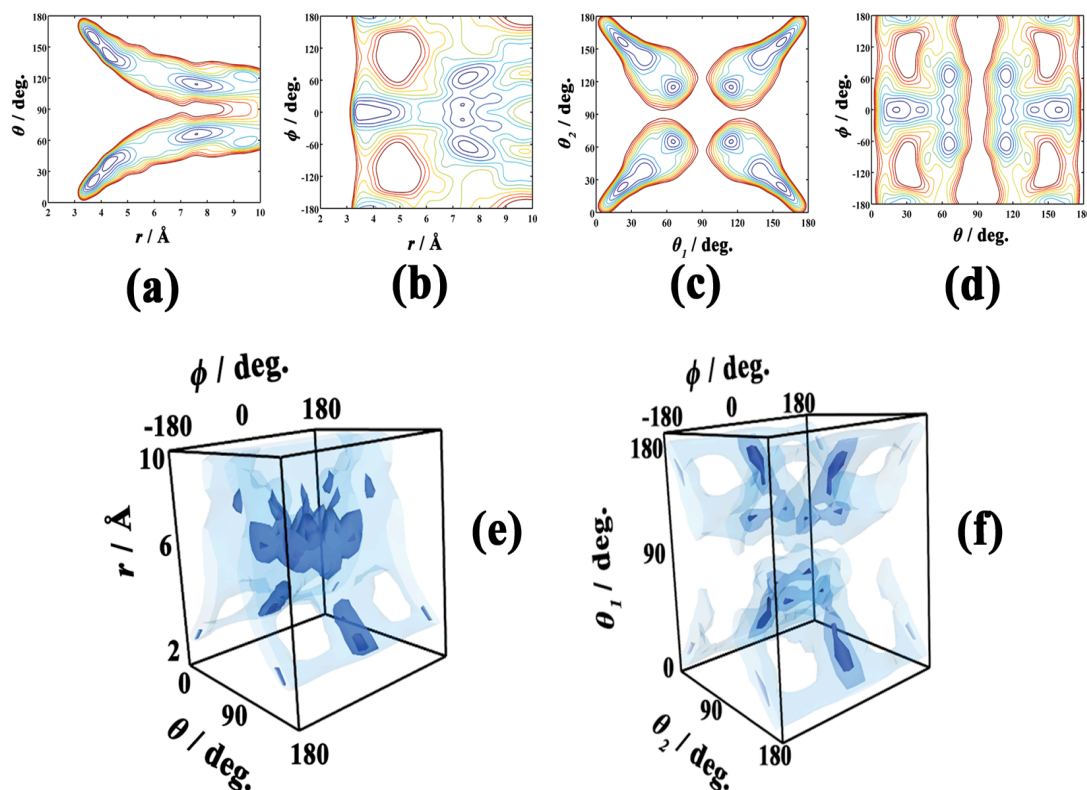


Figure 5.10: The solution TYR dimer 2-D surfaces for (a) $W(r, \theta)$; (b) $W(r, \phi)$; (c) $W(\theta_1, \theta_2)$; and (d) $W(\theta, \phi)$. The 3-D volumes for (e) $W(r, \theta, \phi)$; and (f) $W(\theta_1, \theta_2, \phi)$.

The solvent-shared minimum consists of two configurations, distant-EoE₁/SbS₁ and distant-EoE₂/SbS₂. The distant-EoE₁/SbS₁ configuration is more planar shown by a dihedral angle of ± 20 (Figure 5.10b). The slight offset from planarity allows for OH-OH hydrogen bonding. The distant-EoE₂/SbS₂ (r 7.5 Å; θ_1 120/60; θ_2 60/120; ϕ ± 60) configuration is more offset from than the distant-EoE₁/SbS₁. This corresponds to a structure where the TYR molecules are arranged in such a way that they maximize the hydrogen bonding capability between the OH-OH and OH-water.

Yan and Spangler also found a configuration similar to the distant-EoE₂/SbS₂, where hydrogen bonding was occurring between the OHs of the TYR molecules.⁴⁶ They used the low-frequency portion of the TYR dimer jet-cooled fluorescence excitation spectrum to obtain their results. They also commented on the dynamics of the TYR dimer and suggested that the complete dissociation of the dimer is not in agreement with their results, but appears that it occurs via a photophysical process such as the internal conversion.⁴⁶ Configurational interchange can occur quite easily in each of the legs

observed in the 2-D free energy surfaces, as shown in Figure 5.10a. This supports the internal conversion mechanism of Yan and Spangler.⁴⁶ However, complete dissociation will allow the system to cross from one leg to another.

Table 5.5: The 4-D reaction coordinates for the TYR pair in solution. The PD-cross, -parallel and -anti-parallel arrangements were assigned using Figure 5.6c.

Configuration Name	Reaction Coordinates (r ; θ_1 ; θ_2 ; ϕ)	ΔA (kcal/mol)
Contact Pair		
PD-(and anti-)	3.5 Å; θ_1 160/20°; θ_2 20/160°; 0°	0.00
PD-cross	3.5 Å; 35/145°, 145/35°, 0/±180°	0.50
EoE	4.5 Å; 60/120°; 120/60°; -60 – 60°	0.50
BT	4.5 Å; θ_1 115/35; θ_2 35/115; 0°	1.80
Solvent-Shared		
Distant-EoE/Sbs1	7.5 Å; 120/60; 60/120; ±20	
Distant-EoE/Sbs2	8.0 Å; 60/120°, 120/60°, ±60°	

The contact pair for the PHE-TYR falls in range between 4.0 – 6.0 Å, and solvent-shared minimum begins to appear at 8.5 Å (Figure 5.11a). The amount of 4-D (r , θ_1 ; θ_2 ; ϕ) configurational space available to this pair solution (Figures 5.5e-f) is found to be less than in vacuum (Figures 5.11e-f), in accord with the TYR-TYR pair and benzene dimer.¹⁸ The favourable contact pair configurations are found to be BT and PD configurations. However, the PD configurations are slightly tilted forming an approximate EoE-like configuration (Figure 5.11a-f).

This is seen by the non-linearity in the θ_1 and θ_2 values, i.e. $\theta_1 = 25/155^\circ$ and $\theta_2 = 150/30^\circ$, whereas pure PD configurations display linearity, i.e. $\theta_1 = 30/150^\circ$ and $\theta_2 = 150/30^\circ$. The global free energy minimum configuration for the PHE-TYR pair (Figure 5.11 and Table 5.6) is found to be the PD configurations, with the PD-cross being only slightly more probable than the PD-parallel and PD-anti-parallel (Figure 5.8c). The BT is 1.0 kcal/mol higher in energy than the global minimum and an EoE contact pair configuration is observed, but it is almost 3.0 $k_B T$ higher in energy than the global minimum. The PD and BT configurations have a decreased electrostatic and larger dispersion interaction compared to vacuum, which is found to be in agreement with

Chelli et al.¹³ A distant-EoE/SbS occurs for the solvent-shared minima configurations. This occurs to enhance the OH-to-PHE-ring-centre (RC) interaction, as well as the interaction of PHE and TYR with the surrounding waters (see Appendix Interactions and Solvent-Structuring section for the PHE-TYR pair), and the solvent-shared local minimum is only $1.0 k_B T$ higher in energy than the global minimum. The energy barriers separating the contact pair configurations are found to be approximately $3.0 k_B T$, making the interconversion between configurations accessible at 298.15 K. However, it is more likely that the interconversion occurs via the solvent-shared configuration where rearrangement can occur.

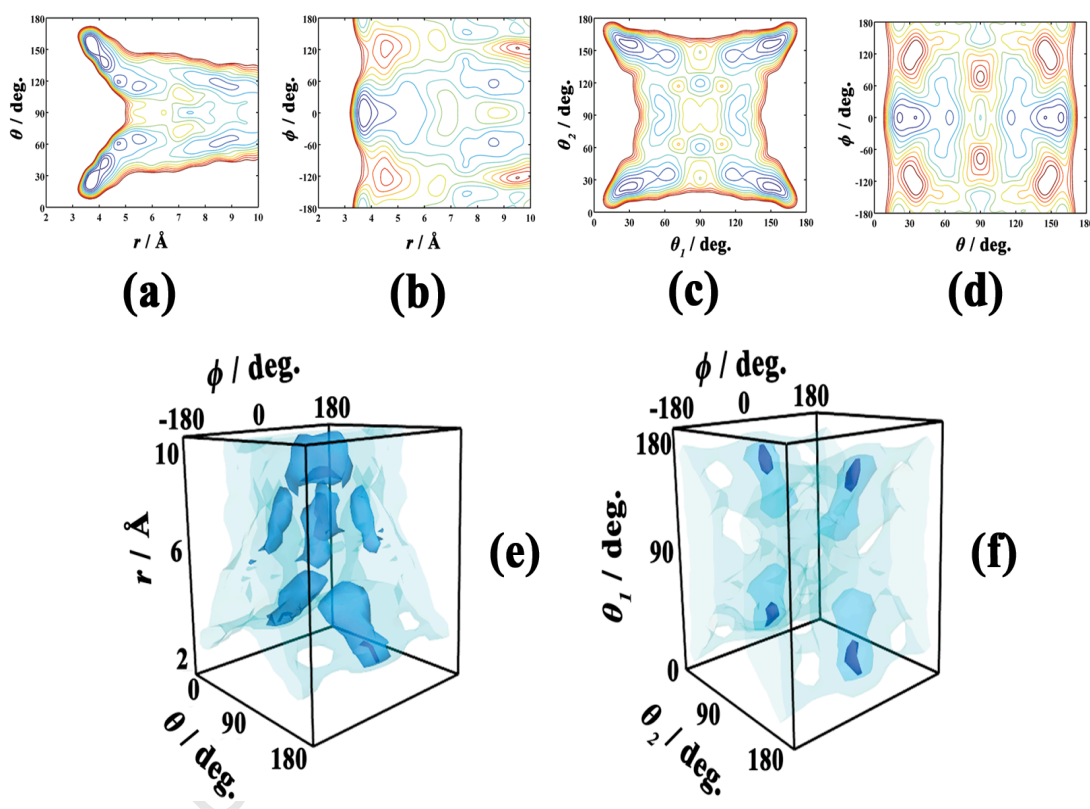


Figure 5.11: The solution PHE-TYR dimer 2-D surfaces for (a) $W(r, \theta)$; (b) $W(r, \phi)$; (c) $W(\theta_1, \theta_2)$; and (d) $W(\theta, \phi)$. The 3-D volumes for (e) $W(r, \theta, \phi)$; and (f) $W(\theta_1, \theta_2, \phi)$.

Fractionation of the 2-D free energy surface (Figure 5.11a) appears to be less dramatic for the PHE-TYR pair compared with the TYR-TYR pair. This indicates that the PHE-TYR is more forgiving with regards to planar solvent-shared configurations. The PHE-TYR also appears to be a mix of the PHE-PHE and TYR-TYR pair interactions.

The solvent-shared minima occurring at $8.0 \text{ \AA}$, as opposed to the $7.5 \text{ \AA}$ for the TYR-TYR pair, can be attributed to lesser attractive electrostatic interactions for the PHE-TYR compared with the TYR-TYR pair. The energy barrier separating the two legs of the PHE-TYR surface (Figure 5.11a), at the solvent separated minima at $8.0 \text{ \AA}$, is observed to

be ≤ 2.0 kcal/mol. A planar solvent-shared distant-EoE/SbS configuration is possible for the PHE-TYR pair because PHE lacks an OH, and the PHE-TYR interaction (in vacuum) is only found to be favourable by 0.3 kcal/mol compared to 0.6 kcal/mol for the TYR-TYR at 8.0 Å (see Appendix Interactions and Solvent-Structuring section for the PHE-TYR pair).

Table 5.6: The 4-D reaction coordinates for the PHE-TYR pair in solution. The PD-cross, -parallel and -anti-parallel arrangements were assigned using Figure 5.8c.

Configuration Name	Reaction Coordinates (r ; θ_1 ; θ_2 ; ϕ)	ΔA (kcal/mol)
Contact Pair		
PD-cross	4.0 Å; 155/25°; 25/155°; 0°	0.00
PD-(and anti-)	4.0 Å; 135/45°; 45/135°; 0°	0.30
BT	4.8 Å; 150/60°; 60/150°; 0/±180°	1.00
EoE	5.5 Å; 90/120°; 90/120°; 0/±180°	1.80
Solvent-Shared		
Distant-SbS	8.8 Å; 120/60°; 60/120°; ±60°	1.00

5.3.3 Aromatic amino acid association as found in proteins

The protein can be considered as a special solvent where several aromatic side chains, rigidly anchored to the backbone, are solvated. Many interacting (neighbour) and non-interacting aromatic residue pairs can be identified in the dataset, and all of these pairs can be considered as a statistical ensemble.¹³ Therefore, an ensemble approach can be applied following the same procedure can be performed as described for calculating the 4-D PMFs in the r , θ_1 , θ_2 and f configurational space in the “protein solvent”.

It has been shown that the interior hydrophobic core does not have an effective dielectric constant of 1, rather it has found to be between 2 – 4.^{47, 48} The apparent dielectric effect experienced by the internal charges is considerably higher than the low dielectric constants of hydrophobic matter used to represent the protein interior in electrostatic continuum models of proteins.⁴⁷ This suggests that the hydrophobic core has a significantly higher dielectric constant than vacuum. The backbone of the protein

also contributes to the high effective dielectric constant via its dipole moment,⁴⁹⁻⁵¹ and water molecules have also been known to penetrate the hydrophobic core,^{48, 52, 53} that is corroborated by the number of solvent inaccessible water molecules in Table 5.7.

Details concerning the number of buried, at or near the protein surface (termed here as interface) and exposed amino acid residues and water molecules are given in Table 5.7. This data, in conjunction with Figure 5.12, provides insight into the environment around associating PHE-PHE, TYR-TYR and PHE-TYR pairs in proteins. The solvent accessible surface area (SASA) for each amino acid is used to calculate the relative SASA, and it is used as a metric to determine if an amino acid is buried, at the interface or exposed to the solvent. A relative SASA value of < 20% is buried, between 20 and 50 % interface, and > 50 % is exposed.

Table 5.7: An analysis of a random subset (100 protein structures) from our larger protein set was used to generate the data in this table. Calculated using data obtained from the Stride program.⁵⁴ This is a rough estimate and more rigorous investigation will be required, however this is beyond the scope of this work and will be studied in future work.

	% Buried	% Interface	% Exposed
Water	<u>56.7</u>	27.1	<u>16.2</u>
Protein	55.6	14.8	29.6
PHE	81.3	12.5	6.2
TYR	28.9	54.5	16.6

Even from such as small database of 100 hundred proteins, the results are already showing that the PHE residues prefer to be buried, or inside the protein where water cannot access it (Figure 5.12a). On the other hand, the TYR residue does not have a particular preference in that it was found to either be buried or at the interface/surface. This is it agrees with the dogma that TYR residues area interfacial residues, occurring at the surface and interface of proteins.⁵⁵ Another interesting note is that there appear to very large number of water molecules buried, or inaccessible to the solvent, inside the protein and at the protein-protein interfaces (Figure 5.12b). These waters have shown to be disordered and entropically favoured, while the waters in are enthalpically favourable.^{56, 57} The large occurrence of PHE-TYR and TYR-TYR pairs at the surface suggests that a number of weakly associating aromatic interactions can add up to stabilize even protein-protein interfaces. Nonetheless, it is clearly evident that the PHE-PHE interactions are predominantly found within the hydrophobic core of the protein, while the TYR-TYR show a preference for interacting near or at the surface of the

protein. The interaction between the hydrophobic-hydrophilic PHE-TYR (*[ENV is hydrophilic]-TYR- π/π -PHE-[ENV is hydrophobic]*) pair also shows a slight bias towards interacting near the surface; the TYR has its non-associated side facing the hydrophilic environment, while the PHE has its non-associating side facing the hydrophobic core.

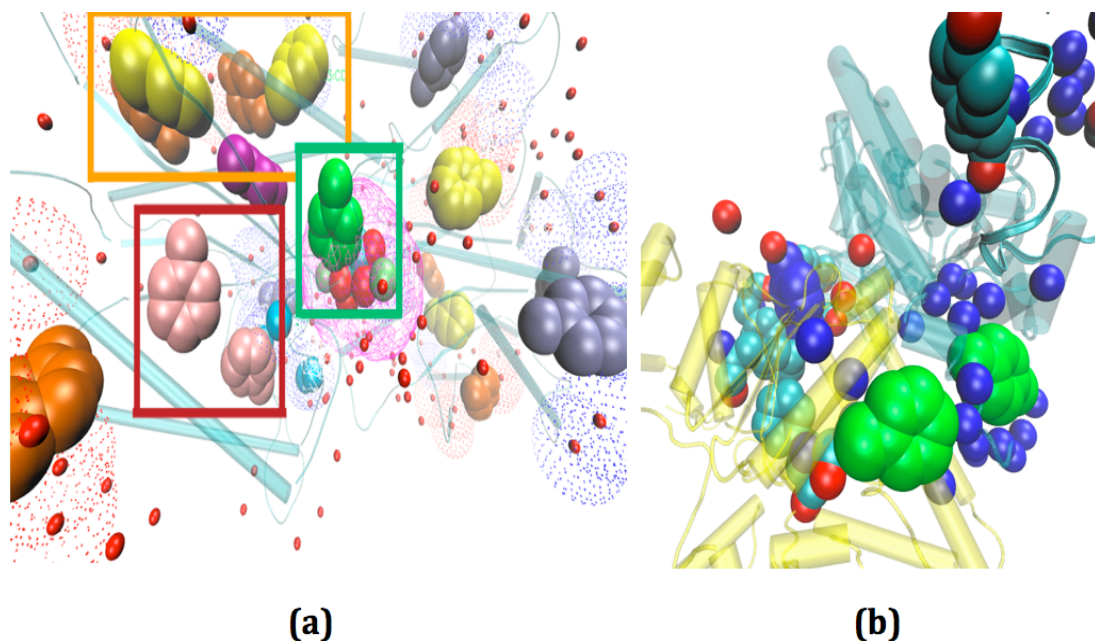


Figure 5.12: The two proteins chosen reflect to vital aspects of proteins. (a) 8ABP.pdb is a transport enzyme and has to bind its substrate with high affinity and specificity. The important interactions here are the buried PHE-PHE residue contact pair (maroon box), the PHE (yellow)-TYR (orange) contact pairs occurring near the surface (orange box) and the substrate-PHE interaction (green box). (b) 1M1N.pdb is hetero-dimeric tetramer involving protein-protein interactions – only chain A and B (cartoon representation) are shown for clarity. The important interactions are the green solvent-shared PHE-TYR configuration (green) and adjacent to it a collection of PHE-PHE and PHE-TYR contact pair configurations occurring within the interface (cyan and blue). Red and blue spheres are exposed and buried waters.

A closer investigation of the PHE-PHE, TYR-TYR and PHE-TYR pairs using 4-D PMF approach, such that we obtained “pseudo” free energy hyper-volumes for these pairs. The 4-D reaction coordinates are defined in the same manner as the theoretically derived multi-dimensional PMFs (Figure 5.2). The interest is in pairs that are closely associated (i.e., $r \leq 8.0$ Å), because pairs that are farther apart may have other amino acid residues or molecules between them. In the r phase space, for $2.0 \leq r \leq 8.0$ Å, the number of PHE-PHE, TYR-TYR and PHE-TYR pairs are found to be 1588064, 729248 and 861408 respectively. From this information, the PHE-PHE pair is the more prevalent aromatic-aromatic interaction in proteins compared with the PHE-TYR and TYR-TYR, respectively. PHE-PHE pairs occur more often than expected by chance, which is in

agreement with others.^{6, 36, 37, 58} The 3-D volume plots, in $W(r, \theta, \phi)$ phase space, for the PHE-PHE, TYR-TYR and PHE-TYR pair associations in proteins and solution are plotted in Figures 5.13 and 5.14, respectively. A quick glance at the volume plots shows the $\sin\theta$ (and $\sin\phi$) phenomenon for $r > 8.0$ Å, described by Singh and Thornton.⁶ The free energy hyper-volumes are tagged pseudo because they represent a biased sampling toward aromatic pair configurations found in proteins, and parameterization from such data will lead to biased simulations. The phase space for aromatic amino acid interaction in proteins is profoundly undersampled and the parameterization of an effective coarse-grain (CG) interaction potential from such data will lead to biased protein structures and dynamics. This was discovered by Liwo et al. using their United Residue (UNRES) protein folding simulation package.⁵⁹ Thus, it is preferable to use converged (i.e., near-complete sampling of configurational phase space) free energy data for parameterizing CG interaction potentials. This notion will be the subject for the remaining chapters of this thesis.

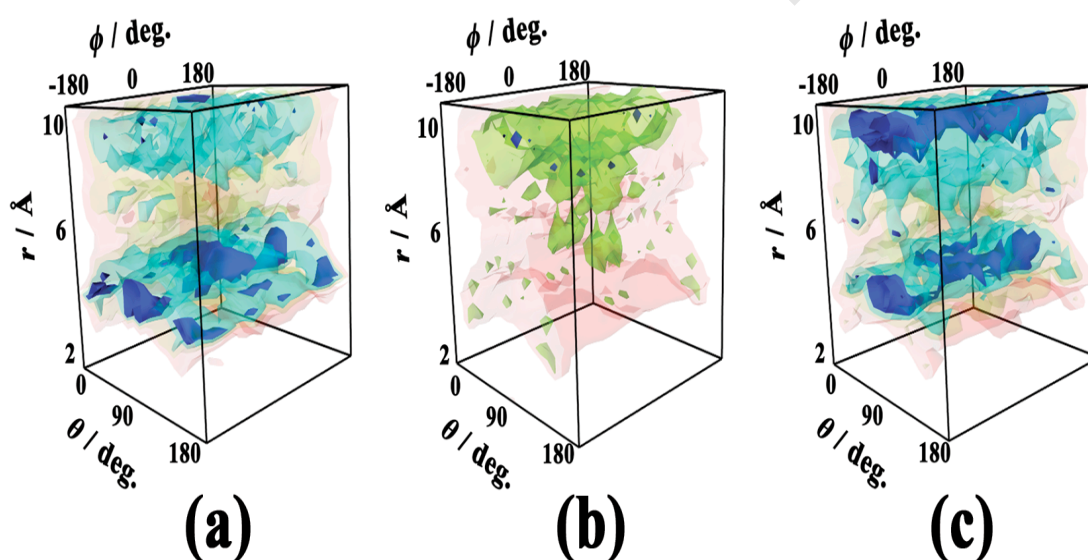


Figure 5.13: The 3-D PMF_{PDB} , $W(r, \theta, \phi)$, volume plots for the (a) PHE-PHE, (b) TYR-TYR and (c) PHE-TYR pairs. Volume values of 1.0, 1.5 and 3.0 $k_B T$ are the blue, yellow/light blue and red ≤ 3.0 $k_B T$, respectively.

The PHE-PHE association in proteins (Figure 5.13a) is composed of an ensemble of contact pair configurations (i.e., $r \leq 6.0$ Å), which corresponds with the PHE-PHE solution PMF (Figure 5.14a). Although this pair interaction occurs in the hydrophobic core of the protein, consideration of the internal effective dielectric constant and water penetration into the hydrophobic must be taken into account. The contact pair ensemble is composed of PD, BT and EoE/CW/V configurations, in agreement with other studies,^{4, 6, 11, 60} and is also evident from Figure 5.12. The large number of types of contact pairs available suggests that the PHE-PHE pair association in proteins contributes to the

stability of the tertiary structure (as well as the process of protein folding) via a favourable entropy contribution (see next section thermodynamics) to the free energy.

On the other hand, TYR-TYR pair association in proteins is found to be more likely for larger separations ($r \geq 6.0$ Å) than contact pairs (Figure 5.13b). This is not unexpected due to the fact that this pair occurs mostly in a hydrophilic environment, such as the solvent exposed surface of proteins.¹³ By being exposed to the solvent the TYR-TYR should display more solvent-shared like behaviour, which is found to be the energetically favourable form in the TYR-TYR solution PMFs (Figure 5.14b). From the solution PMF the solvent-shared configuration is a distant-EoE/SbS geometry, which occurs so as to maximize the TYR-TYR and TYR-water electrostatic interactions through hydrogen bonding (see Appendix Interactions and Solvent-Structuring section for the TYR-TYR pair).

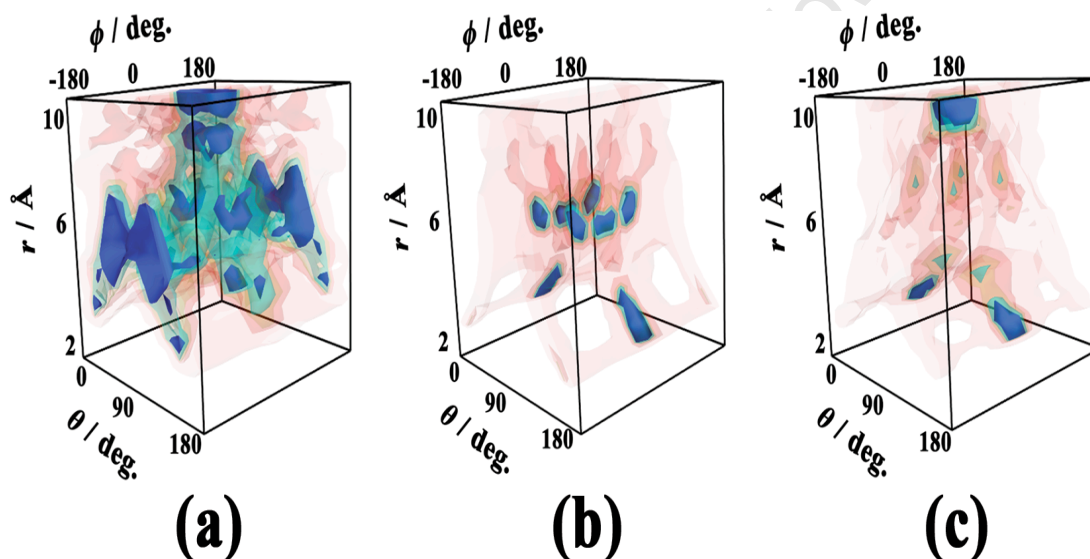


Figure 5.14: The solution 3-D PMF, $W(r, \theta, \phi)$, volume plots for the (a) PHE-PHE, (b) TYR-TYR and (c) PHE-TYR pairs. Volume values of 1.0, 1.5, 3.0 and $8.3 k_B T$ are shown by dark blue, light blue/green, darker red and the light, transparent red respectively.

