Summer 8-15-2019

A Physics-Based Intermolecular Potential for Biomolecular Simulation

Joshua Andrew Rackers
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS
Division of Biology and Biomedical Sciences
Computational and Molecular Biophysics

Dissertation Examination Committee:
Jay Ponder, Chair
Garland Marshall, Co-Chair
Gregory Bowman
Anders Carlsson
Li Yang

A Physics-Based Intermolecular Potential for Biomolecular Simulation

by
Joshua A. Rackers

A dissertation presented to
The Graduate School
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

August 2019
St. Louis, Missouri

Table of Contents

List of Figures

List of Tables

Acknowledgments

It is a privilege to explore the boundaries of science for a living and I need never forget those who make that work possible. I am eternally grateful for the people and institutions that have allowed me nearly six years of dedication to a highly speculative research endeavor. Without them, this dissertation would not exist.

To my wife, Erin: You are the single biggest reason for my success. You believed in me when I said I wanted to leave the teaching profession we both love. You believed in my research even when I didn't. Most of all, you loved me regardless of my achievements or failures. I love you and am forever grateful.

To my Dad: If anyone is responsible for my deep desire to know how things work, it's you. Some of my earliest memories are pulling your fluid mechanics books off the shelf and just savoring how cool I thought you were. You have always believed in my ability to understand anything and I cannot thank you enough for that support.

To my Mom: I owe you the world. It was not until becoming a teacher that I realized how important the simple knowledge that someone loves you is. You have been my source of that from the day I was born. Nothing good I do would be possible without that love.

To my daughter, Charlie: You are the single best thing that has ever happened to me. I hope one day you read this and know that I think you are capable of things one hundred times more important.

To my advisor, Jay: I will never understand why, but from the day I walked through the door of the lab, you trusted me. You treated me as an equal. You gave me space to explore. You've been an incredible mentor and better friend. I will miss our daily conversations dearly. Know that I owe an enormous portion of my present and future success to you.

I am grateful for the funds that have supported this research. The fact that this work is supported by all of our tax dollars is not lost on me. In particular, I acknowledge funding from the National Institutes of Health Grants R01 GM106137 and R01 GM114237, the National Science Foundation Molecular Sciences Software Institute Software Fellows program and the MilliporeSigma Fellowship in memory of Dr. Gerty T. Cori.

Josh Rackers

Washington University in St. Louis

August 2019

Dedicated to my parents.

ABSTRACT OF THE DISSERTATION

A Physics-Based Intermolecular Potential for Biomolecular Simulation

by

Joshua A. Rackers

Doctor of Philosophy in Biology and Biomedical Sciences

Computational and Molecular Biophysics

Washington University in St. Louis, 2019

Professor Jay Ponder, Chair

Professor Garland Marshall, Co-Chair

The grand challenge of biophysics is to use the fundamental laws of physics to predict how biological molecules will move and interact. The atomistic HIPPO (Hydrogen-like Intermolecular Polarizable Potential) force field is meant to address this challenge. It does so by breaking down the intermolecular potential energy function of biomolecular interactions into physically meaningful components (electrostatics, polarization, dispersion, and exchange-repulsion) and using this function to drive molecular dynamics simulations. This force field is able to achieve accuracy within 1 kcal/mol for each component when compared with *ab initio* Symmetry Adapted Perturbation Theory calculations. HIPPO is capable of this accuracy because it introduces a model electron density on every atom in the molecular system. Since the model is built on first-principles physics, it is transferable from small model systems to bulk phase. In the first test case, the HIPPO force field for water was able to reproduce the experimental density, heat of vaporization and dielectric constant to within 1%. Importantly, HIPPO has been shown to be only 10% more computationally expensive than the widely-used AMOEBA force field, meaning that more accurate simulations of larger biological molecules are well within reach

Chapter 1: Introduction

Our understanding of biology is profoundly incomplete. This seems a strange thing to say in an age when we've discovered so much. From the structure of DNA to the recent invention of a cure for Hepatitis C, the past half century has seen a boom in human understanding of biomedicine. Despite this, however, the amount that we *don't* understand about our biological world still far outweighs what we *do* understand. We don't understand how intrinsically disordered proteins contribute to health and disease. We only partially understand how the ribosome functions. We remain in the dark about how cells regulate traffic across their membranes. We are unable to predict the affinity of drug molecules for their biomolecular targets. The list is long and humbling. Central to nearly all of these yet unanswered questions, however, is one unifying theme: the behavior of molecules at the atomic scale. And it is this fact that should give us hope. If we can understand the behavior of molecules, we hold the keys to being able to answer some of the most important questions in biology.

Addressing the behavior of biological molecules is the aim of this dissertation. More specifically, the goal is to predict the behavior of and interactions between biomolecules using physics-based computer simulations. HIPPO (Hydrogen-like Intermolecular Polarizable Potential) is an atomistic model that I have developed in my graduate work specifically for the purpose of making these predictions. It is a set of physical models that determines exactly how a protein or piece of DNA moves in a computer simulation. Although it contains a multitude of approximations, every term of the HIPPO model is derived from first-principles physics. This is what makes it unique. There have been many physics-based models for computer simulations of biomolecules proposed in the past, but none as rooted in the principles of elementary quantum

mechanics. The promise of the HIPPO model is that by staying true to the fundamental laws driving the dynamics of every biomolecular system, it should yield predictive biomolecular simulations. In this sense, this dissertation should be considered both complete model and part of a collaborative work-in-progress effort. The work presented here shows the derivation and validation of the HIPPO model. It does not, unfortunately finish the task of constructing and validating this model on biomolecular systems like proteins. Assessing the ability of the model to predict complicated biomolecular phenomena is an ongoing and broad-ranging research effort.

What makes this effort important is that it strikes at the core of the central problem of biomolecular behavior. It answers the question: "What rules govern the interactions of molecules?" HIPPO is an attempt to define those rules and to the extent it does so successfully, has the power to help us understand some of the most important molecular phenomena in biology.

1.1 The Grand Challenge of Biophysics

To motivate the need for a physics-based model for biomolecular simulations, allow me to start with defining the "Grand Challenge of Biophysics". Imagine that we wanted to completely understand an arbitrary biological molecule. As an example, take the ribosome, pictured in figure 1.1. The structure and function of the ribosome is so important that the 2009 Nobel Prize in Chemistry was awarded in large part for determining an x-ray crystal structure of the molecule.[1] It is a complex molecular machine that we only partly understand despite decades of molecular biology and structural biology research. At a fundamental level, however, understanding the motions of the ribosome is conceptually simple.

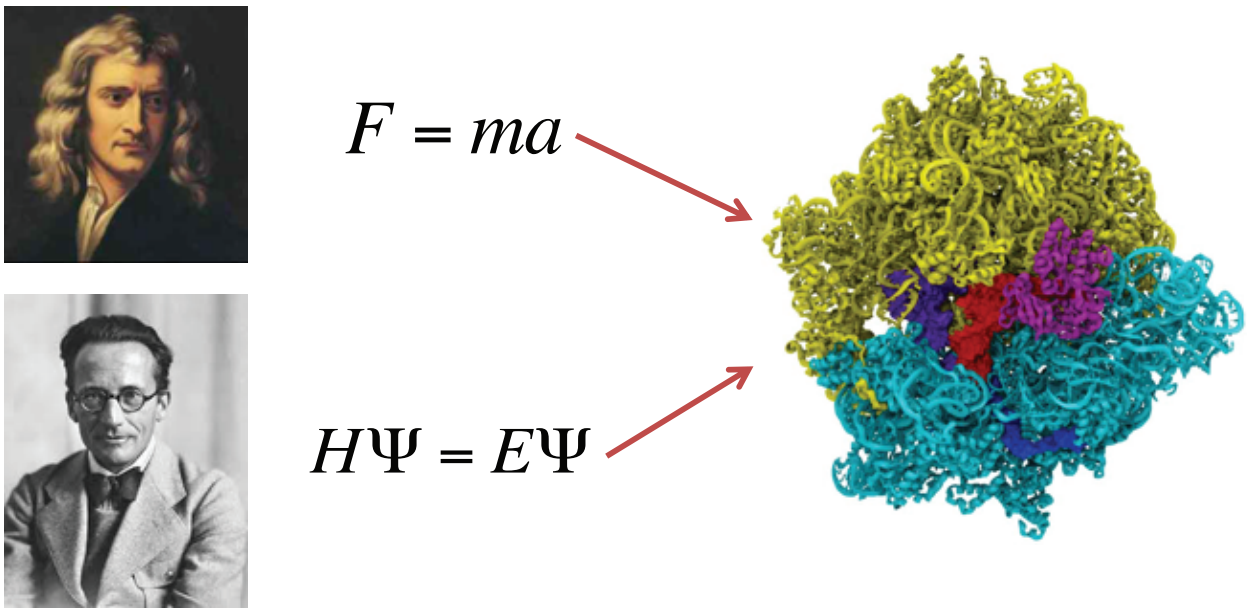


Figure 1.1 The Grand Challenge of Biophysics.

Pictured are the ribosome (right), and Sir Isaac Newton (top left) and Erwin Schrödinger (bottom left) along with the physical laws for which each are known. The challenge, which HIPPO addresses, is to apply these two fundamental laws in an approximate way that is accurate enough to give predictive simulations of biomolecules.

