The explicit polarization theory as a quantum mechanical force field and the development of coarse-grained models for simulating crowded systems of many proteins

A DISSERTATION

SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL

OF THE UNIVERSITY OF MINNESOTA

BY

Michael John Morgan Mazack

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

Doctor of Philosophy

Prof. Jiali Gao, Advisor

January, 2014

Acknowledgements

January 30th, 2014

The journey of a thousand miles begins with a single step. –Laozi

My journey through college began at the age of 16 over a decade ago. It was around that time when I had developed a vague goal to pursue a career which combined mathematics, physics, and computer programming. Today represents the culmination of that era, which never would have come to pass without the help and support from a multitude of individuals.

Although my childhood was filled with many challenges, I am blessed to have had parents who always encouraged me to pursue my education as far as it could take me – for that, I am extraordinarily grateful. I am also thankful for the other members of my family who believed in my pursuit, especially my aunt Elaine.

Elaine graciously supported me through my undergraduate years, and enabled my life-changing study abroad experience in Japan. Without her generous support, I would not be where I am today. As a retired scientist, her insight and advice were of great value and were well-received. I commend and celebrate her distinguished role in my education, for which I am extremely grateful.

A very special thank you and appreciation goes to my advisor, Jiali Gao, for his abundance of patience in enduring my less than perfect attitude and stubbornness. I thank him for the many meetings we had to discuss and debate scientific and technical details of projects, and his willingness to have me drop by his office whenever he was there. I also thank Jiali for funding my studies through his National Institute of Health grants (GM46376 and RC1-GM091445) and allowing me to attend the 2011 WATOC conference in Santiago de Compostela, Spain. I owe Jiali a great thanks for believing in me as a student, and making the sacrifice to take me under his wing when no one else was willing – thank you, Jiali.

I owe many thanks to my former teachers from Conway Middle School (Jennifer Wilson, John Townsend, Linda Talman), former teachers from Mount Vernon High School (Paul Everyhope, Dorian Drury, Mr. Hunt), former instructors from Skagit Valley College (Greta Kocol, Chuck Stevens, Ann Zukoski, Brad Smith), and former professors from Western Washington University (Tjalling Ypma, Branko Ćurgus, Stephen McDowall, Masanori Deguchi) for their quality teaching and inspiration. I also thank my dissertation committee at the University of Minnesota (Yousef Saad, Carlos Sosa, George Karypis) for their insight and help through this final phase of my formal education.

Much gratitude is given to my current and former graduate student colleagues, who struggled with me through the depths of graduate school, especially Brian Heinze, Jennifer Angelosante, and Johnny Hu at Western Washington University, and Carlos Andres Ramirez-Mondragon and Peng Zhang at the University of Minnesota. I'd also like to thank Jiali's former postdoc, Peng Bao, for making the group more entertaining.

My co-authors (Alessandro Cembran, Jaebeom Han, Donald Truhlar) deserve a mention for their collaboration and cooperativeness, despite having much more work of their own to attend to. I also thank Pal Jedlovszky, Sylvia McLain, Matthew McGrath, and J. Ilja Siepmann for providing experimental and computational data for hydrogen fluoride in Chapter 4. My gratitude also extends to Martin Karplus at Harvard

University and his former group members who allowed me fill in for Jiali and present Chapter 6 at the CHARMM developers meeting back in 2010. Similarly, I would also like to thank Klaus Schulten and his group (Jim Phillips in particular) for hosting a visit to the Beckman Institute where I was able to present Chapter 2 and receive insight on improving my X-Pol implementation. My thanks also goes to David Goodsell of the Scripps Research Institute for granting me permission to use his artistic rendition of an *E. coli* cell cross-section in Chapter 5.

I thank the Minnesota Supercomputing Institute for providing half of my first year's funding as well as computer time, technical support, and helpful tutorials throughout my time at the U of M. I also extend my thanks to the organizers and attendees of the 2013 International Summer School on HPC Challenges in Computational Science for making a great week at New York University.

An incalculable amount of gratitude, appreciation, and love goes to my wife, Jane, who graciously endured the long, lonely, and depressing months of my dissertation writing. Her encouragement, selflessness, and great patience throughout that time mean much more to me than words can express – thank you, Jane.

Although it may sound cliché, I would also like to thank God Almighty, The Ancient of Days, whose peace transcends all understanding through Christ Jesus. May I never forget His faithfulness, great mercy, and redemption in my life.

And to the many whom I have not mentioned here, I apologize for the omission, but thank you collectively for being a part of my journey.

The end of one era marks the beginning of another. The journey of a thousand miles begins with a single step. Let us begin again.

Michael J. M. Mazack, Ph.D.

Dedication

To my aunt Elaine, who has graciously supported my interest in science from the beginning.

Abstract

This dissertation consists of two parts.

The first part concerns the use of explicit polarization theory (X-Pol), the semiempirical polarized molecular orbital (PMO) method, and the dipole preserving, polarization consistent (DPPC) charge model as a quantum mechanical force field (QMFF). A detailed discussion of Hartree-Fock theory and X-Pol is provided, along with expressions for the energy and the analytical first derivative of this QMFF. Test cases for this QMFF with extensive comparisons to experimental data and other models are provided for water (XP3P) and hydrogen fluoride (XPHF), showing that the PMO/X-Pol/DPPC approach discussed in this dissertation is competitive with the most accurate models for those two chemical species over a wide range of chemical and physical properties.

The second part of this dissertation concerns the development and application of coarse-grained models for protein dynamics. First, a coarse-grained force field (CGFF) for macromolecules in crowded environments is introduced and described along with a visualization environment for the cartoon-like rendering of biomolecules *in vivo*. This CGFF is tested against experimental diffusion coefficients for myoglobin (Mb) at a wide range of concentrations, including volume fractions as high as 40%, finding it to be surprisingly accurate for its simplicity and level of coarseness. Second, an analytical coarse-grained (ACG) model for mapping the internal dynamics of proteins into a spherical harmonic expansion is described.

Contents

Acknowledgements **i**

Dedication **iv**

Abstract **v**

List of Tables **xii**

List of Figures **xix**

Preface **xxviii**

1 Introduction **1**

1.1 Novelty of Results 4

1.2 Description of Chapters 5

1.2.1 The Explicit Polarization Theory and its Analytical First Derivative 5

1.2.2 Quantum Mechanical Force Field for Water 6

1.2.3 Quantum Mechanical Force Field for Hydrogen Fluoride 7

1.2.4 MACROSHAKER: A Coarse-Grained Force Field for Crowded Systems of Many Proteins . . . 7

1.2.5 Internal Dynamics of an Analytically Coarse-Grained Protein . . . 8

1.2.6 Conclusion & Discussion . . . 9

1.2.7 A Charge-Fitting Procedure for Coarse-Grained Proteins . . . 9

1.2.8 Algorithm for Spherical Harmonic Expansion and Evaluation . . . 9

2 The Explicit Polarization Theory and its Analytical First Derivative 10

2.1 Introduction . . . 10

2.2 Hartree-Fock Theory . . . 13

2.2.1 Born-Oppenheimer Approximation . . . 14

2.2.2 Wave Function Description . . . 16

2.2.3 Electronic Integral Calculations . . . 18

2.2.4 Self-Consistent Field Procedure . . . 21

2.2.5 Analytical First Derivative of Energy . . . 24

2.3 Explicit Polarization Theory . . . 26

2.3.1 Wave Function Description . . . 27

2.3.2 The X-Pol Hamiltonian . . . 28

2.3.3 Double Self-Consistent Field Procedure . . . 30

2.3.4 Analytical First Derivative of Energy . . . 32

2.4 Modified Neglect of Diatomic Overlap Approximation . . . 34

2.4.1 Integral Calculations . . . 35

2.4.2 $\mathbf{H}^{\text{core}}$ Matrix . . . 36

2.4.3 Fock Matrix . . . 37

2.4.4 Core-Core Repulsion . . . 38
2.4.5 X-Pol with MNDO . . . 39
2.4.6 Polarized Molecular Orbital Method . . . 39
2.5 Dipole-Preserving and Polarization-Consistent Charge . . . 41
2.5.1 DPPC Method . . . 42
2.5.2 Modification to X-Pol Fock matrix by DPPC . . . 44
2.5.3 DPPC Gradient . . . 47
2.6 Sample Data . . . 51
2.7 Conclusion . . . 51
2.8 Supporting Information . . . 55

3 Quantum Mechanical Force Field for Water 56

3.1 Introduction . . . 56
3.2 Method . . . 60
3.2.1 Polarized molecular orbital model for water . . . 60
3.2.2 Explicit polarization theory . . . 64
3.2.3 The XP3P model for liquid water . . . 67
3.3 Computational Details . . . 70
3.4 Results and Discussion . . . 73
3.4.1 Gas-phase properties . . . 73
3.4.2 Liquid properties . . . 84
3.5 Conclusions . . . 104
3.6 Supporting Information . . . 106

4 Quantum Mechanical Force Field for Hydrogen Fluoride 107

4.1 Introduction . . . 107
4.2 Semiempirical PMO Method for Hydrogen Fluoride . . . 111

4.2.1 Polarized Molecular Orbital Method 111
4.2.2 Motivation for using PMOw 113
4.2.3 Fluorine Parameters for PMOw 116
4.2.4 Hydrogen Fluoride Clusters 116
4.2.5 $(HF)_n(H_2O)_n$ Complexes 119
4.3 XPHF Model for Liquid HF 124
4.3.1 Wave Function Description 125
4.3.2 Effective Hamiltonian and Energy Expression 126
4.3.3 X-Pol with PMOw 128
4.4 Simulation of Liquid HF with XPHF 128
4.4.1 Simulation Setup 129
4.4.2 Energetic Properties 130
4.4.3 Electronic Properties 131
4.4.4 Density 132
4.4.5 Structural Properties 134
4.4.6 Diffusion 139
4.4.7 Deviation in Supercritical States 141
4.5 Conclusion 142

5 MACROSHAKER: A Coarse-Grained Force Field for Crowded Systems of many Proteins 144

5.1 Introduction 144
5.2 Dynamics in MACROSHAKER 148
5.2.1 Ermak-McCammon Integration 149
5.2.2 Rigid-body Mechanics 151
5.3 Coarse-Grained Force Field 154
5.3.1 Coarse-Grained Model 155

5.3.2 Potential Energy 159

5.4 Graphical User Interface 161

5.5 Visualization in MACROSHAKER 163

5.5.1 Spherical Harmonic Expansions 165

5.5.2 Non-photorealistic Rendering 170

5.6 Illustrative Example: Diffusion of Myoglobin as a Function of Protein Concentration 174

5.7 Conclusion 179

6 Internal Dynamics of an Analytically Coarse-Grained Protein 180

6.1 Introduction 180

6.2 Method 184

6.2.1 Description of a Macromolecular Particle 185

6.2.2 Spherical Harmonic Representation of a Macromolecular Particle 187

6.2.3 Dynamic Motion 189

6.3 Numerical Considerations 198

6.4 Discussion 199

6.5 Conclusion 206

6.6 Supporting Information 208

7 Conclusion & Discussion 209

7.1 Conclusion 209

7.2 Future Work on QMFFs 210

7.2.1 Variational many-body expansion 210

7.2.2 Unpolarizable core in PMO 212

7.2.3 Additional parameters for PMO 213

7.3 Future work on CGFFs 213

7.3.1 Macromolecular Assembly 213
7.3.2 Metabolic Processes 214
7.3.3 Interactive Visualization 216

References **217**

Appendix A. A Charge-Fitting Procedure for Coarse-Grained Proteins **262**
A.1 Introduction 262
A.2 Background 262
A.2.1 Poisson Equation 262
A.2.2 Poisson-Boltzmann Equation 263
A.2.3 Debye-Hückel Equation 264
A.3 Charge Fitting 265
A.3.1 Description 265
A.3.2 Computational Details 266
A.4 Selecting Charge Sites 267
A.4.1 Coarse-Graining of Byron 268
A.4.2 Surface Sites 268
A.4.3 ACG Charges 269
A.5 Conclusion 269

Appendix B. Algorithm for Spherical Harmonic Expansion and Evaluation **270**
B.1 Spherical Harmonic Expansion 272
B.2 Spherical Harmonic Evaluation 274

List of Tables

2.1 Pauling electronegativity for the six most essential elements in living organisms. A list for elements up to atomic number 94 is provided in the CRC handbook. [1] 43

2.2 Cartesian coordinates for the water dimer used in our tests of various combinations of theory, semiempirical method, and population analysis scheme. The dashed line in the table indicates the separation of the fragments for X-Pol calculations. 52

2.3 Heat of formation/X-Pol energy (kcal/mol), dipole moment (Debye), and charges (e) for the water dimer geometry given in Table 2.2 under various combinations of theory, semiempirical method, and population analysis scheme. Calculations are performed without exchange-correlation and dispersion contributions for X-Pol results. “X” denotes the use of X-Pol and “M” and “D” denote the use of Mulliken and DPPC charges respectively. 52

2.4 Analytical first derivatives (in kcal mol^{-1} Å^{-1}) of full semiempirical calculations, X-Pol calculations, and X-Pol calculations with DPPC population analysis for the water dimer geometry given in Table 2.2. The results in this table derive from the use of Eq. 2.30, without exchange-correlation and dispersion contributions for X-Pol results. The dashed line in the table indicates the separation of the fragments for X-Pol calculations. “X” denotes the use of X-Pol and “M” and “D” denote the use of Mulliken and DPPC charges respectively. 53

2.5 Cartesian coordinates for the hydrogen fluoride used in our tests of various combinations of theory, semiempirical method, and population analysis scheme. The dashed line in the table indicates the separation of the fragments for X-Pol calculations. 53

2.6 Heat of formation/X-Pol energy (kcal/mol), dipole moment (Debye), and charges (e) for the hydrogen fluoride dimer geometry given in Table 2.5 under various combinations of theory, semiempirical method, and population analysis scheme. Calculations are performed without exchange-correlation and dispersion contributions for X-Pol results. “X” denotes the use of X-Pol and “M” and “D” denote the use of Mulliken and DPPC charges respectively. 54

2.7 Analytical first derivatives (in kcal mol^{-1} Å^{-1}) of full semiempirical calculations, X-Pol calculations, and X-Pol calculations with DPPC population analysis for the hydrogen fluoride dimer geometry given in Table 2.5. The results in this table derive from the use of Eq. 2.30, without exchange-correlation and dispersion contributions for X-Pol results. The dashed line in the table indicates the separation of the fragments for X-Pol calculations. "X" denotes the use of X-Pol and "M" and "D" denote the use of Mulliken and DPPC charges respectively. 54

3.1 Semiempirical parameters for H and O Atoms in the PMOw model. The derived parameter, h_{pp}, is determined from g_{pp} and $g_{pp'}$ and has been set to a minimum value of 0.1 eV as implemented in the MOPAC program, $h_{pp} = \max\{0.1\text{eV}, (g_{pp} - g_{pp'})/2\}$. 63

3.2 Additional semiempirical parameters for oxygen and hydrogen in the polarized molecular orbital model and the Lennard-Jones parameters in explicit polarization model for liquid water. 63

3.3 Computed equilibrium properties for water monomer and dimer from different polarizable water models and *ab initio* MP2/(CBS) with CCSD(T) corrections along with experimental data. 64

3.4 Computed and experimental properties for water clusters. The angle $\langle\phi\rangle$ is the average O. . .H–O angle of the hydrogen bonds in a given cluster. (a) Ref. [2]. (b) [3]. (c) [4]. (d) MP2/(CBS) limit [5]. 80

3.5 Computed interaction energies in kcal/mol for H$^+$(H$_2$O)$_n$ complexes from the PMOw and MP2 methods. Interaction energies are calculated by $\Delta E = E(\text{cluster}) - \left[E(\text{H}_5\text{O}_2^+) + nE(\text{H}_2\text{O})\right]$, where n is the number of water molecules. 84

3.6 Liquid properties of the XP3P model for water along with those from experiments, and the TIP3P, AMOEBA, and SWM4-NDP models. (a) Refs. [6,7]. (b) Ref. [2]. (c) Ref. [8]. (d) See text for details. (e) The average $E_i(l)$ from molecular dynamics simulations employing the variational Fock operator is -9.99 kcal/mol over 400 ps. This gives a heat of vaporization of 10.52 kcal/mol. 85

3.7 Computed average thermodynamic properties of water at different temperatures between -40 °C and 100 °C (values in parentheses for C_p and α are obtained from the direct derivative calculations). The corresponding units are kcal/mol (ΔH_{vap}), cal/(mol K) (C_p), g/cm^3 (ρ), K^{-1} ($10^5\alpha$), and atm^{-1} ($10^6\kappa$). 94

3.8 Temperature-dependent energy components (units in kcal/mol). 103

4.1 Parameters in the PMOw model. The parameters for F were obtained by stochastic optimization starting from the RM1 parameter for F using a fitness function related to experimental and *ab initio* calculated quantities of HF, $(HF)_2$, $(HF)_3$, $(HF)(H_2O)$, and OF_2. The parameters for H and O come from Chapter 3. 117

4.2 HF monomer properties of PMOw compared to *ab initio* and experimental results; (a): Ref. [9], (b): Ref. [10], (c): Ref. [11], (d): Ref. [12], (e): Ref. [13]. 119

4.3 Properties of HF cyclic clusters at the PMOw and M06-2X/MG3S levels with a combination of experimental and *ab initio* data that forms "best estimates" [14]. (a): Ref. [14]. 120

4.4 Properties of (HF)$_n$(H$_2$O)$_n$ complexes at the PMOw and MP2/TZP levels for $n = 1, 2, 4$. Partial charges were obtained from Mulliken and Löwdin population analysis for PMOw and MP2/TZP respectively. (a) Ref. [15]. 123

4.5 The state 11 points used in our liquid simulations of HF and their corresponding experimental densities. (a) Ref. [16], (b) Ref. [17], (c) Ref. [18], (d) Ref. [19], (e) Ref. [20]. 129

4.6 Energy per molecule of the 11 states in kJ/mol from XPHF compared to several other models and experiment. All values for the other models originate from Table III of Ref. [21], in which uncertainties are also listed. (a) From Visco-Kofke equation of state [22]; (b,c) See Ref. [21] for details; (d) Ref. [23]. 131

4.7 The average dipole moments of the 11 states in Debye for XPHF and JV-P along with the static values from JV-NP and the TIPS models found in Ref. [23]. 132

4.8 Densities of the 11 states in g/cm^3 from XPHF compared to several other models and experiment. Unless otherwise noted, all values for the other models originate from Table IV of Ref. [21], in which uncertainties are also listed. (a) Ref. [16]; (b) Ref. [17]; (c) Ref. [19]; (d) Ref. [20]; (e) Ref. [23].136

4.9 Diffusion coefficients in 10^{-9} m^2/s as predicted by XPHF at the 11 state points with comparison to experiment [24]. Experimental values for states A-H and K were determined from the Arrhenius equation $D_{\text{Expt.}} = D_0 \exp(-E_A/RT)$ where $E_A = 9.92$ kJ/mol and $D_0 = 452$ and $D_0 = 398$ 10^{-9}m^2/s for standard vapor pressure (A-H) and 500 Bar (K) respectively.141

5.1 The net charges of each amino acid residue used by MACROSHAKER in the creation of coarse-grained models. Charges are centered on the α-carbon of each residue and are summed to determine the charge on the bead that contains them. 157

5.2 An eight-site, coarse-grained model of myoglobin [25] using our grid-based method with a spacing of 16 Å. The Cartesian coordinates of each bead are given relative to the center of mass. The ϵ values for use in Eq. 5.22 were fit to reproduce the concentration-dependent diffusion by experiment from a scaling of σ by 0.0105 kcal/(mol Å). 175

5.3 The concentration-dependent diffusion coefficients of Mb, average Mb interaction energy, and the Mb volume fraction for each concentration. The Mb volume fraction was calculated using $N(4/3)\pi\sigma^3/V$ where $N = 216$ and $\sigma = 22.683$Å and V denotes the simulation box volume. 177

6.1 Optimized amplitudes (Å), frequencies (rad/ns), and phase (rad) for the time-dependent quasiharmonic mode projection along the molecular dynamics trajectory of the protein orotidine monophosphate decarboxylase as represented by Eq. 6.16. 196

6.2 Computed mode weighting factors to represent the overall complex protein fluctuations using the first ten lowest quasiharmonic modes 203

7.1 Binding energy in kcal/mol of various water hexamers for PMOw using full QM, X-Pol, VMB2, and VMB3. Mean unsigned deviations (MUD) and root-mean square deviations (RMSD) compared to the full PMOw result show that both VMB2 and VMB3 agree significantly better with the full energy than the X-Pol method alone. All calculations were done without Lennard-Jones terms and used Mulliken charges. 211

A.1 Table showing statistical data for charged sites using the coarse-graining of Byron with 36 sites and geodesically spaced ACG surface charges for 12 and 42 sites respectively. Unites are given in electron charge. 269

List of Figures

2.1 An illustration of the bonded terms in the CHARMM22 force field (Eq. 2.1) on a phenylalanine molecule. 11

2.2 A flow chart showing the steps of the self-consistent field procedure for the Hartree-Fock method. Iterations of Fock matrix diagonalization, density matrix calculation, and reconstruction of the Fock matrix are repeated until the maximum change in density between successive iterations is sufficiently small. 23

2.3 A flow chart showing the steps of the double self-consistent field procedure used in Hartree-Fock based variants of X-Pol. The box labeled "SCF for all fragments" denotes only the cyclic portion of the flow chart from Figure 2.2. 31

3.1 Illustration of the angle between the molecular dipole moment derivative and the O–H bond vector in water monomer. Experimental values are given first, followed by the PMOw results in parentheses. 75

3.2 Potential energy profiles for a water dimer at the hydrogen bonding configuration from the PMOw (black) and the XP3P (blue) models for water along with CCSD(T) results (red). Definition of the geometrical parameters listed in Table 3.3 are given in the structure shown as inset in the upper right-hand corner. The CCSD(T) results are obtained with the aug-cc-pVDZ basis set on fully optimized geometries at various fixed O–O distances. Studies have shown that extrapolation to the complete basis set limit from the current size does not affect the computed energies by more than 0.2 kcal/mol. [5] All other geometric parameters are optimized. 76

3.3 Optimized water clusters with PMOw 78

3.4 Optimized water clusters with XP3P 79

3.5 Optimized geometries of $H^+(H_2O)_n$ clusters from PMOw. 82

3.6 Potential energy profile for $H_5O_2^+$ in the gas phase as a function of the proton transfer coordinate , defined as the distance from the mid-point between the two oxygen atoms, (a) at the minimum geometry , (b) at a fixed O–O separation of Å, and (c) at a fixed O–O distance of Å from PMOw (black), and CCSD(T)/aug-cc-pVDZ (red) calculations. Geometries were optimized with fixed O–O distances. 83

3.7 Distribution of the scalar molecular dipole moment in liquid water from Monte Carlo simulations with the XP3P potential at 25°C and 1 atm. The units for the ordinate are mole percent per Debye. 87

3.8 Distribution of the binding energies of water in the liquid at temperatures ranging from -40°C to 100°C. The binding energy corresponds to the total interaction energy of one water with the rest of the bulk solvent. 89

3.9 Computed (black) and experimental (red, dashed) oxygen-oxygen (a), oxygen-hydrogen (b), and hydrogen-hydrogen (c) radial distribution functions of liquid water at 25°C and 1 atm. 93

3.10 Computed (black) and experimental (red) heats of vaporization for liquid water. The results from the TIP5P model are illustrated in green. . . 94

3.11 Computed (black) and experimental (red) densities for liquid water, along with those from the TIP3P (brown), the TIP4P (maroon), and the TIP5P (green) models. 95

3.12 Computed and experimental coefficients of thermal expansion (α) for liquid water. The α values are determined from numerical derivatives of liquid volume variations with temperature. 95

3.13 Computed heat capacities from the fluctuation formula and direct numerical derivatives from XP3P at temperatures ranging from -40 °C to 100 °C, compared to those from experiment. 96

3.14 Computed average molecular dipole moments for liquid water at different temperatures. 100

3.15 Average total interaction energies (black) per water in the liquid and their contributing components, including vertical interaction energies (blue), polarization energies (green), total electrostatic interaction energies (red), and exchange-dispersion correlation energies (magenta). . . . 103

4.1 Optimized HF dimers using several NDDO-type semiempirical methods, PMOw, DFT at the M06-2X/MG3S level, and that measured from experiment [13]. All non-PMOw semiempirical dimers, with the exception of the PM3 optimized geometry, exhibit a qualitatively incorrect structure compared to *ab initio* and experimental results. 114

4.2 Interaction energy profile of optimized HF dimers with respect to constrained F-F distance for several NDDO-type semiempirical methods, PMOw, XPHF, and DFT at the M06-2X/MG3S level. Notice the odd behavior of the interaction energy curve for PM3, which produced a qualitatively correct optimized HF dimer geometry. 115

4.3 Labeled quantities for the HF dimer corresponding to quantities given in Table 4.2. 118

4.4 Labeled quantities for the HF trimer corresponding to quantities given in Table 4.3. Entries in the table for clusters larger than the trimer have analogous quantities to the labels in the figure. 121

4.5 Cyclic clusters of HF at PMOw level (top) and the M06-2X/MG3S level (bottom). The various monomer colors indicate their respective cluster, and are as follows: dimer (red), trimer (blue), tetramer (yellow), pentamer (cyan), hexamer (magenta), heptamer (orange), octamer (black). The multicolored monomer at the bottom of each pane indicates the first monomer in each cluster. 122

4.6 Three $(HF)_n(H_2O)_n$ complexes: $(HF)(H_2O)$ (top), $(HF)_2(H_2O)_2$ (bottom left), $(HF)_4(H_2O)_4$ (bottom right). The $(HF)(H_2O)$ complex exhibits a hydrogen bond while the $(HF)_2(H_2O)_2$ and $(HF)_4(H_2O)_4$ complexes show an ionic bonding behavior. 124

4.7 The mole fraction of average dipole moments for each of the 11 states tested with XPHF. In supercritical states H and I the distribution appears to take a maximum value near the gas-phase dipole moment, while supercritical states J and K show a wider distribution across many values. 133

4.8 The temperature dependence of liquid HF density at atmospheric pressure from XPHF and other models at states C and D compared to the experimental values of Simons and Bouknight [19]. Notice that the polarizable models (XPHF, PJV-P, JV-P) predict densities in better qualitative and quantitative agreement with experiment than the non-polarizable models (JV-NP, HF-Kr, TIPS). 135

4.9 Total RDFs of the XPHF (red), JV-P (cyan), JV-NP (orange), and TIPS (blue) models with experimental data (black) [18] for the four supercritical states at 473 K. The TIPS model is over-structured for all tested supercritical states but J (473 K, 166 bar). XPHF gives results that appear slightly better than the JV-P model, but with more qualitatively-correct peaks. 137

4.10 A comparison of the partial RDFs of HF at state E (296 K, 1.2 bar), for XPHF (red), PJV-P (blue), BOMD at the BLYP-D2/TZV2P level (orange), and the experiment of McLain and co-workers (black) [26]. XPHF shows agreement with experiment that is comparable to the results of the PJV-P model. Note that the BOMD simulation was performed at a slightly different state of 300 K and 1.0 bar. 140

5.1 An artistic rendition of a cross-section of an *E. coli* cell clearly showing a crowded environment of macromolecules. Reproduced with the permission of D.S. Goodsell, Scripps Research Institute [27]. 145

5.2 The Cartesian coordinates of the center of mass of a myoglobin under Ermak-McCammon dynamics over a period of 1 μs. 152

5.3 The van der Waals surface of a deoxyhemoglobin S (PDB: 2HBS) (left; 8760 heavy atoms), the bounding volume and grid representation formed from using a spacing of $w = 15$ Å(center), and the result of contracting all atoms in a grid entry to a single sphere (right: 107 spheres) 156

5.4 The uniform grid representation with spacings $w =$ 5, 10, 15, 20, and 25, respectively (top), and the corresponding coarse-grained models with 1592, 269, 107, 37, and 21 sites, respectively (bottom). 159

5.5 The component of the MACROSHAKER GUI which selects the level of coarse-graining to be used. A PDB is first loaded (top) and the grid size is specified along with the number of copies to use in the simulation (bottom). 162

5.6 The component of the MACROSHAKER GUI in which the box size is entered and the initial coordinates are randomized; equilibrium steps can be run to reduce steric clashes and visualized in realtime in this part of the GUI. 163

5.7 The visualization component of the MACROSHAKER GUI showing a box of several of the proteins in albumen. Trajectories can be played back within the interface or exported to video files. 164

5.8 The van der Waals surface of myoglobin and the corresponding one-to-one surface from a ray casting procedure. 167

5.9 Spherical harmonic surfaces of a Mb from $L = 0$ (upper-left) to $L = 19$ (lower-right). 169

5.10 Definition of vectors $\vec{V}$, $\vec{N}$, and $\vec{L}$ used in shading (left), and the interpolated normals $\vec{N_i}$ (right). 171

5.11 Gouraud shading for various representations of Mb (top), and cel-shading of the same representations 173

5.12 A visualization of the enzymes involved in glycolysis using a hybrid Phong and cel-shading technique (left) and an artistic rendition of a cross-section of an *E. coli* cell. [27] 173

5.13 Results of MACROSHAKER on a system of 216 myoglobin (Mb) compared to the experimental data of Wittenberg *et al.* [28] and Longeville *et al.* [29] 176

5.14 Radial distribution functions of the 216 myoglobin (Mb) from simulations of the MACROSHAKER CGFF at concentrations ranging from 2 mM to 32 mM. As the concentration increases, the peak positions become shorter and the peak heights become higher. A notable exception from this trend is the slightly shorter peak height at 32 mM, for which the volume of the Stokes' spheres is approximately 94% of the simulation box. 178

6.1 Histogram of the computed projection of instantaneous molecular structure of OMPDC in water onto normalized eigenvector directions of the lowest quansiharmonic mode (black), the second (red), the fourth (purple), the tenth (green), the fiftieth (blue), the one hundredth (orange), and the one thousandth (cyan) modes. 193

6.2 Sinusoidal fit of harmonic frequencies and phases (with respect to the structure used at the start of the molecular dynamic simulation of OMPDC) to the oscillatory structural projections illustrated in Figure 6.1 for modes number 1, 2, 4, and 10. 195

6.3 Spherical harmonic reconstruction of the Lee and Richards surface for the trimer complex of the capsid protein of turnip yellow mosaic virus using representation degrees of L = 3, 4, 5, 10, 12, 15, 20, 25, and 30 numbered from top left to bottom right. 200

6.4 Sinusoidal fit of harmonic frequencies and phase (with respect to the structure used at the start of the molecular dynamic simulation of OMPDC) to the oscillatory structural projections illustrated in Figure 6.1 for modes number 5 (a) and 6 (b). Note that if the trajectory of the first 1 ns is discarded in mode 6 evaluation, the frequency and amplitude are both reasonable, suggesting there is either a conformational jump or change in the first 1 ns. 202

6.5 Snapshots of three structures of the analytically coarse-grained (ACG) protein OMPDC using spherical harmonic basis at a representation degree of $L = 15$ from the 0.25 μs composite fluctuation trajectory using the amplitudes, frequencies, and phases listed in Tables 6.1 and 6.2. The three structures on the right-hand side are the same as the corresponding ones on the left, rotated by 180°. The ACG protein surfaces are colored by the surface charge density for the illustration with red representing negative and blue positive charge densities, respectively. 204

6.6 Histogram of the fluctuation of the excluded volume of the ACG OMPDC protein at various resolutions in L, ranging from 5 to 20 along the internal quasiharmonic fluctuation trajectory. The excluded volume illustrated in this figure is defined as the cavity enclosed by the Lee-Richards surface, which is about one solvent layer larger than the van der Waals surface. 205

7.1 Screenshots from a program showing the virus assembly model of Wales. [30] 214

7.2 A visualization of the 10 enzymes involved in glycolysis as rendered by MACROSHAKER. 215

A.1 The two methods of coarse-graining used for charge fitting. Left to right: Atomistic model, coarse-graining of Byron, ACG model. 268

Preface

This dissertation contains previously published work and is reproduced with permission from the American Institute of Physics and the American Chemical Society.

Chapter 3

"Quantum mechanical force field for water with explicit electronic polarization", J. Han, M. J. M. Mazack, P. Zhang, D. G. Truhlar, J. Gao, *Journal of Chemical Physics*, **2013**, *139*, 054503, http://dx.doi.org/10.1063/1.4816280. ©2013 American Institute of Physics

Chapter 6 & Appendix B

"Internal dynamics of an analytically coarse-grained protein", M. J. M. Mazack, A. Cembran, J. Gao, *Journal of Chemical Theory and Computation*, **2010**, *6*(11), 3601-3612, http://dx.doi.org/10.1021/ct100426m. ©2010 American Chemical Society

Chapter 1

Introduction

Since the first *in silico* study of chemical systems [31] on the primitive MANIAC computer at Los Alamos National Laboratory in the wake of the Manhattan Project and the Second World War, great leaps and bounds have been made in computer simulations of chemical and biophysical systems. At the core of these developments are improved representations for the potential energy surface, or force fields, of chemical systems, which describe the intra- and inter-molecular interactions. The accuracy of these force fields is what ultimately determines the reliability and predictive power of computational studies of biological systems.

This dissertation is divided into two parts. The first part covers the description of an electronic structure method for an entire chemical system with periodic boundary conditions, such that electronic polarization and charge transfer can be explicitly included into the force field. We call this approach the explicit polarization theory (X-Pol) [32,33], and use it as a quantum mechanical force field (QMFF). In the second part, we extend the *in silico* treatment of biological systems to a much greater realm, with the capability to model a section of a biological cell by developing a coarse-grained force field (CGFF), and other simulation techniques, to model crowded systems of many proteins.

X-Pol moves beyond the current, widely used molecular mechanical approximation of condensed-phase systems into a more realistic representation using quantum mechanics, and signals a paradigm shift in the treatment of biological macromolecules, potentially increasing the accuracy of computer simulations. In contrast, traditional *in silico* treatments of chemical and biological systems in the condensed phase have employed empirical force fields, which consist of spring-like potentials to mimic the behavior of the chemical bond, and Coulomb's law with nucleus-centered, fixed, point charges and Lennard-Jones potential to mimic nonbonded interactions [34]. Although such molecular mechanics (MM) models continue to be the *de facto* standard for molecular dynamics (MD) simulations, and are often adequate for many systems, in other cases – such as liquid hydrogen fluoride, which is highly polar, and characterized by its ability to form chains of hydrogen bonds – molecular mechanics is inadequate to model the highly covalent nature of intermolecular interactions. Similar observations for other polar substances, including the peptide bond and many amino acid sidechains, have given rise to interest in the development of *polarizable force fields*, where the local dipoles induced by the surrounding electric field directly affect the potential felt between atoms.

One approach to including polarization effects in a force field is through the application of the variationally optimized X-Pol theory [35]. X-Pol is a fragment-based QM/QM-type method where the bonded interactions within each fragment are fully described by a chosen level of electronic structure theory, and the nonbonded interactions are described by Coulomb integrals, neglecting short-range exchange repulsion. The energy expressions and analytical first derivative for the X-Pol QMFF used in this dissertation are provided in Chapter 2. Chapters 3 and 4 illustrate two specific applications to show the performance of X-Pol when using the semiempirical polarized molecular orbital (PMO) method and a dipole-preserving, polarization consistent

(DPPC) charge method applied to water (XP3P) and hydrogen fluoride (XPHF).

Although our QMFF has been parallelized, the problem of scalability to systems consisting of several million atoms has not yet been addressed. Others have successfully used MM in such cases, and a recent simulation of 64 million atoms, the largest published simulation to date, was accurate enough to predict the structure of the HIV virus capsid [36]. Despite the fact that 64 million atoms is a staggering number, which is orders of magnitude larger than the largest MM simulations of just a decade ago [37,38], it is still dwarfed by the number of atoms inside a single living cell.

It has been estimated that a single *E. coli* cell contains carbon atoms on the order of 10 billion [39]. Such an estimate originates from the observation that nearly half the dry mass of *E. coli* is carbon, suggesting that about one trillion (10^{12}) atoms are needed to model a living cell with an all atom simulation. In addition to the vast number of atoms required to simulate a single living cell and the increased spatial scale that it brings, an increased temporal scale is also required to relate simulation results to relevant biological phenomena, observed experimentally, which can occur in the second or longer time regime. This translates to a simulation length on the order of one quadrillion (10^{15}) simulation steps with standard all-atom MM in a molecular dynamics simulation.

One way these increased spatial and temporal scale requirements have been addressed is through the use of coarse-grained force fields (CGFF), which allow for an increased time step by decreasing the number of degrees of freedom. Historically, these types of models have been based upon introducing effective sites for groups of atoms. The first such model used in a simulation of proteins was that of Levitt and Warshel in 1975 for modeling the folding of the bovine pancreatic trypsin inhibitor (BPTI), using two coarse-grained sites per amino acid residue [40]. Tanaka and Scheraga introduced their own residue-based, coarse-grained model for protein folding the following year [41]. Other notable work in this area includes the efforts of Smit [42] and Voth [43], as well as the MARTINI [44, 45] CGFF proposed by Marrink *et al.* for modeling proteins and lipid bilayers.

In light of the ambitious challenge of cell simulation, we introduce a CGFF called MACROSHAKER for which coarse-grained macromolecules are treated as rigid bodies under the Ermak-McCammon dynamics scheme [46]. The details of this model and a case study applied to studying the diffusion of myoglobin are given in Chapter 5. Expanding on the idea of coarse-graining to reduce the degrees of freedom in a system, Chapter 6 provides an atomless description of a protein where its surface and dynamical fluctuations are mapped into oscillating spherical harmonic coefficients.

1.1 Novelty of Results

The novelty of this dissertation is briefly summarized in the bulleted list below.

- The first detailed description of the energy, energy gradient, integral, and other terms needed to implement the PMO/X-Pol/DPPC QMFF is provided in Chapter 2.

- The first stand-alone library for X-Pol has been developed using the formulas in Chapter 2. A web-based interface for X-Pol using this library has been introduced.

- The first studies of the PMO/X-Pol/DPPC QMFF are presented in Chapters 3 and 4. Although quantitative studies using Monte Carlo have been carried out with X-Pol, the first quantitative studies of MD simulations using the X-Pol potential are presented in Chapters 3 and 4.

- A new CGFF and software package called MACROSHAKER for protein diffusion in crowded environments is introduced in Chapter 5. This CGFF can reproduce concentration-dependent diffusion coefficients of myoglobin for volume fractions as high as 40%.

- A spherical harmonic expansion for a protein surface that fluctuates with time is introduced in Chapter 6. In contrast to previous work in this area, the functional form is still a single spherical harmonic expansion.

1.2 Description of Chapters

More detailed descriptions for each of the subsequent chapters in this dissertation are provided in the following abstracts:

1.2.1 The Explicit Polarization Theory and its Analytical First Derivative

Chapter 2: The variational explicit polarization theory (X-Pol) provides a framework for the development of a quantum mechanical force fields (QMFF) that goes beyond the current molecular mechanics approximation. X-Pol is based on a hierarchy of approximations to increase the computational efficiency of electronic structure calculations, and strives for accuracy by parameterization using an effective Hamiltonian. In this chapter, the theory and implementation of X-Pol as a QMFF and its analytical first derivative are described. In particular, our focus is on the use of neglect of diatomic differential overlap (NDDO)-type semiemprical methods, including our recently-introduced polarized molecular orbital (PMO) approach. In addition, we describe the application of the dipole-preserving, polarization consistent (DPPC) charges on calculating the X-Pol energy and derive an analytical expression for its first derivative. The theoretical method described in this chapter is the basis for developing the

XP3P model for liquid water and the XPHF model for liquid hydrogen fluoride in subsequent chapters. The analytical first derivatives have been implemented into modified versions of CHARMM and NAMD, as well as a library and stand-alone program with an optional web-based interface.

1.2.2 Quantum Mechanical Force Field for Water

Chapter 3[1] : This chapter describes a quantum mechanical force field (QMFF) for water. Unlike traditional approaches that use quantum mechanical results and experimental data to parameterize empirical potential energy functions, the present QMFF uses a quantum mechanical framework to represent intramolecular and intermolecular interactions in an entire condensed-phase system. In particular, the internal energy terms used in molecular mechanics are replaced by a quantum mechanical formalism that naturally includes electronic polarization due to intermolecular interactions and its effects on the force constants of the intramolecular force field. As a quantum mechanical force field, both intermolecular interactions and the Hamiltonian describing the individual molecular fragments can be parameterized to strive for accuracy and computational efficiency. In this chapter, we introduce a polarizable molecular orbital model Hamiltonian for water and for oxygen- and hydrogen-containing compounds, whereas the electrostatic potential responsible for intermolecular interactions in the liquid and in solution is modeled by a three-point charge representation that realistically reproduces the total molecular dipole moment and the local hybridization contributions. The present QMFF for water, which is called the XP3P (explicit polarization with three-point-charge potential) model, is suitable for modeling both gas-phase clusters and liquid water. The chapter demonstrates the performance of the XP3P model for water and proton clusters and the properties of the pure liquid from about 900×10^6 self-consistent-field calculations on a periodic system consisting of 267

[1] This chapter is a result of collaborative efforts between the author, J. Han, and P. Zhang.

water molecules. The unusual dipole derivative behavior of water, which is incorrectly modeled in molecular mechanics, is naturally reproduced as a result of an electronic structural treatment of chemical bonding by XP3P. We anticipate that the XP3P model will be useful for studying proton transport in solution and solid phases as well as across biological ion channels through membranes.

1.2.3 Quantum Mechanical Force Field for Hydrogen Fluoride

Chapter 4: The X-Pol quantum mechanical force field (QMFF) is extended to liquid hydrogen fluoride (HF). The parameterization, called XPHF, is built upon the formalism introduced for the XP3P model of liquid water, based on the polarized molecular orbital (PMO) semiempirical quantum chemistry method, and the dipole-preserving polarization consistent (DPPC) point charge model. We introduce a fluorine parameter set for PMO, and find good agreement for various gas-phase results of small HF clusters compared to experiment and *ab initio* calculations at the M06-2X/MG3S level of theory. The XPHF model employs the newly introduced fluorine parameter set and shows good agreement with experiment for a variety of structural and thermodynamic properties in the liquid state, including radial distribution functions, interaction energy, diffusion coefficients, and densities at various state points.

1.2.4 MACROSHAKER: A Coarse-Grained Force Field for Crowded Systems of Many Proteins

Chapter 5: The ultimate goal of biophysical and biochemical studies is the understanding of how living organisms function, in which macromolecular particles, including proteins, nucleic acids, as well as ions and metabolites are packed in an extremely crowded environment. However, the vast majority of computational studies concerning enzymes, proteins, RNAs, etc. are conducted in idealized conditions rather than in the living cell itself. These studies largely ignore the systematic effects of protein crowding on the diffusion of enzymes and metabolites as well as protein-protein associations in signal transduction pathways, casting doubts on the relevance of extrapolating results directly to living cells. Our goal is to incorporate these effects, among others, by the creation of a computational model, called MACROSHAKER, to realistically model biomolecular processes that are directly comparable to that *in vivo* (i.e. *in life*). We seek not only to validate experimental results with our model, but also to develop a graphical user interface for the construction of initial conditions, the analysis of dynamic trajectories, and interactive visualization.

1.2.5 Internal Dynamics of an Analytically Coarse-Grained Protein

Chapter 6: An analytically coarse-grained model (ACG) is introduced to represent individual macromolecules for the simulation of dynamic processes in cells. In the ACG model, a macromolecular structure is treated as a fully coarse-grained entity with a uniform mass density without the explicit atomic details. The excluded volume and surface of the ACG macromolecular species are explicitly treated by a spherical harmonic representation in the present study (although ellipsoidal, solid, and radial augmented functions can be used), which can provide any desired accuracy and detail depending on the problem of interest. The present chapter focuses on the description of the internal fluctuations of a single ACG macromolecule, modeled by the superposition of low frequency quasiharmonic modes from explicit molecular dynamics simulation. A procedure for estimating the amplitudes, time scales of the quasiharmonic motions, and the corresponding phases is presented and used to synthesize the complex motion. The analytical description and numerical algorithm can provide an adequate representation of the internal protein fluctuations revealed from the corresponding atomistic simulations, although the internal motions of ACG macromolecules do not explore motions not exhibited in the dynamic simulations.

1.2.6 Conclusion & Discussion

Chapter 7: Some of the author's plans for future work on PMO and X-Pol are described with some preliminary results. These descriptions include the variational many-body expansion (VMB), improvements to PMO's core-core term, and discussion about new parameters for PMOw. Additionally, plans for future work on MACROSHAKER are discussed.

1.2.7 A Charge-Fitting Procedure for Coarse-Grained Proteins

Appendix A: We present a procedure for fitting of effective charges on interaction sites of proteins in ionic solution for use in the modeling of many protein simulations. Our model determines effective point-charges for a Debye-Hückel type of ionic screening potential from numerical solutions of the Poisson-Boltzmann equation. We show that this procedure works well when charges are assigned to positions on the protein surface as determined by ACG rather than sites determined from traditional coarse-graining methods. Our procedure can be used to fit any number surface charges, or can even be used to fit a charge density in the form of a spherical harmonic expansion. We present the origin of the Debye-Hückel equation and the charge-fitting procedure in detail.

1.2.8 Algorithm for Spherical Harmonic Expansion and Evaluation

Appendix B: An algorithm for fast computation of the spherical harmonic expansion for use with the techniques described in Chapter 6 is given and explained.

Chapter 2

The Explicit Polarization Theory and its Analytical First Derivative

2.1 Introduction

Because of its computational efficiency, molecular mechanics (MM), also known as force fields, have traditionally been used in molecular dynamics (MD) simulations of macromolecular systems. Eq 2.1 shows a representative functional form as adopted in the CHARMM22 force field (Figure 2.1).

$$\begin{aligned} E_{\text{CHARMM22}} = \sum_{\text{Bonds}} k_b(b-b_0)^2 + \sum_{\text{Angles}} k_\theta(\theta-\theta_0)^2 + \sum_{\text{Dihedrals}} k_\phi\left[1+\cos\left(n\phi-\delta\right)\right] \\ + \sum_{\text{Impropers}} k_\omega(\omega-\omega_0)^2 + \sum_{\text{Urey-Bradley}} k_u(u-u_0)^2 \\ + \sum_{\text{Nonbonded}} \left[\frac{q_i q_j}{\epsilon_l r_{ij}} + \epsilon\left[\left(\frac{R_{\text{min}_{ij}}}{r_{ij}}\right)^{12} - \left(\frac{R_{\text{min}_{ij}}}{r_{ij}}\right)^{6}\right]\right] \end{aligned} \tag{2.1}$$

It is assumed in these types of classical force fields that chemical bonding (b, θ, ϕ, ω, u) is adequately described by harmonic potentials about a set of equilibrium values (b_0, θ_0,

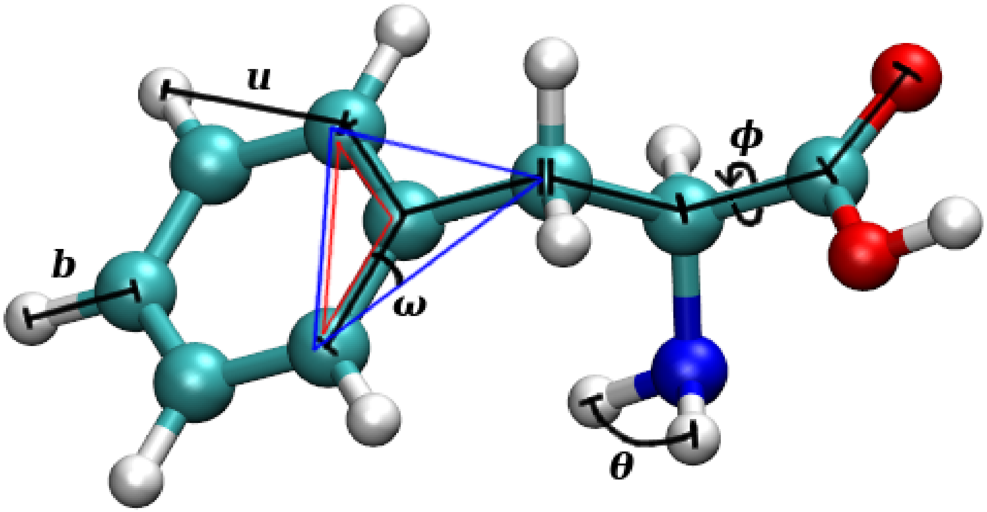


Figure 2.1: An illustration of the bonded terms in the CHARMM22 force field (Eq. 2.1) on a phenylalanine molecule.

ϕ_0, ω_0, u_0), and that nonbonded interactions can be described through a 12-6 Lennard-Jones potential and Coulomb's law of electrically charged particles. Such a simple, empirical description is in stark contrast to the most accurate and rigorous treatment of molecular systems by electronic structure theory using quantum mechanics (QM), in which the electronic wave function is obtained from solutions to a partial differential equation called the Schrödinger equation. However, electronic structure methods are extremely expensive compared to MM, and MD simulations using full QM to describe all interactions are intractable for all but small systems with relatively small basis sets since approximate methods are required to solve the QM problem.

Consequently, there has been much interest in the development of hybrid QM/MM methods which seek to combine both the accuracy of QM and the speed of MM. Although very useful, traditional QM/MM approaches suffer from the ambiguity in partitioning a system into QM and MM spatial regions, and do not explicitly incorporate

QM effects into all atoms.

The explicit polarization (X-Pol) theory was designed to provide a QM/QM-type force field through the decomposition of a molecular system into smaller pieces, called *fragments*. In X-Pol, the internal motions of atoms are described by full QM on a system whose nonbonded interactions are modeled through the polarization of the wave functions of constituent fragments. X-Pol was introduced in a non-variationally optimized form by Gao in 1997 under the name "molecular orbital derived empirical potential for liquid simulations" (MODEL) [32], and was used the following year in Monte Carlo simulations of liquid water using the semiempirical AM1 Hamiltonian [47]. In 2003 it was used again with AM1 for liquid simulations of supercritical hydrogen fluoride [48], and in 2008, a variationally optimized version of X-Pol was introduced by Gao, Truhlar, and co-workers with a fully analytical first derivative for use as a quantum mechanical force field (QMFF) [35].

The first published MD simulation that used X-Pol appeared in 2009 [49], and produced a 50 ps trajectory of a fully solvated bovine pancreatic trypsin inhibitor (BPTI) protein, consisting of 14,281 atoms divided into 4519 fragments, of which 4461 were water molecules and 58 were amino acid residues. This simulation generated 10 ps of simulation time in 75 hours on a single 2.66-GHz processor, demonstrating the feasibility of using X-Pol with a semiempirical Hamiltonian as a type of QMFF.

The X-Pol QMFF for MD simulations has been implemented into modified versions of CHARMM [50] and NAMD [51], incorporating recent improvements, including the semiempirical polarized molecular orbital (PMO) model [52] and the dipole-preserving point charge (DPPC) model [53] with analytical gradients.

The purpose of this chapter is to provide the theoretical background for understanding and programming the X-Pol method for use as a QMFF in MD simulations. We begin with a description of Hartree-Fock theory, proceeding to X-Pol, and arriving at the analytical gradients needed for the simulating water with the XP3P model and hydrogen fluoride with the XPHF model. We also clarify and make note of some of the more subtle details of the X-Pol method, which have not been previously discussed in the literature.

2.2 Hartree-Fock Theory

Hartree-Fock theory [54, 55] is one of the most successful and widely-used *ab inito* (i.e. *first principles*) quantum chemistry methods for solving the electronic Schrödinger equation. Although Hartree-Fock theory is an approximate electronic structure method by its use of a single Slater determinant and a mean-field treatment of electron-electron repulsion, it is of critical importance to modern quantum chemistry due to its theoretical simplicity and computational efficiency. Nevertheless, the computational cost of Hartree-Fock theory scales as $\mathcal{O}(m^4)$ where m is the number of basis functions. A number of increasingly accurate, post-Hartree-Fock methods have been developed for which Hartree-Fock theory serves as the foundation, including the popular MP2 [56] and CCSD(T) [57] methods, which include electron correlation effects. However, these methods scale at $\mathcal{O}(m^5)$ and $\mathcal{O}(m^7)$, respectively. While the success and theory of these methods is duly noted elsewhere [58], they are beyond the scope of our introductory description here.

Rather, we provide an introduction to Hartree-Fock theory as a motivation for its use as a QMFF under X-Pol. For the sake of brevity, our discussion assumes the use of restricted Hartree-Fock theory (RHF), where each orbital is occupied by a pair of electrons of opposite spin. Although less restrictive types of Hartree-Fock theory exist [59, 60], such an assumption is reasonable for a wide range of applications, including MD simulations of proteins solvated in water, for which an X-Pol-based QMFF has been successfully demonstrated [49].

2.2.1 Born-Oppenheimer Approximation

The behavior and interaction of electrons and nuclei are rigorously described by solutions to the Schrödinger equation [61]. Quantum chemistry calculations often employ a non-relativistic, time-independent molecular Hamiltonian operator to describe the stationary states of molecular systems. Under this approach, the Schrödinger equation can be written as Eq. 2.2 where $\mathcal{H}$ denotes the molecular Hamiltonian, Ψ denotes the wave function, and E denotes the associated energy.

$$\mathcal{H}|\Psi\rangle = E|\Psi\rangle \tag{2.2}$$

$\mathcal{H}$ is a Hermitian, linear operator, and is written in terms of electronic and nuclear terms with a coupling interaction. The description of $\mathcal{H}$ is of critical importance to the level of accuracy and computational ease of calculating the wave function Ψ and its associated energy E. Eq. 2.3 gives one such description, in atomic units, where N denotes the number of electrons and M, Z_A, and M_A denote the number of nuclei, the charge of each nucleus, and the mass of each nucleus respectively.

$$\mathcal{H} = \left[-\sum_{i=1}^{N} \frac{1}{2}\nabla_i^2 - \sum_{i=1}^{N}\sum_{A=1}^{M} \frac{Z_A}{r_{iA}} + \sum_{i=1}^{N}\sum_{j>i}^{N} \frac{1}{r_{ij}} \right] + \left[-\sum_{A=1}^{M} \frac{1}{2M_A}\nabla_A^2 + \sum_{A=1}^{M}\sum_{B>A}^{M} \frac{Z_A Z_B}{R_{AB}} \right] \tag{2.3}$$

The individual terms of the molecular Hamiltonian have specific physical descriptions. The three terms in the first set of brackets are, respectively, the kinetic energy of the electrons, the electron-nucleus attraction energy, and the electron-electron repulsion, where r_{iA} denotes the electron-nucleus distance and r_{ij} denotes the electron-electron distance. Similarly, the two terms in the last set of brackets are regarded as the kinetic energy of the nuclei and the nucleus-nucleus repulsion, respectively, where R_{AB} denotes the nucleus-nucleus distance.

The total mass of subatomic particles in the nucleus is orders of magnitude greater than the mass of the electrons [1], causing the nuclear motion in the stationary state to appear relatively small compared to the electronic motion. This observation leads to a simplification of the Hamiltonian known as the Born-Oppenheimer approximation [62], where nuclear positions are held constant when solving the Schrödinger equation. The result is a separation of variables between the electronic and nuclear coordinates, for which the electronic component is solved using the Schrödinger equation, and the nuclear degrees of freedom are held as external parameters; their motions can be solved either classically or quantum mechanically, albeit in a separate manner not depending on varying electronic coordinates.

Such treatment leads to the Hamiltonian operator used in Hartree-Fock theory (Eq. 2.4), in which the only variables are electronic coordinates.

$$\mathcal{H}_{\text{Elec.}} = \left[-\sum_{i=1}^{N} \frac{1}{2}\nabla_i^2 - \sum_{i=1}^{N}\sum_{A=1}^{M} \frac{Z_A}{r_{iA}} + \sum_{i=1}^{N}\sum_{j>i}^{N} \frac{1}{r_{ij}} \right] \tag{2.4}$$

Despite the assumption that the nuclear coordinates remain fixed, the nucleus-nucleus repulsion potential is still present under the Born-Oppenheimer approximation since it only depends on the position of the nuclei and not their motion. However, this term can be added to the total electronic energy after the Schrödinger equation has been solved, since Ψ does not explicitly depend on the nucleus-nucleus distance R_{AB}. The total energy of a chemical system under the Born-Oppenheimer approximation is thus given by Eq. 2.5 where $E_{\text{Elec.}}$ is called the electronic energy.

$$E_{\text{HF}} = E_{\text{Elec.}} + V_{\text{nuc}} \quad , \quad E_{\text{Elec.}} = \langle \Psi | \mathcal{H}_{\text{Elec.}} | \Psi \rangle \quad , \quad V_{\text{nuc}} = \sum_{A=1}^{M} \sum_{B>A}^{M} \frac{Z_A Z_B}{R_{AB}} \tag{2.5}$$

2.2.2 Wave Function Description

Exact, closed-form solutions to the Schödinger equation are in general not known for molecular systems of more than one electron, requiring any solution of the wave function Ψ and its associated electronic energy E to be approximations. The molecular wave function Ψ is approximated in Hartree-Fock theory as a combination of one-electron orbitals, where the solution is exact in that context, such as the dihydrogen cation H_2^+ [63]. The wave functions in these non-interacting systems are refered to as *molecular orbitals* (MOs), which we denote by ϕ_i. An ansatz for how the MOs may be combined into the total wave function Ψ is the Hartree product (Eq. 2.6).

$$\Psi(x_1, x_2, \cdots, x_N) = \prod_{i=1}^{N} \phi_i(x_i) \tag{2.6}$$

It is a trivial exercise to show that the Hartree product indeed satisfies $\mathcal{H}|\Psi\rangle = E|\Psi\rangle$ for $\mathcal{H}$ of Eq. 2.4 under the assumption that $\mathcal{H}_i|\phi_i(x_i)\rangle = E_i|\phi_i(x_i)\rangle$.

The description of Ψ as a Hartree product has its origins in methods for solving separable partial differential equations, and seems to be a reasonable first guess. Despite this reasoning, it is known that in addition to the mathematical constraints imposed on Ψ by the characteristic equation, there is also a physical constraint on Ψ called the anti-symmetry requirement, or Pauli exclusion principle [64] (Eq. 2.7).

$$\Psi(\cdots, x_i, \cdots, x_j, \cdots) = -\Psi(\cdots, x_j, \cdots, x_i, \cdots) \tag{2.7}$$

This requirement states that the interchange of elections (or MOs) must change the sign of the wave function, which is clearly not true for the Hartree product description of Ψ. However, the anti-symmetry requirement can be satisfied using a linear combination of Hartree products of MOs, called a Slater determinant [65] (Eq. 2.8).

$$\Psi = \frac{1}{\sqrt{N!}} \begin{vmatrix} \phi_1(x_1) & \phi_2(x_1) & \cdots & \phi_N(x_1) \\ \phi_1(x_2) & \phi_2(x_2) & \cdots & \phi_N(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_1(x_N) & \phi_2(x_N) & \cdots & \phi_N(x_N) \end{vmatrix} \tag{2.8}$$

The prefactor of $1/\sqrt{N!}$ in Eq. 2.8 normalizes $\Psi^*\Psi = |\Psi|^2$, which for the electronic Schrödinger equation is a probability density function for finding an electron in Ψ in any arbitrary region of space.

In practice, the individual MOs ϕ_i of the molecular system (i.e. Slater determinant) are approximated by a linear combination of atomic orbitals (LCAO), [66] refered to as a basis set $\{\chi\}$, often centered on each nucleus (Eq. 2.9).

$$\phi_i = \sum_{j=1}^{m} C_j^i \chi_j^i \tag{2.9}$$

For reasons of computational tractability, the basis set is taken as a finite set of m functions, although, in general, the basis set can be of infinite dimension in order to span the entire space of possible ϕ_i. In the case of an infinitely large basis set, the result is called the *Hartree-Fock limit*.

Since an analytical solution to the Schrödinger equation for more than one electron is unavailable, the precise functional form of the ideal basis set is also unknown. Instead, basis sets of atomic orbitals are typically in the form of Slater-type orbitals (STOs), [67] of the same form derived from the exact solution of the one-electron hydrogen atom. For computational efficiency, a set of Gaussian-type functions (orbitals) (GTOs) [68], which have more appealing mathematical and computational properties over STOs, are typically used.

2.2.3 Electronic Integral Calculations

The energy derived from the electronic Schrödinger equation $E_{\text{Elec.}}$ is in general obtained by an inner product that results in integration. The Hartree-Fock electronic energy $E_{\text{Elec.}} = \langle \Psi | \mathcal{H}_{\text{Elec.}} | \Psi \rangle$ and our definitions of $\mathcal{H}$ and Ψ in Eqs. 2.4 and 2.8, respectively, suggest that the procedure for calculating the electronic energy requires the integration of several flavors and combinations of MOs. This is indeed the case, and there are two classes of electronic integrals in the Hartree-Fock method: one-electron integrals and two-electron integrals.

The one-electron integrals are those which involve only coordinates of one electron. These one-electron terms have an associated matrix representation $\mathbf{H}^{\text{core}}$ with elements given by Eq. 2.10 where $\phi_i(1)$ denotes that the integral is over the coordinates of one electron.

$$H_{\mu\nu}^{\text{core}} = \int \phi_\mu^*(1) \left[-\frac{1}{2}\nabla^2 + \sum_A \frac{Z_A}{r_{1A}} \right] \phi_\nu(1) d\tau_1 \tag{2.10}$$

There are two types of one-electron integrals; the kinetic energy integral from the

Laplacian operator and the electron-nucleus attraction integral. These *one-electron integrals* are pairwise in orbital indices, resulting in $\mathcal{O}(m^2)$ terms for both types of one-electron integrals.

The second class of integrals in Hartree-Fock theory are the *two-electron integrals*. These integrals arise from the electron-electron repulsion potential of the r_{ij}^{-1} term in the electronic Hamiltonian, and the series of Hartree products from the anti-symmetric Slater determinant. Eqs. 2.11 and 2.12 provide the general forms of the two-electron integrals in the Hartree-Fock method where $\phi_i(j)$ denotes the integration over the coordinates of electron j in orbital ϕ_i and the indices μ, ν, λ, and σ refer to the same orbitals between 2.11 and 2.12.

$$\langle\mu\nu|\lambda\sigma\rangle = \int\int \phi_\mu^*(1)\phi_\nu(1)(1/r_{12})\phi_\lambda^*(2)\phi_\sigma(2)d\tau_1 d\tau_2 \tag{2.11}$$

$$\langle\mu\lambda|\nu\sigma\rangle = \int\int \phi_\mu^*(1)\phi_\lambda(2)(1/r_{12})\phi_\nu^*(1)\phi_\sigma(2)d\tau_1 d\tau_2 \tag{2.12}$$

The two-electron integral of Eq. 2.11 is referred to as the *Coulomb integral* and the two-electron integral of Eq. 2.12 is refered to as the *exchange integral*. The Coulomb integrals appear as direct result of the r_{ij}^{-1} potential, and have a direct correspondence to classical electron-electron repulsion. On the other hand, the exchange term is entirely a quantum effect appearing from the linear combination of Hartree products obtained from the anti-symmetrized wave function that satisfies the Pauli exclusion principle by way of the Slater determinant.

Taking notice of the four orbital indices in the two-electron integrals, we see that there are $\mathcal{O}(m^4)$ terms of both two-electron integral types. The computation of these "four-center", two-electron integrals is the dominating cost in the Hartree-Fock method.

These two-electron integrals are typically computed by Fourier transform for a finite basis representation of $\phi_i(j)$ in the STO case [69], and by recurrence relations in the GTO case [70, 71]. When the basis set is real-valued, as it is for GTOs, an eight-fold symmetry exists allowing for a reduction by a factor of eight in the number of integral calculations required.

Although the two-electron integral terms introduce a dependence on more than one set of electronic coordinates, each orbital involved depends on at most one set of electronic coordinates within the integral. The result of this integration is therefore an average, effective electron-electron repulsion and is called the mean-field approximation. This is equivalent to saying that electronic motion in the Hartree-Fock method is *uncorrelated*. That is, with the exception of the Pauli exclusion principle where two particles of the same spin are forbidden from occupying the same MO due to the anti-symmetric Slater determinant representation of the wave function, Hartree-Fock theory does not consider that multiple electrons cannot simultaneously occupy the same space, and replaces the exact N-body electron-electron repulsion for each electron with a mean-field representation from the sum of $N-1$ two-body terms.

Similar to the one-electron integrals and their associated matrix representation $\mathbf{H}^{\text{core}}$, the two-electron integrals have a matrix representation $\mathbf{G}$ (Eq. 2.13),

$$G_{\mu\nu} = \sum_{\lambda\sigma} P_{\lambda\sigma} \left[\langle \mu\nu | \lambda\sigma \rangle - \frac{1}{2} \langle \mu\lambda | \nu\sigma \rangle \right] \tag{2.13}$$

where $\mathbf{P}$ is called the density matrix and is given by Eq. 2.14

$$P_{\lambda\sigma} = 2 \sum_{i=1}^{N/2} C^*_{\lambda i} C_{\sigma i} \tag{2.14}$$

where $N/2$ denotes the set of doubly-occupied orbitals only.

The sum of the one-electron matrix $\mathbf{H}^{\text{core}}$ and the two-electron matrix $\mathbf{G}$ is called the Fock matrix, and is discretized analog of the electronic Hamiltonian for the Schrödinger equation.

$$\mathbf{F} = \mathbf{H}^{\text{core}} + \mathbf{G} \tag{2.15}$$

2.2.4 Self-Consistent Field Procedure

The goal of Hartree-Fock theory is not just to find an electronic energy and a set of MO coefficients $C_{\mu i}$ such that the Schrödinger equation is solved, but rather to find those $C_{\mu i}$ that are variationally optimized such that the energy of Eq. 2.5 is at minimum. Thus, the partial derivative of the Hartree-Fock electronic energy with respect the MO coefficients is zero (i.e. $\partial E_{\text{Elec.}}/\partial C_{\mu i} = 0$) under the constraint that MOs remain orthonormal (i.e. $\int \phi_\mu^* \phi_\nu d\tau = \delta_{\mu\nu}$).

While the inner product $\langle \Psi | E_{\text{Elec.}} | \Psi \rangle$ simplifies to $E_{\text{Elec.}} \langle \Psi | \Psi \rangle = E_{\text{Elec.}}$ in the infinite dimensional MO space due to the orthonormality of the eigenfunctions of Hermetian operators, the discretization and approximation of the MOs into LCAOs with finite basis set $\{\chi\}$ does not necessarily preserve orthonomallity under the same inner product. The consequence of this discretization is an overlap matrix $\mathbf{S}$ whose elements are given in Eq. 2.16.

$$S_{\mu\nu} = \int \chi_\mu^*(\tau) \chi_\nu(\tau) d\tau = \langle \chi_\mu | \chi_\nu \rangle \tag{2.16}$$

Notice that construction of $\mathbf{S}$ is of computational complexity $\mathcal{O}(m^2)$ and depends only on the basis set $\{\chi\}$ from Eq. 2.9. Regardless of the basis set employed, $\mathbf{S}$ is a symmetric, positive-definite matrix.