The PHE-TYR association shows characteristics of both the PHE-PHE and TYR-TYR pairs, in that solvent-shared and contact pair configurations are observed in the protein (Figure 5.13c). A similar observation is made for the PHE-TYR solution PMF shown in Figure 5.14c. The contact pair is composed of PD and BT configurations, while solvent-shared configurations are a distant-EoE/SbS, clearly seen in Figure 5.12. The similarity between the protein and solution PMF results from the polar OH group of TYR, which can have attractive interactions with the ring centre of the PHE (for the contact pair) and the surrounding solvent (for solvent-shared configuration).

These results suggest that the solution free energy hyper-volumes for the PHE-PHE, TYR-TYR and PHE-TYR pairs can satisfactorily describe these aromatic associations found in proteins. A closer investigation into the thermodynamics will, therefore, provide insight into the major factors that aromatic association has on the 3-D tertiary structure found in proteins.

5.3.4 Thermodynamics analysis

THE AROMATIC AMINO ACID ASSOCIATION

The Helmholtz free energy, ΔA , internal energy, ΔE , and the entropy, $-T\Delta S$, for the PHE-PHE, TYR-TYR and PHE-TYR pairs association in solution is shown in Figure 5.15. The thermodynamic calculations were performed in the same way as detailed for the benzene dimer (Section 4.2.3).¹⁸ The global free energy minima (red curve) for the PHE-PHE, TYR-TYR and PHE-TYR pairs occur at 5.7, 5.0 and 5.3 Å, respectively, and presents as a broad well in agreement with results obtained by others.^{13, 16} Therefore, similar to the benzene dimer, the 3-D volume and 2-D contour plots (Figures 5.9–5.11) deconvolutes the orientational preference for these aromatic pairs indicating that the broad well, seen in the 1D PMF, is a result of a collection of aromatic (PHE-PHE, TYR-TYR and PHE-TYR) contact pair configurations. It is immediately clear that the relative orientation of the aromatic pairs are of significant importance, implying that distance alone is insufficient for describing aromatic association in vacuum and solution. Important to note here is that the critical distance of 6.0 Å for contact pair formation is used to define the distance at which a contact begins to form for all pairs studied here.¹⁸

The curves for the entropy (green) and enthalpy (blue) are also plotted in Figure 5.15. Quite evident from the PHE-PHE profile is that there is no strong enthalpically stabilized PD contact pair ($r \leq 4.0$ Å) found in solution (Figure 5.15a), whereas the TYR-TYR, PHE-TYR and benzene dimer (Figure 5.15b-d) have these present. The cause for the lack of such an enthalpic minimum is due to the (i) the weak association of the PHE's (see Figure A5.4 for the PHE-PHE vacuum 1D free energy profile), and (ii) the symmetry breaking and, hence, structure breaking effect of the methyl group (which is not observed in the benzene dimer case). However, an internal energy (or enthalpy, blue curve) global minimum is observed at 7.0 Å, closer to the free energy minimum than the PD global minima for the TYR-TYR, PHE-TYR and benzene dimer pairs. This would suggest an internal energy-driven association, or an enthalpy driven hydrophobic effect, as it appears that this is responsible for the initial formation of the PHE pair in solution.

This appears to run counter-intuitive to what we consider to be the hydrophobic effect, but there have been a few studies that have corroborated such a result, indicating that the shape of the hydrophobic surface determines the driven force.^{56, 57, 61} Chandler has summed up how the hydrophobic effect changes characteristics as the length scale of the hydrophobic scales.⁶² As the length increases, it causes the water to form an interface decreasing its ability to form all of its hydrogen bond, thus, incurring an enthalpic cost. Upon hydrophobic association, these waters are released back into the bulk where they can form hydrogen bonds resulting in the large enthalpic stabilization, while an entropic cost is incurred as they become more structured. However, the enthalpic contribution is found to be greater than the entropy resulting in an enthalpically driven hydrophobic effect.^{56, 57, 61, 62}

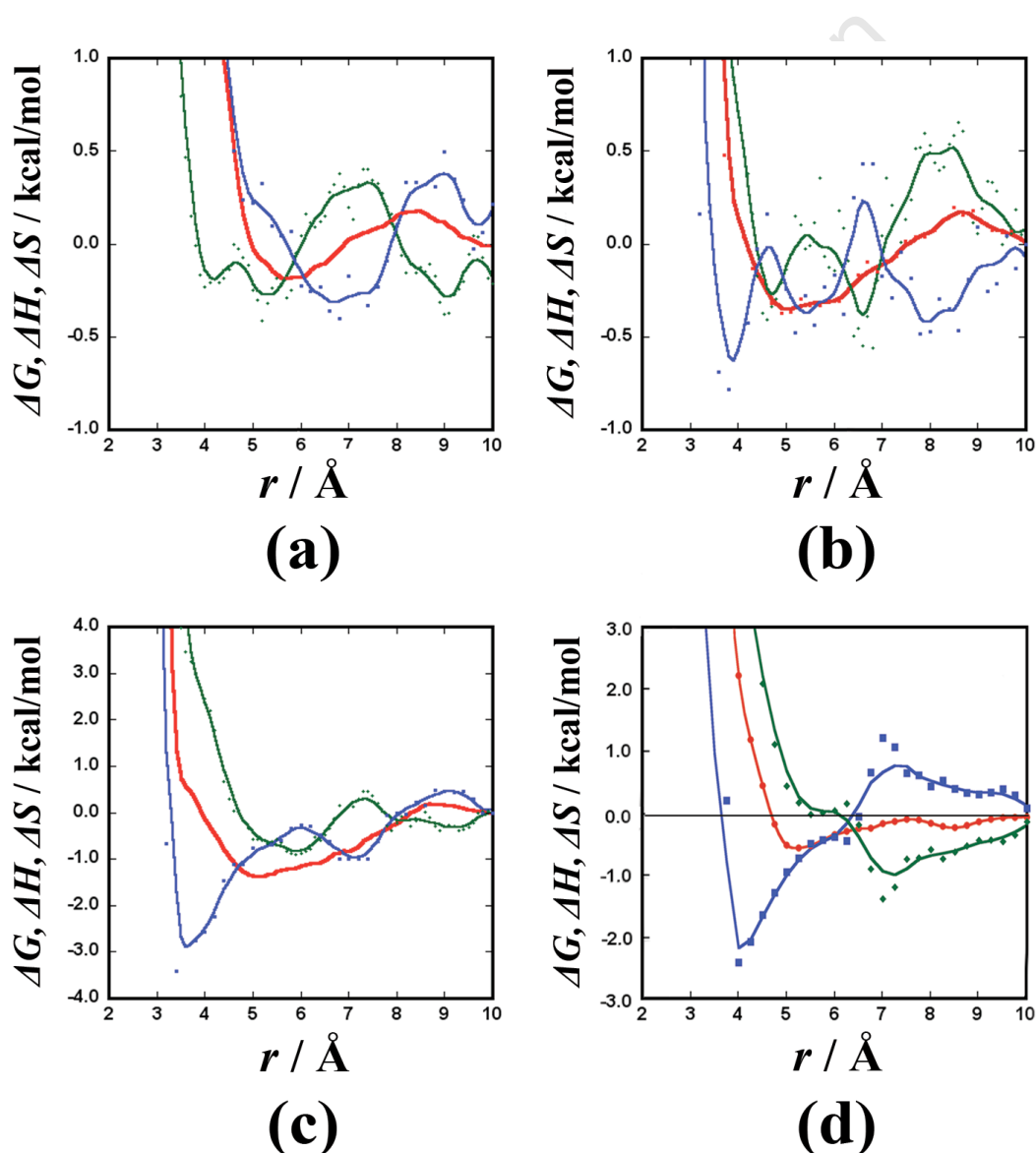


Figure 5.15: The thermodynamic free energy (red), entropy (green), and internal energy/enthalpy (blue) curves for (a) PHE-PHE, (b) TYR-TYR and (c) PHE-TYR pairs in solution; and (d) the benzene dimer curves.

Two local entropy minima are observed for the PHE-PHE contact pair at 4.0 and 5.2 Å, in the PD configuration region, while the TYR-TYR, PHE-TYR and benzene dimer pairs have only a one. The occurrence of two rather than one local entropy minima indicates that the PHE-PHE contact is much less disordered than TYR-TYR, PHE-TYR and the benzene dimer pairs. This is also concurs with the PHE-PHE not possessing a very favourably-stabilized PD near 4.0 Å, which is mainly in part due to the symmetry and structure breaking methyl group. This is further confirmed by the solution PHE-PHE pair in-plane dihedral angle probability distribution, $P(\psi)$, plot, which shows that it does not favour either of the PD-parallel, -anti-parallel or -cross configurations (Figure 5.4c), leading to the increase in entropy observed. However, for the TYR-TYR the parallel and anti-parallel configurations are favoured (Figure 5.6c),¹³ and the PD-cross configurations are favoured for the PHE-TYR pair (Figure 5.8c). The type of PD configuration is null for the benzene dimer, due to the symmetry. The increased probability is a result of a significant interaction, causing that particular configuration to have a larger probability than compared to complete random structure (more or less demonstrated by the PHE-PHE pair). The directionality of these PD configurations indicate that there is a strong electrostatic interaction occurring between the OH and RC (ring centre) for the TYR-TYR and PHE-TYR, resulting in the favourable enthalpy,^{56, 57, 61} for case of the benzene dimer this interaction is mostly composed strong dispersion and higher-order multipole (i.e, quadrupole-quadrupole) interactions.¹⁸ This indicates that the tightly bound contact, i.e. separation distance, $r \leq 4.5$ Å, is not entropically stabilized but rather enthalpically stabilized, which agrees with Marsili et al.⁶⁰

Solvent-shared configurations (at $r = 7.5$ Å) for the PHE-PHE and benzene dimer ($\Delta A = 0.1$ and -0.15 kcal/mol) are much less stable than the TYR-TYR (-0.3 kcal/mol) and PHE-TYR (-0.45 kcal/mol) pairs. The scenario for the pairs are similar but differ only in the magnitude of the interaction between the aromatic and water and the fact that PHE and TYR can only support 5 waters along the ring edge, whereas benzene can accommodate 6. The 1-D free energy interaction profiles for constituent components and solvent distribution functions (SDFs) are shown in the appendix of this chapter. The solvent-shared distant-EoE/SbS is enthalpically stabilized indicating that solvent-shared and solvent-separated configurations are a phenomena caused by the solvent stabilizing these distant pairs via solute-solvent and solvent-solvent interactions.

Interestingly enough is that the global free energy minimum for the all the amino acid aromatic associating pairs appear to occur at or close to the point at which the entropy and enthalpy curves intersect (Figure 5.15a-c), whereas for the benzene dimer

this does not occur (Figure 5.15d). The negative enthalpy gradient for the PHE-PHE and TYR-TYR pair happens in these two pairs for different reasons. For the PHE-PHE pair this occurs due to the stabilization offered by the entropy for the disordered PD configuration, while for the TYR-TYR it is caused by the breaking of OH-OH hydrogen bonding to reform a configuration with shorter separation and OH-RC hydrogen bonding. The PHE-TYR follows the same rationale as the benzene dimer; where the entropy is closer the free energy global minimum indicating that the entropy determines the orientation of the contact pair configuration. These results, in conjunction with the thermodynamic results for the benzene dimer,¹⁸ shows that the entropy (i.e., hydrophobic effect) plays an important role in determining the orientation of the aromatic pairs association in solution. The ability of the PHE-TYR and TYR-TYR pairs to form solvent-shared configurations, allows them to stabilize the protein surface as well as protein-protein interfaces. The associations of the PHE-PHE, TYR-TYR and PHE-TYR pair's contributions to protein structure are mainly driven by entropy. This is in agreement with the notion that the hydrophobic effect is the main driving force for protein folding.^{63, 64}

5.4 Conclusions

In this study, an investigation of the biologically relevant aromatic amino acid residues is conducted. The PHE-PHE, TYR-TYR and PHE-TYR interactions have been studied using multi-dimensional anisotropic free energy hyper-volumes in vacuum, solution and proteins.

The free energy hyper-volume for the association of the PHE-PHE pair shows that only PD configurations in vacuum is formed (stabilized by dispersion interactions), while in solution electrostatic interactions between the PHE-water results in the lower local structuring of the PHE contact pair, as shown by the ensemble of configurations available. The TYR-TYR interaction is dominated by electrostatic interactions, mostly via hydrogen bonding and O-H-- π interactions. This is observed by the formation of EoE/V/CW and BT configurations in vacuum, which is not present for either the benzene dimer or PHE pair. The competition between solute-solute and solute-solvent interactions results in the solution 4-D PMF splitting into two "legs", which are separated by a free energy barrier >5.0 kcal/mol. This is a consequence of the unexpected highly stabilized solvent-shared configuration, which has the an OH capable of hydrogen bonding with either with the adjacent TYR OH, water or both by optimizing the TYR-TYR-water hydrogen bonding network. The latter is inferred from the

configuration located about this intermolecular separation ($r \sim 7.5 \text{ \AA}$). Only a weak shoulder solvent-shared shoulder is observed from the 1D PMF, further substantiating the importance of the orientation. The PHE-TYR anisotropic free energy of association shows a “cross-correlation” between the structure and thermodynamic behaviour of the PHE-PHE and TYR-TYR pairs. Direct interconversion between contact pair global minima is somewhat more possible compared to the TYR pair and the wells are less defined, the PHE-like behaviour; while the electrostatics in the form of hydrogen bonding and O-H $\cdots$ π interactions is found to be important, similar to the TYR pair. The notion that the thermodynamic data of the PHE-TYR pair shows behaviour similar to both the PHE-PHE and TYR-TYR pairs is reverberated in the thermodynamic data.

The PMF_{PDB} for the PHE-PHE, TYR-TYR and PHE-TYR pairs show good correlation with the computed solution free energy hyper-volumes. The ensemble of contact pair configurations for PHE-PHE pair and the thermodynamic analysis shows that these configurations are driven by a favourable entropic contribution to the free energy. On the other hand, the PMF_{PDB} for the TYR-TYR pair indicates that it prefers the entropically favoured solvent-shared configuration, which maximizes the electrostatic interactions, via hydrogen bonding, between the TYR-TYR and TYR-water. The PHE-TYR shows similar behaviour to both the PHE-PHE and TYR-TYR pairs; it has both contact pair and solvent-shared configurations, both driven by a favourable entropic contribution to the free energy of association. However, upon closer inspection of two proteins structures it is observed that the PHE-TYR contact pair and solvent-shared associations occur near the surface of the protein and quite possibly at protein-protein interfaces.

It is found that the computed free energy hyper-volumes in solution are able to describe the association of the PHE-PHE, TYR-TYR and PHE-TYR in proteins adequately. The thermodynamic analysis reveals that the entropy makes a significant contribution to the contact pair configuration and the overall association. This shows that the entropically driven aromatic association (hydrophobic effect) is important for the stability of protein tertiary structure and the process of protein folding. Not only is this an important result in terms of chemistry, but now it also lays a foundation for developing coarse-grain models of these aromatic amino acid residues for coarse-grained simulations of protein dynamics. Future work includes perform a more instructive PDB analysis on larger PDB dataset as well as for certain types of enzymes, e.g. glycosyltransferases. The final part of this thesis details a proof of concept, the use of free energy hyper-volumes for parameterizing coarse-grain interaction potentials.

5.5 Appendix

AROMATIC INTERACTIONS IN PROTEINS

The PHE-PHE 2-D contour plots, contoured at 0.3 kcal/mol, are shown in Figure A5.1.

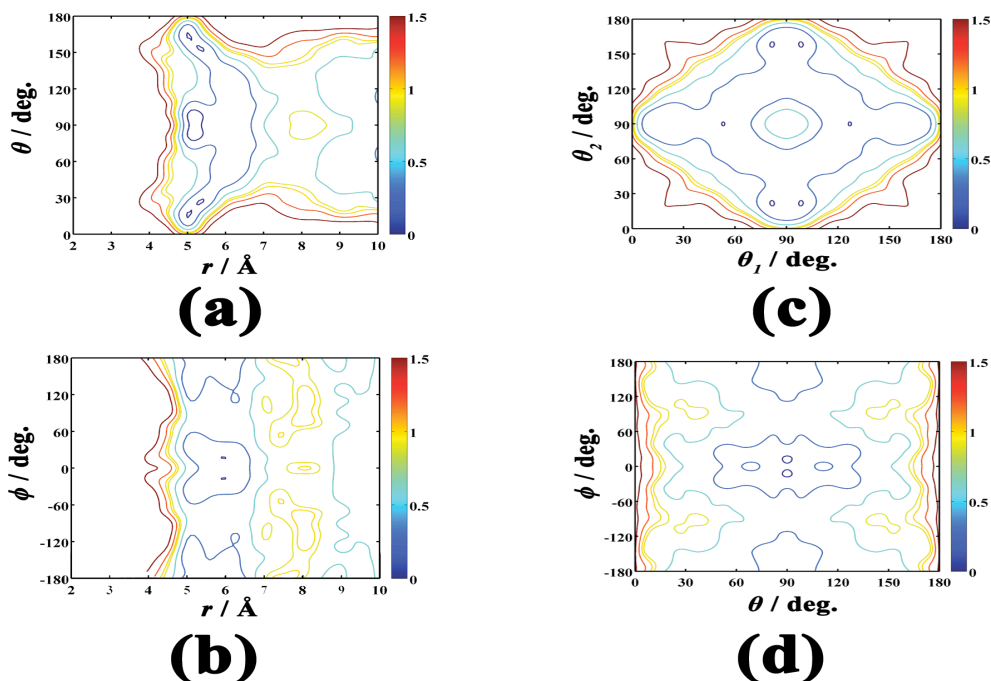


Figure A5.1: The PHE-PHE dimer 2-D surfaces for (a) $W(r, \theta)$; (b) $W(r, \phi)$; (c) $W(\theta_1, \theta_2)$ extracted for distances < 6.0 ; and (d) $W(\theta, \phi)$ extracted for distances < 6.0 Å.

The TYR-TYR 2-D contour plots, contoured at 0.3 kcal/mol, are shown in Figure A5.2.

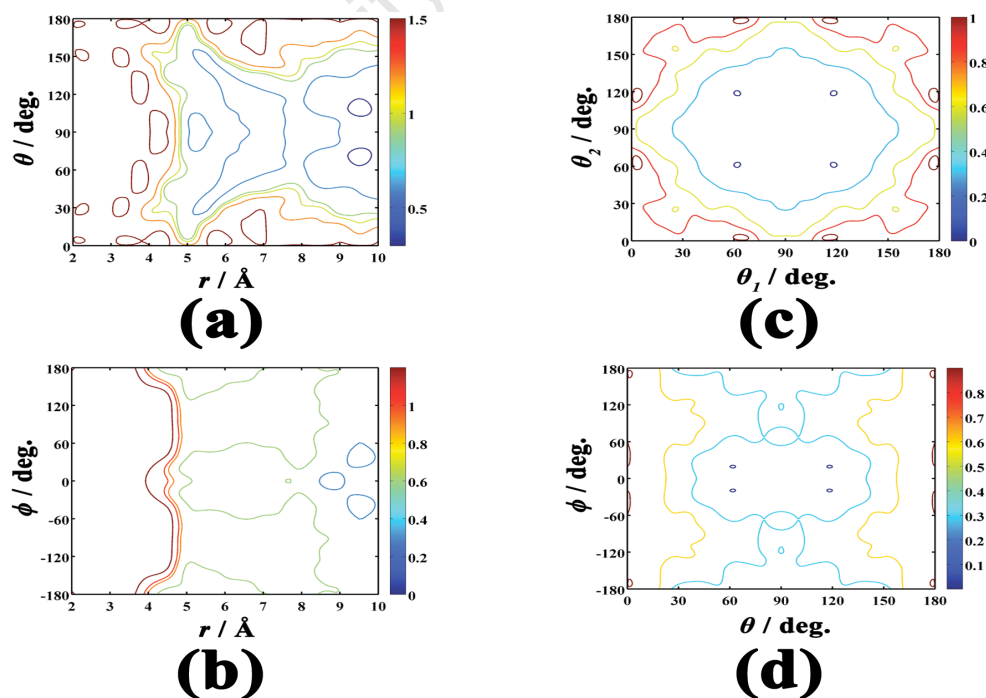


Figure A2: The protein TYR-TYR dimer 2-D surfaces for (a) $W(r, \theta)$; (b) $W(r, \phi)$; (c) $W(\theta_1, \theta_2)$ extracted for distances < 6.0 ; and (d) $W(\theta, \phi)$ extracted for distances < 6.0 Å.

The PHE-TYR 2-D contour plots, contoured at 0.3 kcal/mol, are shown in Figure A5.3.

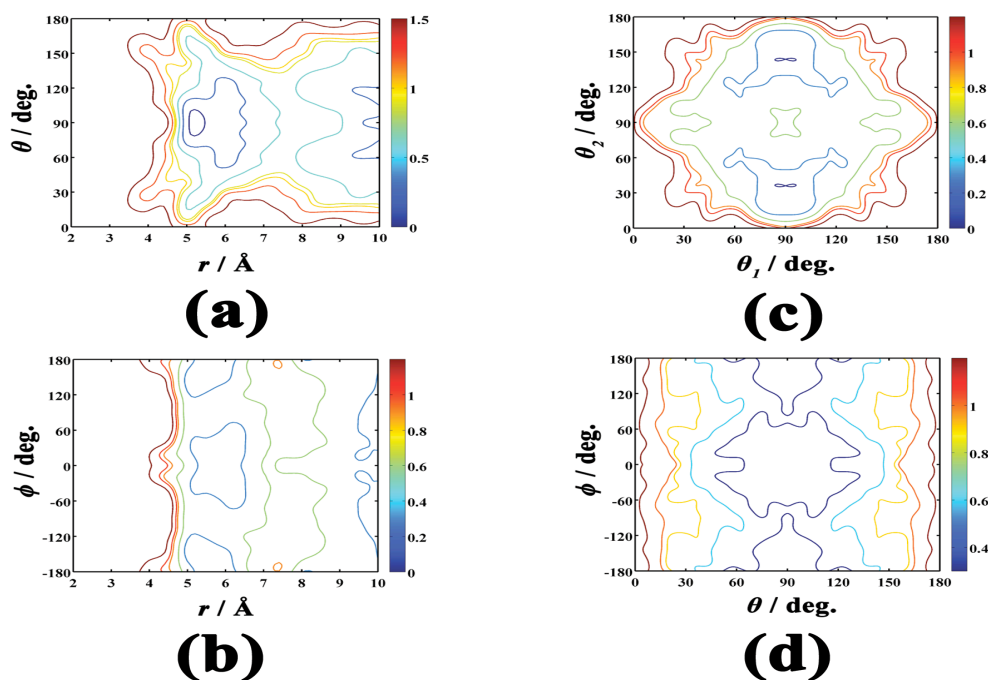


Figure A5.3: The PHE-TYR dimer 2-D surfaces for (a) $W(r, \theta)$; (b) $W(r, \phi)$; (c) $W(\theta_1, \theta_2)$ extracted for distances < 6.0 ; and (d) $W(\theta, \phi)$ extracted for distances < 6.0 .

INTERACTIONS AND SOLVENT-STRUCTURING

The 1D PMFs for the PHE-PHE and PHE-water interactions, in vacuum, are shown in Figure A5.4. The PHE contact pair global minimum (Figure A5.4a) is found to occur at 5.5 Å, in agreement with Chipot et al.¹⁶ and the vacuum free energy surfaces (Figure 3a-d). The PHE-water (Figure A5.4b) interaction shows that the water interacts mostly with the PHE ring-centre, and to a lesser degree with the ring H's shown by the small shoulder at 5.0 Å. The PHE-water interaction is found to be 0.8 kcal/mol, which is 0.2 kcal/mol more favourable than the benzene-water interaction.

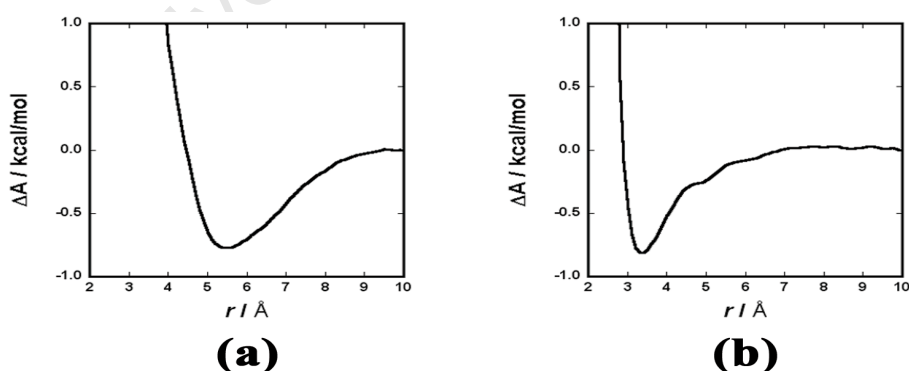


Figure A5.4: The 1D PMF interaction profiles for (a) PHE-PHE; (b) PHE-water in vacuum.

This can be attributed to favourable electrostatic interactions due to the non-zero dipole of PHE. This increased attractive PHE-water interaction results in the destabilization of the contact pair as observed in the free energy surfaces and volumes in Figure 5. Thus, it appears that the solvated PHE pair configurations appear to a competition between the solute-solute and solute-solvent interactions, which have approximately equal free energies (Figure A5.4).

The solvent distribution functions are shown below in Figure A5.5 for the contact pair, solvent-shared and non-interacting for the PHE-PHE pair. A decrease in water probability density (i.e, release of bound water molecules) is observed when going from the non-interacting (Figure A5.5c) to the solvent-shared (Figure A5.5b) to the contact pair (Figure A5.5a) configurations.

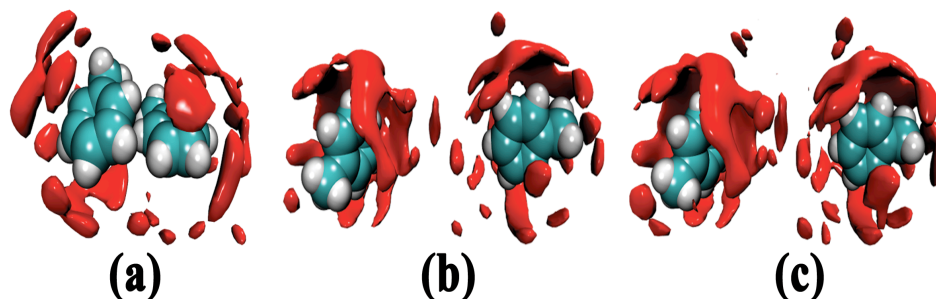


Figure A5.5: Spatial distribution functions (SDFs) for the PHE-PHE (a) contact pair ($r \leq 6.0$ Å); (b) solvent-shared ($6.0 \text{ Å} \leq r \leq 12.0$ Å) and (c) non-interacting ($r \geq 12.0$ Å). All densities are plotted at 50% above bulk.