Like every molecule, the ribosome is made up of atoms, and the laws of physics have already given us all the rules that govern how atoms interact with each other. Namely, we have Schrödinger's equation which determines the electron density of each atom and Newton's 2nd Law which defines how each atom will move under a given force. In principle these two laws give us all the tools necessary to run a simulation of the ribosome (or any other biological molecule) that replicates reality. In practice, however, it is virtually impossible to solve Schrödinger's equation completely for systems bigger than a few dozen atoms (the ribosome has ~250,000), and current computer power limits the time scale for which such simulations can be performed. The Grand Challenge of Biophysics is to find a way to apply these two fundamental

laws in a way that is approximate, but accurate and computationally efficient enough to be predictive.

1.2 The Importance of the Potential Energy Surface

The most important, and consequently the most challenging part of addressing the Grand Challenge is approximating a solution to Schrödinger's equation. This is because the true solution to this equation, for a given molecule, defines the potential energy surface on which the molecule moves. In other words, the exact solution to Schrödinger's equation gives the exact energy (and thus force) of every atom in the molecular system. In order for a model to replicate reality, it must be an accurate approximation of this potential energy surface. For biomolecules there is a long history of using classical functions for this purpose. The current standard in the field is known as the Point Charge Force Field and it defines the potential energy of every atom in a biomolecular system according to a set of classical intramolecular and intermolecular energy terms. As illustrated in figure 1.2 the intramolecular terms are harmonic approximations to the energy of interaction between atoms that are connected by chemical bonds and the intermolecular terms consist of a simple fixed charge model that follows Coulomb's Law and a Lennard-Jones van der Waals model.

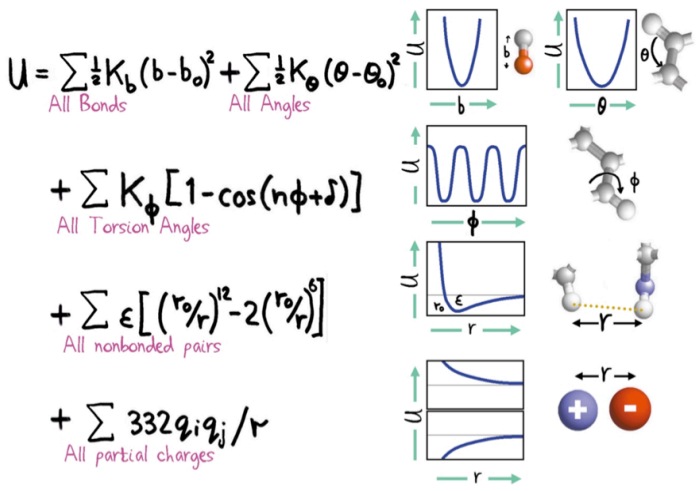


Figure 1.2 The Point Charge Force Field Energy Model.

This model is a classical, empirical potential energy surface with intramolecular (first three terms) and intermolecular (last two terms) terms. With small variations, this model is used in the vast majority of published molecular dynamics simulations of biomolecules. (credit: Michael Levitt)

Despite (or perhaps because of) its simplicity, the Point Charge Force Field has long been the standard for biomolecular simulations. In fact, this functional form, which was used on the very first published simulations of a protein in 1977, remains the most popular choice of model for biomolecular simulation today.[2]

The importance of the accuracy of an approximate potential energy surface is hard to understate. At a conceptual level, if the forces generated by the model do not match reality, then the motion of the atoms in the simulation will likewise be in error. Multiply these errors by the thousands of atoms in a typical protein and the result is simulations that give an incorrect picture of molecular motion and interactions. To make the level of accuracy needed in simulations concrete, a simple example application is helpful. Take the case of using simulation to predict the affinity of a drug for a target protein.

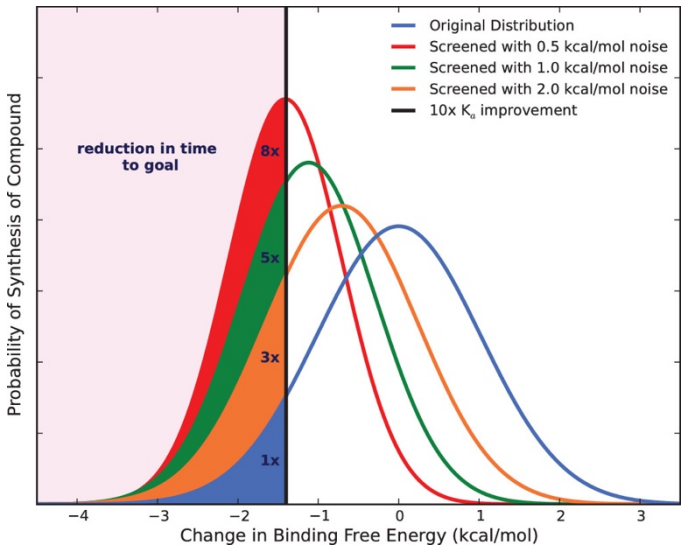


Figure 1.3 Importance of Free Energy Estimate in Drug Design.

Given a particular lead compound, the blue curve shows the probability that an experienced medicinal chemist will synthesize a new compound with tighter (more negative) binding free energy. The orange, green and red curves indicate the probabilities if those compounds are filtered using simulations that have an accuracy of 2.0 kcal/mol, 1.0 kcal/mol and 0.5 kcal/mol respectively. The shaded region indicates compound with a factor of 10 tighter binding than the lead molecule. One can see that a model capable of even 1.0 kcal/mol accuracy can yield five times the number of tighter binding compounds. Reproduced from reference 3.

Figure 1.3 shows the importance of accuracy in predicting the energy of the interaction between drug and protein. A model potential energy surface that is capable of predicting the binding energy to within 1 kcal/mol error can increase the number of potential drug candidates by 5x.[3] Put another way, Because of the relation $\Delta G = -RT \log(K_D)$, at room temperature every order of magnitude in the binding affinity translates into 1.36 kcal/mol in the free energy of binding. This is of great practical importance since a factor of 10 variance in a drug's binding affinity can be the difference between a medicine that hits a specific target *vs*. one that binds non-specifically. These and other considerations lead to the goal of "chemical accuracy": potential energies that

are accurate to within 1 kcal/mol. This level of accuracy matters specifically to binding interactions, but it is also applicable to the veracity of biomolecular simulations, generally.

Unfortunately, in many cases the point charge force field model is not capable of the accuracy necessary to be predictive of biomolecular reality. One concrete example that makes this clear is a recent study examining the performance of current force fields for predicting the fold of the UUCG RNA tetraloop. This particular RNA structure has been extensively studied by NMR (Nuclear Magnetic Resonance) and is known to spend greater than 90% of its time in the conformation represented by cluster 5 in figure 1.4.[4]

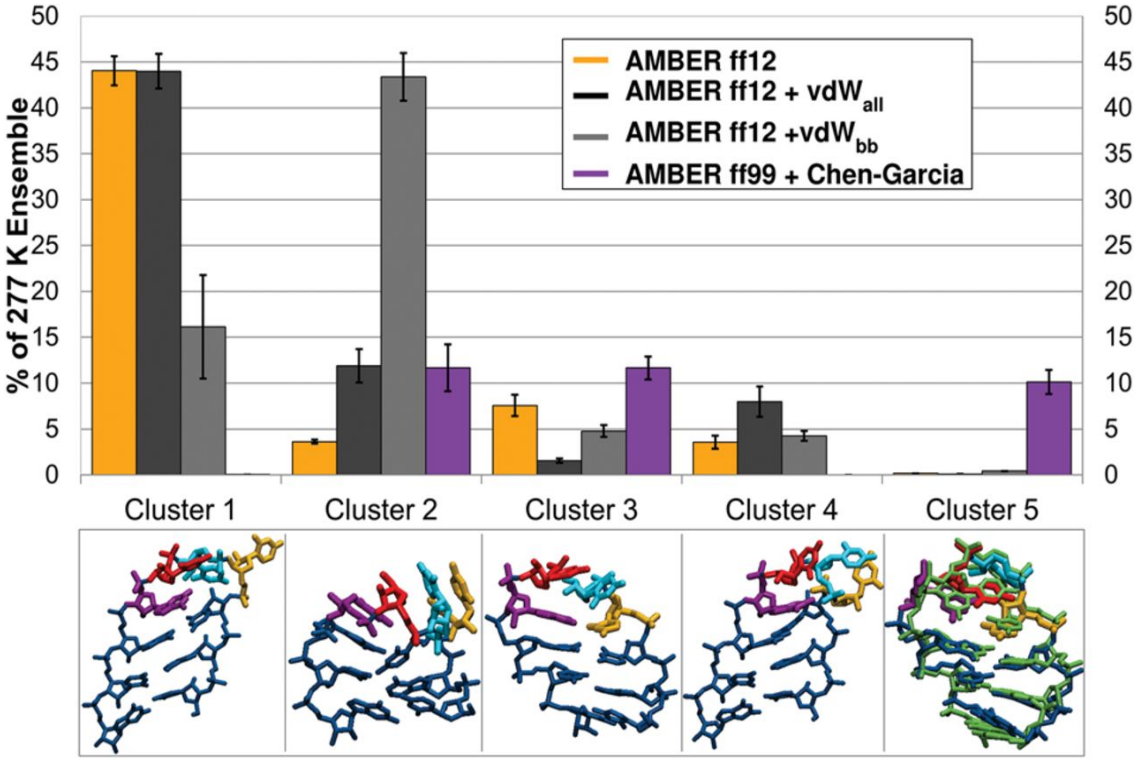


Figure 1.4 Structures of UUCG tetraloop as observed in simulations with various force fields.

No force field predicts the correct structure, cluster 5, shown overlaid in green. Reproduced from reference 4.