The result of the constrained optimization of the electronic energy produces what is known as the Roothaan-Hall equations [66, 72], whose matrix form is given in Eq. 2.17, where ϵ is a diagonal matrix of orbital energies.

$$\mathbf{FC} = \mathbf{SC}\epsilon \tag{2.17}$$

By inspection one can immediately see that the Roothaan-Hall equations are a type of a generalized eigenvalue problem with generalized eigenvectors $\mathbf{C}$ and associated eigenvalues ϵ. Since the overlap matrix $\mathbf{S}$ is symmetric positive-definite, it is also invertible. This implies that the Roothaan-Hall equations can be reduced to a standard eigenvalue problem, and can be solved using a variety of numerical methods and existing codes [73–75].

Although finding the orbital coefficients $\mathbf{C}$ appears to be the cost of merely one generalized eigenvalue problem from Eq. 2.17, on closer inspection one sees that the Fock matrix $\mathbf{F}$ has an explicit dependence on $\mathbf{C}$ through the density matrix $\mathbf{P}$ by its appearance in the two-electron matrix $\mathbf{G}$. This second observation leads to what is known as the self-consistent field (SCF) procedure.

The SCF procedure is an iterative process whereby an initial guess of the density matrix $\mathbf{P}$ is used to construct a Fock matrix for calculating the coefficient matrix $\mathbf{C}$ from the Roothaan-Hall equations. A number of initial guesses can be made with the most common being the result from diagonalizing $\mathbf{H}^{\text{core}}$ or a diagonal matrix.

A flow chart of the Hartree-Fock algorithm is given in Figure 2.2 where the loop over the decision on the density convergence is a single SCF step. The SCF procedure is typically terminated after the maximum magnitude of the difference in the density matrix between successive iterations is less than an *a priori* convergence criterion, typically less than 10^{-5}.

In practice, the SCF procedure can become numerically unstable and difficult to

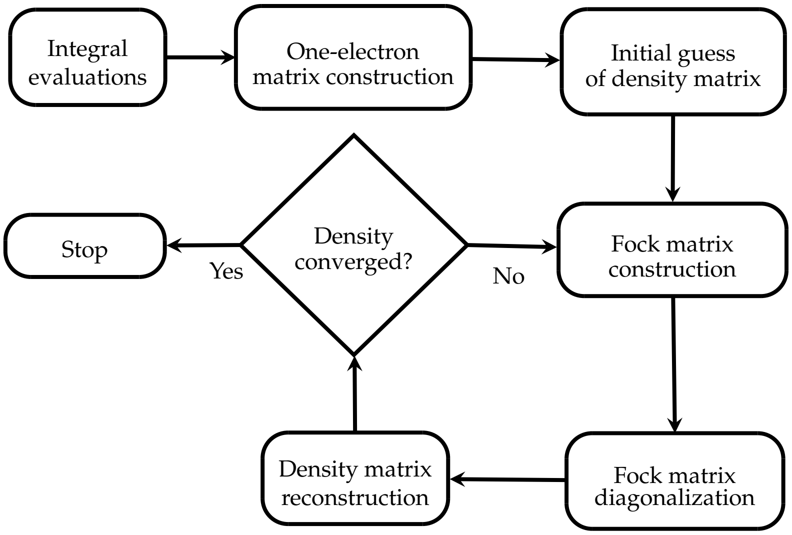

Integral evaluations
One-electron matrix construction
Initial guess of density matrix
Stop
Yes
Density converged?
No
Fock matrix construction
Density matrix reconstruction
Fock matrix diagonalization

Figure 2.2: A flow chart showing the steps of the self-consistent field procedure for the Hartree-Fock method. Iterations of Fock matrix diagonalization, density matrix calculation, and reconstruction of the Fock matrix are repeated until the maximum change in density between successive iterations is sufficiently small.

converge at times. Several methods that improve the stability of convergence have been developed. The most simple of these methods is called *damping*, where the density matrix and/or Fock matrix at each step of the SCF procedure is replaced with a weighted average of itself and that of the previous step. A more general approach of this same persuasion is the direct inversion of iterative subspace (DIIS) method [76] which can significantly improve convergence speed, but has higher storage requirements. A lucid description of several such methods to stabilize and accelerate convergence is available from Schlegel and McDouall [77].

Once the SCF procedure has converged, the density matrix can be analyzed and the electronic energy can be calculated. Although it may seem that the total electronic energy for the Hartree-Fock method should be a simple sum of the individual orbital energies of eigenvalues in ϵ, electron-electron repulsion terms are counted twice, and must be removed. The full, variationally optimized, Hartree-Fock energy under orthonormal constraints for the ground state wave function Ψ represented by a Slater determinant of MOs consisting of LCAOs can be calculated by applying Eq. 2.18.

$$E_{\text{HF}} = \sum_{\mu\nu} P_{\mu\nu} \left[H^{\text{core}}_{\mu\nu} + \frac{1}{2} \sum_{\lambda\sigma} P_{\lambda\sigma} \left[\langle \mu\nu | \lambda\sigma \rangle - \frac{1}{2} \langle \mu\lambda | \nu\sigma \rangle \right] \right] + V_{\text{nuc}}. \tag{2.18}$$

2.2.5 Analytical First Derivative of Energy

Up to this point we have only discussed Hartree-Fock theory for single-point energy calculations on a molecular system. While such an application should not be dismissed as unimportant or uninteresting, there are many problems that require not just the energy, but also the negative gradient of the energy with respect to the nuclear coordinates – the force.

Although the integral calculations, matrix constructions, and the SCF procedure may appear to be a complex and perhaps even unruly method, the analytical first derivative of the Hartree-Fock energy takes on a surprisingly elegant form. The partial derivative of the Hartree-Fock energy E_{HF} for a nuclear coordinate X on an atom A can be thought of as the sum of pieces $\hat{E}$ that depend explicitly on that coordinate, such as electronic integrals, and those which depend implicitly on it through the change of orbital coefficients (Eq. 2.19).

$$\frac{\partial E_{\text{HF}}}{\partial X_A} = \frac{\partial \hat{E}}{\partial X_A} + \sum_{\mu i} \frac{\partial E_{\text{HF}}}{\partial C_{\mu i}} \frac{\partial C_{\mu i}}{\partial X_A} \tag{2.19}$$

The variational optimization of the Hartree-Fock method necessarily implies that all terms $\partial E_{\text{HF}}/\partial C_{\mu i}$ = 0, causing Eq. 2.19 to reduce to Eq. 2.20.

$$\frac{\partial E_{\text{HF}}}{\partial X_A} = \frac{\partial \hat{E}}{\partial X_A} \tag{2.20}$$

This is to say that for the first derivative of the variationally optimized Hartree-Fock energy, there exists an analytical expression that contains only the derivatives of those terms which have explicit dependence on nuclear coordinates.

Through some tricks which are outlined elsewhere [58], the following expression for the analytical first derivative of the Hartree-Fock energy can be obtained (Eq. 2.21), where the matrix $\mathbf{W}$ is called the energy-weighted density matrix and has elements $W_{\mu\nu} = 2\sum_{i=1}^{N/2} \epsilon_i C^*_{\mu i} C_{\nu i}$ with $N/2$ having the same interpretation as it does for the density matrix (Eq. 2.14).

$$\begin{aligned} E_{\text{HF}}^{X_A} = \sum_{\mu\nu} P_{\mu\nu} \left[H_{\mu\nu}^{X_A} + \frac{1}{2} \sum_{\lambda\sigma} P_{\lambda\sigma} \left[\langle\mu\nu|\lambda\sigma\rangle - \frac{1}{2}\langle\mu\lambda|\nu\sigma\rangle \right]^{X_A} \right] \\ - \sum_{\mu\nu} W_{\mu\nu} S_{\mu\nu}^{X_A} + V_{\text{nuc}}^{X_A} \end{aligned} \tag{2.21}$$

Thus, it is sufficient to use only the derivatives of one-electron, two-electron, and overlap integrals in calculating the analytical first derivative of the Hartree-Fock energy. This derivative can be used for finding optimal molecular geometries that minimize total energy via standard optimization techniques such as the conjugate gradient method, or can be used for a host of other applications including molecular dynamics (MD) simulations.

2.3 Explicit Polarization Theory

Full Hartree-Fock treatments of molecular systems consisting of thousands atoms are rarely employed due to the high computational cost and numerical instability of the SCF procedure that becomes worse as the matrix size increases. A notable example of Hartree-Fock on a "large" system was the geometry optimization of a 642 atom and 42 residue protein called crambin by van Alsenoy and co-workers in 1998 [78]. Although the results of that study agreed well with the structure of crambin from X-ray crystalography [79], it is a sobering fact that they were not obtained without employing approximation of the two-electron integrals [80] and the use of the small 4-31G basis set [81].

As seen in the case of crambin, which would be widely-regarded as a very small system for the CHARMM22 force field, further approximations beyond Hartree-Fock theory are necessary for the full quantum mechanical description of "large" systems. In the case of even larger systems, such as a solvated protein, it becomes even more necessary to systematically decrease the computational cost of the quantum mechanical method. Thus, a new paradigm for quantum chemistry is required for tractable use of quantal force fields.

Explicit polarization theory (X-Pol) is one general approach for modeling large systems with quantum mechanical formalism. X-Pol is based on the separation of the system into sets of atoms according to structural or functional arrangement called *fragments*. Fragments may consist of individual molecules, such as water molecules, or may be individual amino acids divided along a covalent bond using the generalized hybridized orbital (GHO) theory [82, 83], as in the case of polypeptide chains [33], or a group of several molecules or amino acids. Here, the bonded interactions within a fragment are fully modeled by the QM theory used whereas non-bonded, interactions between different fragments are approximated by a one-electron integral term specifying Coulomb interactions between charged particles, and empirical terms for short-range repulsion and long-range dispersion (van der Waals) interactions. Such a treatment for a system of N fragments, each with m basis function, reduces the computational cost from formally $\mathcal{O}([Nm]^k)$ for the pure quantum calculation, to a cost of formally $\mathcal{O}(Nm^k)$ for the same level of quantum theory under X-Pol[1] .

2.3.1 Wave Function Description

Unlike the Hartree-Fock method, for which the anti-symmetry principle is satisfied by treating the entire wave function as a Slater determinant, the X-Pol method assumes that electrons within individual fragments are localized, but can be polarized, and that significant charge transfer does not occur between fragments. This assumption allows the treatment of the total molecular system of fragments as the Hartree product of Slater determinants from the N individual fragments (Eq. 2.22).

$$\Psi_{\text{TOT}} = \prod_{A=1}^{N} \Psi_A \tag{2.22}$$

The consequence of this assumption is that electronic exchange-correlation and dispersion effects between fragments are ignored. In condensed phase systems, the neglect of exchange-correlation and dispersion effects can drastically reduce the accuracy of the model; consequently, they must be corrected, albeit in an approximate way. Although treatments exist for modeling these effects in the context of QM/MM-type approaches in which density-dependent exchange and repulsion terms can be incorporated into the effective Hamiltonian [84], for the sake of simplicity, we assume that

[1] As an example, k is formally equal to 4 for Hartree-Fock theory and 5 for MP2.

the exchange-correlation and dispersion effects between fragments can be modeled in X-Pol by an empirical Lennard-Jones potential [85] with combining rules $\epsilon_{ab} = \sqrt{\epsilon_a \epsilon_b}$ and $\sigma_{ab} = \sqrt{\sigma_a \sigma_b}$ for atoms a in fragment A and b in fragment B.

$$E_{AB}^{\mathrm{XD}} = \sum_{a \in A} \sum_{b \in B} 4\epsilon_{ab} \left[\left(\frac{\sigma_{ab}}{r_{ab}} \right)^{12} - \left(\frac{\sigma_{ab}}{r_{ab}} \right)^{6} \right] \tag{2.23}$$

2.3.2 The X-Pol Hamiltonian

X-Pol is a fragment-based, or QM/QM, quantum chemistry method where the MOs of individual fragments are polarized by their surrounding fragments through explicit Coulomb integrals. The Hamiltonian described here is a variant of the X-Pol method that approximates the surrounding fragments as point charges, changing what would normally be a two-electron integral into a one-electron integral[2] . If the QM electrostatic potential due to other fragments is directly used instead of the MM point charge representation, there is no approximation for a given basis set used. However, this would not reduce computational cost much, since the two-electron integrals would be retained, scaling at the same complexity as standard Hartree-Fock theory.

For a predetermined set of fragments, the Hamiltonian of the X-Pol method can be thought of as the sum of two terms (Eq. 2.24).

$$\mathcal{H}^{\mathrm{XP}} = \sum_{A=1}^{N} \mathcal{H}_A^o + \frac{1}{2} \sum_{A=1}^{N} \sum_{B \neq A}^{N} \mathcal{H}_{AB} \tag{2.24}$$

The first of these terms, involving $\mathcal{H}_A^o$, is the sum of the electronic Hamiltonians of the individual fragments, which may be obtained from any level of QM theory, including density functional theory, Hartree-Fock, or semiempirical methods [87].

[2] Although we only consider point charges here, a more general approach using multipole expansions is available [86].

The second term describes the interactions between fragments A and B, where the total interaction is half of the double summation over all $\mathcal{H}_{AB}$ for $A \neq B$. Eq. 2.25 gives the expression for $\mathcal{H}_{AB}$ where m is the number of electrons in fragment A, M_A is the number of atoms in fragment A, Z_α^A denotes the nuclear charge of atom α on fragment A, and E_{AB}^{XD} is the exchange-correlation and dispersion interaction between fragments A and B.

$$\mathcal{H}_{AB} = -\sum_{k=1}^{m} V_k(\Psi_B) + \sum_{\alpha=1}^{M_A} Z_\alpha^A V_\alpha(\Psi_B) + E_{AB}^{\text{XD}} \tag{2.25}$$

$V_x(\Psi_B)$ is the electrostatic potential of fragment B, at position α of a particle (electron or nucleus) of fragment A:

$$V_x(\Psi_B) = -\int \frac{\rho_B(\mathbf{r})d\mathbf{r}}{|\mathbf{r}_x - \mathbf{r}|} + \sum_{\beta=1}^{M_B} \frac{Z_\beta^B}{|\mathbf{r}_x - \mathbf{R}_\beta^B|}, \tag{2.26}$$

where M_B is the number of atoms in fragment B and $x = k$ and $x = \alpha$ denote an interaction at electronic and nuclear positions respectively, and $\rho_B(\mathbf{r})$ denotes the electron density of fragment B derived from Ψ_B. When the electron density $\rho_B(\mathbf{r})$ of Eq. 2.26 is approximated by partial charges at nuclear positions, as is the case for our purposes here, the interaction Hamiltonian becomes

$$\mathcal{H}_{AB} = -\sum_{k=1}^{m}\sum_{\beta=1}^{M_B} \frac{e \cdot q_\beta(\Psi_B)}{|r_k - R_\beta|} + \sum_{\alpha=1}^{M_A}\sum_{\beta=1}^{M_B} \frac{Z_\alpha^A q_\beta(\Psi_B)}{R_{\alpha\beta}} + E_{AB}^{\text{XD}}, \tag{2.27}$$

where q_β is the partial charge of atom β, derived from population analysis of the wave function Ψ_B, and e is a unit electron charge. If this point-charge approximation is not made, the use of the interaction Hamiltonian of Eq. 2.25 will require calculation of explicit two-electron integrals between fragments instead of effective two-electron integrals approximated by one-electron integrals between the electron density of fragment A and the point charges of fragment B.

Within Hartree-Fock theory or Kohn-Sham density functional theory, the variational X-Pol Fock matrix is given by,

$$F^{\text{XP},A}_{\mu\nu} = F^{A}_{\mu\nu} - \frac{1}{2}\sum_{B\neq A}\sum_{b\in B} q^{B}_{b}\langle\mu|\frac{1}{r_{1b}}|\nu\rangle - \frac{1}{2}\frac{\partial q_a}{\partial P_{\mu\nu}}\sum_{B\neq A}\sum_{\lambda\sigma\in B}\left[P^{B}_{\lambda\sigma}\langle\lambda|\frac{1}{r_{1a}}|\sigma\rangle + \sum_{b\in B}\frac{Z^{B}_{b}}{r^{AB}_{ab}}\right], \tag{2.28}$$

and has a total interaction energy of,

$$E_{\text{TOT}} = \langle\Psi_{\text{TOT}}|\mathcal{H}^{\text{XP}}|\Psi_{\text{TOT}}\rangle - \sum_{A=1}^{N}\langle\Psi^{o}_{A}|\mathcal{H}^{o}_{A}|\Psi^{o}_{A}\rangle, \tag{2.29}$$

where the superscript "o" denotes the optimized wave function of a single fragment in the gas phase.

2.3.3 Double Self-Consistent Field Procedure

Similar to the Hartree-Fock method where the Fock matrix has an explicit dependence on the orbital coefficients, which are also the eigenvectors of the Roothaan-Hall equations, the polarization term in the X-Pol method depends on partial charges, which depend on the effects to the density matrix from the polarization term. This problem is solved in a manner similar to the SCF procedure, and yields what is known as the double self-consistent field (DSCF) procedure.

In the DSCF method, outlined in Figure 2.3, an initial guess of the partial charges is made and the polarization and variational terms between all fragments is calculated.

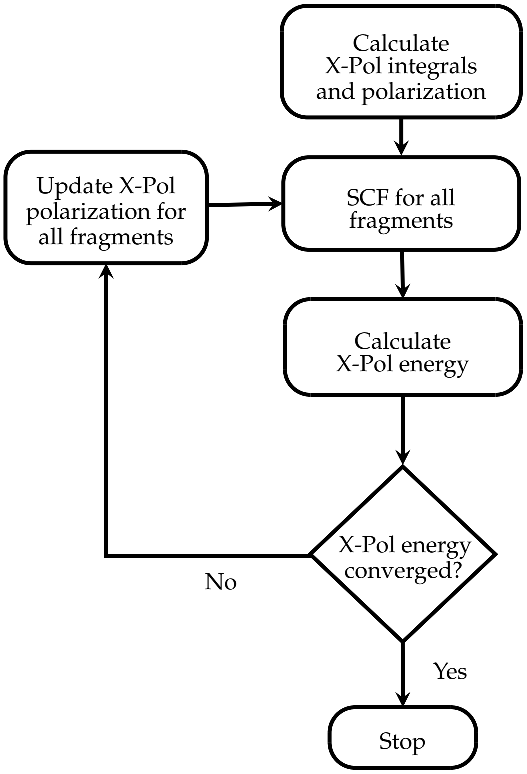

Calculate X-Pol integrals and polarization
Update X-Pol polarization for all fragments
SCF for all fragments
Calculate X-Pol energy
X-Pol energy converged?
No
Yes
Stop

Figure 2.3: A flow chart showing the steps of the double self-consistent field procedure used in Hartree-Fock based variants of X-Pol. The box labeled "SCF for all fragments" denotes only the cyclic portion of the flow chart from Figure 2.2.

Next, the SCF procedure is carried out on all fragments for the polarized X-Pol Hamiltonian, and the total electronic energy from the X-Pol method is calculated. If this energy is determined to be converged, the DSCF procedure stops; otherwise, the DSCF procedure repeats using the newly calculated partial charges as the next guess.

In practice, the step where the single SCF is run for each fragment need not converge entirely. This is because the density matrix and partial charges from each fragment obtained from the single SCF are contaminated at the start of the next DSCF iteration. We have determined empirically that two steps of diagonalization and density matrix reconstruction in the single SCF procedure for each fragment provides the fastest convergence in terms of overall computation time for simulations carried out in Chapters 3 and 4, using semiempirical Hamiltonians.

2.3.4 Analytical First Derivative of Energy

The components of the analytical first derivative of X-Pol can be written as three main terms: gas-phase Hartree-Fock contribution, X-Pol polarization contribution with nucleus-nucleus repulsion, and variational terms related to changing partial charges used for polarization (Eq. 2.30).

$$
\begin{aligned}
E_{\text{X-Pol}}^{X_{A,m}} = E_{\text{HF}}^{X_{A,m}} &+ \frac{1}{2}\sum_{\mu\nu} P_{\mu\nu}\left[\sum_{n\neq m}\sum_{B\in n} q_B I_{\mu\nu}^{X_{A,m}}(B,n)\right] + \frac{1}{2}\sum_{A\in m}\sum_{n\neq m}\sum_{B\in n} q_B Z_A \left(\frac{1}{R_{AB}}\right)^{X_{A,m}} \\
&+\frac{1}{2}\sum_{n\neq m}\sum_{\lambda\sigma\in n} P_{\lambda\sigma}\left[\sum_{A\in m} q_A I_{\lambda\sigma}^{X_{A,m}}(A,m)\right] + \frac{1}{2}\sum_{A\in m}\sum_{n\neq m}\sum_{B\in n} q_A Z_B \left(\frac{1}{R_{AB}}\right)^{X_{A,m}} \\
&+\sum_{\mu\nu}\left[\frac{1}{2}\sum_{n\neq m}\sum_{A\in m}\left[\sum_{\lambda\sigma} I_{\lambda\sigma}(A,m) + \sum_{B\in n}\frac{Z_B}{R_{AB}}\right]\frac{\partial q_A}{\partial S_{\mu\nu}}\right] S_{\mu\nu}^{X_{A,m}}
\end{aligned}
\tag{2.30}
$$

Here q_A and q_B denote the partial charges of atoms $A \in m$ and $B \in n$, Z_A and Z_B denote the nuclear charge, and $I_{\mu\nu}$ and $I_{\lambda\sigma}$ denote the one-electron integrals used to build the X-Pol Fock matrix of Eq. 2.28.

2.4 Modified Neglect of Diatomic Overlap Approximation

The computational complexity of the *ab initio* Hartree-Fock method scales formally at an order of $\mathcal{O}(m^4)$, for a chemical system of m basis functions, due to the four atomic centers in the two-electron integral evaluation[3] . A *semiempirical* way of reducing this cost is through neglect of diatomic differential overlap (NDDO) approximation, first proposed by Pople in 1967 [88].

The NDDO approximation assumes that the differential overlap of two atomic orbitals on different centers is zero; that is, $\int \phi_\mu \phi_\nu dr = S_{\mu\nu} = \delta_{\mu\nu}$. The most direct consequence of the NDDO approximation is the replacement of the overlap matrix **S** by the identity matrix. In addition to the elimination of the overlap matrix, three-center and four-center electronic integrals are eliminated, resulting in an order of $\mathcal{O}(m^2)$ two-center, two-electron integrals.

Although the number of integrals is reduced by two orders, the matrix diagonalization in the SCF procedure, which costs $\mathcal{O}(m^3)$, becomes the dominating expense. Reducing the order of the computational complexity beyond $\mathcal{O}(m^3)$ is difficult, although the number of basis functions m can be reduced to represent only the valence electrons of an atom, or the use of localized orbitals.

The NDDO-based, modified neglect of differential overlap (MNDO) method, introduced by Dewar and Thiel in 1977 [89], assumes a minimal basis set of s and p orbitals to represent only the valence electrons of atoms, while fixing the other electrons to the

[3] We note that most modern electronic structure codes have much better computational scaling due to the use of prescreening of electronic integrals and the use of other linear-scaling techinques.

nuclear center, resulting in an effective core for each atomic center. The approximations used in MNDO result in major changes to the $\mathbf{H}^{\text{core}}$ matrix, the Fock matrix, and the nucleus-nucleus repulsion terms compared to *ab initio* Hartree-Fock theory.

Since we have extensively used MNDO-based methods in this work, we provide an outline of the MNDO method, noting that a more detailed description, suitable for writing an implementation, is presented by Stewart in Ref. [90].

2.4.1 Integral Calculations

The largest set of integrals retained in MNDO are the two-center, two-electron integrals. In MNDO and similar NDDO-type semiempirical methods, the two-electron integrals are calculated through a linear-transformed multipole expansion between two atoms aligned along the Z-axis (local coordinate system) [91]. These transformed, two-electron integrals are also used to approximate the two-center, one-electron integrals $\langle\mu|Z_B/r_{1B}|\nu\rangle \approx \langle\mu\nu|s_B s_B\rangle$.

The transformation matrix from local to laboratory coordinates for the electronic integrals between atoms A and B by multipole expansion [92] is given in Eq. 2.31,

$$\mathbf{T} = \begin{pmatrix} \frac{X}{R_{AB}} & \frac{Y}{R_{AB}} & \frac{Z}{R_{AB}} \\ -\frac{Y}{R_{XY}} & \frac{X}{R_{XY}} & 0 \\ \frac{XZ}{R_{XY}R_{AB}} & \frac{YZ}{R_{XY}R_{AB}} & \frac{R_{XY}}{R_{AB}} \end{pmatrix}, \tag{2.31}$$

where $X = x_A - x_B$, $Y = y_A - y_B$, $Z = z_A - z_B$, and $R_{XY} = \sqrt{X^2 + Y^2}$, $R_{AB} = \sqrt{X^2 + Y^2 + Z^2}$.

For the gradient calculation, the derivative of the rotation matrix is required in conjunction with the product rule, since both the transformation matrix $\mathbf{T}$ and the integrals v have an R_{AB} dependence.

$$\frac{\partial}{\partial u_C}\mathbf{T}v(R_{AB}(x_A,x_B,y_A,y_B,z_A,z_B)) = \frac{\partial \mathbf{T}}{\partial u_C}v + \mathbf{T}\left(\frac{\partial v}{\partial R_{AB}}\right)\left(\frac{\partial R_{AB}}{\partial u_C}\right), \tag{2.32}$$

where u_C represents a chosen Cartesian coordinate and

$$\frac{\partial R_{AB}}{\partial u_C} = -\frac{2u_C}{\sqrt{X^2+Y^2+Z^2}}. \tag{2.33}$$

2.4.2 $\mathbf{H}^{\text{core}}$ Matrix

The $\mathbf{H}^{\text{core}}$ matrix in MNDO is constructed using different expressions for diagonal and off-diagonal blocks. The diagonal blocks of $\mathbf{H}^{\text{core}}$ represent the self-interaction and polarization on a single atom. In MNDO, the diagonal blocks of the $\mathbf{H}^{\text{core}}$ matrix are given by,

$$\begin{aligned} H^{\text{core}}_{\mu\mu} &= U_{\mu\mu} - \sum_{B\neq A} Z_B\langle\mu\mu|s_B s_B\rangle, \\ H^{\text{core}}_{\mu\nu} &= -\sum_{B\neq A} Z_B\langle\mu\nu|s_B s_B\rangle, \end{aligned} \tag{2.34}$$

where the parameters U_{ss} and U_{pp} are the one-center, one-electron integrals, and Z_B is the core charge of atom B (i.e. the net charge of the nucleus and core electrons). There are no U_{sp} or $U_{pp'}$ parameters due to the orthogonality of the atomic orbitals on a single nuclear center. Note that the one-electron two-center integrals are approximated by two-center two-electron integrals in MNDO.

The off-diagonal blocks of $\mathbf{H}^{\text{core}}$ are between two atomic centers, and are called *resonance integrals*. The resonance integrals provide a description of the chemical bond.

In MNDO, the resonance integrals are approximated by,

$$H_{\mu\nu} = \frac{1}{2}(\beta_\mu + \beta_\nu)S_{\mu\nu}, \tag{2.35}$$

where β_μ and β_ν are parameters, and $S_{\mu\nu}$ is an overlap integral determined by $\langle\phi_\mu|\phi_\nu\rangle$. In principle, this violates the NDDO approximation, but is done in MNDO to provide a better description of the chemical bond. In practice, $S_{\mu\nu}$ is determined by employing a STO or GTO basis representation of ϕ_μ and ϕ_ν, using the ζ_s and ζ_p parameters as orbital exponents.

2.4.3 Fock Matrix

Similar to the one-center terms involved in $\mathbf{H}^{\text{core}}$, the one-center, two-electron integrals are parameterized into values $G_{ss} = \langle ss|ss\rangle$, $G_{sp} = \langle ss|pp\rangle$, $H_{sp} = \langle sp|sp\rangle$, $G_{pp} = \langle pp|pp\rangle$, $G_{pp'} = \langle pp|p'p'\rangle$, and $H_{pp'} = \langle pp'|pp'\rangle$. The $G_{\mu\nu}$ parameters are Coulomb-type integrals and the $H_{\mu\nu}$ parameters are exchange-type integrals. Due to symmetry, it has been shown that $H_{pp'} = (G_{pp} - G_{pp'})/2$, which preserves rotational invariance.

The one-center terms contribute to the diagonal blocks of the Fock matrix as below [90],

$$\begin{aligned}
F_{ss} &: P_{ss}G_{ss} + (P_{p_xp_x} + P_{p_yp_y} + P_{p_zp_z})(G_{sp} - H_{sp}) \\
F_{sp} &: 2P_{sp}H_{sp} - P_{sp}(H_{sp} + G_{sp}) \\
F_{pp} &: P_{ss}G_{sp} - P_{ss}H_{sp} + P_{pp}G_{pp} + (P_{p'} + P_{p''})G_{pp'} - (P_{p'} + P_{p''})H_{pp'} \\
F_{pp'} &: 2P_{pp'}H_{pp'} - \frac{1}{2}P_{pp'}(G_{pp} + G_{pp'})
\end{aligned} \tag{2.36}$$

where $P_{\mu\nu}$ are respective elements of the associated density matrix.

The remainder of the two-electron integrals in MNDO are two-center terms, calculated from the multipole expansion [91]. After transformation of the two-electron integrals, the diagonal blocks are further modified by,

$$\begin{aligned} F_{\mu\mu} &: H_{\mu\mu} + \sum_{\nu}^{A} (P_{\nu\nu} \langle \mu\mu|\nu\nu\rangle - P_{\nu\nu}\langle \mu\nu|\mu\nu\rangle) + \sum_{B}^{B} \sum_{\lambda}^{B} \sum_{\sigma}^{B} P_{\lambda\sigma}\langle \mu\mu|\lambda\sigma\rangle, \\ F_{\mu\nu} &: H_{\mu\nu} + 2P_{\mu\nu}\langle \mu\nu|\mu\nu\rangle - P_{\mu\nu}(\langle \mu\nu|\mu\nu\rangle + \langle \mu\mu|\nu\nu\rangle). \end{aligned} \tag{2.37}$$

For ϕ_μ and ϕ_ν on different centers (off-diagonal blocks), the MNDO Fock matrix expression is,

$$F_{\mu\nu} = H_{\mu\nu} - \sum_{\lambda}^{A} \sum_{\sigma}^{B} P_{\lambda\sigma}\langle \mu\lambda|\nu\sigma\rangle \tag{2.38}$$

2.4.4 Core-Core Repulsion

The core-charge analog of the nucleus-nucleus repulsion, called the *core-core repulsion*, does not take the same form as that of *ab initio* Hartree-Fock theory. This is due to the use of the effective core nucleus charges such that the Coulomb interactions are screened by core (as well as valence) electrons.

The MNDO core-core repulsion term has the general form,

$$V_{\text{core}}^{AB} = Z_A Z_B \langle s_A s_A | s_B s_B\rangle \left[1 + e^{-\alpha_A R_{AB}} + e^{-\alpha_B R_{AB}}\right] \tag{2.39}$$

and the special form

$$V_{\text{core}}^{AB} = Z_A Z_H \langle s_A s_A | s_H s_H\rangle \left[1 + R_{AH} e^{-\alpha_A R_{AH}} + e^{-\alpha_H R_{AH}}\right] \tag{2.40}$$

for O-H and N-H interactions.

2.4.5 X-Pol with MNDO

Under the MNDO formalism, the one-center, one-electron integral is replaced by the two-center, two-electron integral in the form $\langle \mu^a \nu^a | s^b s^b \rangle$. Furthermore, Mulliken charges have a density derivative $\partial q_a / \partial P_{\mu\nu} = -1$, due to the approximation of the overlap matrix as identity, which also causes the last term of the X-Pol derivative expression to vanish (Eq. 2.30).

The Fock matrix for MNDO under X-Pol becomes,

$$F^{\mathrm{XP},A}_{\mu\nu} = F^{A}_{\mu\nu} - \frac{1}{2} \sum_{B \neq A} \sum_{b \in B} q^{B}_{b} \langle \mu\nu | s^b s^b \rangle + \frac{1}{2} \sum_{B \neq A} \sum_{\lambda\sigma \in B} \left[P^{B}_{\lambda\sigma} \langle \lambda\sigma | s^a s^a \rangle + \sum_{b \in B} Z^{B}_{b} \langle s^b s^b | s^a s^a \rangle \right] . \tag{2.41}$$

where $F^{A}_{\mu\nu}$ is the typical MNDO Fock matrix, and q^{B}_{b} is the Mulliken charge on atom b of fragment B.

Due to the one-electron integral polarization by partial charges in the X-Pol method, the standard MNDO core-core repulsion is modified such that the partial charge q_a of the MM atom is used for the interfragment interaction, rather than the core charge Z_a, as in the standard MNDO method.

A detailed explanation of the analytical first derivative of the MNDO method is provided in Ref. [92].

2.4.6 Polarized Molecular Orbital Method

The polarized molecular orbital (PMO) method [52,93,94] is a semiempirical quantum chemistry method with three main improvements over the MNDO formalism. The

PMO method was introduced with the aim to improve the description of intermolecular interactions within the NDDO formalism.

First, a set of p-orbitals is added to hydrogen atoms. This greatly enhances molecular polarizabilities and provides significantly improved hydrogen bonding performance over MNDO. The resonance integrals on the additional p-orbitals are damped to compensate for the increase in overlap terms (Eqs. 2.42 and 2.43), and special exponents are used for homonuclear resonance integral calculations.

$$H_{lp}^{\mathrm{HH}} = 0, \tag{2.42}$$

$$H_{lp}^{\mathrm{XH}} = \frac{\beta_l^{\mathrm{X}} + \beta_p^{\mathrm{H}}}{2} S_{lp} A_{lp} e^{\kappa_{lp} R_{\mathrm{XH}}}. \tag{2.43}$$

Second, the one-electron attraction integral from MNDO (which is approximated by a two-electron integral) is modified in PMO when both atoms A and B are hydrogen (Eq. 2.44).

$$\langle pp' | s^H s^H \rangle_{\mathrm{PMO}} = \left[1 - B e^{-\lambda R_{HH'}^2}\right] \langle pp' | s^H s^H \rangle_{\mathrm{MNDO}} \tag{2.44}$$

Note that this is only done in the context of the one-electron integral.

Finally, special exponents $\hat{\alpha}$ are used for homonuclear core-core repulsion in addition to the existing core-core term of MNDO, and dispersion effects are accounted for between all atom pairs by the empirical D1 dispersion correction of Grimme [95].

The PMO method has been used with X-Pol and the DPPC charge model (see next section) to model liquid water (Chapter 3) and liquid hydrogen fluoride (Chapter 4).

2.5 Dipole-Preserving and Polarization-Consistent Charge

A key component in the X-Pol method is having an accurate and effective representation of the electrostatic potential based on the QM wave function of each fragment. The simplest approximation is the use of monopoles only (i.e. atomic partial charges). Mulliken population analysis [96] is a widely-used scheme for calculating partial charges. Under the NDDO approximation, the modification to the overlap matrix $S_{\mu\nu} = \delta_{\mu\nu}$ causes partial charges q_k obtained from Mulliken population analysis to take the form of Eq. 2.45,

$$q_k^{\mathrm{MP}} = Z_k - \sum_{\mu\nu \in k} (P_{\mu\nu} S_{\mu\nu}) = Z_k - \sum_{\mu \in k} P_{\mu\mu} \tag{2.45}$$

where atom k has core charge Z_k and diagonal density matrix elements $P_{\mu\mu}$.

In general, dipole moments calculated using Mulliken point charges fail to reproduce the total molecular dipole moment of a given system under the NDDO approximation. This is because, in general, the molecular dipole moment in the NDDO framework is given by two distinct contributions; one from Mulliken population analysis $\mathbf{D}_{\mathrm{MP}}$ (monopole contribution), and another from one-center sp-hybridized orbitals $\mathbf{D}_{\mathrm{hyb}}$ (dipole contibution) (Eq. 2.46).

$$\mathbf{D}_{\mathrm{NDDO}} = \mathbf{D}_{\mathrm{MP}} + \mathbf{D}_{\mathrm{hyb}} = \sum_{k=1}^{N} q_k^{\mathrm{MP}} \mathbf{r}_k - \sum_{k=1}^{N} R_k (\mathbf{P}_{sp})_k \tag{2.46}$$

$\mathbf{D}_{\mathrm{MP}}$ is the sum of the products of Mulliken charge q_k^{MP} and Cartesian position $\mathbf{r}_k$ of atom k and $\mathbf{D}_{\mathrm{hyb}}$ is the sum of the products of the dipole integral R_k and the one-center sp-hybridized orbital density matrix elements taken as a vector $(\mathbf{P}_{sp})_k \in \mathbb{R}^3$. The dipole integral of atom k is in practice twice the dipole displacement used in the multipole expansion for the semiempirical calculation of two-electron integrals [91].

The dipole-preserving, polarization-consistent (DPPC) population analysis method [53] produces a correction to the Mulliken charge which reproduces the NDDO molecular dipole moment under a point charge treatment. We describe the method below and give an expression for its use with X-Pol and derive its analytical first derivative.

2.5.1 DPPC Method

The DPPC method has its origins in the dipole-preserving charge (DPC) method of Thole and van Duijnen [97]. In both treatments, a correction to the Mulliken charge for each atom from all other atoms (Eq. 2.47) is derived subject to the constraints that the net charge on the system remains constant (Eq. 2.48) and that the hybridized dipole moment in the point-charge regime is preserved by the corrections (Eq. 2.49).

$$q_k = q_k^{\mathrm{MP}} + \sum_{i=1}^{N} \Delta q_i \tag{2.47}$$

$$\sum_{k=1}^{N} \Delta q_k^i = 0 \tag{2.48}$$

$$\mathbf{D}_{\mathrm{hyb}}^i = \sum_{k=1}^{N} \Delta q_k^i \mathbf{r}_k \tag{2.49}$$

By inspection, it is clear that satisfying these constraints will reproduce $\mathbf{D}_{\mathrm{NDDO}}$, using only the product of Cartesian coordinates and point charges of Eq. 2.47 for each atom.

The charge correction is minimized subject to a set of weights that decay as a Gaussian with distance, and are scaled by a term related to the relative difference of electronegativity of the atoms in the system (Eq. 2.50).

$$w_k^i = \left(1 + \left|\frac{\eta_k - \eta_i}{\eta_i}\right|\right) e^{-\lambda|\mathbf{r}_k - \mathbf{r}_i|^2} \tag{2.50}$$

In practice, the employment of this weighting function is the main difference between the DPC and DPPC methods, although we have extended the DPPC method to X-Pol and provide its analytical gradient here. The value $\lambda = 1$ and the Pauling electronegativities [98] η of each element are used, some of which are shown in Table 2.1.

Element	H	C	N	O	P	S
η	2.20	2.55	3.04	3.44	2.19	2.58

Table 2.1: Pauling electronegativity for the six most essential elements in living organisms. A list for elements up to atomic number 94 is provided in the CRC handbook. [1]

The minimization of the weighted charge corrections under the two linear constraints is solved using Lagrange multipliers (Eq. 2.51).

$$L^i = \left(\sum_{k=1}^{N} \frac{\left(\Delta q_k^i\right)^2}{2w_k^i} \right) + \left(0 - \sum_{k=1}^{N} \Delta q_k^i \right) \alpha_i + \left(\mathbf{D}_{\text{hyb}}^i - \sum_{k=1}^{N} \Delta q_k^i \mathbf{r}_k \right)^T \boldsymbol{\beta}_i \tag{2.51}$$

The solution of the minimization problem is given in Eq. 2.52. A full derivation of this expression and its associated pieces (Eqs. 2.53, 2.54, 2.74) is derived in detail elsewhere [53].

$$\Delta q_k^i = \frac{w_k^i}{W^i} \left[\left(\mathbf{r}_k - \langle \mathbf{r} \rangle_i \right)^T \cdot \left(\boldsymbol{\Omega}^i \right)^{-1} \cdot \mathbf{D}_{\text{hyb}}^i \right] \tag{2.52}$$

$$W^i = \sum_{k=1}^{N} w_k^i \tag{2.53}$$

$$\langle \mathbf{r} \rangle_i^T = \frac{1}{W^i} \sum_{k=1}^{N} w_k^i \left(\mathbf{r}_k \right)^T \tag{2.54}$$

$$\mathbf{\Omega}^i = \langle \mathbf{r}\mathbf{r}^T \rangle_i - \langle \mathbf{r} \rangle_i \langle \mathbf{r} \rangle_i^T = \frac{1}{W_i} \left[\sum_{k=1}^{N} w_k^i \mathbf{r}_k \left(\mathbf{r}_k \right)^T \right] - \langle \mathbf{r} \rangle_i \langle \mathbf{r} \rangle_i^T \tag{2.55}$$

An important detail in the DPPC method is the construction of the $\mathbf{\Omega}$ matrix. $\mathbf{\Omega}$ is a real-valued, symmetric 3×3 matrix that requires an inversion in Eq. 2.52. In the case of linear and planar molecules, such as hydrogen fluroide and water, $\mathbf{\Omega}$ is singular and Eq. 2.52 has no unique solution. To prevent this situation, a matrix diagonalization is performed instead of an inversion [97]. The motivation is explained as follows.

Consider the reformulation of the matrix-vector multiplication $\mathbf{\Omega}^{-1}\mathbf{D}_{\text{hyb}} = x$ into the solution of the linear system $\mathbf{\Omega}x = \mathbf{D}_{\text{hyb}}$. Let $\mathbf{\Omega} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^T$ represent a matrix diagonalization of $\mathbf{\Omega}$. Notice that $\mathbf{\Omega}x = \mathbf{D}_{\text{hyb}} \Leftrightarrow \mathbf{U}\mathbf{\Sigma}\mathbf{U}^T x = \mathbf{D}_{\text{hyb}}$ and that $\mathbf{U}\mathbf{\Sigma}\mathbf{U}^T x = \mathbf{D}_{\text{hyb}} \Leftrightarrow \mathbf{U}\mathbf{\Sigma}^{-1}\mathbf{U}^T\mathbf{D}_{\text{hyb}} = x$, provided that $\mathbf{\Sigma}$ is non-singular.

The matrix $\mathbf{\Sigma}$ is singular if and only if it contains a zero eigenvalue, which is true in the case of planar and linear molecules. The generally ill-posed $\mathbf{\Sigma}$ matrix is forced to be non-singular for *all systems* by a small perturbation of the eigenvalues of $\mathbf{\Omega}$. This not only guarantees that $\mathbf{\Sigma}$ is non-singular, but also that the first derivative of the DPPC charge is continuous, although some error is introduced.

Let ω denote the eigenvalues of $\mathbf{\Omega}$. The perturbation of eigenvalues in the DPPC method is given in Eq. 2.56, where $\theta = 10^{-5}$ in our implementation.

$$\omega' = \omega + \theta(\omega_{\text{max}} + \theta) \tag{2.56}$$

2.5.2 Modification to X-Pol Fock matrix by DPPC

The modification to the Fock matrix in the X-Pol method (Eq. 2.57) related to variational optimization contains two derivative terms involving partial charges (Eqs. 2.58

and 2.59).

$$F^{A}_{\mu\nu} = f^{A}_{\mu\nu} - \frac{1}{2}\sum_{j \notin A} \left(I_j\right) q_j + \sum_{j \in A} M_j \left(\Lambda_j\right)_{\mu\nu} \tag{2.57}$$

$$M_j = \frac{\partial E_{QM/MM}}{\partial q_j} = \frac{1}{2} \sum_{B(J \neq B)} \left(- \sum_{\lambda\sigma \in b} P_{\lambda\sigma} I^{j}_{\lambda\sigma} + \sum_{i \in B} L_{ij} \right) \tag{2.58}$$

$$\left(\Lambda_j\right)_{\mu\nu} = \frac{\partial q_j}{\partial P_{\mu\nu}} \tag{2.59}$$

The DPPC charge is fundamentally different from Mulliken charge and thus requires a modification of the term derived from Eq. 2.59.

Since the DPPC method provides a sum of charge corrections to the Mulliken charge, it is clear that the form of $(\Lambda_j)_{\mu\nu}$ will be that of Eq. 2.60, due to the linear nature of the derivative operator.

$$\left(\Lambda_j\right)^{\text{DPPC}}_{\mu\nu} = \left(\Lambda_j\right)^{\text{MP}}_{\mu\nu} + \sum_{i \in A}^{N_A} \left(\Delta\Lambda^{i}_{j}\right)_{\mu\nu} \tag{2.60}$$

Differentiating the Mulliken charge under the NDDO approximation (Eq. 2.45) with respect to $P_{\mu\nu}$ for orbitals μ and ν on atom j results in a simple equation of a negative Kronecker delta (Eq. 2.61).

$$\left(\Lambda_j\right)^{\text{MP}}_{\mu\nu} = -\delta_{\mu\nu}. \tag{2.61}$$

The contribution to the variational term of X-Pol from each DPPC charge correction Δq^{i}_{k} is more complex than the Mulliken case, and requires the differentiation of Eq. 2.52 with respect to $P_{\mu\nu}$ for orbitals μ and ν on atom i. The only density dependent term is the hybridized dipole contribution $\mathbf{D}^{i}_{\text{hyb}}$, causing $(\Delta\Lambda^{i}_{j})_{\mu\nu}$ to take the form of Eq. 2.62

where $\partial \mathbf{D}^{i}_{\text{hyb},X}/\partial P_{\mu\nu}$ is given by 2.63 and X denotes one of the x, y, or z coordinates.

$$\left(\Delta\Lambda^{i}_{j}\right)_{\mu\nu} = \frac{\partial \Delta q^{i}_{j}}{\partial P_{\mu\nu}} = \frac{w^{i}_{j}}{W^{i}} \left[\left(\mathbf{r}_{j} - \langle \mathbf{r} \rangle_{i}\right)^{T} \cdot \left(\Omega^{i}\right)^{-1} \cdot \frac{\partial \mathbf{D}^{i}_{\text{hyb}}}{\partial P_{\mu\nu}} \right] \tag{2.62}$$

$$\frac{\partial \mathbf{D}^{i}_{\text{hyb},X}}{\partial P_{\mu\nu}} = \frac{1}{2} R^{i} \left(\delta_{\mu s}\delta_{\nu p_X} + \delta_{\nu s}\delta_{\mu p_X}\right)^{i} \tag{2.63}$$

The factor of a half in front of the dipole integral of Eq. 2.63 exists due to the symmetry of sp and ps orbitals and prevents double counting.

Eq. 2.63 shows that the DPPC charge density derivative will only modify sp-hybridized orbitals in the X-Pol Fock matrix. In addition, the observation that $(\Lambda_j)^{\text{MP}}_{\mu\nu} = -\delta_{\mu\nu}$ and further examination of Eqs. 2.57 and 2.58 show that the only terms produced by the variational part of X-Pol under the NDDO approximation with Mulliken charges are identical terms along the block diagonals of each atom in a fragment. The combination of these two facts implies that there are only four terms to calculate per atom-atom interaction.

The implication of this observation in the context of a parallel implementation is that these four terms can be calculated on the same processor where the density matrix of the interacting fragment resides, and can be passed to the other processors as needed, rather than passing the much larger density matrix of needed fragments to calculate the term of Eq. 2.58 on the processor that requires it. Such an observation of the variational term greatly increases feasibility of a parallel X-Pol implementation, although it requires the storage of both types of one-electron integrals for A:QM/B:MM and BQM:/A:MM on each processor.

2.5.3 DPPC Gradient

The use of X-Pol as a force field requires the derivative of the charge with respect to the position. Differentiating the DPPC charge correction Δq_k^i of Eq. 2.52 with respect to Cartesian coordinates $\mathbf{r}_j$ where i and j denote atoms within a fragment requires the product rule, which produces Eq. 2.64.

$$\begin{aligned}\frac{\partial \Delta q_k^i}{\partial \mathbf{r}_j} &= \frac{1}{W^i}\frac{\partial}{\partial \mathbf{r}_j}\left[w_k^i\left(\mathbf{r}_k - \langle r\rangle_i\right)^T \cdot \left(\boldsymbol{\Omega}^i\right)^{-1} \cdot \Delta\mathbf{D}_i\right] \\ &- \frac{1}{W_i^2}\frac{\partial W^i}{\partial \mathbf{r}_j}\left[w_k^i\left(\mathbf{r}_k - \langle \mathbf{r}\rangle_i\right)^T \cdot \left(\boldsymbol{\Omega}^i\right)^{-1} \cdot \Delta\mathbf{D}_i\right]\end{aligned} \tag{2.64}$$

With the exception of the factor $\partial W_i/\partial \mathbf{r}_j$, the second term of Eq. 2.64 is immediate from values already computed for the DPPC charge correction. Recall that W_i is the sum of the individual weights w_k^i (Eq. 2.53) and that these weights decay in a Gaussian-like manner with distance (Eq. 2.50). Differentiating w_k^i gives the piecewise equation in 2.65.

$$\frac{\partial w_k^i}{\partial \mathbf{r}_j} = \begin{cases} -2\lambda\left(\mathbf{r}_j - \mathbf{r}_i\right)\left(1 + \left|\dfrac{\eta_j - \eta_i}{\eta_i}\right|\right) e^{-\lambda|\mathbf{r}_j - \mathbf{r}_i|^2} & (j = k) \\ -2\lambda\left(\mathbf{r}_j - \mathbf{r}_k\right)\left(1 + \left|\dfrac{\eta_k - \eta_j}{\eta_j}\right|\right) e^{-\lambda|\mathbf{r}_k - \mathbf{r}_j|^2} & (j = i) \\ \qquad\qquad\qquad\qquad\qquad\qquad = 0 & (j \neq i \cap j \neq k) \end{cases} \tag{2.65}$$

The partial derivative $\partial W_i/\partial \mathbf{r}_j = \sum_{k=1}^{N} \partial w_k^i/\partial \mathbf{r}_j$ is also piecewise, consisting of either a summation or single term and takes the form of Eq. 2.66. Thus, the description of the second term of Eq. 2.64 is complete.

$$\frac{\partial W^i}{\partial \mathbf{r}_j} = \begin{cases} \displaystyle\sum_{k=1}^{N} \frac{\partial w_k^i}{\partial \mathbf{r}_i} & (j = i) \\ -2\lambda\left(\mathbf{r}_j - \mathbf{r}_i\right)\left(1 + \left|\dfrac{\eta_j - \eta_i}{\eta_i}\right|\right) e^{-\lambda|\mathbf{r}_j - \mathbf{r}_i|^2} & (j \neq i) \end{cases} \tag{2.66}$$

The derivation of the derivative of the first term of Eq. 2.64 with respect to $\mathbf{r}_j$ is much more involved than the second term since there are now three factors with explicit dependence on $\mathbf{r}_j$. Through double application of the product rule we obtain Eq. 2.67.

$$\begin{aligned} \frac{\partial}{\partial \mathbf{r}_j}\left[w_k^i \left(\mathbf{r}_k - \langle\mathbf{r}\rangle_i\right)^T \cdot \left(\boldsymbol{\Omega}^i\right)^{-1} \cdot \Delta\mathbf{D}_i\right] = & \frac{\partial w_k^i}{\partial \mathbf{r}_j}\left(\mathbf{r}_k - \langle\mathbf{r}\rangle_i\right)^T \cdot \left(\boldsymbol{\Omega}^i\right)^{-1} \cdot \Delta\mathbf{D}_i \\ & + w_k^i \frac{\partial\left(\mathbf{r}_k - \langle\mathbf{r}\rangle_i\right)^T}{\partial \mathbf{r}_j} \cdot \left(\boldsymbol{\Omega}^i\right)^{-1} \cdot \Delta\mathbf{D}_i \\ & + w_k^i \left(\mathbf{r}_k - \langle\mathbf{r}\rangle_i\right)^T \cdot \frac{\partial\left(\boldsymbol{\Omega}^i\right)^{-1}}{\partial \mathbf{r}_j} \cdot \Delta\mathbf{D}_i \end{aligned} \tag{2.67}$$

All three terms of Eq. 2.67 contain a product of a single derivative with respect to $\mathbf{r}_j$ and three other factors which are used in finding the charge correction Δq_k^i of Eq. 2.52. The first term of Eq. 2.67 is fully accounted for from our derivation of Eq. 2.65, but the second and third terms require further derivations.

The derivative of the second term can be directly calculated to produce Eq. 2.68.

$$\frac{\partial\left(\mathbf{r}_k - \langle\mathbf{r}\rangle_i\right)^T}{\partial \mathbf{r}_j} = \delta_{kj} - \frac{\partial\langle\mathbf{r}\rangle_i^T}{\partial \mathbf{r}_j} \tag{2.68}$$

Differentiating $\langle\mathbf{r}\rangle_i^T$ from Eq. 2.54 with respect to $\mathbf{r}_j$ produces the remaining piece as given in Eq. 2.69.

$$\frac{\partial \langle \mathbf{r} \rangle_i^T}{\partial \mathbf{r}_j} = \begin{cases} \dfrac{1}{W_i}\left(\displaystyle\sum_{k=1}^{N} \frac{\partial w_k^i}{\partial \mathbf{r}_i}\mathbf{r}_k^T + 1\right) - \dfrac{1}{W_i^2}\dfrac{\partial W^i}{\partial \mathbf{r}_i}\displaystyle\sum_{k=1}^{N} w_k^i \mathbf{r}_i^T & (j = i) \\ \dfrac{1}{W_i}\left(\dfrac{\partial w_j^i}{\partial \mathbf{r}_j}\mathbf{r}_j^T + w_j^i\right) - \dfrac{1}{W_i^2}\dfrac{\partial W_i}{\partial \mathbf{r}_j}\displaystyle\sum_{k=1}^{N} w_k^i \mathbf{r}_j^T & (j \neq i) \end{cases} \tag{2.69}$$

The third term involves differentiation of the inverse of the 3×3 matrix $\mathbf{\Omega}$, and is the most involved calculation. The general definition of this inverse is $\mathbf{\Omega^{-1}\Omega} = \mathbf{I}$, for which differentiation of both sides and some rearrangement gives Eq. 2.70.

$$\frac{\partial \left(\mathbf{\Omega}^i\right)^{-1}}{\partial \mathbf{r}_j} = -\left(\mathbf{\Omega}^i\right)^{-1} \frac{\partial \left(\mathbf{\Omega}^i\right)}{\partial \mathbf{r}_j} \left(\mathbf{\Omega}^i\right)^{-1} \tag{2.70}$$

The derivative of the matrix $(\mathbf{\Omega}^i)^{-1}$ with respect to a coordinate X of $\mathbf{r}_j$ is the derivative of the elements of that matrix with respect to that same variable (Eq. 2.71).

$$\left[\frac{\partial \left(\mathbf{\Omega}^i\right)^{-1}}{\partial \mathbf{r}_j}\right]_{mn}^{X} = \frac{\partial \left[\left(\mathbf{\Omega}^i\right)^{-1}\right]_{nm}}{\partial r_j^X} \tag{2.71}$$

The derivative of the matrix-vector product in Eq. 2.67 is,

$$\left[(\mathbf{r}_k - \langle \mathbf{r} \rangle_i)^T \cdot \frac{\partial \left(\mathbf{\Omega}^i\right)^{-1}}{\partial \mathbf{r}_j}\right]_m^X = \sum_{n=1}^{3} \left[(\mathbf{r}_k - \langle \mathbf{r} \rangle_i)^T\right]_n \left[\frac{\partial \left(\mathbf{\Omega}^i\right)^{-1}}{\partial \mathbf{r}_j}\right]_{mn}^X . \tag{2.72}$$

Elements n and m of the partial derivative of the $\mathbf{\Omega}^i$ matrix with respect to the X component of $\mathbf{r}_j$ are given by Eq. 2.73.

$$\left[\frac{\partial\left(\mathbf{\Omega}^i\right)}{\partial \mathbf{r}_j}\right]^X_{nm} = \begin{cases} \begin{array}{r} \dfrac{\partial \mathbf{\Omega}^i_{nm}}{\partial W^i}\dfrac{\partial W^i}{\partial r^X_j} + \dfrac{1}{W^i}\left(\displaystyle\sum_{k=1}^{N}\frac{\partial w^i_k}{\partial r^X_i} r^n_k r^m_k + \delta_{Xn} + \delta_{Xm}\right) \\ -\dfrac{1}{W_i^2}\left[\left(\displaystyle\sum_{k=1}^{N}\frac{\partial w^i_k}{\partial r^X_i} r^n_k + \delta_{Xn}\right)\sum_{k=1}^{N} w^i_k r^m_k + \sum_{k=1}^{N} w^i_k r^n_k \left(\sum_{k=1}^{N}\frac{\partial w^i_k}{\partial r^X_i} r^m_k + \delta_{Xm}\right)\right] \end{array} & (j = i) \\ \begin{array}{r} \dfrac{\partial \mathbf{\Omega}^i_{nm}}{\partial W^i}\dfrac{\partial W^i}{\partial r^X_j} + \dfrac{1}{W^i}\left(\dfrac{\partial w^i_j}{\partial r^X_j} r^n_j r^m_j + w^i_j \delta_{Xn} + w^i_j \delta_{Xm}\right) \\ -\dfrac{1}{W_i^2}\left[\left(\dfrac{\partial w^i_j}{\partial r^X_j} r^n_j + w^i_j \delta_{ln}\right)\displaystyle\sum_{k=1}^{N} w^i_k r^m_k + \sum_{k=1}^{N} w^i_k r^n_k \left(\frac{\partial w^i_j}{\partial r^X_j} r^m_j + w^i_j \delta_{Xm}\right)\right] \end{array} & (j \neq i) \end{cases} \tag{2.73}$$

where,

$$\frac{\partial \mathbf{\Omega}^i}{\partial W_i} = \frac{1}{W_i}\left[\frac{1}{W_i}\left[2\langle \mathbf{r}\rangle_i \langle \mathbf{r}\rangle_i^T - \sum_{k=1}^{N} w^i_k \mathbf{r}_k \left(\mathbf{r}_k\right)^T\right]\right]. \tag{2.74}$$

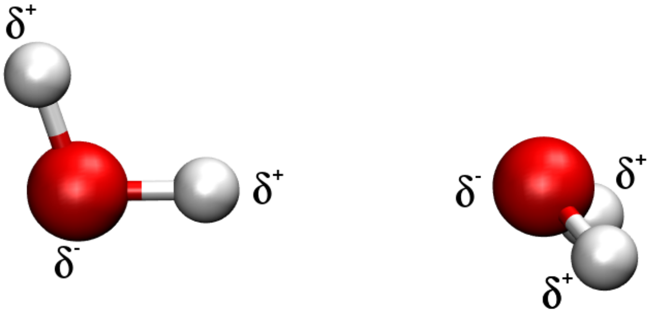
This completes the description of the terms needed in calculating the DPPC gradient.

2.6 Sample Data

To provide an initial test of the derivative methods described in the previous section, we performed several full semiempirical QM and semiempirical X-Pol calculations on the water dimer geometry given in Table 2.2 and the hydrogen fluoride dimer geometry given in Table 2.5. The results of each of these calculations are provided in two pairs of tables for various combinations of semiempirical theory. Tables 2.3 and 2.6 provide the heats of formation or X-Pol energy where appropriate, the dipole moment, and the partial charges on each atom. Tables 2.4 and 2.7 provide the analytical gradients for the dimers with the same coordinates and choices of semiempirical theory. In each case, full QM calculations were done for the AM1 and PMOw Hamiltonians along with their corresponding X-Pol calculations, both with Mulliken charges and DPPC charges. No dispersion energy E_{AB}^{XD} was used in any of the calculations. The values in the tables can be checked against the results of a web-based program linked to in the supporting information section.

2.7 Conclusion

We have described the theoretical background needed for implementing the X-Pol method into molecular dynamics packages. In addition to background information, we have provided the expressions for the analytical first derivative of the variational version of X-Pol and the DPPC population analysis. Sample data has been given for testing new implementations, should the reader be inclined to write one. We believe



Atom	x	y	z
O1$_x$	0.00000	0.00000	0.00000
H1$_y$	0.95231	0.00000	0.00000
H2$_z$	-0.28645	0.41092	-0.80818
O2$_x$	3.46074	0.00000	0.00000
H3$_y$	3.78745	-0.89352	0.00000
H4$_z$	3.92180	0.47778	0.68087

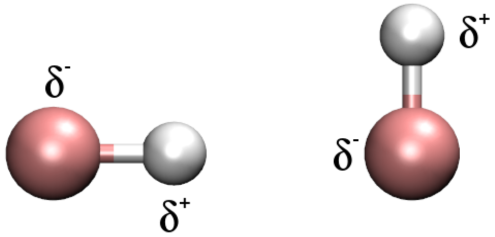
Table 2.2: Cartesian coordinates for the water dimer used in our tests of various combinations of theory, semiempirical method, and population analysis scheme. The dashed line in the table indicates the separation of the fragments for X-Pol calculations.

	AM1	AM1/X/M	AM1/X/D	PMOw	PMOw/X/M	PMOw/X/D
Energy	-119.960	-118.917	-119.822	-140.676	-138.505	-139.928
Dipole	2.476	2.470	2.488	2.599	2.491	2.574
q_{O_1}	-0.40529	-0.40651	-0.69411	-0.31354	-0.32628	-0.70000
q_{H_2}	0.20984	0.21166	0.35624	0.16521	0.18044	0.36848
q_{H_3}	0.19468	0.19485	0.33787	0.14355	0.14583	0.33152
q_{O_2}	-0.40584	-0.40654	-0.69573	-0.33493	-0.33462	-0.70736
q_{H_3}	0.20304	0.20299	0.34744	0.16913	0.16664	0.35292
q_{H_4}	0.20357	0.20355	0.34829	0.17058	0.16798	0.35436

Table 2.3: Heat of formation/X-Pol energy (kcal/mol), dipole moment (Debye), and charges (e) for the water dimer geometry given in Table 2.2 under various combinations of theory, semiempirical method, and population analysis scheme. Calculations are performed without exchange-correlation and dispersion contributions for X-Pol results. "X" denotes the use of X-Pol and "M" and "D" denote the use of Mulliken and DPPC charges respectively.

Atom	AM1	AM1/X/M	AM1/X/D	PMOw	PMOw/X/M	PMOw/X/D
O1$_x$	14.36159	14.64767	14.10543	12.50594	11.34429	12.57168
O1$_y$	9.07860	8.99318	8.95315	5.01112	4.42014	4.96005
O1$_z$	-17.52986	-17.47715	-17.24853	-9.39657	-8.49840	-9.33173
H1$_x$	-10.86555	-9.86415	-10.18750	-9.58385	-7.61349	-9.42117
H1$_y$	-4.11263	-4.13109	-4.15492	-3.35482	-2.87792	-3.46698
H1$_z$	7.74625	7.85911	7.71957	5.86996	5.37791	6.22117
H2$_x$	-6.31026	-6.07657	-6.11781	-5.72054	-5.11807	-6.07247
H2$_y$	-4.95726	-4.91770	-4.88996	-1.79938	-1.65406	-1.71560
H2$_z$	9.62181	9.59901	9.49175	3.47198	3.21093	3.27865
O2$_x$	20.20417	19.41619	20.06744	13.94524	12.70340	15.54289
O2$_y$	-8.93693	-8.99105	-8.70041	-5.41487	-5.02571	-5.23754
O2$_z$	15.41744	15.42969	14.97008	10.01214	9.15550	9.45721
H3$_x$	-8.53955	-8.88609	-8.80590	-5.54104	-5.56129	-6.29024
H3$_y$	5.81383	6.01734	5.88210	3.39648	2.77227	2.80485
H3$_z$	-6.89995	-6.89524	-6.68533	-4.56260	-4.40846	-4.70050
H4$_x$	-8.85040	-9.23705	-9.06165	-5.60574	-5.75484	-6.33068
H4$_y$	3.11439	3.02931	2.91003	2.16147	2.36528	2.65509
H4$_z$	-8.3569	-8.51541	-8.24754	-5.39491	-4.83748	-4.92480

Table 2.4: Analytical first derivatives (in kcal mol^{-1} Å^{-1}) of full semiempirical calculations, X-Pol calculations, and X-Pol calculations with DPPC population analysis for the water dimer geometry given in Table 2.2. The results in this table derive from the use of Eq. 2.30, without exchange-correlation and dispersion contributions for X-Pol results. The dashed line in the table indicates the separation of the fragments for X-Pol calculations. "X" denotes the use of X-Pol and "M" and "D" denote the use of Mulliken and DPPC charges respectively.



Atom	x	y	z
F1$_x$	0.00000	0.00000	0.00000
H1$_y$	0.92500	0.00000	0.00000
F2$_x$	2.75000	0.00000	0.00000
H2$_y$	2.75000	0.92500	0.00000

Table 2.5: Cartesian coordinates for the hydrogen fluoride used in our tests of various combinations of theory, semiempirical method, and population analysis scheme. The dashed line in the table indicates the separation of the fragments for X-Pol calculations.

	AM1	AM1/X/M	AM1/X/D	PMOw	PMOw/X/M	PMOw/X/D
Energy	-136.517	-137.531	-138.169	-160.541	-159.048	-160.699
Dipole	2.688	2.554	2.560	2.875	2.670	2.689
q_{F_1}	-0.30148	-0.29029	-0.40820	-0.19899	-0.20767	-0.42591
q_{H_2}	0.28536	0.29029	0.40820	0.17028	0.20767	0.42591
q_{F_2}	-0.27484	-0.28816	-0.40638	-0.19330	-0.20981	-0.43134
q_{H_2}	0.29096	0.28816	0.40638	0.22201	0.20981	0.43134

Table 2.6: Heat of formation/X-Pol energy (kcal/mol), dipole moment (Debye), and charges (e) for the hydrogen fluoride dimer geometry given in Table 2.5 under various combinations of theory, semiempirical method, and population analysis scheme. Calculations are performed without exchange-correlation and dispersion contributions for X-Pol results. "X" denotes the use of X-Pol and "M" and "D" denote the use of Mulliken and DPPC charges respectively.