The 1D PMFs for the TYR-TYR and TYR-water are shown in Figure A5.6. It is observed that the association of the TYR pair is stronger than the PHE and benzene dimers in vacuum.^{31, 39-41} Figure A5.6b shows the water interacts stronger with the OH at 4.9 Å (1.4 kcal/mol) than the ring centre at 3.3 Å (0.5 kcal/mol). The TYR-TYR interaction is found to be at least 0.5 kcal/mol stronger than the TYR-water interaction. This suggests a reason for the fractionation of the free energy surface (Figure 5.10a) observed in solution and not in vacuum (Figure 5.3a). For the solvent-shared configuration, at 7.5 Å, the TYR pair is attempting to create as many hydrogen bonds as possible (i.e., between the OH-OH and OH-water), which is only possible for the non-planar arrangement. The planar arrangement loses the ability for hydrogen bonding to occur between the OHs and loses a favourable 1.0 kcal/mol interaction (Figure A5.6a). The loss of a favourable interaction and the TYR's inhibited ability to rotate, due to a high solvent-structuring (Figure A5.7b), results in the appearance of the two legs in the solution free energy surface (Figure 5a). The contact pair is separated for similar reasons, highly favourable TYR-water interactions resulting in waters being bound to the contact pair (Figure A5.6b) and the obvious steric effects.

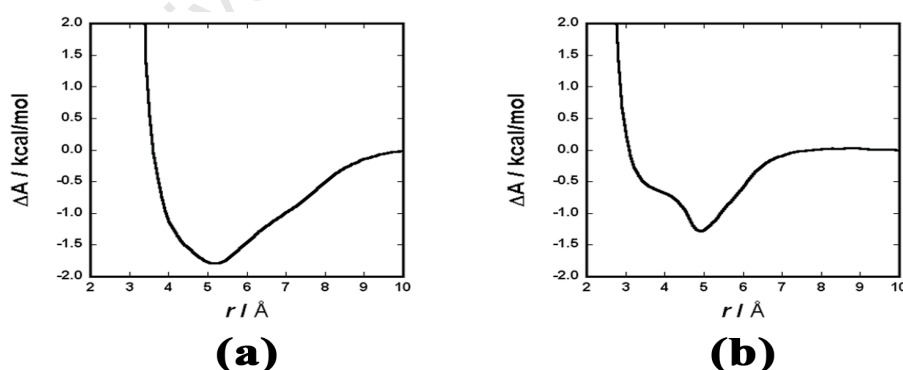


Figure A5.6: The 1D PMF interaction profiles for (a) TYR-TYR; (b) TYR-water in vacuum.

The solvent distribution functions are shown below in Figure A5.7 for the contact pair, solvent-shared and non-interacting for the TYR-TYR pair. A decrease in water probability density (i.e, release of bound water molecules) is observed when going from the non-interacting (Figure A5.7c) to the solvent-shared (Figure A5.7b) to the contact pair (Figure A5.7a) configurations.

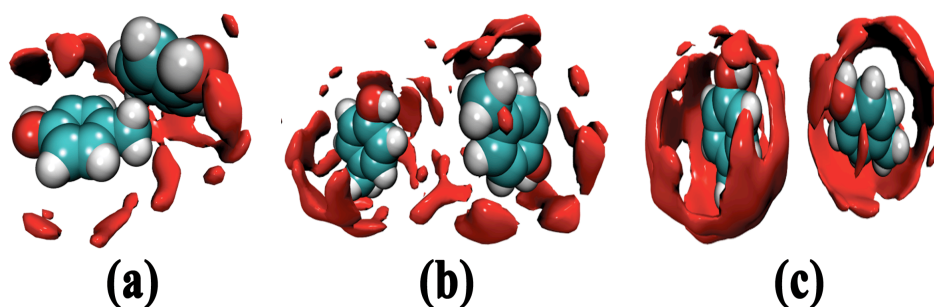


Figure A5.7: Spatial distribution functions (SDFs) for the TYR-TYR (a) contact pair ($r \leq 6.0$ Å); (b) solvent-shared ($6.0 \text{ Å} \leq r \leq 12.0$ Å) and (c) non-interacting ($r \geq 12.0$ Å). All densities are plotted at 50% above bulk.

The solute-solute (PHE-TYR) and solute-solvent (PHE-water and TYR-water) interaction profiles are shown in Figure A5.8. The PHE-TYR interaction is found to be 1.45 kcal/mol, which is more attractive than the benzene dimer. This is similar observation made by Sherrill et al,^{40, 41} and can be attributed to the increased electrostatic attraction. The solvent-shared configuration (at 8.0 Å) is found to have an attractive free energy of 0.4 kcal/mol; the global minimum for the PHE-water and the TYR-water interactions is found to be 0.8 and 1.4 kcal/mol respectively. Therefore, the solvent-shared configuration is dominated by the solute-solvent interaction, while the contact pair is found to a balance between the PHE-TYR (1.45 kcal/mol) and solute-solvent competitions.

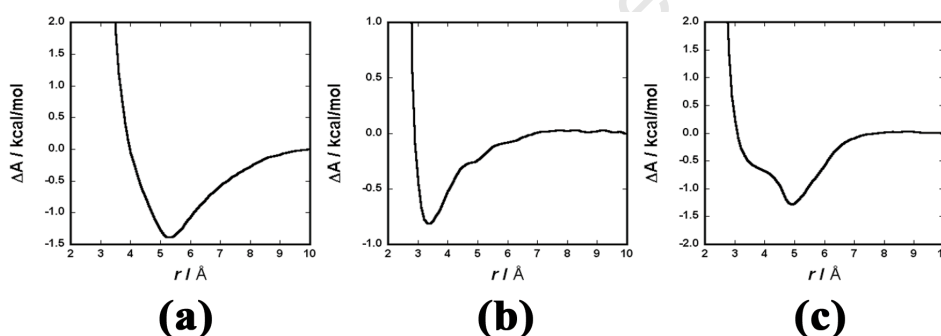


Figure A5.8: The 1D distance PMF interaction profiles for (a) PHE-TYR; (b) PHE-water; and (c) TYR-water in vacuum.

The solvent distribution functions are shown below in Figure A5.9 for the contact pair, solvent-shared and non-interacting for the PHE-TYR pair. A decrease in water probability density (i.e, release of bound water molecules) is observed when going from the non-interacting (Figure A5.9c) to the solvent-shared (Figure A5.9b) to the contact pair (Figure A5.9a) configurations.

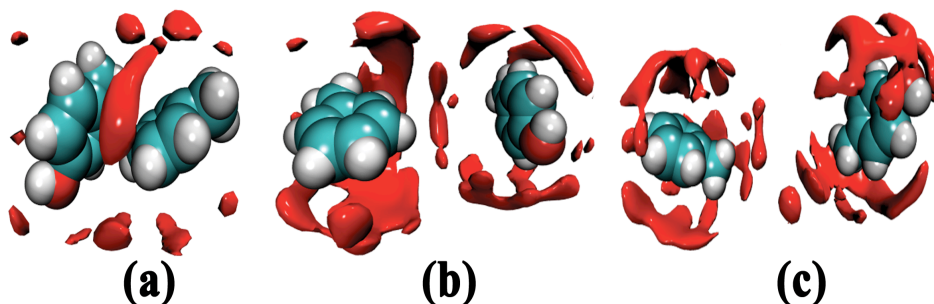


Figure A5.9: Spatial distribution functions (SDFs) for the PHE-TYR (a) contact pair ($r \leq 6.0$ Å); (b) solvent-shared ($6.0 \text{ Å} \leq r \leq 12.0$ Å) and (c) non-interacting ($r \geq 12.0$ Å). All densities are plotted at 50% above bulk.

5.6 References

1. Evans, D. A., *et al.*, Transition State Pi-Solvation by Aromatic Rings: An Electronic Contribution to Diels–Alder Reaction Diastereoselectivity. *Angew. Chem., Int. Ed. Engl.* **1987**, 26 (11), 1184.
2. Muehldorf, A. V., *et al.*, Aromatic-Aromatic Interactions in Molecular Recognition: A Family of Artificial Receptors for Thymine That Shows Both Face-to-Face and Edge-to-Face Orientations. *J. Am. Chem. Soc.* **1988**, 110 (19), 6561.
3. Rebek, J., Model Studies in Molecular Recognition. *Science* **1987**, 235 (4795), 1478.
4. Burley, S. K.; Petsko, G. A., Aromatic-Aromatic Interaction - a Mechanism of Protein-Structure Stabilization. *Science* **1985**, 229 (4708), 23.
5. Hunter, C. A., Sequence-Dependent DNA Structure: The Role of Base Stacking Interactions. *J. Mol. Biol.* **1993**, 230 (3), 1025.
6. Singh, J.; Thornton, J. M., The Interaction between Phenylalanine Rings in Proteins. *FEBS Lett.* **1985**, 191 (1), 1.
7. Saenger, W., *Principles of Nucleic Acid Structure*. Springer-Verlag: New York, 1984.
8. Rebek, J., *et al.*, Molecular Recognition - Hydrogen-Bonding and Stacking Interactions Stabilize a Model for Nucleic-Acid Structure. *J. Am. Chem. Soc.* **1987**, 109 (16), 5033.
9. Rebek, J., *et al.*, Molecular Recognition - Stacking Interactions Influence Watson-Crick Vs Hoogsteen Base-Pairing in a Model for Adenine Receptors. *Angew. Chem., Int. Ed. Engl.* **1987**, 26 (12), 1244.
10. Hunter, C. A., Meldola Lecture. The Role of Aromatic Interactions in Molecular Recognition. *Chem. Soc. Rev.* **1994**, 23 (2), 101.
11. Blundell, T., *et al.*, Aromatic Interactions. *Science* **1986**, 234 (4779), 1005.
12. Mitchell, J. B. O., *et al.*, Amino/Aromatic Interactions in Proteins: Is the Evidence Stacked against Hydrogen Bonding? *J. Mol. Biol.* **1994**, 239 (2), 315.
13. Chelli, R., *et al.*, Stacking and T-Shape Competition in Aromatic-Aromatic Amino Acid Interactions. *J. Am. Chem. Soc.* **2002**, 124 (21), 6133.
14. Brocchieri, L.; Karlin, S., Geometry of Interplanar Residue Contacts in Protein Structures. *Proc. Natl. Acad. Sci. U. S. A.* **1994**, 91 (20), 9297.
15. Jorgensen, W. L.; Severance, D. L., Aromatic Aromatic Interactions - Free-Energy Profiles for the Benzene Dimer in Water, Chloroform, and Liquid Benzene. *J. Am. Chem. Soc.* **1990**, 112 (12), 4768.
16. Chipot, C., *et al.*, Benzene Dimer: A Good Model for Pi-Pi Interactions in Proteins? A Comparison between the Benzene and the Toluene Dimers in the Gas Phase and in an Aqueous Solution. *J. Am. Chem. Soc.* **1996**, 118 (45), 11217.
17. Gervasio, F. L., *et al.*, Determination of the Potential of Mean Force of Aromatic Amino Acid Complexes in Various Solvents Using Molecular Dynamics Simulations: The Case of the Tryptophan-Histidine Pair. *J. Phys. Chem. B* **2001**, 105 (32), 7835.
18. Gamielien, M. R., *et al.*, Hydration-Determined Orientational Preferences in Aromatic Association from Benzene Dimer Free Energy Volumes. *J. Phys. Chem. B* **2012**, 116 (1), 324.
19. Gervasio, F. L., *et al.*, Interaction between Aromatic Residues. Molecular Dynamics and Ab Initio Exploration of the Potential Energy Surface of the Tryptophan-Histidine Pair. *J. Phys. Chem. B* **2000**, 104 (5), 1108.
20. Pitonak, M., *et al.*, Benzene Dimer: High-Level Wave Function and Density Functional Theory Calculations. *J. Chem. Theory Comput.* **2008**, 4 (11), 1829.
21. Dinadayalane, T. C.; Leszczynski, J., Geometries and Stabilities of Various Configurations of Benzene Dimer: Details of Novel V-Shaped Structure Revealed. *Struct. Chem.* **2009**, 20 (1), 11.

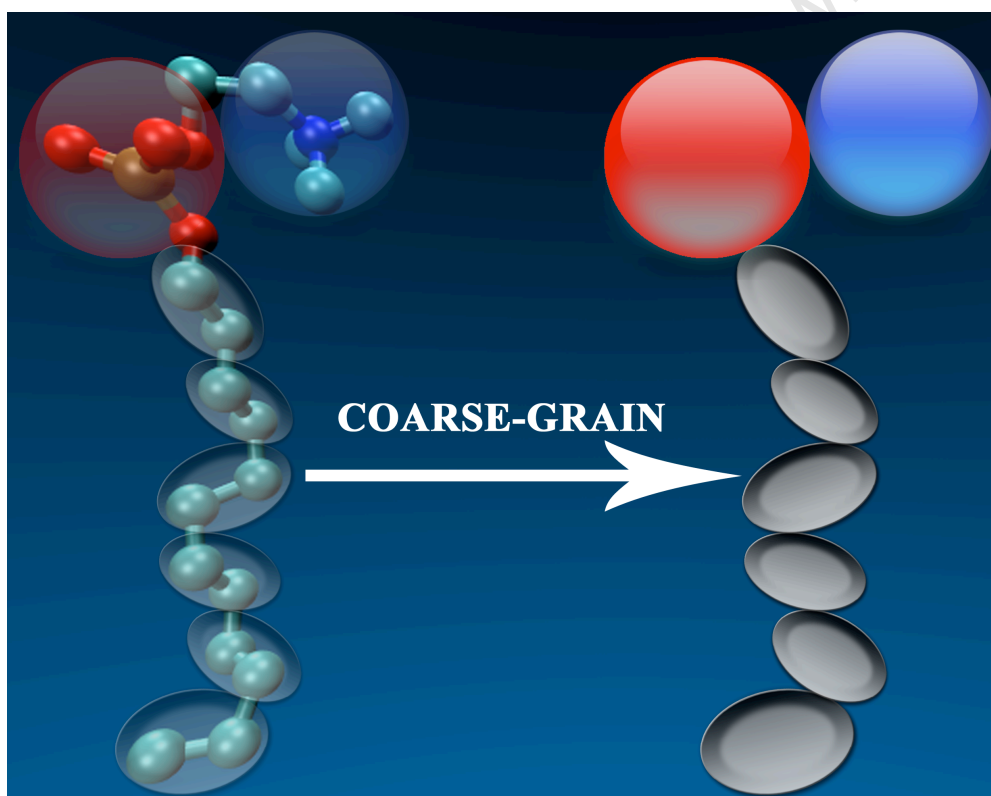
22. Distasio, R. A., *et al.*, On the T-Shaped Structures of the Benzene Dimer. *Chem. Phys. Lett.* **2007**, 437 (4-6), 277.
23. Podeszwa, R., *et al.*, Potential Energy Surface for the Benzene Dimer and Perturbational Analysis of Pi-Pi Interactions. *J. Phys. Chem. A* **2006**, 110 (34), 10345.
24. Sinnokrot, M. O., *et al.*, Estimates of the Ab Initio Limit for Pi-Pi Interactions: The Benzene Dimer. *J. Am. Chem. Soc.* **2002**, 124 (36), 10887.
25. Tsuzuki, S., *et al.*, Origin of Attraction and Directionality of the Pi/Pi Interaction: Model Chemistry Calculations of Benzene Dimer Interaction. *J. Am. Chem. Soc.* **2002**, 124 (1), 104.
26. Rezac, J.; Hobza, P., Benzene Dimer: Dynamic Structure and Thermodynamics Derived from on-the-Fly Ab Initio Dft-D Molecular Dynamic Simulations. *J. Chem. Theory Comput.* **2008**, 4 (11), 1835.
27. Podeszwa, R., *et al.*, Potential Energy Surface for the Benzene Dimer and Perturbational Analysis of Pi-Pi Interactions. *J. Phys. Chem. A* **2006**, 110, 10345.
28. Henson, B. F., *et al.*, Raman-Vibronic Double-Resonance Spectroscopy of Benzene Dimer Isotopomers. *J. Chem. Phys.* **1992**, 97 (4), 2189.
29. Headen, T. F., *et al.*, Structure of Pi-Pi Interactions in Aromatic Liquids. *J. Am. Chem. Soc.* **2010**, 132 (16), 5735.
30. Dinadayalane, T. C.; Leszczynski, J., Geometries and Stabilities of Various Configurations of Benzene Dimer: Details of Novel V-Shaped Structure Revealed. *Struct. Chem.* **2009**, 20, 11.
31. Sinnokrot, M. O.; Sherrill, C. D., High-Accuracy Quantum Mechanical Studies of Pi-Pi Interactions in Benzene Dimers. *J. Phys. Chem. A* **2006**, 110 (37), 10656.
32. Linse, P., Stacked or T-Shaped Benzene Dimer in Aqueous-Solution - a Molecular-Dynamics Study. *J. Am. Chem. Soc.* **1992**, 114 (11), 4366.
33. Linse, P., Orientation-Averaged Benzene Benzene Potential of Mean Force in Aqueous-Solution. *J. Am. Chem. Soc.* **1993**, 115 (19), 8793.
34. Gräfenstein, J.; Cremer, D., An Efficient Algorithm for the Density-Functional Theory Treatment of Dispersion Interactions. *J. Chem. Phys.* **2009**, 130 (12), 124105.
35. Cozzi, F., *et al.*, Dominance of Polar/.Pi. Over Charge-Transfer Effects in Stacked Phenyl Interactions. *J. Am. Chem. Soc.* **1993**, 115 (12), 5330.
36. Gervasio, F. L., *et al.*, The Nature of Intermolecular Interactions between Aromatic Amino Acid Residues. *Proteins: Struct., Funct., Bioinf.* **2002**, 48 (1), 117.
37. Godzik, A., *et al.*, Are Proteins Ideal Mixtures of Amino Acids? Analysis of Energy Parameter Sets. *Protein Sci.* **1995**, 4 (10), 2107.
38. Tsuzuki, S., *et al.*, Ab Initio Calculations of Structures and Interaction Energies of Toluene Dimers Including Ccsd(T) Level Electron Correlation Correction. *J. Chem. Phys.* **2005**, 122 (14), 144323.
39. Ringer, A. L., *et al.*, The Effect of Multiple Substituents on Sandwich and T-Shaped Pi-Pi Interactions. *Chemistry-a European Journal* **2006**, 12 (14), 3821.
40. Sinnokrot, M. O.; Sherrill, C. D., Substituent Effects in Pi-Pi Interactions: Sandwich and T-Shaped Configurations. *J. Am. Chem. Soc.* **2004**, 126 (24), 7690.
41. Sinnokrot, M. O.; Sherrill, C. D., Unexpected Substituent Effects in Face-to-Face Pi-Stacking Interactions. *J. Phys. Chem. A* **2003**, 107 (41), 8377.
42. Biswal, H. S., *et al.*, Structure of the Indole-Benzene Dimer Revisited. *J. Phys. Chem. A* **2011**, 115 (34), 9485.
43. Geng, Y., *et al.*, Accurately Characterizing the Pi-Pi Interaction Energies of Indole-Benzene Complexes. *The Journal of Physical Chemistry A* **2010**, 114 (10), 3576.
44. Braun, J., *et al.*, N-Pi Interactions in Indole-Benzene-H₆D₆ and Indole-Benzene-H₆D₆ Radical Cation Complexes. Mass Analyzed Threshold Ionization Experiments and Correlated Ab Initio Quantum Chemical Calculations. *J. Phys. Chem. A* **2003**, 107 (19), 3918.
45. Levitt, M.; Perutz, M. F., Aromatic Rings Act as Hydrogen-Bond Acceptors. *J. Mol. Biol.* **1988**, 201 (4), 751.

46. Yan, S.; Spangler, L. H., Intermolecular Modes of S1 P-Cresol Dimer; Information Concerning Structure and Dynamics. *J. Phys. Chem.* **1991**, 95 (10), 3915.
47. Isom, D. G., *et al.*, Charges in the Hydrophobic Interior of Proteins. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, 107 (37), 16096.
48. Dwyer, J. J., *et al.*, High Apparent Dielectric Constants in the Interior of a Protein Reflect Water Penetration. *Biophysical Journal* **2000**, 79 (3), 1610.
49. Ripoll, D. R., *et al.*, On the Orientation of the Backbone Dipoles in Native Folds. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, 102 (21), 7559.
50. Felder, C. E., *et al.*, A Server and Database for Dipole Moments of Proteins. *Nucleic Acids Res.* **2007**, 35 (suppl 2), W512.
51. Farzadfard, F., *et al.*, Beta-Sheet Capping: Signals That Initiate and Terminate B-Sheet Formation. *J. Struct. Biol.* **2008**, 161 (1), 101.
52. Denisov, V. P., *et al.*, Stabilization of Internal Charges in a Protein: Water Penetration or Conformational Change? *Biophys. J.* **2004**, 87 (6), 3982.
53. García, A. E.; Hummer, G., Water Penetration and Escape in Proteins. *Proteins: Struct., Funct., Bioinf.* **2000**, 38 (3), 261.
54. Frishman, D.; Argos, P., Knowledge-Based Protein Secondary Structure Assignment. *Proteins: Struct., Funct., Bioinf.* **1995**, 23 (4), 566.
55. Berg, J. M., *et al.*, *Biochemistry*. Fifth ed.; W. H. Freeman and Company: New York, 2002.
56. Snyder, P. W., *et al.*, Mechanism of the Hydrophobic Effect in the Biomolecular Recognition of Arylsulfonamides by Carbonic Anhydrase. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, 108 (44), 17889.
57. Setny, P., *et al.*, How Can Hydrophobic Association Be Enthalpy Driven? *J. Chem. Theory Comput.* **2010**, 6 (9), 2866.
58. Warme, P. K.; Morgan, R. S., A Survey of Amino Acid Side-Chain Interactions in 21 Proteins. *J. Mol. Biol.* **1978**, 118 (3), 289.
59. Liwo, A., *et al.*, Simulation of Protein Structure and Dynamics with the Coarse-Grained Unres Force Field. In *Coarse-Graining of Condensed Phase and Biomolecular Systems*, Voth, G. A., Ed. CRC Press: New York, 2009; p 108.
60. Marsili, S., *et al.*, Thermodynamics of Stacking Interactions in Proteins. *Phys. Chem. Chem. Phys.* **2008**, 10 (19), 2673.
61. Dzubiella, J., How Interface Geometry Dictates Water's Thermodynamic Signature in Hydrophobic Association. *J. Stat. Phys.* **2011**, 145 (2), 227.
62. Chandler, D., Interfaces and the Driving Force of Hydrophobic Assembly. *Nature* **2005**, 437 (7059), 640.
63. Spolar, R. S., *et al.*, Hydrophobic Effect in Protein Folding and Other Noncovalent Processes Involving Proteins. *Proc. Natl. Acad. Sci. U. S. A.* **1989**, 86 (21), 8382.
64. Dill, K. A., Dominant Forces in Protein Folding. *Biochemistry* **1990**, 29 (31), 7133.

PART 2

REDUCTION OF REPRESENTATION – COARSE-GRAIN DYNAMICS

“... most CG methods are best cast within the context of the Helmholtz free energy.” – Gregory A. Voth



(Left) all atom representation of a dodecylphosphocholine (DPC) and (Right) its coarse-grain (or reduced) representation.

Representing a system in a reduced manner, while maintaining connections to the all atom representation and biophysics, will result in the ability to simulate large biomolecules on biologically relevant timescales. The second part of this thesis deals with the parameterization of an effective coarse-grain (CG) interaction potential (from free energy hyper-volumes), development of a CG dynamics code and the subsequent CG dynamics simulation and analysis of liquid benzene.

CHAPTER 6

COARSE-GRAIN DYNAMICS AND CODE DEVELOPMENT

The explosive growth of computing power over the past two decades has led to the development of large-scale simulation techniques whose aim is to directly reproduce (or simulate), on a molecular level, experimental macroscopic observables for small systems. Over the years, these techniques have been extended to study macromolecular and biological systems. Unlike small systems, such as water and other organic molecules, the usefulness of these computer simulation techniques have been limited by the spatial (size) and temporal (time) scale of the system.¹⁻⁴ Large systems such as biomolecules have a greater number of degrees of freedom, which makes reproducing their experimental observables, by employing either mixed Hamiltonian (QM/MM) dynamics or all atom molecular dynamics, nearly impossible. Quantum mechanics solves the Schrodinger equation for motion of electrons, while the nuclei are at rest (Born-Oppenheimer approximation). When the nuclei move, a new wavefunction is generated and the motion of electrons has to be calculated again. This computation involves many second-order differentials, integrals etc and this limits the computation to a small number of atoms and short timescales (femto- to picoseconds), Figure 1.9.⁵ Statistical mechanics techniques, on the other hand, such as molecular dynamics are apt for simulating large systems for longer timescales (nano- to microsecond),⁶ because they do not calculate the motion of electrons (Figure 1.9). The motions of the atoms are treated classical mechanically and solved using Newton's equations of motion.

Grand challenge problems such as protein folding, which has an astronomically large configurational space, cannot currently be investigated using molecular dynamics. Further theoretical approximations, such as coarse-graining, have to be introduced to deal with the high computational cost. Coarse-graining is a method whereby the certain atomic and molecular degrees of freedom are removed by treating them as an average. In this chapter, a brief account of coarse-grain (CG) dynamics implemented with the Gay-Berne anisotropic interaction potential energy function is presented. First the

Hamiltonian is described, followed by the Euler equations for rotational rigid body dynamics and concluded with a description of the Gay-Berne potential, forces and torques.⁷ The derivation of the forces and torques⁸ is given in the Appendix.

6.1 Coarse-Grain Dynamics

By employing the CG dynamics formalism, certain molecular degrees of freedom are removed by treating their effect as an averaged over. This averaging has the consequence of reducing the amount of calculations to perform per MD step, which leads to a significant reduction in the computational cost. Thus, this methodology offers an attractive route for simulations of very large (macromolecular or biological) systems that could reach biologically relevant time–(microsecond)–scales. It has been shown by Orsi et al. that CG dynamics can be approximately 130 times faster than the dynamics of its all atom counterpart.⁹ A molecule can be coarse-grained in several ways,^{3, 10, 11} but the choice has to make physical sense. For example, the alkane chain, hexane (C_6H_{14}), it detracts from the conformational accuracy of the model if the entire molecule is treated as a sphere or ellipsoid.