As shown in figure 1.4, however, no version of the Amber force field is able to correctly predict the fold of the molecule. In fact, only the Amber ff99 + Chen-Garcia model is able to predict that the sequence will spend any amount of its time (~10%) in the known structure. This across-the-

board failure occurs despite the fact that the tested force fields were specifically parameterized for RNA. This points to a serious problem with the point charge force field potential energy surface. Unfortunately, not all problems can be fixed with new parameterizations of the existing functional form. In many cases the problem with the potential energy surface resides in the functional form itself.

1.3 Symmetry Adapted Perturbation Theory

This and other failures of Point Charge Force Fields, of course, beg the question: If the Point Charge Force Field model is not a sufficient functional form to generate an accurate potential energy surface, what is? Because what we need is a classical functional form (solving the equations of quantum mechanics directly is too computationally expensive), there is no exactly correct or unique answer to this question. The fundamental issue is that classical models must split the total energy of interaction into components (e.g. electrostatics and van der Waals in standard force fields), but these quantities are not quantum mechanical (QM) observables. This means that a particular energy component cannot be measured experimentally. It does not mean, however, that such partitions are not possible mathematically. There have been many *ab initio* Energy Decomposition Analysis (EDA) schemes proposed and each partitions the total QM interaction energy into physically meaningful components. They include Kitaura–Morokuma (KM) EDA, Reduced Variational Space (RVS) EDA, Absolutely Localized Molecular Orbital (ALMO) EDA and others. For a useful review of EDA methods, see reference 5. The most powerful for the purposes of determining a force field functional form, however, is Symmetry Adapted Perturbation Theory (SAPT). SAPT uses perturbation theory to decompose the total *ab initio* intermolecular energy into electrostatics, induction, dispersion and exchange-repulsion components.[6-8] Because each of these components, as I will show through the course of

this dissertation, has a natural, classically derived analog, I choose SAPT as the *ab initio* method against which to develop and parameterize the HIPPO model. Using SAPT in this way gives us blueprint for a first principles physics-based functional form to replace the empirical Point Charge function.

An exhaustive review of the SAPT method is beyond the scope of this dissertation. A comprehensive explanation and derivation can be found in reference 6, and a more accessible review article in reference 8. It is instructive, however, for the purposes of understanding the derivation of the HIPPO potential energy model to lay out a brief explanation of each term in the SAPT perturbation theory expansion. Since SAPT is built off of Rayleigh-Schrödinger (RS) Perturbation Theory, I will start with overview of the RS method and then proceed to each of the SAPT energy terms.

1.3.1 Electrostatics

SAPT is fundamentally a perturbation theory method. Generally, the idea of perturbation theory methods is to find the solution to a complex problem by starting from the solution to a nearby simple problem and then perturbing the simple solution to fit the complex one. In the case of SAPT for intermolecular interactions, where we are attempting to find the full intermolecular interaction energy between two molecules, the simple, or unperturbed, problem is monomer wavefunctions. To get from isolated monomer wavefunctions to the full, interacting solution, SAPT starts with RS Perturbation Theory.

In RS Perturbation Theory for intermolecular interactions the goal is to find the solution to the equation,

$$H\Psi = E\Psi \tag{1.1}$$

where Ψ is the wavefunction of the dimer system, H is the Hamiltonian of the dimer system, and E is the energy of the interaction. Since H is complicated for a molecular dimer, we can define it in terms of perturbation theory as,

$$H = H_0 + \lambda V \tag{1.2}$$

where H_0 is the unperturbed Hamiltonian, V is the perturbation and λ is a parameter, between 0 and 1, scaling the magnitude of perturbation. In this case H_0 is taken to be

$$H_0 = H_A + H_B \tag{1.3}$$

where H_A and H_B are the unperturbed Hamiltonians of monomers A and B, respectively. This means that for $\lambda = 0$, $\Psi_0 = \Psi_A\Psi_B$ and $E_0 = E_A + E_B$. In other words, the zeroth order of RS Perturbation theory is the sum of the energies of the two isolated monomers. What we are interested in, however, is the intermolecular energy represented by the remaining orders of the perturbation theory expansion when $\lambda > 0$. In this case we can write the energy and wavefunction as power series expansions in λ,

$$\begin{aligned} \Psi &= \Psi_0 + \lambda\Psi_1 + \lambda^2\Psi_2 + \cdots \\ E &= E_0 + \lambda E_1 + \lambda^2 E_2 + \cdots \end{aligned} \tag{1.4}$$

where the infinite order expansion in energy with terms E_1 through E_n represents the exact intermolecular energy. In practice only the first few terms in the expansion are needed for a very accurate approximation. Unfortunately, the total wavefunction defined by equation 1.4 does not follow the antisymmetry requirement imposed by the Pauli Exclusion Rule. This is the root of the need for Symmetry Adapted Perturbation Theory over canonical RS perturbation theory. The correction that SAPT adds will be explained fully in section 1.3.4. However, the first few terms of the RS expansion can tell us a great deal about the physical nature of intermolecular

interactions. To derive these terms, we combine together terms with like powers of λ (in Dirac notation) to give,

$$\begin{aligned} E_0^{RS} &= \langle \Psi_0 | H_0 | \Psi_0 \rangle \\ E_1^{RS} &= \langle \Psi_0 | V | \Psi_0 \rangle \\ E_2^{RS} &= \langle \Psi_0 | V | \Psi_1 \rangle \\ &\vdots \\ E_n^{RS} &= \langle \Psi_0 | V | \Psi_{n-1} \rangle \end{aligned} \quad , \tag{1.5}$$

the energies of each order according to the corresponding n-1 order wavefunction.

The first and most important term in the RS expansion, E_1^{RS}, represents the electrostatic component of the intermolecular interaction. This is apparent when we consider the explicit physical form of the intermolecular interaction operator, V, in equation 1.2. This is simply the application of Coulombs law, or the operator 1/r. Inserting this into equation 1.5 gives,

$$E_1^{RS} = \langle \Psi_A \Psi_B | V | \Psi_A \Psi_B \rangle = \langle \Psi_A \Psi_A | V | \Psi_B \Psi_B \rangle \tag{1.6}$$

or in integral form,

$$E_1^{RS} = \iint \rho_A \frac{1}{r} \rho_B dv^2 \, . \tag{1.7}$$

This is the statement of Coulomb's Law between two charge densities, namely the unperturbed electronic charge densities of molecules A and B. SAPT naturally terms this the electrostatic energy. It is the portion of the interaction energy due to the Coulomb interaction between the two charge densities before they deform in response to each other.

1.3.2 Induction

Of course, in reality the electronic charge densities of interacting molecules deform in response to each other. This is where the 2nd order and higher of RS Perturbation Theory comes into play. For intermolecular (non-bonding) interactions, the 2nd order term all that it is needed for a nearly-complete picture of the interaction. 3rd order and higher SAPT terms do exist, but

their classical interpretations become much more complex and their contributions to the energy vanishingly small. For these reasons, HIPPO uses only SAPT terms through 2nd order. The full second order energy is defined as,

$$E_2^{RS} = \langle \Psi_0 | V | \Psi_1 \rangle, \tag{1.8}$$

in terms of the 1st order correction to the wavefunction, Ψ_1. This first order correction can be split into two parts. The first, when excitations from the ground state wavefunction Ψ_0 are localized exclusively to either monomer A or B. The contributions from these excitations, termed $\Psi_A^{exc}\Psi_B$ and $\Psi_A\Psi_B^{exc}$ form the SAPT induction energy. The remaining contributions, where excitations occur in concert on both monomers are termed dispersion and will be discussed in the next section. The induction component can be further subdivided into contributions from $\Psi_A^{exc}\Psi_B$ and $\Psi_A\Psi_B^{exc}$ respectively,

$$E_2^{RS}(ind) = E_2^{RS}(ind, B \rightarrow A) + E_2^{RS}(ind, A \rightarrow B). \tag{1.9}$$

The induction energy of the deformation of A in response to B, the first term in equation 1.9, can be written as:

$$E_2^{RS}(ind, B \rightarrow A) = \langle \Psi_A \Psi_B | V | \Psi_A^{exc} \Psi_B \rangle = \langle \Psi_A | \omega | \Psi_A^{exc} \rangle, \tag{1.10}$$

where,

$$\omega = \int \frac{1}{r} \rho_B dv\,. \tag{1.11}$$

This term, ω, represents the electrostatic potential of the unperturbed electron density of monomer B acting on monomer A. The energy associated with the deformation this causes is defined by equation 1.10. The same is true, swapping symbols, for the electrostatic potential of A acting on monomer B.

The sum of these two terms gives the SAPT definition of the induction energy. This is a natural definition because it matches our classical understanding of induction. When an electric field is applied to charge density, that density responds in a predictable way. This is the idea that underlies the concept of polarizability, which gives a description of how easily a given density will deform. In the case of molecules, rather than an external electric field, the field applied to A or B is coming from the other monomer. It is natural to call this component of the intermolecular interaction the induction energy.