Atom	AM1	AM1/X/M	AM1/X/D	PMOw	PMOw/X/M	PMOw/X/D
F1$_x$	-106.90225	-112.05861	-112.85115	-0.66265	-9.05080	-9.68167
F1$_y$	-1.79738	-1.35207	-1.90772	-2.55965	-1.35252	-2.80668
F1$_z$	0.00000	0.00000	0.00000	0.00000	-0.00000	0.00000
H1$_x$	118.55425	109.37669	109.06714	3.73792	6.46075	4.24874
H1$_y$	5.84047	3.90783	5.51206	7.83053	3.28746	6.77464
H1$_z$	0.00000	-0.00000	-0.00000	-0.00000	0.00000	-0.00001
F2$_x$	-5.47239	6.37227	8.98755	4.76430	5.05512	10.45495
F2$_y$	-112.58874	-112.87785	-113.71712	-4.13392	-8.98721	-9.56861
F2$_z$	0.00000	0.00000	0.00000	-0.00000	0.00000	0.00001
H2$_x$	-6.17962	-3.69034	-5.20353	-7.83956	-2.46507	-5.02202
H2$_y$	108.54566	110.32209	110.11278	-1.13696	7.05227	5.60065
H2$_z$	-0.00000	-0.00000	0.00000	0.00000	-0.00000	0.00000

Table 2.7: Analytical first derivatives (in kcal mol^{-1} Å^{-1}) of full semiempirical calculations, X-Pol calculations, and X-Pol calculations with DPPC population analysis for the hydrogen fluoride dimer geometry given in Table 2.5. The results in this table derive from the use of Eq. 2.30, without exchange-correlation and dispersion contributions for X-Pol results. The dashed line in the table indicates the separation of the fragments for X-Pol calculations. "X" denotes the use of X-Pol and "M" and "D" denote the use of Mulliken and DPPC charges respectively.

that the ability to use any desired level of QM theory, its formally linear scaling in the number of fragments, and its more physically-rigorous nature make X-Pol an ideal generalized framework for polarizable force fields.[4]

2.8 Supporting Information

A web-based interface for single-point energy, gradient, and partial charge calculations of NDDO-type semiempirical methods (AM1, AM1-D, MNDO, PM3, PM3-D, PM6, PMO, PMOw, RM1) and the X-Pol method using either Mulliken or DPPC charges is available for free use at `http://mazack.org/cgi-bin/xpol.pl` (accessed on January 30th, 2014).

[4] We would like to thank Peng Zhang for helpful discussions regarding the details of the DPPC method and its gradient. This work has been partially supported by National Institutes of Health grants GM46376 and RC1-GM091445.

Chapter 3

Quantum Mechanical Force Field for Water

This chapter is a result of collaborative efforts between the author, J. Han, and P. Zhang.

3.1 Introduction

Critical to the success of dynamical simulations of chemical and biological systems is the potential energy function used to describe intermolecular interactions. [99,100] Because of the importance of aqueous solution and its unique roles in biomolecular interactions, water has been a subject of extensive and continuous investigation (a review in 2002 included a partial list of 46 water models, [101] while at least two dozen new models have appeared since that time). [102,103] An accurate and efficient model for liquid water also serves as an anchor for developing force fields for proteins, nucleic acids, and carbohydrates. Traditionally, the Lifson-type of effective, pairwise potentials have been used, [99,100,104] and much effort has also been devoted to incorporating many-body polarization effects into such force fields. [105] However, unlike the development of pairwise potentials, there is a great deal of uncertainty in the treatment of polarization effects, both in the choice of the functional form and in the associated parameters.

This is reflected in the fact that simple point charge models such as SPC, [106] TIP3P and TIP4P [107] quickly emerged as the standards in the 1980s for biomolecular force fields, but no standard model for water has emerged although dozens of polarizable potentials for water have been proposed. [2, 101, 102] We have developed a quantum mechanical framework in which each individual molecular fragment is treated by electronic structure theory. [32, 35, 47, 49] Since polarization effects are naturally included in the self-consistent field (SCF) optimization of molecular wave functions, we call this method the explicit polarization (X-Pol) theory. [32,33,35] Recent studies demonstrated the feasibility of X-Pol as a next generation force field for biomolecular simulations, [49] and encouraging results have been obtained using standard semiempirical Hamiltonians. [47, 48] In this chapter we describe a novel model for water, called XP3P, based on X-Pol theory and a three-point charge representation of the electrostatic potential, as a first step in our effort to develop a full quantum mechanical X-Pol force field for biomolecular and materials simulations.

The present quantum mechanical force field (QMFF) may be compared with phenomenological representations of electronic polarization in three commonly used methods in molecular mechanics, namely induced-dipole, Drude-oscillator, and fluctuating-charge models. In the induced-dipole approach, [108–111] atomic polarizabilities are assigned to the interaction sites, typically located on, but not limited to, atomic centers, from which induced point dipoles, representing the total electric field of the system, are obtained. [112] A commonly used method to assign atomic polarizabilities is the dipole interaction model (DIM) popularized by Applequist *et al.* [113] and extended by Thole [114] to incorporate short-range damping functions. Remarkably, the values optimized in DIM are quite transferable, [115] requiring typically one parameter per element. The Drude-oscillator model may be considered as a point-charge equivalent of the induced-dipole method. [116,117] Here, one or a set of point charges are harmonically linked to a polarizable site, in which the directions and distances of the Drude oscillators give rise to the corresponding induced dipole moments. The fluctuating-charge [118–121] approach employs a chemical potential equalization scheme, in which the instantaneous partial charges minimize the energy of the system. The fundamental parameters used in the fluctuating-charge model correspond to the atomic electronegativity and hardness that are rigorously defined in density functional theory. [122]

Each of these classical methods has its advantages and shortcomings in practice. In the fluctuating-charge model, unphysical charge transfer effects between distant monomers can occur. Thus, charge constraints are required. On the other hand, the induced-dipole and the Drude-oscillator model are difficult to use for representing molecular polarization involving a significant charge delocalization such as that across a conjugated polyene chain and the polarization of push-pull compounds (e.g., the crystal of *p*-nitroaniline). The Drude-oscillator model has the advantage of simplicity in practice since any dynamics simulation code can be conveniently adapted to treat polarization effects by that method.

The X-Pol method relies on the partition of a large, condensed-phase system into molecular or submolecular fragments (or blocks), [32, 33, 47] which can be single solvent molecules like water, amino acid residues or nucleotide bases, small ions or enzyme cofactors, or a collection of these small units. The wave function of each molecular fragment is described by a Slater determinant of block-localized molecular orbitals that are expanded over basis functions located on atoms of the fragment. The total molecular wave function is approximated as a Hartree product of these fragmental, determinant functions. Consequently, Coulombic interactions between different fragments are naturally incorporated into the Hamiltonian, but short-range exchange repulsion, charge delocalization (also called charge transfer) and long-range dispersion interactions are not explicitly treated in the quantum chemical formalism. [84,123,124] These effects are included and optimized empirically to strive for accuracy (and efficiency) in X-Pol in the same spirit as that in force field development. The determinantal wave function for each monomer fragment can be approximated by wave function theory (WFT) at either an *ab initio* or a semiempirical level, [47,87] the density may be approximated by density functional theory (DFT), [87,125] or one can combine levels of theory, [126] but in this study we use only semiempirical wave function theory. Although the present work involves only water, we note that the X-Pol theory can be used to model electronic polarization involving conjugated systems and significant charge delocalization contributions, [127] and the X-Pol model is also a reactive force field for modeling systems involving bond-forming and bond-breaking processes.

Semiempirical methods employing neglect diatomic differential overlap (NDDO) [128] are especially suited for QMFF development because of their computational efficiency. However, most such semiempirical models were not optimized to describe intermolecular interactions that are essential for modeling condensed-phase systems. [90,129–131] Part of this problem has been remedied through the incorporation of empirically damped dispersion functions. [52,132–136] Another important deficiency of many semiempirical models for treating non-bonded interactions is that molecular polarization is systematically underestimated. Recently, a polarized molecular orbital (PMO) has been introduced as an alternative, [52,93,94] in which a set of p-orbitals are added to each hydrogen atom. [137] It was found that the computed molecular polarizabilities for a range of compounds containing hydrogen, carbon, and oxygen are significantly improved. [52,94] Employing this strategy, we report here a parametrization of the PMO model for water (PMOw), which can be used in X-Pol for liquid simulations.

In the following, Sec. 3.2 summarizes the PMO parameterization for water and the development of the XP3P model liquid water. Computational details are given in Sec.

3.3. In Sec. 3.4, we present results and discussion. Sec. 3.5 concludes the chapter with a summary of major findings.

3.2 Method

The X-Pol quantum mechanical force field is designed to model condensed phase systems with or without bond-forming and bond-breaking processes. Thus, the X-Pol method can be used as a general-purpose force field in dynamics simulations of solvated proteins or as a reactive force field to model chemical reactions in solutions and in enzymes. In this section, we first describe the quantum chemical model designated as PMOw for water and compounds containing oxygen and hydrogen atoms. The acronym PMO is used to describe the general semiempirical model in which, in addition to a minimal basis set, a set of p-orbitals is added to hydrogen atoms. [52,94] Then, we highlight its incorporation in X-Pol, called the XP3P model, for simulation of liquid water.

3.2.1 Polarized molecular orbital model for water

The PMOw model is a new parameterization of the PMO method, [52] which is based on the MNDO formalism [89] with three key enhancements. First, a set of diffuse p-type basis functions is added on the hydrogen atoms. [93] This greatly improves the quality of the computed molecular polarizabilities and hence the treatment of hydrogen bonding interactions. Second, a damped dispersion function, following the work of Tang and Toennies in wave function theory [138] and Grimme in density functional theory, [139] is included as a post-SCF correction to the electronic energy. In the present implementation, we have adopted the method and parameters proposed by Hillier and co-workers in the PM3-D method. [132–134] The inclusion of the damped dispersion terms further improves the description of intermolecular interactions and the performance of PMO on small molecular clusters. [52, 94, 132–135] Third, the PMOw model is parameterized for general applications to a specific class of compounds (see Sec. 3.4.1 for the set of parametrization data), and the optimization targets include molecular polarizabilities and non-bonded interactions as well as other properties used in the traditional semiempirical parameterization. [52] The parameters presented here are optimized for compounds containing oxygen and hydrogen atoms, especially for studying liquids, aqueous solutions, and proton transport. We note here that, in the same way that atoms are assigned types in molecular mechanics, the parameters for oxygen and hydrogen atoms in functional groups other than water (e.g., peptide bonds) need not be restricted to the same as used for such atoms in water. This departs from the philosophy that has usually been used in semiempirical methods, [140, 141] in which general atomic parameters are used for all functionalities.

In the MNDO formalism, [89, 142] there are 12 atomic parameters for each element, and the PMOw values for water and other compounds containing oxygen and hydrogen are listed in Table 3.1. These values are similar in many respects to the PMOv1 model introduced previously, [52] but they result from a new parametrization presented below. Three exceptions were made to the MNDO functional forms because of the addition of diffuse p basis functions on hydrogen atoms, [52] and they are listed as follows:

1. For the resonance integral involving p orbitals on hydrogen, the following conventions are used:

$$\beta_{lp}^{\mathrm{HH}} = 0, \tag{3.1}$$

$$\beta_{lp}^{\mathrm{OH}} = \frac{\beta_l^{\mathrm{O}} + \beta_p^{\mathrm{H}}}{2} S_{lp} A_{lp} e^{\kappa_{lp} R_{\mathrm{OH}}}, \tag{3.2}$$

where l is the angular momentum quantum number, having the values of 0 (s orbital) and 1 (p orbital), and the subscript p denotes a p-orbital on hydrogen. Notice that Eq. 3.2 is slightly different from the expression used in Ref. [52], in which the exponential function is absent. In Eq. 3.2, β_l^{O} and β_p^{H} are standard MNDO-type parameters, A_{lp} and κ_{lp} are additional parameters introduced in PMO, and R_{OH} is the distance between oxygen and hydrogen atoms. S_{lp} in Eq. 3.2 is an overlap integral ($\langle \mathrm{O}_l | \mathrm{H}_p \rangle$) between oxygen and hydrogen Slater-type orbitals using the parameters listed in Table 3.1, but specific exponents, ζ_{OO} and ζ_{HH}, are used for H–H and O–O pairs, respectively, in PMOw.

2. In standard MNDO, [89, 142] the nucleus-electron attraction integral, $H_{\mu\nu}^{A}$, between electronic charge density on atom A and nucleus B is evaluated on the basis of the two-electron repulsion integral, $\langle \mu_A \nu_A | s_B s_B \rangle$. [91] In PMOw, if both A and B are hydrogen atoms, for a distribution of p orbitals (pp'), this is screened as follows:

$$H_{pp'}^{\mathrm{H}} = \left[1 - B e^{-\lambda R_{\mathrm{HH'}}^2}\right] \left(H_{pp'}^{\mathrm{H}}\right)_{\mathrm{MNDO}} \tag{3.3}$$

3. For the homonuclear core-core repulsion integrals between oxygen-oxygen and a hydrogen-hydrogen atom pairs, [52, 89, 142] the standard values for α^{O} and α^{H} are replaced by $\hat{\alpha}^{\mathrm{O}}$ and $\hat{\alpha}^{\mathrm{H}}$. Note that α^{O} and α^{H} are used as in standard MNDO for core-core repulsion integrals between oxygen and hydrogen atoms.

The parameters in the standard MNDO formalism [89] (Table 3.1) and the additional parameters (Table 3.2) described above were adjusted by iterative optimization using a genetic algorithm for some of the systems and properties listed in Table S1 in the supporting information. In comparison with the results in Ref. [52], the present

	H	O
U_{ss} (eV)	-11.15043	-111.86028
U_{pp} (eV)	-7.35459	-78.64105
β_s (eV)	-6.88125	-25.57063
β_p (eV)	-3.52628	-31.90404
ζ_s (bohr^{-1})	1.17236	3.05303
ζ_p (bohr^{-1})	1.05333	3.12265
α (Å^{-1})	3.05440	3.76880
g_{ss} (eV)	12.73667	17.36659
g_{sp} (eV)	8.04688	13.37288
g_{pp} (eV)	6.98401	14.78196
$g_{pp'}$ (eV)	10.65161	13.49319
h_{sp} (eV)	1.92149	4.42643

Table 3.1: Semiempirical parameters for H and O Atoms in the PMOw model. The derived parameter, h_{pp}, is determined from g_{pp} and $g_{pp'}$ and has been set to a minimum value of 0.1 eV as implemented in the MOPAC program, $h_{pp} = \max\{0.1\text{eV}, (g_{pp} - g_{pp'})/2\}$.

Parameter	Value
A_{sp}	0.03000
A_{pp}	0.15000
B	1.00000
κ_{sp} (Å^{-1})	0.47069
κ_{pp} (Å^{-1})	0.47069
λ (Å^{-2})	1.10000
$\hat{\alpha}^{\mathrm{H}}$ (Å^{-1})	2.52552
$\hat{\alpha}^{\mathrm{O}}$ (Å^{-1})	3.03253
ζ_{HH} (bohr^{-1})	1.28000
ζ_{OO} (bohr^{-1})	2.76400
σ_{H} (Å)	0.800
σ_{O} (Å)	3.225
ϵ_{H} (kcal/mol)	0.05
ϵ_{O} (kcal/mol)	0.15

Table 3.2: Additional semiempirical parameters for oxygen and hydrogen in the polarized molecular orbital model and the Lennard-Jones parameters in explicit polarization model for liquid water.

		PMOw	XP3P	AMOEBA	POL5.TZ	*Ab initio*	Expt. [143]
H_2O	AE (kcal/mol)	233.0	233.0			229.3 [144]	232.2
	IP (eV)	13.20	13.20			12.42	12.68
	r (Å)	0.955	0.957	0.957	0.957	0.9589 [145]	0.9572
	θ (°)	104.6	104.5	108.5	104.5	104.16 [145]	104.52
	α (Å^3)	1.27	1.27	1.41	1.29	1.45 [146]	1.45
	q^{H} (e)	0.16	0.34	0.26		0.35	N/A
	q^{O} (e)	-0.31	-0.67	-0.52		-0.70	N/A
	μ (Debye)	1.88	1.88	1.77	1.85	1.84 [146]	1.86 [147]
$(H_2O)_2$	ΔE_b	-5.1	-5.2	-4.96	-4.96	-5.0 [3]	-5.44
	R_{OO}	2.89	2.90	2.89	2.90	2.92	2.98
	α	6.2	1.3	4.2	4.7	4.8	-1 ± 10
	ϕ	115	165	123	117	125	123 ± 10
	$\langle \mu_{\mathrm{mol}} \rangle$	2.10	2.16				
	μ	2.39	3.85	2.54	2.44	2.65	2.64

Table 3.3: Computed equilibrium properties for water monomer and dimer from different polarizable water models and *ab initio* MP2/(CBS) with CCSD(T) corrections along with experimental data.

parameter set further improves the calculated molecular polarizability and dipole moment of water in the gas phase as well as the binding energy and dipole moment of water dimer (Table 3.3).

3.2.2 Explicit polarization theory

In X-Pol, [32,33,47] the system is partitioned into molecular or submolecular fragments, in which the total wave function of the system is assumed to be a Hartree product of the determinant wave functions of the individual fragments. In the present case, each fragment is simply a single water molecule, and the overall wave function of the system is

$$\Phi = \prod_{a=1}^{N} \Psi_a, \tag{3.4}$$

where N is the number of fragments in the system, and Ψ_a is a Slater determinant of doubly-occupied molecular orbitals (MOs) block-localized on molecule (fragment) a.

The approximation of Eq. 3.4 implies neglect of the short-range exchange repulsion [123] and long-range dispersion interactions [148] between different fragments, which are corrected empirically below. [32, 33, 47] Use of Eq. 3.4 reduces the computational costs, allowing molecular dynamics and Monte Carlo simulations to be carried out for large systems efficiently with sufficient sampling. [47, 49]

The effective Hamiltonian of the system is given by

$$H = \sum_{a=1}^{N} H_a^o + \frac{1}{2} \sum_{a=1}^{N} \sum_{b \neq a} H_{ab}, \tag{3.5}$$

where H_a^o is the electronic Hamiltonian of fragment a in the gas phase and H_{ab} represents the effective interactions between molecules a and b:

$$H_{ab}(\rho_b) = -\sum_{i=1}^{M} V_i(\rho_b) + \sum_{A \neq 1}^{Q} Z_A^a V_A(\rho_b) + E_{ab}^{\mathrm{XD}}, \tag{3.6}$$

where M is the number of electrons and Q is the number of atoms in fragment a, Z_A^a would be the nuclear charge of atom A of fragment a if all electrons were treated explicitly, but here it is the core charge since $1s$ electrons of oxygen atoms are in the core, and E_{ab}^{XD} is the exchange-dispersion, correlation energy. The electrostatic potential $V_x(\rho_b)$, either at the electronic ($x = i$) or at the nuclear ($x = A$) position, due to the instantaneous charge density of fragment b is given by

$$V_x(\rho_b) = -\int \frac{\rho_b(\mathbf{r})\mathrm{d}\mathbf{r}}{|\mathbf{r}_x - \mathbf{r}|} + \sum_{B=1}^{Q} \frac{Z_B^b}{|\mathbf{r}_x - \mathbf{R}_B^b|}. \tag{3.7}$$

Here, $\rho_b(\mathbf{r})$ is the electron density of fragment b derived from the corresponding wave function Ψ_b (or Kohn–Sham Slater determinant), [32, 47] and R_B^b denotes the nuclear coordinates.

We define the total interaction energy of a condensed phase system by

$$E_{\text{tot}} = \langle \Phi | H | \Phi \rangle - \sum_{a=1}^{N} \langle \Psi_a^o | H | \Psi_a^o \rangle. \tag{3.8}$$

The energy defined in Eq. 3.8 corresponds to the total energy of the condensed-phase system relative to that of infinitively separated fragments. Since all molecules are identical in pure liquid water in the present study, the last summation term in Eq. 3.8 is simply NE_a^o with $E_a^o = \langle \Psi_a^o | H_a^o | \Psi_a^o \rangle$ being the energy of an isolated monomer. It is often useful for interpretive purposes to consider the dimeric interaction energies between two fragments even for a potential that includes many-body polarization effects as in the present X-Pol potential. To this end, we define the interaction energy between fragments a and b by [47]

$$E_{ab} = \frac{1}{2} \left(\langle \Psi_a | H_{ab} | \Psi_a \rangle - \langle \Psi_b | H_{ba} | \Psi_b \rangle \right). \tag{3.9}$$

The two terms in Eq. 3.9 corresponds to a embedding in b and b embedding in a, respectively, both in the presence of the rest of the system, and they are not always numerically equivalent in practice [32] even though they describe the same intermolecular interactions. The definition of Eq. 3.9 ensures that $E_{ab} = E_{ba}$.

The exchange-dispersion, correlation energy can be incorporated with an explicit density dependent term and added to the Fock operator as described in the work of York and co-workers. [84, 149] Alternatively, the damped dispersion term that is an intrinsic part of the PMOw model can be used with the addition of a repulsive potential. Here, in the spirit of simplicity for a force field, we adopt a Lennard-Jones potential to approximate the remaining energy contributions [32,33,47] not included in the PMOw electronic structure method. [52] (Thus there are two R^{-6} terms, one in PMOw for intrafragment interactions and one in the Lennard-Jones term associated with interfragment interactions.) The Lennard-Jones term introduces two empirical parameters per atom type:

$$E_{ab}^{\mathrm{XD}} = \sum_{A}^{Q} \sum_{B}^{Q} 4\epsilon_{AB} \left[\left(\frac{\sigma_{AB}}{R_{AB}} \right)^{12} - \left(\frac{\sigma_{AB}}{R_{AB}} \right)^{6} \right], \qquad (3.10)$$

where ϵ_{AB} and σ_{AB} are obtained from the geometric mean of atomic parameters such that $\epsilon_{AB} = (\epsilon_A \epsilon_B)^{1/2}$ and $\sigma_{AB} = (\sigma_A \sigma_B)^{1/2}$. These parameters are also listed in Table 3.2.

3.2.3 The XP3P model for liquid water

The electrostatic potential (ESP) in Eq. 3.7 can be determined explicitly by evaluating the associated one-electron integrals. However, this would have not saved much computational time in SCF calculations since integration of two-electron coordinates is needed, and would have missed the point of developing a fragment-based technique in electronic structure calculations. As we have proposed previously, [32, 33, 87] it is desirable to employ a more computationally efficient method to approximate the external potential $V_x(\Psi_b)$. In the present application to liquid water, we use a simple, three-point-charge approximation to $V_x(\Psi_b)$. Consequently, we call this X-Pol potential with three-point charges for water the XP3P model.

Several methods based on atomic partial charges for approximating the quantum external potential were described originally for the X-Pol potential, [32,47] and some of them were adopted later in other fragment-based molecular orbital models. [150] Although the use of atomic charges obtained from fitting the quantum mechanical $V_x(\Psi_b)$ has been successfully used in several molecular mechanics force fields, [151, 152] it is known that the ESP-fitting method sometimes yields unreasonably large partial charges on structurally buried atoms. [153] In addition, large variations could occur as a result of structural fluctuations to expose buried atoms during a dynamics simulation. A general approach is the multi-center multipole expansion of the quantum mechanical ESP, [154] and this method has been used in the effective fragment potential model; [155] multi-center multipolar representations could also be used with X-Pol. [86] A conceptually simple alternative is to use atomic charges derived from a population analysis such as the Mulliken or Löwdin population method. [156] When used with small, well balanced basis sets, the Mulliken or Löwdin charges can provide a good representation of the relative atomic electronegativity and they are computationally efficient. Scaled Mulliken population charges have been used and shown to be effective in statistical mechanical Monte Carlo simulations of liquid water using an explicit QMFF. [47]

Another way of approximating the external potential for intermolecular interactions is to employ partial atomic charges that are mapped from the density matrix to reproduce experimental dipole moments (in contrast to ESP fitting). This has been called a class IV charge model, and it can be parametrized to show good consistency for a variety of electronic structure methods and basis sets. [157,158] Alternatively, partial atomic charges can be derived to rigorously reproduce the molecular moments to any order of accuracy from a Lagrangian multiplier procedure. Following the method proposed by Thole and van Duijnen [97] and extended by Swart and van Duijnen, [159] we applied the Lagrangian multiplier approach to semiempirical methods, [53] which are known to yield excellent molecular dipole moments in comparison with experiments. In this approach, both the total molecular dipole moment and the local atomic hybridization contributions of the approximate NDDO wave function are reproduced exactly. In the present implementation, we preserve the total and local molecular dipole moments. In addition, we included in the procedure the capability to reproduce experimental molecular polarizability and its atomic decomposition according to the dipole interaction model. [53] We called this method the dipole preserving and polarization consistent (DPPC) charge model. [53]

Specifically, the DPPC charge has two contributions, the Mulliken population charge q_A^{MP} and the residual charges Δq_A^B due to preservation of atomic s and p hybridization dipole moments: [53]

$$q_A^{\text{DPPC}} = q_A^{\text{MP}} + \sum_{B=1}^{Q} \Delta q_A^B, \tag{3.11}$$

where the residual charge Δq_A^B on atom A due to the constraint that the residual moment is identical to the atomic hybridization contribution from atom B:

$$\mu_B^{\text{hyb}} = -(\mathbf{P}_{sp})_B \cdot \mathbf{D}_B = \sum_{A=1}^{Q} \Delta q_A^B R_A, \tag{3.12}$$

where $(\mathbf{P}_{sp})_B$ is a diagonal matrix with the densities $P_{sp_x}^B$, $P_{sp_y}^B$, and $P_{sp_z}^B$, on atom B, $\mathbf{D}_B$ is the corresponding dipole integral, and R_A denotes the coordinates of atom A. The residual charges Δq_A^B that reproduce the hybridization component of molecular dipole moment, μ_B^{hyb}, are predominantly localized on atoms closest to atom B.

Since the molecular dipole moment is determined from

$$\mu_{\text{QM}} = \sum_{A=1}^{Q} q_A^{\text{MP}} R_A + \sum_{B=1}^{Q} \mu_B^{\text{hyb}} \tag{3.13}$$

in semiempirical methods employing the NDDO approximation, [160] it is clear that the atomic charges given in Eq. 3.11 reproduce exactly the full quantum mechanical dipole moment and the local, atomic hybridization contributions:

$$\sum_{A=1}^{Q} q_A^{\mathrm{DPPC}} R_A = \mu_{\mathrm{QM}}. \tag{3.14}$$

The residual charges depend on geometry and atomic electronegativity, and an expression for them was given in Ref. [53]. The advantage of using the DPPC charges over the ESP-fitted ones is that local properties of the dipole integrals are explicitly accounted for and fully utilized to generate the partial atomic charges. The method to generate DPPC charges is applicable both to neutral and ionic molecules, independent of the origin of coordinates. [53]

3.3 Computational Details

The parameterization of the PMOw model was carried out by iterative optimization using a genetic algorithm that has been detailed in Ref. [52]. The PMOv1 set of parameters overestimated the dipole moment of water (2.19 D) and underestimated the interaction energy for the water dimer (4.7 kcal/mol) in comparison with the target values of 1.85 D from experiment [147] and 5.0 kcal/mol from CCSD(T) and MP2/(CBS) calculations. [3] The PMOw parametrization improves these quantities for application to water and its ions.

Statistical mechanical Monte Carlo simulations were performed on a system consisting of 267 water molecules in a cubic box, employing the PMOw Hamiltonian. Based on procedures described previously, [47,48] periodic boundary conditions were used along with the isothermal-isobaric ensemble (NPT) at 1 atm and temperature ranging from -40 to 100°C. As in the development of other empirical potentials including the successful SPC, [106] TIP3P, and TIP4P models [107] and the polarizable AMEOBA, [2] SWM4-NDP [117] and POL5/TZ [161] potentials for water (and many other water models not explicitly compared in this paper), the parameterization was performed only at 25°C. The XP3P model based on the PMOw Hamiltonian has four Lennard-Jones parameters, ϵ_{O}, ϵ_{H}, σ_{O}, and σ_{H}. We have kept the ϵ_{H} and σ_{H} values used in a previous X-Pol simulation of liquid water with the AM1 Hamiltonian (called the MODEL potential for water), and we made small adjustments of the other two values (3.24 Å and 0.16 kcal/mol) [47] to reproduce the liquid density and heat of vaporization within 1% of the experimental values at 25°C. In the parameterization stage, spherical cutoff with a switching function between 8.5 Å and 9.0 Å based on oxygen-oxygen separations was employed, and a long-range correction to the Lennard-Jones potential was included. (The SPC and TIP3P/TIP4P models [106, 107] and later the TIP5P model [6] were also developed using cutoff distances, which were as small as 7.5 Å with a box of 125 or 216 water molecules.) Although it is possible to use Ewald sums to treat long-range electrostatic interactions, [162] we have not used the particle-mesh Ewald implementation in the present Monte Carlo calculation.

In Monte Carlo simulations, new configurations were generated by randomly translating and rotating a randomly selected water molecule within ranges of ±0.13 Å and ±13°. In addition, the volume of the system was changed randomly within the limit of ±150 Å^3 on every 550th attempted move, and the coordinates of oxygen atoms were scaled accordingly. (Note that in the Monte Carlo calculations, the waters are rigid, so the hydrogen positions also adjust when the oxygen positions are adjusted.) These options were slightly adjusted to maintain an acceptance rate of about 45% at each temperature in the Metropolis sampling. In each simulation, at least 5×10^6 configurations were discarded for equilibration, which was followed by an additional 1×10^7 to 1.1×10^8 configurations for averaging. About 6×10^6 configurations can be executed per day on a 6-core Intel Xeon X7542 Westmere processor at 2.66 GHz.

The XP3P model was further examined in molecular dynamics simulations for 500 ps in the NVT ensemble, using the Lowe-Andersen thermostat [163, 164] and a volume fixed at the average value from the Monte Carlo simulation; the number of water molecules in the dynamics simulations was also 267. The monomer geometries were enforced by the SHAKE/RATTLE procedure. [165] Although long-range electrostatic interactions can be computed using the particle-mesh Ewald summation that has been extended for the X-Pol potential, [162, 166] we have used a 9.0Å cutoff in the present study. The velocity Verlet integration algorithm was used with a 1fs time step. The total energy of the system was obtained from fully converged wave functions for each water molecule for each microscopic configuration, although different procedures were utilized in the Monte Carlo sampling and in molecular dynamics simulations. In Monte Carlo, an initial set of DPPC charges, derived from an initial guess of the X-Pol wave function, e.g., that from the previous configuration (with random perturbation to some randomly selected elements in the density matrix), are incorporated into the Fock matrix in terms of one-electron integrals (as in combined QM/MM schemes) in the subsequent iteration step during the self-consistent field (SCF) optimization. Then, a new set of orbital coefficients is obtained to generate updated DPPC charges for the next iteration until the electronic energy is converged to 5×10^{-5} eV for each monomer and to 10^{-5} for the partial atomic charges (in atomic units) between consecutive iterations.

In Monte Carlo simulations, the Fock operator is constructed analogously to a combined QM/MM scheme, [167] which is not fully variational with respect to the change of the charge density; the external potential does incorporate the complete electrostatic effects in a self-consistent manner. [32, 47] The procedure is efficient in Monte Carlo simulations since the electronic integrals are not required from all other molecular fragments, and it does not pose problems because gradients are not needed. This is the method proposed in the original development of the method for Monte Carlo calculations, [32,47] and it was used a few years later in the fragment molecular orbital model of Kitaura and co-workers. [168] For molecular dynamics simulations, a fully variational Fock operator for each monomer was used in which the external potential consists of contributions both from the DPPC charges and the explicit electron densities of all other fragments. [35, 87] Here, analytic gradients can be directly obtained from the optimized X-Pol wave function. In molecular dynamics simulations, the criteria for energy and density conversion were set as 10^{-9} eV for energy and 10^{-6} for density matrix elements. The average energy difference from the two approaches in Monte Carlo and molecular dynamics is less than 1.5% in the computed heat of vaporization.

The Monte Carlo simulations were performed using the MCSOL program for X-Pol simulations, [169] while molecular dynamics simulations were carried out using a newly developed X-Pol program [170] written in C++ which has been interfaced both with CHARMM [50] and NAMD. [51] All *ab initio* electronic structure calculations were performed using GAUSSIAN 09. [171] All calculations were run on a constellation of clusters at the Minnesota Supercomputing Institute.

3.4 Results and Discussion

3.4.1 Gas-phase properties

Properties for the optimized water monomer and dimer using the PMOw and XP3P models are listed in Table 3.3 along with experimental data and the results from two empirical polarizable potentials that have been examined by Ren and Ponder. [2] The PMOw parameters were optimized against experimental or high-level *ab initio* data for a series of small molecules containing hydrogen and oxygen atoms (supporting information), including the properties listed in Table 3.3. In particular, the computed atomization energy (233.0 kcal/mol) and dipole moment (1.88 D) for water from PMOw agree with the corresponding experimental data that have been summarized in Ref. [52] (232.6 kcal/mol and 1.85 D, respectively). The Mulliken population charges from the PMOw wave function and the DPPC charges used in the XP3P potential are also listed in Table 3.3; the latter yields exactly the same molecular dipole moment as that from the QM calculation. An important quantity critical to describing hydrogen-bonding interactions is the molecular polarizability, which also shows good agreement with experiment (a deviation of 14%). This represents a major improvement over all previous NDDO-based models, which typically have errors more than 60% for water. Nevertheless, a question arises on whether or not the somewhat smaller polarizability would affect liquid properties. To address this issue, it is interesting to consider polarizable potential functions for water, in which the experimental gas-phase electrostatic properties are not always enforced. This is illustrated by the use of smaller molecular polarizabilities in these empirical force fields, and this was justified as to reflect the relatively larger electric field than the mean field of the bulk due to the highly inhomogeneous environment in the first solvation shell; [172] for example, polarizabilities are set to 1.41, 1.29, and 0.98 $Å^3$ in the AMOEBA, [2] POL5/TZ [2], [161] and SWM4-NDP [117] models, respectively, all of which yield similar heats of vaporization and similar densities of liquid water at ambient conditions.

The optimized bond length and bond angle for water are 0.9552 Å and 104.61° using PMOw; these values are in excellent agreement with the experimental values of 0.9572 Å and 104.54° [173] Thus, either the optimized or the experimental monomer geometry can be used in the XP3P potential for liquid simulations discussed below. The change of the molecular dipole moment with geometry variation for the water monomer has an intriguing nonlinear dependence, which is not correctly reproduced in nearly all polarizable and non-polarizable potentials for water, except the TTM2-F model [174] that was specifically fitted with a function to reproduce an accurate *ab initio*

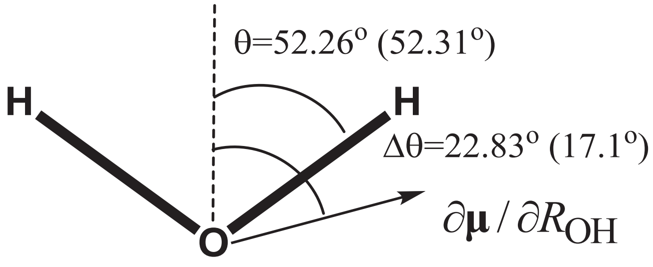


Figure 3.1: Illustration of the angle between the molecular dipole moment derivative and the O–H bond vector in water monomer. Experimental values are given first, followed by the PMOw results in parentheses.

dipole moment surface. [175] This is illustrated in Figure 3.1, which shows that the dipole derivative with respect to an O–H stretch, $\partial\mu/\partial R_{\text{OH}}$, lies significantly outside of the two O–H bonds of water. An angle of $\Delta\theta = 22.8^\circ$ was obtained based on the vibrational absorption intensities. [174, 176, 177] For comparison, the present PMOw model yields a value of $\Delta\theta = 17.1^\circ$, in reasonable agreement with experiment. This is encouraging since this information was not included in the PMOw parametrization process; it is purely a result of the qualitatively correct treatment of chemical bonding interactions in the present quantum mechanical model.

The potential energy profile for the water dimer along the O–O separation is illustrated in Figure 3.2, and the computed binding energies from PMOw and the XP3P potential are -5.1 and -5.2 kcal/mol, respectively, slightly greater than the best estimate of -5.0 kcal/mol from *ab initio* calculations using MP2/(CBS)+ CCSD(T) with the 6-311++G(d,p) basis set, [3] but somewhat smaller than an estimated value (-5.4 kcal/mol) based on measured molecular vibrations. [178] For comparison, both the POL5/TZ [161] model and the AMOEBA model yield a binding energy of -5.0 kcal/mol. [2] The equilibrium structures optimized using the full PMOw Hamiltonian and the fragmental XP3P potential are listed in Table 3.3. [5, 143–146] The O–O distances from the PMOw and XP3P models agree well with those from POL5/TZ and AMOEBA,

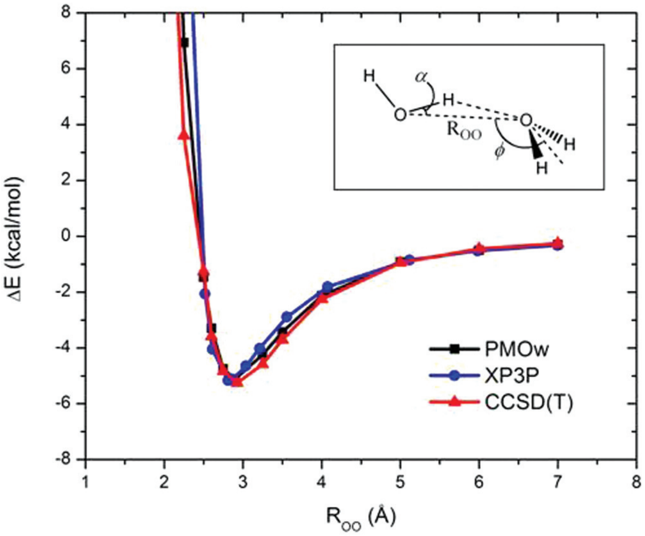
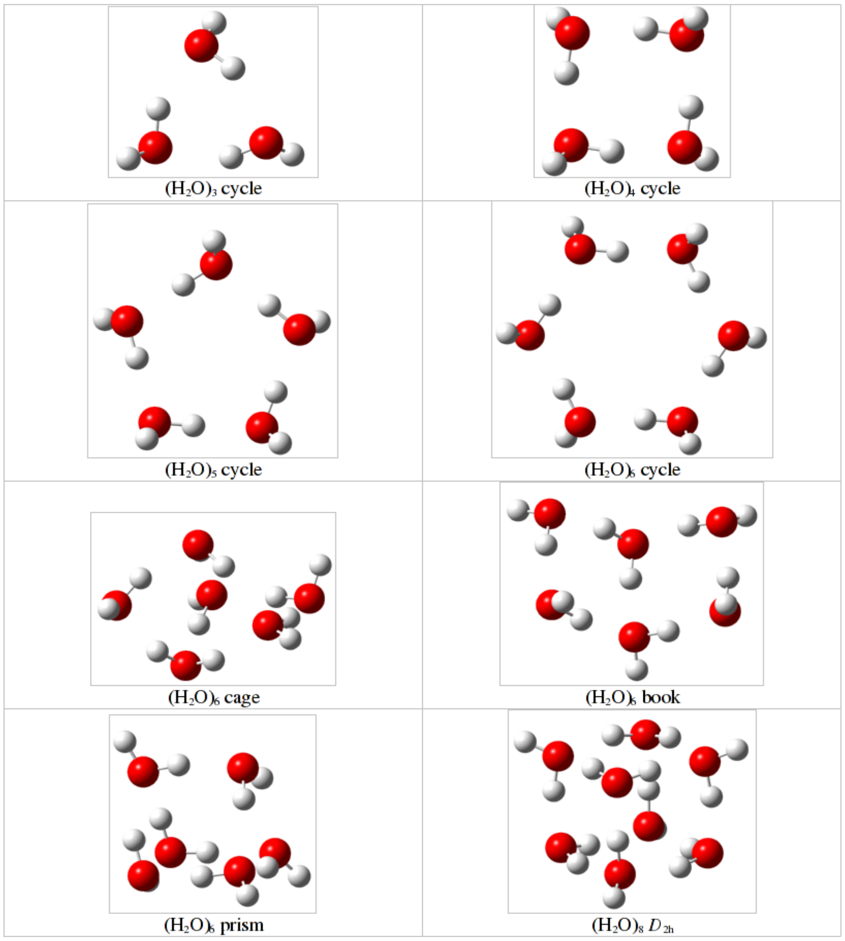


Figure 3.2: Potential energy profiles for a water dimer at the hydrogen bonding configuration from the PMOw (black) and the XP3P (blue) models for water along with CCSD(T) results (red). Definition of the geometrical parameters listed in Table 3.3 are given in the structure shown as inset in the upper right-hand corner. The CCSD(T) results are obtained with the aug-cc-pVDZ basis set on fully optimized geometries at various fixed O–O distances. Studies have shown that extrapolation to the complete basis set limit from the current size does not affect the computed energies by more than 0.2 kcal/mol. [5] All other geometric parameters are optimized.

which yield 2.89 Å and with the *ab initio* value of 2.91 Å. [3] Ren and Ponder found that the flap angle ϕ (the flap angle is defined as the angle between the C_2 axis of the hydrogen bond acceptor monomer and the O–O distance vector, depicted in the inset of Figure 3.2) is dependent on the monomer quadrupole moment, and that it was necessary to use explicit quadrupole terms in the AMOEBA model to yield a flap angle in agreement with the *ab initio* results. The results on the flap angle in the water dimer from the PMOw and XP3P models are also good, and the small tilt angle, α, from the hydrogen bond donor is also predicted. However, the large flap angle is not preserved in the XP3P model. The structures and energies on other stationary points of water dimer are given in the supporting information.

Small water clusters (Figures 3.3 and 3.4), including the cyclic configurations of the trimer, tetramer, and pentamer, four configurations of the hexamer, and the cubic D_{2h} arrangement of the octamer have been examined (Table 3.4). All clusters were fully optimized with PMOw using the conjugated gradient method with NAMD. [51,170] A configuration was considered optimized when its gradient norm fell below 0.5 kcal mol^{-1} Å^{-1}. The best theoretical estimates for these systems are from the work of Bryantsev *et al.*, who performed single-point MP2/(CBS) along with a CCSD(T) correction (simply called CCSD(T) results in this discussion) at the B3LYP/6-311++G(2d,2p) optimized structures. [3] As in the work of Ren and Ponder, [2] we list in Table 3.4 the total binding energies, the average O–O distances (R_{OO}), average O. . .H–O hydrogen bond angles ($\langle \phi \rangle$), and the total (μ) and average monomer ($\langle \mu_{\text{mol}} \rangle$) dipole moments. Of all water clusters, the average monomer dipole moments from the POL5/TZ and AMOEBA models [2] fall between the values computed using the PMOw and the XP3P method, and the trends are in accord with that estimated by Gregory *et al.* [4] using a portioning scheme for the electron density. Overall, the computed binding energies from PMOw and XP3P methods are in good agreement with the CCSD(T) results, with root-mean-square (RMS) deviations of 1.2 and 2.4 kcal/mol, respectively. The performance of the AMOEBA force field is excellent, whereas the POL5/TZ model slightly underestimates the binding energies. [2,161] For the hexamers, the ordering of relative stability is cage $>$ book $>$ prism $>$ cyclic from CCSD(T), and cage $=$ prism $>$ book $>$ cyclic from PMOw. For comparison, the ordering from the MP2/CBS+CCSD(T) calculations with 6-311++G(d,p) basis [3] and AMOEBA optimizations is prism $>$ cage $>$ book $>$ cyclic. [2] In any event, the three non-cyclic structures of the water hexamer are energetically similar in binding, whereas the cyclic configuration is noticeably less stable than the other three.

We have also examined several configurations of micro-solvated proton $H^+(H_2O)_n$,



$(H_2O)_3$ cycle

$(H_2O)_4$ cycle

$(H_2O)_5$ cycle

$(H_2O)_6$ cycle

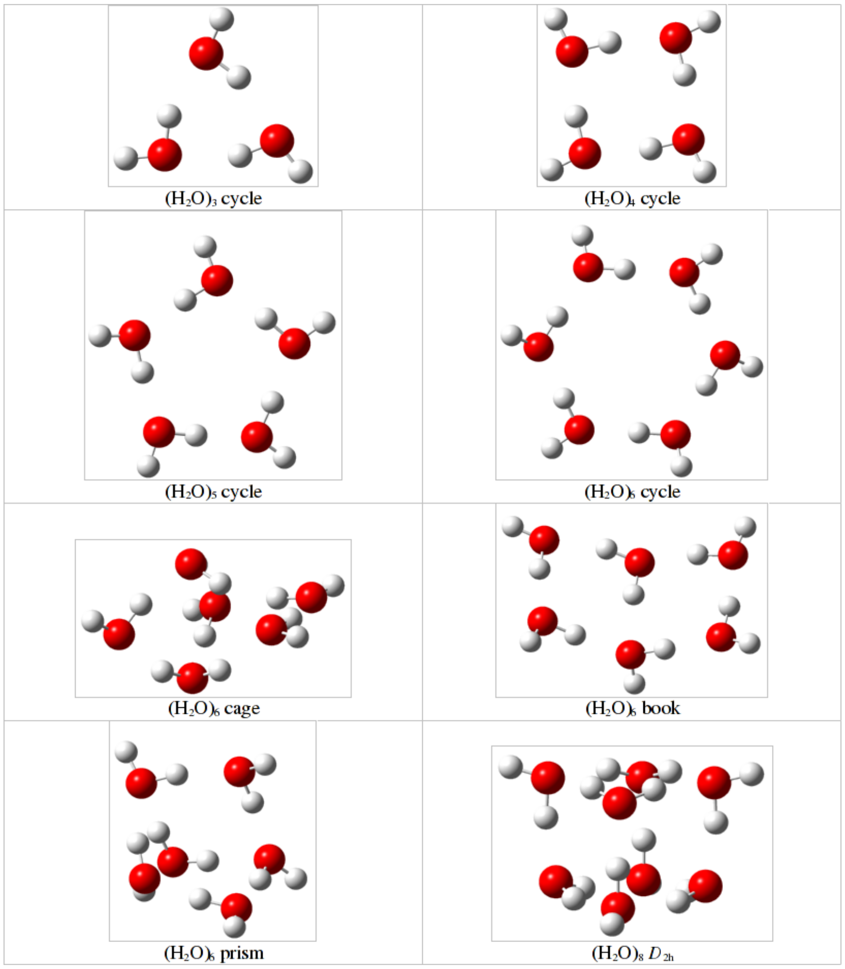
$(H_2O)_6$ cage

$(H_2O)_6$ book

$(H_2O)_6$ prism

$(H_2O)_8$ D_{2h}

Figure 3.3: Optimized water clusters with PMOw



$(H_2O)_3$ cycle

$(H_2O)_4$ cycle

$(H_2O)_5$ cycle

$(H_2O)_6$ cycle

$(H_2O)_6$ cage

$(H_2O)_6$ book

$(H_2O)_6$ prism

$(H_2O)_8$ D_{2h}

Figure 3.4: Optimized water clusters with XP3P

		PMOw	XP3P	POL5/TZ^a	AMOEBA^a	*Ab initio*^{b,c}	Expt.
Trimer cyclic	ΔE_b	-14.8	-15.7	-13.4	-15.3	-15.8	
	$\langle R_{\mathrm{OO}} \rangle$	2.87	2.77	2.90	2.81	2.81	2.845
	$\langle \phi \rangle$	105.1	125.6		151.5	110.4	152
	$\langle \mu_{\mathrm{mol}} \rangle$	2.14	2.46	2.22	2.29	2.3	
	μ	1.19	0.01	1.21	1.09	1.07	
Tetramer cyclic	ΔE_b	-27.5	-28.9	25.5	27.7	-27.4	
	$\langle R_{\mathrm{OO}} \rangle$	2.74	2.68	2.769	2.76	2.75	2.79
	$\langle \phi \rangle$	116.5	145.9		168.0	121.6	
	$\langle \mu_{\mathrm{mol}} \rangle$	2.22	2.71	2.47	2.55	2.6	
	μ	0.0	0.0	0.0	0.0	0.0	
Pentamer cyclic	ΔE_b	-35.7	-39.7	34.1	36.5	-35.9	
	$\langle R_{\mathrm{OO}} \rangle$	2.73	2.66	2.74	2.76	2.73	2.76
	$\langle \phi \rangle$	126	159		176	132	
	$\langle \mu_{\mathrm{mol}} \rangle$	2.26	2.82	2.57	2.64	2.7	
	μ	1.17	0.02	1.19	0.92	0.93	
Hexamer cyclic	ΔE_b	-43.3	-49.0	41.8	44.8	-44.3	
	$\langle R_{\mathrm{OO}} \rangle$	2.72	2.65	2.74	2.75	2.72	2.76
	$\langle \phi \rangle$	130	167		179	139	
	$\langle \mu_{\mathrm{mol}} \rangle$	2.28	2.86	2.62	2.70	2.7	
	μ	0.0	0.0	0.02	0.0		
Hexamer prism	ΔE_b	-47.8	-44.4	41.9	45.9	-45.3	
	$\langle R_{\mathrm{OO}} \rangle$	2.84	2.76	2.79	2.80	2.86	
	$\langle \phi \rangle$	121.0	128.7			123.1	
	$\langle \mu_{\mathrm{mol}} \rangle$	2.24	2.72	2.52	2.60		
	μ	2.40	3.29	2.91	2.57	2.70	
Hexamer cage	ΔE_b	-47.8	-45.2	41.8	45.9	-46.0	
	$\langle R_{\mathrm{OO}} \rangle$	2.80	2.76	2.78	2.80	2.83	2.82
	$\langle \phi \rangle$	118	126			121	
	$\langle \mu_{\mathrm{mol}} \rangle$	2.22	2.72	2.49	2.58	2.6	
	μ	2.05	2.01	2.44	2.16	1.90	1.82–2.07^c
Hexamer book	ΔE_b	-46.2	-48.3	42.5	45.8	-45.8	
	$\langle R_{\mathrm{OO}} \rangle$	2.75	2.70	2.79	2.78	2.78	
	$\langle \phi \rangle$	121	144			127	
	$\langle \mu_{\mathrm{mol}} \rangle$	2.24	2.79	2.55	2.63		
	μ	2.40	2.22	2.45	2.29		
Octamer	ΔE_b	-77.7	-69.5			-72.64^d	
	$\langle R_{\mathrm{OO}} \rangle$	2.74	2.72			2.81	
	$\langle \phi \rangle$	163	164			163	
	$\langle \mu_{\mathrm{mol}} \rangle$	2.20	2.86				
	μ	0.0	0.0			0.0	

Table 3.4: Computed and experimental properties for water clusters. The angle $\langle \phi \rangle$ is the average O…H–O angle of the hydrogen bonds in a given cluster. (a) Ref. [2]. (b) [3]. (c) [4]. (d) MP2/(CBS) limit [5].

where $n = 2, 3, 4$, and 6 (Figure 3.5). Depicted in Figure 3.6 are the potential energy profile for a proton migration between two water molecules at fixed O–O distances of the global minimum $R_{\text{min}}(\text{OO})$, $R_{\text{min}}(\text{OO})$ + 0.2 Å, and $R_{\text{min}}(\text{OO})$ + 0.4 Å from PMOw, MP2/aug-cc-pVDZ, B3LYP/aug-cc-pVTZ, and M06-2X/aug-cc-pVTZ optimizations. The equilibrium structure has an $R_{\text{min}}(\text{OO})$ separation of 2.46, 2.40, 2.41, and 2.39 Å, respectively, from these theoretical models. With a basis set comparable to aug-cc-pVDZ, the MP2 results on these proton clusters are very close to CCSD(T)-F12 results with jun-cc-pVTZ basis. [179] The PMOw O–O distance is about 0.05 Å longer than the MP2 result, while DFT values are in close agreement with MP2. In all cases, the proton is essentially symmetrically located between the two water molecules (Figure 3.6a). A small barrier appears when the O–O distance is stretched by 0.2 Å. The PMOw model yields a barrier of 1.9 kcal/mol, compared to 1.9, 1.4, and 1.3 kcal/mol from MP2, B3LYP, and M06-2X. Further stretching the O–O distance to $R_{\text{min}}(\text{OO})$ + 0.4 Å increases the barrier heights to 7.9, 7.5, 6.7, and 6.9 kcal/mol, respectively. There are numerous studies of proton-water clusters and proton transfer barriers with a variety of computational methods; [179–182] a thorough comparison with earlier studies is beyond the scope of the present work.

The binding energies between additional water molecules and $H_5O_2^+$ are listed in Table 3.5, along with the MP2/aug-cc-pVDZ results. Overall, the agreement is good, with a mean-signed deviation of 1.6 kcal/mol. Note that unconstrained optimization of the structure $H^+(H_2O)_6$ (IV) using PMOw collapses to isomer (III). Thus, the value in Table 3.5 was obtained by fixing the relative torsion angles of the hydrogen atoms of the central $H_5O_2^+$ unit to the MP2 values. Overall, the results from the PMOw model are in good accord with MP2 calculations and other theoretical models.

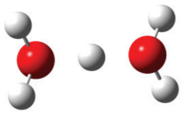
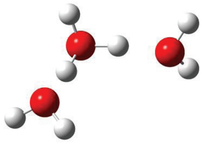
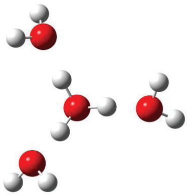
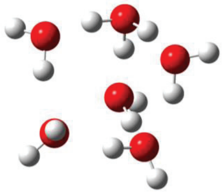
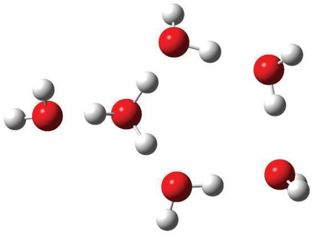
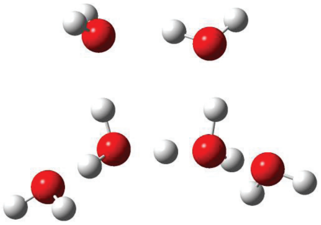
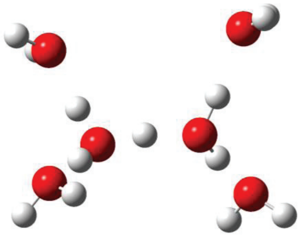
	
$H^+(H_2O)_2$	$H^+(H_2O)_3$
	
$H^+(H_2O)_4$	$H^+(H_2O)_6$ (I)
	
$H+(H_2O)_6$ (II)	$H^+(H_2O)_6$ (III)
	
$H^+(H_2O)_6$ (IV)	

Figure 3.5: Optimized geometries of $H^+(H_2O)_n$ clusters from PMOw.

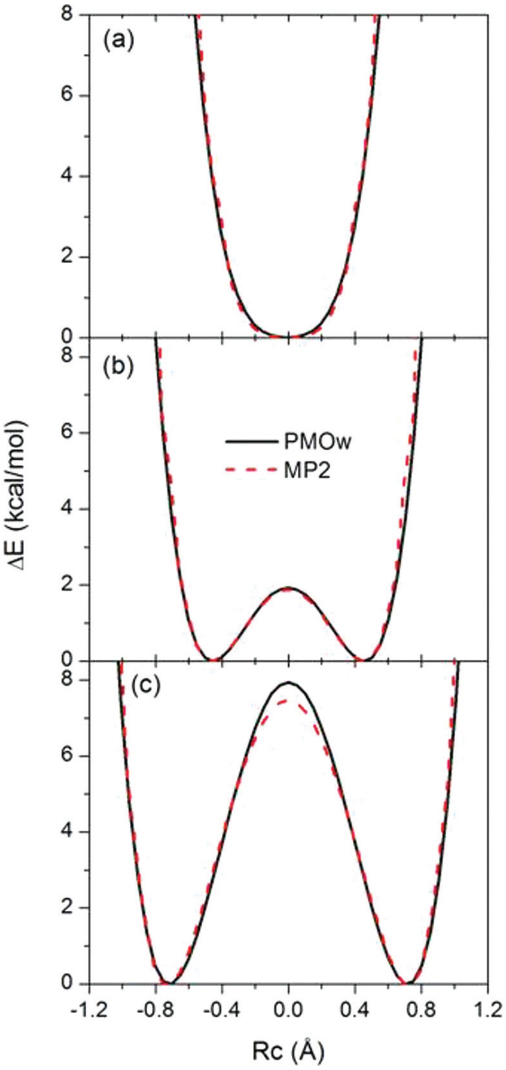


Figure 3.6: Potential energy profile for $H_5O_2^+$ in the gas phase as a function of the proton transfer coordinate , defined as the distance from the mid-point between the two oxygen atoms, (a) at the minimum geometry , (b) at a fixed O–O separation of Å, and (c) at a fixed O–O distance of Å from PMOw (black), and CCSD(T)/aug-cc-pVDZ (red) calculations. Geometries were optimized with fixed O–O distances.

Complex	PMOw	MP2
$H_5O_2^+ \ldots H_2O$	-21.4	-23.8
$H_5O_2^+ \ldots (H_2O)_2$	-39.8	-43.8
$H_5O_2^+ \ldots (H_2O)_4$ (Isomer I)	-68.9	-71.8
$H_5O_2^+ \ldots (H_2O)_4$ (Isomer II)	-67.3	-71.8
$H_5O_2^+ \ldots (H_2O)_4$ (Isomer III)	-66.6	-71.0
$H_5O_2^+ \ldots (H_2O)_4$ (Isomer IV)	-60.5	-69.7

Table 3.5: Computed interaction energies in kcal/mol for $H^+(H_2O)_n$ complexes from the PMOw and MP2 methods. Interaction energies are calculated by $\Delta E = E(\text{cluster}) - \left[E(H_5O_2^+) + nE(H_2O)\right]$, where n is the number of water molecules.

3.4.2 Liquid properties

Properties at 25°C

The computed and experimental thermodynamic and dynamic properties of liquid water at 25°C and 1 atm are listed in Table 3.6, along with the results from TIP3P, [6,7,107] AMOEBA, [2] and SWM4-NDP. [117] The standard errors ($\pm 1\sigma$) were obtained from fluctuations of separate averages over blocks of $2-4\times 10^5$ configurations. A correction, by integrating the Lennard-Jones potential beyond the cutoff distance, for the Lennard-Jones potential neglected by the cutoff has been included, and this contributes to the total computed heat of vaporization by about 1%. Long-range electrostatic interactions were not corrected in the Monte Carlo simulations. Previous studies using empirical force fields indicate that there is little size dependency of the computed properties for liquid water, and these effects will be investigated in a future study. (The TIP3P and TIP4P potential functions were developed with 125 water molecules with a cutoff of 7.5 Å without long-range corrections. [6,7,107,183])

The average density of XP3P is 0.996 ± 0.001 g/cm^3 at 25°C, which is within 1% of the experimental value and is similar to results obtained with other polarizable and

	XP3P	TIP3P[a]	AMOEBA[b]	SWM4-NDP[c]	Expt.[d]
$E(l)$ (kcal/mol)	-9.83 ± 0.01[e]	-9.82	-9.89	-9.92	-9.98
ΔH_v (kcal/mol)	10.42 ± 0.01[e]	10.41	10.48	10.51	10.51
d (g/cm^3)	0.996 ± 0.001	1.002	1.000	1.000	0.997
C_p (cal mol^{-1} K^{-1})	21.8 ± 1.0	20.0	20.9		18.0
$10^6\kappa$ (atm^{-1})	25 ± 2	60			46
$10^5\alpha$ (K^{-1})	37 ± 3	75			26
μ_{gas} (D)	1.88	2.31	1.77	1.85	1.85
μ_{liq} (D)	2.524 ± 0.002	2.31	2.78	2.33	2.3–2.6
$10^5 D$ (cm^2/s)	2.7	5.1	2.02	2.3	2.3
ϵ	97 ± 8	92	82	79 ± 3	78
τ_D (ps)	8.8			11 ± 2	8.3
τ_{NMR} (ps)	2.6			1.87 ± 0.03	2.1

Table 3.6: Liquid properties of the XP3P model for water along with those from experiments, and the TIP3P, AMOEBA, and SWM4-NDP models. (a) Refs. [6,7]. (b) Ref. [2]. (c) Ref. [8]. (d) See text for details. (e) The average $E_i(l)$ from molecular dynamics simulations employing the variational Fock operator is -9.99 kcal/mol over 400 ps. This gives a heat of vaporization of 10.52 kcal/mol.

non-polarizable force fields (Table 3.6). [2, 6, 7, 117] The total energy per monomer of liquid water, $E_i(l)$, is related to the heat of vaporization by

$$\Delta H_v = -E_i(l) + P(V_{\text{gas}} - V_{\text{liq}}) + \Delta Q - (H^o - H), \tag{3.15}$$

where V_{gas} and V_{liq} are the molar volumes of water in the gas phase (ideal) and in the liquid, P is the pressure, ΔQ is the quantum corrections to inter and intramolecular degrees of freedom between the gas and liquid, and the last term, $(H^o - H)$, is the enthalpy departure function. [184] Although ΔQ and $(H^o - H)$ has been tabulated and can be explicitly included [47, 107, 183] and this amount to a total correction of -0.06 kcal/mol at 25°C, they have typically been neglected. [2, 6, 117] In this case, ΔH_v is simply approximated by $-E_i(l) + RT$, which is also adopted in the present study (Table 3.6). The calculated heat of vaporization from the XP3P model is 10.42 ± 0.01

kcal/mol using the non-variational approximation in Monte Carlo simulations, [32,47] and the value is increased to 10.58 kcal/mol using the variational Fock operator in molecular dynamics. [35, 87, 170] The variational X-Pol approach used in molecular dynamics simulations lowers the interaction energy of the liquid by about 1.5% relative to the non-variational approach used in Monte Carlo. Overall, the agreement with experiment [185,186] is good, although there is greater deviation in the non-variational approach. The quality of the XP3P quantum mechanical potential for these two critical thermodynamic properties is comparable to that of the widely used SPC, TIP3P and TIP4P models for water [106, 107] and to that of the recent polarizable models. [2, 117, 161, 187]

The distribution of the magnitudes of monomer dipole moments from polarized wave functions in the liquid is shown in Figure 3.7; these dipole moments span a range from 2.1 to 2.9 D, and they yield an average $\langle \mu_{\mathrm{liq}} \rangle$ of 2.524 ± 0.002 D. The width at half maximum in the dipole distribution is 0.30 D (a half-width of 0.8 D was reported for the AMOEBA model, [2] which seems to be unrealistically large). Clearly, there is a major enhancement of the molecular dipole moment in the liquid, amounting to an increase over 35% relative to the gas phase value. For comparison, the AMOEBA model produced a much greater average, 2.78 D, or 50% greater than its gas phase value. The SWM4-NDP model yielded an average of 2.46 D, [117] similar to the present XP3P quantum mechanical model. Our previous investigation, employing the AM1 Hamiltonian to represent water monomers in X-Pol, resulted in an average dipole moment of 2.29 D; [47] however, the smaller value in that study is partly due to the much smaller molecular polarizability from AM1, and the weak polarization effect was corrected by scaling Mulliken population charges. There is no experimental value for the dipole moment of liquid water (and in fact this quantity is not well defined), but values ranging from 2.3 to 2.6 D have been cited based on an estimate for ice Ih. [188, 189] Finally,

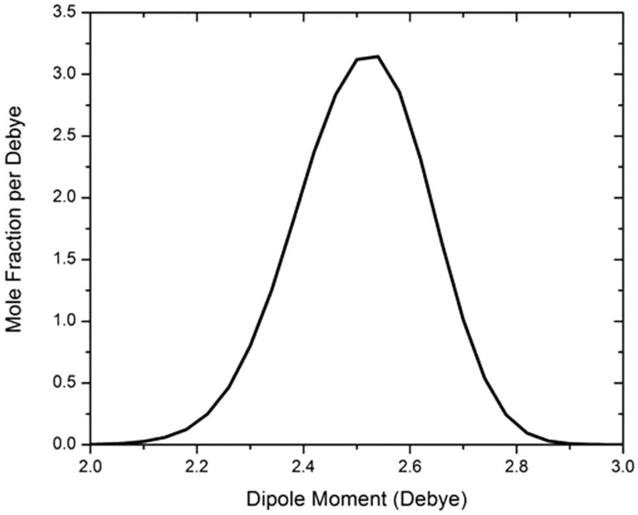


Figure 3.7: Distribution of the scalar molecular dipole moment in liquid water from Monte Carlo simulations with the XP3P potential at 25°C and 1 atm. The units for the ordinate are mole percent per Debye.

we note that *ab initio* molecular dynamics simulations yielded dipole moments ranging from 2.3 D to 3.8 D, depending on the method and functional used in DFT. [190] *ab initio* molecular dynamics simulations seem to produce greater average dipole moments than polarizable force fields and the present XP3P model.

The dielectric constant of the liquid is related to the fluctuations of the total dipole moment of the simulation box and it is dependent on the boundary conditions used to treat long-range electrostatics. [191,192] We employed the reaction field approximation in the NVT ensemble at 25°C and experimental density, where intermolecular interactions are truncated at R_{cut} = 9.0 Å. Under these conditions, a reaction field contribution is added to the electrostatic potential in Eq. 3.7: [191,193]

$$V_x^{\mathrm{RF}} = V_x(\rho_b)\left[1 + \frac{2(\epsilon_{\mathrm{RF}}-1)}{2\epsilon_{\mathrm{RF}}+1}\left(\frac{\left|\mathbf{r}_x - \mathbf{R}_B^b\right|}{R_{\mathrm{cut}}}\right)\right] \tag{3.16}$$

where ϵ_{RF} is the dielectric constant of the continuum. The static dielectric constant ϵ is determined from Eq. 3.17. [193–195]

$$\frac{(\epsilon - 1)(2\epsilon_{\mathrm{RF}} + 1)}{2\epsilon_{\mathrm{RF}} + \epsilon} = \frac{4\pi}{3k_BT}\frac{\langle \mathbf{M}^2 \rangle}{\langle V \rangle} \tag{3.17}$$

where M is the total dipole moment of the simulation box and $\langle V \rangle$ is the average volume per monomer. Ideally the reaction field dielectric ϵ_{RF} should be the same as that of the liquid in the cutoff sphere, although previous studies suggest that a choice of ϵ_{RF} in the range of $\epsilon \leq \epsilon_{\mathrm{RF}} < \infty$ typically yields consistent results, [196] and a value of 160 has been used in the present study. The liquid dipole fluctuation converges slowly, and we have carried out 16 separate simulations, each lasting about 15×10^6 configurations at 25°C. An average value of 97 ± 8 was obtained by removing the two highest and two lowest values from the 16 samples; the present average is greater than the experimental value of 78. Interestingly, Sprik argued that an average dipole moment of 2.5–2.6 D in liquid water would lead to the correct dielectric constant at room temperature, [197] and a similar observation was used in the parameter optimization process by Lamoureux *et al.* [172] In view of the average dipole moment from the XP3P liquid, which falls in the middle of this range, it is likely that a better agreement with experiment could be obtained if the simulations were further converged by extending the simulation to 100×10^6 configurations or more in each simulation. It is interesting to note that Ren and Ponder obtained a static dielectric constant of 82, in spite of a significantly larger dipole moment of 2.78 D of the liquid from the AMEOBA potential. [2] In that work, the authors argued that the correct average H–O–H angle was responsible for the good agreement between experimental and calculated liquid dielectric constant. [2, 198, 199]

Displayed in Figure 3.8 are the distributions of the binding energies per monomer in liquid water at a temperature range of -40°C to 100°C. The binding energies in Figure

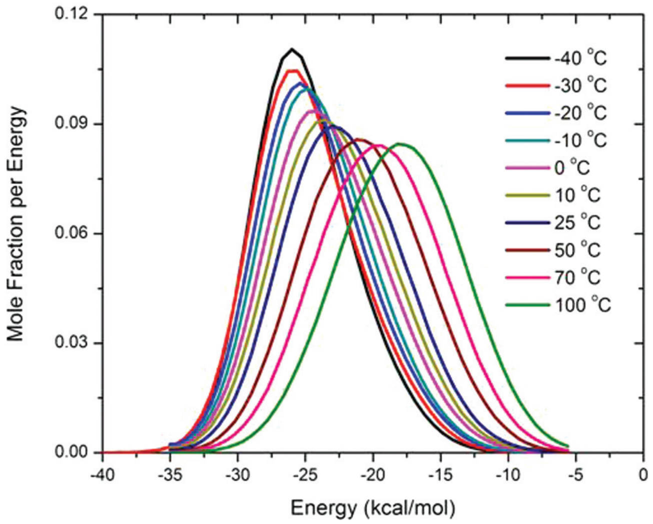


Figure 3.8: Distribution of the binding energies of water in the liquid at temperatures ranging from -40°C to 100°C. The binding energy corresponds to the total interaction energy of one water with the rest of the bulk solvent.