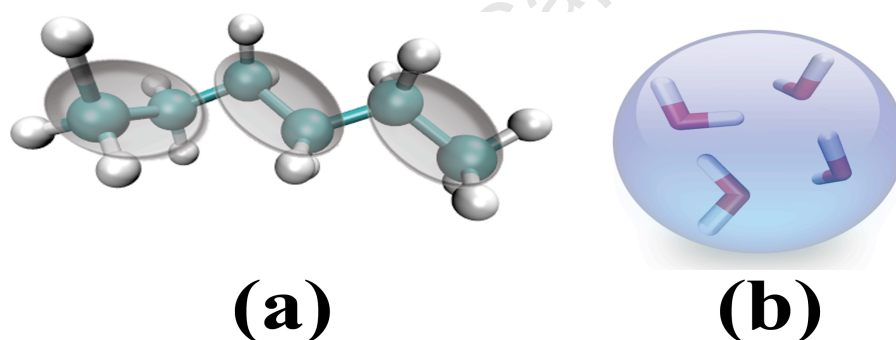


Figure 6.2: (a) A hexane (20 atoms) molecule coarse-grained to 3 bonded ellipsoids; and (b) coarse-grained quartet of water molecules represented as a sphere.

Rather, a combination of two bonded carbons (CH_2-CH_2), uniting the hydrogen and carbon atoms, into a single “bead” or site can be done, as shown in Figure 6.2a. The result is that the hexane is represented as a chain with 3 ellipsoidally shaped beads.^{9, 12, 13} Here, atomic sites have been reduced to coarse-grained sites resulting in more a 6-fold decrease in the systems degrees of freedom (from 20 to 3). This approximation has been shown to be reasonable for the ethylene group and has been applied to long aliphatic chains found in tail moiety of lipids.^{9, 12, 13} Other approximations include conflating 4 or 5 water molecules into a single spherical CG site, Figure 6.2b,¹⁴ and representing a single benzene molecule as an ellipsoid,⁸ which is the subject of this work (Chapter 8).

A reduced representation of the system is also advantageous in terms of overcoming problems such as heavy-atom-to-hydrogen-bond vibrations, which occur on very short time scales. By removing these vibrations, by fixing the bond length, allows the use of larger timesteps without too much sacrifice in the accuracy and stability of the simulation.¹ Usually CG particles are represented as ellipsoidally shaped particles. This infers that the interaction potential, between non-bonded particles, should not only depend on the inter-particle distance but also on the relative orientation. An anisotropic interaction potential, such as the Gay-Berne potential,⁷ which treats the particles as being rigid, is often used in CG dynamics simulations. Such a potential can be incorporated into a rigid body dynamics algorithm, Figure 6.3. The rigid body Hamiltonian, and algorithm, has a similar form to that of atomistic molecular dynamics, but includes an additional orientational/rotational term. The total potential energy function remains similar to molecular mechanics in that the bonded and non-bonded terms are functions of the coordinates only, as pointed out earlier (Chapter 2, section 2.1.1).

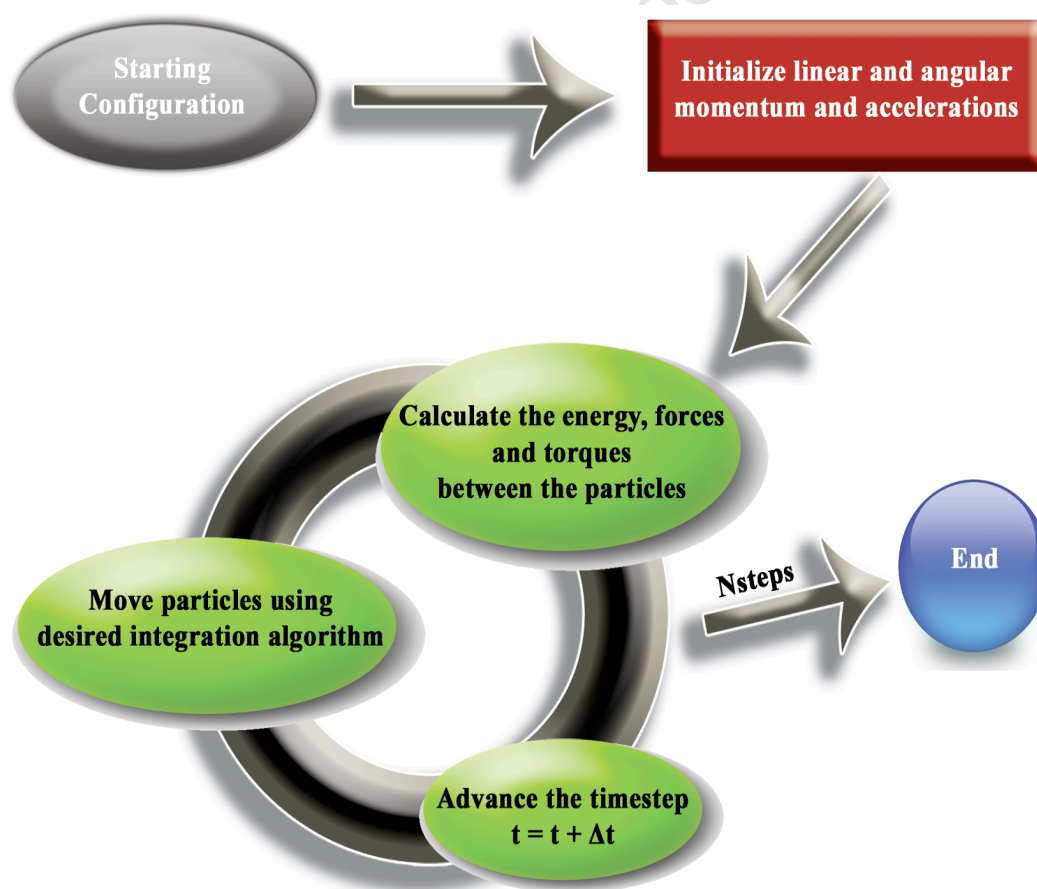


Figure 6.3: Simplified scheme showing the molecular (and rigid body) dynamics algorithm. Begin with some starting configuration of the system, proceed to dynamics for specified number of steps (Nsteps) and then end dynamics.

The Gay-Berne potential differs from the Lennard-Jones potential by being dependent on the inter-particle orientation, which requires the additional orientation term. The Euler equations of motion, which allows for the integration of the rotational kinematic equations, is similar to the Newtonian equations of motion used in molecular dynamics.

6.2 The Hamiltonian

The Hamiltonian for CG mechanics is similar to the one used in molecular mechanics (MM). However, the orientational and rotational terms are incorporated into the potential (V) and kinetic (T) energy terms respectively. This is effectively the Hamiltonian for a rigid body (H_{RBD}), with an anisotropic inter-particle potential energy function (V^{RGB}), and can be expressed as:

$$\mathbf{H}_{RBD} = \sum_{i=1}^N T_i^{RBD} + \sum_{i=1}^N V_i^{RBD} \quad 6.1$$

where T_i^{RBD} is the kinetic energy contribution to the total energy for i -th particle, which includes translation and rotational velocities, and V_i^{RBD} is the anisotropic interaction energy contribution from the i -th particle to the total energy. The linear momentum and spatial coordinates have been defined previously (Chapter 2), while the angular momentum is given by

$$\mathbf{i} = \frac{\partial H}{\partial \phi} \frac{\partial H}{\partial \theta} \frac{\partial H}{\partial \psi} = \nabla_{\phi, \theta, \psi} = \frac{\boldsymbol{\tau}}{I}, \quad 6.2$$

where $\mathbf{i}$ is the angular momentum, $\boldsymbol{\tau}$ the torque and I is the diagonal of inertia matrix. The first derivative of the Euler angles are given by

$$\dot{\phi} = \frac{\omega_{x,y,z}}{I}, \quad 6.3.1$$

$$\dot{\theta} = \frac{\omega_{x,y}}{I} \quad 6.3.2$$

$$\dot{\psi} = \frac{\omega_{x,y}}{I}. \quad 6.3.3$$

6.2.1 Coarse-grain (CG) mechanics

Much like the empirical potential energy function used in MM, the CG potential energy function is composed of bonded and non-bonded interaction terms. Thus, two types of potential energy functions exist: (1) a “full” potential energy that has both bonded and non-bonded interactions; and (2) a “partial” potential energy function

for simulating *only* non-bonded CG particles using only non-bonded interactions. Only the “partial” potential energy function is discussed, as this thesis is only concerned with the CG dynamics of non-bonded interacting ellipsoidal particles.

The “partial” potential energy function, described by *only* non-bonded interactions, is composed of a dispersion-like interaction potential, such as the Gay-Berne potential, and an electrostatic potential determined using the electrostatic multipole expansion. The forces ($\mathbf{F} = -\frac{\partial V^{RBD}}{\partial \xi}$) and torques ($\boldsymbol{\tau} = \mathbf{r} \times \mathbf{F} = \mathbf{r} \times \left(-\frac{\partial V^{RBD}}{\partial \xi}\right)$) are derived in the Appendix.

The total potential energy of the non-bonded interactions is expressed in a similar fashion as in MM, in that it is the sum of a dispersion-like (i.e., Gay-Berne) and electrostatic potential terms⁸

$$V^{RBD} = E_{GB}(r, \theta_1, \theta_2, \phi) + E_{Elec}. \quad 6.4$$

The first term is the Gay-Berne anisotropic inter-particle potential energy function and has the following form:

$$E_{GB}(r, \theta_1, \theta_2, \phi) = 4\varepsilon(\theta_1, \theta_2, \phi) \left[\left(\frac{\sigma^o}{r_{ij} - \sigma(\theta_1, \theta_2, \phi) + \sigma^o} \right)^{12} - \left(\frac{\sigma^o}{r_{ij} - \sigma(\theta_1, \theta_2, \phi) + \sigma^o} \right)^6 \right]. \quad 6.5$$

The Gay-Berne potential is discussed later in Section 6.5.1, and with greater detail in Chapter 7. The electrostatics, which have not been incorporated into the CG dynamics code, can be approximated by the multipole expansion

$$E_{Elec} = \frac{1}{4\pi\epsilon_o} \sum_{n=1}^{\infty} \frac{1}{R^{(a+b+1)}} \int_{all\ space} \mathbf{r}^n P_n(\cos\theta) \rho(\mathbf{r}) d\tau, \quad 6.6$$

where R is the distance between the particles, $P_n(x)$ is a Legendre polynomial of order n , the angular term $\cos\theta$ is given by the dot product of the two vectors $\hat{\mathbf{r}}$ and $\hat{\mathbf{R}}$, and a and b are the multipole expansion coefficients (see Chapter 1).

6.3 Rotational And Rigid Body Dynamics

6.3.1 Euler equations of motion

The linear forces in system result in the translation motion, but how do we deal with the rotation of a molecule or a CG particle? Undoubtedly the rotational force, also known as torque, is required to describe the rotational motion of an object about its

centre of mass. The torque, τ , is calculated by taking the cross product of the rotation axis, around which rotation occurs, and the force¹⁵

$$\tau = r \times F. \quad 6.7$$

Here r is the positions of atoms relative to the centre of mass, and F is the force (calculated from molecular positions and orientations). The torques enters the rotational equations of motion (rotational kinematics) in the same way as the forces enter the translation equations of motion. The nature of the orientation space, however, guarantees that the equations of reorientational motion will not be as simple as the translational kinematics. We present here the necessary mechanics and mathematics to describe the rotational motion of a rigid body.

6.3.2 Orientation matrix

The orientation of a rigid body specifies the relation between an axis system fixed in space and one fixed with respect to the system, as well as where the inertia tensor is diagonal.¹⁵ For example, in the case of benzene, the vector normal to the plane can be used to describe its orientation, Figure 6.4.

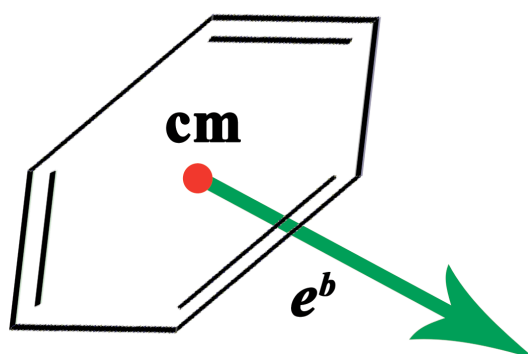


Figure 6.4: The vector (e^b), normal to plane of the benzene molecule, can be used to define orientation of the benzene molecule.

The 3D orientation of a molecule or any object, in space, can be completely described by the orientation matrix, which is a 9-component matrix

$$A = \begin{pmatrix} \cos\phi\cos\psi - \sin\phi\cos\theta\sin\psi & \sin\phi\cos\psi + \cos\phi\cos\theta\sin\psi & \sin\theta\sin\psi \\ -\cos\phi\sin\psi - \sin\phi\cos\theta\cos\psi & -\sin\phi\sin\psi + \cos\phi\cos\theta\cos\psi & \sin\theta\cos\psi \\ \sin\phi\sin\theta & -\cos\phi\sin\theta & \cos\theta \end{pmatrix}. \quad 6.8$$

Only 3 independent quantities (or generalized coordinates) are required to define the orientation (or rotation) matrix, referred to as A . These generalized coordinates are referred to as the Euler angles, ϕ - θ - ψ , Figure 6.5.¹⁶

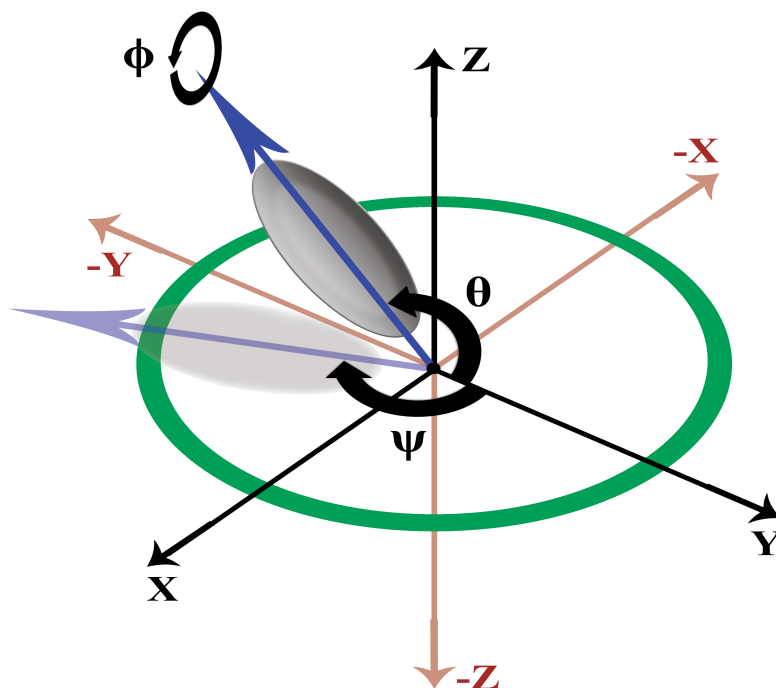


Figure 6.5: Pictorial description of the Euler angles; θ is the elevation, ψ the heading and ϕ the bank.¹⁵ First a rotation of ψ degrees is made about the original z-axis, followed by a θ degrees rotation around the x-axis. Finally the orientation is complete by rotating by ϕ degrees around the new z-axis.

6.3.3 Rotational frames of reference

If $\mathbf{e}^b$ is a vector fixed in the molecular frame (e.g. a vector defined to be normal to the benzene molecular plane, Figure 6.4), then $\mathbf{e}^b$ will not change with time.¹⁵ However, in the space-fixed coordinates the components of the vector, $\mathbf{e}^s$, will change¹⁵

$$\dot{\mathbf{e}}^s = \dot{\mathbf{e}}^b + \boldsymbol{\omega}^s \times \mathbf{e}^s = \boldsymbol{\omega}^s \times \mathbf{e}^s. \quad 6.9$$

6.3.4 Angular momentum and velocity

The time derivative of the angular velocity is the angular force, known as the torque. It dictates the magnitude and direction of the angular velocity, much in the same way as the translational force acts on the translation velocity. It is most convenient to connect the angular velocity ($\boldsymbol{\omega}$) with the angular momentum ($\mathbf{l}$), in the body-fixed principal axis system, where the moment of inertia tensor ($\mathbf{I}$) is diagonal¹⁵

$$\dot{\mathbf{l}}^s = \boldsymbol{\tau}^s, \quad 6.10$$

$$\dot{\mathbf{l}}^b + \boldsymbol{\omega} \times \mathbf{l}^b = \boldsymbol{\tau}^b, \quad 6.11$$

where the angular momentum is defined as

$$l = I\omega . \quad 6.12$$

Therefore, the resulting equations for the components of the ω , in the body-fixed frame, are shown in Equations 6.13.1 – 6.13.3.

$$\dot{\omega}_x = \frac{\tau_x^b}{I_{xx}} \left(\frac{I_{yy} - I_{zz}}{I_{xx}} \right) \omega_y^b \omega_z^b, \quad 6.13.1$$

$$\dot{\omega}_y = \frac{\tau_y^b}{I_{yy}} \left(\frac{I_{zz} - I_{xx}}{I_{yy}} \right) \omega_z^b \omega_x^b, \quad 6.13.2$$

$$\dot{\omega}_z = \frac{\tau_z^b}{I_{zz}} \left(\frac{I_{xx} - I_{yy}}{I_{zz}} \right) \omega_x^b \omega_y^b, \quad 6.13.3$$

where I_{xx} , I_{yy} and I_{zz} are the diagonals of the inertia tensor, i.e. the 3 principals of the moment of inertia in the body-fixed frame. Conversion to space-fixed space and back is quite easily done with knowledge of the orientation matrix and its inverse (or transpose).

$$\tau^b = \tau^s \times A, \quad 6.14$$

$$\omega^s = A^{-1} \cdot \omega^b = A^T \cdot \omega^b, \quad 6.15$$

since the inversion of an orthogonal matrix (A) is its transpose.¹⁵ The only piece of information that is left to be determined is how the orientation changes with time. This proceeds in a similar fashion as the calculated components of the angular velocity, ω , in terms of the Euler angles. The first derivative of the orientation ($\dot{\phi}$, $\dot{\theta}$ and $\dot{\psi}$) is given in Equations 6.16.1 – 6.16.3.

$$\dot{\phi} = -\omega_x^s \frac{\sin \phi \cos \theta}{\sin \theta} + \omega_y^s \frac{\cos \phi \cos \theta}{\sin \theta} + \omega_z^s, \quad 6.16.1$$

$$\dot{\theta} = \omega_x^s \cos \theta + \omega_y^s \sin \phi \quad 6.16.2$$

$$\dot{\psi} = \omega_x^s \frac{\sin \phi}{\sin \theta} - \omega_y^s \frac{\cos \phi}{\sin \theta}. \quad 6.16.3$$

Much like in the regular translation kinematics equations, the above rotational equations may also be solved in a stepwise fashion. However, the above equations have a major drawback in that it contains a $\sin \theta$ term in the denominator, and as we approach 0 or π the orientation equations become divergent.¹⁵ Of course the molecular motions are unaffected, but because of the choice of axes the ϕ and ψ angles become degenerate.¹⁵ Thus, it is unsatisfactory to have such an event occur if we are interested in simulating rigid CG molecules.¹⁷ We could overcome this drawback by decreasing the

size of the dynamics timestep, but this would defeat the purpose of coarse-graining where the objective is to simulate at larger timescales ($\geq$ microseconds).

6.3.5 Quaternions: A substitute for Euler angles

A much more elegant solution to the problem of divergence in the orientational equations of motion has been proposed, in place of the original Euler angles.¹⁵ It was recognized by Evans¹⁸ that singularity-free equations could not be obtained in terms of 3 independent variables. This led to the suggestion that the generalized coordinates can be described using 4 quaternions. Quaternions fulfill all the requirements of having well-behaved equations of motion, i.e. there are no divergent or degenerate terms. The 4 quaternions are linked by one algebraic equation, such that there is only one “redundant” variable.¹⁹

A quaternion, $\mathbf{Q}$, is a set of four scalar quantities

$$\mathbf{Q} = (q_0, q_1, q_2, q_3), \quad 6.17$$

where it is often useful to think of the last three elements as constituting a vector.¹⁵ The property of interest is that they satisfy the following constraint

$$q_0^2 + q_1^2 + q_2^2 + q_3^2 = 1. \quad 6.18$$

Therefore, quaternions can be used to represent the orientation of a rigid body. In the Euler convention, it is most convenient to define them as

$$q_0 = \cos \frac{1}{2} \theta \cos \frac{1}{2} (\phi + \psi), \quad 6.19.1$$

$$q_1 = \sin \frac{1}{2} \theta \cos \frac{1}{2} (\phi - \psi), \quad 6.19.2$$

$$q_2 = \sin \frac{1}{2} \theta \sin \frac{1}{2} (\phi - \psi), \quad 6.19.3$$

$$q_3 = \cos \frac{1}{2} \theta \sin \frac{1}{2} (\phi + \psi), \quad 6.19.4$$

and the orientation (or rotation) matrix becomes

$$\mathbf{A} = \begin{pmatrix} q_0^2 + q_1^2 - q_2^2 - q_3^2 & 2(q_1q_2 + q_0q_3) & 2(q_1q_3 - q_0q_2) \\ 2(q_1q_2 - q_0q_3) & q_0^2 - q_1^2 + q_2^2 - q_3^2 & 2(q_2q_3 + q_0q_1) \\ 2(q_1q_3 + q_0q_2) & 2(q_2q_3 - q_0q_1) & q_0^2 - q_1^2 - q_2^2 + q_3^2 \end{pmatrix}. \quad 6.20$$

The quaternions, for each particle, can then satisfy the equations of motion,¹⁵ as shown by their time derivatives

$$\begin{pmatrix} \dot{q}_0 \\ \dot{q}_1 \\ \dot{q}_2 \\ \dot{q}_3 \end{pmatrix} = \frac{1}{2} \begin{pmatrix} q_0 & -q_1 & -q_2 & -q_3 \\ q_1 & q_0 & -q_3 & q_2 \\ q_2 & q_3 & q_0 & -q_1 \\ q_3 & -q_2 & q_1 & q_0 \end{pmatrix} \begin{pmatrix} 0 \\ \omega_x^b \\ \omega_y^b \\ \omega_z^b \end{pmatrix}. \quad 6.21$$

For further information on quaternions, their properties and conversion from quaternions to Euler angles refer to the Appendix.

6.4 Rotational Kinematics

The equations of motion using quaternions with knowledge of the orientation matrix, to transform between body- and space-fixed coordinates, contain no unpleasant singularities.¹⁹ They are a system of first-order differential equations and can be quite nicely solved with the Gear-Predictor-corrector method.¹⁵ However, this involves a lot of variable storage. It requires storage of all 4 quaternions together with their 4 time derivatives and also the angular velocities with its 4 time derivatives, for each particle.¹⁵ It is possible to eliminate the angular velocities from the equations of motion all together; thus, obtaining second-order differential equations for the quaternions.²⁰ An advantage of this approach is that, for multipolar potentials, the torques may be expressed very easily as derivatives of the potential with respect to the quaternions, thereby simplifying the geometry.¹⁵

As in the translational case an alternative to the Gear-Predictor-corrector algorithm would be preferable, and a leapfrog formulation for quaternions has been proposed.²¹ The simple leapfrog previously described (Chapter 2) cannot be used directly, because the quaternion derivatives ($\dot{q}$) appearing in the algorithm depend not only the angular velocity but also on $\mathbf{q}$ itself. However, a modified leapfrog may be applied in this case.²² The method is based on stored values of $\mathbf{F}(t - \frac{1}{2}\Delta t)$, $\mathbf{q}(t)$ and the torques $\boldsymbol{\tau}(t)$, computed from the positions and orientations only.¹⁵ The leapfrog algorithm then proceeds as usual. Here, we discuss a rigid body velocity Verlet algorithm implemented using quaternions describing the orientation.

6.4.1 Integration schemes II

An alternative to the leapfrog algorithm is the velocity Verlet algorithm. As the name implies, it not only calculates the coordinates, orientation, forces and torques at each time step but also the velocities. This algorithm allows the velocities to be calculated at each timestep, which is most desirable for the determining diffusion constants and other velocity related properties such as viscosity. Martyna et al. have

developed this method,²³ and this section briefly describes the method and its applications to rigid body dynamics.

The standard velocity Verlet algorithm²⁴⁻²⁶ involves the following equations to calculate the position and momentum at $t = t + \Delta t$ from $t = t$

$$\mathbf{r}(t + \Delta t) = \mathbf{r}(t) + \dot{\mathbf{r}}(t)\Delta t + \frac{\ddot{\mathbf{r}}(t)}{2}(\Delta t)^2 \quad 6.22$$

$$\dot{\mathbf{r}}(t + \Delta t) = \dot{\mathbf{r}}(t) + \frac{\ddot{\mathbf{r}}(t)\Delta t + \ddot{\mathbf{r}}(t + \Delta t)}{2}\Delta t \quad 6.23$$

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

$$iL = \sum_i^N \left(\dot{\mathbf{r}}_i \frac{\partial}{\partial \mathbf{r}_i} + \dot{\mathbf{p}}_i \frac{\partial}{\partial \mathbf{p}_i} \right), \quad 6.24$$

and the time evolution operator that propagates the system from $t = 0$ to $t = t$ is then $\exp(iL\Delta t)$. The time evolution of any dynamic variable, starting at $\xi(0)$, is given by

$$\xi(t) = \exp(iLt)\xi(0) = [\exp(iL\Delta t)]^{N_t} \xi(0) + O(N_t\Delta t^4) \quad 6.25$$

where $t = N_t\Delta t$. A useful technique is to split the operator into two parts, namely the position and momentum operators ($iL = iL_r + iL_p$):

$$iL_r = \sum_i^N \dot{\mathbf{r}}_i \frac{\partial}{\partial \mathbf{r}_i} \quad 6.26$$

$$iL_p = \sum_i^N \dot{\mathbf{p}}_i \frac{\partial}{\partial \mathbf{p}_i}. \quad 6.27$$

The time evolution of the system is then approximated by:

$$\exp(iL\Delta t) = \exp\left(iL_p \frac{\Delta t}{2}\right) \exp(iL_r\Delta t) \exp(iL_p\Delta t). \quad 6.28$$

This approach has the effect of propagating the positions and momenta separately, and can be easily seen in the form of

$$\exp(iL_r\Delta t)\xi(\mathbf{r}^N, \mathbf{p}^N) = \xi(\mathbf{r}^N + \dot{\mathbf{r}}^N, \mathbf{p}^N) \quad 6.29$$

$$\exp(iL_p\Delta t)\xi(\mathbf{r}^N, \mathbf{p}^N) = \xi(\mathbf{p}^N + \dot{\mathbf{p}}^N, \mathbf{r}^N). \quad 6.30$$

For systems that have forces that act on different time scales this approach is also very useful as the Liouville operators can be split into multiple terms, where multiple timesteps can be used. These are not in this thesis but it is worth noting for future applications to multiscale CG dynamics simulation methods. Different parts of the system can be treated with different timesteps, with a smaller timestep for the finer motions and a larger timestep for the environment. Martyna et al. describe the above form of the Liouville operator and splitting technique.²³ They present the Martyna, Tuckerman and Klein (MTK) equations of motion, which incorporates a chain of Nose-Hoover^{27, 28} thermostats for NVT dynamics, as well as the Liouville operator and propagators used to integrate them.