1.3.3 Dispersion

The induction component of the 2nd order RS Perturbation theory expansion only covers part of the full 2nd order energy. In the simplest terms, dispersion is defined as what is left over after the induction energy is calculated,

$$E_2^{RS}(disp) = E_2^{RS} - E_2^{RS}(ind). \tag{1.12}$$

This definition, however, does not give us any physical meaning behind the term. To extract physical meaning, we can derive the "left over" component. As described in the previous section, this is the contribution to the energy arising from the component of Ψ_1 that involve excitations on both monomers, $\Psi_A^{exc}\Psi_B^{exc}$. This component of the RS 2nd order energy,

$$E_2^{RS}(disp) = \langle \Psi_A \Psi_B | V | \Psi_A^{exc} \Psi_B^{exc} \rangle, \tag{1.13}$$

cannot be decomposed into any terms involving simple, unperturbed monomer densities. The form of equation 1.13 does, however, give us an understanding of this component. This energy is coming from the correlation of instantaneous changes in the wavefunctions (and thus densities) of monomers A and B. This is the classical definition of the dispersion energy. In the classical Drude oscillator model of London dispersion (see section 4.2.1) the dispersion energy is the

energy due to the interaction of instantaneous fluctuations in electron density. SAPT follows this rationale and names the "left over" part of the 2nd order RS energy dispersion as well.

1.3.4 Exchange-Repulsion

The electrostatics, induction, and dispersion components of the SAPT decomposition of intermolecular interaction energies are derived from straightforward Rayleigh-Schrödinger Perturbation Theory. Where SAPT differs, and from whence it draws its name, is its treatment of exchange-repulsion. As stated above, the problem with RS Perturbation Theory is that it does not yield a final wavefunction (equation 1.4) that is antisymmetric. This cannot be correct, as the Pauli Exclusion Principle specifically demands that all valid electronic wavefunctions be antisymmetric. To remedy this problem, SAPT introduces an operator call an antisymmetrizer, $\mathcal{A}$, which appropriately permutes all pairs of electron labels to yield an antisymmetric wavefunction. Operating on the zeroth order H_2 wavefunction, for instance, the antisymmetrizer gives,

$$\mathcal{A}[\Psi_A(1)\Psi_B(2)] = [\Psi_A(1)\Psi_B(2) - \Psi_A(2)\Psi_B(1)]. \tag{1.14}$$

In the more general case of arbitrary intermolecular interactions, SAPT applies the antisymmetrizer to each order wavefunction from RS Perturbation Theory. This gives,

$$\Psi_n^{SAPT} = \mathcal{A}\Psi_n^{RS}, \tag{1.15}$$

the symmetry adapted, corrected wavefunction for each order. Each component of the RS expansion is then recalculated with these symmetry adapted wavefunctions and the difference between these two energies defines the SAPT exchange-repulsion energy:

$$E_{exch} = \sum_{i=0}^{2} E_i^{SAPT} - E_i^{RS}\,. \tag{1.16}$$

In this way the total energy from SAPT is guaranteed to be the energy associated with an antisymmetric total wavefunction, but we still retain the physical motivation behind the electrostatics, induction and dispersion terms from RS Perturbation Theory. The physical intuition behind this energy which SAPT calls "exchange-repulsion" (also referred to as "Pauli repulsion") is less obvious than the other terms in SAPT. A full interpretation is given in Chapter 5 of this dissertation, but in short, the energy due to forcing antisymmetrization of the wavefunction arises from the overlap between the noninteracting monomer densities. The overlap in these unperturbed densities violates the Pauli Exclusion Rule and therefore, relative to this reference, the density in the overlap region is reduced to accommodate the rule. This reduction in density in the internuclear region de-screens the nuclei, resulting in internuclear repulsion. This affect is seen clearly if one plots the densities corresponding to Ψ and $\mathcal{A}\Psi$. An example of this for the helium dimer is shown in section 5.2.1.

1.3.5 Why SAPT is a Natural Fit

There are a large number of legitimate *ab initio* energy decomposition analysis methods available. A natural question is, why use SAPT? Although there are many similarities amongst the various EDAs, the structure of SAPT makes it particularly well-suited for the purpose of constructing a classical force field. Specifically, it is the use of perturbation theory and how it is applied in SAPT that makes it a natural fit for building the HIPPO force field.

As described above, the reference, or unperturbed, state of the SAPT calculation is the two non-interacting monomer wavefunctions. Although this reference state is technically not a valid system wavefunction, it does correspond directly to the strategy used to build the electrostatics portion of the HIPPO force field. In HIPPO the multipole moments of each atom (charge, dipole and quadrupole) are derived directly from the isolated, gas-phase monomer

electron density. In other words, HIPPO and SAPT start from the same reference state. The other, higher-order terms are in direct correspondence as well. The polarization model of HIPPO matches the definition from SAPT of electron density deformation in the presence of an external field. The dispersion and Pauli repulsion components likewise correspond directly between SAPT and HIPPO. The depth of these associations will be explored in Chapters 2-5 for each component, but they stand on the same conceptual foundation. This matters to having an interpretable force field. Because HIPPO is a natural fit with SAPT, it allows us to describe in quantum mechanical language what the force field is approximating classically.

1.3.6 A SAPT Example

Because SAPT is both a highly accurate *ab initio* method and because it aligns closely with the design principles of classical force field models, we can use it evaluate model performance. The body of this dissertation with be full of examples comparing SAPT components for molecular interactions to classical models (HIPPO and otherwise), but one example at this point will clarify how SAPT can be used.

A good example of the utility of SAPT in evaluating force fields is the RNA tetraloop mentioned in Section 1.2. A primary driver of nucleic acid structure is the base stacking interaction. Thus, work by Parker and Sherrill in reference 9 set out to evaluate how well standard force fields performed against SAPT for this particular interaction. Their findings are illustrated in figure 1.5.

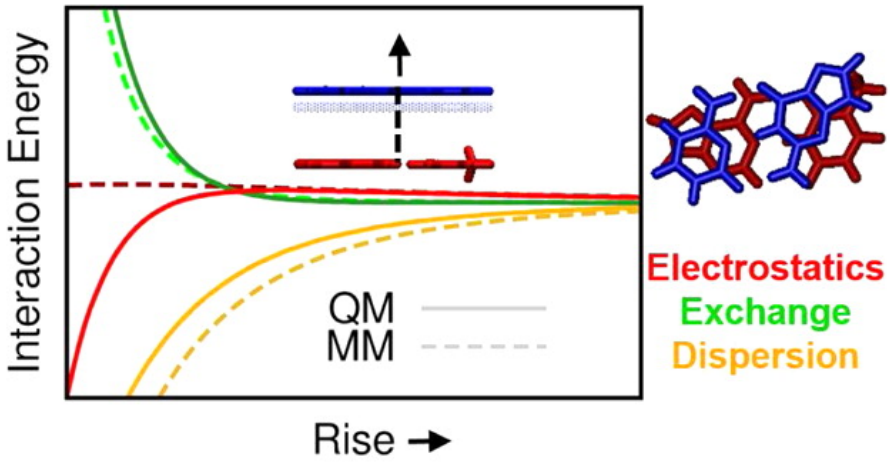


Figure 1.5 SAPT vs. Point Charge Force Field Energy Components for RNA Base Stacking.

The figure is an illustration of the general trends observed in reference 9. "QM" represents SAPT. The SAPT Induction component is omitted since there is no corresponding component in point charge force fields.

Clearly, there are some major issues with the components of the standard force field models. First, there is an entire component missing. Point charge force fields do not include polarization, so they miss this component of the SAPT EDA. Second, although the components for electrostatics, exchange, and dispersion match well at long-range, each diverges at short-range. This is particularly true for electrostatics where the divergence is sharp. As I will describe in the body of the dissertation, these problems are not unique to RNA. They occur across the space of chemical interactions and they are a specific result of point charge force fields neglecting the molecular charge density.

Point charge force fields typically rely on cancellation of errors in order to cover up these two issues, resulting in two further problems. First, the cancellation of errors that works for one system may not necessarily work for another system. Second, and more importantly in the case of RNA, the functional form of the point charge force field is not flexible enough to cancel errors in all distance ranges simultaneously. This is what was observed by Parker and Sherrill. They

found that without a charge density model included, the point charge force field model could not be reparametrized to obtain accurate total energies across a range of intermolecular interaction distances and conformations.

This analysis of the atomic-level interactions that drive RNA structure is enlightening, particularly in light of the failure of point charge models in reproducing RNA structure shown in figure 1.4. One cannot conclusively say that the inability to model base stacking correctly is the cause of the structural errors seen in figure 1.4, but the suggestion is strong. The SAPT analysis hints that the pathway to more accurate simulations of RNA and biomolecules in general does not lie in reparametrizing existing models; it lies in building new models that explicitly include the missing physics that is causing the standard models to fail in the first place.

1.3.7 The S101x7 Database

In order to use SAPT to build a model that approximates the EDA energy components, the first step was to construct a database of reference data. Since the goal of the HIPPO force field is to simulate biomolecules, a database was assembled that included a wide variety of intermolecular interactions that are prevalent in biomolecular systems. A total of 101 molecular dimers where chosen, as enumerated in figure 1.6.[10]

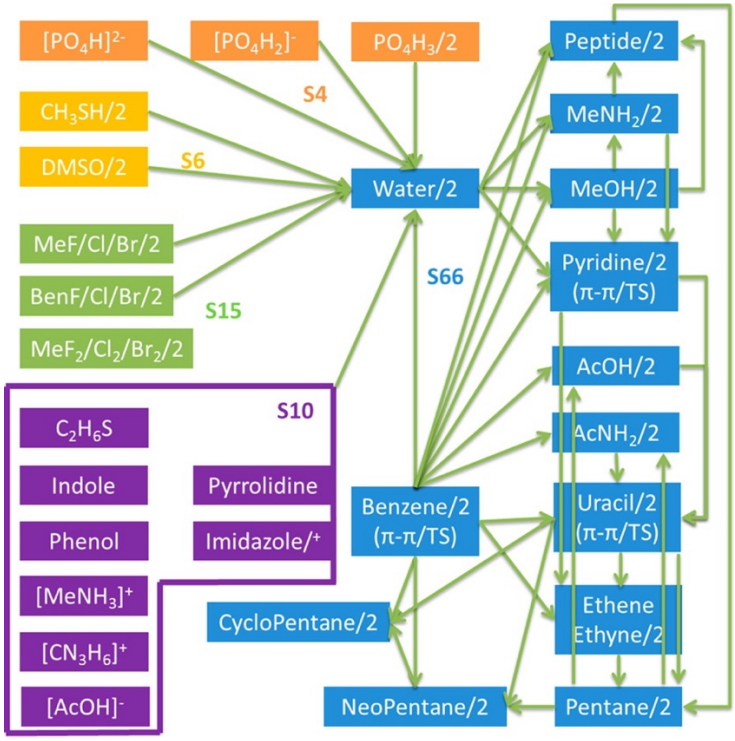


Figure 1.6 Dimers in the S101 Database.