3.8 correspond to the interaction energy of one monomer with the rest of the system. In a polarizable model, the total energy of the liquid also includes the energy cost needed to polarize the electronic wave function (also called self-energy, see below). Thus, in contrast to the use of a pairwise potential, the average energy, $E_i(l)$, per monomer in Table 3.6 is not exactly equal to half of the binding energy at 25°C from Figure 3.8. but it is smaller by the amount of the self-energy. This is a reflection of the non-additive nature of a polarizable force field. [200] Note that such a self-energy term has been used to develop the SPC/E model. [201]

We have estimated several thermodynamic properties involving molecular fluctuations. The intermolecular contribution to the isobaric heat capacity C_P of water is defined below and can also be computed from the enthalpy fluctuations by

$$C_P = \left(\frac{\partial\langle H_i(l)\rangle}{\partial T}\right)_P + 3R = \frac{\langle H_i(l)^2\rangle - \langle H_i(l)\rangle^2}{RT^2} + 3R \qquad (3.18)$$

where $H_i(l) = E_i(l) + PV_{\text{liq}}$ is the average enthalpy of the system per monomer. The total heat capacity of the liquid C_P for a rigid monomer model is determined by adding the classical kinetic energy contributions from translation and rotation of a water molecule (3R). [6] The average from the fluctuation formula in Eq. 3.18 is 22 ± 1 cal mol-1 K-1, which is greater than the experimental value at 25°C. [202, 203] Path integral simulations by Vega *et al.* showed that inclusion of nuclear quantum effects lowers the computed heat capacity by up to 6 cal/mol. [204] Quantities based on the fluctuation formula, including C_P (isobaric heat capacity), α (coefficient of thermal expansion), and κ (isothermal compressibility) are difficult to converge; they can also be estimated from the numerical derivatives of their definitions. The derivative estimate from liquid enthalpies vs. T yields a C_P of 19 cal mol-1 K-1 at 25°C. The coefficient of thermal expansion (α) and the isothermal compressibility (κ) are determined from fluctuations of volume and enthalpy, with a computed value of $37 \times 10^{-5} K^{-1}$ for α and 25×10^{-6} atm^{-1} for κ, respectively. These quantities show relatively large deviations from experiment ($\alpha = 25.6 \times 10^{-5}\text{K}^{-1}$ and $\kappa = 45.8 \times 10^{-6}$ atm^{-1}) [202] due to their convergence.

The self-diffusion coefficient of liquid water was determined using the Einstein formula [195] from molecular dynamics simulations with constant volume and temperature:

$$D = \lim_{t \to \infty} \frac{1}{6t} \langle |\mathbf{r}(t) - \mathbf{r}(0)|^2 \rangle, \tag{3.19}$$

where $\mathbf{r}(t)$ is the position of the oxygen atom of water at time t. The diffusion coefficient was obtained as the slope from a linear fit of $\langle |\mathbf{r}(t) - \mathbf{r}(0)|^2 \rangle / 6$ as a function of t, and we obtained a value of 2.7×10^5 cm^2 s^{-1}, which agrees with experiment. [205] It is known that non-polarizable potentials for water, such as SPC, TIP3P, and TIP4P, tend to overestimate the self-diffusion coefficient, while most polarizable force fields, including the present XP3P model, show significant improvement. [2,117,161,187] The computed diffusion coefficient is also affected by finite size of the simulation box, and extrapolation to infinity will further increase the value of the diffusion coefficient. [206]

The rotational correlation times, τ_2^α, of water with respect to the H–H and O–H axes are obtained from least-square fits of the orientational time-correlation function to a single exponential function, $C_2^\alpha(t) = Ae^{-t/\tau_2^\alpha}$, where α specifies the rotation axis. The orientation time-correlation function is defined as follows: [195]

$$C_2^\alpha(t) = \langle P_2\left[\mathbf{u}_i^\alpha(t)\mathbf{u}_i^\alpha(0)\right]\rangle, \tag{3.20}$$

where P_2 is the second-order Legendre polynomial, and $\mathbf{u}_i^\alpha(t)$ is the unit vector along the α rotation axis of molecule i at time t. The time-integral of Eq. 3.20, $A\tau_2^{\mathrm{HH}}$, corresponds to the NMR rotational relaxation time of H_2O, τ_{NMR}; [207] the present XP3P model yields a value of 2.6 ps, which may be compared with the experimental value (2.1 ps). [208] For comparison, the SWM4-NDP model predicts a τ_{NMR} value of 1.9 ps. [117] Similarly, the Debye dielectric relaxation time was determined from an exponential fit to the normalized autocorrelation function of the total dipole moment $\mathbf{M}$ of the system: [195]

$$C_D(t) = \frac{\langle \mathbf{M}(t)\mathbf{M}(0)\rangle}{\langle \mathbf{M}^2(0)\rangle}. \tag{3.21}$$

The Debye relaxation time characterizes the relaxation time of the hydrogen bonding network in the liquid. The XP3P model shows that the Debye relaxation time is about 6% faster than the observed values (8.3 ps). [209] In comparison with other models, the present XP3P model performs well for these dynamic properties. [2,117,187]

The structure of liquid water is characterized by radial distribution functions (RDFs),

$g_{xy}(r)$, which gives the probability of finding an atom of type y at a distance r from an atom of type x relative to the bulk. The RDFs computed at 25°C from Monte Carlo simulations are shown in Figure 3.9 along with the neutron diffraction data. Overall, the agreement with experimental results is excellent. For the XP3P potential, the location of the maximum of the first peak of the O–O RDF is 2.78 ± 0.05 Å with a peak height of 3.0 (Figure 3.9a). For comparison, the corresponding experimental values are 2.73 Å and 2.8 from neutron diffraction. [210, 211] Integration of the O–O RDF to the first minimum at 3.30 Å yields an estimated coordination number of 4.5, which is in good agreement with the neutron diffraction result of 4.51 (integrated to 3.36 Å), but somewhat smaller than the X-ray diffraction result (4.7). The oxygen-hydrogen and hydrogen-hydrogen radial distribution functions are also in accord with experiments.

Temperature-dependent liquid properties

The computed liquid properties for ΔH_v, C_P, ρ, α, and κ, at different temperatures ranging from -40 to 100°C are listed in Table 3.7, and some of these are compared with experimental data in Figures 3.10,3.11,3.12. The formulas involving fluctuations of enthalpy and volume for C_P, α, and κ are known to have slow convergence even when Monte Carlo simulations were extended to over hundreds of millions of configurations. In the present simulations, C_P and α can also be determined directly from the enthalpy and volume derivatives with respect to temperature. For the isothermal compressibility, the fluctuation formula was used since the pressure was not changed in the present study.

The heats vaporization from -40 to 100°C were obtained from the average energies plus RT for the PV term of an ideal gas; here, we have ignored the small corrections for the quantum vibrational energy difference and enthalpy departure function. For comparison, we have included the computed heats of vaporization in Figure 3.10 from

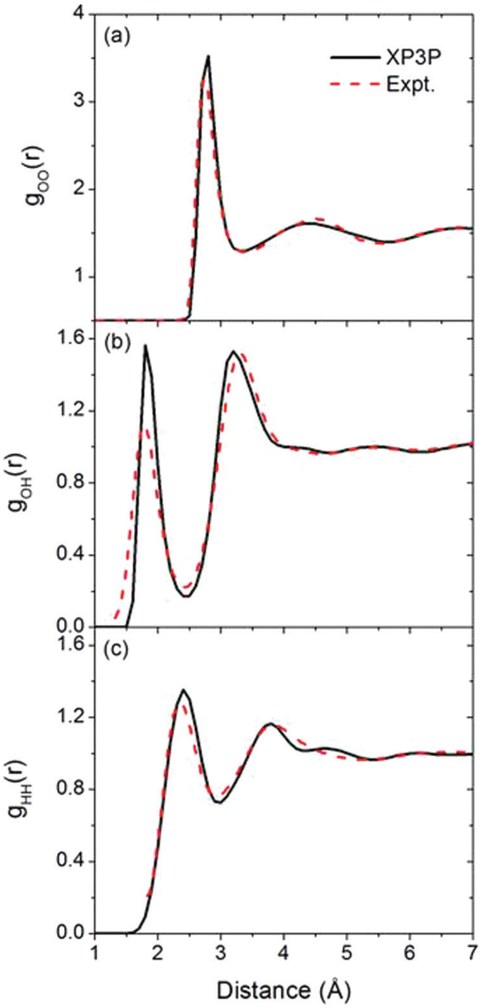


Figure 3.9: Computed (black) and experimental (red, dashed) oxygen-oxygen (a), oxygen-hydrogen (b), and hydrogen-hydrogen (c) radial distribution functions of liquid water at 25°C and 1 atm.

	ΔH_{vap}	C_p	ρ	$10^5\ \alpha$	$10^6\ \kappa$
-40 °C	11.36 ± 0.01	20.8 ± 0.7 (14)	1.008 ± 0.001	1.9 ± 3.5 (-77)	31.7 ± 2.2
-30 °C	11.30 ± 0.01	17.4 ± 0.4 (18)	1.016 ± 0.001	29.8 ± 2.0 (-117)	16.5 ± 0.6
-20 °C	11.16 ± 0.01	19.3 ± 0.5 (22)	1.032 ± 0.001	56.6 ± 3.6 (-39)	22.4 ± 1.2
-10 °C	11.02 ± 0.01	22.2 ± 0.8 (24)	1.024 ± 0.001	30.2 ± 2.2 (50)	23.9 ± 1.3
0 °C	10.83 ± 0.01	21.7 ± 0.7 (27)	1.022 ± 0.001	43.7 ± 2.6 (44)	39.3 ± 2.8
10 °C	10.64 ± 0.01	21.2 ± 1.0 (25)	1.015 ± 0.001	35.0 ± 4.0 (95)	28.2 ± 2.0
25 °C	10.42 ± 0.01	21.8 ± 1.0 (25)	0.996 ± 0.001	36.6 ± 3.0 (105)	25.0 ± 1.6
50 °C	9.96 ± 0.01	25.5 ± 1.4 (28)	0.975 ± 0.001	79.3 ± 6.2 (101)	33.8 ± 2.3
70 °C	9.54 ± 0.01	22.9 ± 1.1 (27)	0.953 ± 0.002	141.3 ± 13.8 (111)	78.3 ± 8.1
100 °C	9.03 ± 0.01	21.8 ± 0.9 (25)	0.923 ± 0.002	107.6 ± 8.0 (105)	76.2 ± 6.8

Table 3.7: Computed average thermodynamic properties of water at different temperatures between -40 °C and 100 °C (values in parentheses for C_p and α are obtained from the direct derivative calculations). The corresponding units are kcal/mol (ΔH_{vap}), cal/(mol K) (C_p), g/cm^3 (ρ), K^{-1} ($10^5\alpha$), and atm^{-1} ($10^6\kappa$).

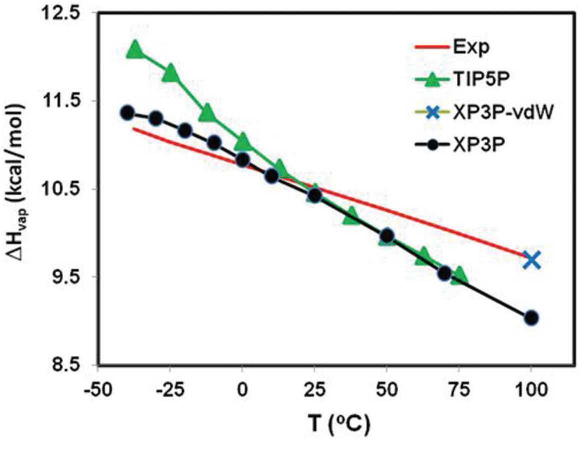


Figure 3.10: Computed (black) and experimental (red) heats of vaporization for liquid water. The results from the TIP5P model are illustrated in green.

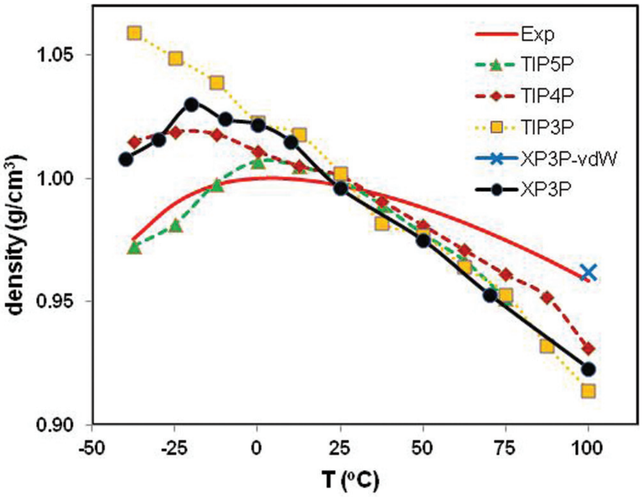


Figure 3.11: Computed (black) and experimental (red) densities for liquid water, along with those from the TIP3P (brown), the TIP4P (maroon), and the TIP5P (green) models.

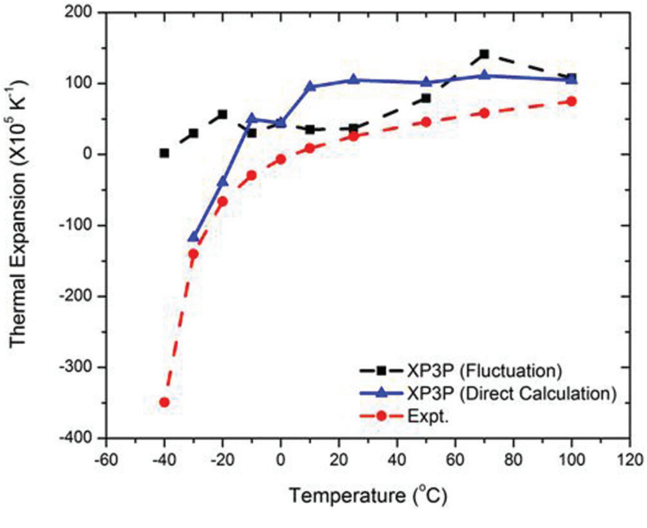


Figure 3.12: Computed and experimental coefficients of thermal expansion (α) for liquid water. The α values are determined from numerical derivatives of liquid volume variations with temperature.

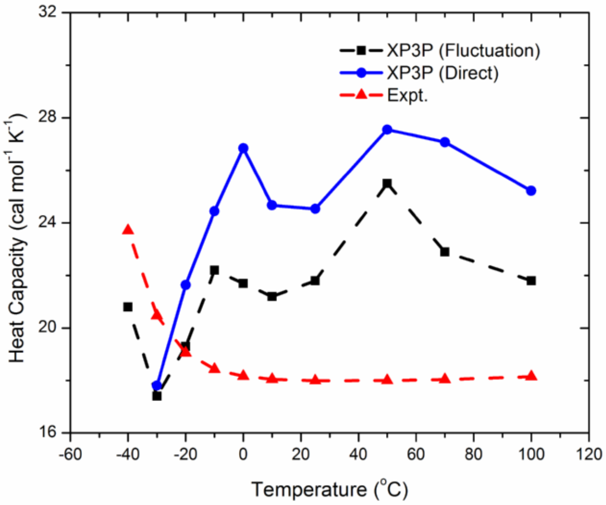


Figure 3.13: Computed heat capacities from the fluctuation formula and direct numerical derivatives from XP3P at temperatures ranging from -40 °C to 100 °C, compared to those from experiment.

the TIP5P model. The XP3P model agrees with the results from TIP5P quantitatively at temperature above 25°C. Both XP3P and TIP5P overestimate ΔH_v at temperature lower than 25°C, but the TIP5P model yielded a greater deviation on supercooled water. Figure 3.10 shows that the change in ΔH_v is nearly linear over the entire temperature range considered. This agrees with the experimental results on heat capacity, which is nearly constant at about 20 cal mol^{-1} K^{-1}. [202] The changes of heat capacity with temperature are given in Figure 3.13. The trends are in reasonable agreement with experiment at temperatures above 0°C, although the sharp increase of C_P below 20°C is not reproduced by the present simulations.

The liquid density as a function of temperature is presented in Figure 3.11 along with the experimental density of liquid water. The XP3P model, which is optimized to reproduce the heat of vaporization and density at 25°C, yields a maximum density at about -20°C. Although the density maximum is significantly lower than the experimental value at 4°C, [202] it is in fact remarkable in that there is a density maximum at all from the present model because other three-point-charge models do not possess this property with a reasonable temperature (except the SPC/E with much enhanced electrostatics). The computed density at temperature greater than 25°C shows more rapid decline with increasing temperatures than experimental results. [202] This trend is similar to that found in the TIPxP series of models. [6] The densities for supercooled water are overestimated by 2%–5% compared with the experimental data. [202] For comparison, among the non-polarizable models that do possess a density maximum, SPC/E [201] has a density maximum at -38°C, [212] TIP4P at -15°C, [7] and TIP5P at about 0°C; the TIP5P model was optimized to reproduce the temperature dependence of liquid density of water. [6] The AMOEBA model has a density maximum at 17°C. [199]

The temperature dependences of the computed density and ΔH_v from the non-polarizable TIPxP series of models [6,7] and the polarizable AMEOBA potential [199] indicate that it is difficult, with fixed empirical parameters, to obtain good agreement (within 1%) with experiment for the entire temperature range from the supercooled liquid to the boiling point. This difficulty has been pointed out by Siepmann and co-workers, who used a charge-dependent van der Waals radius for oxygen in a fluctuating charge model for water. [213] Giese and York [84] developed a density-dependent van der Waals potential that can be directly incorporated into QM/MM style simulations. We have further optimized σ_{O} at 100°C to yield a better agreement with the experimental liquid density ρ. We found that a small change in σ_{O} from 3.225 to 3.205 Å is sufficient to produce a liquid density (0.962 g/cm^3) in good agreement with experiment (0.958 g/cm^3). This is shown by the blue cross point in Figure 3.11. Interestingly, the computed ΔH_v (9.70 kcal/mol) was also found to be in excellent agreement with experiment (9.72 kcal/mol) [185] after this small adjustment (blue cross point in Figure 3.10). With this change, the average dipole moment is computed to be 2.470 ± 0.001 D, representing an increase of 0.042 D from 2.428 D computed with the original Lennard-Jones parameters in Table 3.2.

In view of the small change in the $\sigma_{\rm O}$ value, we suggest a simple temperature-dependent relationship for $\sigma_{\rm O}$,

$$\sigma_{\rm O}(T) = 3.225 - 2.667 \times 10^{-4}(T - 298.15) \tag{3.22}$$

in Å^3 where T is the absolute temperature. Alternatively, Eq. 3.22 may be rewritten in terms molecular dipole moment, which translates the expression to an aesthetically appealing, density-dependent one. In any event, it is straightforward to use Eq. 3.22 in Monte Carlo simulations, while it can be conveniently incorporated into a thermostat algorithm in molecular dynamics simulations. [164,214,215] However, a thorough examination of the performance of temperature-dependent van der Waals parameters is beyond the scope of the present work.

The computed coefficient of thermal expansion, α, follows the experimental trends nicely in Figure 3.12, and the negative values for supercooled water are consistent with the experimental values as a result of the existence of a density maximum vs. temperature.

The average dipole moment from the XP3P model decreases monotonically with increasing temperature (inset of Figure 3.14). The distributions of scalar dipole moment in the liquid at different temperatures are given in Figure 3.14. Consistent with Figure 3.8, the maximum positions are shifted towards smaller values as temperature increases, and this shift is accompanied by an increase in half width from about 0.26 D to about 0.32 D. The broader distribution of molecular dipole moment in liquid water at higher temperature reflects greater variations in the local hydrogen bonding networks and reduced average binding energies (Figure 3.8) and heats of vaporization (Figure 3.10). It is interesting to notice that the maximum dipole values in the distributions are not shifted at different temperatures (Figure 3.14); it is the population of the molecular dipole moment in the liquid that is broadened. This results in a shift of the maximum position towards smaller average values as the temperature increases. In a recent study, Raabe and Sadus suggested that the introduction of bond and angle flexibility in a water model is responsible for the decrease in the dipole moment with increased temperature and for the good performance on computed dielectric constant and pressure-temperature-density behavior using a flexible water model. [216] However, the water geometry was severely distorted from the gas-phase structure and the average bond lengths and angles in the liquid states are both significantly larger than commonly accepted values of liquid water. [216, 217] The results displayed in Figure 3.14 show that the change in electronic polarization at different thermodynamic state points also makes critical contributions to the variation of the molecular dipole moment.

Computed radial distribution functions, which exhibit the expected trends as functions of temperature, are given in the supporting information. The loss of the liquid structure is observed with increasing temperature, and the height of the first peak in $g_{\mathrm{OO}}(r)$ declines with broadening of the peak as the first minimum disappears at high temperature (Figure S7 of the supporting information). On the contrary, $g_{\mathrm{OO}}(r)$ at low temperatures exhibits more structured RDFs. Similar trends are observed in both $g_{\mathrm{OH}}(r)$ and $g_{\mathrm{HH}}(r)$ as functions of temperature (Figures S8 and S9 of the supporting information).

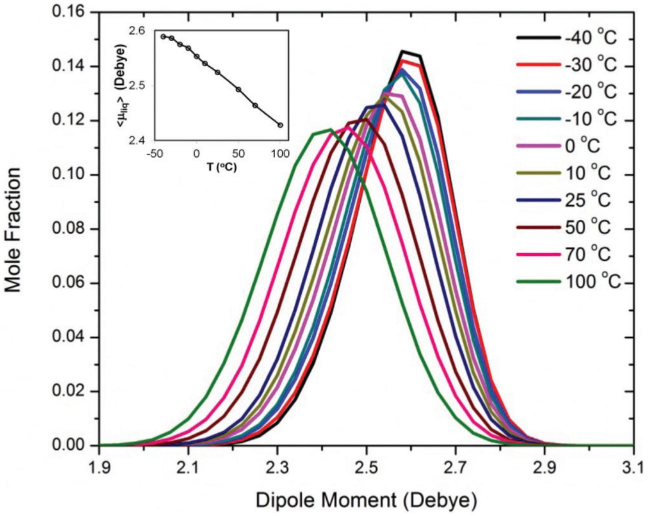


Figure 3.14: Computed average molecular dipole moments for liquid water at different temperatures.

Energy decomposition analysis of liquid water

The total binding energy, $E_i(l)$, from the XP3P water can be decomposed into specific contributing factors, [47,167,200] including vertical interaction energy and polarization energy. This analysis is useful for understanding the energy terms that are implicitly fitted in the development of polarizable or non-polarizable empirical potentials.

The vertical interaction energy represents the total energy of the liquid in which the wave function of each water molecule is not polarized, corresponding to that in the gas phase,

$$\Delta E_{\text{vert}} = \frac{1}{2}\sum_{a=1}^{N}\sum_{b\neq a}^{N}\langle\Psi_a^o|H_{ab}^o\left(\rho_b^o\right)|\Psi_a^o\rangle + E^{\text{XD}} \tag{3.23}$$

where $H_{ab}^o(\rho_b^o)$ is the interaction Hamiltonian between molecules a and b, in which the electrostatic potential defined in Eqs. 3.6 & 3.7 is obtained using the density of molecule b in the gas phase, ρ_b^o, and $E^{\text{XD}} = \sum_{a>b} E_{ab}^{\text{XD}}$ is the total van der Waals (i.e., the exchange-correlation term approximated by the Lennard-Jones potential in Eq. 3.10).

We emphasize that the term "vertical interaction energy" in energy decomposition analysis (EDA) is used to describe the interaction energy of the solvent molecules with their gas-phase, non-polarized electronic wave function relative to that of non-interacting molecules (Eq. 3.23). [167,200,218] This differs from the meaning of "vertical" that is associated with processes such as ionization and electronic excitation, where the geometries of the solute and the surrounding solvent are hypothetically kept in the un-ionized or the ground-state equilibrium configuration. In both cases the electronic wave function of the solute does change. In condensed-phase simulations, however, the energy accompanying the change of the electronic wave function is called polarization energy. Therefore, the term vertical is used to specify the interaction energy from an electronic state that is kept to remain in its gas-phase (electronic) configuration, prior to polarization.

The wave functions of the solvent molecules are polarized in the liquid, and the energy change induced by the mutual interactions with the rest of the system corresponds to the polarization interaction energy, which is defined by Eq. 3.24. [47,167,200]

$$\Delta E_{\text{pol}} = (\langle \Phi | H | \Phi \rangle - N E_a^o) - \Delta E_{\text{vert}} = E_{\text{tot}} - \Delta E_{\text{vert}} \tag{3.24}$$

The polarization energy can be further separated into two physically significant terms, corresponding to the so-called self-energy, ΔE_{self}, which is an energy cost (also called energy penalty) needed to pay for distorting the molecular wave function, and a net stabilizing contribution, ΔE_{stab}, which is responsible for polarizing the electronic wave function to lower the total energy of the system. These energy terms are given below,

[47,127,167,200]

$$\Delta E_{\text{self}} = \sum_{a=1}^{N} \left[\langle \Psi_a | H_a^o | \Psi_a \rangle - \langle \Psi_a^o | H_a^o | \Psi_a^o \rangle \right] = \sum_{a=1}^{N} \Delta E_a \tag{3.25}$$

$$\Delta E_{\text{stab}} = \frac{1}{2} \sum_{a=1}^{N} \sum_{b \neq a}^{N} \left[\langle \Psi_a | \hat{H}_{ab}(\rho_b) | \Psi_a \rangle - \langle \Psi_a^o | \hat{H}_{ab}^o(\rho_b^o) | \Psi_a^o \rangle \right] = \frac{1}{2} \sum_{a=1}^{N} \sum_{b \neq a}^{N} \Delta\Delta E_{ab}. \tag{3.26}$$

Shown in Table 3.8 and Figure 3.15 are the XP3P energy components at different temperatures. The vertical interaction energy contributes an almost constant percentage of the total binding energy, ranging from 60.8% at -40°C to 65.0% at 100°C. The increase of the percentage with increasing temperature can be attributed to the increased volume of the system and reduced polarization effects at higher temperatures. At all temperatures used in the simulations, polarization effects are significant, contributing 35.0%–39.2% of the total binding energies. At 25°C, the average polarization energy is -3.66 kcal/mol (37.2% of $E_i(l)$). The van der Waals (or exchange-dispersion) term E^{XD} is dominated by the repulsive potential. The total electrostatic (non-van der Waals) component of the binding energy, $E_i(l)$, is the sum of the vertical and polarization interaction energies less the E^{XD} term, and it is about 20%–30% greater than the total binding energy in the 140°C temperature range.

Table 3.8 shows that the average energy cost, i.e., self-energy (Eq. 3.25), needed to polarize the molecular wave function, is 3.10 ± 0.01 kcal/mol from the XP3P mode at 25°C. This value is somewhat greater than the value estimated using the AM1 Hamiltonian (3.03 ± 0.01 kcal/mol). [47] If the classical expression for the self-energy, [201]

$$\Delta E_{\text{self}}^{\text{cl}} = \Delta \mu_{\text{ind}}^2 / 2\alpha \tag{3.27}$$

T (°C)	$E_i(l)$	E_{vert}	E_{pol}	ΔE_{stab}	ΔE_{self}	E^{XD}	E_{ele}
-40	-10.89	-6.62	-4.27	-7.98	3.71	3.17	-14.06
-30	-10.81	-6.57	-4.24	-7.92	3.68	3.12	-13.93
-20	-10.66	-6.52	-4.14	-7.71	3.57	2.97	-13.63
-10	-10.50	-6.44	-4.06	-7.56	3.50	2.90	-13.40
0	-10.29	-6.35	-3.94	-7.30	3.36	2.74	-13.03
10	-10.08	-6.26	-3.82	-7.07	3.25	2.62	-12.70
25	-9.83	-6.17	-3.66	-6.76	3.10	2.49	-12.32
50	-9.32	-5.90	-3.42	-6.26	2.84	2.20	-11.52
70	-8.86	-5.69	-3.17	-5.78	2.61	1.97	-10.83
100	-8.28	-5.38	-2.90	-5.23	2.33	1.69	-9.97

Table 3.8: Temperature-dependent energy components (units in kcal/mol).

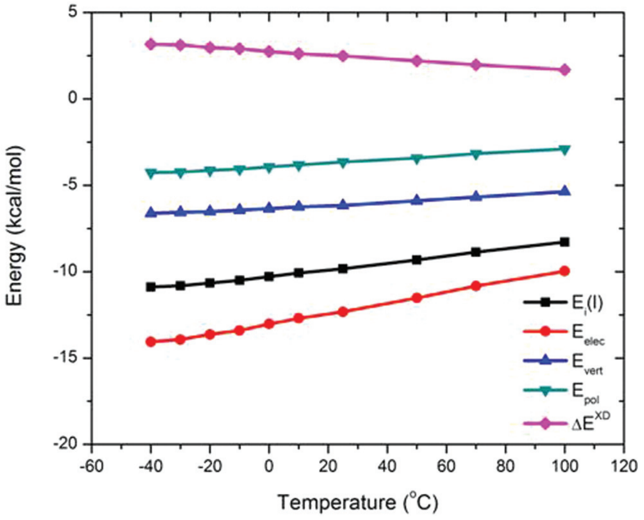


Figure 3.15: Average total interaction energies (black) per water in the liquid and their contributing components, including vertical interaction energies (blue), polarization energies (green), total electrostatic interaction energies (red), and exchange-dispersion correlation energies (magenta).

is used, where $\Delta\mu_{\text{ind}}$ is the induced dipole moment in the liquid, which is 0.64 D at 25°C, and α is the molecular polarizability (1.27 Å^3) from the XP3P model, we obtain a self-energy of 2.35 kcal/mol, somewhat smaller than the quantum mechanical result (Eq. 3.24). The self-energy was used to correct the total energy of liquid water in the SPC/E model, [201] which has an effective dipole of 2.35 D ($\Delta\mu_{\text{ind}} = 0.50$D). In that work, an estimate of $\Delta E^{\text{cl}}_{\text{self}}$ = 1.25 kcal/mol was used as an energy correction based on experimental polarizability of water. Table 3.8 shows that over the temperature range of -40 to 100°C, ΔE_{self} varies from 3.69 kcal/mol to 2.33 kcal/mol, and the corresponding total polarization energies change from -4.25 to -2.90 kcal/mol.

3.5 Conclusions

A quantum mechanical force field (QMFF) for water with the explicit treatment of electronic polarization (X-Pol) has been described. Moving beyond the current Lifson-type, molecular mechanics force fields (MMFF) that have been under continuous development in the past half century, [219–221] the present QMFF represents the condensed-phase system explicitly by an electronic structure method. Consequently, the internal energy terms in the traditional MMFF are replaced by a quantum mechanical formalism that naturally includes electronic polarization. An important aspect of the present procedure is the partition of a solution into molecular fragments such that the total wave function of the system is approximated as a Hartree product of antisymmetric, fragment wave functions. This approximation requires an empirical treatment of short-range intermolecular exchange repulsion and long-range dispersion interactions between different molecular fragments; however, one can model these effects using customary empirical formalisms. To this end, we have introduced a polarizable molecular orbital (PMO) model in the framework of the neglect diatomic differential overlap approximation. The present study represents a first step towards the goal of developing a full QMFF for the dynamic simulations of macromolecular systems as traditionally carried out with MMFF.

In this work, we introduce the first generation of a QMFF for water, making use of the PMO model specifically parameterized for compounds composed of hydrogen and oxygen, i.e., PMOw. The electrostatic potential responsible for the interactions among different fragments is model by a three-point charge representation that reproduces the total molecular dipole moment and the local hybridization contributions exactly. Consequently, the present QMFF for water, suitable for modeling gas-phase clusters, pure liquids, solid isomorphs, aqueous solutions, and the self-dissociation along with proton and anion transport, is called the XP3P model. The work in this chapter highlights the performance of the PMOw model for small water and proton clusters and simple proton transfer reactions, and the properties of liquid water using XP3P from a conglomeration of about 900×10^6 self-consistent-field calculations on a periodic system consisting of 267 water molecules. It is no exaggeration to say that this is the longest quantum mechanical simulation performed to date. More significantly, the unusual dipole derivative behavior of water, which is incorrectly modeled in molecular mechanics, but is critical for a flexible water model, is naturally reproduced as a result of an electronic structural treatment of chemical bonding by XP3P. Much remains to be tested and investigated in future studies with the combined use of large clusters treated by PMOw embedded the XP3P liquid water. We anticipate that the present model is useful for studying proton transport in solution and solid phases as well as across biological membranes through ion channels.[1]

[1] Jaebeom Han and Peng Zhang provided the simulation results from Monte Carlo and the PMOw parameterizations, respectively. This work has been partially supported by the National Institutes of Health, Grant Nos. GM46376 and RC1-GM091445, and by the National Science Foundation Grant Nos. CHE09-56776 and CHE09-57162. Part of the computations has been performed on an SGI Altix cluster acquired through an NIH grant (Grant No. S10-RR029467). We thank an anonymous referee for several insightful comments.

3.6 Supporting Information

See supplementary material for optimized geometries and computed properties for water clusters and proton-water clusters using the PMOw and XP3P method and various *ab initio* molecular orbital and density functional theory approaches mentioned in the text, and average thermodynamic properties for liquid water at temperature ranging from -40 to 100 °C. In addition, figures depicting optimized structures for water clusters, computed reorientation and molecular dipole time-correlation functions, root-of-mean square displacement, heat capacities, isothermal compressibilities, and radial distributions functions for liquid water are provided. This information is available free of charge via the Internet at `http://dx.doi.org/10.1063/1.4816280`.

Chapter 4

Quantum Mechanical Force Field for Hydrogen Fluoride

4.1 Introduction

Hydrogen fluoride (HF) is a highly corrosive and toxic compound with a boiling point temperature of 19.5 °C at atmospheric pressure [222]. It is commonly used in industrial applications such as glass etching, where it reacts with silicon dioxide to produce hexafluorosilic acid [223].

$$\begin{aligned} SiO_2 + 6HF &\rightarrow H_2SiF_6 + 2H_2O \\ H_2SiF_6 &\rightarrow SiF_4 + 2HF \end{aligned} \tag{4.1}$$

Although progress has been made [9, 10, 13, 16, 18–20, 26, 222, 224–227], the amount of experimental data concerning the structure and physical properties of HF is scarce in comparison to other simple fluids of small molecules. As a result, there is great interest in developing computational models for HF, which can produce accurate dynamic and thermodynamic properties across a wide range of temperatures and pressures.

The first computational studies of HF in the liquid state date back to the works of Cournoyer and Jorgensen [228–230] and the works of Klein and McDonald [231, 232] in the late 1970s. The approach of Cournoyer and Jorgensen used in Monte Carlo simulations of the liquid employed a pairwise 12-6-3-1 interaction potential between rigid monomers of the gas-phase geometry. The potential function was fitted to reproduce binding energies of molecular complexes from *ab initio* Hartree-Fock theory [54,55] using the STO-3G [233] and 6-31G [81] basis sets. Similarly, Klein and McDonald used molecular dynamics simulations to model the liquid state, in which a mixture of exponential and inverse power functions were parameterized to reproduce *ab initio* energies of the HF dimer [234].

Both groups employed three-site models in their studies, but did not make use of condensed phase experimental data in their parameterizations. In 1984 Cournoyer and Jorgensen introduced the three-site TIPS model for liquid HF [17], which was greatly improved over their previous models and parameterized to reproduce experimentally-measured thermodynamic properties of the liquid state [19,20,224]. The following year, the first neutron diffraction study of deuterium fluoride (DF) was published [225], providing structural information, including radial distribution functions at 293 K. A direct comparison to predictions from the TIPS model was provided in that study, showing TIPS to be surprisingly accurate at 293 K.

In 1997, two three-site models for liquid HF were introduced by Jedlovszky and Vallauri [235,236]. These models consisted of a non-polarizable potential called JV-NP and a polarizable alternative called JV-P, both of which employed the experimental H-F bond length of 0.973Å, determined by electron diffraction in the vapor phase [226]. This was a departure from all previous models, which employed an H-F bond length of 0.917Å for the monomer in the gas phase [9]. In 2000, a more extensive set of experimental data for two liquid and four supercritical states of DF was reported by

Pfleiderer and co-workers [18], which was followed by a study from Jedlovsky and co-workers using the TIPS, JV-NP, and JV-P potentials. Comparison with the new experimental data revealed that the polarizable JV-P model was clearly superior compared to the non-polarizable potentials [23].

In 2003, Wierzchowski *et al.* introduced a quantum mechanical potential for liquid HF, which incorporated electronic polarization directly in the molecular wave function [48]. This method, which was initially called a molecular orbital derived empirical potential for liquids (MODEL) [32, 47], has been subsequently called the explicit polarization (X-Pol) theory [33, 35, 87], and has been used in studies of liquid water [237] and a molecular dynamics simulation of a solvated protein [49]. In the study of Wierzchowski *et al.*, the semiempirical AM1 model [129] was used to represent the individual monomers in the liquid, and the polarization of the molecular wave function of each HF molecule by the surrounding monomers was directly incorporated into the one-electron Hamiltonian. To account for short-range exchange repulsion and long-range dispersion interactions, Lennard-Jones terms were used. Parameters for both the H-F bond lengths of 0.917Å and 0.973Å were provided, and simulation results suggested that the AM1 model was not sufficiently polarized for liquid HF.

In 2005, Kreitmeir *et al.* published a paper where the JV-NP model was used with a slightly shorter H-F bond length of 0.950Å to obtain improved results over the original JV-NP model [238]. This observation led to a reparameterization of the JV-P model by Pártay, Jedlovszky, and Vallauri, called PJV-P, resulting in a bond length of 0.930Å [21]. Pártay and co-workers provided an extensive comparison of the PJV-P model with several other models for liquid HF at many different states in that study. Additionally, comparisons were made for 11 different state points using the newly available experimental results of McLain and co-workers [16, 26].

Recent years have seen numerous studies using *ab initio* molecular dynamics (AIMD) techniques, including Born-Oppenheimer MD (BOMD) and Car-Parrinello MD (CPMD) [239], to simulate the vapor and liquid phases of HF [240–244]. In these studies, density functional theory (DFT) with the BLYP exchange-correlation functional [245, 246], which is known to lack an adequate description of dispersion effects, is typically used. The AIMD studies have shown that although HF monomers are flexible, dissociation of the H-F bond does not occur in the anhydrous liquid, agreeing with experiments [240]. However, results from these simulations, even with dispersion corrections, have not been on the level of accuracy of the empirical models, such as PJV-P and JV-P, and have produced densities that are siginificantly higher than those observed experimentally, leading McGrath *et al.* to conclude that BLYP with the D2 dispersion correction [247] is unsatisfactory for describing HF [244].

While the modeling of liquid water by X-Pol under the AM1 method with Mulliken charges to represent the intermolecular electrostatic potential yielded accurate thermodynamic results [47], a similar attempt by Wierzchowski *et al.* at modeling liquid HF was not as fruitful [48]. However, they recognized that the H-F bond length could greatly affect the simulation results and should be considered as a parameter in the potential function optimization for a rigid model, which was not attempted in any HF model until a few years later. Additionally, it was noted that the AM1 method yielded poor molecular polarizability and that the Mulliken population charges used for polarization were too small.

The explicit polarization model for hydrogen fluoride introduced here seeks to rectify these problems. As in earlier studies [48, 242], a two-site monomer is used in the X-Pol model for hydrogen fluoride (XPHF), as opposed to the more common three-site approach. In contrast to the approach by Wierzchowski *et al.*, in which AM1 was adopted, we employ a recently introduced semiempirical quantum chemistry model called the polarized molecular orbital (PMO) method [52, 93, 94]. By introducing a set of p-orbitals onto hydrogen atoms in PMO, it was found that the performance of molecular polarization and hydrogen-bonding was significantly improved over existing semiempirical models. The PMO Hamiltonian has been successfully used to develop a quantum mechanical force field for liquid water, called XP3P [237]. In addition, we have employed an alternative population analysis for calculating partial charges, called the dipole-preserving polarization consistent (DPPC) charge method [53], which reproduces the total molecular dipole moment of each monomer, eliminating the need for the charge scaling parameter in previous X-Pol simulations. Finally, the model has been parameterized extensively using experimental data not available at the time of the previous study [16,26], and has employed an optimized H-F bond length of 0.930Å.

In this work, a set of parameters for fluorine is incorporated into the PMOw method, originally developed for oxygen and hydrogen containing compounds. These parameters were optimized by fitting against experimental and *ab initio* data on the HF monomer, HF dimer, HF trimer, (HF)(H_2O) complex, and OF_2. In the present study, the PMO formalism is identical to that of the XP3P model for liquid water. As in the case of XP3P, the XPHF method has been implemented into the MCSOL Monte Carlo program [248] and a modified version of NAMD [51].

4.2 Semiempirical PMO Method for Hydrogen Fluoride

4.2.1 Polarized Molecular Orbital Method

The polarized molecular orbital method (PMO) is based on the formalisms of the MNDO method [89], which makes use of the neglect of diatomic differential overlap approximation (NDDO) [88,91]. Three key modifications were introduced in PMO.

Since the method has been reported in detail previously [52, 93, 94, 237], we only provide a brief summary of its key departures from MNDO.

First, a set of p-orbitals is introduced on hydrogen atoms. The addition of the p-orbitals greatly improves the performance on calculated molecular polarizabilities for a range of compounds, and provides an excellent description of hydrogen-bonding interactions. To prevent unphysical bonding interactions from the additional p-orbitals, the resonance integrals involving hydrogen are damped (Eqs. 4.2 and 4.3).

$$\beta_{lp}^{\mathrm{HH}} = 0, \tag{4.2}$$

$$\beta_{lp}^{\mathrm{FH}} = \frac{\beta_l^{\mathrm{F}} + \beta_p^{\mathrm{H}}}{2} S_{lp} A_{lp} e^{\kappa_{lp} R_{\mathrm{FH}}}. \tag{4.3}$$

The S_{lp} of Eq. 4.3 is the overlap integral $\langle \mathrm{F}_l | \mathrm{H}_p \rangle$ between fluorine and hydrogen p-type orbitals, which uses the typical MNDO exponent ζ. In the more general case that this integral needs to be evaluated for homonuclear pairs (e.g. $\langle \mathrm{F}_l | \mathrm{F}_l \rangle$), the specialized ζ_{PMO} exponents are used.

Second, as is done in the MNDO formalism [89,142], the nucleus-electron attraction integral, $H_{\mu\nu}^{A}$, between the electronic charge density of atom A and the nucleus of atom B, is evaluated by the two-electron repulsion integral $\langle \mu_A \nu_A | s_B s_B \rangle$, where s_B denotes an s-orbital on nucleus B [91]. This attraction integral is modified in the PMO model when both A and B are hydrogen atoms (Eq. 4.4).

$$H_{pp'}^{\mathrm{H}} = \left[1 - B e^{-\lambda R_{HH'}^2} \right] \left(H_{pp'}^{\mathrm{H}} \right)_{\mathrm{MNDO}} \tag{4.4}$$

The third and final departure from MNDO is concerned with core-core interactions.

Similar to homonuclear overlap integrals, special exponents $\hat{\alpha}$ are used for homonuclear core-core repulsion. In addition, as is commonly used in DFT, the pairwise D1 dispersion correction of Grimme [95] is used between all atom pairs.

At present, three different variations of the PMO method have been reported, PMOv1 [52], PMO2 [94], and PMOw [237], none of which contain parameters for fluorine. The variant of PMO for which we have decided to introduce the F parameters is the PMOw model, having kept all other parameters fixed to those reported in Chapter 3.

4.2.2 Motivation for using PMOw

The starting point for parameterization of empirical models for vapor and liquid phase simulations is an accurate description of dimer interactions. Here, the PMOw model has been parameterized with the goal of accurately describing the HF dimer, as well as other HF clusters.

As a motivation for introducing a new PMOw parameter set for F, we performed several single-point energy calculations and geometry optimizations on the HF dimer using the MOPAC [90] software with the NDDO-type MNDO [89], AM1 [129], RM1 [130], PM3 [249], and PM6 [131] semiempirical methods. Interaction energies for fixed F-F distances with optimized H coordinates for each method were tabulated. In addition to the MOPAC calculations, analogous PMOw and *ab initio* calculations at the M06-2X/MG3S level [250–252] were performed with an in-house code [253] and NWChem version 5.1.1 [254], respectively.

This series of calculations showed that with the exception of the PM3 method, all existing semiempirical methods tested produced an optimized geometry for the HF dimer that is qualitatively incorrect in comparison with experimental and M06-2X/MG3S derived results (Figure 4.1). In addition, these methods failed to accurately reproduce the HF dimer interaction energy profile (Figure 4.2) and its minimum interaction energy, which is -4.54 kcal/mol with F-F distance 2.72Å from experiment [12,13], -4.94 kcal/mol with F-F distance 2.7316Å at the CCSD(T)/TZ2P(f,d) level [11] and -5.13 kcal/mol with F-F distance 2.7242Å at the M06-2X/MG3S level.

It is worth noting that although the PM3 optimized geometry appears qualitatively correct, the HF dimer interaction energy profile was found to be qualitatively incorrect. Thus, none of the existing NDDO-type semiempirical methods tested are adequate for accurately describing the HF dimer interaction.

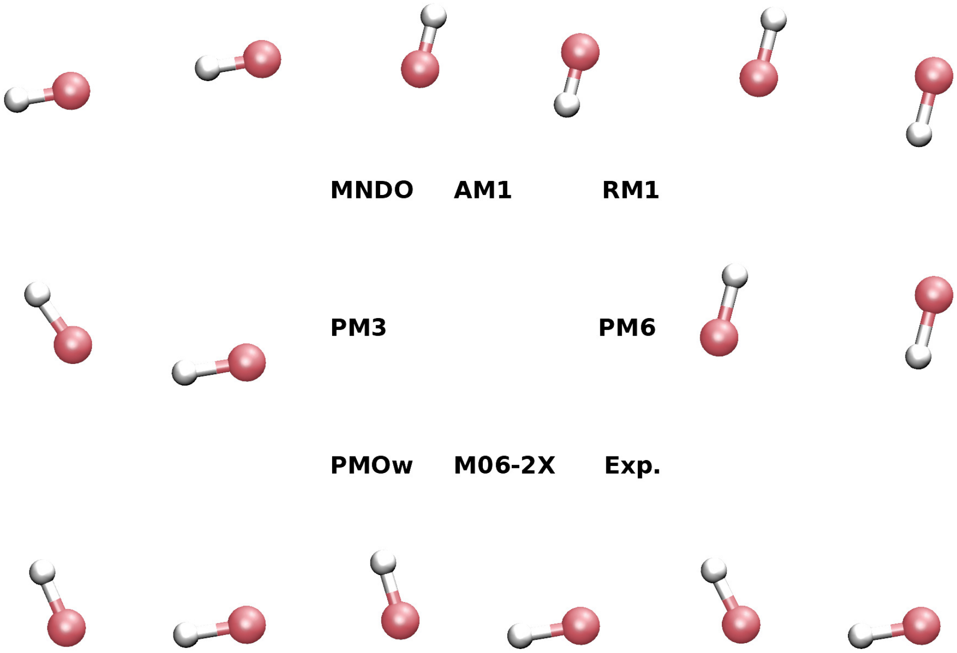


Figure 4.1: Optimized HF dimers using several NDDO-type semiempirical methods, PMOw, DFT at the M06-2X/MG3S level, and that measured from experiment [13]. All non-PMOw semiempirical dimers, with the exception of the PM3 optimized geometry, exhibit a qualitatively incorrect structure compared to *ab initio* and experimental results.

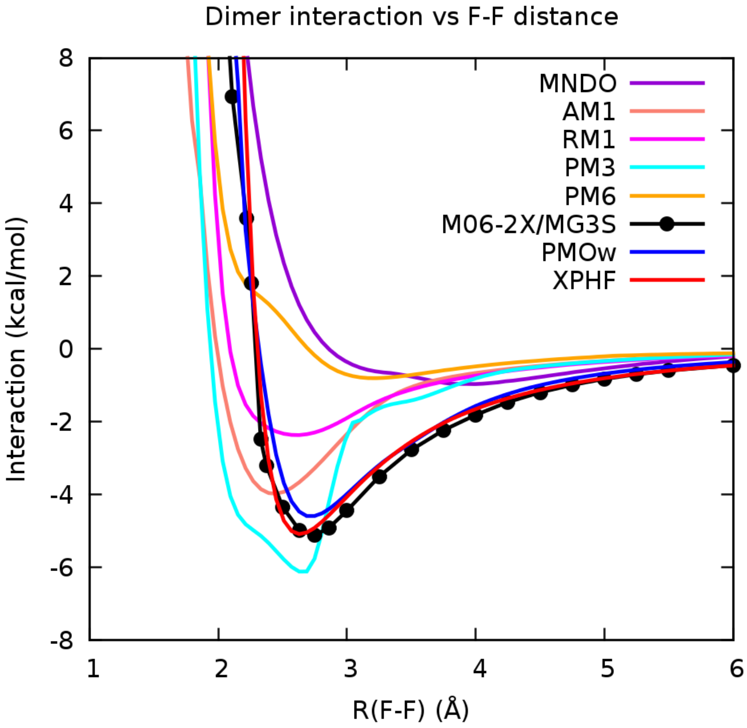


Figure 4.2: Interaction energy profile of optimized HF dimers with respect to constrained F-F distance for several NDDO-type semiempirical methods, PMOw, XPHF, and DFT at the M06-2X/MG3S level. Notice the odd behavior of the interaction energy curve for PM3, which produced a qualitatively correct optimized HF dimer geometry.

4.2.3 Fluorine Parameters for PMOw

The poor HF dimer descriptions of existing semiemprical methods and the success of the PMOw-based XP3P water model led to the decision to include fluorine in the PMOw model in connection with the original parameters for O and H for use with XPHF. The parameters for F in PMOw were obtained from the minimization of a fitness function that is defined as a weighted sum of absolute differences for properties of the HF monomer, HF dimer, HF trimer, (HF)(H_2O) complex, and OF_2 molecule; these properties include bond length, bond angle, dipole moment, and interaction energies of optimized geometries.

Target values for the fitness function were set to experimental values where available and to *ab initio* derived values when experimental data were absent. Minimization of the fitness function was performed using stochastic optimization, starting from the RM1 parameter for F [130] and the PMOw specific parameters for O as the initial guess with A_{sp} and A_{pp} held fixed. All previous parameters of PMOw for H and O were kept fixed to ensure that the PMOw results for water remained reproducible. The results of the optimized parameters are listed in Table 4.1, where the columns for H and O are reproduced from Chapter 3.

4.2.4 Hydrogen Fluoride Clusters

We tested our PMOw parameter set for F on a series of cyclic HF clusters from the trimer to the octamer as well as the monomer and dimer. We present a comparison to experimental results where available, previously published *ab initio* data, and additional DFT calculations at the M06-2X level with the MG3S basis set.

Table 4.2 lists bond lengths, angles, and dipole moments of the HF monomer and dimer at the PMOw, M06-2X/MG3S, and CCSD(T)/TZ2P(f,d) levels of theory as well

	H	O	F
U_{ss} (eV)	-11.15043	-111.86028	-139.42406
U_{pp} (eV)	-7.35459	-78.64105	-109.03911
β_s (eV)	-6.88125	-25.57063	-69.32684
β_p (eV)	-3.52628	-31.90404	-34.08908
ζ_s (Bohr^{-1})	1.17236	3.05303	5.60791
ζ_p (Bohr^{-1})	1.05333	3.12265	3.11602
α (Å^{-1})	3.05440	3.76880	4.29492
g_{ss} (eV)	12.73667	17.36659	16.39526
g_{sp} (eV)	8.04688	13.37288	18.38443
g_{pp} (eV)	6.98401	14.78196	16.67384
$g_{pp'}$ (eV)	10.65161	13.49319	14.77192
h_{sp} (eV)	1.92149	4.42643	4.30118
$\hat{\alpha}$ (Å^{-1})	2.52552	3.03253	3.48493
ζ_{PMO} (Bohr^{-1})	1.280	2.764	2.786
A_{sp}	NA	0.03	0.03
A_{pp}	NA	0.15	0.15
κ_{sp} (Å^{-1})	NA	0.47069	0.48605
κ_{pp} (Å^{-1})	NA	0.47069	0.38879

Table 4.1: Parameters in the PMOw model. The parameters for F were obtained by stochastic optimization starting from the RM1 parameter for F using a fitness function related to experimental and *ab initio* calculated quantities of HF, $(HF)_2$, $(HF)_3$, $(HF)(H_2O)$, and OF_2. The parameters for H and O come from Chapter 3.

as corresponding experimental values. The definitions for lengths and angles in Table 4.2 are shown in Figure 4.3. The experimental results of Table 4.2 are the same values used in the training set for parameterization of F.

The optimized HF monomer bond length and dipole moment from PMOw are 0.917Å and 1.80 Debye respectively, showing excellent agreement with the experimental values of 0.917Å [9] and 1.80 Debye [10]. The F-F separation of the optimized dimer structure using PMOw was 2.72Å, in good accord with the experimental value of 2.72 ± 0.03Å. The optimized tilt (θ) and flap (ϕ) angles of $8.5°$ and $67.3°$ for the dimer fall within the uncertainty range of the experimental data $10° \pm 6°$ and $63° \pm 6°$ [13], respectively. The computed binding energy was -4.64 kcal/mol for the HF dimer and the corresponding total dipole moment was 3.26 Debye, which is in agreement with the CCSD(T)/TZ2P(f,d) results of -4.94 kcal/mol and 3.33 Debye [11]. The corresponding experimental dipole moment has been reported to be 2.99 Debye [13].

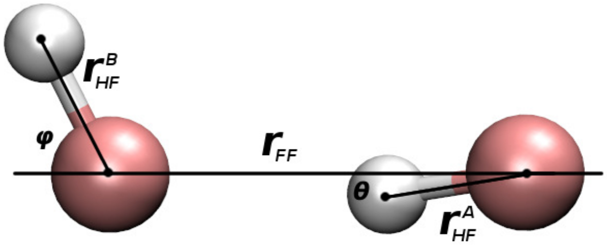


Figure 4.3: Labeled quantities for the HF dimer corresponding to quantities given in Table 4.2.

The results of geometry optimization on the larger clusters (HF)$_n$ for $n = 3, \cdots, 8$ using PMOw and M06-2X/MG3S along with the "best estimates" of Maerker and coworkers [14] are given in Table 4.3. The quantities listed in the table are labeled in Figure 4.4 for the trimer, and are similar for larger clusters. As in the case of the monomer and dimer, the trimer was included in the parameterization training set, and

	PMOw	M06-2X/MG3S	CCSD(T)/TZ2P(f,d)	Expt.
HF				
r_{HF} (Å)	0.917	0.918	0.918[c]	0.917[a]
μ (Debye)	1.80	1.88	1.82[c]	1.80[b]
(HF)$_2$				
E_{int} (kcal/mol)	-4.64	-5.13	-4.94[c]	-4.54[d]
r_{HF}^{A} (Å)	0.925	0.924	0.923[c]	–
r_{HF}^{B} (Å)	0.924	0.921	0.921[c]	–
r_{FF} (Å)	2.72	2.72	2.73[c]	2.72 ± 0.03[e]
θ (°)	8.5	11.3	6.4[c]	10 ± 6[e]
ϕ (°)	67.3	71.8	68.8[c]	63 ± 6[e]
μ (Debye)	3.27	3.26	3.33[c]	2.99[e]

Table 4.2: HF monomer properties of PMOw compared to *ab initio* and experimental results; (a): Ref. [9], (b): Ref. [10], (c): Ref. [11], (d): Ref. [12], (e): Ref. [13].

the optimized properties are in good agreement with the M06-2X/MG3S results and the "best estimates". Overall, the larger clusters, which were not included in the training set, also exhibit good agreement between PMOw, M06-2X/MG3S, and the "best estimates". Figure 4.5 illustrates the cyclic clusters obtained from geometry optimization using PMOw and M06-2X/MG3S, color-coded and displayed side-by-side.

4.2.5 (HF)$_n$(H$_2$O)$_n$ Complexes

In addition to HF clusters, we also examined three different (HF)$_n$(H$_2$O)$_n$ complexes where $n = 1, 2, 4$ and compared the optimized structures with those obtained at the MP2/TZP level [15] (Table 4.4). The complex with $n = 1$ was included in the parameterization training set. Partial charges were computed by Mulliken population analysis [96] using PMOw, whereas partial charges for the MP2/TZP results were computed by Löwdin population analysis [255].

The complexes for the cases with $n = 2$ and $n = 4$ exhibit ionic bonding behavior in the optimized geometries using PMOw (Figure 4.6). However, these structures are

	PMOw	M06-2X/MG3S	"Best Estimate"[a]
$(HF)_3$			
E_{int} (kcal/mol)	-13.92	-17.50	–
$\langle r_{\text{HF}} \rangle$ (Å)	0.935	0.934	0.933
$\langle r_{\text{FF}} \rangle$ (Å)	2.66	2.60	2.59
$\langle \theta \rangle$ (°)	25.6	23.4	24
$(HF)_4$			
E_{int} (kcal/mol)	-26.56	-29.73	–
$\langle r_{\text{HF}} \rangle$ (Å)	0.952	0.943	0.944
$\langle r_{\text{FF}} \rangle$ (Å)	2.54	2.54	2.51
$\langle \theta \rangle$ (°)	12.8	11.2	12
$(HF)_5$			
E_{int} (kcal/mol)	-36.32	-40.02	–
$\langle r_{\text{HF}} \rangle$ (Å)	0.957	0.947	0.948
$\langle r_{\text{FF}} \rangle$ (Å)	2.51	2.50	2.48
$\langle \theta \rangle$ (°)	5.7	5.1	6
$(HF)_6$			
E_{int} (kcal/mol)	-43.95	-49.04	–
$\langle r_{\text{HF}} \rangle$ (Å)	0.957	0.949	0.949
$\langle r_{\text{FF}} \rangle$ (Å)	2.51	2.49	2.47
$\langle \theta \rangle$ (°)	1.5	1.5	3
$(HF)_7$			
E_{int} (kcal/mol)	-50.44	-57.23	–
$\langle r_{\text{HF}} \rangle$ (Å)	0.956	0.948	–
$\langle r_{\text{FF}} \rangle$ (Å)	2.51	2.49	–
$\langle \theta \rangle$ (°)	1.3	0.8	–
$(HF)_8$			
E_{int} (kcal/mol)	-56.32	-64.97	–
$\langle r_{\text{HF}} \rangle$ (Å)	0.954	0.946	–
$\langle r_{\text{FF}} \rangle$ (Å)	2.51	2.50	–
$\langle \theta \rangle$ (°)	3.2	2.4	–

Table 4.3: Properties of HF cyclic clusters at the PMOw and M06-2X/MG3S levels with a combination of experimental and *ab initio* data that forms "best estimates" [14]. (a): Ref. [14].

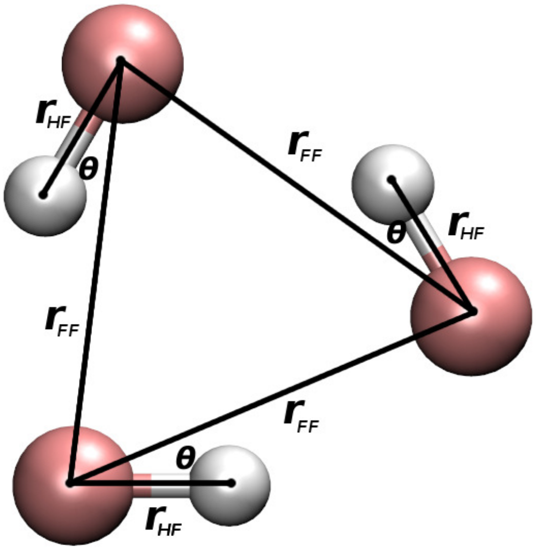


Figure 4.4: Labeled quantities for the HF trimer corresponding to quantities given in Table 4.3. Entries in the table for clusters larger than the trimer have analogous quantities to the labels in the figure.

not global minima in the MP2/TZP optimization as reported by Chaban and Gerber. For the $(HF)_2(H_2O)_2$ optimization, the relevant chemical bonds were purely covalent in character and no local minimum with ionic bonds (shared proton) was reported. In contrast to $n = 2$, Chaban and Gerber report both covalent and ionic structures for the case of $n = 4$, noting that the ionic structure is a local minimum. For comparison, we have included data for the covalent structures in the MP2/TZP column of Table 4.4 for $n = 1, 2$ and data for the ionic structure for $n = 4$. The large charges on the oxygen atoms at the PMOw level in the $n = 2$ and $n = 4$ complex suggest that the PMOw oxygen parameter may be too electronegative for the current hydrogen parameter.

To understand the behavior of the $(HF)_n(H_2O)_n$ clusters, we included OF_2 in the F-parameter optimization; however, the optimized geometrical parameters show relatively large deviations from the experimental targets. In particular, an FOF bond angle

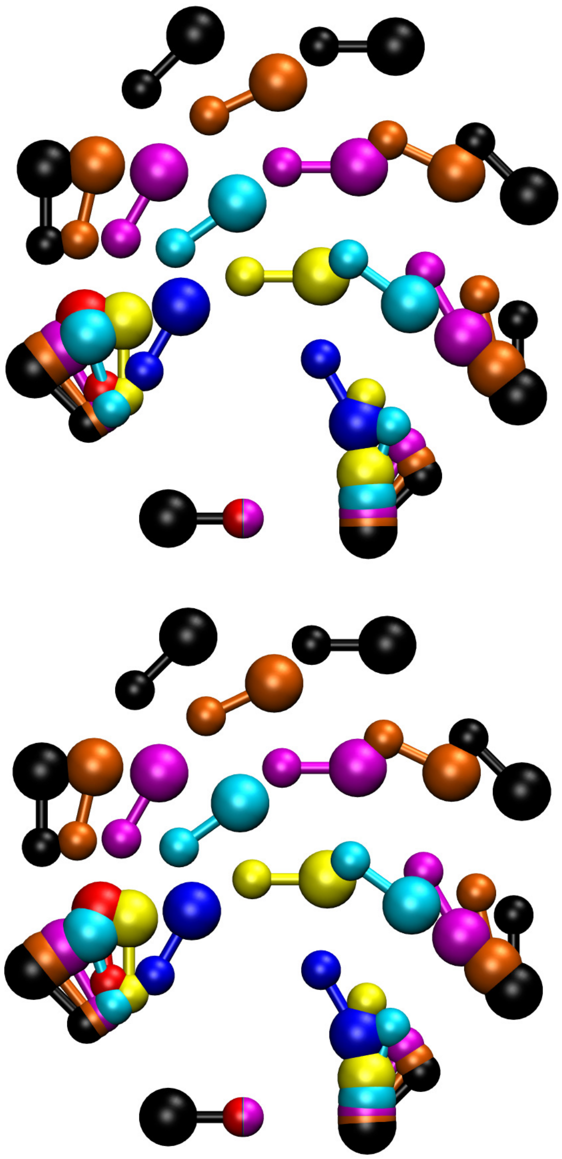


Figure 4.5: Cyclic clusters of HF at PMOw level (top) and the M06-2X/MG3S level (bottom). The various monomer colors indicate their respective cluster, and are as follows: dimer (red), trimer (blue), tetramer (yellow), pentamer (cyan), hexamer (magenta), heptamer (orange), octamer (black). The multicolored monomer at the bottom of each pane indicates the first monomer in each cluster.

	PMOw	MP2/TZP^a
(HF)(H_2O)		
r_{HF} (Å)	0.925	0.933
r_{OH} (Å)	0.961	0.958
r_{OF} (Å)	2.63	2.64
q_{H}, HF (e)	0.19	0.20
q_{F}, HF (e)	-0.22	-0.28
q_{O} H_2O (e)	-0.40	-0.36
q_{H}, H_2O (e)	0.21	0.22
$(HF)_2(H_2O)_2$		
r_{HF} (Å)	1.09	0.957
r_{OH1} (Å)	1.18	0.969
r_{OH2} (Å)	0.959	0.957
r_{OF} (Å)	2.24	1.82
q_{H}, HF (e)	0.27	0.21
q_{F}, HF (e)	-0.22	-0.28
q_{O}, H_2O (e)	-0.53	-0.36
q_{H1}, H_2O (e)	0.27	0.22
q_{H2}, H_2O (e)	0.21	0.21
$(HF)_4(H_2O)_4$		
r_{HF} (Å)	1.24	1.02
r_{OH} (Å)	1.08	1.49
r_{OF} (Å)	2.26	–
q_{H} (e)	0.27	0.22
q_{O} (e)	-0.48	-0.13
q_{F} (e)	-0.34	-0.54

Table 4.4: Properties of $(HF)_n(H_2O)_n$ complexes at the PMOw and MP2/TZP levels for $n = 1, 2, 4$. Partial charges were obtained from Mulliken and Löwdin population analysis for PMOw and MP2/TZP respectively. (a) Ref. [15].