Now, when including the rotational equations (in the form of quaternions) into the kinematics equations we proceed in the following manner.¹⁵ The first step in the algorithm is to bring the angular momentum up to the half step

$$\mathbf{l}^s\left(t + \frac{\Delta t}{2}\right) = \mathbf{l}^s(t) + \frac{\Delta t}{2} \boldsymbol{\tau}^s(t). \quad 6.31$$

From this, along with transformation equations, we can calculate the body-fixed angular velocity and, hence, the time derivative of the quaternions, $\dot{\mathbf{q}}$. Then the guess of quaternion at the half step, $(\mathbf{q}(t + \frac{1}{2}\Delta t))$, can be made¹⁵

$$\mathbf{q}\left(t + \frac{\Delta t}{2}\right) = \mathbf{q}(t) + \frac{1}{2} \dot{\mathbf{q}}\left(t + \frac{\Delta t}{2}\right). \quad 6.32$$

Note, none of these values need be stored permanently. The sole aim is to provide auxiliary equations for calculating the $\mathbf{q}(t + \frac{1}{2}\Delta t)$, so that transformations from space- to body-fixed angular momentum, and the calculation of $\dot{\mathbf{q}}$, can be calculated at the half step.¹⁵ The next procedure is to update the angular momentum and orientation to complete the time step. This is performed when the forces ($\mathbf{F}(t + \Delta t)$) and torques ($\boldsymbol{\tau}^s(t + \Delta t)$) are calculated at $t = t + \Delta t$ giving

$$\mathbf{l}^s(t + \Delta t) = \mathbf{l}^s\left(t + \frac{\Delta t}{2}\right) + \frac{\Delta t}{2} \boldsymbol{\tau}^s(t + \Delta t). \quad 6.33$$

Finally, updating the quaternions completes the integration algorithm

$$\mathbf{q}(t + \Delta t) = \mathbf{q}\left(t + \frac{\Delta t}{2}\right) + \frac{1}{2} \dot{\mathbf{q}}\left(t + \frac{\Delta t}{2}\right). \quad 6.34$$

Special care must be taken at every step to ensure that the orientation (or rotation) matrix remains orthogonal such that the angular velocity converges.¹⁵ If these requirements are not met the angular momentum, and hence the total momentum, will not be conserved in the simulation.¹⁵

6.5 Gay-Berne Potential and Forces

6.5.1 The Gay-Berne interaction potential

The Gay-Berne potential,^{7, 8, 29, 30} developed by Gay and Berne, is a generalized anisotropic form of the Lennard-Jones 12-6 interaction potential. The Gay-Berne potential is described by the distance and three angular variables describing the relative orientation of the molecules, Figure 1.xxx. It was originally adapted from the Berne-Pechukas potential,³¹ and over the years has been adapted to include biaxial molecules.³⁰ The potential was initially intended to describe the interaction between liquid crystals, but since has also been used to describe the non-bonded interactions in CG molecular dynamics simulations of macromolecules.^{9, 32}

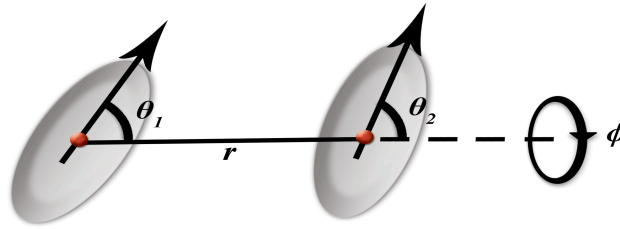


Figure 6.6: Generalized molecular anisotropic definition that describes the distance (r) between and relative orientation ($\theta_1 = \mathbf{r} \cdot \mathbf{u}_{1(i)}$, $\theta_2 = \mathbf{r} \cdot \mathbf{u}_{2(j)}$ and $\phi = \mathbf{u}_{1(i)} \cdot \mathbf{u}_{2(j)}$) of two non-bonded coarse-grain particles. The vectors ($\mathbf{u}_{1(i)}$ and $\mathbf{u}_{2(j)}$) are defined along the molecular principal axis, unless otherwise stated.

The Gay-Berne interaction potential energy function incorporates the orientational contributions to the energy, forces and torques. The original uniaxial Gay-Berne potential energy function has the following form:

$$U_{GB} = 4\epsilon_{ij} \left(\hat{u}_i, \hat{u}_j, \hat{r}_{ij} \right) \left[x_{ij}^{12} - x_{ij}^6 \right] \quad 6.35$$

where the inter-particle distance, x_{ij} , has been modified (shifted) in the following way

$$x_{ij} = \frac{\sigma_{ij}^o}{r_{ij} - \sigma_{ij}(\hat{u}_i, \hat{u}_j, \hat{r}_{ij}) + \sigma_{ij}^o}. \quad 6.36.1$$

The combining rules for the interaction strength, ϵ_{ij} , and inter-particle range, σ_{ij}^o , parameters have been modified as follows

$$\epsilon_{ij}(\hat{u}_i, \hat{u}_j, \hat{r}_{ij}) = \epsilon_{ij}^o \left[\epsilon_{ij}^1(\hat{u}_i, \hat{u}_j) \right]^v \left[\epsilon_{ij}^2(\hat{u}_i, \hat{u}_j, \hat{r}_{ij}) \right]^\mu \quad 6.36.2$$

$$\sigma_{ij}^o = \sqrt{(\sigma_{i,\perp})^2 + (\sigma_{j,\perp})^2}. \quad 6.36.3$$

A more in depth discussion of this potential is been provided in Chapter 7, while only the important details are given here. As can be seen from the above equations that the interaction strength (ϵ_{ij}) and range (σ_{ij}) parameters are directly coupled to the orientation of the particles as pointed out in Chapter 7.

6.5.2 Forces and torques

The force, in any computational simulation, is calculated as the negative derivative of the potential energy function. In all atom MD simulations the Lennard-Jones potential and Coulomb energies are used. They are quite easily differentiated due to their univariate nature. However, in liquid crystal and macromolecular dynamics simulations the orientation of the molecules has to be taken into account as well. In other words a different orientation results in different energies and forces. To determine the forces and torques requires knowledge of the partial derivative of the interaction potential, which is a made more complicated by having angular dependence. The force for this potential has been derived is by Golubkov and Ren,⁸ while Allen and Germano provide generalized expressions for these Corner type potentials.³³

The total translation force for the Gay-Berne potential, in terms of the partial derivatives for the distance and angular variables, is obtained as follows:^{8, 33}

$$F_i = -\frac{\partial U}{\partial \hat{r}_{ij}} \hat{r}_{ij} - \frac{\partial U}{\partial (\hat{r}_{ij} \cdot \hat{u}_i)} \hat{u}_i - \frac{\partial U}{\partial (\hat{r}_{ij} \cdot \hat{u}_j)} \hat{u}_j = -F_j. \quad 6.37$$

And, the torque exerted on each molecule, based on their relative orientation, is obtained as follows:^{8, 33}

$$\tau_i = \frac{\partial U}{\partial (\hat{r}_{ij} \cdot \hat{u}_i)} (\hat{r}_{ij} \times \hat{u}_i) + \frac{\partial U}{\partial (\hat{u}_i \cdot \hat{u}_j)} (\hat{u}_j \times \hat{u}_i) \quad 6.38.1$$

$$\tau_j = \frac{\partial U}{\partial (\hat{r}_{ij} \cdot \hat{u}_j)} (\hat{r}_{ij} \times \hat{u}_j) - \frac{\partial U}{\partial (\hat{u}_i \cdot \hat{u}_j)} (\hat{u}_j \times \hat{u}_i), \quad 6.39.2$$

where τ_i is the torque for the i -th molecule and τ_j is the torque for the j -th molecule.

6.8 References

1. Nielsen, S. O., *et al.*, Coarse Grain Models and the Computer Simulation of Soft Materials. *J. Phys.: Condens. Matter* **2004**, *16* (15), R481.
2. Marrink, S. J.; Mark, A. E., The Mechanism of Vesicle Fusion as Revealed by Molecular Dynamics Simulations. *J. Am. Chem. Soc.* **2003**, *125* (37), 11144.
3. Tries, V.; Paul, W., Modeling Polyethylene with the Bond Fluctuation Model. *J. Chem. Phys.* **1997**, *106* (2), 738.
4. Derreumaux, P.; Mousseau, N., Coarse-Grained Protein Molecular Dynamics Simulations. *J. Chem. Phys.* **2007**, *126* (2), 025101.
5. Barnett, C. B., *et al.*, Molecular Details from Computational Reaction Dynamics for the Cellobiohydrolase I Glycosylation Reaction. *J. Am. Chem. Soc.* **2011**, *133* (48), 19474.
6. Haliloglu, T.; Bahar, I., Coarse-Grained Simulations of Conformational Dynamics of Proteins: Applications to Apomyoglobin. *Proteins: Struct., Funct., Bioinf.* **1998**, *31*, 271.
7. Gay, J. G.; Berne, B. J., Modification of the Overlap Potential to Mimic a Linear Site-Site Potential. *J. Chem. Phys.* **1981**, *74* (6), 3316.
8. Golubkov, P. A.; Ren, P. Y., Generalized Coarse-Grained Model Based on Point Multipole and Gay-Berne Potentials. *J. Chem. Phys.* **2006**, *125* (6), 064103.
9. Orsi, M., *et al.*, A Quantitative Coarse-Grain Model for Lipid Bilayers. *J. Phys. Chem. B* **2007**, *112* (3), 802.
10. Paul, W., *et al.*, Structure-Property Correlation of Polymers, a Monte Carlo Approach. *Macromolecules* **1991**, *24* (23), 6332.
11. Binder, K., *Monte Carlo and Molecular Dynamics Simulations in Polymer Science*. Oxford University Press: New York, 1995.
12. Whitehead, L., *et al.*, Molecular Dynamics Simulation of the Hydrocarbon Region of a Biomembrane Using a Reduced Representation Model. *J. Comput. Chem.* **2001**, *22* (14), 1622.
13. Fukunaga, H., *et al.*, A Coarse-Graining Procedure for Flexible Polymer Chains with Bonded and Nonbonded Interactions. *J. Chem. Phys.* **2002**, *116* (18), 8183.
14. Darre, L., *et al.*, Another Coarse Grain Model for Aqueous Solvation: Wat Four? *J. Chem. Theory Comput.* **2010**, *6* (12), 3793.
15. Allen, M. P.; Tildesly, D. J., *Computer Simulation of Liquids*. Oxford University Press: Oxford, 1987.
16. Goldstein, H., *Classical Mechanics*. Addison-Wesley: Reading, MA, 1980.
17. Barojas, J., *et al.*, Simulation of Diatomic Homonuclear Liquids. *Phys. Rev. A* **1973**, *7* (3), 1092.
18. Evans, D. J., On the Representation of Orientation Space. *Mol. Phys.* **1977**, *34* (2), 317.
19. Evans, D. J.; Murad, S., Singularity Free Algorithm for Molecular Dynamics Simulation of Rigid Polyatomics. *Mol. Phys.* **1977**, *34* (2), 327.
20. Powles, J. G., *et al.*, A Computer Simulation for a Simple Model of Liquid Hydrogen Chloride. *Mol. Phys.* **1979**, *38* (3), 893.
21. Fincham, D., An Algorithm for the Rotational Motion of Rigid Molecules. *CCP5 Quarterly* **1981**, *8*, 45.
22. Potter, D., *Computational Physics*. Wiley: New York, 1972.
23. Martyna, G. J., *et al.*, Explicit Reversible Integrators for Extended Systems Dynamics. *Mol. Phys.* **1996**, *87* (5), 1117.
24. Swope, W. C., *et al.*, A Computer Simulation Method for the Calculation of Equilibrium Constants for the Formation of Physical Clusters of Molecules: Application to Small Water Clusters. *J. Chem. Phys.* **1982**, *76* (1), 637.
25. Verlet, L., Computer "Experiments" On Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. *Phys. Rev.* **1967**, *159* (98), 98.

26. Verlet, L., *Phys. Rev.* **1967**, 165 (201), 201.
27. Hoover, W. G., Canonical Dynamics: Equilibrium Phase-Space Distributions. *Phys. Rev. A* **1985**, 31 (3), 1695.
28. Nose, S., A Molecular Dynamics Method for Simulations in the Canonical Ensemble. *Mol. Phys.* **1984**, 52 (2), 255.
29. Berardi, R., *et al.*, A Generalized Gay-Berne Intermolecular Potential for Biaxial Particles. *Chem. Phys. Lett.* **1995**, 236 (4-5), 462.
30. Berardi, R., *et al.*, A Gay-Berne Potential for Dissimilar Biaxial Particles. *Chem. Phys. Lett.* **1998**, 297 (1-2), 8.
31. Berne, B. J.; Pechukas, P., Gaussian Model Potentials for Molecular Interactions. *J. Chem. Phys.* **1972**, 56, 4213.
32. Liwo, A., *et al.*, A United-Residue Force Field for Off-Lattice Protein-Structure Simulations .1. Functional Forms and Parameters of Long-Range Side-Chain Interaction Potentials from Protein Crystal Data. *J. Comput. Chem.* **1997**, 18 (7), 849.
33. Allen, M. P.; Germano, G., Expressions for Forces and Torques in Molecular Dynamics Simulations Using Rigid Bodies. *Mol. Phys.* **2006**, 104 (20-21), 3225.

CHAPTER 7

ANALYSIS OF THE GAY-BERNE INTERACTION POTENTIAL AND STRATEGY FOR PARAMETERIZATION

The Gay-Berne potential^{1, 2} is a multivariate potential described by 4 degrees of freedom (or independent variables) and 4 (or 6) parameters. Figure 3.4 the variables are defined the benzene dimer and the parameters are concisely described in Table 7.1.

Table 7.1: The parameters and their explanations.^{1, 2}

ϵ_{fac}	Ratio of well depth of edge-to-edge to sandwich configurations
$\sigma_{\perp}$	Size of semi-major (or long) axis of the ellipsoid
$\sigma_{ }$	Size of semi-minor (or short) axis of the ellipsoid
ϵ	Interaction strength or well depth
μ	First exponent of the interaction strength
ν	Second exponent of the interaction strength

The aim of this section is to provide a full description of the Gay-Berne potential and to describe the method used to parameterize this potential from free energy hyper-volumes. The four reaction coordinates defining the free energy hyper-volumes (in Chapters 4 and 5) were chosen, by no coincidence, to have the same description for the 4 degrees of freedom as the Gay-Berne potential. A method for reducing the dimensionality of the hyper-volumes, from 4 *degrees of freedom* to 3-(*volumes*) to 2-(*surfaces*) *degrees of freedom* is introduced.^{3, 4} This was done to allow for the interpretation and visualization of the 5D data, which would otherwise be remarkably difficult to analyze. The bulk of this chapter proceeds in a step-wise fashion by first describing the Gay-Berne in complete detail, and showing the affect of changing each of

the parameters on the Gay-Berne interaction potential surfaces. The optimization tool employed, the Genetic Algorithm (GA), is described in brief detail. The GA was chosen because of its ability to sample a large range of possible parameters with a low probability of getting stuck in local minima. The GA algorithm is also inherently parallel. Finally, the free energy hyper-volumes obtained from the *FEARCF*^{3, 5} and *WHAM*⁶ will then be used to parameterize the Gay-Berne potential.

7.1 The Gay-Berne Potential: The Full Story

7.1.1 Gay-Berne potential

In Chapter 6 a more brief discussion of the Gay-Berne potential (and derivation of its forces) in the context of its degrees of freedom and parameters was discussed. Here, a more in depth look at the potential is given. The Gay-Berne potential is essentially an anisotropic form of the 12-6 Lennard-Jones potential; in other words it not only has distance dependence but includes angular dependence.^{1, 2} The Gay-Berne potential has the following form:

$$U_{GB} = 4\varepsilon_{ij}(\theta_1, \theta_2, \phi) \left[x_{ij}^{12} - x_{ij}^6 \right], \quad 7.1$$

where

$$\varepsilon_{ij} = \varepsilon_{ij}^0 \left[\varepsilon_{ij}^1(\phi) \right]^v \left[\varepsilon_{ij}^2(\theta_1, \theta_2, \phi) \right]^\mu, \quad 7.2$$

the anisotropic contribution to the interaction strength and

$$x_{ij} = \frac{\sigma_{ij}^0}{r_{ij} - \sigma_{ij}(\theta_1, \theta_2, \phi) + \sigma_{ij}^0}, \quad 7.3$$

the anisotropic contribution to the range. The angular variables (θ_1 , θ_2 and ϕ) can also be written in the following form:

$$\theta_1 = \hat{\mathbf{r}} \cdot \hat{\mathbf{u}}_i, \quad 7.4.1$$

$$\theta_2 = \hat{\mathbf{r}} \cdot \hat{\mathbf{u}}_j. \quad 7.4.2$$

$$\phi = \hat{\mathbf{u}}_i \cdot \hat{\mathbf{u}}_j, \quad 7.4.3$$

It is quite clear to see that both the interaction strength and range are based on the molecular shape and orientation.

The usual combining rules applied for Lennard-Jones potential do not apply for the Gay-Berne potential, due to its orientational dependence. The manner in which we obtain the inter-particle interaction strength (ϵ_{ij}) and range, or closest approach (σ_{ij}), are detailed below.^{1, 2}

To make the Gay-Berne easily legible a few terms, based on the shape and orientation of the particles, have to be defined. The first few factors relate the ratio of the semi-major (or long/length, $\sigma_{\perp}$) and semi-minor (or short/breadth, $\sigma_{\parallel}$) axis, of the ellipsoid.

$$\chi = \left[\frac{(l_i^2 - b_i^2)(l_j^2 - b_j^2)}{(l_j^2 + b_i^2)(l_i^2 + b_j^2)} \right]^{1/2}, \quad 7.5.1$$

$$\alpha^2 = \left[\frac{(l_i^2 - b_i^2)(l_j^2 - b_j^2)}{(l_j^2 - b_j^2)(l_i^2 + b_j^2)} \right]^{1/2} \quad 7.5.2$$

where l and b describe the length and breadth of the ellipsoids, respectively (Figure 7.2).

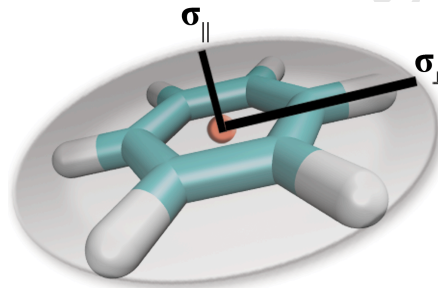


Figure 7.2: The description of an ellipsoid used in the Gay-Berne potential using a benzene molecule as an example. The semi-major axis ($\sigma_{\perp}$ or l) that is perpendicular to the molecular vector and the semi-minor axis ($\sigma_{\parallel}$ or d) that is parallel to the molecular vector.

To avoid imaginary numbers the terms are combined in the following way²

$$\chi\alpha^2 = \frac{l_i^2 - b_i^2}{l_i^2 + d_j^2}, \quad 7.6.1$$

$$\chi\alpha^{-2} = \frac{l_j^2 - b_j^2}{l_j^2 + d_i^2}, \quad 7.6.2$$

$$\chi^2 = \frac{(l_i^2 - b_i^2)(l_j^2 - b_j^2)}{(l_j^2 + b_i^2)(l_i^2 + b_j^2)}. \quad 7.6.3$$

The next two terms are the anisotropic factors that affect the interaction strength and are based on which particle configuration is preferred, which is controlled by ϵ_{fac} . The

ε_{fac} parameter is the ratio of the interaction strength for edge-to-edge (ε_{--}) to sandwich ($\varepsilon_{//}$) configurations, Equation 7.7.1.

$$\varepsilon_{fac} = \frac{\varepsilon_{--}}{\varepsilon_{//}} \quad 7.7.1$$

$$\chi' = \frac{1 - (\varepsilon_{fac})^{1/\mu}}{1 + (\varepsilon_{fac})^{1/\mu}} \quad 7.7.2$$

$$\alpha' = \left[1 + (\varepsilon_{fac})^{1/\mu} \right]^{-1}. \quad 7.7.3$$

Along with orientation between the Gay-Berne particles, these factors give the interaction strength (ε_{ij}) its anisotropic form. Note that if $l = d$, then the Gay-Berne potential simply reduces to the Lennard-Jones intermolecular potential.^{1,2}

Finally, two more terms are required to ensure that reading of the Gay-Berne potential is simple. The first, is the anisotropic factor affecting the interaction strength (ε_{ij})

$$H' = \frac{\chi' \alpha'^2 (u_i \cdot r_{ij}) + \chi' \alpha'^2 (u_j \cdot r_{ij}) - 2\chi'^2 (u_i \cdot r_{ij})(u_j \cdot r_{ij})(u_i \cdot u_j)}{1 - \chi'^2 (u_i \cdot u_j)} \quad 7.8$$

and, the second, is the anisotropic factor affecting the range (σ_{ij})

$$H = \frac{\chi \alpha^2 (u_i \cdot r_{ij}) + \chi \alpha^2 (u_j \cdot r_{ij}) - 2\chi^2 (u_i \cdot r_{ij})(u_j \cdot r_{ij})(u_i \cdot u_j)}{1 - \chi^2 (u_i \cdot u_j)}, \quad 7.9$$

where both factors are based on the relative particles (ellipsoid) orientation. Now that we have introduced all the necessary terms, a much better description of the Gay-Berne potential can be provided.

The most favourable interaction energy, ε_{ij}^0 , is found when the particle vectors are orthogonal. This can be obtained from either quantum mechanical or 1D potential of mean force calculations. The anisotropic contribution to the interaction strength (ε_{ij}^1), which is a function of the dihedral angle only,^{1,2} is given by

$$\varepsilon_{ij}^1(\phi) = \left[1 - \chi^2 (u_i \cdot u_j)^2 \right]^{1/2}. \quad 7.10$$

In addition, the contribution from all three of the angular degrees of freedom (θ_1 , θ_2 and ϕ) is given by the ε_{ij}^2 term

$$\varepsilon_{ij}^2(\theta_1, \theta_2, \phi) = 1 - H', \quad 7.11$$

The closet approach distance, σ_{ij}^0 , along the semi-major (or long) axis of the two interacting ellipsoids is obtained by using the geometric combining rule:

$$\sigma_{ij}^0 = \sqrt{\left[(\sigma_{\perp,i})^2 + (\sigma_{\perp,j})^2\right]}. \quad 7.12$$

Finally, the anisotropic range parameter (σ_{ij}), which is also dependent on all the angular degrees of freedom, is defined as follows:

$$\sigma_{ij}(\theta_1, \theta_2, \phi) = \sigma_{ij}^0 [1 - H]^{-1/2}, \quad 7.13$$

Combining Equations 7.1 to 7.13 results in all the details necessary for the full definition of the Gay-Berne potential, which is an anisotropically shifted form of the Lennard-Jones potential.^{1, 2}

7.1.2 Visualizing the Gay-Berne datasets

The Gay-Berne is a multi-dimensional, orientationally dependent form of the Lennard-Jones soft sphere potential. Viewing this potential by including the 4 independent variables, or degrees of freedom, ($r, \theta_1, \theta_2, \phi$) and the dependent variable (U_{GB}) is impractical and cumbersome to interpret the data. It is for this reason that we reduce the dimensionality of the problem, by averaging over the independent variables to form either a 3- or 2-variable subset of the original 4-variable ($U(r, \theta_1, \theta_2, \phi)$) dataset.^{3, 4} The 3-variable subsets are denoted by $U^{3,1}(r, \theta_1, \theta_2; \phi_i)$ and $U^{3,2}(r, \theta_{1or2}, \phi; \theta_{2or1})$:

$$U^{3,1}(r, \theta_1, \theta_2; \phi_i); U^{3,2}(r, \theta_{1or2}, \phi; \theta_{2or1}) \subseteq U(r, \theta_1, \theta_2, \phi). \quad 7.14$$

For similar particles it should be noted that θ_1 and θ_2 are interchangeable, and the two reduced forms on the left of Equation 7.14 can sufficiently describe the original 4-variable superset ($U(r, \theta_1, \theta_2, \phi)$).⁴ It also follows that the 2-variable form is a subset of the 3-variable subset, and, by corollary, a subset of the original data. The 2-variable subsets are denoted by $U^{2,1}(r, \theta_1; \theta_2, \phi); U^{2,2}; U^{2,3}$ and $U^{2,4}$:

$$U^{2,1}(r, \theta_1; \theta_2, \phi); U^{2,2}; U^{2,3}; U^{2,4} \subseteq U^{3,1}(r, \theta_1, \theta_2; \phi). \quad 7.15$$

The reduction in dimensionality can be done in several ways but the method employed to average over the degrees of freedom we wish to view along (much like viewing down the z-axis of a crystal where a 3D view is reduced to 2D view) is by using Boltzmann averaging^{3, 4}. The Boltzmann averaging is performed in the following manner:

$$U^{3,1}(\xi_1, \xi_2, \xi_3) = -k_B T \ln \left[\sum_{\xi_4} P^{3,1}(\xi_1, \xi_2, \xi_3, \xi_4) \right] \quad (3D \text{ volume}) \quad 7.16.1$$

$$U^{2,1}(\xi_1, \xi_2) = -k_B T \ln \left[\sum_{\xi_3, \xi_4} P^{2,1}(\xi_1, \xi_2, \xi_3, \xi_4) \right] \quad (2D \text{ contour}), \quad 7.16.2$$

where the probability, P , is given by the usual Boltzmann distribution

$$P^{3,1}(\xi_1, \xi_2, \xi_3, \xi_4)_{\xi_4} = \frac{\exp(-\beta U_{\xi_4})}{\sum \exp(-\beta U_{\xi_1, \xi_2, \xi_3, \xi_4})} \quad (3D \text{ volume}) \quad 7.17.1$$

$$P^{2,1}(\xi_1, \xi_2, \xi_3, \xi_4)_{\xi_3, \xi_4} = \frac{\exp(-\beta U_{\xi_3, \xi_4})}{\sum \exp(-\beta U_{\xi_1, \xi_2, \xi_3, \xi_4})} \quad (2D \text{ contour}). \quad 7.17.2$$

Using these reductions, it is possible to view the characteristics of the Gay-Berne potential and conceive its topology as it relates to select configurations. The plots produced from the *3-variable* subsets are 4 dimensional (i.e., volume plots), while the *2-variable* subsets are 3 dimensional (i.e., contour plots). A comparison between the Lennard-Jones (Figure 7.3) and the Gay-Berne potentials, presented as *3D* (volume) and *2D* (contour) plots (Figure 7.4), are used to demonstrate the anisotropic nature of the latter interaction potential energy function.