Arrows connecting molecules indicate a dimer. The "/2" designation indicates a homodimer. The "/+(-)" designation indicates that both a neutral and charged species are included. Reproduced from reference 10.

The database is derived from the widely-used S22 and S66 databases of Hobza and co-workers.[11,12] It includes additions of halogenated systems, phosphates, sulfur-containing compounds and amino acid side-chain analogs.

In addition to including the equilibrium structure for each dimer, the database was expanded to incorporate six additional points along each dimer's dissociation curve. This yields the S101x7 database, which has been used extensively in the development of the HIPPO. The points along the dissociation curve are 0.7, 0.8, 0.9, 0.95, 1.00, 1.05, and 1.10 times the equilibrium distance. These represent the range of intermolecular distances typically seen in condensed phase simulations. The monomer geometries are optimized at the equilibrium distance at the MP2/aug-cc-pVTZ level of theory, then held fixed for the other distances.

For each dimer structure, we performed SAPT2+ calculations using the Psi4 quantum mechanics software package.[13] SAPT2+, as illustrated in figure 1.7, includes electrostatics, induction, dispersion and exchange terms as well as some correction terms.[14]

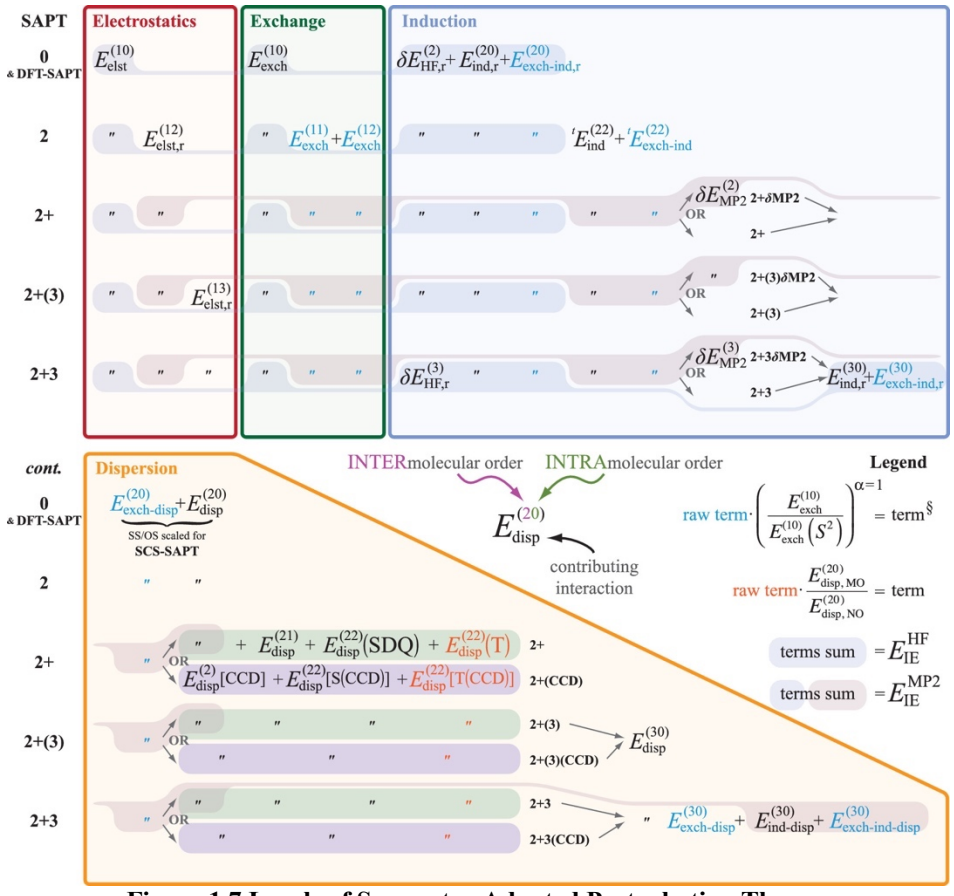


Figure 1.7 Levels of Symmetry Adapted Perturbation Theory.

Reproduced from reference 14.

We performed SAPT2+ calculations with two different basis sets: aug-cc-pVDZ and aug-cc-pVTZ. The former constitutes the "silver" standard according to the reference 14 and the latter was computed to estimate complete basis set (CBS) limits. The output of these calculations were an electrostatics, induction, dispersion and exchange data point for each of the 707 dimers in the S101x7 database.

1.4 A New Kind of Force Field

For more than three decades the standard of biomolecular simulation has been the point charge force field. This model has had many notable successes. Molecular dynamics simulations using point charge force fields have successfully folded proteins and reproduced enzyme-inhibitor binding interactions.[15,16] As illustrated with the RNA example above and as will be further demonstrated in the body of this dissertation, however, this standard model is missing some key physics. The hypothesis of this work is that rather than attempting to cover over these insufficiencies, a more productive approach is to derive a model that includes the most relevant and important physics from the start. It is this design strategy that has driven the development of the model presented in this work. The HIPPO model is a new class of force field. Although it is a classical potential energy function, every term is derived, in some fashion, from first-principles physics.

The HIPPO model is different from the point charge force field model in two subtle, but important ways. First, the HIPPO model is derived and parameterized to explicitly reproduce the *ab initio* energy components from Symmetry Adapted Perturbation Theory. This stands in stark contrast to the strategy of the standard force fields, which are parameterized empirically, based on condensed phase properties. Second, the HIPPO model abandons the atoms-as-points model of the standard force field and introduces a model electron density around every atom. As I will show, every intermolecular potential energy term is related in some way to the overlap of atomic charge densities. Including a model density on each atom allows derivation of the first-principles-based energy terms of the HIPPO model.

These two profound changes that make HIPPO a new class of force field yield a host of improvements over conventional models. HIPPO is able to reproduce each separate component of the intermolecular energy relative to SAPT within chemical accuracy, or ~1 kcal/mol.

Including a charge density model solves the longstanding "charge penetration problem" in molecular modeling (see Chapter 2). The resulting polarization model yields better molecular polarizabilities than the leading polarizable force fields (see Chapter 3). The first-principle derived dispersion model produces a damping function with true physical meaning (see Chapter 4). The exchange-repulsion model describes the anisotropy of halogen bonding with drug molecules more accurately than any alternative force field (see Chapter 5). And, when all these parts are added together, the whole model works naturally for simulating water and a host of other organic molecules (see Chapter 6). These successes, as I will show through the course of this dissertation are a direct consequence of the physics-first design of the HIPPO model. This is not to say that HIPPO is a purely "*ab initio*" model. The data clearly show that the parameters must be tuned to reproduce experimental reality in the same way biomolecular force fields have always done. The fact that the model is rooted in its derivation from first principles, however, means that it makes the stubborn optimization problem of empirical force fields tractable. The direct connection between HIPPO and SAPT acts as a strong set of guidelines for model development. We have developed water and small organic molecule HIPPO force fields, but work is also underway to build HIPPO models for full protein and nucleic acid simulations. The work presented in this dissertation lays the groundwork for HIPPO to produce a new class biomolecular simulation.

1.5 Structure of the Dissertation

The dissertation will be laid out in the following manner. This first chapter has served to give a motivation, background information and overview of the HIPPO model. Chapters 2-5 will explain in detail how each term of the function was derived and parameterized. These will go in order of their place in the perturbation theory expansion. Chapter 2 is devoted to electrostatics,

Chapter 3 to polarization (induction), Chapter 4 to dispersion and Chapter 5 to exchange-repulsion (Pauli repulsion). Each chapter will detail how the component was derived and how well it matches the SAPT data for that component. Chapters 2, 4, and 5 are taken directly from published works of which I am the first author. These chapters will contain an introduction to put the paper in context, the full body of the paper and a "Further Work" section that explains any parts that subsequently changed. Finally, Chapter 6 will tie the dissertation together by presenting simulation results of the full HIPPO model on water. Taken together, this should provide a full picture of where the model comes from, how accurate it is, and how we can expect it to perform in future applications.

1.5 References

1 Ban, N., Nissen, P., Hansen, J., Moore, P. B. & Steitz, T. A. The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. *Science* **289**, 905-920 (2000).

2 McCammon, J. A., Gelin, B. R. & Karplus, M. Dynamics of folded proteins. *Nature* **267**, 585 (1977).

3 Mobley, D. L. & Klimovich, P. V. Perspective: Alchemical free energy calculations for drug discovery. *The Journal of Chemical Physics* **137**, 230901 (2012).