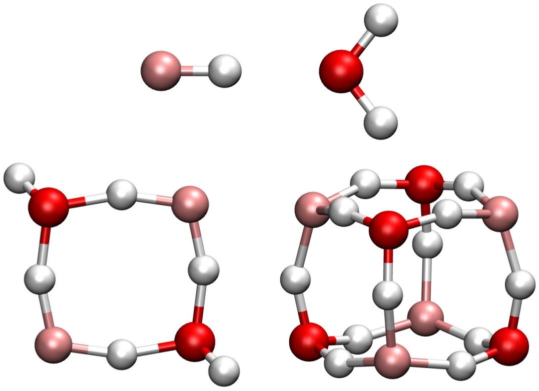


Figure 4.6: Three $(HF)_n(H_2O)_n$ complexes: $(HF)(H_2O)$ (top), $(HF)_2(H_2O)_2$ (bottom left), $(HF)_4(H_2O)_4$ (bottom right). The $(HF)(H_2O)$ complex exhibits a hydrogen bond while the $(HF)_2(H_2O)_2$ and $(HF)_4(H_2O)_4$ complexes show an ionic bonding behavior.

of 115.7° and an OF bond length of 1.16Å were obtained, which may be compared to the experimental values of 103.2° and 1.41Å, respectively [256]. This suggests that re-optimization of the parameters associated with the F and O pair may be necessary to correctly reproduce the qualitative results of $(HF)_n(H_2O)_n$ complexes using the PMOw method.

4.3 XPHF Model for Liquid HF

The explicit polarization model for liquid HF (XPHF) follows the same formalism of X-Pol reported previously [32, 33, 35, 87]. The X-Pol method has been applied to liquid water [47, 237], liquid HF [48], and a solvated bovine pancreatic trypsin inhibitor protein using the semiempirical AM1 model to represent individual fragments [49].

X-Pol is a fragment-based electronic structure method that incorporates electronic polarization by wave function theory for a condensed-phase system. For this reason, X-Pol is regarded as a QM/QM-type method or a QM force field (QMFF) if relevant parameters are optimized to reproduce experimental properties of liquids and solutions.

In earlier studies, aimed at demonstrating the feasibility of the idea of a QMFF for fluid and biomolecular simulations, the semiempirical AM1 Hamiltonian [129] and Mulliken charges [96] were used to model intermolecular interactions. The XP3P model for water and the XPHF model for hydrogen fluoride in the present study adopt a new approach, where the PMO method is used as the QM model for individual fragments and the dipole-preserving polarization consistent (DPPC) charge method [53] is used to represent partial charges in calculating the interfragment potential, providing a more rigorous and accurate framework for QMFF development.

We now provide a brief description of the X-Pol method, noting that it is covered in greater detail elsewhere [32,33,35,87].

4.3.1 Wave Function Description

The X-Pol method approximates the total wave function of a chemical system Φ as the Hartree product of the wave functions Ψ_i of N smaller subsystems called fragments.

$$\Phi = \prod_{i=1}^{N} \Psi_i \tag{4.5}$$

In the present case, Ψ_i is considered to be a Slater determinant [65] of each HF monomer, which takes the form of Eq. 4.6,

$$\Psi_i = \frac{1}{\sqrt{N!}} \begin{vmatrix} \phi_1(x_1) & \phi_2(x_1) & \cdots & \phi_N(x_1) \\ \phi_1(x_2) & \phi_2(x_2) & \cdots & \phi_N(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_1(x_N) & \phi_2(x_N) & \cdots & \phi_N(x_N) \end{vmatrix} \tag{4.6}$$

where ϕ_k is a molecular orbital formed by a linear combination of m atomic orbitals (LCAO) described by the basis set $\{\chi\}$ (Eq. 4.7).

$$\phi_k = \sum_{\mu}^{m} C_{\mu}^{k} \chi_{\mu}^{k} \tag{4.7}$$

The molecular orbitals are subjected to the orthonormalization condition (Eq. 4.8).

$$\Lambda_{ij} = \sum_{\mu}^{m} C_{\mu}^{i} C_{\mu}^{j} - \delta_{ij} = 0 \tag{4.8}$$

The Hatree-product approximation (Eq. 4.5) greatly reduces the computational cost of of the quantum calculation from a formally $\mathcal{O}([Nm]^k)$ scaling to $O(N^2m^k)$, where k depends on the level of quantum theory used. Note that the cost may be reduced to $O[Nm \log{(Nm)}]$ by using particle mesh Ewald for electrostatics between monomers [162]. The caveat of the Hartree-product approximation is the neglect of the exchange-correlation and dispersion interactions between fragments. However, it is worth noting that the intrafragment exchange-correlation and dispersion are preserved at the level of quantum theory employed in calculating Ψ_i.

4.3.2 Effective Hamiltonian and Energy Expression

The Hamiltonian used in X-Pol is defined as the sum of the electronic Hamiltonians of individual fragments $\mathcal{H}_i^o$ plus all interfragment interactions $\mathcal{H}_{ij}$.

$$\mathcal{H}^{\text{XP}} = \sum_{i=1}^{N} \mathcal{H}_i^o + \frac{1}{2} \sum_{i=1}^{N} \sum_{j \neq j}^{N} \mathcal{H}_{ij} \tag{4.9}$$

$\mathcal{H}_{ij}$ is defined as the interaction between fragment i and fragment j. Eq. 4.10 gives the expression for $\mathcal{H}_{ij}$

$$\mathcal{H}_{ij} = -\sum_{k=1}^{m} V_k(\Psi_j) + \sum_{\alpha=1}^{A} Z_\alpha^i V_\alpha(\Psi_j) + E_{ij}^{\text{XD}}, \tag{4.10}$$

where m is the number of electrons in fragment i, A is the number of atoms in fragment i, Z_α^i denotes the core charge of atom α on fragment i, and E_{ij}^{XD} is the exchange-correlation and dispersion interaction between fragments i and j.

The term $V_x(\Psi_j)$ describes the electrostatic potential at position x due to the j-th QM fragment, and is given by Eq. 4.11

$$V_x(\Psi_j) = -\int \frac{\rho_j(\mathbf{r})d\mathbf{r}}{|\mathbf{r}_x - \mathbf{r}|} + \sum_{\beta=1}^{B} \frac{Z_\beta^j}{|\mathbf{r}_x - \mathbf{R}_\beta^j|}, \tag{4.11}$$

where B is the number of atoms in fragment j and $x = k$ and $x = \alpha$ denote an interaction at electronic and nuclear positions respectively, and $\rho_j(\mathbf{r})$ denotes the electron density of fragment j derived from Ψ_j.

The total interaction energy of the system is defined by Eq. 4.12

$$E_{\text{tot}}^{\text{XP}} = \langle \Phi | \mathcal{H}^{\text{XP}} | \Phi \rangle - \sum_{i=1}^{N} \langle \Psi_i^o | \mathcal{H}_i^o | \Psi_i^o \rangle, \tag{4.12}$$

where Ψ_i^o is the optimized wave function associated with the geometry of fragment i in the gas phase.

The energy calculation can be done in a variational way with respect to partial charges obtained from Mulliken population analysis, leading to an analytical expression for its gradient to be used in molecular dynamics simulations [35].

4.3.3 X-Pol with PMOw

The XPHF model employs the semiempirical PMOw Hamiltonian with the parameters introduced in the previous section to determine the wave functions Ψ_i for individual fragments. Under the MNDO formalism [89, 92], which is used in PMOw, all one-electron integrals are approximated as two-electron integrals where a charge density is represented by an s-type distribution. In our implementation, these two-electron integrals are computed by a multipole expansion [91] and partial charges are calculated by the DPPC population analysis [53]. The functional form and methods used in the XPHF model are identical to those employed in the XP3P model, and only differ by parameters and the system of question.

In the present model, exchange-correlation and dispersion are empirically described by the Lennard-Jones 12-6 potential (Eq. 4.13), though explicit density dependence can be incorporated into the Fock matrix in a manner described by York and co-workers [84].

$$E_{ij}^{\mathrm{XD}} = \sum_{\alpha=1}^{A} \sum_{\beta=1}^{B} 4\epsilon_{\alpha\beta} \left[\left(\frac{\sigma_{\alpha\beta}}{r_{\alpha\beta}} \right)^{12} - \left(\frac{\sigma_{\alpha\beta}}{r_{\alpha\beta}} \right)^{6} \right] \tag{4.13}$$

Standard combining rules are used such that $\epsilon_{\alpha\beta} = \sqrt{\epsilon_\alpha \epsilon_\beta}$ and $\sigma_{\alpha\beta} = \sqrt{\sigma_\alpha \sigma_\beta}$ for interactions between different atom types, and the Lennard-Jones parameters used in the XPHF model for F and H are $\epsilon_{\mathrm{F}} = 0.145$ kcal/mol, $\epsilon_{\mathrm{H}} = 0.05$ kcal/mol, $\sigma_{\mathrm{F}} = 2.97$Å, and $\sigma_{\mathrm{H}} = 0.80$Å.

4.4 Simulation of Liquid HF with XPHF

We performed Monte Carlo simulations under the isothermal-isobaric ensemble (NPT) at temperature and pressure conditions listed in Table 4.5. Subsequently, molecular dynamics simulations under the isothermal-isochoric ensemble (NVT) at experimental density for the same 11 state points were performed. The 11 state points were chosen to correspond with other simulation studies of liquid HF, for which the most comprehensive comparison of various models with experiment is given by Pártay and coworkers [21].

State	T (K)	P (bar)	ρ (g/cm^3)
A	195	0.1	1.058^a
B	246	0.1	1.038^a
C	203	1.0	1.176^{b,d}
D	273	1.0	1.015^{b,e}
E	296	1.2	0.997^a
F	300	2.0	0.962^c
G	373	12	0.796^c
H	473	78	0.236^c
I	473	84	0.398^c
J	473	166	0.647^c
K	473	319	0.796^c

Table 4.5: The state 11 points used in our liquid simulations of HF and their corresponding experimental densities. (a) Ref. [16], (b) Ref. [17], (c) Ref. [18], (d) Ref. [19], (e) Ref. [20].

4.4.1 Simulation Setup

All Monte Carlo simulations were performed using the MCSOL program [248] which implements the non-variational X-Pol potential [32,47]. Each simulation box contained 267 rigid HF monomers with an H-F bond length of 0.930 Å. A switching function was used to smooth intermolecular interactions to zero in the region of 8.5 to 9.5 Å. Each Monte Carlo move was performed by randomly selecting an HF monomer, randomly translating along a randomly chosen axis at a maximum distance of 0.18 Å, and randomly rotating the monomer about a randomly selected axis centered at the fluorine atom a maximum angle of 17°. Volume moves were attempted every 500 steps with a maximum volume change of 150 Å^3. All states were initialized with random position and orientation and given at least 10^8 configurations for equilibration. The averaging of quantities in the Monte Carlo simulations for each state was carried out over at least an additional 10^7 configurations after equilibration. About 6×10^6 configurations can be executed per day on a 6-core 2.66 GHz Intel Xeon X7542 Westmer processor for a system of this size using the current version of MCSOL.

All molecular dynamics simulations of the same 11 states were carried out using a version of NAMD modified to incorporate the variational X-Pol potential. Each simulation box was modeled under the NVT ensemble at experimental density, and contained 267 rigid HF monomers that were constrained at an H-F bond length of 0.930 Å by RATTLE [257]. A switching function between 8.5 Å and 9.5 Å was employed to evaluate Lennard-Jones interactions, and electronic interfragment interactions were truncated at 9.0 Å. Each state was equilibrated for no less than 10^6 steps from a set of random coordinates. A time step of 2 fs was used for simulations of each state, and trajectories of 500 ps were used for calculating diffusion coefficients.

4.4.2 Energetic Properties

Energies per molecule of each of the 11 states are given in Table 4.6 for XPHF and the PJV-P, JV-P, JV-NP, HF-Kr and TIPS models along with experimental values obtained from the Visco-Kofke equation [22]. XPHF appears to have similar performance compared with other polarizable models both in the non-supercritical states A-G and in the supercritical states H-K. In some cases (states C, D, and G) the XPHF model slightly underestimates the energy per molecule compared to experiment, while the other polarizable models consistently overestimate the energy per molecule of every state.

State	XPHF	PJV-P	JV-P	JV-NP	HF-Kr	TIPS	Expt.
A	-34.32 ± 0.106	-32.32	-31.33	-32.41	-29.58	-30.04	-38.55^{a,b}
B	-33.89 ± 0.513	-29.97	-31.03	-32.66	-27.36	-28.03	-34.40^{a,b}
C	-34.01 ± 0.081	-31.87	-31.36	-32.24	-28.90	-29.64	-31.94^{c}
D	-32.55 ± 1.103	-28.53	-27.68	-29.18	-26.00	-26.77	-29.01^{c}
E	-29.20 ± 0.151	-27.17	-26.29	-28.36	-24.98	-25.30	-30.36^{a,b}
F	-29.11 ± 0.368	-27.02	-26.11	-28.10	-24.74	-24.90	-28.22^{a,d}
G	-24.50 ± 0.173	-22.84	-22.21	-14.49	-11.10	-20.96	-23.08^{a,d}
H	-6.89 ± 0.702	-6.82	-5.98	-10.27	-7.32	-8.81	-19.75^{a,d}
I	-7.33 ± 0.947	-7.73	-7.28	-11.32	-7.48	-9.17	-19.83^{a,d}
J	-16.30 ± 0.695	-16.77	-14.08	-16.40	-12.59	-16.01	-23.39^{a,d}
K	-18.98 ± 0.186	-18.23	-17.24	-19.14	-15.70	-17.52	-27.49^{a,d}

Table 4.6: Energy per molecule of the 11 states in kJ/mol from XPHF compared to several other models and experiment. All values for the other models originate from Table III of Ref. [21], in which uncertainties are also listed. (a) From Visco-Kofke equation of state [22]; (b,c) See Ref. [21] for details; (d) Ref. [23].

4.4.3 Electronic Properties

The average dipole moment of the 11 states with uncertainties for XPHF are given in Table 4.7 along with results from the polarizable JV-P model [23] and static values from JV-NP and TIPS. It is seen from the table that XPHF is in good agreement with the JV-P model despite its reported large uncertainties.

Average dipole moments from a previous study using the X-Pol formalism by Wierzchowski *et al.* employing the AM1 model and Mulliken charges were provided for states F and H. The values were 1.99 and 1.79 Debye respectively, with a bond length of 0.917Å, and values of 2.03 and 1.79 Debye for a bond length of 0.973Å. When compared to the average values from XPHF of 2.27 and 1.96 Debye for states F and H, respectively, and the mean values of JV-P of 2.17 and 2.02 Debye, the poor polarizability of the AM1 model is clearly seen.

State	XPHF	JV-P	JV-NP	TIPS
A	2.34 ± 0.0011	–	1.83	2.04
B	2.34 ± 0.0078	–	1.83	2.04
C	2.34 ± 0.0013	–	1.83	2.04
D	2.32 ± 0.0015	–	1.83	2.04
E	2.27 ± 0.0019	–	1.83	2.04
F	2.27 ± 0.0045	2.17 ± 0.49	1.83	2.04
G	2.21 ± 0.0025	2.11 ± 0.42	1.83	2.04
H	1.96 ± 0.0105	2.02 ± 0.08	1.83	2.04
I	1.97 ± 0.0140	2.04 ± 0.24	1.83	2.04
J	2.09 ± 0.0092	2.06 ± 0.32	1.83	2.04
K	2.13 ± 0.0024	2.07 ± 0.08	1.83	2.04

Table 4.7: The average dipole moments of the 11 states in Debye for XPHF and JV-P along with the static values from JV-NP and the TIPS models found in Ref. [23].

Figure 4.7 shows the distribution of dipole moments for each of the 11 states simulated with XPHF. Dipole moments in each state are clearly enhanced beyond the gas-phase dipole of 1.80 Debye, with the least enhancement occurring in the states with the lowest density (states H and I). General trends with respect to pressure and temperature are apparent, such as the positive correlation of the dipole distribution width with respect to increasing pressure, and the negative correlation in the enhancement of the dipole moment beyond the gas phase value with respect to temperature.

4.4.4 Density

The densities of the 11 states from Monte Carlo simulations are tabulated in Table 4.8 and compared to the PJV-P, JV-P, JV-NP, HF-Kr, and TIPS models as well as experiment. Uncertainties for all models besides XPHF are given by Pártay and co-workers [21]. The XPHF model tends to underestimate the densities of the states when compared to experiment, but appears to be comparable to the polarizable PJV-P and JV-P models in terms of overall trend.

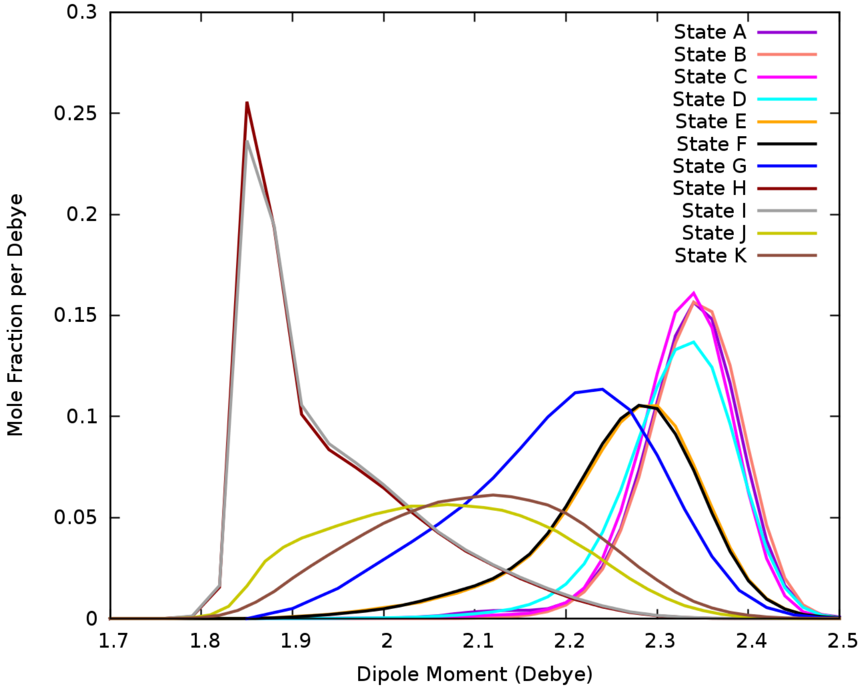


Figure 4.7: The mole fraction of average dipole moments for each of the 11 states tested with XPHF. In supercritical states H and I the distribution appears to take a maximum value near the gas-phase dipole moment, while supercritical states J and K show a wider distribution across many values.

A comparison of experimental densities from Sim and Bouknight [19] at temperatures ranging from 199.3 K to 277.4 K and atmospheric pressure to state points C (203 K, 1 bar) and D (273 K, 1 bar) of the HF models is shown in Figure 4.8. It can be seen in the figure that the polarizable models (XPHF, PJV-P, and JV-P) are in better agreement with experiment than the non-polarizable models (JV-NP, HF-Kr, TIPS). The non-polarizable models tend to predict densities much higher than experiment, especially in the case of the TIPS model. However, it is noteworthy that the densities given by Pártay and co-workers for the TIPS model at states C (1.276 ± 0.021 g/cm^3) and D (1.300 ± 0.041 g/cm^3) are within uncertainty of each other, suggesting that the qualitatively incorrect trend of the TIPS model may be due to convergence issues.

Although the trend is clearly linear in the temperature dependence of density for the experimental data set shown in Figure 4.8, the linear interpolation between states C and D of the computational models may be an over-simplification, and more investigation is required to make a definitive conclusion on the linearity of isobaric temperature dependence of density under the various models.

4.4.5 Structural Properties

Historically, experimental radial distribution functions (RDFs) for HF have been reported as a total RDF, for which a weighted decomposition of the three pair correlation functions was suggested by Pfleiderer and co-workers [18] (Eq. 4.14).

$$G(r) = 0.4966 g_{\mathrm{HF}}(r) + 0.2104 g_{\mathrm{FF}}(r) + 0.2930 g_{\mathrm{HH}}(r) \tag{4.14}$$

Using these weights, we constructed the total RDFs the supercritical states H-K from NVT molecular dynamics trajectories of XPHF at the experimental densities, and compared them to RDFs of other models and experiment [18]. These total RDFs are plotted

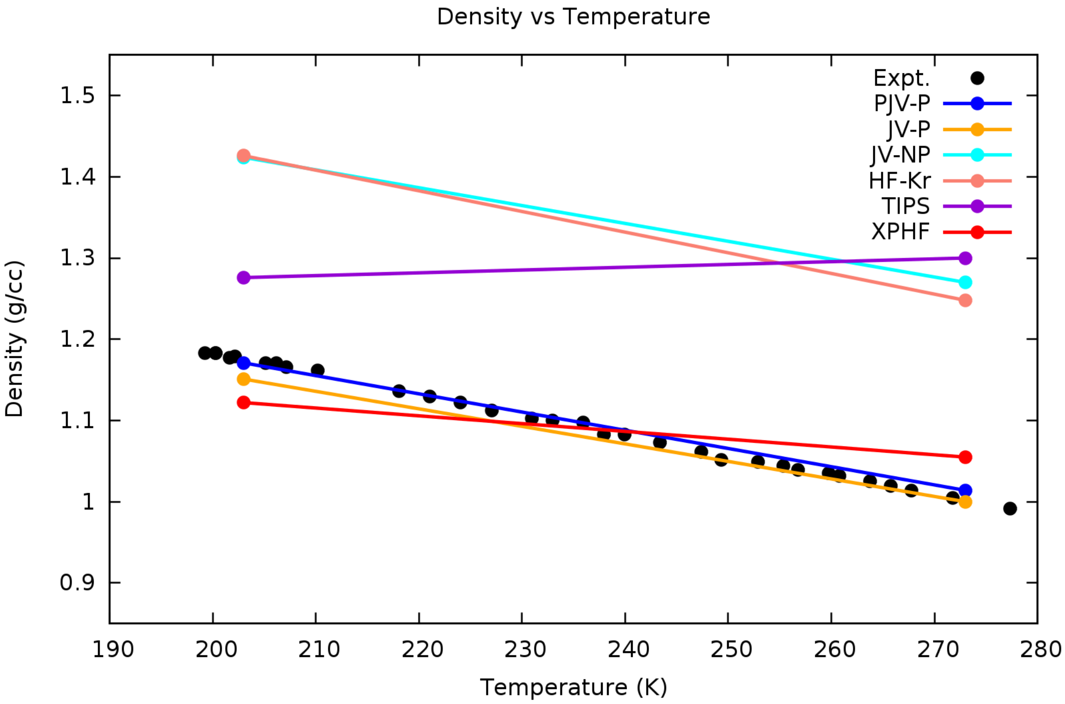


Figure 4.8: The temperature dependence of liquid HF density at atmospheric pressure from XPHF and other models at states C and D compared to the experimental values of Simons and Bouknight [19]. Notice that the polarizable models (XPHF, PJV-P, JV-P) predict densities in better qualitative and quantitative agreement with experiment than the non-polarizable models (JV-NP, HF-Kr, TIPS).

State	XPHF	PJV-P	JV-P	JV-NP	HF-Kr	TIPS	Expt.
A	1.112 ± 0.024	1.182	1.143	1.350	1.507	1.300	1.058[a]
B	1.081 ± 0.015	1.085	1.101	1.336	1.371	1.224	1.038[a]
C	1.122 ± 0.021	1.171	1.151	1.424	1.426	1.276	1.176[b,c]
D	1.055 ± 0.055	1.014	1.000	1.270	1.248	1.300	1.015[b,d]
E	0.911 ± 0.030	0.950	0.923	1.234	1.190	1.019	0.997[a]
F	0.902 ± 0.041	0.951	0.924	1.230	1.171	0.971[e]	0.962[f]
G	0.748 ± 0.015	0.769	0.774[e]	0.019[e]	0.029	0.633[e]	0.796[f]
H	0.068 ± 0.006	0.065	0.068[e]	0.073[e]	0.068	0.081[e]	0.236[f]
I	0.078 ± 0.011	0.083	0.081[e]	0.097[e]	0.070	0.091[e]	0.398[f]
J	0.392 ± 0.050	0.490	0.334[e]	0.294[e]	0.240	0.423[e]	0.647[f]
K	0.636 ± 0.018	0.634	0.584[e]	0.615[e]	0.511	0.579[e]	0.796[f]

Table 4.8: Densities of the 11 states in g/cm^3 from XPHF compared to several other models and experiment. Unless otherwise noted, all values for the other models originate from Table IV of Ref. [21], in which uncertainties are also listed. (a) Ref. [16]; (b) Ref. [17]; (c) Ref. [19]; (d) Ref. [20]; (e) Ref. [23].

in Figure 4.9, showing that XPHF produces slightly better agreement with experiment than the JV-P and JV-NP models for each state. However, it is worth noting that some of the same qualitative differences of those models compared with experiments are observed in XPHF. In particular, state J shows peaks that are much shorter than their values from experiment.

The partial RDFs of HF (i.e. g_{FF}, g_{FH} and g_{HH}) were first resolved experimentally in 2004 by McLain and co-workers [26], and reported for a temperature and pressure of 296 K and 1.2 bar (state E). Figure 4.10 shows a comparison of XPHF and the PJV-P model [21] to these partial RDFs and the BOMD result of McGrath and co-workers [244] at the BLYP-D2/TZV2P level with temperature and pressure of 300 K and 1.0 bar.

Figure 4.10 shows that the partial RDFs of the XPHF model are in agreement with those of the PJV-P model and they are consistent with experiment. The first peak of the $g_{\text{FF}}(r)$ RDF for XPHF appears to be in slightly better agreement with experiment

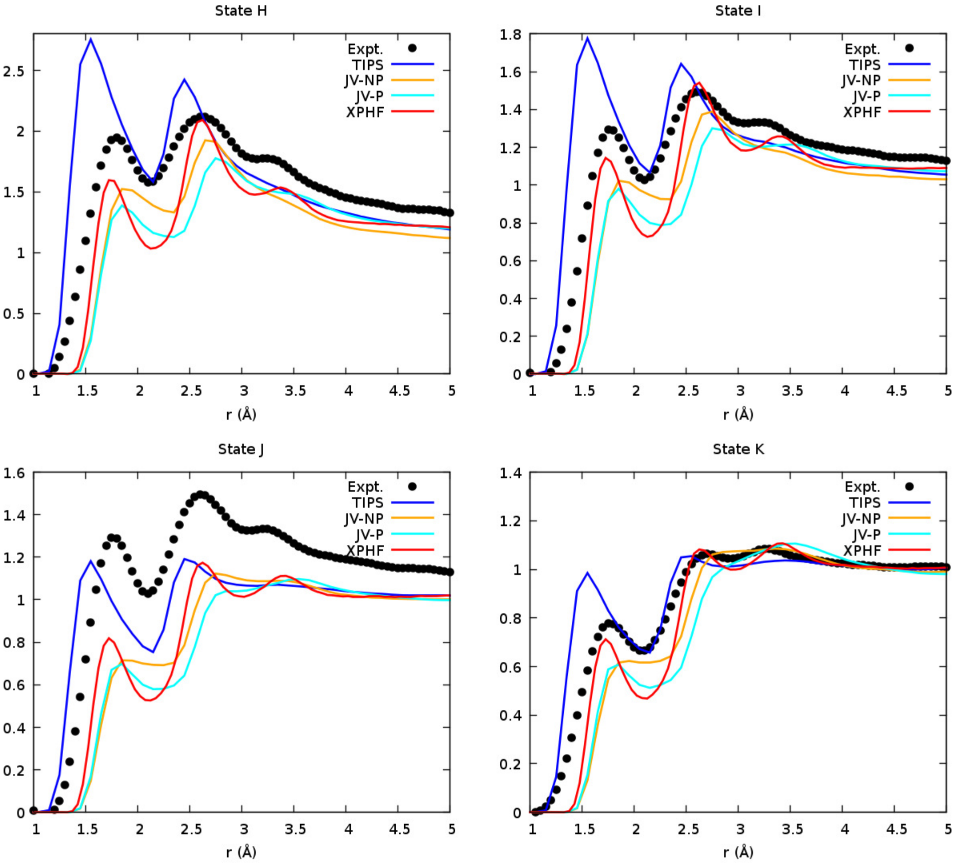


Figure 4.9: Total RDFs of the XPHF (red), JV-P (cyan), JV-NP (orange), and TIPS (blue) models with experimental data (black) [18] for the four supercritical states at 473 K. The TIPS model is over-structured for all tested supercritical states but J (473 K, 166 bar). XPHF gives results that appear slightly better than the JV-P model, but with more qualitatively-correct peaks.

than that of PJV-P in terms of peak position and height. The $g_{\mathrm{FH}}(r)$ RDFs behave similarly between the XPHF and PJV-P models in terms of position, with XPHF displaying somewhat higher first and second peaks. Finally, the $g_{\mathrm{HH}}(r)$ RDF peak positions are nearly half an Ångstrom too long for both the XPHF and PJV-P models compared to experiment, with the peak heights being higher for XPHF and lower for PJV-P.

The partial RDFs of the BOMD simulation in the NPT ensemble by McGrath and co-workers at a slightly different state point show somewhat better agreement with experiment than the XPHF and PJV-P models at peak positions, particularly in the case of g_{HH}. It has been suggested by Pártay and co-workers that the drop to zero in the experimental g_{HH} RDF is the result of measurement anomaly. This is further supported by the BOMD-derived RDFs, which with that exception have similar first and second peak heights and positions compared to experiment. Although McGrath and co-workers stated that BLYP-D2 was unsatisfactory for accurately describing HF at that state point based upon the calculated density, the partial RDFs are in agreement with experiment and exhibit qualitatively correct behavior for the HH pair correlation function.

Due to the Hartree-product approximation in X-Pol and the use of a two-site model for the HF monomer, the optimized HF dimer in the XPHF model is a linear structure similar to the full MNDO result, but with a better interaction energy profile, as indicated in Figure 4.2. Although the optimized dimer structure is completely linear, the hydrogen-bonded monomer displays an enhanced charge on the hydrogen atom. Since each monomer must maintain a neutral net charge, the fluorine atom associated with the enhanced charge on the hydrogen atom is also enhanced. This is equivalent to stating that the XPHF model does not incorporate or consider charge transfer effects. On the basis that a bent dimer similar to the full PMOw result, which includes charge transfer effects, has a shorter H-H distance than an analogous linear structure with the same F-F distance, we suggest that the first peak in the g_{HH} RDF could be shortened by the incorporation of charge transfer effects into XPHF.

4.4.6 Diffusion

The self-diffusion coefficients of hydrogen fluoride for the 11 states were determined using the Einstein formula [195] on 500 ps trajectories from MD simulations in the NVT ensemble at experimental densities. Eq. 4.15 gives the Einstein formula,

$$D = \lim_{t \to \infty} \frac{1}{6t} \langle |\mathbf{r}(t) - \mathbf{r}(0)|^2 \rangle, \tag{4.15}$$

where $\mathbf{r}(t)$ denotes the position of the fluorine atom at time t. Diffusion coefficients were measured using a linear fit of $\langle |\mathbf{r}(t) - \mathbf{r}(0)|^2 \rangle / 6$ in the middle region of the curve from the MD trajectories. Diffusion coefficients from XPHF and experiments [24] are given in Table 4.9. Experimental values for states A-H and K were determined from the Arrhenius equation $D_{\mathrm{Expt.}} = D_0 \exp(-E_A/RT)$ where $E_A = 9.92$ kJ/mol and the values for D_0 are 452×10^9 and 398×10^{-9} m^2/s for standard vapor pressure (A-H) and 500 Bar (K), respectively.

The computed diffusion coefficients for states D, E, F, G, and K show good agreement with the Arrhenius equations derived from experiments, while those from states A, B, C, and H show noticeable discrepancies. In the case of state H, the predicted diffusion coefficient is nearly three times that of experiment, indicating that state H is approaching the vapor phase under XPHF.

As for states A, B, and C, further inspection of the MD trajectories shows the presence of a series of antiparallel chains of HF. Such chain-forming behavior is consistent with observations of HF in the solid state by X-ray diffraction [227], for which a freezing point of -83.4 °C was reported. These antiparallel chains are the most rigid in states

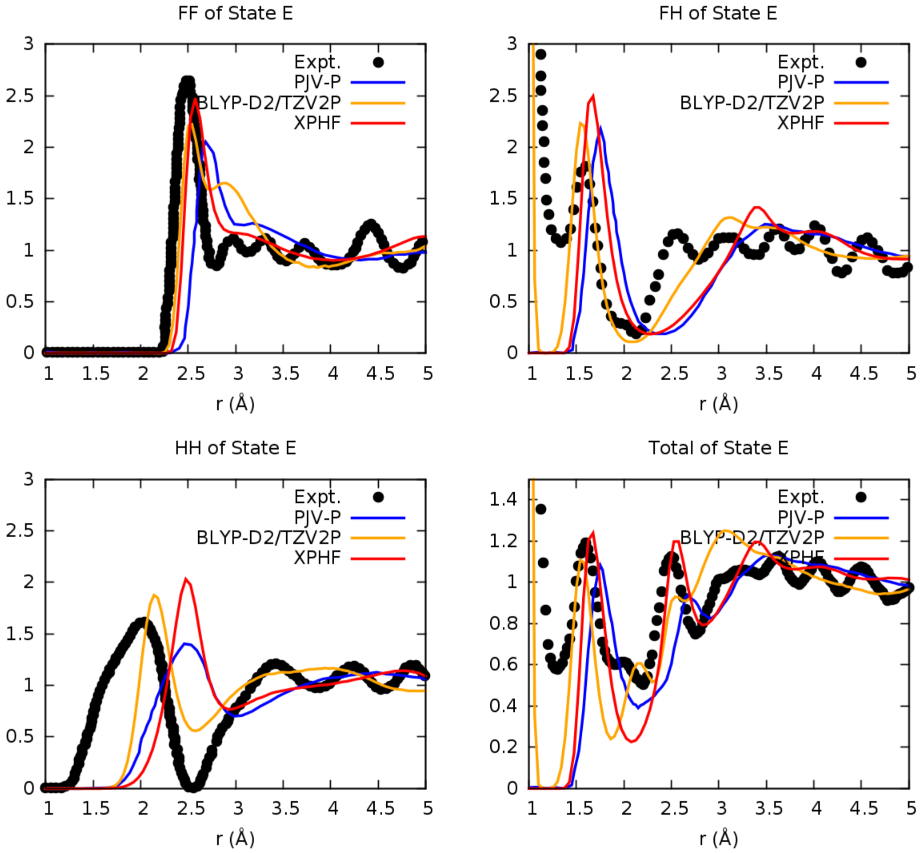


Figure 4.10: A comparison of the partial RDFs of HF at state E (296 K, 1.2 bar), for XPHF (red), PJV-P (blue), BOMD at the BLYP-D2/TZV2P level (orange), and the experiment of McLain and co-workers (black) [26]. XPHF shows agreement with experiment that is comparable to the results of the PJV-P model. Note that the BOMD simulation was performed at a slightly different state of 300 K and 1.0 bar.

A and C, while state B shows some flexibility. Since the temperatures of states A, B, and C are near the experimental freezing point, we suggest that the freezing point temperature predicted by XPHF model might be too high.

State	T (K)	ρ (g/cm^3)	D_{XPHF}	$D_{\text{Expt.}}$
A	195	1.058	0.07	1.00
B	246	1.038	0.63	3.54
C	203	1.176	0.03	1.27
D	273	1.015	5.24	5.72
E	296	0.997	7.19	8.03
F	300	0.962	7.28	8.47
G	373	0.796	21.43	18.45
H	473	0.236	111.10	36.28
I	473	0.398	71.02	–
J	473	0.647	45.37	–
K	473	0.796	34.53	31.95

Table 4.9: Diffusion coefficients in 10^{-9} m^2/s as predicted by XPHF at the 11 state points with comparison to experiment [24]. Experimental values for states A-H and K were determined from the Arrhenius equation $D_{\text{Expt.}} = D_0 \exp(-E_A/RT)$ where $E_A = 9.92$ kJ/mol and $D_0 = 452$ and $D_0 = 398$ 10^{-9}m^2/s for standard vapor pressure (A-H) and 500 Bar (K) respectively.

4.4.7 Deviation in Supercritical States

Pártay and co-workers stated that the poor performance of energetic, density, and structural properties of HF in the supercritical states H-K compared to experiments may be due to an overestimation of the pressure at which the maximum value of the isothermal compressibility curve occurs [21]. This observation was made by noting that the deviation in these properties compared to experiments is much larger at lower pressures than at higher pressures (see Tables 4.6 and 4.8). Shortly following that study, Baburao and Visco suggested that the isothermal compressibility of HF in the supercritical region may exhibit more than one maximum [258], further complicating the understanding of HF in the supercritical states.

We estimated the isothermal compressibility κ for the supercritical states H-K by using

$$\kappa = \frac{1}{\rho}\left(\frac{\partial \rho}{\partial P}\right)_T \approx \frac{1}{\rho}\frac{\Delta \rho}{\Delta P}, \tag{4.16}$$

where ρ and P denote the density and pressure, respectively, at each state for T = 473 K. Taking a one-sided finite difference for the end points (states H and K) and an average of two one-sided finite differences for the middle points (states I and J), we found κ to be 0.0245, 0.0352, 0.007, and 0.003 bar^{-1} for states H, I, J, and K, respectively. This behavior of κ is in accord with the other models for HF (see Figure 5 of Ref. [21]), suggesting that the accuracy of the XPHF model in the supercritical states may also be affected by this type of overestimation.

4.5 Conclusion

We have introduced a fluorine parameter into the PMOw model for use with X-Pol as a QMFF for HF – the XPHF model. XPHF shows good agreement with experiments, and is comparable to the JV-P model and its PJV-P re-parameterization in terms of radial distribution functions, energies per molecule, average dipole moments, and density profiles.

Although the results of the XPHF model are comparable to the PJV-P and JV-P models for most quantities tested, several of the same limitations of those models are present in XPHF. Pártay and co-workers suggested that the r^{-12} term in the Lennard-Jones potential of the PJV-P and JV-P models limits the accuracy of the predicted densities. The XPHF model could be improved by incorporating these exchange-correlation and dispersion effects directly into the Fock matrix using an approach similar to that used by York and co-workers [84]. Further, a two-body correction to the X-Pol method which incorporates charge transfer effects into individual fragments by examining all interactions of fragment pairs could be used [148]. In addition to including charge transfer effects, such an "XP2HF" model could immediately incorporate exchange-correlation and dispersion effects through PMO's pairwise D1 dispersion [95], though such treatment would be empirical. Finally, the semiempirical parameters could be further optimized to give more accurate results, although as in the case of the JV-P and PJV-P models, the improvement would likely not be drastic.

The XPHF model is greatly improved over the previous attempt of modeling HF with X-Pol based on the AM1 model, and our results show that a two-site model for HF can be as accurate as three-site, polarizable models such as PJV-P and JV-P. This further demonstrates the utility of the PMO/X-Pol/DPPC methods for use in force field development beyond the XP3P model for liquid water, suggesting that this approach may be useful as a general framework for a polarizable force field.[1]

[1] We thank Pal Jedlovszky for providing us with the total RDFs of the supercritical states for the TIPS, JV-NP, and JV-P models, and Sylvia McLain for the partial RDFs of the experiment at 296 K and 1.2 bar. We also thank Matthew McGrath and J. Ilja Siepmann for providing us with the BLYP-D2/TZV2P trajectory of 64 HF under the NPT ensemble at 300 K and 1 bar, allowing us to calculate the partial RDFs and make comparisons with that result. This work has been partially supported by National Institutes of Health grants GM46376 and RC1-GM091445. Computations were performed on an SGI Altix cluster acquired through National Institutes of Health grant S10-RR029467.

Chapter 5

MACROSHAKER: A Coarse-Grained Force Field for Crowded Systems of many Proteins

5.1 Introduction

Macromolecular crowding is a characteristic feature of living cells, due to the large size of the confined macromolecules compared to the relatively small, finite cellular compartments that hold them. Even at low concentrations of individual proteins, the excluded volume of the cellular compartments can be high, leading to exceptionally large activity coefficients up to 10^6 [259]. Experimental studies [259–264] have examined the effects of macromolecular crowding on numerous properties. For example, reaction rates *in vivo* can be drastically different from those *in vitro*. According to Ellis [262], "it can be stated with some confidence that many estimates of reaction rates and equilibria made with uncrowded solutions in the test tube differ by orders of magnitude from those of the same reactions operating under crowded conditions within cells." Therefore, it is of paramount importance to include the effects of macromolecular crowding when modeling the biochemical and biophysical processes occurring in living cells.

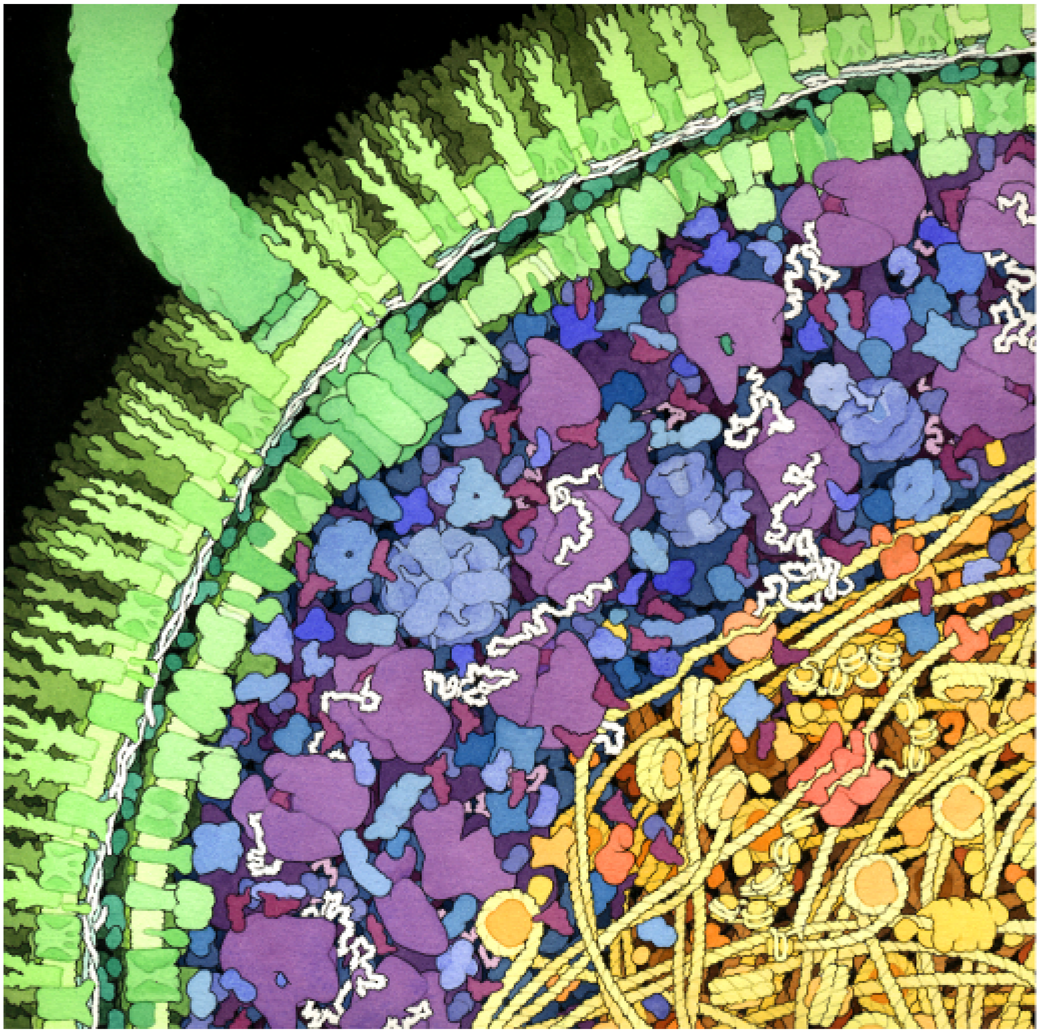


Figure 5.1: An artistic rendition of a cross-section of an *E. coli* cell clearly showing a crowded environment of macromolecules. Reproduced with the permission of D.S. Goodsell, Scripps Research Institute [27].

The vast majority of molecular dynamics (MD) simulations of biological systems to date have focused on the explicit modeling all atoms, which limits the integration time step in the equations of motion to a value on the order of 1 fs. The conditions in such simulations are often close to *in vitro* experiments, making them convenient for comparison with observed data at very low concentrations of proteins. However, such conditions are far from the reality of the crowded compartments in a living cell, as vividly depicted by Goodsell's artistic perception (Figure 5.1) [27]. Furthermore, all-atom MD simulations have prohibitively high computational cost for a system at even a fraction of the size of a cell. Therefore, there is an urgent need for the development of a "mesoscopic" computational model for specific biological processes taking place in the cell under conditions closely resembling the real system that includes all relevant macromolecular particles[1] .

The goal of this study is to develop a theoretical and computational model for the representation of macromolecular particles that can be conveniently used to model a section of the cell, which can adequately describe intermolecular interactions and the execution of the dynamic and reactive trajectories of cellular processes, such as metabolism and signal transduction. As an initial step towards this goal, we introduce a model to reproduce concentration-dependent diffusion coefficients obtained from experiment.

One approach to model macromolecular diffusion and transport is to use Brownian dynamics. Studies such as those described in Refs. [265–273] have demonstrated that Brownian dynamics can be useful for modeling macromolecular transport in cellular organelles. Consequently, our computational model is based on stochastic processes, and uses the Brownian dynamics scheme of Ermak and McCammon [46]. Another

[1] We refer to "large" molecules other than metabolites, enzyme cofactors, single molecules, ions and solute, or a small fragment of polypeptide as macromolecular particles. They include folded proteins, nucleic acids, protein-nucleic acid complexes, protein complexes, ribosomes, and lipid bilayers. For simplicity, at present time, we do not consider intrinsically disordered proteins.

approach we take is the reduced complexity of the macromolecules through the contraction of several groups of atoms into single “beads”, or coarse-graining. Our model incorporates a coarse-graining scheme using a superimposed “uniformly”-spaced grid similar to, but also different from, the approach of Byron [274]. The approximation scheme is nearly volume-preserving, and preserves the center of mass of the original macromolecule. In addition to Brownian dynamics and coarse-graining, we treat the macromolecules as rigid bodies, and use quaternions and the techniques of Bulgac and Adamuţi-Trache for rotational dynamics [275].

Our model, called MACROSHAKER[2] , has been implemented into a software package of the same name, and has a graphical user interface for the generation of configuration files and an interactive visualization environment that renders a scene similar to the artistic works of Goodsell [27]. The remainder of this chapter describes our implementation of dynamics for the diffusive process, our coarse-grained force field, and our graphical user interface with visualization environment. We also provide a comparison of concentration-dependent diffusion coefficients for myoglobin obtained from MACROSHAKER with those obtained from experiments.

[2] MACROSHAKER is a portmanteau of “macromolecule” and “shaker”. The “macromolecule” part comes from the types of molecules we use in simulations, while the “shaker” part comes from the random movements of these particles while undergoing Brownian dynamics. The name MACROSHAKER is not an acronym, but rather is capitalized solely for dramatic effect, in the same fashion as many other scientific programs.

5.2 Dynamics in MACROSHAKER

In all-atom simulations of many proteins, the number of solvent molecules dwarfs the number of proteins by several orders of magnitude, and the detailed, short-range repulsive force requires the use of a small integration time step for the equations of motion. MACROSHAKER attempts to address the solvent problem by employing Brownian dynamics.

The equations of motion for Brownian dynamics have been established since the early works of Einstein in 1905 [276] and Langevin in 1908 [277]. The most widely used form of Brownian dynamics is Langevin's stochastic differential equation, which is often written as Eq. 5.1,

$$m\frac{d^2\vec{r}}{dt^2} = -\xi\frac{d\vec{r}}{dt} - \nabla V(\vec{r}) + \vec{S}(t), \tag{5.1}$$

where m is the mass of the particle, ξ is a damping coefficient related to the diffusion coefficient and solvent viscosity, $V(\vec{r})$ is a potential energy function, and $\vec{S}(t)$ is a stochastic force satisfying the conditions,

$$\left\langle \vec{S}(t) \right\rangle = 0 \quad \text{and} \quad \left\langle \vec{S}(t) \cdot \vec{S}(t') \right\rangle = 6\xi k_B T \delta(t - t'), \tag{5.2}$$

where T is absolute temperature and k_B is Boltzmann's constant.

By inspection, one can see that the Langevin equation is indeed a form of Newton's equation of motion with a stochastic term, which, in our case, represents the effects of the solvent on the macromolecules. The computational benefit of such an approach is clear, and several numerical integration schemes for the Langevin equation have been proposed.

5.2.1 Ermak-McCammon Integration

The most widely used numerical method to perform Brownian dynamics simulations with a large time step was introduced by Ermak and McCammon in 1978 [46]. The Ermak-McCammon scheme arises from the integration of the Langevin equation over a time step $\Delta t >> m/\xi$, which implies $\left\langle md^2\vec{r}/dt^2\right\rangle_{\Delta t} \approx 0$.

Integrating the Langevin equation under the assumption that $\Delta t >> m/\xi$ produces the translational Brownian dynamics scheme of Ermak and McCammon in Eq. 5.5, where $F = -\nabla V$ is the systematic force due to intermolecular interactions of Brownian particles and D_t is the translational diffusion constant.

$$\vec{r}(t+\Delta t) \quad = \quad \vec{r}(t) - \nabla V\frac{\Delta t}{\xi} + \int_{t'}^{t'+\Delta t}\frac{\vec{S}(t)}{\xi}dt' \quad = \quad \vec{r}(t) + \frac{D_t\vec{F}}{k_BT}\Delta t + \vec{R}(\Delta t) \tag{5.3}$$

Here, $F = -\nabla V$ is derived from the total potential energy of the system describing intermolecular interactions, and $R(\Delta t)$ is a term describing the random displacement at each step due to collisions by the implicitly treated solvent molecules. Similar to the Langevin equation, the stochastic displacement term $R(\Delta t)$ of Eq. 5.5 satisfies the conditions given in Eq. 5.4.

$$\left\langle \vec{R}(\Delta t)\right\rangle = 0 \quad \text{and} \quad \left\langle \vec{R}(\Delta t)\cdot\vec{R}(\Delta t)\right\rangle = 6D_t\Delta t. \tag{5.4}$$

The value $D_t = k_BT/\xi$ represents the translational diffusion constant of a given macromolecule where $\xi = 6\pi\eta R$ for a solvent with a viscosity η and the macromolecule with a Stokes' radius R. D_t is molecule dependent and tends to decrease as the mass and/or Stokes' radius of the molecule increases [278].

Strictly speaking, D_t is a $3N \times 3N$ matrix representing a more generalized pairwise-distance dependent *diffusion tensor* used for hydrodynamic interactions [46]. In this more detailed case, the equation describing translational motion becomes

$$\vec{r}_i(t + \Delta t) = \vec{r}_i(t) + \sum_j \frac{\partial D^o_{ij}}{\partial r_j}\Delta t + \sum_j \frac{D^o_{ij} F^o_j}{k_B T}\Delta t + \vec{R}_i(\Delta t), \tag{5.5}$$

where the superscript "o" indicates evaluation at the beginning of each time step. While hydrodynamic effects have been used extensively by others such as Skolnick *et al.* [269, 279–282], the diffusion tensor is regarded as a $3N \times 3N$ diagonal matrix in this work (Eq. 5.6).

$$D^o_{ij} = (k_B T/\xi)\mathbf{I}, \tag{5.6}$$

Nevertheless, an implementation of the commonly-used Rotne-Prager-Yamakawa (Eq. 5.7) [283] and Rotne-Prager (Eq. 5.8) [284] diffusion tensors for non-overlapping ($r_{ab} \geq (R_a + R_b)$) and overlapping ($r_{ab} < (R_a + R_b)$) particles, respectively, is available to use with MACROSHAKER.

$$D^o_{ij} = \left[\mathbf{I} + (\vec{r}_{ab}\vec{r}^{\,T}_{ab})/r^2_{ab} + (2R^2/3r^2_{ab})(\mathbf{I} - 3(\vec{r}_{ab}\vec{r}^{\,T}_{ab})/r^2_{ab})\right]/(8\pi\eta r_{ab}), \tag{5.7}$$

for non-overlapping spheres, and

$$D^o_{ij} = (k_B T/\xi)\left[(1 - 9r_{ab}/32R)\mathbf{I} + (3/32R)(\vec{r}_{ab}\vec{r}^{\,T}_{ab})/r^2_{ab}\right], \tag{5.8}$$

for overlapping spheres.

The derivation and equation for rotational Brownian dynamics under the Ermak-McCammon scheme is analogous to that of translational motions with the substitution of the angular counterparts, notably the torque $\tau = r \times \vec{F}$. Eq. 5.9 gives the Ermak-McCammon scheme for rotational motions satisfying the conditions in Eq. 5.10, where the value $D_r = k_B T / 8\pi\eta R^3$ is called the rotational diffusion constant.

$$\vec{\theta}(t + \Delta t) = \vec{\theta}(t) + \frac{D_r \vec{\tau}}{k_B T} \Delta t + \vec{\Theta}(\Delta t) \tag{5.9}$$

$$\left\langle \vec{\Theta}(\Delta t) \right\rangle = 0 \quad \text{and} \quad \left\langle \vec{\Theta}(\Delta t) \cdot \vec{\Theta}(\Delta t) \right\rangle = 6 D_r \Delta t. \tag{5.10}$$

At this point, we have not addressed the details of the stochastic terms used in the dynamics scheme. These terms are determined such that the mean value is zero and the variance satisfies the fluctuation-dissipation theorem (Eqs. 5.4 and 5.10). Numerically, they are represented as independent random variables from a normal distribution.

We have used a Box-Muller transform to generate random normally-distributed numbers from random, uniformly-distributed numbers [285]. In particular, for independent random variables U_1 and U_2 from the same uniform distribution on the interval $(0, 1)$, X_1 and X_2 give a pair of independent random variables from the same normal distribution with mean zero and unit variance (Eq. 5.11).

$$\begin{aligned} X_1 &= \sqrt{-2 \ln U_1} \cos(2\pi U_2) \\ X_2 &= \sqrt{-2 \ln U_1} \sin(2\pi U_2) \end{aligned} \tag{5.11}$$

After generation, X_1 and X_2 can be multiplied by an appropriate scale-factor to enforce the necessary constraints on the stochastic terms seen in Eqs. 5.4 and 5.10.

5.2.2 Rigid-body Mechanics

In MACROSHAKER, the coarse-grained macromolecules are treated as rigid bodies during simulations. The data structure storing the coarse-grained model contains the center of mass of the macromolecule and the coordinates of the beads relative to that

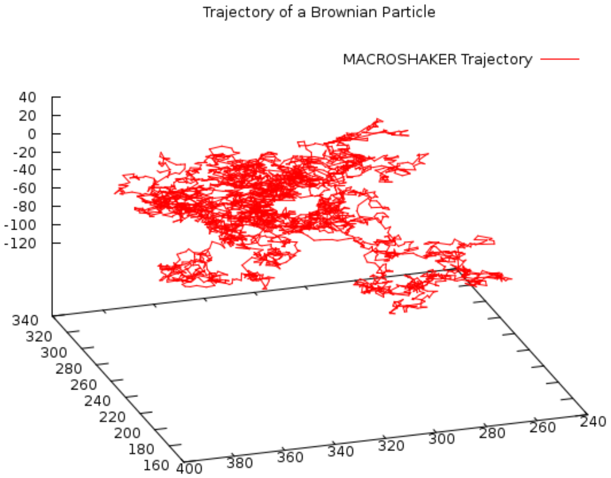


Figure 5.2: The Cartesian coordinates of the center of mass of a myoglobin under Ermak-McCammon dynamics over a period of 1 μs.

center. This allows for translational motions described by simple center of mass dynamics, as accomplished by Eq. 5.5.

Rotational motions of the macromolecules are more complicated than translational motions. We use a conversion from $Z - X' - Z''$ Euler angles to quaternions at each step to account for the rotational motion seen in Eq. 5.9. The rotation matrices and quaternion conversion formulas follow the pattern of those in Ref. [275]. The $Z - X' - Z''$ Euler angles α, β, and γ describing the initial orientation of a molecule can be converted to a quaternion $\vec{q} = q_0 + iq_1 + jq_2 + kq_3$, where $\sum_{m=0}^{3}(q_m)^2 = 1$, using Eq. 5.12.

$$\begin{aligned}
q_0 &= \cos(\beta/2)\cos\left[(\alpha+\gamma)/2\right] \\
q_1 &= \sin(\beta/2)\cos\left[(\alpha-\gamma)/2\right] \\
q_2 &= \sin(\beta/2)\sin\left[(\alpha-\gamma)/2\right] \\
q_3 &= \cos(\beta/2)\sin\left[(\alpha+\gamma)/2\right]
\end{aligned} \tag{5.12}$$

Since rotational motion accumulates with each rotation, it is necessary to consider only the change in angular orientation at each time step of the Brownian dynamics simulation, rather than the actual orientation of the rigid molecule. This can be done by using Eq. 5.13 as opposed to Eq. 5.9 which describes the actual orientation at each t.

$$d\vec{\theta}(t + \Delta t) = \frac{D_r \vec{\tau}}{k_B T} \Delta t + \vec{\Theta}(\Delta t) \tag{5.13}$$

These changes in angular orientation can be transformed to a change in quaternions using techniques similar to those found in Ref. [275]. The result is the matrix-vector product below.

$$\begin{pmatrix} dq_0 \\ dq_1 \\ dq_2 \\ dq_3 \end{pmatrix} = \frac{1}{2} \begin{pmatrix} q_0 & -q_1 & -q_2 & -q_3 \\ q_1 & q_0 & -q_3 & q_2 \\ q_2 & q_3 & q_0 & -q_1 \\ q_3 & -q_2 & q_1 & q_0 \end{pmatrix} \begin{pmatrix} 0 \\ d\theta_x \\ d\theta_y \\ d\theta_z \end{pmatrix} \tag{5.14}$$

Eq. 5.14 is applied for each rigid-body macromolecule at each time step, producing a time-dependent quaternion, describing its orientation, by $q(t + \Delta t) = \vec{q}(t) + d\vec{q}$. This results in an extra step to ensure that the updated time-dependent quaternion has been normalized by applying the constraint $\sum_{m=0}^{3} (q_m(t + \Delta t))^2 = 1$.

The application of Eqs. 5.13 and 5.14 and the renormalization of quaternions allows us to produce a 3×3 rotation matrix for the rotation of the bead position vectors in $\mathbb{R}^3$, relative to the center of mass, of each coarse-grained macromolecule.

$$\begin{pmatrix} q_0^2 + q_1^2 - q_2^2 - q_3^2 & 2(q_1q_2 + q_0q_3) & 2(q_1q_3 - q_0q_2) \\ 2(q_1q_2 - q_0q_3) & q_0^2 - q_1^2 + q_2^2 - q_3^2 & 2(q_2q_3 + q_0q_1) \\ 2(q_1q_3 + q_0q_2) & 2(q_2q_3 - q_0q_1) & q_0^2 - q_1^2 - q_2^2 + q_3^2 \end{pmatrix} \tag{5.15}$$

This rotation matrix is orthogonal, and can be used to rotate each bead position vector as needed. Such a scheme is preferred to the permanent rotation of the position vectors which can lead to numerical instability [275]. It is also useful in practice considering that one coarse-grained model can be stored in the memory while multiple instantiations of the model can be created by only describing its center of mass location and angular orientation.

5.3 Coarse-Grained Force Field

Due to the n-body nature of atomistic MD simulations, the cost of calculating the intermolecular forces in an MD simulation is naïvely $\mathcal{O}(n^2)$. For large n, the calculation of these forces is expensive. One method of reducing the cost of the n-body problem is to reduce n by replacing clusters of atoms with single "beads". This technique, called *coarse graining*, has become popular in recent years for representation of lipids, macromolecules, proteins, and nanoparticles. Levitt and Warshel introduced the first coarse-grained model for proteins, based on amino acid residues, in 1975 to model the folding process of the bovine pancreatic trypsin inhibitor (BPTI) [40]. The following year, Tanaka and Scheraga used their own residue based model for studying protein folding [41]. Other work includes that of Smit *et al.*, who developed a simple coarse-graining approach for the representation of oil and water particles for simulating mixing at their interface in 1990 [42]. More recent work has focused on coarse-graining with the explicit preservation of physical characteristics of the original particles such as center of mass, moment of inertia, and certain thermodynamic properties by researchers such as Voth *et al.* [43]

Our coarse-graining method in MACROSHAKER is based upon the work of Byron [274], and uses the superimposition of a "uniformly"-spaced grid to determine the atomic clusters to be replaced by a single interaction site called a *bead*. However, the present approach is fundamentally different in that it explicitly preserves the center of mass of the original particle while heuristically preserving the volume of the molecular system.

The structural information of macromolecules in MACROSHAKER is obtained from PDBs [286] hosted by the RCSB Protein Data Bank [287]. Each atom is treated as a sphere of its van der Waals radius obtained from the X-ray diffraction experiments of Bondi [288]. For the purpose of our rigid-body dynamics simulations, the relevant information needed for each atom consists of its position $\vec{p}$, its van der Waals radius r, its mass m, and its charge q. Our coarse-graining method seeks to re-create these properties for a single coarse-grained bead representing a cluster of atoms.

5.3.1 Coarse-Grained Model

We now describe the current coarse-graining method in MACROSHAKER. A nearly uniformly-spaced grid of spacing w is superimposed on the smallest bounding box containing the molecule in question. The bottom-left-far corner coordinates and the top-right-near corner coordinates of the bounding box are stored as vectors $\vec{B}$ and $\vec{T}$, respectively; $\vec{B}$ stores the minimum x, y, and z coordinates from all atoms, while $\vec{T}$ stores the maximum x, y, and z coordinates from all atoms. The number and types of atoms in each grid entry are recorded, and a resulting coarse-grained approximation is formed. This process is illustrated in Figure 5.3 on the crystal structure of a

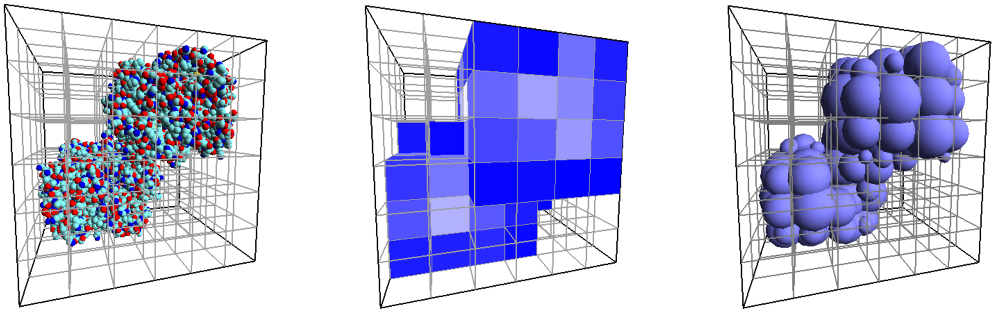


Figure 5.3: The van der Waals surface of a deoxyhemoglobin S (PDB: 2HBS) (left; 8760 heavy atoms), the bounding volume and grid representation formed from using a spacing of $w = 15$ Å(center), and the result of contracting all atoms in a grid entry to a single sphere (right: 107 spheres)

deoxyhemoglobin S dimer [289].

Due to the use of a fixed bounding box, which can have different integer numbers of elements in each dimension, the grid is said to be nearly uniformly-spaced. For a desired grid spacing w, we determine the number of grid entries in the x, y, and z directions using Eq. 5.16. After the dimensions of the grid are calculated, the new grid spacings in each direction d_x, d_y, and d_z are computed using Eq. 5.17. The use of the floor function in Eq. 5.16 ensures that the grid dimensions are of an integer number, and results in spacings d_x, d_y, and d_z all greater than or equal to w.

$$n_x = \lfloor \frac{T_x - B_x}{w} \rfloor, \quad n_y = \lfloor \frac{T_y - B_y}{w} \rfloor, \quad n_z = \lfloor \frac{T_z - B_z}{w} \rfloor. \tag{5.16}$$

$$d_x = \frac{T_x - B_x}{n_x}, \quad d_y = \frac{T_y - B_y}{n_y}, \quad d_z = \frac{T_z - B_z}{n_z}. \tag{5.17}$$

After the computation of the dimensions and new spacings, each entry of the grid is visited and a list of atoms whose centers lie in each entry are recorded. These lists

ALA	ARG	ASN	ASP	CYS	GLU	GLN	GLY	HIS	ISO
0	1	0	-1	0	-1	0	0	0	0
LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
0	1	0	0	0	0	0	0	0	0

Table 5.1: The net charges of each amino acid residue used by MACROSHAKER in the creation of coarse-grained models. Charges are centered on the α-carbon of each residue and are summed to determine the charge on the bead that contains them.

are used to compute the mass contained in each grid entry, the net charge of each grid entry, as well as a bounding volume formed from the sum of the volume of those grid entries which contain atoms (center pane of Figure 5.3). These three quantities are given in Eq. 5.18 as M, Q, and V_{tot} respectively, where N is the number of atoms in a particular grid entry and N_G is the total number of grid entries which contain at least one atom.

$$M = \sum_{k=1}^{N} m_k, \quad Q = \sum_{k=1}^{N} q_k, \quad V_{\text{tot}} = d_x d_y d_z N_G \tag{5.18}$$

For the sake of simplicity, charges on the beads are currently determined by summing the charges on the α-carbon sites[3] located within each bead according to Table 5.1 A more sophisticated scheme for determining charges on the beads is discussed in Appendix A.

The recording of the lists of atoms in each grid entry not only allows us to find the quantities M and Q for each entry in the grid, but also the location of the center of mass for each grid entry containing at least one atom. By placing each coarse-grained bead with mass M at the center of mass of its grid entry $\vec{P}$, we preserve the center of mass of the entire system. The center of mass is calculated for each non-empty grid entry

[3] At the present time, histidine (HIS) is considered to be neutrally charged. In reality, HIS can be either neutral or positive depending on the residues it interacts with and its exposure to the solvent. This remains an issue to be addressed in the future.

using Eq. 5.19, where $\vec{p}_k$ denotes the position of an individual atom.

$$\vec{P} = \frac{1}{M} \sum_{k=1}^{N} m_k \vec{p}_k \tag{5.19}$$

The next step in the method is to determine the radius of the coarse-grained bead. The radius R of the coarse-grained bead is related to the total mass M_{tot} and total grid-based bounding volume V_{tot} of the macromolecule. The radius R is determined by relating the total volume derived from the non-empty grid entries to the percentage of mass contained within a particular grid entry, given by M/M_{tot}. We compute R using Eq. 5.20.

$$R = \left(\left(\frac{3}{4\pi} \right) \left(\frac{M}{M_{\text{tot}}} V_{\text{tot}} \right) \right)^{(1/3)} \tag{5.20}$$

The bead radius R is identical to the σ parameter used in our dispersion potential (see next section).

Once the quantities of Eqs. 5.18, 5.19, and 5.20 are known, the final step is to initialize and allocate the memory for a data structure containing them. The total number of resulting coarse-grained beads is N_G.

Figure 5.4 shows several renderings of the grid representation and coarse-grained representation of a deoxyhemoglobin S dimer containing 8760 heavy atoms [289]. The results seen in Figure 5.4 indicate that the present coarse-graining method not only significantly reduces the number of interaction sites compared to the number of atoms, but also preserves molecular shape for large grid spacings.