This comparison exposes the superiority of the Gay-Berne potential is for describing molecular anisotropic interactions. It should be noted that the parameters used in Figure 7.4 are fictitious and serve only to demonstrate the anisotropic nature of the Gay-Berne potential. A value for ϵ_{fac} representing an edge-to-edge configuration is used.² Variations in the character of the interaction potential energy hyper-volume, while varying each of the parameters, will be discussed in the next section.

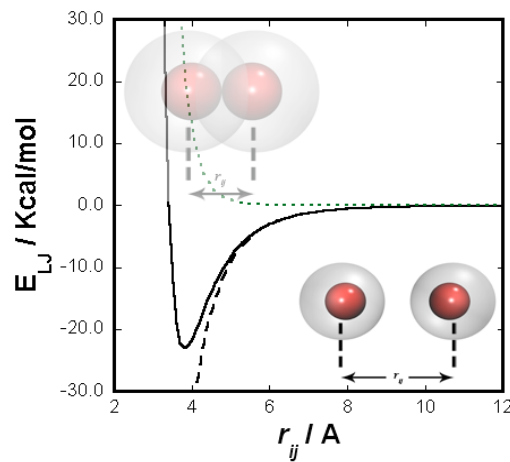


Figure 7.3: The Lennard-Jones potential with a $\sigma = 3.4 \text{ \AA}$ and $\epsilon = 4 \text{ kcal/mol}$, obtained from an isotropic Gay-Berne potential, i.e. where $l = d$.

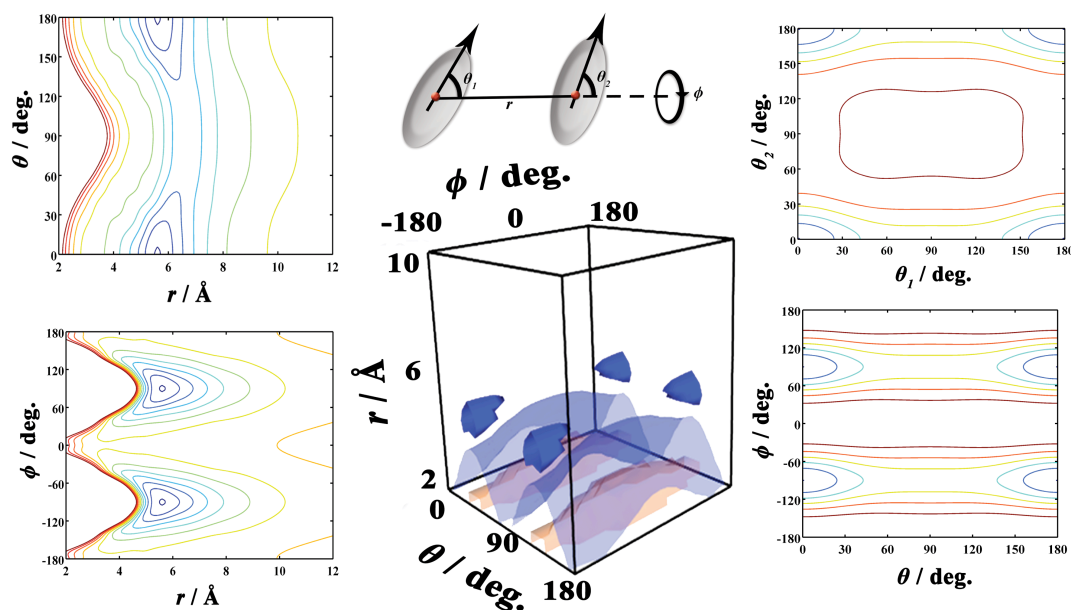


Figure 7.4: The Gay-Berne potential with fictitious parameters: $\sigma_{\perp} = 3.4 \text{ \AA}$, $\sigma_{\parallel} = 2.4 \text{ \AA}$, $\epsilon_{fac} = 5$, $\epsilon = 1.0 \text{ kcal/mol}$, $\mu = 2$ and $\nu = 1$. The variable definition (top middle), 3D volume (middle; volumes at 0.6 kcal/mol (dark blue), 1.8 kcal/mol (opaque blue) and 5.0 kcal/mol (red)) and 2D contour plots (sides, contoured at 0.5 kcal/mol from 0.0 to 5.0 kcal/mol).

7.1.3 Parametric analysis

The affects of varying the parameters (listed in Table 7.1) on the shape and topology of the Gay-Berne potential will be discussed. To illustrate the effects of changing the values for the parameters the use of the reduced 2D plots are used, showing the distance and one orientation variable ($\theta = \theta_1 = \theta_2$ for similar molecules), for the sake of brevity. The affect will be shown for all 6 Gay-Berne parameters (Table 7.1). Values of θ that are 0 or 180° correspond to a stacked arrangement, while a θ value of 90° corresponds to a side-by-side configuration, Figure 7.5.

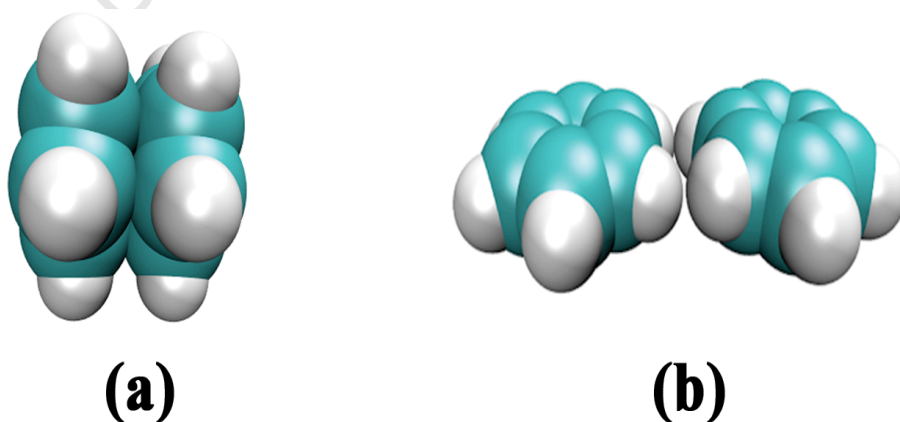


Figure 7.5: (a) Sandwich/stacked arrangement, resulting in a short inter-particle distance, and (b) edge-to-edge configuration, that has a longer inter-particle distance.

ϵ_{fac} : Controls the arrangement of the ellipsoidal particles as either sandwich or edge-to-edge arrangement, Figure 7.5. It is essentially a ratio of the interaction strength between the edge-to-edge to sandwich configurations. For an $\epsilon_{fac} > 1.0$ will favour edge-to-edge arrangements (Figure 7.6b) and a value of < 1.0 will favour sandwich arrangements (Figure 7.6a). $\epsilon_{fac} = 1.0$ induces no particular preference for either arrangement (Figure 7.6c), giving rise to an ensemble of possible configurations.

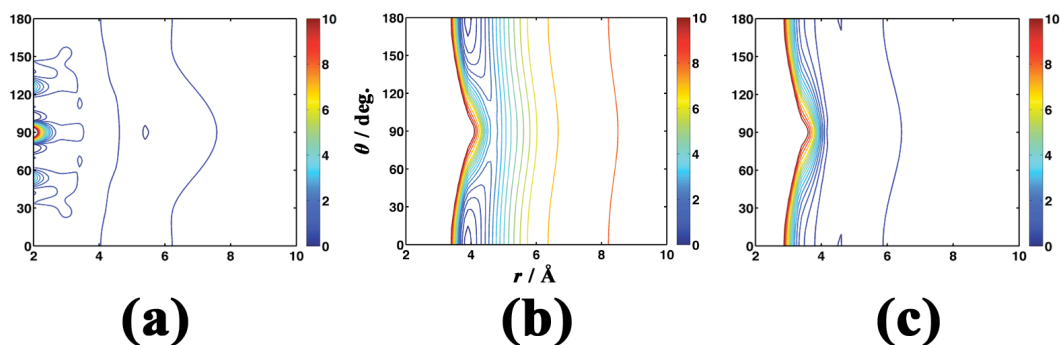


Figure 7.6: The affect of ϵ_{fac} parameter on the Gay-Berne potential energy function, shown for the 2D surface r - θ only. The ϵ_{fac} values are (a) 0.1, (b) 5.0 (as in Figure 7.3) and (c) 1.0.

Length($l, \sigma_{\perp}$):breadth($d, \sigma_{\parallel}$): When we have the length longer than the breadth by a factor of 2, i.e. a ratio of 2:1, then the surface protrudes with increasing distance, reaching maximum of the curvature occurs at 90° (Figure 7.7a and b). The opposite is observed when the breadth is longer than the length, i.e. a ratio of 0.75:1 – the surface recedes with a decrease in the distance (Figure 7.7c). When the length and breadth are equal or tends toward 1, the surface approaches Lennard-Jones type behaviour.

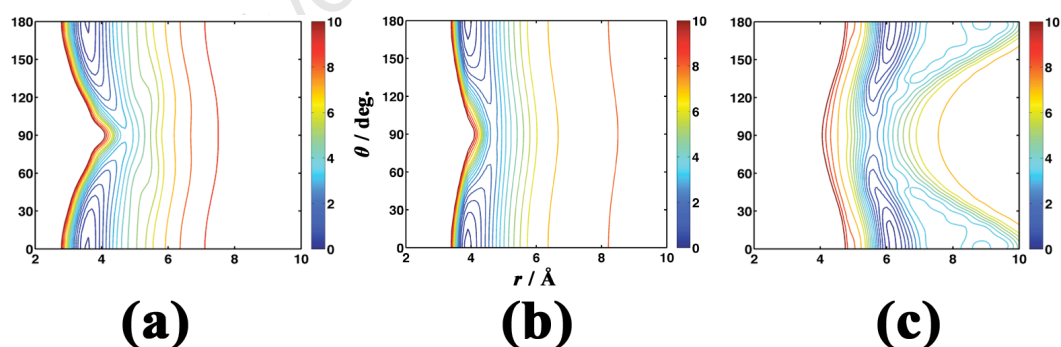


Figure 7.7: The affect of changing the $l:b$ ($\sigma_{\perp}:\sigma_{\parallel}$) ratio. The ratios are (a) 2:1, (b) 1.42:1 and (c) 0.75:1.

ϵ^o : This parameter controls the interaction strength, or well depth, affecting the overall interaction energy. An illustration of for ϵ values of 0.1, 1.0 and 10.0 are shown in Figure 7.8.

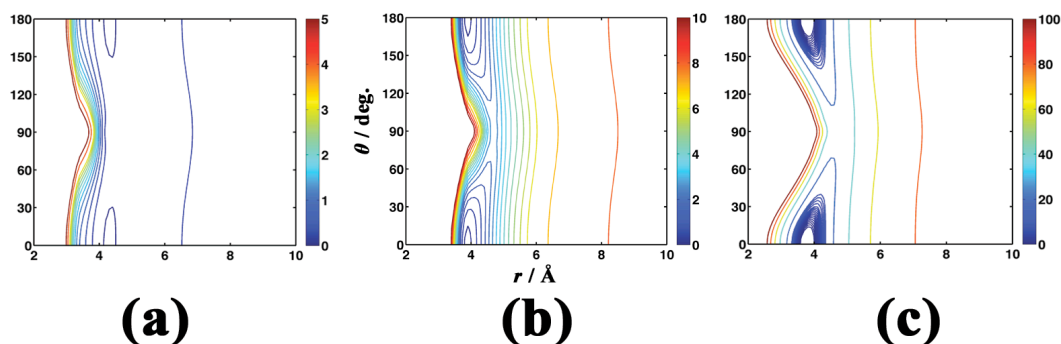


Figure 7.8: The various values of the ε parameter (a) 0.1, (b) 1.0 and (c) 10.0 kcal/mol.

μ and ν : These two parameters are the exponents for ε^1 and ε^2 interaction strength factors, respectively. They work in conjunction ε to control the overall interaction strength between the particles (Figure 7.9).

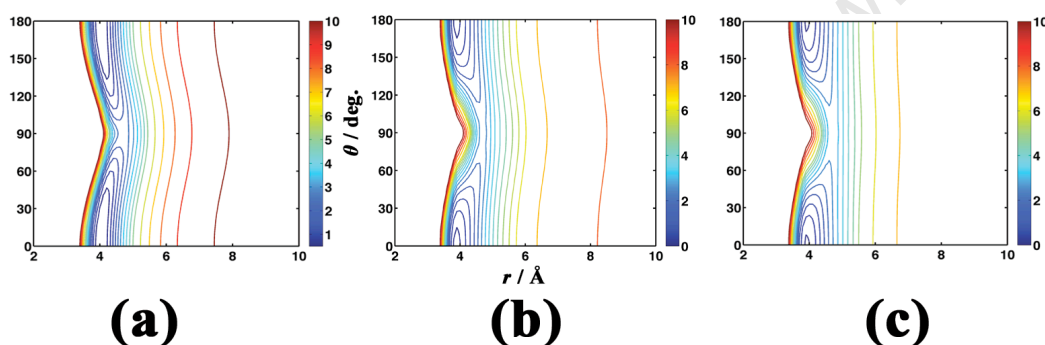


Figure 7.9: The affect of changing the exponents' μ and ν on the Gay-Berne potential energy surface. The values for μ and ν are (a) 1.0 and 2.0, (b) 2.0 and 1.0 and (c) 6.0 and 2.0 respectively.

By swapping the values of μ and ν of the original data (Figure 7.9b), results in the minima becoming more distorted and less well defined, Figure 7.9a. A larger ratio of $\mu:\nu$, i.e. 6:2, results decrease in the anisotropy decrease with increasing distance, Figure 7.9c.

7.2 Parameterization Using a Genetic Algorithm (GA)

7.2.1 Fitness function

The fitness function, used in a Genetic Algorithm (GA), can be as simple as an error function or as complex as a heuristic neural network system. In this thesis, we make use of the least squares function as the fitness function to determine the optimal parameters for the Gay-Berne interaction potential using free energy hyper-volumes. The least squares function is defined in the following way

$$S(x; p) = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i(x) - f_i(x; p))^2}. \quad 7.18$$

where $S(x; p)$ is the error function, $y_i(x)$ the i -th observable and $f_i(x; p)$ the i -th function value for the point x and the parameter (p) . This can be extended for the fitting of the Gay-Berne potential to the free energy hyper-volumes in the following manner

$$S(r, \theta_1, \theta_2, \phi; \varepsilon_{fac}, \varepsilon, \sigma_{||}, \sigma_{\perp}) = \sqrt{\sum_i \left[y_i(r, \theta_1, \theta_2, \phi) - f_i(r, \theta_1, \theta_2, \phi; \varepsilon_{fac}, \varepsilon, \sigma_{||}, \sigma_{\perp}) \right]^2}, \quad 7.19$$

where $S(r, \theta_1, \theta_2, \phi; \varepsilon_{fac}, \varepsilon, \sigma_{||}, \sigma_{\perp})$ is the error function, $y_i(r, \theta_1, \theta_2, \phi)$ the i -th free energy observable and $f_i(r, \theta_1, \theta_2, \phi; \varepsilon_{fac}, \varepsilon, \sigma_{||}, \sigma_{\perp})$ the i -th Gay-Berne interaction potential energy value at the point $(r, \theta_1, \theta_2, \phi)$ and the parameters $(\varepsilon_{fac}, \varepsilon, \sigma_{||}, \sigma_{\perp})$. It should be noted that no reliable global offset methods are available at present, and this due to the high-dimensionality of the free energy hyper-volume. Instead, the global minimum of the Gay-Berne potential, $\min[U_{GB}(r, \theta_1, \theta_2, \phi)]$, was used to offset the Gay-Berne potential energy value, $f_i(r, \theta_1, \theta_2, \phi; \varepsilon_{fac}, \varepsilon, \sigma_{||}, \sigma_{\perp})$, to match the free energy, such that the value is actual given by

$$f_i(r, \theta_1, \theta_2, \phi; \varepsilon_{fac}, \varepsilon, \sigma_{||}, \sigma_{\perp}) = f_i(r, \theta_1, \theta_2, \phi; \varepsilon_{fac}, \varepsilon, \sigma_{||}, \sigma_{\perp}) - \min[U_{GB}(r, \theta_1, \theta_2, \phi)].$$

7.2.2 Crossing boundaries

In describing the algorithm used in this thesis, we have to first realize that it is a method drawn from the principles of evolutionary biology, and more specifically from Darwin's theory of evolution, to a new realm in all forms of science. This is analogous to the evolution from simple unicellular bacteria to multicellular complex organisms. Darwin's theory is based on the process of natural selection. It states that the organism with a set of genes, which allows it to better survive in its environment, is allowed to reproduce. This ensures its genes survival in the population pool. In other words organisms that are less adapted, or have less fit genes, will be eliminated and their genes removed from the population pool. By this, and not strictly speaking, nature is always trying to produce the perfect organisms to survive in a dynamically changing environment.

Concisely stated, the GA is an optimization tool that allows the solutions to a problem to be analyzed using a metric called a fitness function, and each solution (or candidate) is qualitatively evaluated. The poor solutions are removed from the pool of solutions, while the better solutions are allowed to produce offspring (or reproduce). The

offspring of the remaining solutions can be mutated, to allow for a better solution, and go on to be evaluated by the fitness function in the next generation. Once again, poor solutions are removed and better ones are retained. This cycle goes on until optimal solutions to the problem are obtained.⁷ Further change in the individuals can be introduced by the process of mutation (exploring the genetic diversity, or solutions, by introducing a random change in the individual) and/or crossover (exploiting the genes, or solutions, of the fitter individuals).⁷

7.2.3 *Methods of representation*

As for any search and learning method, the way in which the candidate solutions are encoded is a central factor in the success of a GA. Most GA applications use fixed-length, fixed-order bit strings to encode the candidate solutions.⁸ There are three general methods for encoding the candidate solutions such as (1) binary strings, (2) many-character and real-values and (3) tree encodings.

The binary string method is the most commonly used method for many reasons, but mostly for historical reasons. Holland and others have concentrated mostly on this type of encoding.⁹ Much of the GA is based on the assumption of fixed-length, fixed-order binary encodings where much of that theory can be extended to non-binary encodings. However, non-binary encodings are not as well developed as the original binary encoding.⁸ In addition, heuristics about appropriate parameter settings (e.g. mutation and crossover rates) have been developed in the context of binary encodings.⁸ There has also been other extensions to binary encoding over the years such as gray coding^{10, 11} and Hillis's diploid binary encoding scheme.^{9, 12, 13}

The many-character and real-values encoding is a more natural-feeling choice for encoding entities such as chromosomes. It uses characters or real numbers both of which are more abundant than 0s and 1s.⁸ Examples include Kitano's many-character representation for graph-generation grammars,¹⁴ Meyer and Packard's real-valued representation for condition sets,¹⁵ Montana and Davis's real-valued representation for neural-network weights¹⁶ and Schulze-Kremer's real-valued representation for torsional angles in proteins.¹⁷

Tree encoding schemes, such as John Koza's scheme for representing computer programs^{18, 19} have several advantages including the fact that they allow the search space to be open-ended (in principle, any size tree could be formed by mutation and crossover).⁸ This open-endedness also leads to some disadvantages; trees can grow very

large, in an uncontrolled manner, preventing the formation of more structured, hierarchical candidate solutions. The resulting trees can be very difficult to understand and simplify by virtue of their size.⁸

7.2.4 *Methods of selection*

There are many methods for selecting which solutions will be retained in the next generation. In brief, a selection method is required to choose which individual (or solution) will create offspring for the next generation and how many offspring it will produce. The purpose of selection is to emphasize the fitter individuals in the population in the hope that their offspring will bring about better fitness.⁸ Selection has to be balanced with the variation introduced by mutation and crossover (exploration versus exploitation). Too strong of a selection will result in suboptimal highly fit individuals that will take over the population. This will result in a decrease in the diversity and hamper the need for further change and progress. However, too weak of a selection will result in evolution that is too slow.⁸ A brief explanation of the common selection methods are provided below. More in depth discussions can be found in the following references [8, 20-23].^{8, 20-23}

*Roulette Wheel:*⁷⁻⁹ Conceptualized as a roulette wheel, where each individual is given a slice of the wheel. The size of the slice is determined by the fitness of the individual. The wheel is spun and the slice on which it lands is chosen to spawn the next generation.

Sigma Scaling:^{7, 8, 12, 24} Keeps the selection pressure relatively constant throughout the run rather than depending on the fitness variances in the population. An individual's fitness expected value is a function of its fitness, the population mean and standard deviation. At the beginning of the run the selection pressure does not allow only the fittest individuals to spawn but less fit individuals are also allowed to produce offspring, while near the end of the run the selection becomes more discriminating.

Elitism:^{7, 8, 25} The fittest members of each generation are guaranteed to be selected. Such individuals can be lost if not selected to reproduce or destroyed by mutation or crossover.

Boltzmann Selection:^{7, 8, 26} An approach similar to simulated annealing. A continuously varying "temperature" controls the rate of selection according to the preset selection scheme. The temperature starts out high, with a low selection pressure, and is slowly cooled resulting in an increase in the selection pressure.

Rank Selection:^{7, 8, 27} Each individual in the population is assigned a numerical rank based on fitness. The selection is based on this ranking and not on absolute differences in fitness. It is an alternative method to prevent convergence occurring too quickly.

Tournament Selection:^{7, 8, 28} Two individuals are selected at random from the population and a random number (*rand*) between 0 and 1 is chosen. If the $rand < k$ (where *k* is a parameter) the fitter of the two individual's is chosen to be the parent, else the less fit individual is chosen. Both individuals are retained in the population and can be selected again.

Steady State Selection:^{7, 8, 25} Only a few individuals are replaced in each generation. Usually less fit individuals are replaced by the offspring of the fittest individuals, which have been subjected to mutation and crossover.

7.2.5 Methods of change

Once the selection scheme has selected the fit individuals, they must be randomly altered in the hope that the alteration will improve the fitness in the next generation. There are basically two types of strategies to accomplish this alteration.^{7, 8} The first is conceptually the simplest and is called mutation – just as mutation occurs in living organisms – it changes one gene to another. Mutation in a GA causes small alterations at single points in an individual's code, thus mutating the individual into something new.^{7, 8}

The second method is called crossover and entails choosing two individuals to swap parts of their genes. This swapping produces artificial offspring that are combinations of the parents – much like sexual reproduction does in nature. Common forms of crossover include single-point crossover, in which a point of exchange is set at a random location in the genetic code of the two individuals. All the code before the set “point” in the one individual goes to the offspring, while the second individual contributes code after the set “point”. This is called uniform crossover, as the offspring has a 50/50 split of the two parents.^{7, 8} To reduce positional bias, which results in an “endpoint” anomaly, another crossover method has been developed that uses two points, called the two-point crossover.^{7, 8} Here two positions are chosen at random and the segments between them are exchanged. The segments that are exchanged do not necessarily contain the endpoints of the strings from the two parents. A number of crossover points have been experimented with for each parent. In one method the crossover points for each of the parents is chosen from a Poisson distribution, whose mean is a function of the length of the chromosome.^{7, 8}

It is important to note that it is not the choice between mutation or crossover, but rather the balance between mutation, crossover and selection that is important. The correct balance also depends on the details of the fitness function and type of encoding used.^{7,8}

7.2.6 Strengths

The first and most important note about the GA is that it is inherently parallel.²³ Most algorithms are serial and can only explore solution space in one direction at a time. If the solution is found to be suboptimal, which is highly dependent on the initial guess, all the work to that point has to be discarded and the algorithm has to started over. Since the GA spawns many offspring, it can explore many different solutions at once. If one path turns out to be a poor solution, it can be removed and continue along a more promising route. This gives a greater chance of finding the optimal solution.⁷

In the beginning, GAs generates a vast population of possible solutions thereby casting a “net” over the fitness landscape.⁷ The strength of this is that it performs well when fitting problems with rugged and complex landscapes. In other words landscapes where the fitness function is discontinuous, noisy, changes over time and has many local minima. It is here that GAs have found their niche and proven to be effective. GAs have the powerful ability to escape local minima and discover the global minimum in a very rugged and complex fitness function landscape. The major aspects of the GA – parallelism, selection, mutation and crossover – work together for finding the optimal solution.

Another area where the GA supersedes other optimization algorithms is that it excels in the ability to manipulate many parameters simultaneously.²⁴ Many real-world problems cannot be expressed as a single value problem that requires minimization or maximization, but rather depend on multiple variables and parameters. This is especially true in the case of this thesis in that the Gay-Berne interaction potential has to be fitted to a free energy hyper-volume.