4 Bergonzo, C., Henriksen, N. M., Roe, D. R. & Cheatham, T. E. Highly sampled tetranucleotide and tetraloop motifs enable evaluation of common RNA force fields. *RNA* **21**, 1578-1590 (2015).

5 Phipps, M. J., Fox, T., Tautermann, C. S. & Skylaris, C.-K. Energy decomposition analysis approaches and their evaluation on prototypical protein–drug interaction patterns. *Chemical Society Reviews* **44**, 3177-3211 (2015).

6 Jeziorski, B., Moszynski, R. & Szalewicz, K. Perturbation theory approach to intermolecular potential energy surfaces of van der Waals complexes. *Chemical Reviews* **94**, 1887-1930 (1994).

7 Rybak, S. a., Jeziorski, B. & Szalewicz, K. Many-body symmetry-adapted perturbation theory of intermolecular interactions. H2O and HF dimers. *The Journal of Chemical Physics* **95**, 6576-6601 (1991).

8 Szalewicz, K. Symmetry-adapted perturbation theory of intermolecular forces. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2**, 254-272 (2012).

9 Parker, T. M. & Sherrill, C. D. Assessment of Empirical Models versus High-Accuracy Ab Initio Methods for Nucleobase Stacking: Evaluating the Importance of Charge Penetration. *Journal of Chemical Theory and Computation* **11**, 4197-4204 (2015).

10 Wang, Q. *et al.* A General Model for Treating Short-Range Electrostatic Penetration in a Molecular Mechanics Force Field. *Journal of Chemical Theory and Computation* (2015).
11 Jurečka, P., Šponer, J., Černý, J. & Hobza, P. Benchmark database of accurate (MP2 and CCSD (T) complete basis set limit) interaction energies of small model complexes, DNA base pairs, and amino acid pairs. *Physical Chemistry Chemical Physics* **8**, 1985-1993 (2006).
12 Rezác, J., Riley, K. E. & Hobza, P. S66: A well-balanced database of benchmark interaction energies relevant to biomolecular structures. *Journal of Chemical Theory and Computation* **7**, 2427-2438 (2011).
13 Parrish, R. M. *et al.* Psi4 1.1: An Open-Source Electronic Structure Program Emphasizing Automation, Advanced Libraries, and Interoperability. *Journal of Chemical Theory and Computation* (2017).
14 Parker, T. M., Burns, L. A., Parrish, R. M., Ryno, A. G. & Sherrill, C. D. Levels of symmetry adapted perturbation theory (SAPT). I. Efficiency and performance for interaction energies. *Journal of Chemical Physics* **140**, 094106 (2014).
15 Lindorff-Larsen, K., Piana, S., Dror, R. O. & Shaw, D. E. How fast-folding proteins fold. *Science* **334**, 517-520 (2011).
16 Buch, I., Giorgino, T. & De Fabritiis, G. Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations. *Proceedings of the National Academy of Sciences* **108**, 10184-10189 (2011).

Chapter 2: Electrostatics

In many ways the electrostatics portion of the HIPPO model is the most important component of the whole. There are two reasons for this primacy over the other components. The first is conceptual. According to the perturbation theory strategy of Symmetry Adapted Perturbation Theory, electrostatics is the first order contribution to the total intermolecular interaction energy. Because of this, all of the other terms are built on top of it, making it essential that the model be both simple and accurate. The second reason is practical. Decades of biomedical research has shown the importance of charged interactions in biomolecular systems. This ranges from partial charges interacting between protein sidechains to ions neutralizing charged nucleic acid backbones. For these reasons, electrostatics was the first portion of the HIPPO model that I addressed. The work in the following published paper lays out the foundational idea of the HIPPO model: the atomic charge density.

At the time the following paper was published, I did not know that the charge density would end up being so integral to the overall model. The paper was specifically meant to address the narrower problem of "charge penetration" in molecular mechanics force fields. There had been some work on this in the literature, but nothing comprehensive for a biomolecular force field. This showed that within the context of the AMOEBA (Atomic Multipole Optimized Energetics for Biomolecular Applications) force field, a simple charge density model could solve the charge penetration problem by applying that charge density to all orders of the AMOEBA multipole expansion (charge, dipole, and quadrupole). As I will show in "Further Work" (Section 2.8), this model was changed slightly for the final HIPPO model. The conclusions of the paper remain, nonetheless, unchanged.

2.1 Introduction

A grand challenge of molecular mechanics (MM) force fields is modeling the physics of molecular interactions with an accuracy and efficiency that allows realistic, tractable simulations of large systems. The goal is not only to correctly capture the physics of molecular interactions, but also to be able to answer important practical questions posed by biology, materials science and a number of other fields. To do this, MM models make classical approximations to the 1st principles quantum mechanics driving the true dynamics of a molecular system. Typically, this is done via a set of classical harmonic potential terms describing the intramolecular interactions of bonded atoms in the system and a separate set of non-bonded terms to describe intermolecular interactions. In particular, the electrostatic nonbonded terms are especially important for accurately modeling both short and long range molecular interactions.[1]

The AMOEBA force field is unique in its treatment of these important intermolecular electrostatic interactions. Most MM force fields use point charges to approximate the charge distribution around atoms in a system and parameterize these point charges based on thermodynamic measurements. AMOEBA takes a more physically realistic approach. The AMOEBA model approximates the charge distribution around atoms as a point multipole expansion of the charge distribution obtained from *ab initio* quantum mechanics (QM) calculations.[2,3] Using a multipole expansion derived from *ab initio* QM calculations provides a much more accurate description of electrostatic interactions at medium-range (~2 to 4 times the vdW radius), and has been shown to yield satisfactory results for simulations of water, proteins, nucleic acids and small molecules.[1,2,4,5]

The multipole approximation of electrostatics, however, starts to break down at short-range. While the multipole expansion is rigorously correct for interactions of atoms at sufficient distance, it is no longer strictly valid once the electron clouds of interacting atoms start to overlap.

This phenomenon is known as charge penetration. Charge penetration is simply the change in the electrostatic interaction between two atoms due to their electron cloud overlap and the associated loss of nuclear screening. It is a simple accounting for the fact that atoms in a system are not points; they represent finite charge distributions. Accurately modeling electrostatics has been a priority with AMOEBA since its inception. The importance of these interactions was a key motivation for the original AMOEBA multipole model. Qualitatively, accounting for charge penetration is the logical next step in improving this model.

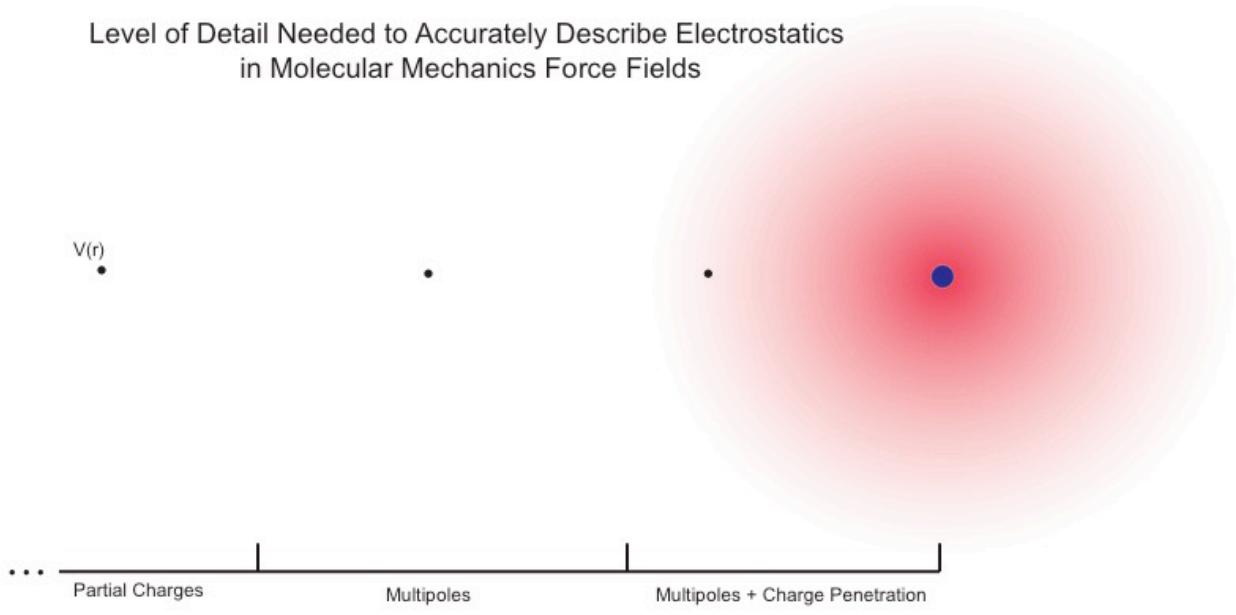


Figure 2.1 Electrostatic potential as a function of distance.

An increasing level of theory is needed as the radial distance from an atom of interest decreases.

As depicted in figure 2.1, the current model covers the accuracy of long- and medium-range electrostatic interactions. What is needed is a description of charge penetration to accurately model short-range interactions.

In addition to being physically relevant, charge penetration has been shown to be an important factor in many intermolecular interactions. A particularly instructive set of examples lies with what are commonly called "pi-pi" stacking interactions.[6] The benzene sandwich dimer, as illustrated in figure 2.2, should classically be considered electrostatically repulsive since like charges are lined up across from one another. High level *ab initio* quantum mechanical calculations, however, show the counterintuitive result that the benzene sandwich dimer is electrostatically attractive.[7] This is almost entirely due to charge penetration.