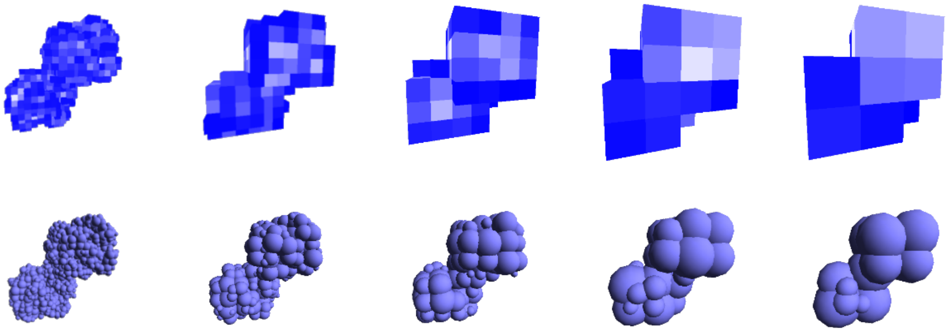


Figure 5.4: The uniform grid representation with spacings $w = 5$, 10, 15, 20, and 25, respectively (top), and the corresponding coarse-grained models with 1592, 269, 107, 37, and 21 sites, respectively (bottom).

5.3.2 Potential Energy

A force field is only as good as the potential energy function which describes it. It is well known that the repulsive wall between two particles A and B in close proximity, originating from Pauli exchange, has a great influence on the overall structure of condensed-phase systems [290]. In classical MD simulations, this repulsive wall is often represented by a 12-6 Lennard-Jones potential, which has a singularity at $r_{AB} = 0$. While traditional MD simulations use a small enough time step to avoid this singularity by forcing a gradual climb of the repulsive wall at very short distances, the random displacement in Brownian dynamics along with its larger time step make $r_{AB} \approx 0$ accessible during the simulation. Consequently, the component of the potential which describes the repulsive force must be softer and have a finite value at $r_{AB} = 0$.

Although the ideal functional form of the potential is unknown in our case, we suggest a simple harmonic potential based upon the overlap of Stokes' spheres of macromolecules A and B to represent Pauli exchange-repulsion,

$$V^{\text{Exch.}}(r_{AB}) = \begin{cases} \frac{c}{2}\left(\frac{r_{AB}-(R_A+R_B)}{(R_A+R_B)}\right)^2 & \text{when} \quad r_{AB} \leq (R_A+R_B), \\ 0 & \text{when} \quad r_{AB} > (R_A+R_B), \end{cases} \tag{5.21}$$

where c is a constant fit to reproduce the concentration-dependent diffusion coefficients, and R_A and R_B are the Stokes' radii determined by the Stokes-Einstein relation $D_t = k_B T / 6\pi\eta R$.

In contrast to the case of the repulsive potential, the attractive component of the 12-6 Lennard-Jones potential, which begins at $r_{ab} = 2^{1/6}\sigma_{ab}$, is free of singularity. MACROSHAKER retains the attractive part of the Lennard-Jones potential as a description of dispersion between beads a and b,

$$V^{\text{Disp.}}(r_{ab}) = \begin{cases} 4\epsilon_{ab}\left[\left(\frac{\sigma_{ab}}{r_{ab}}\right)^{12} - \left(\frac{\sigma_{ab}}{r_{ab}}\right)^{6}\right] & \text{when} \quad r_{ab} \geq 2^{\frac{1}{6}}\sigma_{ab}, \\ -\epsilon & \text{when} \quad r_{ab} < 2^{\frac{1}{6}}\sigma_{ab}, \end{cases} \tag{5.22}$$

where $\sigma_{ab} = \sigma_a + \sigma_b$ with σ_a and σ_b equal to the radius of the beads from the grid-based coarse-graining model (i.e. $\sigma = R$ of Eq. 5.20) and $\epsilon_{ab} = (\epsilon_a + \epsilon_b)/2$ with ϵ_a and ϵ_b chosen to fit the concentration-dependent diffusion coefficients.

Together, Eqs. 5.21 and 5.22 give a description of the short-range repulsive and long-range dispersion interactions between two coarse-grained macromolecules. What remains is a description of an intermolecular electrostatic potential that takes the screening effects of the implicit solvent into consideration. Using the assumption that a solution of water and ions surrounds each macromolecule and a series of approximations to the Poisson equation outlined in Appendix A, the Debye-Hückel equation (Eq. 5.23) [291] emerges,

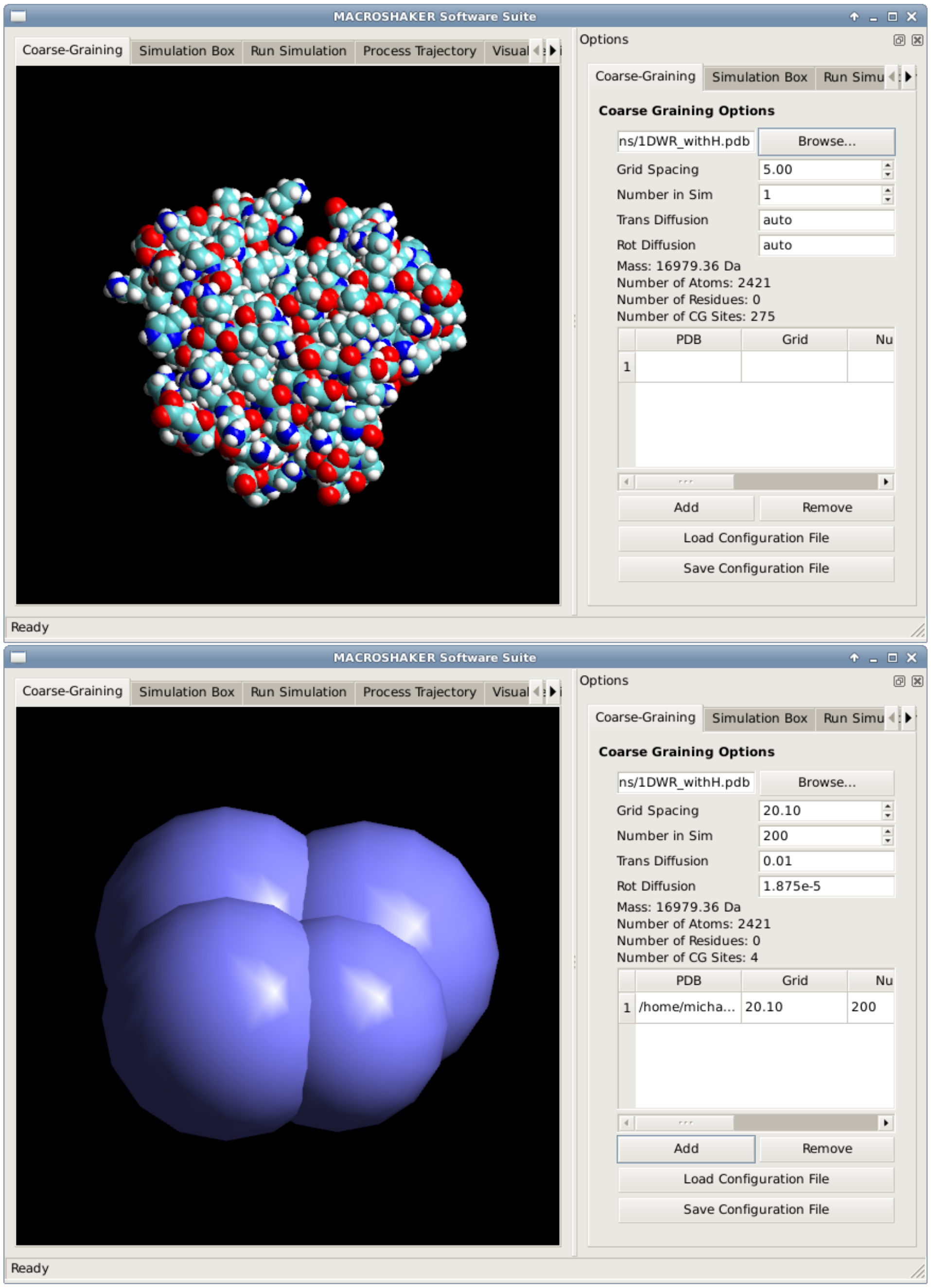
$$V^{\text{Elec.}}(r_{ab}) = \frac{q_a q_b e^{-\kappa r_{ab}}}{4\pi\epsilon_r\epsilon_0\sqrt{r_{ab}^2 + \alpha}} \tag{5.23}$$

where q_a and q_b represent the charges of beads a and b, ϵ_r and ϵ_0 are the dielectric constant of the solvent and vacuum permittivity, κ^{-1} is the Debye-Hückel screening length, which is related to the ionic concentration, and $\alpha = 0.1\text{Å}^2$ is used to avoid the singularity caused by overlapping beads.

In the end, the net force acting on each bead is projected into components in the direction to the center of mass for the force involved in translational motion, and an orthogonal component for the torque involved in rotational motion.

5.4 Graphical User Interface

A graphical user interface (GUI) for generating the initial configurations and setting up simulation conditions and for visualization of the Brownian dynamics trajectories was written in C++ using OpenGL and the Qt toolkit. The graphical interface has the capability to load PDB files for constructing coarse-grained models (Figure 5.5). The user can select a protein or a nucleic acid structure from the PDB, and can choose the number of molecules to be placed in the simulation box. The size of the simulation box can be specified in the GUI, for which randomized positions and orientations of the macromolecular particles can be generated (Figure 5.6). Equilibrium steps can be run within the GUI prior to saving the initial coordinates if desired. Simulations using



Coarse Graining Options

ns/1DWR_withH.pdb

Grid Spacing 5.00

Number in Sim 1

Trans Diffusion auto

Rot Diffusion auto

Mass: 16979.36 Da
Number of Atoms: 2421
Number of Residues: 0
Number of CG Sites: 275

	PDB	Grid	Nu
1			

Coarse Graining Options

ns/1DWR_withH.pdb

Grid Spacing 20.10

Number in Sim 200

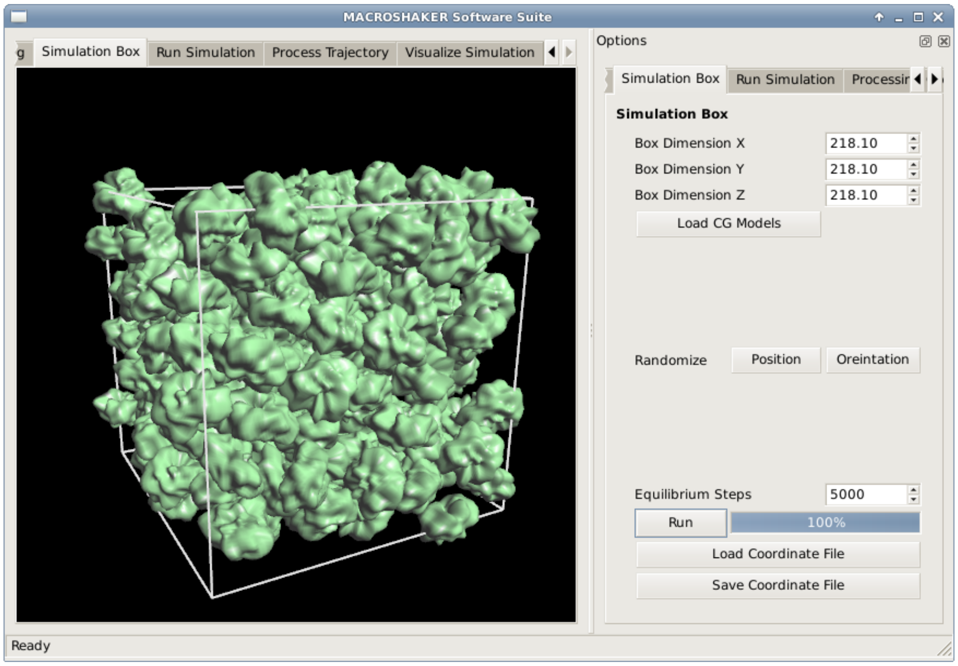
Trans Diffusion 0.01

Rot Diffusion 1.875e-5

Mass: 16979.36 Da
Number of Atoms: 2421
Number of Residues: 0
Number of CG Sites: 4

	PDB	Grid	Nu
1	/home/micha...	20.10	200

Figure 5.5: The component of the MACROSHAKER GUI which selects the level of coarse-graining to be used. A PDB is first loaded (top) and the grid size is specified along with the number of copies to use in the simulation (bottom).



Simulation Box

Box Dimension X 218.10

Box Dimension Y 218.10

Box Dimension Z 218.10

Randomize

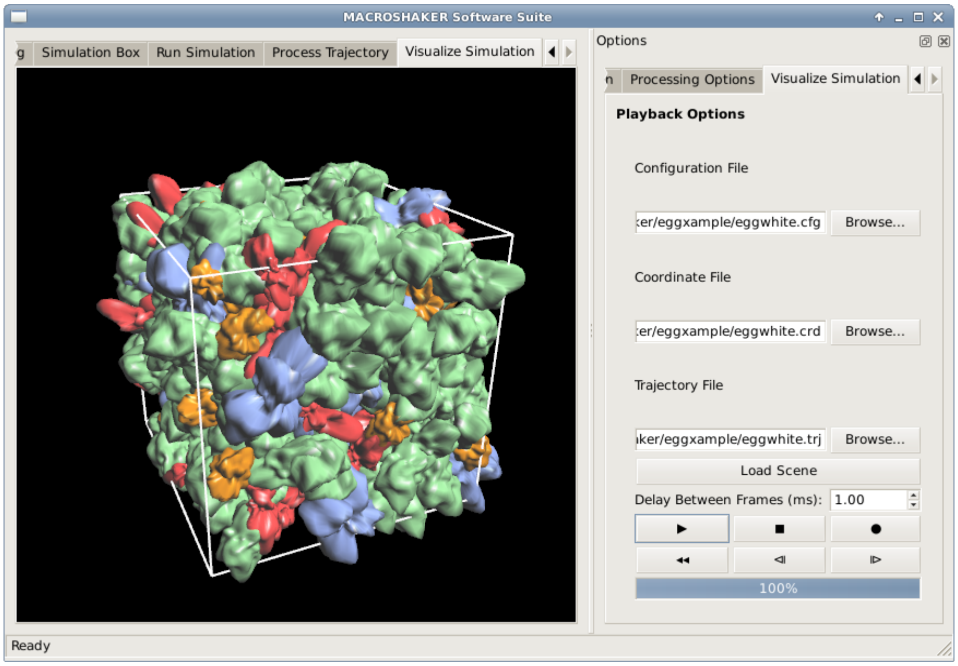
Equilibrium Steps 5000

Figure 5.6: The component of the MACROSHAKER GUI in which the box size is entered and the initial coordinates are randomized; equilibrium steps can be run to reduce steric clashes and visualized in realtime in this part of the GUI.

the MACROSHAKER CGFF can be run within the GUI or on the command line after saving the configuration files. Trajectories of completed simulations can be visualized in the GUI and saved to a video file if desired (Figure 5.7).

5.5 Visualization in MACROSHAKER

The Brownian dynamics trajectory generated from simulations using MACROSHAKER can be visualized by using spherical harmonic expansions of the macromolecular surfaces. Such representation of macromolecules provides not only a smooth and aesthetically pleasing surface, but also allows for evaluating surface properties such as normal vectors and principal curvature. Our implementation makes use of a modified



Configuration File: er/eggxample/eggwhite.cfg

Coordinate File: er/eggxample/eggwhite.crd

Trajectory File: ker/eggxample/eggwhite.trj

Delay Between Frames (ms): 1.00

Ready

Figure 5.7: The visualization component of the MACROSHAKER GUI showing a box of several of the proteins in albumen. Trajectories can be played back within the interface or exported to video files.

approach of Duncan and Olson [292] with non-photorealistic, cartoon-like rendering techniques similar to those of Decaudin [293]. The coupling of spherical harmonic surfaces and cartoon-like rendering techniques allows us to achieve a final result visually similar to the widely known artistic works of Goodsell [27].

5.5.1 Spherical Harmonic Expansions

Arising from the solution of Laplace's equation in spherical coordinates, the spherical harmonic expansion is the spherical coordinate analog of the widely used Fourier series expansion. As Fourier series expansions require a one-to-one set of input points in Cartesian coordinates for an accurate expansion, the same is true for spherical harmonic expansions in the case of spherical coordinates. For any one-to-one surface S in the spherical coordinate system, often called a *star-like surface*, there exists a spherical harmonic expansion (Eq. 5.24).

$$S(\theta, \phi) = \sum_{l=0}^{\infty} \sum_{m=-l}^{m=l} c_{lm} Y_l^m(\theta, \phi) \tag{5.24}$$

The c_{lm} and the Y_l^m of Eq. 5.24 are the coefficients and basis functions of the expansion respectively. The basis functions Y_l^m are defined in Eq. 5.25, where P_l^m is a real-valued, m-th order Legendre polynomial of the l-th kind.

$$Y_l^m(\theta, \phi) = \sqrt{\frac{2l+1}{4\pi} \frac{(l-m)!}{(l+m)!}} P_l^m(\cos\theta) e^{im\phi} \tag{5.25}$$

Several methods exist for evaluating the Legendre polynomials of Eq. 5.25, one of which is discussed in Ref. [294].

As the spherical coordinate analog of the Fourier series expansion, the spherical harmonic expansion retains the useful property that its basis functions are orthogonal under the L^2 inner product. This allows for the determination of an arbitrary coefficient c_{lm} by computing the integral of Eq. 5.26.

$$c_{lm} = \int_0^{2\pi} \int_0^{\pi} S(\theta, \phi) \overline{Y_l^m}(\theta, \phi) \sin\theta d\theta d\phi \tag{5.26}$$

In general, the coefficients c_{lm} and the basis Y_l^m functions are complex-valued. However, since the original surface $S \in \mathbb{R}^3$ is real-valued, it is often desirable to use what are called the *real spherical harmonic basis functions*. These real basis functions are merely linear combinations of the complex-valued basis functions Y_l^m. Consequently, the above analysis remains the same, albeit with a different definition of Y_l^m [292].

Due to the finiteness of computing resources, the infinite sum of Eq. 5.24 is truncated to a relatively small value L which produces an approximation R to the original surface S. Thus, it is only necessary to determine $(L+1)^2$ coefficients, all of which may be precomputed and stored for the evaluation of points on the approximated surface given by R.

$$S(\theta, \phi) \approx R(\theta, \phi) = \sum_{l=0}^{L} \sum_{m=-l}^{m=l} c_{lm} Y_l^m(\theta, \phi) \tag{5.27}$$

The approximation R to the original surface S can be used to obtain vertices for use in a triangular mesh by using the usual formula to convert between spherical and Cartesian coordinates.

$$\begin{pmatrix} x(\theta, \phi) \\ y(\theta, \phi) \\ z(\theta, \phi) \end{pmatrix} = \begin{pmatrix} R(\theta, \phi) \sin\theta \cos\phi \\ R(\theta, \phi) \sin\theta \sin\phi \\ R(\theta, \phi) \cos\theta \end{pmatrix}. \tag{5.28}$$

In order to compute the coefficients using Eq. 5.26, a one-to-one input surface S with respect to θ and ϕ is required. Typically, solvent-accessible surfaces are used; here,

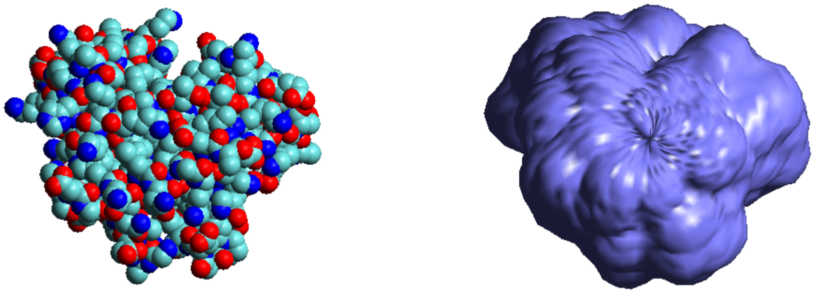


Figure 5.8: The van der Waals surface of myoglobin and the corresponding one-to-one surface from a ray casting procedure.

we propose a technique in the spirit of ray casting [295] for determining the outermost point of a macromolecule by finding the furthest ray-sphere intersection on its van der Waals surface using a ray starting from the center of mass.

A ray $\vec{r}(t)$ starting at the center of mass of a macromolecule p with unit-length direction $\vec{d}$ takes the form of Eq. 5.29 for $t \geq 0$. Similarly, a sphere of radius ρ centered about the point s takes the form of Eq. 5.30. For a given pair of angles (θ, ϕ), Eq. 5.28 can be used to find $\vec{d} = (d_x, d_y, d_z)$ by taking $R(\theta, \phi) = 1$, and ρ can be determined from the van der Waals radius of each atom.

$$\vec{r}(t) = t\vec{d} + p = t(d_x, d_y, d_z) + (p_x, p_y, p_z) \tag{5.29}$$

$$(x - s_x)^2 + (y - s_y)^2 + (z - s_z)^2 = \rho^2 \tag{5.30}$$

The composition of Eq. 5.29 and Eq. 5.30 results in Eq. 5.31.

$$(td_x + (p_x - s_x))^2 + (td_y + (p_y - s_y))^2 + (td_z + (p_z - s_z))^2 = \rho^2 \tag{5.31}$$

Eq. 5.31 is quadratic in t, and can be solved to yield,

$$t = \frac{-B \pm \sqrt{B^2 - 4AC}}{2A}, \tag{5.32}$$

where

$$A = \|\vec{d}\|^2 \quad , \quad B = 2(\vec{d} \cdot (p - s)) \quad , \quad C = \|p - s\|^2 - \rho^2. \tag{5.33}$$

For all possible (θ, ϕ), the greatest intersection time over all atoms is stored. The intersection point corresponding to the greatest time is guaranteed to lie on the van der Waals surface. The maximum ray-sphere intersection time t_{max} is obtained for each (θ, ϕ) pair, and the corresponding intersection point $(\|\vec{r}(t_{\text{max}})\|, \theta, \phi)$ is stored. If no valid intersection is found, the intersection point is set to $(0, \theta, \phi)^4$. This scheme yields a one-to-one surface $S(\theta, \phi) \in \mathbb{R}^3$ which is suitable for use in obtaining a spherical harmonic expansion.

Figure 5.9 shows several renderings of spherical harmonic expansions of a Mb [25] from the RCSB Protein Data Bank. The L values increase from left to right and top to bottom respectively. All renderings were created with MACROSHAKER. As can be seen, the characteristic shape of the protein is clearly identifiable when $L = 5$, and greater details are well presented with $L = 10$. At this level, a total of 121 coefficients are needed in the spherical harmonic expansion.

[4] This step can be modified to trace a sphere along those rays which have no intersection, stopping at a value of t_{max} where no atom is partially contained in the casted sphere. This results in a much smoother surface, but has a larger volume and surface area. Such surfaces were used in the creation of the figures.

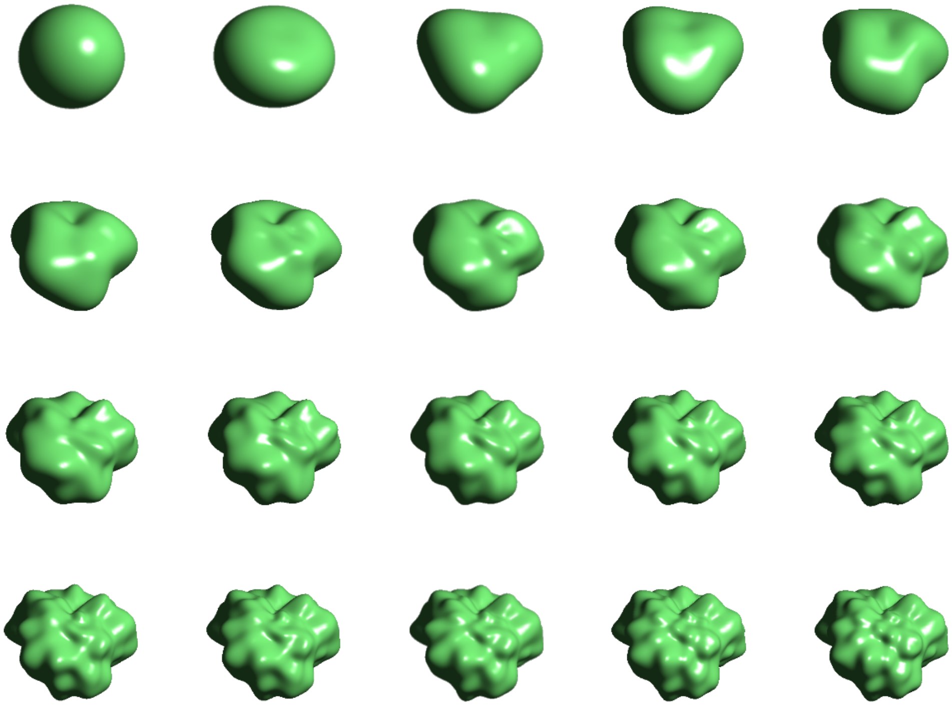


Figure 5.9: Spherical harmonic surfaces of a Mb from $L = 0$ (upper-left) to $L = 19$ (lower-right).

5.5.2 Non-photorealistic Rendering

Macromolecules are quite large by molecular standards, but are still small enough to be invisible to the naked eye, and lack intrinsic color. Consequently, visualizations of macromolecules which do not shade models based upon electrostatic potential or similar properties merely attempt to produce an aesthetically pleasing rendering with respect to color and shading. As a result, we seek to develop a shading model similar to the hand drawings of macromolecules by Goodsell [27].

Many classical reflectance models make use of Lambertian reflectance for diffuse shading [296]. Such shading tends to be very smooth for smooth surfaces due to a varying surface normal. However, cartoon-like drawings tend to use very few colors and relatively little shading variation despite the curved nature of surfaces and varying surface normals. In 1996, Decaudin [293] proposed a technique for creating cartoon-like renderings by discretizing a smoothly shaded surface into regions of specific color intensities. Such techniques are commonly called *cel-shading*. Our cel-shading technique is similar to that of Decaudin, but differs in the criteria for discretization and ignores other aspects such as shadows.

According to the Lambertian reflectance model [296], the diffuse reflectance intensity is given by the magnitude of the projection of the normalized light vector $\vec{L}$ onto the surface normal $\vec{N}$. Since the two vectors are normalized, the magnitude of this projection can be thought of as the percentage of light reflecting off the surface back to the eye. An easy way to achieve a cartoon-like rendering is through binning of these percentages. For example, if a reflectance percentage is greater than 50%, the percentage is set to 100%, and if a reflectance percentage is less than or equal to 50%, the percentage is set to 50%. Such a scheme guarantees only two possible intensity values for the reflectance, which results in a cartoon-like rendering. This can be represented

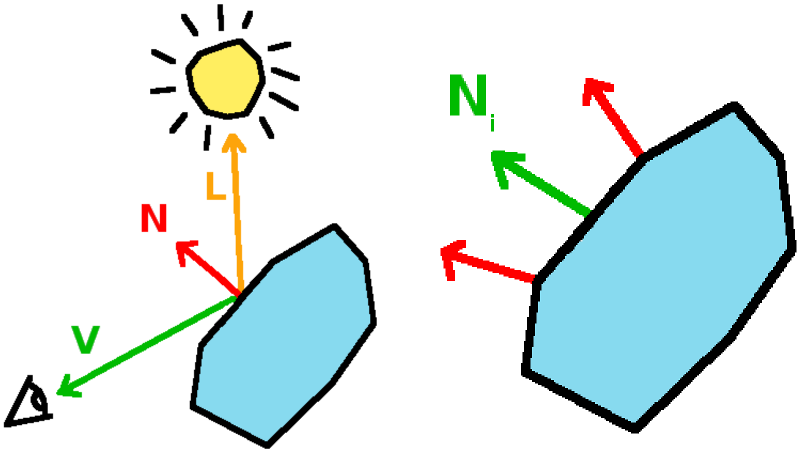


Figure 5.10: Definition of vectors $\vec{V}$, $\vec{N}$, and $\vec{L}$ used in shading (left), and the interpolated normals $\vec{N}_i$ (right).

mathematically by Eq. 5.34.

$$I_d = \lfloor \max\{\vec{L}\cdot\vec{N}, 0\} + 0.5 \rfloor \tag{5.34}$$

Another aspect of hand drawn figures is that of outlines. While there are several methods for determining outlines [293], a simple (but naïve) way is to consider the dot-product of the view vector $\vec{V}$ and the surface normal $\vec{N}$. Points that directly form the edge of a smooth surface satisfy the condition $\vec{V}\cdot\vec{N} = 0$, while those not on the edge will have $\vec{V}\cdot\vec{N} \neq 0$. However, since surfaces are drawn from a discrete number of polygons and the interpolation of vertex normals is not exact, it is wise to use a threshold value to better determine where edges are present.

Consequently, we create a multiplier taking a 0 or 1 value based upon a threshold h which we call the "edge fraction". This edge fraction can be multiplied by the diffuse reflectance intensity to create dark places where there are edges, while leaving the surrounding non-edge places unaffected. We also note that the choice of h is related to the deviation allowed from the orthogonality of $\vec{V}$ and $\vec{N}$, and is given by the expression for ψ in Eq. 5.36. Choosing a value of $h = 0.25$ gives $\psi = 90^\circ - \cos^{-1}(0.25) \approx 14.48^\circ$.

$$E_f = \lfloor \max\{\vec{V} \cdot \vec{N}, 0\} + (1 - h) \rfloor \tag{5.35}$$

$$\psi = 90^\circ - \cos^{-1}(h) \tag{5.36}$$

Our implementation is for a scene with M light sources where I_a is an ambient intensity is given below.

$$I \quad = \quad E_f(I_a + I_d) \quad = \quad \lfloor \max\{\vec{V} \cdot \vec{N}, 0\} + 0.75 \rfloor \left(I_a + \lfloor 0.5 + \sum_{j=1}^{M} \max\{\vec{L_j} \cdot \vec{N}, 0\} \rfloor \right) \tag{5.37}$$

In practice, this formula is "clamped" between the values of 0 and 1 at various stages to avoid calculating intensities greater than $I = 1$.

We implemented our cartoon-like cel-shading model using the OpenGL Shading Language (GLSL) [297]. Figure 5.11 compares our cel-shading reflectance model to the standard OpenGL Gouraud shading. The discretization of shading intensities is clearly visible in the cel-shaded models, giving the renderings a hand-drawn appearance. The outlines are also noticeably present in the figure, but are more difficult to see due to the distortion from the embedding of a rasterized images into the document.

Combining the techniques of spherical harmonic expansions and cel-shading results in the screen capture from MACROSHAKER's trajectory playback program illustrated in the left pane of Figure 5.12. Through the use of spherical harmonic expansions and cel-shading, we achieve a result similar to the hand drawings of Goodsell.

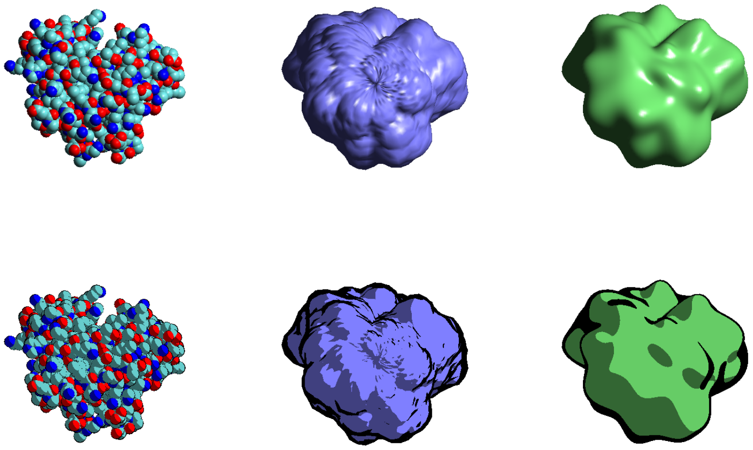


Figure 5.11: Gouraud shading for various representations of Mb (top), and cel-shading of the same representations

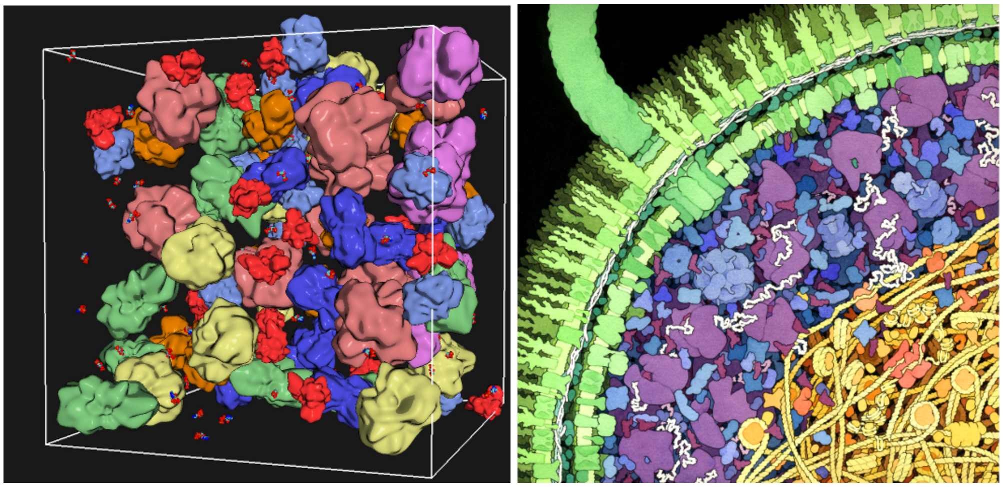


Figure 5.12: A visualization of the enzymes involved in glycolysis using a hybrid Phong and cel-shading technique (left) and an artistic rendition of a cross-section of an *E. coli* cell. [27]

5.6 Illustrative Example: Diffusion of Myoglobin as a Function of Protein Concentration

To illustrate the feasibility of using the MACROSHAKER CGFF described in this work to model a crowded macromolecular environment, we carried out a total of 16 Brownian dynamics simulations on systems of 216 myoglobin (Mb) in an implicit ionic solution of D_2O at 310 K for concentrations ranging from 2 mM to 32 mM. Each simulation used a time step of 10 ps, lasting a total simulation time of 1 μs, after an equilibration phase of 1 μs from random starting coordinates.

A grid spacing of 16 Å was used on coordinates from a PDB file containing the crystal structure of Mb [25] for the creation of the coarse-grained model, resulting in the eight-site model presented in Table 5.2. The Stokes' radius used in the simulation had a value of 22.68 Å, which was derived from a translational self-diffusion constant of Mb in dilute D_2O solution at 310 K of $D_t = 10 \times 10^{-7}$ cm^2/s, which is close to experimental values 9.38-11.3 $\times 10^{-7}$ cm^2/s under similar conditions [28]. A value for D_r of 1.6158×10^{-5} ps^{-1} was used, derived from the Stokes' radius and the solvent viscosity $\eta = 1.001$ mPa·s [298]. The Debye-Hükel screening length κ^{-1} and dielectric constant ϵ_r of Eq. 5.23 were set to 7.952 Å^{-1} and 74.32, respectively. Finally, the c parameter of Eq. 5.21 was set to 119.5 kcal/mol.

The Mb-Mb center of mass radial distribution functions (RDFs) for each of the 16 concentrations are shown in Figure 5.14. From the figure, it is clearly seen that the peak positions became shorter and the peak heights became higher with an increase in Mb concentration. At the very high concentration of 32 mM, for which the volume fraction of the protein Mb from Stokes' spheres was greater than 94% of the total box volume (Table 5.3), the first peak of the RDF is lower, deviating from the trend [5] . This

[5] Experiments have indicated that the Mb volume fraction at 30 mM is around 40% [29,299], which in our case corresponds to a removal of about three solvation layers from the Stokes' sphere.

Bead	x (Å)	y (Å)	z (Å)	σ (Å)	ϵ (kcal/mol)	Charge (e)	Mass (Da)
1	-5.702	-4.998	-3.886	10.784	0.113	-2.0	1224.32
2	-7.506	-8.667	10.371	13.769	0.145	-2.0	2548.01
3	-8.968	7.492	-6.488	15.266	0.161	4.0	3473.01
4	-8.708	4.910	8.553	11.663	0.123	1.0	1548.79
5	7.979	-4.907	-5.277	12.716	0.134	-1.0	2007.32
6	6.252	-8.808	8.361	12.575	0.132	-1.0	1941.05
7	11.052	5.459	-7.550	14.907	0.157	0.0	3233.76
8	6.828	3.862	6.378	10.091	0.106	1.0	1003.11

Table 5.2: An eight-site, coarse-grained model of myoglobin [25] using our grid-based method with a spacing of 16 Å. The Cartesian coordinates of each bead are given relative to the center of mass. The ϵ values for use in Eq. 5.22 were fit to reproduce the concentration-dependent diffusion by experiment from a scaling of σ by 0.0105 kcal/(mol Å).

suggests that model applicability at very high concentrations is of some concern, which is further seen by the small change in interaction energy E_i between concentrations 30 mM and 32 mM compared to the very linear trend for lower concentrations (Table 5.3).

In addition to RDFs, the concentration-dependent diffusion coefficients of Mb were computed for each of the 16 trajectories and compared to values from tracer particle and neutron spin-echo experiments [28, 29]. Figure 5.13 shows a scatter plot of the concentration-dependent diffusion coefficients of Mb obtained from experiments to those obtained from the MACROSHAKER CGFF. It is clearly seen that values from MACROSHAKER agree well with experiment for the concentrations tested, although the computed values are slightly higher than experiments. Wittenberg *et al.* (squares of Figure 5.13) suggested that systematic error that underestimates the self-diffusion coefficients of Mb was present in their experimental data [28].

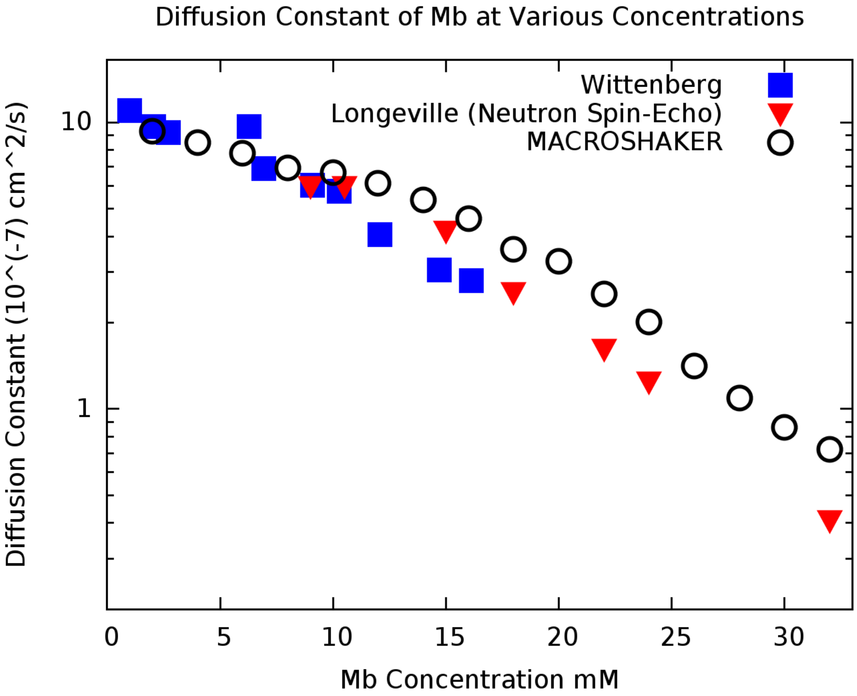


Figure 5.13: Results of MACROSHAKER on a system of 216 myoglobin (Mb) compared to the experimental data of Wittenberg *et al.* [28] and Longeville *et al.* [29]

Concentration (mM)	10^7 D_t (cm^2/s)	E_i (kcal/mol)	Stokes' Volume Fraction
2	9.33	-1.00	0.059
4	8.51	-1.93	0.118
6	7.79	-2.69	0.177
8	6.91	-3.48	0.236
10	6.69	-4.16	0.294
12	6.11	-4.88	0.353
14	5.35	-5.62	0.412
16	4.60	-6.40	0.471
18	3.61	-7.28	0.530
20	3.28	-8.21	0.589
22	2.52	-9.19	0.648
24	2.01	-10.12	0.707
26	1.41	-11.01	0.765
28	1.09	-11.79	0.824
30	0.086	-12.45	0.883
32	0.072	-12.94	0.942

Table 5.3: The concentration-dependent diffusion coefficients of Mb, average Mb interaction energy, and the Mb volume fraction for each concentration. The Mb volume fraction was calculated using $N(4/3)\pi\sigma^3/V$ where $N = 216$ and $\sigma = 22.683$Å and V denotes the simulation box volume.

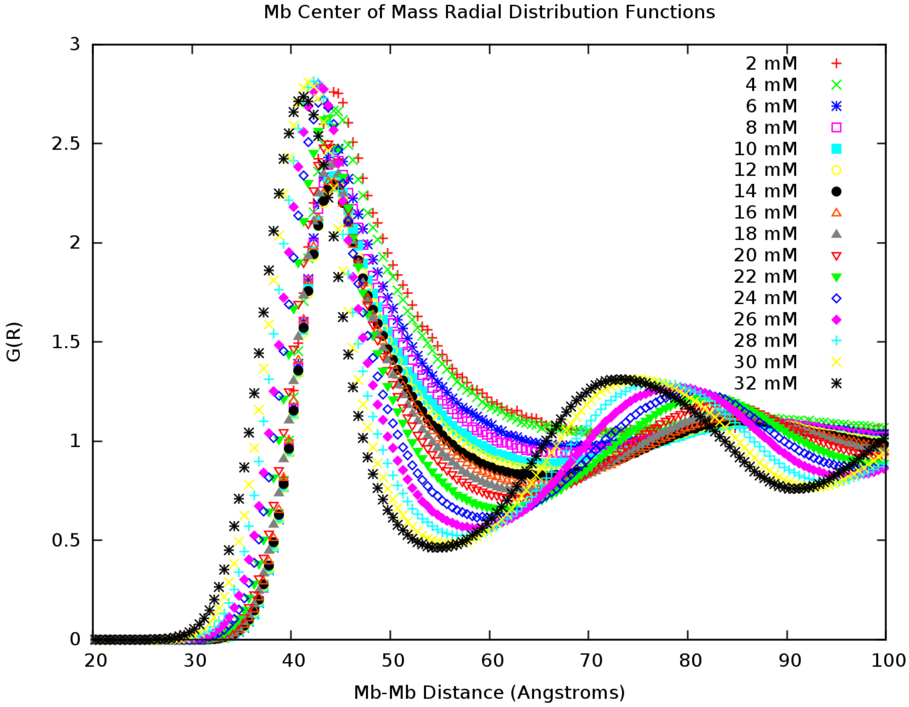


Figure 5.14: Radial distribution functions of the 216 myoglobin (Mb) from simulations of the MACROSHAKER CGFF at concentrations ranging from 2 mM to 32 mM. As the concentration increases, the peak positions become shorter and the peak heights become higher. A notable exception from this trend is the slightly shorter peak height at 32 mM, for which the volume of the Stokes' spheres is approximately 94% of the simulation box.

5.7 Conclusion

We have shown that the Ermak-McCammon scheme of Brownian dynamics provides promising results for effectively modeling the concentration-dependent translational diffusion of myoglobin along with the use of the present coarse-grained force field. In principle, the coarse-graining method used in MACROSHAKER is applicable to other macromolecules. In this chapter, we showed that MACROSHAKER provides a first-step towards a computational model for reaction-diffusion systems of crowded proteins, RNAs, polysaccharides, and other macromolecules *in vivo*, although much remains to be accomplished, especially in the coupling of reactions. In future work, we plan to parameterize MACROSHAKER for many more proteins, including those involved in the reaction-diffusion process of metabolic pathways, such as glycolysis.[6]

[6] We thank David S. Goodsell of the Scripps Research Institute for granting permission to reproduce his artistic rendition of the cross-section of an *E. coli* cell. This work has been partially supported by National Institutes of Health grant GM46736. Computations were performed on an SGI Altix cluster acquired through National Institutes of Health grant S10-RR029467.

Chapter 6

Internal Dynamics of an Analytically Coarse-Grained Protein

6.1 Introduction

Molecular dynamics simulations of biological macromolecules in explicit aqueous solution offer the most detailed information at the atomic level, which is essential for the understanding of dynamics, binding, and activity of these systems. [300] Tremendous progress has been made both in the advance of computer architecture and in the development of computational algorithms, enabling atomistic dynamics simulations to treat systems containing millions of atoms [38] and the dynamics lasting up to milliseconds. [301, 302] However, the spatial and temporal scales needed to address questions relevant to cellular processes such as protein-protein and protein-nucleic acid interactions and macromolecular assembly dwarfs the most sophisticated atomistic approaches available today and perhaps in the distant future. [303] In such a mesoscopic system, it is necessary to use a coarse-grained approach to describe the individual macromolecular components. [304, 305] We present an analytically coarse-grained (ACG) model to represent macromolecular entities such as proteins and nucleic acids as single building blocks that can be used to study macromolecular interactions and assembly in biological cells.

In the past decade, significant efforts have been devoted to the development of coarse-grained models to circumvent the need for describing molecular systems with increasing demands in size and time. [43] Nevertheless, the concepts of atoms and molecules are deeply rooted in our perception of intermolecular interactions; not surprisingly, coarse-grained models typically involve interaction sites in terms of reduced representation of the detailed atomic features of the target system. The early united-atom force field is an example of this type of coarse graining, [306, 307] and recent advances have enabled a much larger number of atoms to be grouped into a united site along with continuum elastic network models. [43] However, to model macromolecular interactions and assembly such as the mechanism of a virus capsid formation, the detailed sequence and the "united-atom" constituents are no longer critical, and it becomes unnecessary to enumerate the specific pairwise interactions between coarse-grained groups among proteins. On the other hand, the use of regular geometrical shapes seems too crude. Indeed, the key structural components are the specific shape and the excluded volume of each capsid protein along with their intrinsic dynamic fluctuations and the accompanying surface electrostatic potential and surface tension. The detailed atomistic interactions, of course, are essential for recognition and binding when two macromolecular particles are in close contact, but these are not the types of details needed for transport processes in the cell. It appears to be desirable to develop a theoretical method that is not restricted to the detailed features of atoms and molecules or coarse-grained interacting groups in a large system; yet, the individual macromolecular species still retain the information of, and are constructed based on the detailed atomistic coordinates determined experimentally, which also provide all physical and biological properties needed to model the dynamic system. Furthermore, the intrinsic fluctuations of the coarse-grained macromolecules relevant to the time scale of the dynamic model for the mesoscopic system need to be taken into account. [43] To limit the scope of discussion, the present chapter is only concerned with the representation of the internal dynamic fluctuations of an ACG macromolecule itself, which are derived from explicit molecular dynamics simulations.

To this end, we make use of the mathematical tools of spherical harmonic analysis to represent the macromolecular particles of interest; spherical harmonic analysis has been extensively applied in a wide range of areas such as geopotential, [308] topography, [309] and physics as well as motion picture and gaming animation. Since analytical harmonic basis functions are used, the method that we design for modeling cellular processes is called the analytical coarse-graining (ACG) of macromolecules. The harmonic representation can also be used directly to describe the physical interactions and to model the dynamics of the system, which will be detailed in later studies. Our strategy provides a single, unifying theory and computational algorithm to study macromolecular systems consisting of thousands of macromolecular particles and entities. In such an approach, each macromolecular unit, such as a protein, is "coarse-grained" as a single moiety of uniform mass density whose excluded volume is encompassed by its solvent-accessible surface that is represented by a set of analytical harmonic basis functions. Furthermore, its physical and biological properties can be treated by exactly the same mathematical procedure as the representation of the macromolecule. Here, we describe the treatment of the intrinsic dynamic fluctuations of a single ACG protein using spherical harmonic functions.

Although the mathematical tools of spherical harmonic computation have been established since the 1780s and modern numerical techniques have greatly enhanced the computational speed, [310, 311] Max and Getzoff, [312] and Olson and co-workers, were among the first to apply spherical harmonic functions to the visualization of molecular surfaces as a graphics rendering tool. [292, 313–316] Recent years have seen the increased usage of this technique to model protein-ligand interactions and protein docking. [317–321] Buchete *et al.* used spherical harmonics to analyze coarse-grained potentials for folding calculations. [8] In a subsequent publication, Duncan and Olson described the possibility of animating the dynamic motion of a protein to render real-time graphics visualization; [322] however, it does not appear that specific investigations have been reported. The approach described by Olson and co-workers was aimed to follow the time-dependent Cartesian coordinates of a protein surface as modeled by normal mode dynamics; the method provides an efficient procedure to generate graphics rendering, but it does not guarantee single-valued properties on the surface, nor is it suitable for modeling protein dynamic motions. [315, 322] In contrast, the method described in this chapter focuses on radial fluctuations of an ACG protein surface that concerns no atomic details, but it is designed to model the most significant dynamic motions revealed from an explicit molecular dynamics trajectory, in which the radial fluctuations are decomposed based on quasiharmonic dynamic analysis. The latter has been extensively explored to characterize large amplitude motions and quasiharmonic vibrational modes of proteins and nucleic acids; [323–327] it provides an adequate analytical procedure to describe the global large amplitude motions of a fully coarse-grained macromolecule with spherical harmonic representation. In addition, for a set of well-chosen numerical quadrature points in the spherical harmonic analysis, our approach provides an efficient procedure for evaluation of molecular properties.

In the following, we first present the analytically coarse-grained (ACG) model and computational details, focusing on the use of spherical harmonic basis functions. Then, we describe a procedure for constructing a mathematical approach to describe the intrinsic quasiharmonic dynamic fluctuations of a protein based on the information from molecular dynamics simulations in explicit solvent. This is followed by an illustrative example to show the feasibility of modeling protein fluctuations without explicit atomic details using the ACG model. Finally, we summarize the key findings of the present study.

6.2 Method

Throughout this chapter, we mainly use the homodimeric enzyme called orotidine 5'-monophosphate decarboxylase (OMPDC) as an illustrative example, which consists of two β-barrels of eight strands of β-sheet and eight α-helices. [328] In this chapter, for convenience of discussion, we focus on the use of a spherical harmonic basis, with which the ACG method is applicable to any macromolecular systems that have star-like topology. [312] We note here that this is not a restriction because any geometrical shapes including non-star-like macromolecules can be represented in the ACG model by augmenting radial functions such as the Zernike function or Slater-type radial functions, [318–321, 329, 330] but we shall not discuss these approaches here. For the rest of this chapter, we interchangeably use the terms of "protein" and "macromolecule", which include proteins, nucleic acids, lipids, and other components of a macromolecular assembly, without specific distinction. In this section, we first outline a qualitative description of the coarse-grained macromolecular model. Then, we provide a brief summary of spherical harmonic representation of the surface of a macromolecular structure. This is followed by the description of incorporating internal dynamic fluctuations of the protein based on the information obtained from principal component analysis (PCA) [324, 325] of the atomistic molecular dynamic trajectory of OMPDC in water. [328] The purpose in this chapter is not aimed at studying the dynamics of proteins using quasiharmonic dynamics; the latter has been thoroughly investigated and its applications and limitations have been characterized. [324–327] The goal here is to illustrate the capability and procedure of incorporating low-frequency, large-amplitude dynamic fluctuations, as revealed by an explicit dynamics simulation, into a fully coarse-grained protein model.

6.2.1 Description of a Macromolecular Particle

We consider a macromolecular structure, which can be a protein or a domain of a protein complex, a segment of nucleic acids, or a protein-nucleic acid complex, as a single entity of uniform mass density. The excluded volume of a given macromolecular structure is defined as the cavity enclosed by the solvent-accessible surface (or the van der Waals surface depending on needs), originating from the detailed three-dimensional atomic structure determined experimentally by X-ray crystallography or NMR, or generated computationally by homology modeling and protein-folding prediction in the absence of experimental data. Note that the solvent-accessible surface encloses a molecular volume which may be significantly greater than that defined by its van der Waals surface, the latter of which is more appropriate for evaluating the macromolecular density. All biochemical functions and physical properties of the macromolecule are fully encoded in the three-dimensional structure, necessary for microscopic and mesoscopic modeling of intermolecular interactions, including electrostatics and hydrophobic surface tension. The characteristic features of a macromolecular structure are considered to have distinguishing features both in size and property from small molecules, peptide fragments, ligands and cofactors, ions and solvent molecules, although the specific criteria depends on a particular application.

From the onset, we do not consider the detailed atomic coordinates or interaction sites; the entire macromolecular unit is a single coarse-grained entity. This is justified as a result of statistical averaging over the spatial and temporal scales to be used to model the dynamic system, which is the cell. However, the size, as defined by the excluded volume occupied by the individual atoms, ligands, and perhaps a small number of buried or surface solvent molecules, and the shape, as represented by the solvent-accessible surface, of a given macromolecule are necessary and critical to a physics-based approach; they are well-defined in our ACG model (*vide infra*) by a single mathematical approach for all types of macromolecular particles of different sizes and shapes, which can be used to systematically provide any desired accuracy and detail of the coarse-grained macromolecule. The definition of a uniform mass density within its excluded volume is akin to the use of a continuum solvent model and a single interior low dielectric constant for a protein in Poisson-Boltzmann calculations and is consistent with our goal of modeling the dynamics of the entire system, which involves the integration of equations of motion at a time step in the order of tens of picoseconds to nanoseconds per iteration. Thus, the representation of the macromolecular species is an average of the system over the time series of the coarse-grained model, [43,331,332] involving the internal atomic fluctuation and spatial orientation when the center of mass of the macromolecule is chosen as a reference point.

Throughout this chapter, the method of Lee and Richards is used to define the macromolecular surface and the excluded volume, [333] although other approaches are available. [334]

6.2.2 Spherical Harmonic Representation of a Macromolecular Particle

The method of using spherical harmonics functions to represent the surface of globular proteins was described by Max and Getzoff [312] and later by Duncan and Olson and by others. [292, 313, 315, 316] Here, we briefly summarize the key elements of this approach and highlight the numerical details employed in our implementation.

Arising from the solution of Laplace's equation in spherical coordinates, the spherical harmonic expansion is the spherical coordinate analog of the widely used Fourier series expansion. Spherical harmonic representations of macromolecules provide not only a smooth and aesthetically pleasing surface, but also the ability for evaluating surface properties such as normal vectors and principal curvature. For any star-like surface, which is single valued in the radial direction of (θ, ϕ) with respect to an origin, there exists a spherical harmonic expansion given as follows:

$$S(\theta, \phi) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} a_{lm} Y_l^m(\theta, \phi) \tag{6.1}$$

where θ denotes the latitudinal or zenith angle $(0 \leq \theta \leq \pi)$, specifies the longitudinal or azimuth angle $(0 \leq \phi < 2\pi)$, $S(\theta, \phi)$ is the radial distance of the surface at angular coordinate (θ, ϕ), a_{lm} are the expansion coefficients and $Y_l^m(\theta, \phi)$ are the real spherical harmonic basis functions, which are orthonormal under the L^2 inner product. We have used the center of mass as the origin in all calculations, and the local axis is generally chosen to coincide with the principal moments of inertia. In general, $S(\theta, \phi)$ can be any scalar physical or chemical property mapped on to the surface of a unit sphere.

The real spherical harmonic basis functions are defined by

$$Y_l^m(\theta,\phi) = \begin{cases} \frac{1}{\sqrt{2\pi}}\overline{P}_l^0(\cos\theta) & m = 0 \\ \frac{1}{\sqrt{\pi}}\overline{P}_l^m(\cos\theta)\cos m\phi & m > 0 \\ \frac{1}{\sqrt{\pi}}\overline{P}_l^{-m}(\cos\theta)\sin m\phi & m < 0 \end{cases} \tag{6.2}$$

where $\overline{P}_l^m(\cos\theta)$ denotes the normalized associated Legendre polynomial of the first kind, of order m and degree l. Methods exist for evaluating the Legendre polynomials of Eq. 6.2, one of which is discussed in Ref. [294], and Ref. [318] lists a technique yielding more numerically stable results.

Making use of the orthonormal property of the real spherical harmonic basis functions under the $\mathcal{L}^2$ inner product, the expansion coefficients in Eq. 6.1 are given by

$$a_{lm} = \int_0^{2\pi}\int_0^{\pi} S(\theta,\phi)Y_l^m(\theta,\phi)\sin\theta d\theta d\phi \tag{6.3}$$

In practice, the infinite sum of Eq. 6.1 is truncated to a relatively small value L, which produces an approximate surface S. Thus, it is only necessary to determine $(L+1)^2$ coefficients and function values for the evaluation of points on the surface:

$$S(\theta,\phi) \approx \sum_{l=0}^{L}\sum_{m=-l}^{l} a_{lm}Y_l^m(\theta,\phi) \tag{6.4}$$

The Cartesian coordinates of vertices used to form a triangulated mesh for graphics display are obtained from the corresponding values in the polar spherical coordinates. This differs from the approach of Duncan and Olson [292,313], who used spherical harmonics expansions to directly approximate the Cartesian coordinates of surface points. Although the direct representation of the surface Cartesian coordinates is convenient for graphics display, it is not suitable for computation of molecular properties of the system, including intermolecular interactions.

6.2.3 Dynamic Motion

Quasiharmonic Dynamics

We use the lowest frequency modes from principal component analysis (PCA) of a molecular dynamics trajectory of a solvated protein to represent its dynamic fluctuations; the dimeric enzyme OMPDC is employed as an illustrative example. The PCA results show the directionality and frequency of protein dynamic motions, in which the lowest frequency modes are typically correlated with protein conformational changes and have been used to interpret conformational variations observed experimentally. [335] Although other approaches such as the continuum elastic network model can be used, [336] for an analytically represented coarse-grained protein without the explicit details of atomic structure, the PCA modes provide the most direct connection to the dynamic motions sampled during an explicit molecular dynamics simulation. The animation of atomic motions following a given normal mode and quasiharmonic dynamics simulation of the complex motions of a macromolecule have been widely used in structure and dynamics analyses at the atomic details. Voth and co-workers described a method to map coarse-grained sites on the basis of PCA modes, [337] and the model has been extended to using the low-frequency normal modes of an elastic network model for the protein. [338] Our method follows a different route of representation than that of Zhang *et al.* ; [337, 338] in ACG, the model is used to represent and animate the low-frequency PCA modes, rather than being derived from PCA. We apply the approach of quasiharmonic dynamics to model the internal fluctuations of coarse-grained macromolecules.

The overall protein fluctuation is obtained by the superposition of individual quasiharmonic modes:

$$\mathbf{R}_j(t) = \mathbf{R}_j(0) + \sum_{k=1}^{K} \mathbf{Q}_{jk}\sigma_k \cos\left(\omega_k t + \lambda_k\right) \tag{6.5}$$

where $\mathbf{R}_j(0)$ and $\mathbf{R}_j(t)$ are the coordinates of atom j at time 0 and t, respectively, and K is the number of quasiharmonic modes used to animate the total dynamic fluctuation of the system. In Eq. 6.5, the parameters associated with mode k, ω_k, $\mathbf{Q}_k$, λ_k, and σ_k are the frequency, mode direction eigenvector, phase, and amplitude, respectively. The phase λ_k is associated with the initial atomic positions, and the thermal average of the second moment of the amplitude distribution is given by $\sigma_k^2 = k_B T/\omega_k^2$, where k_B is Boltzman's constant and T is temperature. The value of K in Eq. 6.5 restricted by the integration time increment, τ, used to propagate the dynamic equations of the coarse-grained system such that $\tau > 2\pi/\omega_K$. Typically, the inclusion of the lowest 10 to 20 modes is more than sufficient to represent the most significant large-amplitude motions.

Here, we use the lowest frequency quasiharmonic motions to represent the internal dynamic fluctuations of analytically coarse-grained macromolecule particles. The limitation of this approach is that it will not produce information for even larger amplitude motions that have not been uncovered in the explicit molecular dynamics simulation. Thus, if the protein undergoes folding and unfolding exchange, it is not appropriate to use the present model; however, it is possible to incorporate into the present treatment conformational transitions for which structures in different conformation substates have been determined experimentally (e.g., by X-ray crystallography or NMR). Nevertheless, our approach is not a simple reproduction of the fluctuation of the molecular dynamics trajectory itself because collision between different coarse-grained macromolecular species can cause random changes in the amplitude and phase of each quasiharmonic mode, resulting in different combinations of modes and trajectories.

Previously, Duncan and Olson proposed a method for shape analysis of protein dynamic surfaces. [322] In that approach, the Cartesian coordinate displacements of surface points corresponding to triangulation vertices were obtained from the static surface and the expansion coefficients for the normal mode eigenvectors projected to these points. Although the method is extremely useful for fast visualization of surface motion, it is not designed to model real-time dynamics. Furthermore, the displacements of triangulation vertices in such a shape analysis algorithm are not suited for property evaluation because the quadrature points and weights will have to be recomputed, which is impractical for real time dynamics simulations. In our approach, the surface deformation is restricted to the direction along the radial vector, consequently preserving the angular coordinates (θ_i, ϕ_j) and the precomputed numerical weights.

Definition of Surface Displacement Vector

We begin with a molecular dynamics trajectory $\mathbf{R}(t_n); n = 0, 1, \cdots, N$ that was saved at time slice t_n, where $\mathbf{R}(t_n)$ is a vector of all atomic coordinates. Principal component analysis of this trajectory yields a set of quasiharmonic vibrations with frequencies $\{\omega_k\}$ and eigenvectors $\{\mathbf{Q}_k\}$. The K lowest frequency modes will be used to model the total dynamic motions that have been sampled by the original molecular dynamics simulation.

For each mode k, we use two distorted configurations generated by following the eigenvector direction stretched to $-2\sigma_k$ and $+2\sigma_k$ from its mean, denoted by $\mathbf{R}_k^{-2\sigma}$ and $\mathbf{R}_k^{+2\sigma}$, to represent approximately the "lower" and "upper" bound of an amplitude, respectively. Let $\{\mathbf{S}_k^{-2\sigma_k}\}$ and $\{\mathbf{S}_k^{+2\sigma}\}$ be the solvent-accessible surfaces (SAS) for the two extreme configurations associated with mode k. Given a set of surface points, $\{u_{ij} = (\theta_i, \phi_j); i = 1, \cdots, M_\theta; j = 1, \cdots, M_\phi\}$, where M_θ and M_ϕ are the number of quadrature points, the radial displacement, due to quasiharmonic vibration of mode k, at the surface point $u_{ij} = (\theta_i, \phi_j)$ is defined as the $2\sigma_k$ variance:

$$q_{ij}^k = \frac{1}{2}\left[S_k^{2\sigma}(\theta_i, \phi_j) - S_k^{-2\sigma}(\theta_i, \phi_j)\right] \tag{6.6}$$

The vector $\mathbf{q}^k$, which can be considered as a property of a unit sphere, represents the approximate amplitude (see below) and direction of surface deformation associated with PCA quasiharmonic mode k:

$$\mathbf{q}^k = \begin{pmatrix} q_{11}^k \\ q_{21}^k \\ \vdots \\ q_{P_\theta P_\phi}^k \end{pmatrix} \tag{6.7}$$

Consequently, the radial displacement vector can also be expressed by a spherical harmonic expansion whose coefficients are determined using the same procedure as for the molecular surface itself (Eq. 6.3):

$$q_{ij}^k = \sum_{l=0}^{L} \sum_{m=-l}^{l} c_{lm}^k Y_l^m(\theta_i, \phi_j) \tag{6.8}$$

Frequency and Phase

Although the rank of the low frequency modes from principal component analysis is very reasonable, the quantitative values of the lowest quasiharmonic vibrational frequencies and the associated time scales are not expected to be accurate to represent the real time dynamic motion. [324–327] Thus, one needs to seek a different way to

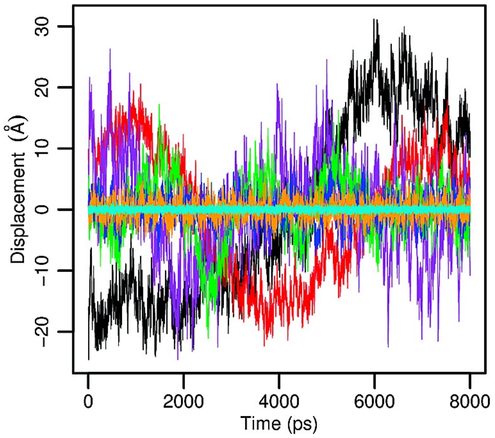


Figure 6.1: Histogram of the computed projection of instantaneous molecular structure of OMPDC in water onto normalized eigenvector directions of the lowest quansiharmonic mode (black), the second (red), the fourth (purple), the tenth (green), the fiftieth (blue), the one hundredth (orange), and the one thousandth (cyan) modes.

obtain the desired oscillatory frequencies. We examined the time dependence of the projections of the instantaneous coordinate vector onto the normalized quasiharmonic eigenvectors, seven of which are shown in Figure 6.1, corresponding to PCA mode numbers 1, 2, 4, 10, 50, 100, and 1000 over a total of 8 ns MD trajectory. Although not unexpected, we are pleased to see the oscillatoroy behavior of each mode, and the amplitudes and frequencies of these oscillations roughly coincide with the order of the PCA modes. For modes above number 50, the fluctuations illustrated in Figure 6.1 can be considered as noise (friction) with respect to the motions of the lowest frequency modes. Importantly, it appears that the PCA mode-projection results can be used to estimate the quantitative frequencies as well as the phase with respect to the structure at time $t_0 = 0$ needed to animate the complex motion of the superimposed fluctuations.

To this end, we used a sinusoidal fitting procedure to optimize the amplitude A_k (not used for mode animation, see below), frequency ω_k, and phase λ_k in Eq. 6.9 for each mode to best reproduce the time-dependent quasiharmonic mode projection data.

$$M_k(t) = A_k \cos\left(\omega_k t + \lambda_k\right) \tag{6.9}$$

In Figure 6.2, we depict the fitted curves against the raw data for modes 1, 2, 4, and 10, whereas the results for the first 20 modes are given as Supporting Information. 6.1 lists the optimized amplitudes, frequencies, and phases for the first 20 modes, the first 10 of which are used to represent the overall protein dynamics fluctuations as an illustrative example in this chapter. An alternative approach is to use the spectral transform of the autocorrelation function of the quasiharmonic mode fluctuations.