Finally evolutionary algorithms, like the GA, are neither aware of nor concerned with whether a solution runs counter some established beliefs. It is only concerned with solutions that work.⁷

7.2.7 Limitations

Although the GA has proven to be an efficient and powerful method for finding the optimal solution to a problem, it is not the be all and end all of optimization schema. The

first and most important consideration for the GA is defining a representation of the problem. The language used to stipulate the candidate solutions must be robust and not prone to fatal errors when random changes are introduced.⁷

The problem of defining the fitness function must be performed with careful consideration and has to be appropriate for the problem at hand. In other words an appropriate fitness function has to be defined such that a higher fitness actually equates to a better solution, rather than not being able to find a solution or, even worse, solving the wrong problem.⁷

The rate of mutation and crossover, size of the population and the strength of the selection have to be considered carefully as well. If the rate of genetic change is too high or the selection scheme is chosen poorly beneficial schema may be disrupted and the system will enter an error catastrophe, resulting in change that occurs too fast to bring about convergence. Population sizes that are too small will result in the GA failing to explore enough of solution space to consistently yield good solutions. The most common method to deal with these problems involve controlling the strength of selection, so as not to give excessively fit individuals too great of an advantage.⁷

In spite of its shortcomings, as well as how counterintuitive and astonishing as it may seem, GAs have proven to be a very efficient and powerful tool for problem solving. This demonstrates the power of evolutionary principles,⁷ as proposed by Charles Darwin in the *Origin of Species* published in 1859. GAs have found their way into a vast array of scientific disciplines from physics²² to chemistry,⁷ and it often comes up with solutions efficiently, elegantly or more complex than any human engineer can produce.⁷

7.3 References

1. Gay, J. G.; Berne, B. J., Modification of the Overlap Potential to Mimic a Linear Site-Site Potential. *J. Chem. Phys.* **1981**, 74 (6), 3316.
2. Golubkov, P. A.; Ren, P. Y., Generalized Coarse-Grained Model Based on Point Multipole and Gay-Berne Potentials. *J. Chem. Phys.* **2006**, 125 (6), 064103.
3. Strumpfer, J.; Naidoo, K. J., Computing Free Energy Hypersurfaces for Anisotropic Intermolecular Associations. *J. Comput. Chem.* **2010**, 31 (2), 308.
4. Gamielien, M. R., *et al.*, Hydration Determined Orientational Preferences in Aromatic Association from Benzene Dimer Free Energy Volumes. *J. Phys. Chem. B* **2011**.
5. Barnett, C. B.; Naidoo, K. J., Free Energies from Adaptive Reaction Coordinate Forces (Fearcf): An Application to Ring Puckering. *Mol. Phys.* **2009**, 107 (8-12), 1243.
6. Kumar, S., *et al.*, The Weighted Histogram Analysis Method for Free-Energy Calculations on Biomolecules. I. The Method. *J. Comput. Chem.* **1992**, 13 (8), 1011.
7. Marczyk, A. Genetic Algorithms and Evolutionary Computation. <http://www.talkorigins.org/faqs/genalg/genalg.html> (accessed 21 November 2011).
8. Mitchell, M., *An Introduction to Genetic Algorithms*. The MIT Press: Cambridge, Massachusetts, 1998.
9. Holland, J. H., *Adaptation in Natural and Artificial Systems*. University of Michigan Press: 1975.
10. Bethke, A. D. Genetic Algorithms as Function Optimizers. University of Michigan, Ann Arbor, 1980.
11. Caruana, R. A.; Schaffer, J. D. In *Representation and Hidden Bias: Gray Vs. Binary Encoding for Genetic Algorithms*, Proceedings of the Fifth International Conference on Machine Learning, Morgan Kaufmann: 1988.
12. Goldberg, D. E., *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley: 1989.
13. Hillis, W. D., Co-Evolving Parasites Improve Simulated Evolution as an Optimization Procedure. *Physica D* **1990**, 42, 228.
14. Kitano, H., Designing Neural Networks Using Genetic Algorithms with Graph Generation System. *Complex Systems* **1990**, 4, 461.
15. Meyer, T. P.; Packard, N. H., *Local Forecasting of High-Dimensional Chaotic Dynamics*. Addison-Wesley: 1992.
16. Montana, D. J.; Davis, L. D. In *Training Feedforward Networks Using Genetic Algorithms*, Proceedings of the International Joint Conference on Artificial Intelligence, Morgan Kaufmann: 1989.
17. Schulze-Kremer, S., *Genetic Algorithms for Protein Tertiary Structure Prediction*. North-Holland: 1992.
18. Koza, J. R., *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. MIT Press: 1992.
19. Koza, J. R., *Genetic Programming II: Automatic Discovery of Reusable Programs*. MIT Press: 1994.
20. Rudas, I. J., *et al.*, *Computational Intelligence in Engineering*. Springer-Verlag: Berlin, Heidelberg, 2010; Vol. 313.
21. Back, T., *Evolutionary Algorithms in Theory and Practice*. Oxford University Press: Oxford, 1996.
22. Haupt, R. L.; Werner, D. H., *Genetic Algorithms in Electromagnetics*. John Wiley & Sons, Inc.: New Jersey, 2007.
23. Sivanandam, S. N.; Deepa, S. N., *Introduction to Genetic Algorithms*. Springer-Verlag: Berlin, Heidelberg, 2008.
24. Forrest, S., Genetic Algorithms: Principles of Natural Selection Applied to Computation. *Science* **1993**, 261, 872.

25. De Jong, K. A. An Analysis of the Behaviour of a Class of Genetic Adaptive Systems. University of Michigan, Ann Arbor, 1975.
26. Goldberg, D. E., A Note of Boltzmann Tournament Selection for Genetic Algorithms and Population Oriented Simulated Annealing. *Complex Systems* **1990**, 4, 445.
27. Baker, J. E. In *Adaptive Selection Methods for Genetic Algorithms*, Proceedings of the First International Conference on Genetic Algorithms and Their Applications, Erlbaum: 1985.
28. Goldberg, D. E.; Deb, K. In *A Comparative Analysis of Selection Schemes Used in Genetic Algorithms*, Morgan Kaufmann, Rawlins, G., Ed. 1991.

University of Cape Town

CHAPTER 8

COARSE-GRAIN DYNAMICS OF LIQUID BENZENE – A PROOF OF CONCEPT

8.1 Introduction

The simulation of the condensed phase and biological systems has resulted in profound new insight into the molecular-scale phenomena that occur in these complex systems.¹ However, many processes that occur in liquids and biological systems occur over spatial (size) and temporal (time) scales that are well beyond the current capabilities of all atom molecular dynamics (MD) simulations.¹ Therefore, there has been a drive to produce new and novel approaches that can access larger size and longer time scales. One such approach is coarse-grain (CG) dynamics methods, which groups atoms together to form so-called CG sites. Coarse-graining is the reduction of the molecular degrees of freedom into lower resolution models, while retaining a connection with the all atom description.^{1, 2} The grouping of atoms, and averaging over other molecular degrees of freedom, results in a decrease in the amount of computations required. These fewer computations paves the way for investigations of liquids and biological systems well into the microsecond (or greater) range. An example of coarse-graining is given for liquid benzene in Figure 8.1.

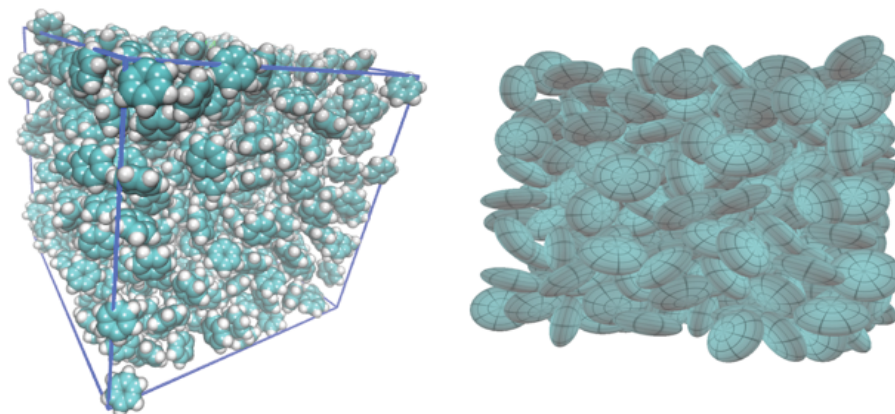


Figure 8.1: The reduction of degrees of freedom for liquid benzene; where the all atom model (left) is converted into a single ellipsoidal site for each benzene molecule (right) for the reduced representation.

The lowering of the resolution of atomic systems into CG sites requires specialized effective interaction potential energy functions. These functions have to be parameterized by being fitted to data obtained from either molecular dynamics (MD), Monte Carlo (MC), quantum mechanics (QM), free energy (FE) simulations or experiments.³

Coarse-graining must be understood, at some level, within the context of statistical mechanics. This remarkable theoretical framework provides the connection between the microscopic world of molecular interactions and the macroscopic world of thermodynamics.¹ In that light, most CG methods are best cast within the context of the following formula for the Helmholtz free energy, A , (derived from the canonical ensemble):

$$\begin{aligned} \exp\left(-\frac{A}{k_B T}\right) &= \text{const.} \int d\mathbf{x} \exp\left[\frac{-V(\mathbf{x})}{k_B T}\right] \\ &\approx \text{const.}' \int d\mathbf{x}_{CG} \exp\left[\frac{-V_{CG}(\mathbf{x}_{CG})}{k_B T}\right]. \end{aligned} \quad 8.1$$

Here $V(\mathbf{x})$ is the system potential energy as a function of the coordinates, $\mathbf{x}$, of all the atoms in the system, T is the temperature, k_B is Boltzmann's constant and "const." the normalization constant (where the prime in the second line represents a different constant).¹ Importantly, in the second line of Equation 8.1, the expression for the free energy is rewritten in terms of the CG variables, $\mathbf{x}_{CG}$, and the effective CG potential, $V_{CG}(\mathbf{x}_{CG})$. The CG variables, by definition, are fewer in number than the atomistic variables such that $N_{xCG} < N_x$.¹

CG simulations have been applied to a multitude of chemical and biological systems² that include the liquid benzene⁴ and water systems,^{5, 6} peptides and proteins,⁷⁻⁹ dendrimers,¹⁰ nucleic acids^{11, 12} and lipids.¹³⁻¹⁶ The development of reduced representations also has a long history of success in the study of condensed matter.² The Gay-Berne (GB) potential was developed to describe the short-range repulsive and attractive interactions (dispersion) of anisotropic molecules.¹⁷ The GB potential has been widely used for liquid crystals,^{18, 19} protein^{7, 8} and nucleic acid^{11, 12} simulations. Although these simulations have been successful, they have often been parameterized from crystal or static quantum mechanical data, which results in a bias for certain non-bonded configurations and also does not include the effects of entropy. The latter

difficulty has been fairly well resolved by fitting the CG simulation to experimentally derived thermodynamic data.³

In the present study, a unique approach to the parameterization of the GB interaction potential is performed using a fully converged benzene dimer anisotropic free energy hyper-volume in vacuum.²⁰ This free energy hyper-volume was found to be sufficient to describe the anisotropic nature of the benzene dimer interaction.²⁰ There have been others who have parameterized the CG interaction potential derived from fitting to *ab initio* simulations.²¹⁻²⁴ However, this work is within the context of statistical mechanics as described above in Equation 8.1. The parameters obtained by parameterization have been used in a CG dynamics code for the simulation of liquid benzene. This CG model has been tested for the parameters derived from our free energy hyper-volume²⁰ and the parameters reported by Golubkov and Ren (from hereto referred to as GR).⁴ Differences arise between our simulations and that of GR, due to GR using a CG model that incorporated the electrostatic interactions as a point multipole expansion.⁴ The structural analysis and diffusion constant reveals the differences between the simulations and parameter sets. We find that the CG simulations using the parameters obtained from the free energy fits the structural and dynamical data from molecular dynamics simulations and experiments quite well.

8.2 Methods

8.2.1 The Gay-Berne potential, forces and torques

The Gay-Berne (GB) interaction potential^{4, 17} has been completely defined in the previous chapter. In this study, the GB has been used to calculate the intermolecular interaction between the CG particles. The range (σ) and strength (ϵ) parameters, for pairwise interactions, are a function of the orientation of the particles. Each uniaxial particle is described by a set of parameters that describe its ellipsoidal shape and the orientation of its principal axis in the molecular frame (Chapter 7, Figure 7.1).

The translation and rotational (torque) forces are given by the partial first derivatives, of the potential energy function, with respect to each of the variables. The translational force and torques^{4, 25} are given in section 6.5 and derived in the appendix of chapter 6 equations.

8.2.2 Coarse-grain dynamics simulation details

All the theory necessary for performing a CG molecular dynamics simulation is described in Chapter 6 and only the necessary details about the simulation are given here.

Each molecule is represented by a single site (or molecular site), its molecular orientation matrix (or unit vector) and by its dimensions along the semi-major and semi-minor axes. The GB interaction potential is used to calculate the interactions between the sites, and the resulting partial first derivatives of the GB are used to determine the translational and rotational (torques) forces. The GB potential is provided in more detail in Chapter 7 and the derivation of the force and torque are given in the appendix of Chapter 6. By using quaternion mathematics,²⁶ we have modified the Euler equations of motion implemented in the velocity Verlet integration algorithm (Chapter 6).^{26, 27} A timestep of 10 fs was used in all the CG dynamics simulations. The reason for choosing to implement quaternion rotational kinematics is two fold: (1) there are no trigonometric functions, especially in the denominator, which would result in singularities; and (2) as a result of the first point, this provides stable and well-behaved dynamics.²⁶ The CG simulations were run in a cubic box with a side length of 41.9 Å and the potential energy cutoff was set to 12.0 Å. The method of Berendsen et al.²⁸ was used to maintain a constant temperature of 298.15 K.

8.2.3 Conversion from coarse-grain particles to all atom benzene molecules

A point in space and orientation, determined by the orientation matrix (A), completely defines the CG benzene particles. The point represents the centre of mass of the benzene molecule and the orientation matrix (A), which defines its orientation, is built up from the quaternions as follows:²⁶

$$A = \begin{pmatrix} q_0^2 + q_1^2 - q_2^2 - q_3^2 & 2(q_1q_2 + q_0q_3) & 2(q_1q_3 - q_0q_2) \\ 2(q_1q_2 - q_0q_3) & q_0^2 - q_1^2 + q_2^2 - q_3^2 & 2(q_2q_3 + q_0q_1) \\ 2(q_1q_3 + q_0q_2) & 2(q_2q_3 - q_0q_1) & q_0^2 - q_1^2 - q_2^2 + q_3^2 \end{pmatrix}. \quad 8.6$$

Conversion from CG particles to the all atom description of benzene requires that we have to build a base benzene molecule from which we can work. Therefore, we build the all atom benzene in the yz-plane, with the centre of mass (**cm**) at the origin and the normal vector (**u_i**) along the x-axis, Figure 8.2.

To obtain the correct orientation of the benzene during the simulation, we multiply the pre-built benzene atomic coordinates (1×3 matrix) by the corresponding orientation (3×3) matrix. For example, the orientation of the C_1 - H_1 atoms is given by

$$C_1' = C_1 A = \begin{bmatrix} x_{C_1} & y_{C_1} & z_{C_1} \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{bmatrix}, \quad 8.7$$

$$H_1' = H_1 A = \begin{bmatrix} x_{H_1} & y_{H_1} & z_{H_1} \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{bmatrix}, \quad 8.8$$

where the prime (') represents the rotated coordinates in the molecular frame. This procedure is applied to all the atoms of the pre-built benzene molecule. It also applied to the conversion of all the CG particles to the all atom representation of the system. Concurrently, each benzene molecule's centre of mass is translated to coincide with its respective CG particle position.

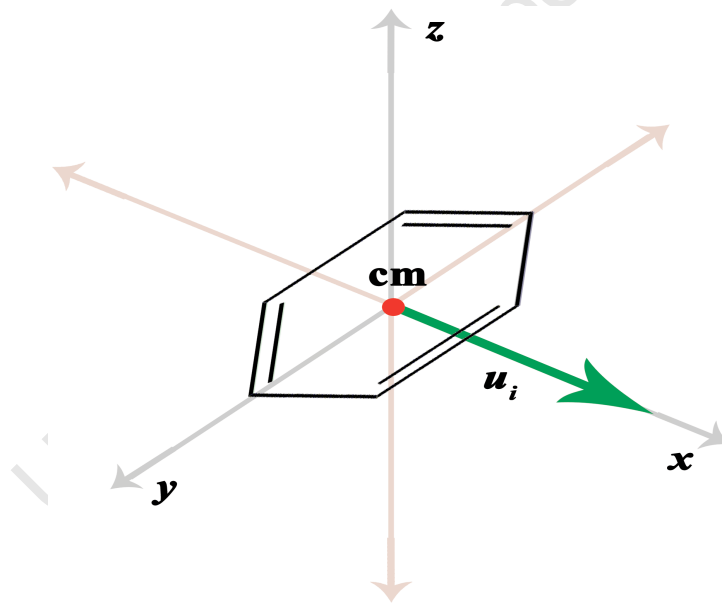


Figure 8.2: The details for the pre-built benzene molecule, which is built in the yz -plane. The centre of mass (cm) is placed at the origin and the relative orientation (u_i) is directed along the x -axis.

8.2.4 All atom simulation details

All atom molecular dynamics simulations were performed using the *CHARMM v33b2* program.²⁹ The liquid benzene simulation, containing 500 benzene molecules in a periodic cubic box with a side length of 42.3 Å, was carried out using the Leapfrog integration algorithm with a timestep of 1.0 fs.^{27, 30} The simulation was run for 10.0 ns in the canonical (NVT) ensemble. The temperature was maintained at 298.15 K using the

extended Hoover thermostat, with a mass for the thermal piston of 1000 kcal.ps²/mol (where *ps* = *picosecond*).³¹ The non-bonded cutoffs were set at 10.0, 12.0 and 14.0 Å on an atom-by-atom basis with force and potential switching methods for truncating the potential. The particle mesh Ewald (*PME*) method^{32, 33} was used to calculate the long-range interactions, with a grid size of 36 x 36 x 36 and 6th order B-spline using a kappa value of 0.34. All bonds involving hydrogen atoms were fixed via the constraint algorithm, *SHAKE*.³⁴

8.2.5 Parameterization

The parameterization of the Gay-Berne potential was performed using the module PYGENE (a python based Genetic Algorithm).³⁵ Here the parameters are set between a user-defined range with a probability and degree of mutation set at a default of 0.1. The genes are represented using floating point numbers, which allows for a wide-ranging description for genes compared to using binary strings. This GA method is based on Mendel genetics, which employs both mutation and crossover as methods of change. The selection is accomplished using the tournament selection method.

An initial population size of 100 was used, with a child count (i.e. the number of children to produce after each generation) of 50 and a child culling number of 5 (the number of children to be selected from each generation). The fitness function that was employed is given by the sum of the squares of the deviations (*S*), which is the function that has to be minimized

$$S(r, \theta_1, \theta_2, \phi; \varepsilon_{fac}, \varepsilon, \sigma_{||}, \sigma_{\perp}) = \sqrt{\sum_i \left[y_i(r_i, \theta_{1_i}, \theta_{2_i}, \phi_i) - F(r_i, \theta_{1_i}, \theta_{2_i}, \phi_i; \varepsilon_{fac}, \varepsilon, \sigma_{||}, \sigma_{\perp}) \right]^2}. \quad 8.9$$

For the parameterization of our CG benzene Gay-Berne particles from the benzene dimer vacuum free energy hyper-volume we used a large range for the parameters, given in Table 8.1.

Table 8.1: The range for the parameters used in the PYGENE genetic algorithm.³⁵ Large ranges were chosen to ensure proper sampling of parameter space.

	Minimum	Maximum
ε_{fac}	0.0	5.0
ε^o	0.1	3.0
$\sigma_{ }$	1.0	4.0
$\sigma_{\perp}$	3.0	7.0

8.2.6 Calculation of the diffusion constant

The diffusion constants for the all atom MD and CG dynamics simulations were calculated using the Einstein relation:²⁶

$$D = \frac{1}{6} \frac{\partial}{\partial t} [MSD]. \quad 8.10$$

The MSD is the mean square displacement, which is given as

$$MSD = \left\langle \left| \mathbf{r}_i(t) - \mathbf{r}_i(0) \right|^2 \right\rangle, \quad 8.11$$

where $\mathbf{r}_i$ is the molecule/particle position. These averages are computed for each of the N particles in the simulation, the results added together and divided by N to improve the statistical accuracy. The trajectories were unfolded (i.e., periodic boundary conditions removed) in order to perform the calculation.²⁶

8.3 Results and Discussion

Previously, we obtained a free energy hyper-volume that consisted of 4 reaction coordinates to comply with the 4 variables of the Gay-Berne interaction potential function.¹⁷ These variables are the distance between the molecules (r); the relative orientation of the molecular normal vectors with respect to the intermolecular distance vector (θ_1 and θ_2); and the relative orientation of the two molecular normal vectors (the dihedral angle, ϕ), Figure 7.1.²⁰

The genetic algorithm was chosen because of its robustness for fitting complicated landscapes and its ability to not get caught in local minima (Chapter 7).³⁶ The parameters required are the molecules semi-major ($\sigma_{\perp}$) and semi-minor ($\sigma_{\parallel}$) axis lengths, i.e. length and/or breadth of the molecule, interaction strength (ϵ^o), and the interaction strength ratio between the edge-to-edge to face-stacking ($\epsilon_{fac} = \epsilon_{face}/\epsilon_{edge}$) configurations. The values for μ and ν exponential parameters have been set to their canonical values, for μ and ν , of 2.0 and 1.0 respectively in this work. However, the μ values can range between -10.0 to 7.6 and the ν values can range from -12.5 to 0.4.³⁷

Can parameters for the Gay-Berne potential obtained from a free energy hyper-volume result in CG dynamics that is comparable with all atom models? Benzene coarse-grain (CG) parameters for use in liquid benzene CG dynamics simulations, obtained from the benzene dimer free energy hyper-volume in vacuum,²⁰ is summarized in Table 8.2. Calculations of the structural properties (in the form of probability distribution

functions) and the diffusion constant are performed. The results are compared with other CG studies performed Golubkov and Ren,⁴ using a Gay-Berne and point multipole expansion CG dynamics, all atom MD simulations done by Headen et al., using molecular dynamics refined with 6D neutron diffraction scattering data,³⁸ and experiment. It should be noted that electrostatic interactions has not implemented in any of the CG simulations. This serves two purposes: (i) to ascertain whether or not parameterization from a free energy hyper-volume produces results that are in agreement with all atom MD simulations and experiment; and (ii) to check the effect that the electrostatic multipole expansion, used by Golubkov and Ren,⁴ has on the CG dynamics simulations.

Table 8.2: The GR⁴ and SCRUI parameters used for the CG dynamics. ϵ_{fac} is the ratio of sandwich to edge-to-edge configurations, ϵ^o the usual interaction strength, $\sigma_{||}$ and $\sigma_{\perp}$ are the ellipsoidal lengths parallel and perpendicular to the vector, respectively.

	ϵ_{fac}	ϵ^o	$\sigma_{ }$	$\sigma_{\perp}$
GR⁴	3.67	0.54	2.03	4.65
SCRUI	1.87	1.31	2.85	4.05

8.3.1 Radial probability distribution function

A radial distribution function (RDF) is a statistical tool used to determine the radially averaged structure of the solvent (or solvent sites) around a solute (or solute site), as well as for pure solvent systems. It is used to investigate the structure of liquids in terms of nearest neighbour shells and how many molecules constitute each shell. RDFs are commonly used as a metric to check the validity of many force-field water model potentials³⁹⁻⁴³. Similarly, a check of the validity of the liquid benzene CG dynamics simulations is achieved by comparing the centre-of-mass (CoM) and carbon-carbon (C-C) RDFs with the corresponding all atom MD RDFs, shown in Figure 8.4.

The parameterization scheme used by GR was to adjust the parameters by matching physical properties to the all atom MD simulation and experimental results.⁴ The GR parameters (Figure 8.4, green curve) gives RDFs that are less structured in the first solvation shell (at $r = 5.5 \text{ \AA}$) compared to the all atom simulations (Figure 8.4, red curve). The broadening and decrease in probability can be attributed to the fact an all atom representation has been replaced by a single particle, as well as the rotation around the C6 symmetry axis that occurs.^{4, 24, 44} Unlike GR, the CG simulations performed in this study did not restrict the aforementioned rotation, which resulted in a slightly

more defined peak centred at 5.5 Å. Also, the lower structuring can be attributed to the small interaction strength parameter (ϵ^o) value used by the GR parameter set. The GR RDF_{C-C} (Figure 8.4a) appears to rise at 2.0 Å, whereas GR found that it rises at 2.8 Å.⁴ This can be explained by the fact that they used electrostatic (more specifically quadrupole-quadrupole) interactions in their simulations. They found that the electrostatic interaction to be repulsive for all parallel configurations, which exist at separation distances ≤ 4.0 Å. Therefore, the 0.8 Å difference is a result of the electrostatic repulsion that becomes hyperbolically increasing for distances less than 4.0 Å (approximately repulsive by 4.2 kcal/mol at 3.0 Å).⁴

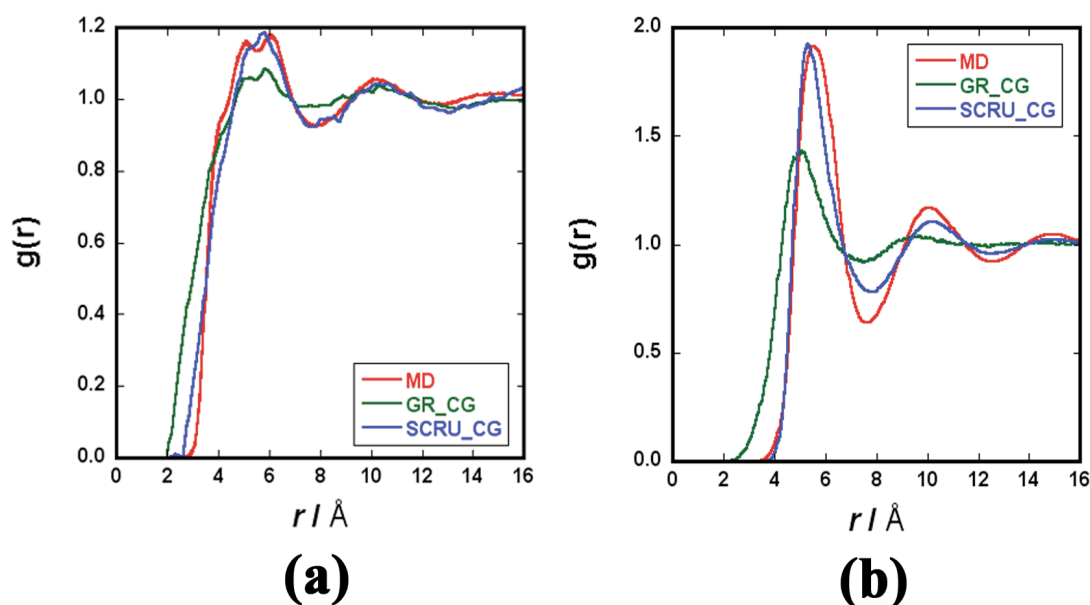


Figure 8.4: The (a) carbon-carbon, C-C (RDF,) and (b) centre of mass (RDF_{COM}) pair distribution functions (RDF) for the all atom MD (MD, red) and GR (green) and SCR (blue) CG simulations.

The notion that the parallel configurations result in the 0.8 Å difference is confirmed by the centre-of-mass-to-centre-of-mass RDF_{COM} (Figure 8.4b), which rises at a shorter distance, approximately 2.5 Å, compared to the all atom model. Thus, the electrostatics included in the GR CG dynamics simulations ensures that very closely stacked configurations do not form. The number of nearest neighbours (or number of molecules) for the first solvation ($r = 5.5$ Å) is found to be 9.7, which is less than that obtained by MD and experiments ($N = 12$). GR maintain that extensive parameterization should yield a unique set of parameters capable of more closely approximating the liquid benzene properties.⁴

Parameterization of the GB interaction potential from the benzene dimer free energy hyper-volume in vacuum (Table 8.2) results in a unique set of benzene GB parameters.²⁰

This has been termed as the SCRUI parameter set. The advantage of using the free energy hyper-volume is that it not only incorporates the benzene configuration orientational preferences, but inherently contains the configurational entropy. Liwo et al. have alluded to the fact that the entropy is an important factor at room temperature.⁴⁵ Another advantage of using free energies for parameterization is that it already includes all the intermolecular forces (i.e., dispersion and electrostatic interactions) that were present during the all atom free energy MD simulation. Therefore, the use of an anisotropic free energy hyper-volume is anticipated to provide a more accurate representation of liquid benzene. The RDF_{C-C} and RDF_{CoM} (Figure 8.4, blue curve), obtained using the SCRUI parameter set, shows a very good correlation with the all atom MD simulations. The RDF begins to rise at 3.0 Å, which is much closer to the all atom RDF than the GR simulation. The first peak centred around 5.5 Å appears to map very closely to the all atom MD simulations, in that two nearby peaks are observed in this region. The same is true for the location and the shape of the second peak at 10.0 Å, finally reaching bulk behaviour for $r \geq 14.0$ Å. The RDF_{CoM} (Figure 8.4b) shows slightly less correlation but the maxima and minima appear in the more or less the same location. However, the general shape agrees with the all atom MD RDF_{CoM} . The number of nearest neighbours in the first solvation ($r = 5.8$ Å) was found to be 11.8 (≈ 12), which agrees with others⁴⁶ and experiments.⁴⁷

8.3.2 Radial-angular probability distribution

The RDF only provides a one-dimensional radially averaged picture of the liquid structure, but to gain better insight into the anisotropic liquid structuring, a radial-angular distribution function have been calculated for the various simulations and are shown in Figure 8.5.