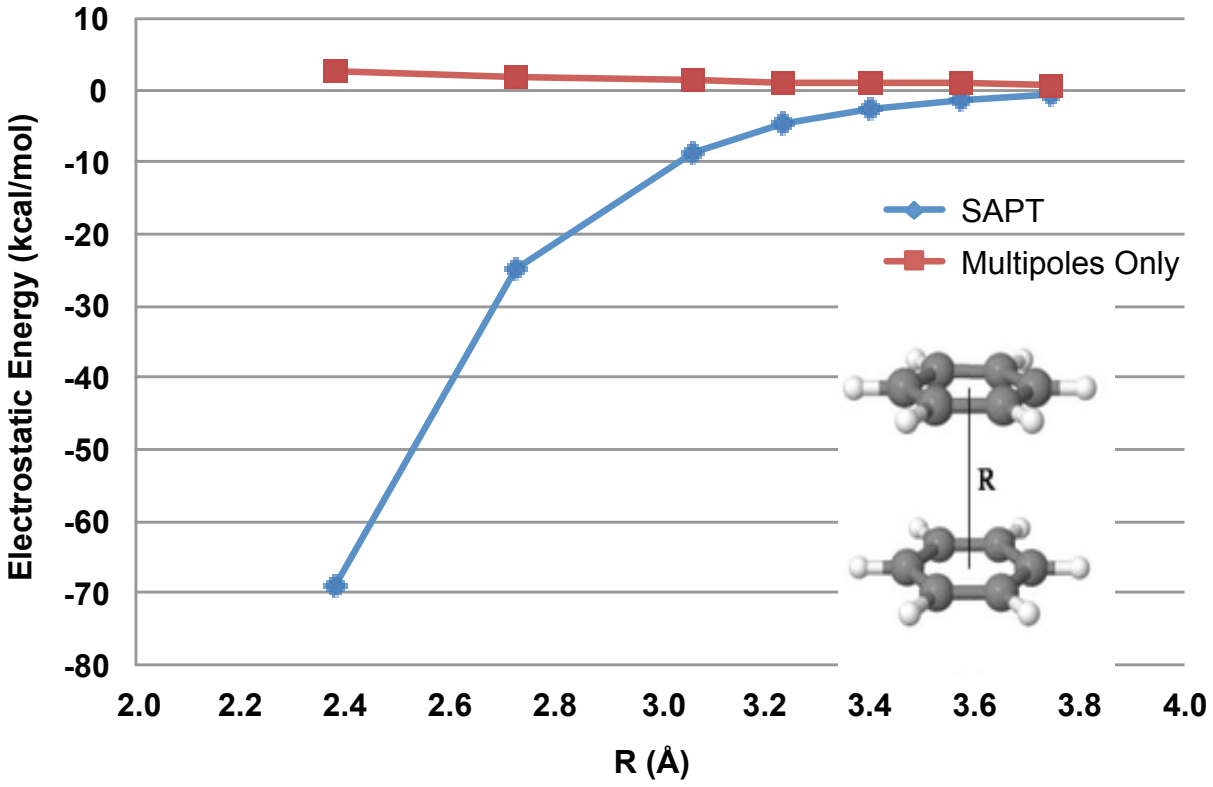


Figure 2.2 Electrostatic energy of the benzene sandwich dimer.

AMOEBA overestimates the electrostatic energy of the interaction compared with the benchmark QM calculations. The error gets progressively worse at short-range.

Figure 2.2 shows that the overlap of electron clouds causes the electrostatic energy of the interaction to become more negative as the two monomers get closer together. This same phenomenon is observed with stacking interactions between nucleobases. Parker and Sherrill have recently shown that without charge penetration, it is difficult, if not impossible to accurately capture the electrostatics of interacting nucleobases.[8] These considerations show that if AMOEBA is to be successful in accurately modeling biologically relevant interactions such as nucleic acid folding or ligand binding, we must account for the short-range electrostatics of charge penetration.

A number of studies have suggested functions for incorporating charge penetration into existing molecular mechanics force fields.[9-20] The derivation of most of these functions has followed the same basic strategy. The electrostatic description of each atom in the system is split into two parts. The first is the core charge (often, but not necessarily simply the nuclear charge), treated as a point and second a smeared electron cloud charge representing the remaining charge of the atom. This splits what was a single interaction into four interactions, as illustrated in figure 2.3.

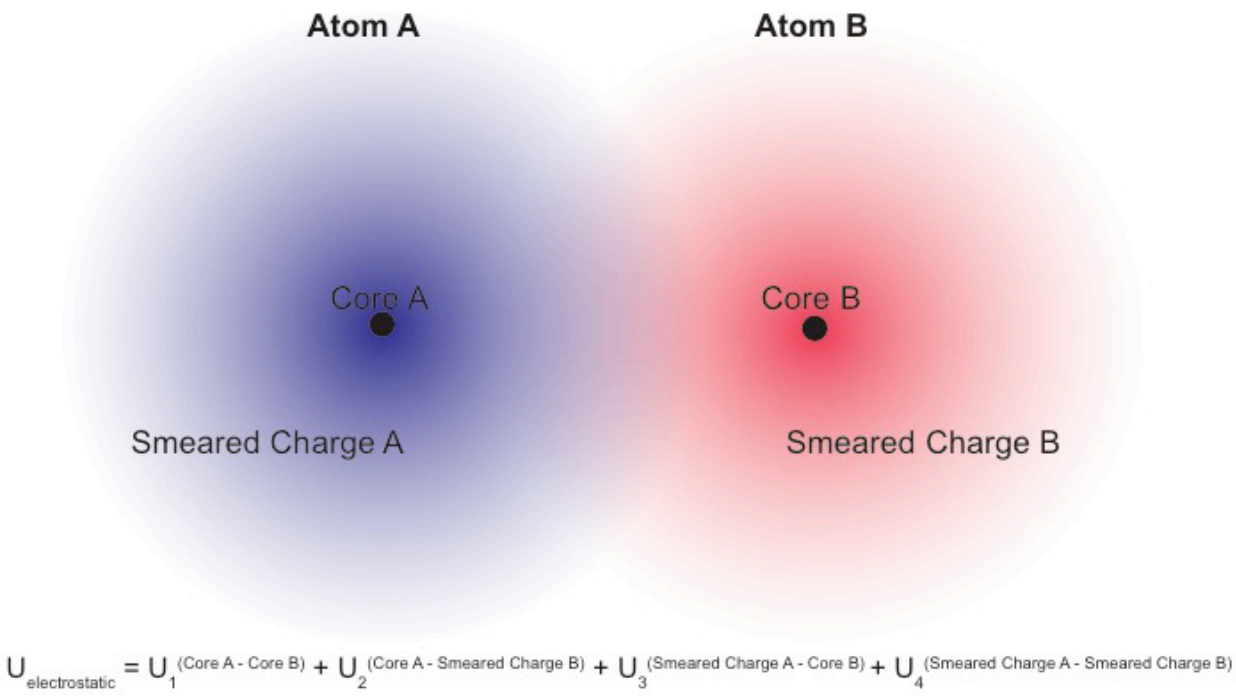


Figure 2.3 Electrostatic energy of charge penetration-corrected, smeared-charge atomic interactions. The total electrostatic energy is split into four parts. The first term is the energy of the core-core, point-point interaction. The second and third terms are the energies of each core in the electrostatic potential of the opposing smeared charge. The fourth term is the energy of the overlap between smeared charge distributions.

The functions listed in table 2.1 are four methods suggested for how best to handle this four-part interaction between atoms.

Model	Core A – Core B	Core A – Smeared Charge B	Smeared Charge A – Core B	Smeared Charge A – Smeared Charge B
Engels; Cisneros	$\frac{Z_A Z_B}{r}$	$\int_{-\infty}^{\infty} \frac{Z_A \rho_B(r_2)}{\lvert R_A - r_2 \rvert} dr_2$	$\int_{-\infty}^{\infty} \frac{Z_B \rho_A(r_1)}{\lvert R_B - r_1 \rvert} dr_1$	$\iint_{-\infty}^{\infty} \frac{\rho_A(r_1)\rho_B(r_2)}{\lvert r_1 - r_2 \rvert} dr_1 dr_2$
Gordon	$\frac{Z_A Z_B}{r}$	$\frac{Z_A q_B}{r} f_{damp}(r)$	$\frac{Z_B q_A}{r} f_{damp}(r)$	$\frac{q_A q_B}{r} f_{damp}^{overlap}(r)$
Piquemal	$\frac{V_A V_B}{r}$	$\frac{V_A(c_B - V_B)}{r} f_{damp}(r)$	$\frac{V_B(c_A - V_A)}{r} f_{damp}(r)$	$\frac{(c_A - V_A)(c_B - V_B)}{r} f_{damp}^{overlap}(r)$
Truhlar	$\frac{(c_A + n_A)(c_B + n_B)}{r}$	$\frac{(c_A + n_A) n_B}{r} f_{damp}(r)$	$\frac{(c_B + n_B) n_A}{r} f_{damp}(r)$	$\frac{n_A n_B}{r} f_{damp}^{overlap}(r)$

Table 2.1 Proposed methods for incorporating charge penetration into molecular mechanics electrostatic energy.

For consistency, Z is the nuclear charge, ρ is the total charge density of the electrons, q is the total charge of the electron cloud, V is the number of valence electrons, c is the partial charge, n is the number of "screening electrons", and r is the internuclear distance. In the first row, the charge density is either a promolecular charge density (Engels) or a density from hermite gaussians in the GEM model (Cisneros).