Time Evolution of the Dynamic Fluctuation

The SAS surface $\mathbf{S}(t_0)$ corresponding to the structure $\mathbf{R}(t_0)$ at time $t_0 = 0$ in the dynamic trajectory is chosen as the starting configuration and is expressed in terms of spherical harmonics basis as follows:

$$S_{ij}(0) \equiv S(t=0, \theta_i, \phi_j) = \sum_{l=0}^{L} \sum_{m=-l}^{l} a_{lm}^{o} Y_l^m(\theta_i, \phi_j) \tag{6.10}$$

where $S_{ij}(0)$ is the radial distance at an angular coordinate $uij = (\theta_i, \phi_j)$, and the coefficients $\{a_{lm}^{o}\}$ are determined according to Eq. 6.3. We assume that the dynamic modulation of the protein surface associated with mode k also has the same frequency. Thus, the atomic coordinates in Eq. 6.5 are replaced by protein surface points, and we

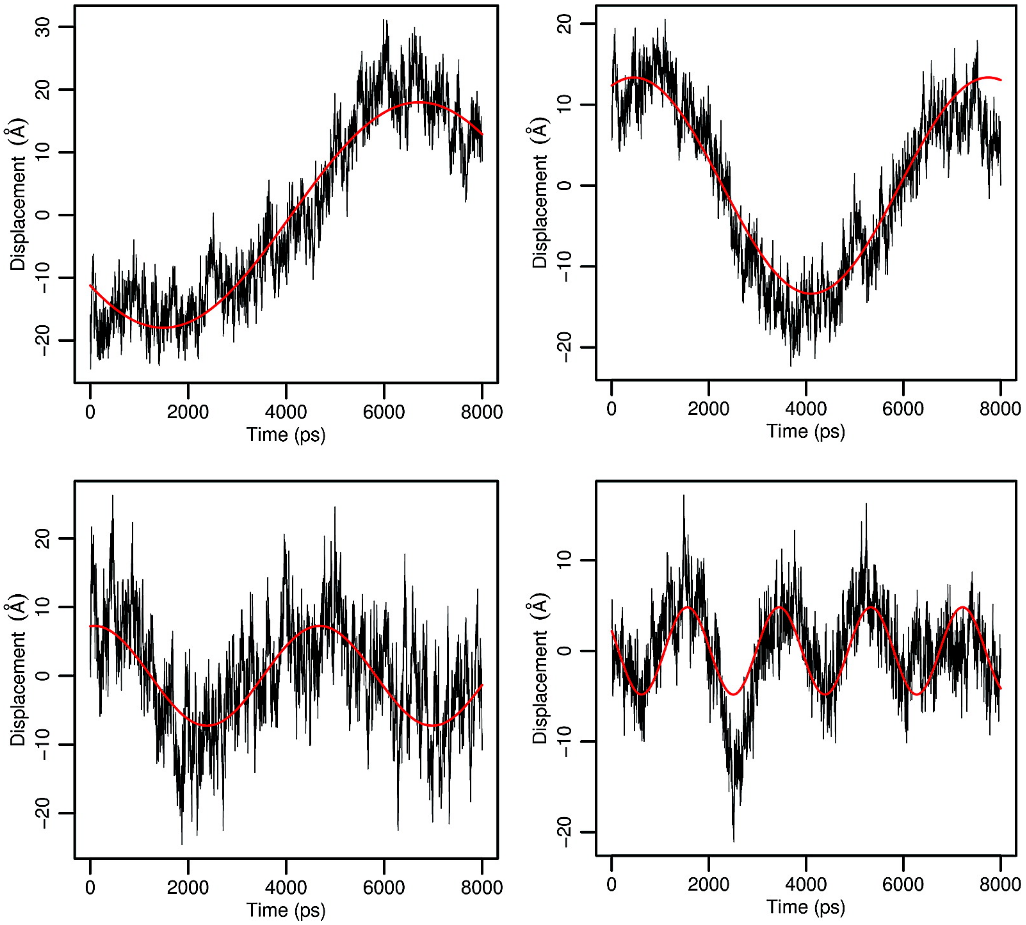


Figure 6.2: Sinusoidal fit of harmonic frequencies and phases (with respect to the structure used at the start of the molecular dynamic simulation of OMPDC) to the oscillatory structural projections illustrated in Figure 6.1 for modes number 1, 2, 4, and 10.

Mode	A_k	ω_k	λ_k
1	18.0	0.60	2.25
2	13.4	0.86	5.89
3	9.8	1.55	1.05
4	7.3	1.37	6.17
5	3.0	2.71	2.50
6	1.3	7.37	4.02
7	6.5	2.12	2.97
8	5.0	2.72	1.62
9	3.9	2.66	3.32
10	4.8	3.33	1.10
11	3.2	3.37	6.00
12	3.3	2.65	1.22
13	0.4	3.79	0.69
14	2.7	4.10	5.47
15	2.6	4.17	2.66
16	0.9	5.01	6.21
17	2.1	4.72	1.33
18	2.5	4.22	5.78
19	1.7	7.59	1.71
20	1.2	7.59	1.60

Table 6.1: Optimized amplitudes (Å), frequencies (rad/ns), and phase (rad) for the time-dependent quasiharmonic mode projection along the molecular dynamics trajectory of the protein orotidine monophosphate decarboxylase as represented by Eq. 6.16.

write the total surface radial displacement at point $u_{ij} = (\theta_i, \phi_j)$ as follows:

$$\begin{aligned} S_{ij}(t) &= S_{ij}(0) + \sum_{k=1}^{K} W_k q_{ij}^k \cos\left(\omega_k t + \lambda_k\right) \\ &= \sum_{l=0}^{L} \sum_{m=-l}^{l} \left[a_{lm}^o + \sum_{k=1}^{K} W_k c_{lm}^k \cos\left(\omega_k t + \lambda_k\right) \right] Y_l^m(\theta_i, \phi_j) \end{aligned} \tag{6.11}$$

where W_k is a mode weighting factor to be determined by least-squares fit to the instantaneous surfaces $\hat{S}(t_n)$ of the structures sampled by the explicit molecular dynamics simulations in the entire trajectory, $\{\mathbf{R}(t_n) \to \hat{S}(t_n); n = 0, 1, \cdots, N\}$. Equation 6.11 preserves the angular coordinates, consequently all precomputed values of the spherical harmonic functions and quadrature weights needed for property evaluations (as well as for real-time graphics animation). [339]

The mode weighting factors in Eq. 6.11 are determined by minimizing the following error function:

$$\epsilon = \frac{1}{N} \sum_{n=1}^{N} \sum_{ij} \left[S_{ij}(t_n) - \hat{S}_{ij}(t_n) \right]^2 \tag{6.12}$$

It is straightforward to show that the minimization yields a linear equation that can be conveniently solved.

$$\sum_{k=1}^{K} B_{qk} W_k = D_q; \quad q = 1, \cdots, K \tag{6.13}$$

where the matrix elements are defined as follows:

$$D_q = \frac{1}{N} \sum_{n=1}^{N} \sum_{ij} \left[\hat{S}_{ij}(t_n) - \hat{S}_{ij}(0) \right] U_{ij}^q(t_n) \tag{6.14}$$

$$B_{qk} = \frac{1}{N} \sum_{n=1}^{N} \sum_{ij} U_{ij}^{q}(t_n) U_{ij}^{k}(t_n) \tag{6.15}$$

$$U_{ij}^{q}(t_n) = \sum_{l=0}^{L} \sum_{m=-l}^{l} c_{lm}^{q} \cos{(\omega_q t_n + \lambda_q)} Y_l^m(\theta_i, \phi_j) \tag{6.16}$$

6.3 Numerical Considerations

The spherical harmonic expansion coefficients are determined by sampling surface values (coordinates) to approximate the integral of Eq. 6.3. A number of methods for optimizing surface point distribution are available including the use of a geodesic unit sphere. [340] In the present application, realizing that the integral in θ is formally a Fourier transform, the numerical integration can be evaluated using M equispaced points to take advantage of fast Fourier transform (FFT) at a computing scaling of $\mathcal{O}(M \log{(M)})$. There are two ways of selecting points in ϕ; the first is to use Gauss-Legendre quadrature nodes and weights, which needs only $M/2$ points, whereas a set of equally spaced points can be selected, which is equivalent to Chebychev nodes in $\cos\phi$. [339,341] The latter is convenient to use but less efficient computationally and requires a total of M points for the same accuracy. The numerical scaling for integrating in is $\mathcal{O}(M^3)$.

If the sampling points used in the evaluation of the integral in Eq. 6.3 are chosen to coincide with the numerical quadrature values, $\{(\theta_p, \phi_q); p = 1, \cdots, M_\theta; q = 1, \cdots, M_\phi\}$, the numerical procedure for property calculation can be greatly simplified. In all cases, the associated Legendre function values are precomputed along with the measure, $\sin\theta_i$, and quadrature weights (see Appendix B) for a given structure and stored. The use of $(\sin\theta_i)^{1/2}\overline{P}_l^m(\theta_i, \phi_j)$ preconditioning in property calculations can greatly increase numerical stability by keeping the product roughly constant. [341]

All computations and illustrations are performed using a software package written in our laboratory.

6.4 Discussion

Figure 6.3 illustrates the spherical harmonic rendering of the trimeric structure of the capsid protein (2FZ2) of turnip yellow mosaic virus at various degrees of representation up to $L = 30$. By simple inspection, the domed triangular shape is not directly associated with a unit sphere, but the 3-fold symmetry of the complex is already represented at $L = 3$, and the domed structural feature is clearly visible with $L = 4$ and 5. As the degree of spherical harmonic basis increases, the molecular shape and detail is well described with an L above 10, while greater local features can be found using higher degrees. In principle, the spherical harmonic representation can yield any desired accuracy by increasing the value of L. However, it should be kept in mind that the protein or macromolecular structure that we model is a coarse-grained representation of a distribution over the time scale of the integration step used in Brownian dynamics simulations. Thus, there is no reason to use a very high degree of L to generate an "accurate" surface that is in fact beyond the variance of the surface amplitude fluctuation over the time interval in Brownian dynamics simulations. In fact, a certain degree of fuzziness is especially desired for these computations, a subject to be addressed in the future. We have found that a value of $L = 10 - 15$ is adequate to provide a compromise of quantitative shape and volume description and sufficient distinguishing details of different proteins. At this level of representation, a total of 121 to 256 terms is needed in the spherical harmonic expansion in Eq. 6.1.

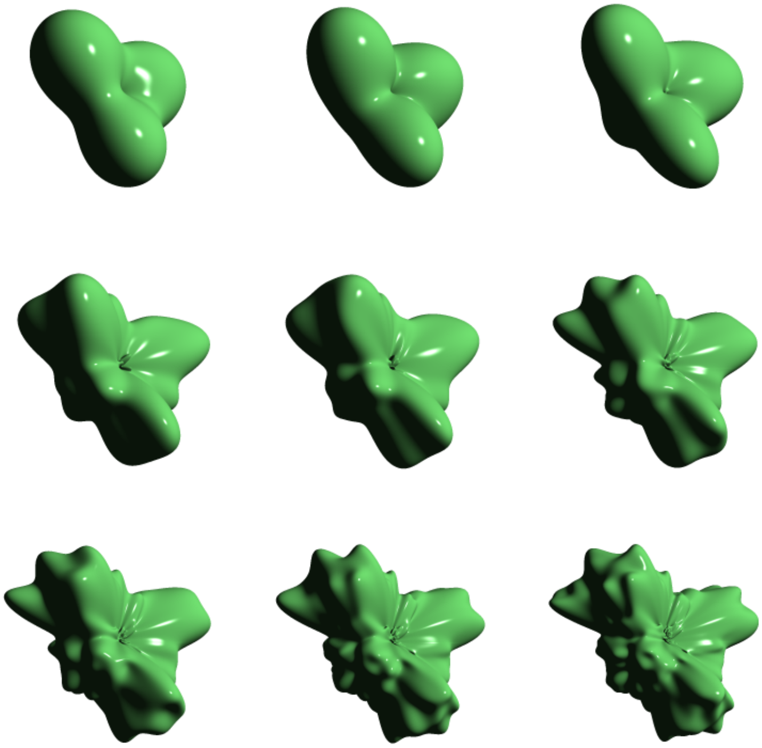


Figure 6.3: Spherical harmonic reconstruction of the Lee and Richards surface for the trimer complex of the capsid protein of turnip yellow mosaic virus using representation degrees of L = 3, 4, 5, 10, 12, 15, 20, 25, and 30 numbered from top left to bottom right.

To animate the dynamic fluctuation of spherical harmonics coarse-grained OM-PDC, we have determined the surface radial displacement amplitudes of the ten lowest quasiharmonic modes of vibration from principal component analysis, which contribute to the overall fluctuation throughout the 8 ns molecular dynamics simulation. The optimized mode weighting factors in Eq. 6.11 for the first ten modes are listed in Table 6.2. The weighting factors are about 0.5 for these low frequency modes, which is a reflection that the approximate "amplitudes" used to define the surface radial displacement vectors by stretching the quasiharmonic deformation to $\pm 2\sigma$ limits is used so as to obtain a large contrast in the analysis. The minimization procedure reduces the initial large variations to about $\pm 1\sigma$, further suggesting that the procedure employed in the present study is a reasonable approximation to represent the overall protein fluctuation. However, as the frequency (mode number) increases, one standard deviation is not a good measure of the dynamic contributions due to stochastic collision and coupling with fast motions. The small weighting factors for modes 5 and 6 indicate that the associated fluctuations from the principal component analysis may not be well described by quasiharmonic vibrational motions (6.4), perhaps due to conformational jumps, or a longer equilibration that is needed in the original MD simulation, or the fact that the explicit molecular dynamics simulation is rather short.

Using the frequencies, phases, and amplitudes optimized using the procedure outlined in section 2 by means of principal component analysis of a molecular dynamics trajectory to train the large amplitude dynamic behavior of the ACG model for OM-PDC, we carried out quasiharmonic dynamics animation of the compounded motion of the ten lowest frequency quasiharmonic modes for 0.25 μs at an integration increment of 25 ps per step, which took about 1 min on a desktop workstation. 6.5 shows three structures from the trajectory using the initial conditions listed in Tables 6.1 and 6.2. Although it is difficult to distinguish the relatively small surface variations in the

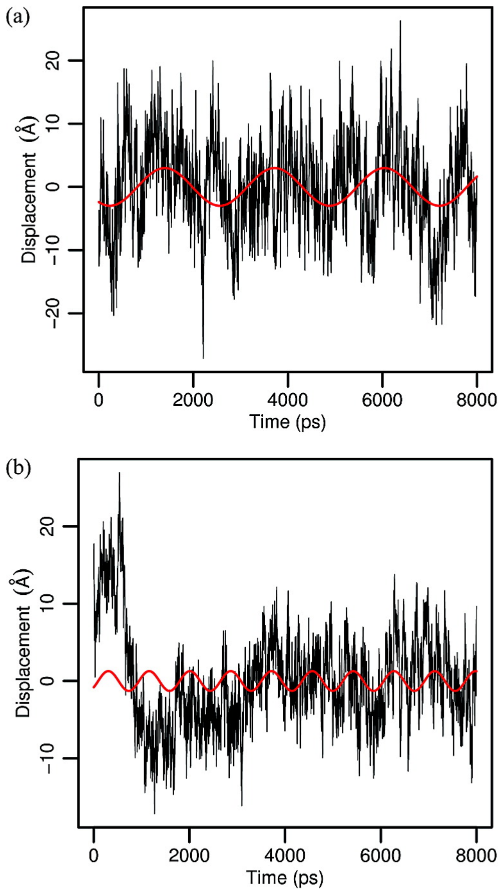


Figure 6.4: Sinusoidal fit of harmonic frequencies and phase (with respect to the structure used at the start of the molecular dynamic simulation of OMPDC) to the oscillatory structural projections illustrated in Figure 6.1 for modes number 5 (a) and 6 (b). Note that if the trajectory of the first 1 ns is discarded in mode 6 evaluation, the frequency and amplitude are both reasonable, suggesting there is either a conformational jump or change in the first 1 ns.

Mode	W_k
1	0.53892
2	0.48497
3	0.48692
4	0.42098
5	0.09645
6	0.05428
7	0.42884
8	0.29127
9	0.33406
10	0.37536

Table 6.2: Computed mode weighting factors to represent the overall complex protein fluctuations using the first ten lowest quasiharmonic modes

static pictures, a movie that is given as Supporting Information (SMovie 1) does provide a more vivid depiction of the dynamic fluctuations of the trajectory. 6.6 shows the computed volume histogram of the ACG protein. Not surprisingly, the primary periodicity is dictated by the lowest frequency mode, which has the largest amplitude contributions to the complex motion (Tables 6.1 and 6.2), and the spectral transform of 6.6 shows frequencies that coincide with the input listed in Table 6.1.

Note that although the surface radial displacement vectors were obtained by considering the corresponding quasiharmonic modes of atomic vibrations, the radial vectors do not possess an orthogonality relationship, and the least-squares fitting procedure used to optimize the displacement amplitudes also introduces contributions from other modes not specifically characterized purely by each quasiharmonic mode. Further, the amplitude for each quasiharmonic mode represents an average fluctuation sampled in the original molecular dynamics simulation; however, the maximum fluctuations can be significantly greater than the individual averages due to stochastic collisions with solvent molecules as well as mode coupling. Consequently, stochastic effects may be included in mode synthesis by randomly increasing and decreasing the

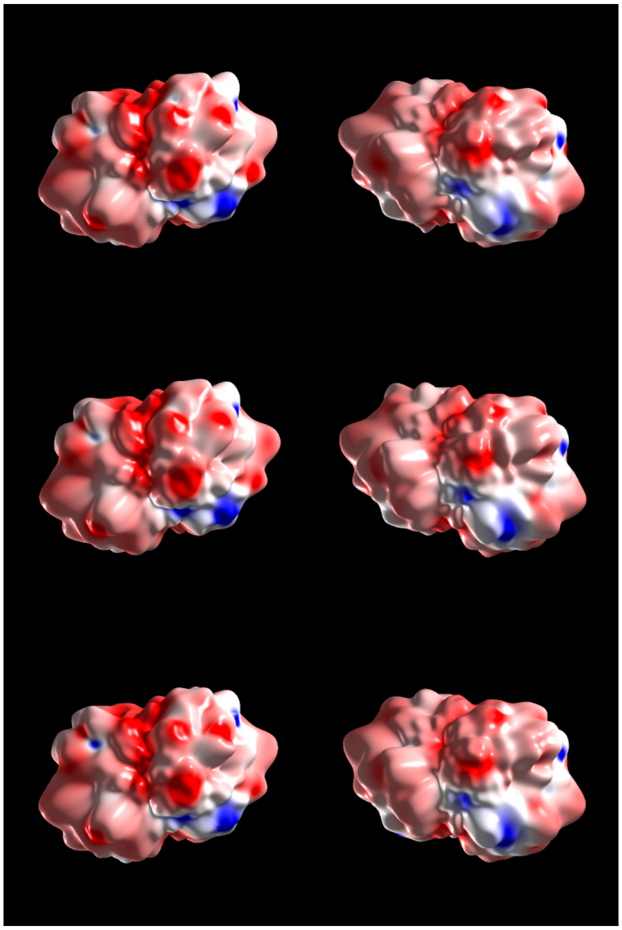


Figure 6.5: Snapshots of three structures of the analytically coarse-grained (ACG) protein OMPDC using spherical harmonic basis at a representation degree of $L = 15$ from the 0.25 μs composite fluctuation trajectory using the amplitudes, frequencies, and phases listed in Tables 6.1 and 6.2. The three structures on the right-hand side are the same as the corresponding ones on the left, rotated by $180°$. The ACG protein surfaces are colored by the surface charge density for the illustration with red representing negative and blue positive charge densities, respectively.

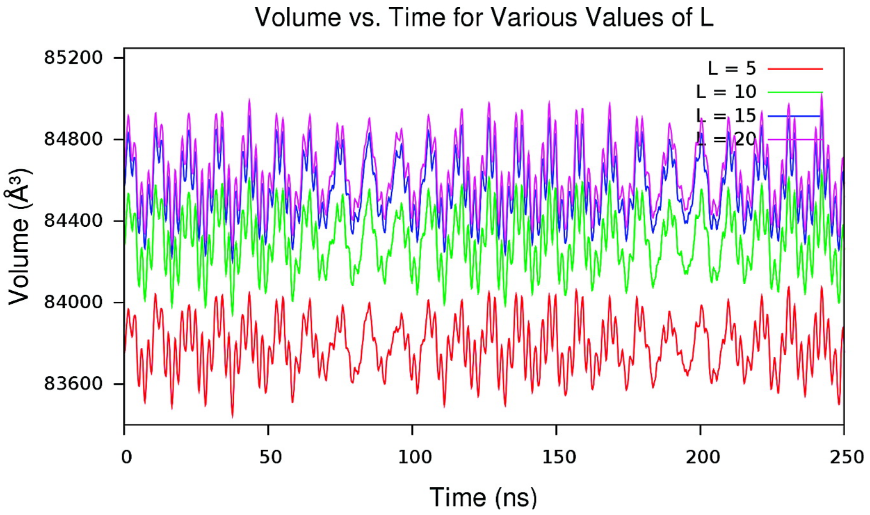


Figure 6.6: Histogram of the fluctuation of the excluded volume of the ACG OMPDC protein at various resolutions in L, ranging from 5 to 20 along the internal quasiharmonic fluctuation trajectory. The excluded volume illustrated in this figure is defined as the cavity enclosed by the Lee-Richards surface, which is about one solvent layer larger than the van der Waals surface.

amplitudes that yield the correct means over a long trajectory and satisfy the condition of the second dissipation theorem.

The fluctuation of the excluded volume defined by the protein surface for OMPDC, which can be conveniently determined by

$$V = \int_0^{2\pi} \int_0^{\pi} \int_0^{S(\theta,\phi)} r^2 \sin\theta d\theta d\phi dr = \frac{1}{3} \int_0^{2\pi} \int_0^{\pi} S^3(\theta,\phi) \sin\theta d\theta d\phi \approx \frac{1}{3} \sum_{ij} S^3(\theta_i,\phi_j) \sin\theta_i w_i w_j \tag{6.17}$$

is shown in Table 6.6 at different degrees of approximation from $L = 5$ to $L = 20$. In Eq. 6.17, w_i and w_j are the quadrature weights and the values $\{S(\theta_i,\phi_j)\}$ are already determined during the dynamics animation. There is no major difference for the results obtained using $L = 15$ and $L = 20$, suggesting that the use of a spherical harmonic representation of degree 15 is sufficiently accurate to model the molecular volume. At $L = 10$, the average volume is about 84.3 ± 0.2 nm^3, which is less than 0.4% smaller than an average value of about 84.6 ± 0.2 nm^3 at higher orders. The lower order at $L = 5$ introduces an error of about 1% in volume. Note that the excluded volume determined by solvent-accessible surface is significantly larger than that encompassed by the van der Waals surface of the macromolecule. Using the initial structure in the molecular dynamics simulation, the ratio between the volumes defined by the solvent-accessible surface and the Bondi van der Waals surface (scaled by 1.20 as proposed by Luque and Orozco [342] in the calculation of solvation free energies treating the solvent as a polarizable dielectric continuum) is 1.46. If this factor is taken into account, the average molecular density of dry (without solvent molecules) OMPDC is estimated to be 1.315 ± 0.002 g/cm^3, in excellent agreement with the typical protein density $(1.35 - 1.40$ g/cm$^3)$ estimated experimentally. [343, 344] If the extra volume enclosed by the solvent-accessible surface is filled with water molecules (about 850 water molecules) at the bulk density, the average density of cavity included in the spherical harmonics coarse-grained protein is estimated to be about 1.20 g/cm^3, which may be considered as the macromolecular structure solvated by one solvent shell.

6.5 Conclusion

An analytical coarse-graining (ACG) model has been introduced to represent biological macromolecules, making use of a spherical harmonic basis in the present study. In our approach, a macromolecular structure is treated as a fully coarse-grained entity with a uniform mass density without the explicit description of atomic details or "coarse-grained" interaction sites. The use of a uniform density of the ACG macromolecule is justified because the model represents an ensemble average relevant to the time series used in the dynamics simulation of cellular processes. However, the excluded volume and specific shape of the ACG macromolecule species are critical, which are explicitly treated by a spherical harmonic representation. In principle, spherical harmonic analysis can provide any desired accuracy and detail of the macromolecular surface. The present chapter focuses on the first issue in a fully coarse-grained protein model, that is, the description of the internal fluctuation of the ACG macromolecule. Here, we make use of the dimeric enzyme OMPDC, consisting of 416 amino acids and 2 substrate molecules in the active site, as an illustrative example.

The internal fluctuation of the ACG protein is modeled by the superposition of a selected number of lowest frequency quasiharmonic modes of vibration, which are derived from an explicit molecular dynamics simulation of the fully solvated protein in water. A procedure for estimating the amplitudes, time scales (frequencies) of the quasiharmonic motions, and the corresponding phase is presented and used to synthesize the complex motion (note that the eigenvalues of the lowest quasiharmonic modes are close to zero and they are not quantitative for description of the time scales of the corresponding motions). In principle, all modes up to a frequency, limited by the time interval of the coarse-grained dynamics, can be included, but as numerous studies have shown, when employing principal component analysis and quasiharmonic essential dynamics, only a fraction of the lowest frequency modes are important in such a representation. The analytical description and numerical algorithm presented here can in principle provide a representation of the internal protein fluctuations as closely as needed in comparison with the atomistic molecular dynamics simulation; however, the internal motion is restricted by the short-time nature of molecular dynamic trajectories, and the present method is not designed for the description of unfolding events unless such transitions occur during the molecular dynamics simulation.[1]

[1] This work has been partially supported by National Institutes of Health grant GM46736.

6.6 Supporting Information

Projected structural fluctuations on to the first 20 lowest frequency modes and a movie animating the trajectory of the dynamic fluctuation of the ACG protein OMPDC represented by using spherical harmonic basis functions. This material is available free of charge via the Internet at `http://dx.doi.org/10.1021/ct100426m`.

Chapter 7

Conclusion & Discussion

7.1 Conclusion

The description of the potential energy surface, or force field, is of paramount importance to the accurate modeling of chemical systems *in silico*. The PMO/X-Pol/DPPC quantum mechanical force field (QMFF) and MACROSHAKER coarse-grained force field (CGFF) represent new paradigms in computational chemistry beyond the *de facto* molecular mechanics force field (MMFF).

In the first part of this dissertation, we showed that QMFFs based upon the explicit polarization theory (X-Pol) can provide an accurate description of polar liquids. Chapter 2 presented the Hartree-Fock and the X-Pol theories with their analytical first derivatives, as well as the class IV DPPC charge model. Chapters 3 and 4 showed that the addition of p-orbitals onto the MNDO Hamiltonian and the use of genetic and stochastic optimization algorithms for semiempirical parameterization were sufficient for accurately modeling the behavior of small clusters of water and hydrogen fluoride compared to results from experiments and *ab initio* calculations. In addition, through the use of DPPC charges and a 12-6 Lennard-Jones potential for modeling exchange-correlation and dispersion interactions, we found that X-Pol could accurately model energetic, thermodynamic, and dynamical properties of liquid water and liquid hydrogen fluoride compared to experiments and other polarizable models.

The second part of this dissertation examined coarse-grained models for simulating crowded systems of many proteins. We showed in Chapter 5 that the Brownian dynamics of Ermak and McCammon [46] can be used with rigid, coarse-grained models to accurately reproduce the concentration-dependent diffusion coefficients of myoglobin at volume fractions as large as 40%. Although the approach we used involving beads was successful, Chapter 6 introduced the notion of representing coarse-grained proteins as atomless, analytical functions, whose internal dynamics are described by quasiharmonic fluctuations, for simulations in the cellular environment. In principle, these analytical functions could be used with formalism similar to that found in electronic structure methods to produce a CGFF for proteins.

We believe that the force fields introduced in this dissertation will be of great use as they are further refined and parameterized. What follows are some of our ideas for future work.

7.2 Future Work on QMFFs

7.2.1 Variational many-body expansion

The many-body expansion for QM calculations proposed by Stoll and Preuß [345] was recently extended to the variational X-Pol theory under the name *variational many-body expansion* (VMB) [148]. The VMB corrects the X-Pol energy E_1 through summing all energy differences between E_1 and the energy computed through permutations of the repartitioning of fragments into a single dimer, trimer, or higher-order fragment with all other monomer fragments from the X-Pol calculation. The benefit of this approach

	Full PMOw	X-Pol	VMB2	VMB3
Prism	46.9	24.3	45.9	46.8
Cage	46.6	25.0	45.7	46.5
Book	44.9	26.7	43.6	44.8
Cyclic	41.8	28.6	40.0	41.6
MUD	0.0	18.9	1.2	0.1
RMSD	0.0	19.2	1.2	0.1

Table 7.1: Binding energy in kcal/mol of various water hexamers for PMOw using full QM, X-Pol, VMB2, and VMB3. Mean unsigned deviations (MUD) and root-mean square deviations (RMSD) compared to the full PMOw result show that both VMB2 and VMB3 agree significantly better with the full energy than the X-Pol method alone. All calculations were done without Lennard-Jones terms and used Mulliken charges.

is that charge-transfer effects within the higher-order fragments are fully described by the level of QM theory employed, but at an increased cost. The total energy of the VMB system is written as

$$E_{\text{TOT}} = E_1 + \Delta E_2 + \Delta E_3 + \cdots + \Delta E_N \tag{7.1}$$

where

$$\Delta E_2 = \sum_{I<J}^{N} \Delta E_{IJ} = \sum_{I<J}^{N} (E_{IJ} - E_1), \tag{7.2}$$

and

$$\Delta E_3 = \sum_{I<J<K}^{N} \Delta E_{IJK} = \sum_{I<J<K}^{N} (E_{IJK} - E_1 - \Delta E_{IJ} - \Delta E_{IK} - \Delta E_{JK}) \tag{7.3}$$

with analogous definitions of higher-order terms. Eq. 7.1 is often truncated at the two-body level (VMB2) or three-body level (VMB3) to reduce cost.

Table 7.1 gives the binding energies of several water hexamers for the full PMOw method, X-Pol, VMB2, and VMB3 all without Lennard-Jones terms and using Mulliken partial charges. These preliminary results show that VMB2 and VMB3 are in significantly better agreement with the full PMOw result than the single-body version of X-Pol.

Since E_{TOT} is a sum of terms (Eq. 7.1), each with an analytical gradient, E_{TOT} also has a straightforward analytical gradient and may be used for geometry optimization or MD simulations. A parallel version of the VMB2 method with analytical gradient has been implemented into our modified version of NAMD [51], which we hope to use for studying liquid hydrogen fluoride.

7.2.2 Unpolarizable core in PMO

It was found by Stewart during the development of the PM6 method [131] that in rare instances nearly "nuclear-fused" geometries were the optimal molecular structures, due to what Stewart called the "complete neglect of the unpolarizable core" of the atoms in that method. During our parameterization of fluorine for PMOw, we observed similar fusion-like behavior, which, through decomposition of gradient terms, we found to be originating from the inclusion of the D1 dispersion correction of Grimme [95] (also used by PM6) on hydrogen atoms. This same behavior was also observed for a gas-phase water molecule when the HOH bond angle became very small during a vapor-phase Monte Carlo simulation using PMOw.

In the case of PM6, Stewart proposed the addition of a simple function to the core-core term of atoms designed to be vanishingly small at normal bonding distances, but very large at shorter, unphysical distances. This term takes the form of Eq. 7.4

$$f_{AB} = c \left(\frac{\left(Z_A^{1/3} + Z_B^{1/3} \right)}{R_{AB}} \right)^{12} \tag{7.4}$$

where Z_A and Z_B denote the core charges of atoms A and B and $c = 10^{-8}$ [131].

A correction of this type to the PMO method will be absolutely essential for VMB2 MD simulations, due to the full QM treatment of dimer pairs and the likely possibility that integration error could cause two hydrogen atoms to have a close approach.

7.2.3 Additional parameters for PMO

The success of the PMO-based XP3P and XPHF models provides encouraging progress towards a general framework for polarizable force fields. At the time of this writing, PMO has only been parameterized for elements H and O in its PMOv1 form [52], H, O, and F in its PMOw form [237], and H, C, and O in its PMO2 form [94].

Similar to oxygen and fluorine, nitrogen is a highly-electronegative atom that heavily participates in hydrogen bonding. Recognizing this, optimization of a nitrogen parameter for PMOw using the existing formalism is currently underway. In addition to being a stepping stone for protein simulations which at minimum requires parameters for H, C, N, and O, we plan to introduce an X-Pol model for liquid ammonia called XPNH$_3$. Beyond nitrogen, chlorine, bromine, and iodine parameterizations for PMOw are under investigation.

7.3 Future work on CGFFs

7.3.1 Macromolecular Assembly

Our goal in building MACROSHAKER is not only to model the diffusion process of macromolecules by a CGFF, but also to model the assembly process of large complexes such as the dynamics of macromolecules in a biological cell, and the capsid self-assembly of viruses. Progress has been made on the implementation of the coarse-grained virus assembly model of Wales [30] (see Figure 7.1), which was later extended

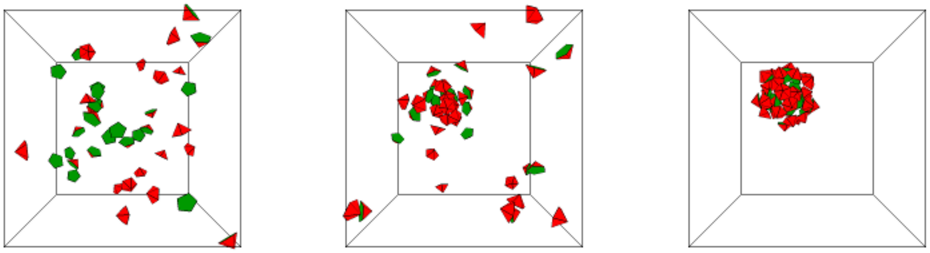


Figure 7.1: Screenshots from a program showing the virus assembly model of Wales. [30]

to a model involving spherical crowders [346].

7.3.2 Metabolic Processes

An immediate target application of the coarse-grained models in this dissertation is to study the direct and secondary (feed-back) control and regulation of ATP production in glycolysis under realistic conditions in a cell. To this end, we plan on using the glycolysis process taking place in the red blood cell as a model, since it represents one of the simplest cases that we can test the influence of crowding effects on the modeling of a sequence of enzymatic processes. Furthermore, the glycolytic process of the red blood cell is well understood, and all rate constants and properties of the ten enzymes are known. The dominant (more than 95% of dry mass [347]) component of proteins in a red blood cell is hemoglobin. Thus, it is possible to construct a model with ten copies of each glycolytic enzyme immersed in a bath of hemoglobin at a relative ratio of 1:10, 1:15, and 1:20 to examine crowding effects. A key question is crowding effects on the effective concentration (activity) of enzymes, which has been a focal point in the study of crowding effects.

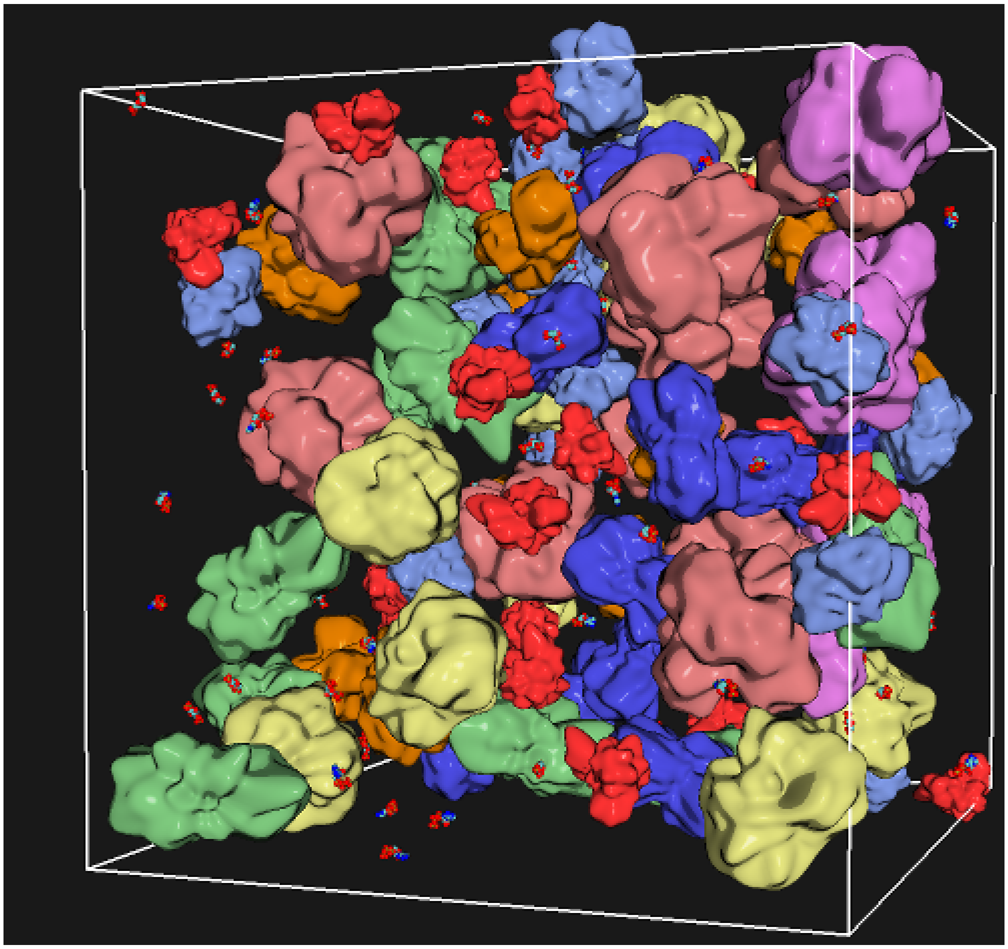


Figure 7.2: A visualization of the 10 enzymes involved in glycolysis as rendered by MACROSHAKER.

7.3.3 Interactive Visualization

With a slight modification to the visualization component of MACROSHAKER, we can employ the use of shutter-glasses to view a crowded cellular environment in 3D. Motion tracking technology could be coupled into the visualization code so that the cellular environment could be explored through walking around within a room and the turning of one's head. Additionally, force-feedback devices could be used to drag and/or rotate individual proteins during the 3D walk-through experience.

References

[1] *CRC Handbook of Chemistry and Physics, 94th Edition*. CRC Press, 2013.

[2] P. Ren and J.W. Ponder. Polarizable Atomic Multipole Water Model for Molecular Mechanics Simulation. *The Journal of Physical Chemistry B*, 107(24):5933–5947, 2003.

[3] V.S. Bryantsev, M.S. Diallo, A.C.T. van Duin, and W.A. Goddard. Evaluation of B3LYP, X3LYP, and M06-Class Density Functionals for Predicting the Binding Energies of Neutral, Protonated, and Deprotonated Water Clusters. *Journal of Chemical Theory and Computation*, 5(4):1016–1026, 2009.

[4] J.K. Gregory, D.C. Clary, K. Liu, M.G. Brown, and R.J. Saykally. The Water Dipole Moment in Water Clusters. *Science*, 275(5301):814–817, 1997.

[5] K. Szalewicz, K. Patkowski, and B. Jeziorski. Intermolecular Interactions via Perturbation Theory: From Diatoms to Biomolecules. In D.J. Wales, editor, *Intermolecular Forces and Clusters II*, volume 116 of *Structure and Bonding*, pages 43–117. Springer Berlin Heidelberg, 2005.

[6] M.W. Mahoney and W.L. Jorgensen. A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions. *The Journal of Chemical Physics*, 112(20):8910–8922, 2000.

[7] W.L. Jorgensen and C. Jenson. Temperature dependence of TIP3P, SPC, and TIP4P water from NPT Monte Carlo simulations: Seeking temperatures of maximum density. *Journal of Computational Chemistry*, 19(10):1179–1186, 1998.

[8] N.-V. Buchete, J.E. Straub, and D. Thirumalai. Continuous anisotropic representation of coarse-grained potentials for proteins by spherical harmonics synthesis. *Journal of Molecular Graphics and Modelling*, 22(5):441 – 450, 2004.

[9] K.P. Huber and G. Herzberg. *Molecular Spectra and Molecular Structure*. Van Nostrand Reinhold (NY), 1979.

[10] J.S. Muenter and W. Klemperer. Hyperfine Structure Constants of HF and DF. *The Journal of Chemical Physics*, 52(12):6033–6037, 1970.

[11] G.S. Tschumper, Y. Yamaguchi, and H.F. Schaefer III. A high level theoretical investigation of the cyclic hydrogen fluoride trimer. *The Journal of Chemical Physics*, 106(23):9627–9633, 1997.

[12] A.S. Pine and B.J. Howard. Hydrogen bond energies of the HF and HCl dimers from absolute infrared intensities. *The Journal of Chemical Physics*, 84(2):590–596, 1986.

[13] B.J. Howard, T.R. Dyke, and W. Klemperer. The molecular beam spectrum and the structure of the hydrogen fluoride dimer. *The Journal of Chemical Physics*, 81(12):5417–5425, 1984.

[14] C. Maerker, P. von R. Schleyer, K.R. Liedl, T.-K. Ha, M. Quack, and M.A. Suhm. A critical analysis of electronic density functionals for structural, energetic, dynamic, and magnetic properties of hydrogen fluoride clusters. *Journal of Computational Chemistry*, 18(14):1695–1719, 1997.

[15] G.M. Chaban and R.B. Gerber. *Ab initio* calculations of anharmonic vibrational spectroscopy for hydrogen fluoride (HF)$_n$ ($n = 3, 4$) and mixed hydrogen fluoride/water (HF)$_n$(H2O)$_n$ ($n = 1, 2, 4$) clusters. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 58(4):887 – 898, 2002.

[16] S.E. McLain, C.J. Benmore, J.E. Siewenie, J.J. Molaison, and J.F.C. Turner. On the variation of the structure of liquid deuterium fluoride with temperature. *The Journal of Chemical Physics*, 121(13):6448–6455, 2004.

[17] M.E. Cournoyer and W.L. Jorgensen. An improved intermolecular potential function for simulations of liquid hydrogen fluoride. *Molecular Physics*, 51(1):119–132, 1984.

[18] T. Pfleiderer, I. Waldner, H. Bertagnolli, K. Todheide, and H.E. Fischer. The structure of liquid and supercritical deuterium fluoride from neutron scattering using high-pressure techniques. *The Journal of Chemical Physics*, 113(9):3690–3696, 2000.

[19] J.H. Simons and J.W. Bouknight. The Density and Surface Tension of Liquid Hydrogen Fluoride. *Journal of the American Chemical Society*, 54(1):129–135, 1932.

[20] I. Sheft, A.J. Perkins, and H.H. Hyman. Anhydrous hydrogen fluoride: Vapor pressure and liquid density. *Journal of Inorganic and Nuclear Chemistry*, 35(11):3677 – 3680, 1973.

[21] L. Partay, P. Jedlovszky, and R. Vallauri. Development of a new polarizable potential model of hydrogen fluoride and comparison with other effective models in liquid and supercritical states. *The Journal of Chemical Physics*, 124(18):184504, 2006.

[22] D.P. Visco and D.A. Kofke. Improved Thermodynamic Equation of State for Hydrogen Fluoride. *Industrial & Engineering Chemistry Research*, 38(10):4125–4129, 1999.

[23] P. Jedlovszky, M. Mezei, and R. Vallauri. Comparison of polarizable and nonpolarizable models of hydrogen fluoride in liquid and supercritical states: A Monte Carlo simulation study. *The Journal of Chemical Physics*, 115(21):9883–9894, 2001.

[24] N. Karger, T. Vardag, and H.-D. Lüdemann. p,T-dependence of self-diffusion in liquid hydrogen fluoride. *The Journal of Chemical Physics*, 100(11):8271–8276, 1994.

[25] K. Chu, J. Vojtchovský, B.H. McMahon, R.M. Sweet, J. Berendzen, and I. Schlichting. Structure of a ligand-binding intermediate in wild-type carbonmonoxy myoglobin. *Nature*, 403(6772):923–923, 2000.

[26] S.E. McLain, C.J. Benmore, J.E. Siewenie, J. Urquidi, and J.F.C. Turner. On the Structure of Liquid Hydrogen Fluoride. *Angewandte Chemie International Edition*, 43(15):1952–1955, 2004.

[27] D.S. Goodsell. Illustrations for public use. `http://mgl.scripps.edu/people/goodsell/illustration/public`, 1999. Accessed on January 30th, 2014.

[28] V. Riveros-Moreno and J.B. Wittenberg. The Self-Diffusion Coefficients of Myoglobin and Hemoglobin in Concentrated Solutions. *Journal of Biological Chemistry*, 247(3):895–901, 1972.

[29] S. Longeville, W. Doster, M. Diehl, R. Gähler, and W. Petry. Neutron Resonance Spin Echo: Oxygen Transport in a Crowded Protein Solutions. In *Lecture Notes in Physics*, volume 601 of *Lecture Notes in Physics*, pages 325–335. 2003.

[30] D.J. Wales. The energy landscape as a unifying theme in molecular science. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 363(1827):357–377, 2005.

[31] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

[32] J. Gao. Toward a Molecular Orbital Derived Empirical Potential for Liquid Simulations. *The Journal of Physical Chemistry B*, 101(4):657–663, 1997.

[33] W. Xie and J. Gao. Design of a Next Generation Force Field: The X-POL Potential. *Journal of Chemical Theory and Computation*, 3(6):1890–1900, 2007. PMID: 18985172.

[34] M. Bixon and S. Lifson. Potential functions and conformations in cycloalkanes. *Tetrahedron*, 23(2):769 – 784, 1967.

[35] W. Xie, L. Song, D.G. Truhlar, and J. Gao. The variational explicit polarization potential and analytical first derivative of energy: Towards a next generation force field. *The Journal of Chemical Physics*, 128(23):234108, 2008.

[36] G. Zhao, J.R. Perilla, E.L. Yufenyuy, X. Meng, B. Chen, J. Ning, J. Ahn, A.M. Gronenborn, K. Schulten, C. Aiken, and P. Zhang. Mature HIV-1 capsid structure by cryo-electron microscopy and all-atom molecular dynamics. *Nature*, 497(7451):643–646, 2013.

[37] A. Aksimentiev and K. Schulten. Imaging α-Hemolysin with Molecular Dynamics: Ionic Conductance, Osmotic Permeability, and the Electrostatic Potential Map. *Biophysical Journal*, 88(6):3745–3761, 2005.

[38] P.L. Freddolino, A.S. Arkhipov, S.B. Larson, A. McPherson, and K. Schulten. Molecular Dynamics Simulations of the Complete Satellite Tobacco Mosaic Virus. *Structure*, 14(3):437 – 449, 2006.

[39] R. Phillips and R. Milo. A feeling for the numbers in biology. *Proceedings of the National Academy of Sciences*, 106(51):21465–21471, 2009.

[40] M. Levitt and Warshel A. Computer simulation of protein folding. *Nature*, 253(5494):694–698, 1975.

[41] S. Tanaka and H.A. Scheraga. Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. *Macromolecules*, 9(6):945–950, 1976.

[42] B. Smit, P.A.J. Hilbers, K. Esselink, L.A.M. Rupert, N.M. van Os, and A.G. Schlijper. Computer simulations of a water/oil interface in the presence of micelles. *Nature*, 348(6302):624–625, 1990.

[43] G. Voth. *Coarse-Graining of Condensed Phase and Biomolecular Systems*. Taylor & Francis Group: Boca Raton, FL, 2008.

[44] S.J. Marrink, A.H. de Vries, and A.E. Mark. Coarse Grained Model for Semiquantitative Lipid Simulations. *The Journal of Physical Chemistry B*, 108(2):750–760, 2004.

[45] S.J. Marrink, H.J. Risselada, S. Yefimov, D.P. Tieleman, and A.H. de Vries. The MARTINI Force Field: Coarse Grained Model for Biomolecular Simulations. *The Journal of Physical Chemistry B*, 111(27):7812–7824, 2007. PMID: 17569554.

[46] D.L. Ermak and J.A. McCammon. Brownian dynamics with hydrodynamic interactions. *The Journal of Chemical Physics*, 69(4):1352–1360, 1978.

[47] J. Gao. A molecular-orbital derived polarization potential for liquid water. *The Journal of Chemical Physics*, 109(6):2346–2354, 1998.

[48] S.J. Wierzchowski, D.A. Kofke, and J. Gao. Hydrogen fluoride phase behavior and molecular structure: A QM/MM potential model approach. *The Journal of Chemical Physics*, 119(14):7365–7371, 2003.

[49] W. Xie, M. Orozco, D.G. Truhlar, and J. Gao. X-Pol Potential: An Electronic Structure-Based Force Field for Molecular Dynamics Simulation of a Solvated Protein in Water. *Journal of Chemical Theory and Computation*, 5(3):459–467, 2009.

[50] B.R. Brooks, C.L. Brooks, A.D. Mackerell, L. Nilsson, R.J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A.R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R.W. Pastor, C.B. Post, J.Z. Pu, M. Schaefer, B. Tidor, R.M. Venable, H.L. Woodcock, X. Wu, W. Yang, D.M. York, and M. Karplus. CHARMM: The biomolecular simulation program. *Journal of Computational Chemistry*, 30(10):1545–1614, 2009.

[51] J.C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R.D. Skeel, L. Kalé, and K. Schulten. Scalable molecular dynamics with NAMD. *Journal of Computational Chemistry*, 26(16):1781–1802, 2005.

[52] P. Zhang, L. Fiedler, H.R. Leverentz, D.G. Truhlar, and J. Gao. Polarized Molecular Orbital Model Chemistry. 2. The PMO Method. *Journal of Chemical Theory and Computation*, 7(4):857–867, 2011.

[53] P. Zhang, P. Bao, and J. Gao. Dipole preserving and polarization consistent charges. *Journal of Computational Chemistry*, 32(10):2127–2139, 2011.

[54] D.R. Hartree. The Wave Mechanics of an Atom with a Non-Coulomb Central Field. Part I. Theory and Methods. *Mathematical Proceedings of the Cambridge Philosophical Society*, 24(01):89–110, 0 1928.

[55] V. Fock. Näherungsmethode zur Lösung des quantenmechanischen Mehrkörperproblems. *Zeitschrift für Physik*, 61(1-2):126–148, 1930.

[56] C. Møller and M.S. Plesset. Note on an Approximation Treatment for Many-Electron Systems. *Phys. Rev.*, 46(7):618–622, Oct 1934.

[57] K. Raghavachari, G.W. Trucks, J.A. Pople, and M. Head-Gordon. A fifth-order perturbation comparison of electron correlation theories. *Chemical Physics Letters*, 157(6):479 – 483, 1989.

[58] A. Szabo and N.S. Ostlund. *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory*. Mcgraw-Hill (Tx), 1989.

[59] C.C.J. Roothaan. Self-Consistent Field Theory for Open Shells of Electronic Systems. *Rev. Mod. Phys.*, 32(2):179–185, Apr 1960.

[60] J.A. Pople and R.K. Nesbet. Self-Consistent Orbitals for Radicals. *The Journal of Chemical Physics*, 22(3):571–572, 1954.

[61] E. Schrödinger. An Undulatory Theory of the Mechanics of Atoms and Molecules. *Phys. Rev.*, 28(6):1049–1070, Dec 1926.

[62] M. Born and R. Oppenheimer. Zur Quantentheorie der Molekeln. *Annalen der Physik*, 389(20):457–484, 1927.

[63] Ø. Burrau. Berechnung des Energiewertes des Wasserstoffmolekel-Ions (H_2^+) im Normalzustand. *Naturwissenschaften*, 15(1):16–17, 1927.

[64] W. Pauli. Über den Zusammenhang des Abschlusses der Elektronengruppen im Atom mit der Komplexstruktur der Spektren. *Zeitschrift für Physik*, 31(1):765–783, 1925.

[65] J.C. Slater. The Theory of Complex Spectra. *Phys. Rev.*, 34(10):1293–1322, Nov 1929.

[66] C.C.J. Roothaan. New Developments in Molecular Orbital Theory. *Rev. Mod. Phys.*, 23(2):69–89, Apr 1951.

[67] J. C. Slater. Atomic Shielding Constants. *Phys. Rev.*, 36(1):57–64, Jul 1930.

[68] S.F. Boys. Electronic Wave Functions. I. A General Method of Calculation for the Stationary States of Any Molecular System. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 200(1063):pp. 542–554, 1950.

[69] M. Geller. Two-Center Coulomb Integrals. *The Journal of Chemical Physics*, 41(12):4006–4007, 1964.

[70] E. Clementi and D.R. Davis. Electronic structure of large molecular systems. *Journal of Computational Physics*, 1(2):223 – 244, 1966.

[71] T. Petersson and B. Hellsing. A detailed derivation of Gaussian orbital-based matrix elements in electron structure calculations. *European Journal of Physics*, 31(1):37, 2010.

[72] G.G. Hall. The Molecular Orbital Theory of Chemical Valency. VIII. A Method of Calculating Ionization Potentials. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 205(1083):541–552, 1951.

[73] Y. Beppu and I. Ninomiya. HQRII: A fast diagonalization subroutine. *Computers & Chemistry*, 6(2):87–91, 1982.

[74] Z.A. Tomašić. SHQRII: An improved, fast, portable diagonalization routine. *Computers & Chemistry*, 9(2):123–132, 1985.

[75] A.V. Bunge and C.F. Bunge. HQRII1: An accurate, portable and fast diagonalization routine. *Computers & Chemistry*, 10(4):259–268, 1986.

[76] P. Pulay. Convergence acceleration of iterative sequences. the case of SCF iteration. *Chemical Physics Letters*, 73(2):393–398, 1980.

[77] H.B. Schlegel and J.J.W. McDouall. Do You Have SCF Stability and Convergence Problems? In C. Ögretir and I. G. Csizmadia, editors, *Computational Advances in Organic Chemistry: Molecular Structure and Reactivity*, volume 330 of *NATO ASI Series*, pages 167–185. Springer Netherlands, 1991.

[78] C. van Alsenoy, C.-H. Yu, A. Peeters, J.M.L. Martin, and L. Schäfer. *Ab Initio* Geometry Determinations of Proteins. 1. Crambin. *The Journal of Physical Chemistry A*, 102(12):2246–2251, 1998.

[79] M.M. Teeter, S.M. Roe, and N.H. Heo. Atomic Resolution (0.83 Å) Crystal Structure of the Hydrophobic Protein Crambin at 130 K. *Journal of Molecular Biology*, 230(1):292–311, 1993.

[80] C. van Alsenoy. *Ab initio* calculations on large molecules: The multiplicative integral approximation. *Journal of Computational Chemistry*, 9(6):620–626, 1988.

[81] R. Ditchfield, W.J. Hehre, and J.A. Pople. Self-Consistent Molecular-Orbital Methods. IX. An Extended Gaussian-Type Basis for Molecular-Orbital Studies of Organic Molecules. *The Journal of Chemical Physics*, 54(2):724–728, 1971.

[82] J. Gao, P. Amara, C. Alhambra, and M.J. Field. A Generalized Hybrid Orbital (GHO) Method for the Treatment of Boundary Atoms in Combined QM/MM Calculations. *The Journal of Physical Chemistry A*, 102(24):4714–4721, 1998.

[83] P. Amara, M.J. Field, C. Alhambra, and J. Gao. The generalized hybrid orbital method for combined quantum mechanical/molecular mechanical calculations: formulation and tests of the analytical derivatives. *Theoretical Chemistry Accounts*, 104(5):336–343, 2000.

[84] T.J. Giese and D.M. York. Charge-dependent model for many-body polarization, exchange, and dispersion interactions in hybrid quantum mechanical/molecular mechanical calculations. *The Journal of Chemical Physics*, 127(19):194101, 2007.

[85] J.E. Jones. On the Determination of Molecular Fields. II. From the Equation of State of a Gas. *Proceedings of the Royal Society of London. Series A*, 106(738):463–477, 1924.

[86] H.R. Leverentz, J. Gao, and D.G. Truhlar. Using multipole point charge distributions to provide the electrostatic potential in the variational explicit polarization (X-Pol) potential. *Theoretical Chemistry Accounts*, 129(1):3–13, 2011.

[87] L. Song, J. Han, Y.-L. Lin, W. Xie, and J. Gao. Explicit Polarization (X-Pol) Potential Using *ab Initio* Molecular Orbital Theory and Density Functional Theory. *The Journal of Physical Chemistry A*, 113(43):11656–11664, 2009. PMID: 19618944.

[88] J.A. Pople, D.L. Beveridge, and P.A. Dobosh. Approximate Self-Consistent Molecular-Orbital Theory. V. Intermediate Neglect of Differential Overlap. *The Journal of Chemical Physics*, 47(6):2026–2033, 1967.

[89] M.J.S. Dewar and W. Thiel. Ground states of molecules. 38. The MNDO method. Approximations and parameters. *Journal of the American Chemical Society*, 99(15):4899–4907, 1977.

[90] J.J.P. Stewart. MOPAC: A semiempirical molecular orbital program. *Journal of Computer-Aided Molecular Design*, 4(1):1–103, 1990.

[91] M.J.S. Dewar and W. Thiel. A semiempirical model for the two-center repulsion integrals in the NDDO approximation. *Theoretica chimica acta*, 46(2):89–104, 1977.

[92] M.J.S. Dewar and Y. Yamaguchi. Analytical first derivatives of the energy in MNDO. *Computers & Chemistry*, 2(1):25–29, 1978.

[93] L. Fiedler, J. Gao, and D.G. Truhlar. Polarized Molecular Orbital Model Chemistry. 1. *Ab Initio* Foundations. *Journal of Chemical Theory and Computation*, 7(4):852–856, 2011.

[94] M. Isegawa, L. Fiedler, H.R. Leverentz, Y. Wang, S. Nachimuthu, J. Gao, and D.G. Truhlar. Polarized Molecular Orbital Model Chemistry 3. The PMO Method Extended to Organic Chemistry. *Journal of Chemical Theory and Computation*, 9(1):33–45, 2013.

[95] S. Grimme. Accurate description of van der Waals complexes by density functional theory including empirical corrections. *Journal of Computational Chemistry*, 25(12):1463–1473, 2004.

[96] R.S. Mulliken. Electronic Population Analysis on LCAO[Single Bond]MO Molecular Wave Functions. I. *The Journal of Chemical Physics*, 23(10):1833–1840, 1955.

[97] B.T. Thole and P.T. van Duijnen. A general population analysis preserving the dipole moment. *Theoretica chimica acta*, 63(3):209–221, 1983.

[98] L. Pauling. The Nature of the Chemical Bond. IV. The Energy of Single Bonds and the Relative Electronegativity of Atoms. *Journal of the American Chemical Society*, 54(9):3570–3582, 1932.

[99] A.D. Mackerell. Empirical force fields for biological macromolecules: Overview and issues. *Journal of Computational Chemistry*, 25(13):1584–1604, 2004.

[100] J.W. Ponder and D.A. Case. Force Fields for Protein Simulations. In V. Daggett, editor, *Protein Simulations*, volume 66 of *Advances in Protein Chemistry*, pages 27 – 85. Academic Press, 2003.

[101] B. Guillot. A reappraisal of what we have learnt during three decades of computer simulations on water. *Journal of Molecular Liquids*, 101(13):219 – 260, 2002. Molecular Liquids. Water at the New Millenium.

[102] C. Vega, J.L.F. Abascal, and P.G. Debenedetti. Physics and chemistry of water and ice. *Phys. Chem. Chem. Phys.*, 13(44):19660–19662, 2011.

[103] B. Schropp and P. Tavan. The Polarizability of Point-Polarizable Water Models: Density Functional Theory/Molecular Mechanics Results. *The Journal of Physical Chemistry B*, 112(19):6233–6240, 2008. PMID: 18198859.

[104] J.D. Bernal and R.H. Fowler. A Theory of Water and Ionic Solution, with Particular Reference to Hydrogen and Hydroxyl Ions. *The Journal of Chemical Physics*, 1(8):515–548, 1933.

[105] W.L. Jorgensen. Special Issue on Polarization. *Journal of Chemical Theory and Computation*, 3(6):1877–1877, 2007.

[106] H. J. C. Berendsen, J. P. M. Postama, W. F. van Gunsteren, and J. Hermans. In B. Pullmann, editor, *Intermolecular Forces*, page 331. D. Reidel Publishing Company, Dordrecht, 1981.

[107] W.L. Jorgensen, J. Chandrasekhar, J.D. Madura, R.W. Impey, and M.L. Klein. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics*, 79(2):926–935, 1983.

[108] F.J. Vesely. N-particle dynamics of polarizable Stockmayer-type molecules. *Journal of Computational Physics*, 24(4):361 – 371, 1977.

[109] A.E. Howard, U.C. Singh, M. Billeter, and P.A. Kollman. Many-body potential for molecular interactions. *Journal of the American Chemical Society*, 110(21):6984–6991, 1988.

[110] D.N. Bernardo, Y. Ding, K. Krogh-Jespersen, and R.M. Levy. An Anisotropic Polarizable Water Model: Incorporation of All-Atom Polarizabilities into Molecular Mechanics Force Fields. *The Journal of Physical Chemistry*, 98(15):4180–4187, 1994.

[111] J. Gao, D. Habibollazadeh, and L. Shao. A Polarizable Intermolecular Potential Function for Simulation of Liquid Alcohols. *The Journal of Physical Chemistry*, 99(44):16460–16467, 1995.

[112] J.M. Stout and C.E. Dykstra. A Distributed Model of the Electrical Response of Organic Molecules. *The Journal of Physical Chemistry A*, 102(9):1576–1582, 1998.

[113] J. Applequist, J.R. Carl, and K.-K. Fung. Atom dipole interaction model for molecular polarizability. Application to polyatomic molecules and determination of atom polarizabilities. *Journal of the American Chemical Society*, 94(9):2952–2960, 1972.

[114] B.T. Thole. Molecular polarizabilities calculated with a modified dipole interaction. *Chemical Physics*, 59(3):341 – 350, 1981.

[115] P.T. van Duijnen and M. Swart. Molecular and Atomic Polarizabilities: Thole's Model Revisited. *The Journal of Physical Chemistry A*, 102(14):2399–2407, 1998.

[116] H. Yu, T. Hansson, and W.F. van Gunsteren. Development of a simple, self-consistent polarizable model for liquid water. *The Journal of Chemical Physics*, 118(1):221–234, 2003.

[117] G. Lamoureux, E. Harder, I.V. Vorobyov, B. Roux, and A.D. MacKerell Jr. A polarizable model of water for molecular dynamics simulations of biomolecules. *Chemical Physics Letters*, 418(1-3):245 – 249, 2006.

[118] A.K. Rappe and W.A. Goddard. Charge equilibration for molecular dynamics simulations. *The Journal of Physical Chemistry*, 95(8):3358–3363, 1991.

[119] S.W. Rick, S.J. Stuart, and B.J. Berne. Dynamical fluctuating charge force fields: Application to liquid water. *The Journal of Chemical Physics*, 101(7):6141–6156, 1994.

[120] G.A. Kaminski, H.A. Stern, B.J. Berne, and R.A. Friesner. Development of an Accurate and Robust Polarizable Molecular Mechanics Force Field from *ab Initio* Quantum Chemistry. *The Journal of Physical Chemistry A*, 108(4):621–627, 2004.

[121] S.M. Valone. Quantum Mechanical Origins of the Iczkowski–Margrave Model of Chemical Potential. *Journal of Chemical Theory and Computation*, 7(7):2253–2261, 2011.

[122] W. Kohn, A.D. Becke, and R.G. Parr. Density Functional Theory of Electronic Structure. *The Journal of Physical Chemistry*, 100(31):12974–12980, 1996.

[123] A. Cembran, P. Bao, Y. Wang, L. Song, D.G. Truhlar, and J. Gao. On the Interfragment Exchange in the X-Pol Method. *Journal of Chemical Theory and Computation*, 6(8):2469–2476, 2010.

[124] J. Gao, A. Cembran, and Y. Mo. Generalized X-Pol Theory and Charge Delocalization States. *Journal of Chemical Theory and Computation*, 6(8):2402–2410, 2010.

[125] J. Han, D.G. Truhlar, and J. Gao. Optimization of the explicit polarization (X-Pol) potential using a hybrid density functional. *Theoretical Chemistry Accounts*, 131(3):1–15, 2012.

[126] Y. Wang, C.P. Sosa, A. Cembran, D.G. Truhlar, and J. Gao. Multilevel X-Pol: A Fragment-Based Method with Mixed Quantum Mechanical Representations of Different Fragments. *The Journal of Physical Chemistry B*, 116(23):6781–6788, 2012.

[127] Y. Mo, P. Bao, and J. Gao. Energy decomposition analysis based on a block-localized wavefunction and multistate density functional theory. *Phys. Chem. Chem. Phys.*, 13(15):6760–6775, 2011.

[128] J.A. Pople, D.P. Santry, and G.A. Segal. Approximate Self-Consistent Molecular Orbital Theory. I. Invariant Procedures. *The Journal of Chemical Physics*, 43(10):S129–S135, 1965.

[129] M.J.S. Dewar, E.G. Zoebisch, E.F. Healy, and J.J.P. Stewart. Development and use of quantum mechanical molecular models. 76. AM1: a new general purpose quantum mechanical molecular model. *Journal of the American Chemical Society*, 107(13):3902–3909, 1985.

[130] G.B. Rocha, R.O. Freire, A.M. Simas, and J.J.P. Stewart. RM1: A reparameterization of AM1 for H, C, N, O, P, S, F, Cl, Br, and I. *Journal of Computational Chemistry*, 27(10):1101–1111, 2006.

[131] J.J.P. Stewart. Optimization of parameters for semiempirical methods V: Modification of NDDO approximations and application to 70 elements. *Journal of Molecular Modeling*, 13(12):1173–1213, 2007.

[132] C.A. Morgado, J.P. McNamara, I.H. Hillier, N.A. Burton, and M.A. Vincent. Density Functional and Semiempirical Molecular Orbital Methods Including Dispersion Corrections for the Accurate Description of Noncovalent Interactions Involving Sulfur-Containing Molecules. *Journal of Chemical Theory and Computation*, 3(5):1656–1664, 2007.

[133] J.P. McNamara and I.H. Hillier. Semi-empirical molecular orbital methods including dispersion corrections for the accurate prediction of the full range of intermolecular interactions in biomolecules. *Phys. Chem. Chem. Phys.*, 9(19):2362–2370, 2007.

[134] J.P. McNamara, R. Sharma, M.A. Vincent, I.H. Hillier, and C.A. Morgado. The non-covalent functionalisation of carbon nanotubes studied by density functional and semi-empirical molecular orbital methods including dispersion corrections. *Phys. Chem. Chem. Phys.*, 10(1):128–135, 2008.

[135] T. Tuttle and W. Thiel. OMx-D: semiempirical methods with orthogonalization and dispersion corrections. Implementation and biochemical application. *Phys. Chem. Chem. Phys.*, 10(16):2159–2166, 2008.

[136] M. Korth and W. Thiel. Benchmarking Semiempirical Methods for Thermochemistry, Kinetics, and Noncovalent Interactions: OMx Methods Are Almost As Accurate and Robust As DFT-GGA Methods for Organic Molecules. *Journal of Chemical Theory and Computation*, 7(9):2929–2936, 2011.