The all atom radial-angular probability distribution (Figure 8.5b) shows that the first solvation shell ($r = 5.5$ Å) is composed of parallel (more specifically PD, $\theta = 0^\circ$) and perpendicular ($\theta = 90^\circ$) configurations. At separation distances ≥ 8.0 Å, it shows the $\sin\theta$ dependence observed by Singh and Thornton for PHE pairs in proteins.⁴⁸

The GR CG simulation (Figure 8.5c) shows a very weak first solvation shell ($r = 5.5$ Å), which corresponds with the RDF (Figure 8.4), as well as a very low preference for either parallel or perpendicular configurations. The small ϵ^o value and the fact that no electrostatic interactions are used results in rather weak interactions between the GR CG benzene particles. This manifests itself as the low probability distributions that are observed.

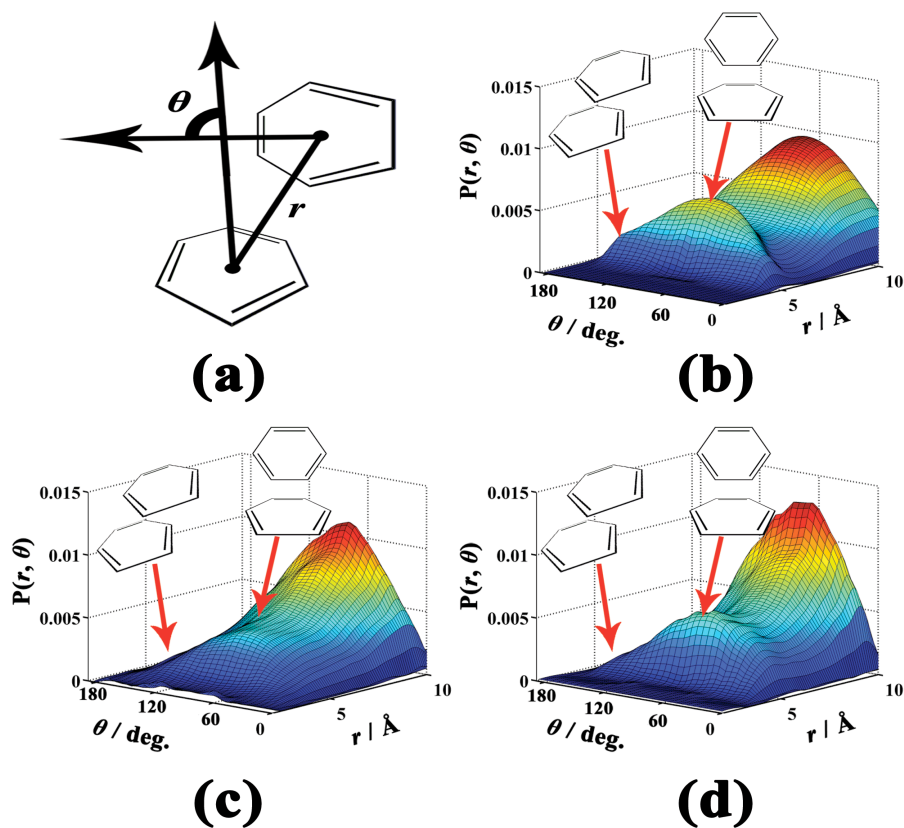


Figure 8.5: (a) Diagram depicting the definition of the angle between the normals, θ . Parallel orientations are defined by $\theta = 0^\circ (\pm 20^\circ)$ and perpendicular orientations by $\theta = 90^\circ (\pm 20^\circ)$. The 2D radial-angular probability distributions, $P(r, \theta)$, are shown for the (b) all atom MD, (c) GR and (d) SCRUCG simulations.

The SCRUCG parameters have been obtained from the free energy and, thus, already include all intermolecular interactions. This results in the SCRUCG simulation (Figure 8.5d) showing greater structuring in the first solvation shell ($r = 5.5 \text{ \AA}$), which agrees with the all atom model better than the GR parameters. Like the all atom model, the SCRUCG probability shows a greater tendency to form parallel and perpendicular configurations compared to the GR model. Both CG models display the $\sin\theta$ dependence at large separation distances (i.e., $\geq 8.0 \text{ \AA}$). Therefore, the SCRUCG parameters perform better than the GR parameters in mimicking the 2D structure of liquid benzene.

8.3.3 Three-dimensional solvent structuring

RDFs only provide a radially averaged description of the liquid structure, while the 2D radial-angular provides better insight into the anisotropic solvent structuring. However, spatial distribution functions (SDF) provide fully anisotropic 3D description of the liquid structure.

Liquids do not often structure isotropically but, rather, are inherently anisotropic due to the anisotropic nature of intermolecular forces. Headen et al. showed that the SDF for

benzene liquid displays a lantern-like pattern, with 6-lobes, around benzene.³⁸ This is indicative of a variety of parallel and perpendicular (Figure 8.5a) found for the first solvation shell.³⁸ The all atom MD simulations performed here display similar 3D anisotropic structuring (Figure 8.6a).

The SDF for the GR CG simulation appears to have a similar overall shape, but shows lower probability density indicating less structuring in this system. The 6-lobes of the lantern appear to be less pronounced and more disordered, Figure 8.6b, in comparison with the all atom model (Figure 8.6a). This corresponds with the distribution functions (Figures 8.4 and 8.5c) and indicates that the inclusion of electrostatic interactions may be important for CG simulations, which do not obtain GB parameters from free energies. Strong densities above and below the benzene ring indicate the existence of parallel configurations. The density around the rim of the benzene is indicative edge-to-edge configurations, which is substantiated by the fact the GR used an ϵ_{fac} value (= 3.67) that favours edge-to-edge arrangements.

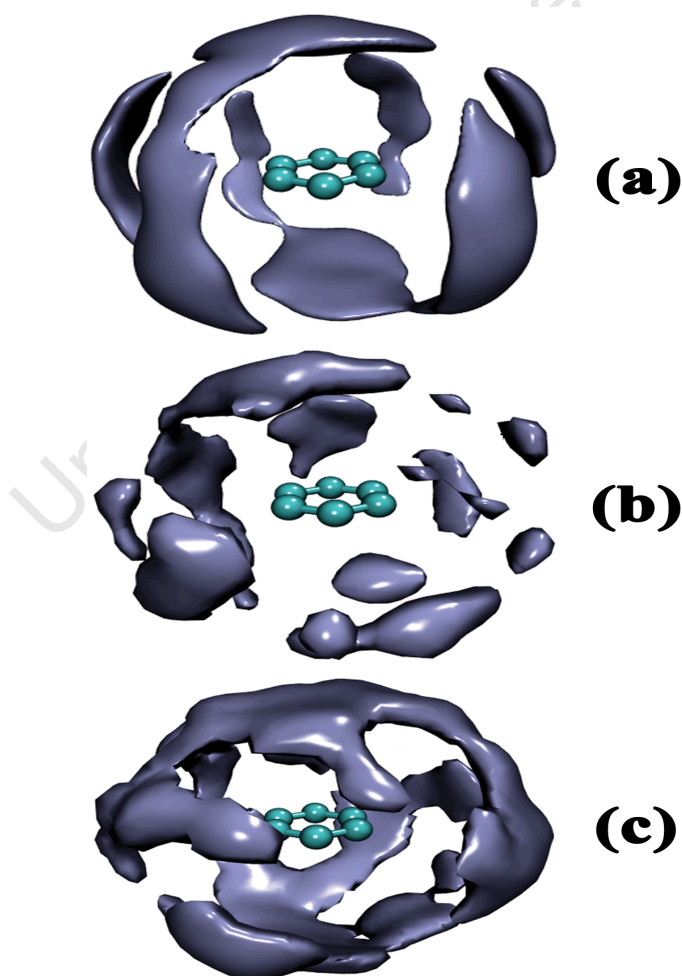


Figure 8.6: The centre-of-mass-to-centre-of-mass (CoM-CoM) spatial distribution functions (SDFs) for the (a) all atom MD; (b) GR and (c) SCRUG dynamics simulations.

The CG simulation using the SCRU parameters resulted in an SDF that shows the shape of a lantern with 6-lobes (Figure 8.6c), which is in agreement with our MD simulations (Figure 8.6a) and the results obtained by Headen et al.³⁸ Although the lobes appear to be slightly disordered, it appears to show greater structuring compared to the GR-CG model. The 6-lobes are indicative of the perpendicular configurations, while the distribution above and below the benzene point to parallel configurations. This agrees with the 2D radial-angular probability functions plotted in Figure 8.5d. Thus, the parameterization scheme presented here (i.e., by obtaining Gay-Berne parameters from an anisotropic free energy hyper-volume) appears to provide a fairly accurate anisotropic description of liquid benzene. This can be attributed to the fact the free energy already incorporates all the intermolecular interactions and, more importantly, the entropy.

8.3.4 Diffusion constant

The diffusion constant for the all atom MD simulations was found to reproduce the experimental value quite well (Table 8.3). The GR parameter set gave a self-diffusion constant slightly larger ($2.29 \times 10^{-9} \text{ m}^2/\text{s}$) than MD and experimental values (2.15×10^{-9} and $2.20 \times 10^{-9} \text{ m}^2/\text{s}$, respectively), while the SCRU parameter set underestimated ($2.05 \times 10^{-9} \text{ m}^2/\text{s}$) the self-diffusion constant.

Table 8.3: The density and self-diffusion constant for the all atom MD, GR CG and SCRU CG dynamics simulations compared with experiment.

	Density (g/cm ³)	Diffusion Constant (10 ⁻⁹ m ² /s)
GR⁴	0.879	2.29
SCRU	0.879	2.05
MD (CHARMM)	0.857	2.15
Exp.	0.870	2.20 ⁴⁹

The difference between the MD and CG diffusion constants are expected due to the coarse-graining.⁴ The difference between the GR and SCRU diffusion is due to the differences in the parameter sets. The GR $\sigma_{||}$ and $\sigma_{\perp}$ values indicate a flat disk-like particle and the ϵ_{fac} value of 3.67 results in edge-to-edge arrangements being preferred. The SCRU parameters describe particles that are less disk-like with an ϵ_{fac} value of 1.87, which has no preference for either stacked or edge-to-edge arrangements. The diffusion constant differences result in the GR model having a higher diffusion constant because

(due the aforementioned facts, weaker interactions small, determined by the small ϵ^o value, and no electrostatics) the benzene particles are allowed “slip” between edge-to-edge arrangements more easily than the SCRUI model. Nonetheless, the CG diffusion constants show good correlation with the all atom MD simulation and experiment.

8.3.5 Comparative timings

Finally, the point of coarse-graining is to not only to reduce the size-scale of the problem but also to attain simulations timescales that are reach biologically relevant. Therefore, the main motivation for obtaining a unique set of CG GB parameters, which yields an accurate anisotropic description of the system, is to increase the computational efficiency without a sacrifice in accuracy. Table 8.4 shows the time taken to run the all atom MD and CG dynamics simulation for 100 ps, as well as the relative speed up. The MD timestep is 1 fs and the CG timestep is 10 fs. The timing data has been extrapolated to determine the time it would require to run a 1 μ s simulations.

Table 8.4: Timings for a 100 ps for the all atom MD and CG dynamics simulations. The estimated time to run a 1 μ s simulation is provided for comparison. **Note:** 1 μ s is equivalent to 10^9 MD dynamics steps, and 10^8 CG dynamics steps – MD timestep is 1 fs, while the CG timestep is 10 fs.

Simulation Methodology	Timestep Δt (fs)	Time taken to run 100 ps (s)	Steps / second	Estimated time for 1 μ s simulation (days)
MD (CHARMM);	$\Delta t=1$ fs	16200	6.2	187.5
CG (SCRUI-CG);	$\Delta t=10$ fs	47	212.8	5.5
Relative Speed Up;	$\Delta t=1$ fs	345	--	--

The all atom MD simulation was composed of 500 benzene molecules in a cubic box with a side length of 42.3 Å. The CG dynamics simulation was composed of 500 GB particles in cubic box with a side length of 41.8 Å. The simulations were run on a single first generation Intel i5 core (clocked at 2.8 GHz) and the speed up obtained is approximately 345 times faster for the CG over the MD simulations with a timestep of 1.0 fs. For biological systems a timescale of at least 1.0 μ s is required. Therefore, an all atom MD simulation would require 188 days (~6 months), while the CG dynamics simulations would only need 5.5 days to complete a 1.0 μ s simulation. The drawback of CG dynamics is that the molecule is described by an ellipsoid and, thus, all atomic information is lost. However, converting the CG particles back into its all atom representation and

performing a short simulation, using an all atom potential energy function, can return the all atom detail and remove any bad contacts. This is a method that Liwo et al. have employed in their CG simulations of proteins.⁴⁵

8.4 Conclusions

A multi-dimensional anisotropic free energy hyper-volume has been used to parameterize the Gay-Berne interaction potential for benzene. The advantage of this method, over others, is that all the intermolecular interactions and entropy are inherently included in the free energy. The application to liquid benzene demonstrates that the parameters obtained by this method reproduces the MD and experimental data quite well, and is found to be superior to the Gay-Berne parameters reported by Golubkov and Ren.⁴

The structural properties, based on probability distribution functions, and the diffusion constant analysis reveals good correlation between the (SCRU) CG dynamics, with parameters derived from a free energy hyper-volumes, and the all atom MD simulations and experimental results. The first solvation shell ($r = 5.5 \text{ \AA}$) is observed to be composed of both parallel and perpendicular configurations, shown by the radial-angular and spatial distribution functions, in agreement with Headen et al., who used MD simulations refined with 6D neutron diffraction data.³⁸

In contrast, the Gay-Berne parameters, reported by Golubkov and Ren,⁴ produced CG dynamics simulation that were found to have distribution functions that are significantly less structured for the first solvation shell ($r = 5.5 \text{ \AA}$) and a diffusion constant that is higher than the experimentally observed value. These discrepancies can be attributed to the weak CG particle interaction, determined by the small ε^o value in Golubkov and Ren parameter set, as well as the lack of electrostatic interactions. Therefore, electrostatic interactions play a crucial role in CG dynamics simulations if the parameterization of the interaction potential was not performed using a free energy.

The computational efficiency of the CG dynamics is greatly increased compared to the all atom simulations. We found that the CG dynamics is found to be approximately 35 times faster than the corresponding all atom MD simulation. Eliminating high-frequency motions, heavy-atom-hydrogen-atom bond vibrations, by coarse-graining also allowed for a larger timestep (10 fs for CG compared to 1 fs for the all atom MD simulations) without detrimental loss of accuracy, which is the goal of this simulation technique

This study shows a proof of concept for the parameterization of a relevant CG intermolecular interaction potential from an anisotropic free energy hyper-volume. This method appears to be superior (in terms of structure and dynamics) compared to other parameterization scheme used by Golubkov and Ren. Therefore, parameterization of a CG benzene model from an atomistically detailed anisotropic benzene dimer vacuum free energy hyper-volume shows that this method allows a CG model to retain its connection with an all atom benzene model, providing strong credence to the theoretical notion that relates the all atom potential and coarse-grain potential energies via the Helmholtz free energy:

$$\begin{aligned} \exp\left(-\frac{A}{k_B T}\right) &= \text{const.} \int d\mathbf{x} \exp\left[\frac{-V(\mathbf{x})}{k_B T}\right] \\ &\approx \text{const.} \int d\mathbf{x}_{CG} \exp\left[\frac{-V_{CG}(\mathbf{x}_{CG})}{k_B T}\right] \end{aligned} \quad 8.12$$

However, the drawback here is that the method is only applicable at the temperatures that the free energy hyper-volumes were calculated (i.e., preventing direct transferability to other temperatures), and by being dependent on the explicit solvent (water) model used, inhibiting the transferability between the various solvent system models. This would require a further development in the underlying theory to understand and relate the free energy hyper-volumes with respect to temperature. Other future work includes: inclusion of the Nose-Hoover thermostat, which generates canonically distributed molecular velocities, and Brownian (an extreme of Langevin) dynamics.

Owing to the success of this parameterization scheme, proposed further work along this direction could include: the parameterization of the Gay-Berne interaction potential for biologically relevant aromatic amino acids for use in the protein folding simulations; to perform CG dynamics simulations of other aromatic molecules applicable for studying liquid crystal biophysics; and develop an accurate CG water model, parameterized from the water free energy hyper-volumes,⁵⁰ that could possibly perform structurally as well as TIP5P-Ew.

8.5 References

1. Voth, G. A., Introduction. In *Coarse-Graining of Condensed Phase and Biomolecular Systems*, Voth, G. A., Ed. CRC Press: New York, 2009; p 1.
2. Nielsen, S. O., *et al.*, Coarse Grain Models and the Computer Simulation of Soft Materials. *J. Phys.: Condens. Matter* **2004**, *16* (15), R481.
3. Matysiak, S.; Clementi, C., Characterization of Protein-Folding Landscapes by Coarse-Grained Models Incorporating Experimental Data. In *Coarse-Graining of Condensed Phase and Biomolecular Systems*, Voth, G. A., Ed. CRC Press: New York, 2009; p 157.
4. Golubkov, P. A.; Ren, P. Y., Generalized Coarse-Grained Model Based on Point Multipole and Gay-Berne Potentials. *J. Chem. Phys.* **2006**, *125* (6), 064103.
5. Darre, L., *et al.*, Another Coarse Grain Model for Aqueous Solvation: Wat Four? *J. Chem. Theory Comput.* **2010**, *6* (12), 3793.
6. Egorov, S. A., Hydrophobic Interactions with Coarse-Grained Model for Water. *J. Chem. Phys.* **2011**, *134* (23), 234509.
7. Liwo, A., *et al.*, Calculation of Protein Backbone Geometry from Alpha-Carbon Coordinates Based on Peptide-Group Dipole Alignment. *Protein Sci.* **1993**, *2*, 1697.
8. Liwo, A., *et al.*, Prediction of Protein Conformation on the Basis of a Search for Compact Structures: Test on Avian Pancreatic Polypeptide. *Protein Sci.* **1993**, *2*, 1715.
9. Derreumaux, P.; Mousseau, N., Coarse-Grained Protein Molecular Dynamics Simulations. *J. Chem. Phys.* **2007**, *126* (2), 025101.
10. Cagin, T., *et al.*, Multiscale Modeling and Simulation Methods with Applications to Dendritic Polymers. *Computational and Theoretical Polymer Science* **2001**, *11* (5), 345.
11. Schug, A., *et al.*, Coarse-Grained Structure-Based Simulations of Proteins and Rna. In *Coarse-Graining of Condensed Phase and Biomolecular Systems*, Voth, G. A., Ed. CRC Press: New York, 2009; p 123.
12. Olson, W. K., *et al.*, Insights into the Sequence-Dependent Macromolecular Properties of DNA from Base-Pair Level Modeling. In *Coarse-Graining of Condensed Phase and Biomolecular Systems*, Voth, G. A., Ed. CRC Press: New York, 2009; p 205.
13. Fukunaga, H., *et al.*, A Coarse-Graining Procedure for Flexible Polymer Chains with Bonded and Nonbonded Interactions. *J. Chem. Phys.* **2002**, *116* (18), 8183.
14. Marrink, S. J.; Mark, A. E., The Mechanism of Vesicle Fusion as Revealed by Molecular Dynamics Simulations. *J. Am. Chem. Soc.* **2003**, *125* (37), 11144.
15. Orsi, M., *et al.*, A Quantitative Coarse-Grain Model for Lipid Bilayers. *J. Phys. Chem. B* **2007**, *112* (3), 802.
16. Whitehead, L., *et al.*, Molecular Dynamics Simulation of the Hydrocarbon Region of a Biomembrane Using a Reduced Representation Model. *J. Comput. Chem.* **2001**, *22* (14), 1622.
17. Gay, J. G.; Berne, B. J., Modification of the Overlap Potential to Mimic a Linear Site-Site Potential. *J. Chem. Phys.* **1981**, *74* (6), 3316.
18. Allen, M. P., *et al.*, Computer Simulation of Liquid Crystals. *J. Phys.: Condens. Matter* **1996**, *8* (47), 9433.
19. Allen, M. P., *et al.*, Molecular Dynamics Calculation of Elastic Constants in Gay-Berne Nematic Liquid Crystals. *J. Chem. Phys.* **1996**, *105* (7), 2850.
20. Gamielien, M. R., *et al.*, Hydration-Determined Orientational Preferences in Aromatic Association from Benzene Dimer Free Energy Volumes. *J. Phys. Chem. B* **2012**, *116* (1), 324.
21. Cacelli, I., *et al.*, Modeling Benzene with Single-Site Potentials from Ab Initio Calculations: A Step toward Hybrid Models of Complex Molecules. *J. Chem. Phys.* **2004**, *120* (8), 3648.
22. Zacharopoulos, N., *et al.*, Coarse Graining Using Pretabulated Potentials: Liquid Benzene. *J. Chem. Phys.* **2005**, *122* (24), 11.

23. Walsh, T. R., Towards an Anisotropic Bead-Spring Model for Polymers: A Gay-Berne Parametrization for Benzene. *Mol. Phys.* **2002**, *100* (17), 2867.
24. Lorenz, S., *et al.*, Towards an Anisotropic Bead-Spring Model for Ribbonlike Polymers: Monte Carlo Simulations of Liquid Benzene. *J. Chem. Phys.* **2003**, *119* (5), 2903.
25. Allen, M. P.; Germano, G., Expressions for Forces and Torques in Molecular Dynamics Simulations Using Rigid Bodies. *Mol. Phys.* **2006**, *104* (20-21), 3225.
26. Allen, M. P.; Tildesly, D. J., Computer Simulations of Liquids. *Oxford University Press* **1987**, London, Chapter 2.
27. Hockney, R. W.; Eastwood, J. W., *Computer Simulations Using Particles*. McGraw-Hill: New York, 1981.
28. Berendsen, H. J. C., *et al.*, Molecular-Dynamics with Coupling to an External Bath. *J. Chem. Phys.* **1984**, *81* (8), 3684.
29. Brooks, B. R., *et al.*, Charmm: The Biomolecular Simulation Program. *J. Comput. Chem.* **2009**, *30* (10), 1545.
30. Amini, M., *et al.*, Time Integration in Particle Models. *Comput. Phys. Commun.* **1987**, *44* (1-2), 83.
31. Hoover, W. G., Canonical Dynamics: Equilibrium Phase-Space Distributions. *Phys. Rev. A* **1985**, *31* (3), 1695.
32. York, D. M., *et al.*, The Effect of Long-Range Electrostatic Interactions in Simulations of Macromolecular Crystals: A Comparison of the Ewald and Truncated List Methods. *J. Chem. Phys.* **1993**, *99* (10), 8345.
33. Essman, U., *et al.*, A Smooth Particle Mesh Ewald Method. *J. Chem. Phys.* **1995**, *103* (19), 8577.
34. Andersen, H. C., A "Velocity" Version of the Shake Algorithm for Molecular Dynamics Simulations. *J. Comp. Phys.* **1983**, *52* (1), 24.
35. Mcnab, D. *Pygene - Simple Python Genetic Algorithms/Programming Library*, 2011.
36. Mitchell, M., *An Introduction to Genetic Algorithms*. The MIT Press: Cambridge, Massachusetts, 1998.
37. Gordon, M. S., *et al.*, Accurate Methods for Large Molecular Systems. *J. Phys. Chem. B* **2009**, *113* (29), 9646.
38. Headen, T. F., *et al.*, Structure of Pi-Pi Interactions in Aromatic Liquids. *J. Am. Chem. Soc.* **2010**, *132* (16), 5735.
39. Jorgensen, W. L., Quantum and Statistical Mechanical Studies of Liquids. 10. Transferable Intermolecular Potential Functions for Water, Alcohols, and Ethers. Application to Liquid Water. *J. Am. Chem. Soc.* **1981**, *103* (2), 335.
40. Jorgensen, W. L., *et al.*, Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79* (2), 926.
41. Horn, H. W., *et al.*, Development of an Improved Four-Site Water Model for Biomolecular Simulations: Tip4p-Ew. *J. Chem. Phys.* **2004**, *120* (20), 9665.
42. 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* (20), 8910.
43. Berendsen, H. J. C., *et al.*, The Missing Term in Effective Pair Potentials. *J. Phys. Chem.* **1987**, *91* (24), 6269.
44. Harvey, S. C., *et al.*, The Flying Ice Cube: Velocity Rescaling in Molecular Dynamics Leads to Violation of Energy Equipartition. *J. Comput. Chem.* **1998**, *19* (7), 726.
45. Liwo, A., *et al.*, Simulation of Protein Structure and Dynamics with the Coarse-Grained Unres Force Field. In *Coarse-Graining of Condensed Phase and Biomolecular Systems*, Voth, G. A., Ed. CRC Press: New York, 2009; p 108.
46. Jorgensen, W. L.; Severance, D. L., Aromatic Aromatic Interactions - Free-Energy Profiles for the Benzene Dimer in Water, Chloroform, and Liquid Benzene. *J. Am. Chem. Soc.* **1990**, *112* (12), 4768.

47. Narten, A. H., X-Ray-Diffraction Pattern and Models of Liquid Benzene. *J. Chem. Phys.* **1977**, 67 (5), 2102.
48. Singh, J.; Thornton, J. M., The Interaction between Phenylalanine Rings in Proteins. *FEBS Lett.* **1985**, 191 (1), 1.
49. Collings, A. F.; Woolf, L. A., Self-Diffusion in Benzene under Pressure. *Journal of the Chemical Society, Faraday Transactions 1: Physical Chemistry in Condensed Phases* **1975**, 71.
50. Strumpfer, J.; Naidoo, K. J., Computing Free Energy Hypersurfaces for Anisotropic Intermolecular Associations. *J. Comput. Chem.* **2010**, 31 (2), 308.

University of Cape Town