[137] K. Jug and G. Geudtner. Treatment of hydrogen bonding in SINDO1. *Journal of Computational Chemistry*, 14(6):639–646, 1993.

[138] K.T. Tang and J.P. Toennies. An improved simple model for the van der Waals potential based on universal damping functions for the dispersion coefficients. *The Journal of Chemical Physics*, 80(8):3726–3741, 1984.

[139] S. Grimme, J. Antony, S. Ehrlich, and H. Krieg. A consistent and accurate *ab initio* parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *The Journal of Chemical Physics*, 132(15):154104, 2010.

[140] J.J.P. Stewart. *Semiempirical Molecular Orbital Methods*, pages 45–81. John Wiley & Sons, Inc., 2007.

[141] M.C. Zerner. *Semiempirical Molecular Orbital Methods*, pages 313–365. John Wiley & Sons, Inc., 2007.

[142] M.J.S. Dewar and W. Thiel. Ground states of molecules. 39. MNDO results for molecules containing hydrogen, carbon, nitrogen, and oxygen. *Journal of the American Chemical Society*, 99(15):4907–4917, 1977.

[143] K. Liu, M.G. Brown, and R.J. Saykally. Terahertz Laser Vibration-Rotation Tunneling Spectroscopy and Dipole Moment of a Cage Form of the Water Hexamer. *The Journal of Physical Chemistry A*, 101(48):8995–9010, 1997.

[144] M. Piris, J.M. Matxain, X. Lopez, and J.M. Ugalde. Communications: Accurate description of atoms and molecules by natural orbital functional theory. *The Journal of Chemical Physics*, 132(3):031103, 2010.

[145] G.S. Tschumper, M.L. Leininger, B.C. Hoffman, E.F. Valeev, H.F. Schaefer III, and M. Quack. Anchoring the water dimer potential energy surface with explicitly

correlated computations and focal point analyses. *The Journal of Chemical Physics*, 116(2):690–701, 2002.

[146] G. Maroulis. Static hyperpolarizability of the water dimer and the interaction hyperpolarizability of two water molecules. *The Journal of Chemical Physics*, 113(5):1813–1820, 2000.

[147] S.A. Clough, Y. Beers, G.P. Klein, and L.S. Rothman. Dipole moment of water from Stark measurements of H_2O, HDO, and D_2O. *The Journal of Chemical Physics*, 59(5):2254–2259, 1973.

[148] J. Gao and Y. Wang. Communication: Variational many-body expansion: Accounting for exchange repulsion, charge delocalization, and dispersion in the fragment-based explicit polarization method. *The Journal of Chemical Physics*, 136(7):071101, 2012.

[149] T.J. Giese, H. Chen, T. Dissanayake, G.M. Giambau, H. Heldenbrand, M. Huang, E.R. Kuechler, T.-S. Lee, M.T. Panteva, B.K. Radak, and D.M. York. A Variational Linear-Scaling Framework to Build Practical, Efficient Next-Generation Orbital-Based Quantum Force Fields. *Journal of Chemical Theory and Computation*, 9(3):1417–1427, 2013.

[150] T. Nakano, T. Kaminuma, T. Sato, K. Fukuzawa, Y. Akiyama, M. Uebayasi, and K. Kitaura. Fragment molecular orbital method: use of approximate electrostatic potential. *Chemical Physics Letters*, 351(56):475 – 480, 2002.

[151] W.D. Cornell, P. Cieplak, C.I. Bayly, I.R. Gould, K.M. Merz, D.M. Ferguson, D.C. Spellmeyer, T. Fox, J.W. Caldwell, and P.A. Kollman. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *Journal of the American Chemical Society*, 117(19):5179–5197, 1995.

[152] P. Cieplak, J. Caldwell, and P. Kollman. Molecular mechanical models for organic and biological systems going beyond the atom centered two body additive approximation: aqueous solution free energies of methanol and N-methyl acetamide, nucleic acid base, and amide hydrogen bonding and chloroform/water partition coefficients of the nucleic acid bases. *Journal of Computational Chemistry*, 22(10):1048–1057, 2001.

[153] J. Wang, P. Cieplak, and P.A. Kollman. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *Journal of Computational Chemistry*, 21(12):1049–1074, 2000.

[154] A.J. Stone. *The theory of intermolecular forces*. Oxford University Press, Oxford, 1996.

[155] M.S. Gordon, L. Slipchenko, H. Li, and J.H. Jensen. Chapter 10 The Effective Fragment Potential: A General Method for Predicting Intermolecular Interactions. volume 3 of *Annual Reports in Computational Chemistry*, pages 177 – 193. Elsevier, 2007.

[156] W.J. Hehre, L. Radom, P.v.R. Schleyer, and J.A. Pople. Ab initio *Molecular Orbital Theory*. John Wiley & Sons, New York, 1986.

[157] J. Li, T. Zhu, C.J. Cramer, and D.G. Truhlar. New Class IV Charge Model for Extracting Accurate Partial Charges from Wave Functions. *The Journal of Physical Chemistry A*, 102(10):1820–1831, 1998.

[158] A.V. Marenich, S.V. Jerome, C.J. Cramer, and D.G. Truhlar. Charge Model 5: An Extension of Hirshfeld Population Analysis for the Accurate Description of

Molecular Interactions in Gaseous and Condensed Phases. *Journal of Chemical Theory and Computation*, 8(2):527–541, 2012.

[159] M. Swart, P.T. van Duijnen, and J.G. Snijders. A charge analysis derived from an atomic multipole expansion. *Journal of Computational Chemistry*, 22(1):79–88, 2001.

[160] J.A. Pople and G.A. Segal. Approximate Self-Consistent Molecular Orbital Theory. II. Calculations with Complete Neglect of Differential Overlap. *The Journal of Chemical Physics*, 43(10):S136–S151, 1965.

[161] H.A. Stern, F. Rittner, B.J. Berne, and R.A. Friesner. Combined fluctuating charge and polarizable dipole models: Application to a five-site water potential function. *The Journal of Chemical Physics*, 115(5):2237–2251, 2001.

[162] P. Zhang, D.G. Truhlar, and J. Gao. Fragment-based quantum mechanical methods for periodic systems with Ewald summation and mean image charge convention for long-range electrostatic interactions. *Phys. Chem. Chem. Phys.*, 14(21):7821–7829, 2012.

[163] H.C. Andersen. Molecular dynamics simulations at constant pressure and/or temperature. *The Journal of Chemical Physics*, 72(4):2384–2393, 1980.

[164] E.A. Koopman and C.P. Lowe. Advantages of a Lowe-Andersen thermostat in molecular dynamics simulations. *The Journal of Chemical Physics*, 124(20):204103, 2006.

[165] S. Miyamoto and P. A. Kollman. Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. *Journal of Computational Chemistry*, 13(8):952–962, 1992.

[166] K. Nam, J. Gao, and D.M. York. An Efficient Linear-Scaling Ewald Method for Long-Range Electrostatic Interactions in Combined QM/MM Calculations. *Journal of Chemical Theory and Computation*, 1(1):2–13, 2005.

[167] J. Gao and X. Xia. *A priori* evaluation of aqueous polarization effects through Monte Carlo QM-MM simulations. *Science*, 258(5082):631–635, 1992.

[168] K. Kitaura, E. Ikeo, T. Asada, T. Nakano, and M. Uebayasi. Fragment molecular orbital method: an approximate computational method for large molecules. *Chemical Physics Letters*, 313(34):701 – 706, 1999.

[169] J. Gao, J. Han, and P. Zhang. *MCSOL, version 2012xp*. Minneapolis, 2012.

[170] M.J.M. Mazack and J. Gao. *X-Pol, version 2013a1*. University of Minnesota, 2013.

[171] M.J. Frisch, G.W. Trucks, and H.B. Schlegel *et al. GAUSSIAN 09, Rev A.02*. Gaussian, Inc., Wallingford, CT, 2009.

[172] G. Lamoureux, Jr. A.D. MacKerell, and B. Roux. A simple polarizable model of water based on classical Drude oscillators. *The Journal of Chemical Physics*, 119(10):5185–5197, 2003.

[173] W.S. Benedict, N. Gailar, and E.K. Plyler. Rotation-Vibration Spectra of Deuterated Water Vapor. *The Journal of Chemical Physics*, 24(6):1139–1165, 1956.

[174] C.J. Burnham and S.S. Xantheas. Development of transferable interaction models for water. IV. A flexible, all-atom polarizable potential (TTM2-F) based on geometry dependent charges derived from an *ab initio* monomer dipole moment surface. *The Journal of Chemical Physics*, 116(12):5115–5124, 2002.

[175] H. Partridge and D.W. Schwenke. The determination of an accurate isotope dependent potential energy surface for water from extensive *ab initio* calculations and experimental data. *The Journal of Chemical Physics*, 106(11):4618–4639, 1997.

[176] E. Whalley and D.D. Klug. Effect of hydrogen bonding on the direction of the dipole-moment derivative of the O–H bond in the water molecule. *The Journal of Chemical Physics*, 84(1):78–80, 1986.

[177] L.S. Rothman, C.P. Rinsland, A. Goldman, S.T. Massie, D.P. Edwards, J.-M. Flaud, A. Perrin, C. Camy-Peyret, V. Dana, J.-Y. Mandin, J. Schroeder, and A. McCann. Reprint of: The HITRAN molecular spectroscopic database and HAWKS (HITRAN Atmospheric Workstation): 1996 edition. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 111(11):1568 – 1613, 2010. 50 Years of JQSRT.

[178] L.A. Curtiss, D.J. Frurip, and M. Blander. Studies of molecular association in H_2O and D_2O vapors by measurement of thermal conductivity. *The Journal of Chemical Physics*, 71(6):2703–2711, 1979.

[179] S. Nachimuthu, J. Gao, and D.G. Truhlar. A benchmark test suite for proton transfer energies and its use to test electronic structure model chemistries. *Chemical Physics*, 400:8 – 12, 2012.

[180] S. Sadhukhan, D. Mu noz, C. Adamo, and G.E. Scuseria. Predicting proton transfer barriers with density functional methods. *Chemical Physics Letters*, 306(1-2):83 – 87, 1999.

[181] R. Kumar, R.A. Christie, and K.D. Jordan. A Modified MSEVB Force Field for Protonated Water Clusters. *The Journal of Physical Chemistry B*, 113(13):4111–4118, 2009.

[182] P. Goyal, M. Elstner, and Q. Cui. Application of the SCC-DFTB Method to Neutral and Protonated Water Clusters and Bulk Water. *The Journal of Physical Chemistry B*, 115(20):6790–6805, 2011.

[183] W.L. Jorgensen and J.D. Madura. Temperature and size dependence for Monte Carlo simulations of TIP4P water. *Molecular Physics*, 56(6):1381–1392, 1985.

[184] B.G. Kyle. *Chemical and Process Thermodynamics*. Prentice Hall PTR, 1999.

[185] W. Wagner and A. Pruss. The IAPWS Formulation 1995 for the Thermodynamic Properties of Ordinary Water Substance for General and Scientific Use. *Journal of Physical and Chemical Reference Data*, 31(2):387–535, 2002.

[186] L. Haar, E. Gallagher, and G. Kell. *NBS/NRC Steam Tables: Thermodynamic and Transport Properties and Computer Programs for Vapor and Liquid States of Water in SI Units*. Hemisphere Publishing Corporation, Washington, 1984.

[187] J. Wang, P. Cieplak, Q. Cai, M.-J. Hsieh, J. Wang, Y. Duan, and R. Luo. Development of Polarizable Models for Molecular Mechanical Calculations. 3. Polarizable Water Models Conforming to Thole Polarization Screening Schemes. *The Journal of Physical Chemistry B*, 116(28):7999–8008, 2012.

[188] C.A. Coulson and D. Eisenberg. Interactions of H_2O Molecules in Ice. I. The Dipole Moment of an H_2O Molecule in Ice. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 291(1427):445–453, 1966.

[189] J.W. Caldwell and P.A. Kollman. Structure and Properties of Neat Liquids Using Nonadditive Molecular Dynamics: Water, Methanol, and N-Methylacetamide. *The Journal of Physical Chemistry*, 99(16):6208–6219, 1995.

[190] P.L. Silvestrelli and M. Parrinello. Water Molecule Dipole in the Gas and in the Liquid Phase. *Phys. Rev. Lett.*, 82(16):3308–3311, Apr 1999.

[191] M. Neumann. The dielectric constant of water. Computer simulations with the MCY potential. *The Journal of Chemical Physics*, 82(12):5663–5672, 1985.

[192] J.L. Aragones, L.G. MacDowell, and C. Vega. Dielectric Constant of Ices and Water: A Lesson about Water Interactions. *The Journal of Physical Chemistry A*, 115(23):5745–5758, 2011.

[193] H.E. Alper and R.M. Levy. Computer simulations of the dielectric properties of water: Studies of the simple point charge and transferrable intermolecular potential models. *The Journal of Chemical Physics*, 91(2):1242–1251, 1989.

[194] J.A. Barker and R.O. Watts. Monte Carlo studies of the dielectric properties of water-like models. *Molecular Physics*, 26(3):789–792, 1973.

[195] M.P. Allen and D.J. Tildesley. *Computer Simulations of liquids*. Oxford University Press, Oxford, 1987.

[196] S.-B. Zhu and C.F. Wong. Sensitivity analysis of water thermodynamics. *The Journal of Chemical Physics*, 98(11):8892–8899, 1993.

[197] M. Sprik. Hydrogen bonding and the static dielectric constant in liquid water. *The Journal of Chemical Physics*, 95(9):6762–6769, 1991.

[198] P. Hochtl, S. Boresch, W. Bitomsky, and O. Steinhauser. Rationalization of the dielectric properties of common three-site water models in terms of their force field parameters. *The Journal of Chemical Physics*, 109(12):4927–4937, 1998.

[199] P. Ren and J.W. Ponder. Temperature and Pressure Dependence of the AMOEBA Water Model. *The Journal of Physical Chemistry B*, 108(35):13427–13437, 2004.

[200] J. Gao. Energy components of aqueous solution: Insight from hybrid QM/MM simulations using a polarizable solvent model. *Journal of Computational Chemistry*, 18(8):1061–1071, 1997.

[201] H.J.C. Berendsen, J.R. Grigera, and T.P. Straatsma. The missing term in effective pair potentials. *The Journal of Physical Chemistry*, 91(24):6269–6271, 1987.

[202] G.S. Kell. Density, thermal expansivity, and compressibility of liquid water from 0° to 150°. Correlations and tables for atmospheric pressure and saturation reviewed and expressed on 1968 temperature scale. *Journal of Chemical & Engineering Data*, 20(1):97–105, 1975.

[203] C.A. Angell, W.J. Sichina, and M. Oguni. Heat capacity of water at extremes of supercooling and superheating. *The Journal of Physical Chemistry*, 86(6):998–1002, 1982.

[204] C. Vega, M.M. Conde, C. McBride, J.L.F. Abascal, E.G. Noya, R. Ramirez, and L.M. Sese. Heat capacity of water: A signature of nuclear quantum effects. *The Journal of Chemical Physics*, 132(4):046101, 2010.

[205] K. Krynicki, C.D. Green, and D.W. Sawyer. Pressure and temperature dependence of self-diffusion in water. *Faraday Discuss. Chem. Soc.*, 66(0):199–208, 1978.

[206] S. Tazi, A. Boțan, M. Salanne, V. Marry, P. Turq, and B. Rotenberg. Diffusion coefficient and shear viscosity of rigid water models. *Journal of Physics: Condensed Matter*, 24(28):284117, 2012.

[207] A. Abragam. *The Principles of Nuclear Magnetism.* Clarendon Press, Oxford England, 1961.

[208] D.J. Wilbur, T. DeFries, and J. Jonas. Self-diffusion in compressed liquid heavy water. *The Journal of Chemical Physics*, 65(5):1783–1786, 1976.

[209] J. Barthel, K. Bachhuber, R. Buchner, and H. Hetzenauer. Dielectric spectra of some common solvents in the microwave region. Water and lower alcohols. *Chemical Physics Letters*, 165(4):369 – 373, 1990.

[210] A.K. Soper. The radial distribution functions of water and ice from 220 to 673 K and at pressures up to 400 MPa. *Chemical Physics*, 258(23):121 – 137, 2000.

[211] T. Head-Gordon and M.E. Johnson. Tetrahedral structure or chains for liquid water. *Proceedings of the National Academy of Sciences*, 103(21):7973–7977, 2006.

[212] L.A. Baez and P. Clancy. Existence of a density maximum in extended simple point charge water. *The Journal of Chemical Physics*, 101(11):9837–9840, 1994.

[213] B. Chen, J. Xing, and J.I. Siepmann. Development of Polarizable Water Force Fields for Phase Equilibrium Calculations. *The Journal of Physical Chemistry B*, 104(10):2391–2401, 2000.

[214] S. Nose. A unified formulation of the constant temperature molecular dynamics methods. *The Journal of Chemical Physics*, 81(1):511–519, 1984.

[215] W.G. Hoover. Canonical dynamics: Equilibrium phase-space distributions. *Phys. Rev. A*, 31(3):1695–1697, Mar 1985.

[216] G. Raabe and R.J. Sadus. Molecular dynamics simulation of the dielectric constant of water: The effect of bond flexibility. *The Journal of Chemical Physics*, 134(23):234501, 2011.

[217] K. Ichikawa, Y. Kameda, T. Yamaguchi, H. Wakita, and M. Misawa. Neutron-diffraction investigation of the intramolecular structure of a water molecule in the liquid phase at high temperatures. *Molecular Physics*, 73(1):79–86, 1991.

[218] Y. Mo and J. Gao. Polarization and Charge-Transfer Effects in Aqueous Solution via *Ab Initio* QM/MM Simulations. *The Journal of Physical Chemistry B*, 110(7):2976–2980, 2006.

[219] S. Lifson. *J. Chim. Phys. Physicochim. Biol.*, 65(40), 1968.

[220] M. Levitt and S. Lifson. Refinement of protein conformations using a macromolecular energy minimization procedure. *Journal of Molecular Biology*, 46(2):269 – 279, 1969.

[221] M. Levitt. The birth of computational structural biology. *Nature Structural Biology*, 8(5):392 – 393, 1969.

[222] A.G. Streng. Miscibility and compatibility of some liquefied and solidified gases at low temperatures. *Journal of Chemical & Engineering Data*, 16(3):357–359, 1971.

[223] R. McIntosh, T.-S. Kuan, and E. Defresart. Hydrogen fluoride vapor etching for Pre-Epi silicon surface preparation. *Journal of Electronic Materials*, 21(1):57–60, 1992.

[224] A.L. Horvath. Heat Capacity of Liquid Hydrogen Fluoride-A Discrepancy. *Zeitschrift für Physikalische Chemie*, 78:209–210, 1972.

[225] M. Deraman, J.C. Dore, J.G. Powles, J.H. Holloway, and P. Chieux. Structural studies of liquid hydrogen fluoride by neutron diffraction. *Molecular Physics*, 55(6):1351–1367, 1985.

[226] J. Janzen and L.S. Bartell. Electron-Diffraction Structural Study of Polymeric Gaseous Hydrogen Fluoride. *The Journal of Chemical Physics*, 50(8):3611–3618, 1969.

[227] M. Atoji and W. N. Lipscomb. The crystal structure of hydrogen fluoride. *Acta Cryst.*, 7(2):173–175, 1954.

[228] W.L. Jorgensen and M.E. Cournoyer. Quantum and statistical studies of liquids. 1. An intermolecular potential function for the hydrogen fluoride dimer from *ab initio* 6-31G computations. *Journal of the American Chemical Society*, 100(16):4942–4945, 1978.

[229] W.L. Jorgensen. Quantum and statistical mechanical studies of liquids. 2. Monte-Carlo simulations of liquid hydrogen fluoride. *Journal of the American Chemical Society*, 100(25):7824–7831, 1978.

[230] W.L. Jorgensen. Basis set dependence of the structure and properties of liquid hydrogen fluoride. *The Journal of Chemical Physics*, 70(12):5888–5897, 1979.

[231] M.L. Klein, I.R. McDonald, and S.F. O'Shea. An intermolecular force model for $(HF)_2$. *The Journal of Chemical Physics*, 69(1):63–66, 1978.

[232] M.L. Klein and I.R. McDonald. Structure and dynamics of associated molecular systems. I. Computer simulation of liquid hydrogen fluoride. *The Journal of Chemical Physics*, 71(1):298–308, 1979.

[233] W.J. Hehre, R.F. Stewart, and J.A. Pople. Self-Consistent Molecular-Orbital Methods. I. Use of Gaussian Expansions of Slater-Type Atomic Orbitals. *The Journal of Chemical Physics*, 51(6):2657–2664, 1969.

[234] D.R. Yarkony, S.V. O'Neil, H.F. Schaefer III, C.P. Baskin, and C.F. Bender. Interaction potential between two rigid HF molecules. *The Journal of Chemical Physics*, 60(3):855–865, 1974.

[235] P. Jedlovszky and R. Vallauri. Computer simulation study of liquid HF with a new effective pair potential model. *Molecular Physics*, 92(2):331–336, 1997.

[236] P. Jedlovszky and R. Vallauri. Computer simulations of liquid HF by a newly developed polarizable potential model. *The Journal of Chemical Physics*, 107(23):10166–10176, 1997.

[237] J. Han, M.J.M. Mazack, P. Zhang, D.G. Truhlar, and J. Gao. Quantum mechanical force field for water with explicit electronic polarization. *The Journal of Chemical Physics*, 139(5):054503, 2013.

[238] M. Kreitmeir, G. Heusel, H. Bertagnolli, K. Todheide, C.J. Mundy, and G.J. Cuello. Structure of dense hydrogen fluoride gas from neutron diffraction and molecular dynamics simulations. *The Journal of Chemical Physics*, 122(15):154511, 2005.

[239] R. Car and M. Parrinello. Unified Approach for Molecular Dynamics and Density-Functional Theory. *Phys. Rev. Lett.*, 55(22):2471–2474, Nov 1985.

[240] U. Röthlisberger and M. Parrinello. *Ab initio* molecular dynamics simulation of liquid hydrogen fluoride. *The Journal of Chemical Physics*, 106(11):4658–4664, 1997.

[241] M. Kreitmeir, H. Bertagnolli, J.J. Mortensen, and M. Parrinello. *Ab initio* molecular dynamics simulation of hydrogen fluoride at several thermodynamic states. *The Journal of Chemical Physics*, 118(8):3639–3645, 2003.

[242] S. Izvekov and G.A. Voth. Effective Force Field for Liquid Hydrogen Fluoride from *Ab Initio* Molecular Dynamics Simulation Using the Force-Matching Method. *The Journal of Physical Chemistry B*, 109(14):6573–6586, 2005. PMID: 16851738.

[243] M.J. McGrath, J.N. Ghogomu, C.J. Mundy, I.-F.W. Kuo, and J.I. Siepmann. First principles Monte Carlo simulations of aggregation in the vapor phase of hydrogen fluoride. *Phys. Chem. Chem. Phys.*, 12(27):7678–7687, 2010.

[244] M.J. McGrath, I.-F.W. Kuo, and J.I. Siepmann. Liquid structures of water, methanol, and hydrogen fluoride at ambient conditions from first principles molecular dynamics simulations with a dispersion corrected density functional. *Phys. Chem. Chem. Phys.*, 13(44):19943–19950, 2011.

[245] A.D. Becke. Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys. Rev. A*, 38(6):3098–3100, Sep 1988.

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

[247] S. Grimme. Semiempirical GGA-type density functional constructed with a long-range dispersion correction. *Journal of Computational Chemistry*, 27(15):1787–1799, 2006.

[248] J. Gao, J. Han, P. Zhang, and M.J.M. Mazack. *MCSOL, version 2013xp*. Minneapolis, 2013.

[249] J.J.P. Stewart. Optimization of parameters for semiempirical methods I. Method. *Journal of Computational Chemistry*, 10(2):209–220, 1989.

[250] Y. Zhao and D.G. Truhlar. The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other functionals. *Theoretical Chemistry Accounts*, 120(1-3):215–241, 2008.

[251] P.L. Fast, M.L. Sanchez, and D.G. Truhlar. Multi-coefficient Gaussian-3 method for calculating potential energy surfaces. *Chemical Physics Letters*, 306(5):407–410, 1999-06-18T00:00:00.

[252] C.M. Tratz, P.L. Fast, and D.G. Truhlar. Improved coefficients for the scaling all correlation and multi-coefficient correlation methods. *PhysChemComm*, 2(14):70–79, 1999.

[253] M. J. M. Mazack and J. Gao. *X-Pol, version 2014a1*. University of Minnesota, 2014.

[254] E.J. Bylaska, W.A. de Jong, N. Govind, K. Kowalski, T.P. Straatsma, M. Valiev, D. Wang, E. Apra, T.L. Windus, J. Hammond, P. Nichols, S. Hirata, M.T. Hackler, Y. Zhao, P.-D. Fan, R.J. Harrison, M. Dupuis, D.M.A. Smith, J. Nieplocha, V. Tipparaju, M. Krishnan, A. Vazquez-Mayagoitia, Q. Wu, T. Van Voorhis, A.A. Auer, M. Nooijen, L.D. Crosby, E. Brown, G. Cisneros, G.I. Fann, H. Fruchtl, J. Garza, K. Hirao, R. Kendall, J.A. Nichols, K. Tsemekhman, K. Wolinski, J. Anchell, D. Bernholdt, P. Borowski, T. Clark, D. Clerc, H. Dachsel, M. Deegan, K. Dyall, D. Elwood, E. Glendening, M. Gutowski, A. Hess, J. Jaffe, B. Johnson, J. Ju, R. Kobayashi, R. Kutteh, Z. Lin, R. Littlefield, X. Long, B. Meng, T. Nakajima, S. Niu, L. Pollack, M. Rosing, G. Sandrone, M. Stave, H. Taylor, G. Thomas, J. van Lenthe, A. Wong, and Z. Zhang. *NWChem, A Computational Chemistry Package for Parallel Computers, Version 5.1.1*. Pacific Northwest National Laboratory, Richland, Washington 99352-0999, USA, 2009.

[255] P.-O. Löwdin. On the Nonorthogonality Problem. volume 5 of *Advances in Quantum Chemistry*, pages 185 – 199. Academic Press, 1970.

[256] L. Pierce, N. di Cianni, and R.H. Jackson. Centrifugal Distortion Effects in Asymmetric Rotor Molecules. I. Quadratic Potential Constants and Average Structure of Oxygen Difluoride from the Ground-State Rotational Spectrum. *The Journal of Chemical Physics*, 38(3):730–739, 1963.

[257] H. C. Andersen. Rattle: A velocity version of the shake algorithm for molecular dynamics calculations. *J. Comput. Phys.*, 52(1):24 – 34, 1983.

[258] B. Baburao and D. P. Visco. Isothermal compressibility maxima of hydrogen fluoride in the supercritical and superheated vapor regions. *J. Phys. Chem. B*, 110(51):26204–26210, 2006.

[259] Allen P. Minton. The influence of macromolecular crowding and macromolecular confinement on biochemical reactions in physiological media. *Journal of Biological Chemistry*, 276(14):10577–10580, 2001.

[260] Minton A.P. Influence of excluded volume upon macromolecular structure and associations in 'crowded' media. *Current Opinion in Biotechnology*, 8(1):65–69, 1997.

[261] G.B. Ralston. Effects of "crowding" in protein solutions. *Journal of Chemical Education*, 67(10):857, 1990.

[262] R.J. Ellis. Macromolecular crowding: obvious but underappreciated. *Trends in Biochemical Sciences*, 26(10):597–604, 2001.

[263] P.M. Haggie and A.S. Verkman. Diffusion of Tricarboxylic Acid Cycle Enzymes in the Mitochondrial Matrix *in vivo* : Evidence for Restricted Mobility of a Multienzyme Complex. *Journal of Biological Chemistry*, 277(43):40782–40788, 2002.

[264] A.S. Verkman. Solute and macromolecule diffusion in cellular aqueous compartments. *Trends in Biochemical Sciences*, 27(1):27 – 33, 2002.

[265] J.D. Dwyer and V.A. Bloomfield. Brownian dynamics simulations of probe and self-diffusion in concentrated protein and DNA solutions. *Biophysical Journal*, 65(5):1810–1816, 1993.

[266] J.D. Dwyer and V.A. Bloomfield. Brownian dynamics simulation of probe diffusion in DNA: effects of probe size, charge and DNA concentration. *Biophysical Chemistry*, 57(1):55 – 64, 1995. Macromolecular Crowding.

[267] D.J. Bicout and M.J. Field. Stochastic Dynamics Simulations of Macromolecular Diffusion in a Model of the Cytoplasm of Escherichia coli. *The Journal of Physical Chemistry*, 100(7):2489–2497, 1996.

[268] T. Ando and J. Skolnick. Brownian dynamics simulation of macromolecule diffusion in a protocell. *Proceedings of the International Conference of the Quantum Bio-Informatics IV*, 28:413–426, 2011.

[269] T. Ando and J. Skolnick. Crowding and hydrodynamic interactions likely dominate in vivo macromolecular motion. *Proceedings of the National Academy of Sciences of the United States of America*, 107:18457–18462, 2010.

[270] S.R. McGuffee and A.H. Elcock. Diffusion, Crowding & Protein Stability in a Dynamic Molecular Model of the Bacterial Cytoplasm. *PLoS Computational Biology*, 6(3):e1000694, 03 2010.

[271] T. Frembgen-Kesner and A.H. Elcock. Striking Effects of Hydrodynamic Interactions on the Simulated Diffusion and Folding of Proteins. *Journal of Chemical Theory and Computation*, 5(2):242–256, 2009.

[272] A.H. Elcock. Atomic-level observation of macromolecular crowding effects: Escape of a protein from the GroEL cage. *Proceedings of the National Academy of Sciences*, 100(5):2340–2344, 2003.

[273] Q. Wang and M.S. Cheung. A physics-based approach of coarse-graining the cytoplasm of escherichia coli (cgcyto). *Biophysical Journal*, 102(10):2353 – 2361, 2012.

[274] O. Byron. Construction of hydrodynamic bead models from high-resolution X-ray crystallographic or nuclear magnetic resonance data. *Biophysical Journal*, 72:408–415, 1997.

[275] A. Bulgac and M. Adamuţi-Trache. Molecular dynamics of rigid molecules. *The Journal of Chemical Physics*, 105(3):1131–1141, 1996.

[276] A. Einstein. Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen. *Annalen der Physik*, 17:549–560, 1905.

[277] P. Langevin. Sur la théorie du mouvement brownien. *C.R. Acad. Sci. (Paris)*, 146:530–533, 1908.

[278] M.T. Tyn and T.W. Gusek. Prediction of diffusion coefficients of proteins. *Biotechnology and Bioengineering*, 35(4):327–338, 1990.

[279] L. Accardi, W. Freudenberg, and M. Ohya, editors. *Importance of excluded volume and hydrodynamic interactions on macromolecular diffusion in vivo*, volume 30, Tokyo, Japan, 2013. WORLD SCIENTIFIC.

[280] T. Ando and J. Skolnick. On the Importance of Hydrodynamic Interactions in Lipid Membrane Formation. *Biophysical Journal*, 104:96–105, 1/2013 2013.

[281] T. Ando, E. Chow, Y. Saad, and J. Skolnick. Krylov subspace methods for computing hydrodynamic interactions in Brownian dynamics simulations. *The Journal of Chemical Physics*, 137:064106, 08/2012 2012.

[282] T. Ando, E. Chow, and J. Skolnick. Dynamic simulation of concentrated macromolecular solutions with screened long-range hydrodynamic interactions: Algorithm and limitations. *The Journal of Chemical Physics*, 139(12):121922–1, 2013.

[283] H. Yamakawa. Transport properties of polymer chains in dilute solution: Hydrodynamic interaction. *The Journal of Chemical Physics*, 53(1):436–443, 1970.

[284] J. Rotne and S. Prager. Variational treatment of hydrodynamic interaction in polymers. *The Journal of Chemical Physics*, 50(11):4831–4837, 1969.

[285] G.E.P. Box and M.E. Muller. A note on the generation of random normal deviates. *Annals of Mathematical Statistics*, 29(2):610–611, 1958.

[286] J. Callaway, M. Cummings, B. Deroski, P. Esposito, A. Forman, P. Langdon, M. Libeson, J. McCarthy, J. Sikora, D. Xue, E. Abola, F. Bernstein, N. Manning, R. Shea, D. Stampf, and J. Sussman. Protein Data Bank contents guide: Atomic coordinate entry format description. *Brookhaven National Laboratory*, 1996.

[287] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, 2000.

[288] A. Bondi. van der Waals Volumes and Radii. *The Journal of Physical Chemistry*, 68(3):441–451, 1964.

[289] D.J. Harrington, K. Adachi, and W.E. Royer Jr. The high resolution crystal structure of deoxyhemoglobin S. *Journal of Molecular Biology*, 272(3):398 – 407, 1997.

[290] J.D. Weeks, D. Chandler, and H.C. Andersen. Role of repulsive forces in determining the equilibrium structure of simple liquids. *The Journal of Chemical Physics*, 54(12):5237–5247, 1971.

[291] P. Debye and E. Hückel. Zur Theorie der Elektrolyte. *Physikalische Zeitschrift*, 24:185–206, 1923.

[292] B.S. Duncan and A.J. Olson. Approximation and characterization of molecular surfaces. *Biopolymers*, 33(2):219–229, 1993.

[293] P. Decaudin. Cartoon-looking rendering of 3D-scenes. *INRIA*, Technical Report 2919, 1996.

[294] W.H. Press, B.P. Flannery, S.A. Teukolsky, and W.T. Vetterling. *Numerical Recipes*. Cambridge University Press, New York, NY, third edition, 2007.

[295] F.S. Hill. *Computer Graphics using OpenGL*. Prentice Hall (NJ), second edition edition, 2001.

[296] J.H. Lambert. *Photometria, sive de Mensura et Gradibus Luminis, Colorum et Umbrae*. Augsburg, 1760.

[297] R. Rost. *The OpenGL Shading Language*. Addison-Wesley (MA), 2006.

[298] R.C. Hardy and R.L. Cottington. Viscosity of deuterium oxide and water in the range 5° to 125° C. *Journal of Research of the National Bureau of Standards*, 42(6):573–578, 1949.

[299] S. Longeville, W. Doster, and G. Kali. Myoglobin in crowded solutions: structure and diffusion. *Chemical Physics*, 292(2-3):413–424, 2003.

[300] M. Karplus and McCammon J.A. Molecular dynamics simulations of biomolecules. *Nature Structural Biology*, 9:646–652, 2002.

[301] R.O. Dror, M.Ø. Jensen, D.W. Borhani, and D.E. Shaw. Exploring atomic resolution physiology on a femtosecond to millisecond timescale using molecular dynamics simulations. *The Journal of General Physiology*, 135(6):555–562, 2010.

[302] V.A. Voelz, G.R. Bowman, K. Beauchamp, and V.S. Pande. Molecular Simulation of *ab initio* Protein Folding for a Millisecond Folder NTL9(1-39). *Journal of the American Chemical Society*, 132(5):1526–1528, 2010.

[303] T.H. Dunning, K. Schulten, J. Tromp, J.P. Ostriker, K. Droegemeier, Ming Xue, and P. Fussell. Science and Engineering in the Petascale Era. *Computing in Science Engineering*, 11(5):28–37, 2009.

[304] A. Arkhipov, W.H. Roos, G.J.L. Wuite, and K. Schulten. Elucidating the Mechanism behind Irreversible Deformation of Viral Capsids. *Biophysical Journal*, 97(7):2061 – 2069, 2009.

[305] L. Lu, S. Izvekov, A. Das, H.C. Andersen, and G.A. Voth. Efficient, Regularized, and Scalable Algorithms for Multiscale Coarse-Graining. *Journal of Chemical Theory and Computation*, 6(3):954–965, 2010.

[306] J.A. McCammon, B.R. Gelin, and M. Karplus. Dynamics of folded proteins. *Nature*, 267:585–590, 1977.

[307] W.L. Jorgensen and J. Tirado-Rives. The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *Journal of the American Chemical Society*, 110(6):1657–1666, 1988.

[308] C. Jekeli, J. Lee, and J. Kwon. On the computation and approximation of ultra-high-degree spherical harmonic series. *Journal of Geodesy*, 81(9):603–615, 2007.

[309] M.A. Wieczorek. Gravity and Topography of the Terrestrial Planets. In Gerald Schubert, editor, *Treatise on Geophysics*, pages 165 – 206. Elsevier, Amsterdam, 2007.

[310] O. Colombo. *Numerical Methods for Harmonic Analysis on the Sphere*. PhD thesis, Ohio State University, 1981.

[311] N. Sneeuw. Global spherical harmonic analysis by least-squares and numerical quadrature methods in historical perspective. *Geophysical Journal International*, 118(3):707–716, 1994.

[312] N.L. Max and E.D. Getzoff. Spherical harmonic molecular surfaces. *Computer Graphics and Applications, IEEE*, 8(4):42 –50, July 1988.

[313] B.S. Duncan and A.J. Olson. Shape analysis of molecular surfaces. *Biopolymers*, 33(2):231–238, 1993.

[314] B.S. Duncan and A.J. Olson. Texture mapping parametric molecular surfaces. *Journal of Molecular Graphics*, 13(4):258 – 264, 1995.

[315] M.F. Sanner and A.J. Olson. Reduced surface: An efficient way to compute molecular surfaces. *Biopolymers*, 38(3):305–320, 1996.

[316] D.W. Ritchie and G.J.L. Kemp. Fast computation, rotation, and comparison of low resolution spherical harmonic molecular surfaces. *J. Computational Chemistry*, 20(4):383–395, 1999.

[317] W. Cai, X. Shao, and B. Maigret. Protein-ligand recognition using spherical harmonic molecular surfaces: towards a fast and efficient filter for large virtual throughput screening. *Journal of Molecular Graphics and Modelling*, 20(4):313 – 328, 2002.

[318] R.J. Morris, R.J. Najmanovich, A. Kahraman, and J.M. Thornton. Real spherical harmonic expansion coefficients as 3D shape descriptors for protein binding pocket and ligand comparisons. *Bioinformatics*, 21(10):2347–2355, 2005.

[319] L. Mavridis, B.D. Hudson, and D.W. Ritchie. Toward High Throughput 3D Virtual Screening Using Spherical Harmonic Surface Representations. *Journal of Chemical Information and Modeling*, 47(5):1787–1796, 2007.

[320] L. Mak, S. Grandison, and R.J. Morris. An extension of spherical harmonics to region-based rotationally invariant descriptors for molecular shape description and comparison. *Journal of Molecular Graphics and Modelling*, 26(7):1035 – 1045, 2008.

[321] V. Venkatraman, L. Sael, and D. Kihara. Potential for Protein Surface Shape Analysis Using Spherical Harmonics and 3D Zernike Descriptors. *Cell Biochemistry and Biophysics*, 54(1):23–32, 2009.

[322] B.S. Duncan and A.J. Olson. Approximation and visualization of large-scale motion of protein surfaces. *Journal of Molecular Graphics*, 13(4):250 – 257, 1995.

[323] B. Brooks and M. Karplus. Normal modes for specific motions of macromolecules: application to the hinge-bending mode of lysozyme. *Proceedings of the National Academy of Sciences of the United States of America*, 82(15):4995–4999, 1985.

[324] A. Amadei, A.B.M. Linssen, and H.J.C. Berendsen. Essential dynamics of proteins. *Proteins: Structure, Function, and Bioinformatics*, 17(4):412–425, 1993.

[325] B.R. Brooks, D. Janežič, and M. Karplus. Harmonic analysis of large systems. I. Methodology. *Journal of Computational Chemistry*, 16(12):1522–1542, 1995.

[326] D. Janežič and B.R. Brooks. Harmonic analysis of large systems. II. Comparison of different protein models. *Journal of Computational Chemistry*, 16(12):1543–1553, 1995.

[327] D. Janežič and B.R. Brooks. Harmonic analysis of large systems. III. Comparison with molecular dynamics. *Journal of Computational Chemistry*, 16(12):1554–1556, 1995.

[328] N. Wu, Y. Mo, J. Gao, and E.F. Pai. Electrostatic stress in catalysis: Structure and mechanism of the enzyme orotidine monophosphate decarboxylase. *Proceedings of the National Academy of Sciences of the United States of America*, 97(5):2017–2022, 2000.

[329] T. Dey, J. Giesen, and S. Goswami. Shape Segmentation and Matching with Flow Discretization. In *Algorithms and Data Structures*, volume 2748 of *Lecture Notes in Computer Science*, pages 25–36. Springer Berlin / Heidelberg, 2003.

[330] R. Klees, R. Tenzer, I. Prutkin, and T. Wittwer. A data-driven approach to local gravity field modelling using spherical radial basis functions. *Journal of Geodesy*, 82(8):457–471, 2008.

[331] A. Arkhipov, P.L. Freddolino, K. Imada, K. Namba, and K. Schulten. Coarse-Grained Molecular Dynamics Simulations of a Rotating Bacterial Flagellum. *Biophysical Journal*, 91(12):4589 – 4597, 2006.

[332] S. Izvekov, J.M.J. Swanson, and G.A. Voth. Coarse-Graining in Interaction Space: A Systematic Approach for Replacing Long-Range Electrostatics with Short-Range Potentials. *The Journal of Physical Chemistry B*, 112(15):4711–4724, 2008.

[333] B. Lee and F.M. Richards. The interpretation of protein structures: Estimation of static accessibility. *Journal of Molecular Biology*, 55(3):379 – 400, IN3–IN4, 1971.

[334] M.L. Connolly. Solvent-accessible surfaces of proteins and nucleic acids. *Science*, 221(4612):709–713, 1983.

[335] K.A. Henzler-Wildman, M. Lei, V. Thai, S.J. Kerns, M. Karplus, and D. Kern. A hierarchy of timescales in protein dynamics is linked to enzyme catalysis. *Nature*, 450(7171):913–916, 2007.

[336] A. Arkhipov, Y. Yin, and K. Schulten. Four-Scale Description of Membrane Sculpting by BAR Domains. *Biophysical Journal*, 95(6):2806 – 2821, 2008.

[337] Z. Zhang, L. Lu, W.G. Noid, V. Krishna, J. Pfaendtner, and G.A. Voth. A Systematic Methodology for Defining Coarse-Grained Sites in Large Biomolecules. *Biophysical Journal*, 95(11):5073 – 5083, 2008.

[338] Z. Zhang, J. Pfaendtner, A. Grafmüller, and G.A. Voth. Defining Coarse-Grained Representations of Large Biomolecules and Biomolecular Complexes from Elastic Network Models. *Biophysical Journal*, 97(8):2327 – 2337, 2009.

[339] A.H. Stroud and Secrest D. *Gaussian Quadrature Formulas*. Prentice-Hall, Englewood Cliffs, NJ, 1966.

[340] I. Sloan and R. Womersley. Extremal Systems of Points and Numerical Integration on the Sphere. *Advances in Computational Mathematics*, 21(1):107–125, 2004.

[341] M.J. Mohlenkamp. A fast transform for spherical harmonics. *Journal of Fourier Analysis and Applications*, 5(2):159–184, 1999.

[342] F.J. Luque, M. Bachs, and M. Orozco. An optimized AM1/MST method for the MST-SCRF representation of solvated systems. *Journal of Computational Chemistry*, 15(8):847–857, 1994.

[343] H. Fischer, I. Polikarpov, and A.F. Craievich. Average protein density is a molecular-weight-dependent function. *Protein Science*, 13(10):2825–2828, 2004.

[344] E.T. White, W.H. Tan, J.M. Ang, S. Tait, and J.D. Litster. The density of a protein crystal. *Powder Technology*, 179(1-2):55 – 58, 2007.

[345] H. Stoll and H. Preuß. On the direct calculation of localized HF orbitals in molecule clusters, layers and solids. *Theoretica chimica acta*, 46(1):11–21, 1977.

[346] I.G. Johnston, A.A. Louis, and J.P.K. Doye. Modelling the self-assembly of virus capsids. *Journal of Physics: Condensed Matter*, 22(10):104101, 2010.

[347] R.I. Weed, C.F. Reed, and G. Berg. IS HEMOGLOBIN AN ESSENTIAL STRUCTURAL COMPONENT OF HUMAN ERYTHROCYTE MEMBRANES?*. *The Journal of Clinical Investigation*, 42(4):581–588, 4 1963.

[348] R.R. Gabdoulline and R.C Wade. Effective Charges for Macromolecules in Solvent. *Journal of Physical Chemistry*, 100:3868–3878, 1996.

[349] B.H. Besler, K.M. Merz, and P.A. Kollman. Atomic Charges Derived from Semiempirical Methods. *Journal of Computational Chemistry*, 11(4):431–439, 1990.

[350] J.D. Jackson. *Classical Electrodynamics*. John Wiley & Sons, Inc., 1999.

[351] F. Fogolari, P. Zuccato, G. Esposito, and P. Viglino. Biomolecular Electrostatics with the Linearized Poisson-Boltzmann Equation. *Biophysical Journal*, 76:1–16, 1999.

[352] S. Jo, T. Kim, V.G. Iyer, and W. Im. CHARMM-GUI: A web-based graphical user interface for CHARMM. *Journal of Computational Chemistry*, 29(11):1859–1865, 2008.

[353] S. Jo, M. Vargyas, J. Vasko-Szedlar, B. Roux, and W. Im. PBEQ-Solver for online visualization of electrostatic potential of biomolecules. *Nucleic Acids Research*, 36(suppl 2):W270–W275, 2008.

[354] M.J.M. Mazack, A. Cembran, and J. Gao. Internal Dynamics of an Analytically Coarse-Grained Protein. *Journal of Chemical Theory and Computation*, 6(11):3601–3612, 2010.

[355] J. Blais, D. Provins, and M. Soofi. Spherical harmonic transforms for discrete multiresolution applications. *The Journal of Supercomputing*, 38(2):173–187, 2006.

[356] J. Blais. Discrete Spherical Harmonic Transforms: Numerical Preconditioning and Optimization. In Marian Bubak, Geert van Albada, Jack Dongarra, and Peter Sloot, editors, *Computational Science - ICCS 2008*, volume 5102 of *Lecture Notes in Computer Science*, pages 638–645. Springer Berlin / Heidelberg, 2008.

[357] C. Jekeli. Spherical harmonic analysis, aliasing, and filtering. *Journal of Geodesy*, 70(4):214–223, 1996.

[358] M.J. Mohlenkamp. *A Fast Transform for Spherical Harmonics*. PhD thesis, Yale University, 1997.

[359] J.R. Driscoll and D.M. Healy. Computing Fourier Transforms and Convolutions on the 2-Sphere. *Advances in Applied Mathematics*, 15(2):202 – 250, 1994.

Appendix A

A Charge-Fitting Procedure for Coarse-Grained Proteins

A.1 Introduction

The accurate modeling of electrostatics in simulating protein-protein interactions is a critically important component in the construction of a force field for use in many protein simulations. In this appendix, we present the physical description of electrostatic potential in the presence of an ionic solution, and how to fit effective, point charges using numerical solutions of the Poisson-Boltzmann equation. The method presented here is nearly identical to the method of Gabdoulline and Wade [348], and uses the linear, least-squares charge-fitting procedure of Besler, Merz, and Kollman [349].

A.2 Background

A.2.1 Poisson Equation

Electrostatic potential can be described by Poisson's equation. Poisson's equation is an elliptic partial differential equation arising as a direct consequence of Gauss' law (Eq. A.1) and Faraday's law of induction in the absence of a changing magnetic field (Eq.

A.2) [350].

$$\nabla \cdot E = \frac{\rho}{\epsilon_0} \tag{A.1}$$

$$\nabla \times E = -\frac{\partial B}{\partial t} = 0 \quad \Leftrightarrow \quad E = -\nabla U \tag{A.2}$$

Together, these two equations give rise to Poisson's equation (Eq. A.3), which determines the electrostatic potential U for a given charge density ρ where ϵ_0 is the permittivity of free space.

$$\nabla^2 U = \frac{\rho}{\epsilon_0} \tag{A.3}$$

In general, the principal difficulty in solving Poisson's equation is the large computational cost required for using the charge density of the atomistic system with explicit solvent. In practice, the solvent is treated implicitly through the use of a Boltzmann distribution, which results in the more frequently used Poisson-Boltzmann equation.

A.2.2 Poisson-Boltzmann Equation

Interactions between charged particles, such as proteins in ionic solution, can be described by the Poisson-Boltzmann equation (PBE) (Eq. A.4). In this approach, the proteins are treated as low dielectric cavities containing partial charges described by the charge density ρ^f. The remaining $\bar{\rho}^m$ term, further detailed in Eq. A.5, describes the charge density of the solvent, where c_i^∞ is the concentration of the ion type i at a distance of infinity from the solute. Additionally, $\epsilon(\vec{r})$ describes the position-dependent distribution of the dielectric media. The PBE is essentially a modified version of Poisson's equation, describing the ions in solution by a Boltzmann distribution.

$$\nabla \cdot [\epsilon(\vec{r})\nabla U(\vec{r})] = \bar{\rho}^m(\vec{r}) - \rho^f(\vec{r}) \tag{A.4}$$

$$\bar{\rho}^m = \sum_i c_i^\infty z_i q e^{\frac{-z_i q U(\vec{r})}{kT}} \tag{A.5}$$

The PBE is nonlinear due to the exponential term in Eq. A.5. To avoid this nonlinearity, the exponential term is expanded into a two-term Taylor series under the assumption that $(z_i qU/kT) << 1$, resulting in Eq. A.6, which is called the linearized Poisson-Boltzmann equation (LPBE) [351].

$$\nabla \cdot [\epsilon(\vec{r}) \nabla U(\vec{r})] = \left(\sum_i c_i^\infty \frac{z_i^2 q^2}{kT} \right) U(\vec{r}) - \rho^f \tag{A.6}$$

While Eq. A.6 is much easier to solve than Eq. A.4, its solution requires the use of numerical techniques such as finite difference, finite element, or boundary element methods. Such methods can produce very accurate results, but have a high computational cost, making a faster, approximate method desirable for coarse-grained MD simulations of many proteins.

A.2.3 Debye-Hückel Equation

The Debye-Hückel equation is the result of approximating the solution to the LPBE by ignoring the effects from the charge density ρ^f and forcing the distribution of the dielectric media to be uniform (i.e. $\epsilon(\vec{r}) = \epsilon_r \epsilon_0$) [291]. Such assumptions result in Eq. A.7 – a form of the Helmholtz equation.

$$\nabla^2 U(\vec{r}) = \frac{1}{\epsilon_r \epsilon_0} \left(\sum_i c_i^\infty \frac{z_i^2 q^2}{kT} \right) U(\vec{r}) \tag{A.7}$$

In the spherical coordinate system, the general solution to this Helmholtz equation is simply a linear combination of two exponential functions divided by the radial distance r.

$$U(\vec{r}) = A\frac{e^{-\kappa r}}{r} + B\frac{e^{\kappa r}}{r} = A\frac{e^{-\kappa r}}{r} \tag{A.8}$$

For electrostatic potential problems, the arbitrary constant B is 0 due to the condition that electrostatic potential decays as $r \to \infty$. In general, the other arbitrary constant A depends on the properties of the ionic solution as well as the units. The constant κ^{-1}, called the Debye-Hückel screening length, can be interpreted as the distance needed for significant charge separation to occur and is equal to the reciprocal square root of the summation in Eq. A.7.

Because of its simplicity, the Debye-Hückel equation provides a convenient means for quickly approximating the solution of the LPBE, and is therefore desirable to use as a potential for charge fitting and modeling protein-protein interactions.

A.3 Charge Fitting

The goal of any approximation is to increase the simplicity in solving the problem while minimizing the deviation from the ideal solution. Such is the philosophy behind what is called *charge fitting*.

A.3.1 Description

In charge fitting, n coarse-grained charge sites are chosen for a system of N atoms, where $n << N$. The benefit of such an approach is clear in that the computational cost of computing the pairwise interaction forces between charged sites is $\mathcal{O}(n^2) << \mathcal{O}(N^2)$. Charges are fit to each site using a linear, least-squares procedure to minimize the error in electrostatic potential between that produced by the fit charges using the Debye-Hückel equation and a numerical solution to the PBE on a Cartesian grid. Due to the approximation made when deriving the Debye-Hückel equation that ignores the charge density on the inside of the protein, those grid points which are inside the protein are removed from the least-squares fitting procedure.

A.3.2 Computational Details

The charge fitting procedure is based upon finding the charges $q_1, q_2, \ldots, q_n$ such that the function z of Eq. A.9 reaches a global minimum.

$$z = \gamma + \lambda g \tag{A.9}$$

$$\gamma(q_1, q_2, \ldots, q_n) = \sum_{i=1}^{m} (V_i - E_i)^2 \tag{A.10}$$

$$E_i = \sum_{j=1}^{n} \frac{q_j e^{-\kappa r_{ij}}}{r_{ij}} \tag{A.11}$$

$$g = \sum_{j=1}^{n} (q_j - q_{tot}) = 0 \tag{A.12}$$

The fitting procedure contains one linear constraint, g of Eq. A.12, requiring the summation of the individual charges to be equal to the charge of the protein. After substitution of all variables and constraints, we obtain the following expression for z.

$$z = \sum_{i=1}^{m} \left(V_i - \sum_{j=1}^{n} \frac{q_j e^{-\kappa r_{ij}}}{r_{ij}} \right)^2 + \lambda \sum_{j=1}^{n} (q_j - q_{tot}) \tag{A.13}$$

Upon finding the roots of the $n+1$ partial derivatives of z, we arrive at the system of equations $Aq = B$ seen in Eqs. A.14, A.15, and A.16, which can be solved by matrix inversion due to the fact that rank(A) $= n+1$. Note that $\sum_{j=1}^{n} q_j = q_{tot}$ is satisfied, and the last element of the q vector is the Lagrange multiplier for the total charge constraint.

$$\begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1n} & 1 \\ A_{11} & A_{12} & \cdots & A_{1n} & 1 \\ \vdots & \vdots & \ddots & \vdots & 1 \\ A_{n1} & A_{n2} & \cdots & A_{nn} & 1 \\ 1 & 1 & 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} q_1 \\ q_2 \\ \vdots \\ q_n \\ \lambda \end{pmatrix} = \begin{pmatrix} B_1 \\ B_2 \\ \vdots \\ B_n \\ q_{tot} \end{pmatrix} \tag{A.14}$$

$$A_{jk} = \sum_{i=1}^{m} \frac{e^{-\kappa(r_{ij}+r_{ik})}}{r_{ij}r_{ik}} \tag{A.15}$$

$$B_k = \sum_{i=1}^{m} \frac{V_i e^{-\kappa r_{ik}}}{r_{ik}} \tag{A.16}$$

The above provides a general procedure for determining effective charge at a given position for an ionically-screened system modeled by the Debye-Hückel equation. Note that all charge positions and charge distances r_{ij} and r_{ik} are known *a priori*. We discuss their selection in the following section.

A.4 Selecting Charge Sites

In principle, all charge sites can be selected arbitrarily. However, it is important to choose sites which give physically plausible charges. In order to make some generalizations about how the sites should be chosen, we tried selecting charge sites based upon two different coarse-graining models applied to the crystal structure of a hemoglobin (PDB:2DN2) having a net charge of +2.

The numerical solution of the PBE was calculated by a finite difference method using CHARMM c36a5 [50] and automatically-generated scripts from `http://charmm-gui.org` [352,353]. A Cartesian grid was placed at the center of mass of the crystal structure of the hemoglobin, reaching an additional 5Å on each of the six sides of the grid for

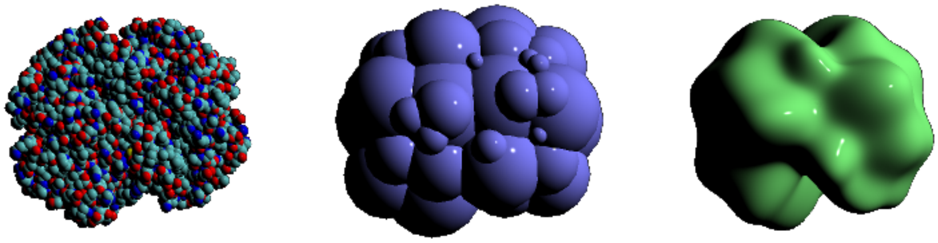


Figure A.1: The two methods of coarse-graining used for charge fitting. Left to right: Atomistic model, coarse-graining of Byron, ACG model.

additional grid points outside of the protein. The spacing between grid points was 1Å and the grid had dimensions of 77x77x69.

A.4.1 Coarse-Graining of Byron

Our first selection of sites for fitting used a coarse-graining model similar to the uniform-grid coarse-graining approach of Byron [274]. In this approach, the center of each of the 36 beads we used for approximating the hemoglobin was assigned a charge using the fitting procedure described here. We observed from this fit that charges centered on beads in the interior of the hemoglobin were very large in magnitude and had a high standard deviation of 19.9 (Table A.1).

A.4.2 Surface Sites

Next, fit charges on 12 and 42 sites on the ACG surface [354] of the protein using the angular grid of a geodesic sphere. The results of the fitting showed the charges to be of a much more plausible value than those of the spatial/volume approximating coarse-graining model of Byron, having a standard deviation of around 4.4 and 3.9 for the 12 and 42 surface charges respectively (Table A.1). This result suggests that surface charges are more important than the buried charges in the determination of the overall

Statistic	Byron	ACG (42)	ACG (12)
Mean	0.0556	0.0476	0.1667
StdDev	19.914	3.932	4.400
Min	-88.66	-14.39	-7.442
Max	37.94	8.923	8.457
Range	126.60	23.32	15.90
Sum	2.000	2.000	2.000

Table A.1: Table showing statistical data for charged sites using the coarse-graining of Byron with 36 sites and geodesically spaced ACG surface charges for 12 and 42 sites respectively. Unites are given in electron charge.

potential from the numerically solved PBE.

A.4.3 ACG Charges

One possible improvement of the surface model, which uses a finite number of sites for charges, could be achieved through the assignment of a continuous surface-charge density, which interacting proteins could feel through an integration over space. Similar to ACG, this could be represented as a spherical harmonic expansion (Eq. A.17).

$$\rho(\theta, \phi) = \sum_{l=0}^{L} \sum_{m=-l}^{l} c_{lm} Y_l^m(\theta, \phi) \tag{A.17}$$

A.5 Conclusion

Our goal in using this charge-fitting procedure is ultimately to use the results for modeling protein-protein interactions in a cellular environment. One conclusion that can be drawn from the fact that fit surface-charges are more physically plausible than the fit charges on buried sites is that the correct modeling of the properties on the protein surface are more relevant to accurately modeling protein-protein interactions than the properties of buried atoms. This principle is well-fitting with the analytical coarse-graining philosophy of our ACG model for proteins.

Appendix B

Algorithm for Spherical Harmonic Expansion and Evaluation

In this appendix, we summarize a numerical procedure for spherical harmonic expansion and evaluation; additional details may be found in Ref. [310] (see also Refs. [341, 355]). The numerical methods are widely used in geopotential modeling and an expansion of degree, and order up to 3800 has been reported corresponding to a ground resolution of about 5 km [356] (for the protein OMPDC considered here, it would correspond to an astonishing surface resolution of about 0.005 Å). It is useful to recast Eq. 6.1 for the surface function $S(\theta, \phi)$ in terms of harmonic coefficients as follows:

$$S(\theta, \phi) = \sum_{m=0}^{L} \sum_{l=m}^{L} a_{lm}^{c} \overline{P}_l^m(\cos\theta) \cos m\phi + a_{lm}^{s} \overline{P}_l^m(\cos\theta) \sin m\phi \tag{B.1}$$

Notice also that the order of the summations over degree and order has been switched. This is especially important in modern spherical harmonic analysis because the latitude and longitude data can now be treated independently, resulting in a two-step computational algorithm. [310,311,341] The expansion coefficients are determined based on a set of sampling data $\{S(\theta_i, \phi_j)\}$ on a grid of equispaced $2M$ points in (for Fourier transform) and M Gauss-Legendre quadrature nodes in $\cos\theta$ (for integration):

$$a_{lm}^{c} = \int_0^{\pi} \left[\int_0^{2\pi} \frac{1}{\sqrt{(1+\delta_{m0})\pi}} S(\theta,\phi) \cos{(m\phi)} d\phi \right] \overline{P}_l^m(\cos\theta) \sin\theta d\theta \tag{B.2}$$

$$a_{lm}^{s} = \int_0^{\pi} \left[\int_0^{2\pi} \frac{1}{\sqrt{(1+\delta_{m0})\pi}} S(\theta,\phi) \sin{(m\phi)} d\phi \right] \overline{P}_l^m(\cos\theta) \sin\theta d\theta \tag{B.3}$$

The grid points are defined as follows:

$$\phi_j = j\Delta\phi = j\frac{\pi}{M}; \quad j = 0, 1, \ldots, 2M-1 \tag{B.4}$$

$$P_{M+1}(\cos\theta_i) = 0; \quad i = 1, 2, \ldots, M+1 \tag{B.5}$$

where $P_{M+1}(\cos\theta_i)$ is the Legendre polynomial. The Gauss-Legendre quadrature weights can be determined using the expression [294, 339]

$$w_i = 2\left[\frac{\sin{(\theta_i)}}{(M+1)P_M(\cos\theta_i)}\right]^2 \tag{B.6}$$

From this set of data, the maximum degree of the expansion coefficients that can be determined is $L = M$ because the maximum number of terms in the summation over l in Eq. B.1, which is $L+1$ when $m = 0$, must be less than or equal to the number of sampling points in the longitudinal direction θ. Here, we do not address the issue of aliasing, [357] and we typically use a larger number of sampled grid points than the maximum degree used in the expansion.

Alternatively, if equal spaced points are used in θ, which is equivalent to the Chebychev nodes in $\cos\theta$, at least $2M+1$ sampling points are needed for degree L in the expansion coefficients as opposed to a minimum of $M = L + 1$ points with the Gauss-Legendre quadrature. In this case, the Chebychev weights are obtained using the formula below: [341,358,359]

$$w_i = \frac{\sqrt{2}}{M} \sum_{l=0}^{M-1} \frac{1}{2l+1} \sin\left([2l+1]\theta_i\right) \tag{B.7}$$

For convenience in the rest of the discussion, we use the degree of the expansion L to define the grid divisions throughout.

B.1 Spherical Harmonic Expansion

For the spherical harmonics expansion, a two-step computation algorithm is used. The first step corresponds to a Fourier transform in the inner integrals in Eqs. B.2 and B.3. For a given value m, the discretized Fourier series in the inner integrals of Eqs. B.2 and B.3 are expressed as follows:

$$U(\theta_i, \hat{m}) = \frac{1}{\sqrt{(1+\delta_{m0})\pi}} \sum_{j=0}^{2L-1} S(\theta_i, \phi_j) \cos m\phi_j, \quad i = 1, \ldots, L+1 \tag{B.8}$$

$$V(\theta_i, \hat{m}) = \frac{1}{\sqrt{(1+\delta_{m0})\pi}} \sum_{j=0}^{2L-1} S(\theta_i, \phi_j) \sin m\phi_j, \quad i = 1, \ldots, L+1 \tag{B.9}$$

where δ_{m0} is the Kronecker delta, and the notation $\hat{m}$ is used to emphasize that the Fourier series can be efficiently performed by Fast Fourier Transform (FFT).

The second step involves integration by Gauss-Legendre quadrature (or equally spaced Chebychev quadrature which requires twice as many sampling points) to yield the $2 \times (L-m+1)$ expansion coefficients:

$$a_{lm}^{c} = \sum_{i=1}^{L+1} w_i \bar{P}_m^l(\cos\theta_i) U(\theta_i, \hat{m}); \quad l = m, \ldots, L \tag{B.10}$$

$$a_{lm}^{s} = \sum_{i=1}^{L+1} w_i \bar{P}_m^l(\cos\theta_i) V(\theta_i, \hat{m}); \quad l = m, \ldots, L \tag{B.11}$$

If the values $U(\theta_i, \hat{m})$ and $V(\theta_i, \hat{m})$ are arranged as column vectors $\mathbf{u}(\hat{m})$ and $\mathbf{v}(\hat{m})$, respectively, the above equations can be conveniently written in matrix form:

$$\mathbf{a}_m^c = \mathbf{P}(m)^T \mathbf{W} \mathbf{u}(\hat{m}) \tag{B.12}$$

$$\mathbf{a}_m^s = \mathbf{P}(m)^T \mathbf{W} \mathbf{v}(\hat{m}) \tag{B.13}$$

where $\mathbf{a}_m^c$ and $\mathbf{a}_m^s$ are the expansion coefficients of Eqs. B.10 and B.11 arranged as column vectors, $\mathbf{W} = \text{diag}\{w_i\}$ is an $(L+1) \times (L+1)$ diagonal matrix consisting of the quadrature weights and the matrix for the precomputed values of the normalized associated Legendre polynomial is arranged as follows:

$$\mathbf{P}(m) = \begin{pmatrix} \bar{P}_m^m(\cos\theta_1) & \cdots & \bar{P}_L^m(\cos\theta_1) \\ \vdots & \ddots & \vdots \\ \bar{P}_m^m(\cos\theta_{L+1}) & \cdots & \bar{P}_L^m(\cos\theta_{L+1}) \end{pmatrix} \tag{B.14}$$

The operation for the first step has a computation scale of $O(L\log{(L)})$ using FFT, whereas the second step is of $O(L^2)$ for each order m. Thus, the overall procedure scales $O(L^2\log{(L)}) + O(L^3)$. Obviously, the $L+1$ parallels can be fully distributed over different processors, each having an overall computational scaling of $O(L^2)$; this is particularly suited for GPUs by choosing the number of parallels equal to that of the processors. Note that a fast spherical harmonic transform algorithm similar to that of FFT has been described. [341,358]

B.2 Spherical Harmonic Evaluation

Evaluation (or spherical harmonic synthesis) of surface function values also involves two computational steps. For a fixed colatitude θ_i, the first step is to compute intermediate vectors $\hat{\mathbf{u}}(\theta_i)$ and $\hat{\mathbf{v}}(\theta_i)$ over l for $0 \leq m \leq L$:

$$\hat{\mathbf{u}}(\theta_i) = \mathbf{P}_i(m)\mathbf{a}_m^c \tag{B.15}$$

$$\hat{\mathbf{v}}(\theta_i) = \mathbf{P}_i(m)\mathbf{a}_m^s \tag{B.16}$$

In the second step, the $2L$ longitudinal values are computed by Fast Fourier Transform for the following discrete series:

$$S(\theta_i, \phi_j) = \sum_{m=0}^{L} \hat{U}(\theta_i, m) \cos m\phi_j + \hat{V}(\theta_i, m) \sin m\phi_j, \quad j = 0, \ldots, 2L-1 \tag{B.17}$$

The overall computational scaling is also $O(L^3)$, which can be distributed to L processors as the two computational steps are fully independent. The use of parities in the construction of the associated Legendre polynomials reduces computation by a factor of 2. [310]