Tafipolsky and Engels took a more direct approach and calculated a numerical integral between spherical pro-molecule charge densities.[17] This is similar in spirit to the approach of the GEM (Gaussian Electrostatic Model) force field, where hermite gaussians are used to reproduce the *ab initio* electron density.[9,21,22] While being physically straightforward, these methods currently lack the efficiency needed for simulating large systems. The other three methods use damping functions

to approximate how the electrostatic potential of an atom changes in its electron cloud and use those damping functions to approximate the value of the overlap integral for U_4.

In a previous proof-of-principle study, we implemented the form of Piquemal and co-workers in the AMOEBA force field.[23] The study showed that accounting for charge penetration can start to recover the true nature of short-range electrostatic interactions between molecules. A follow-up study extended the model for use with smooth particle mesh Ewald.[24] In the present work we seek to develop a comprehensive model based on the previous work that best captures the physics of electrostatic intermolecular interactions and the aims of the AMOEBA force field. Given the potential improvement our previous work has shown possible in such a model, the question becomes: what features would we like the AMOEBA charge penetration model to have? In the work presented here we aim to implement a charge penetration function that best meets the following criteria:

1. The model should be physically derived.
2. The model should be computationally efficient to compute.
3. The model should be numerically stable.
4. The model should accurately reproduce *ab initio* QM measurements for relevant molecular interactions.
5. The model should be consistent with the AMOEBA multipole model.

In section 2, we present the physical derivation of the models that were considered and derive corresponding damping terms for higher-order multipoles. In section 3, the scheme for parameterizing the models is presented. Section 4 lays out results comparing the performance of the models. Section 5 shows validation that the charge penetration model is capturing physical reality. And lastly, section 6 draws our conclusions.

2.2 Theory

Stone illustrated the phenomenon of charge penetration with a simple example.[25] Consider the interaction of a proton with a hydrogen-like atom with nuclear charge Z. From quantum mechanics we know that the wave function of a hydrogen-like atom is

$$\psi(r) = \sqrt{\frac{Z^3}{\pi}} e^{-Zr}. \tag{2.1}$$

This gives us the electron density of the atom,

$$\rho(r) = -\frac{Z^3}{\pi} e^{-Zr}. \tag{2.2}$$

This tells us how dense the electron distribution of the atom is as a function of the radial distance (r) from its nucleus. To get the potential this density generates, we must apply Poisson's equation,

$$\nabla^2 V = \frac{-\rho}{\epsilon_0}, \tag{2.3}$$

where $\epsilon_0$is the permittivity of free space. Applying Eq. (2.3) to Eq. (2.2) we obtain

$$V(r) = -\frac{1}{r} + \left(Z + \frac{1}{r}\right) e^{-2Zr}, \tag{2.4}$$

the familiar potential due to the electron density of a hydrogen-like atom. At large distances from the atom, the first term in Eq. 2.4 dominates the second term due to the second's exponential decay and we have the classical point charge coulomb approximation of the potential. At closer distances, however, as shown in figure 2.4, the second term becomes non-negligible. This second term represents the charge penetration.

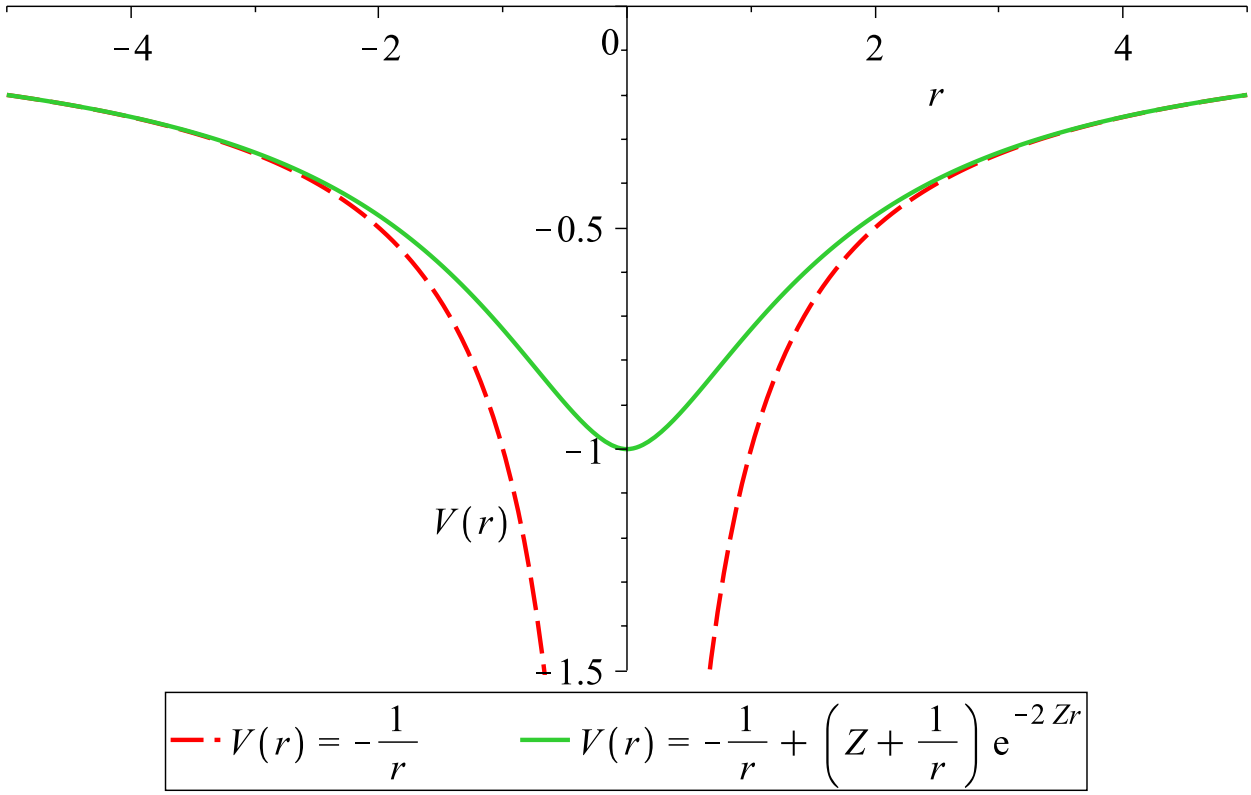


Figure 2.4 Classical coulomb potential *vs.* Hydrogen-like atom potential.

Plotted is the electrostatic potential of a point electron *vs*. the hydrogen-like electron (Z=2 to emphasize the distinction). The classical potential diverges from the hydrogen-like result at short-range.

We can exploit the fact that V(r) converges to -1/r at large distances and rewrite Eq. 2.4 as

$$V(r) = -\frac{1}{r}(1 - (1 + Zr)e^{-2Zr}) = -\frac{1}{r} \cdot f_{damp}(r) \tag{2.5}$$

where,

$$f_{damp}(r) = 1 - (1 + Zr)e^{-2Zr}. \tag{2.6}$$

The potential in this form is represented simply as the point charge coulomb potential multiplied by a damping function. This is convenient because the damping function has the following straightforward properties:

1. It approaches a value of one as r becomes large.
2. It approaches a value of zero as r approaches zero.
3. It is a direct multiplication of the classical point-charge coulomb potential.
4. It describes charge penetration as a deviation from the classical potential.

To this point there are no approximations made in our derivation. Crucially, however, most atoms in systems of interest for molecular simulation are not strictly hydrogen-like. This means that f_{damp} for non-hydrogen-like atoms is not exactly given by Eq. 2.6. The properties and form of Eq. 2.6 are instructive, however. To capture the physics more generally, we introduce a parameter, α, in place of the 2Z and remove the prefactor in front of the exponential to obtain

$$f_{damp}(r) = 1 - e^{-\alpha r}. \tag{2.7}$$

This more general construction of f_{damp} retains all of the relevant damping function properties listed above and allows us to tune the parameter, α, to reproduce *ab initio* electrostatic energies. This is identical to the damping function proposed separately by both Gordon and co-workers[11] and Piquemal and co-workers.[10]

Using the damping formulation of Eq. 2.7, we have now effectively changed the potential due to every atom in a given system. The potential at any point in the system is described by,

$$V(r) = \frac{Z}{r} + f_{damp}(r) \cdot V_{classical} = \frac{Z}{r} + (1 - e^{-\alpha r}) \cdot V_{classical} \tag{2.8}$$

where the potential due to the nucleus is unchanged, but the potential due to the electrons now accounts for the charge penetration effect. This, however, is not quite enough to get the interaction energy between two atoms. Recall from figure 2.3 that although the second and third terms of the charge penetration corrected electrostatic interaction energy involve simple point charges interacting with the potential due to smeared charge distributions, the fourth term has two smeared

charge distributions interacting with each other. In this unique case, we must derive a second "overlap" damping function to account for this interaction.

For the fourth, overlap term we are attempting to approximate the overlap integral between the two charge distributions,

$$U_4 = \int \frac{\rho_A \rho_B}{r} dv_A dv_B = \frac{1}{2}\left(\int \rho_A V_B(\boldsymbol{A})\, dv_A + \int \rho_B V_A(\boldsymbol{B})\, dv_B\right), \tag{2.9}$$

where V_A and V_B are the charge penetration corrected potentials due to atoms A and B respectively. Gordon and co-workers approximate this integral using the one-center method given by Coulson[26] to yield

$$U_4 = \frac{q_A q_B}{r}\left(1 - \frac{\alpha_B^2}{(\alpha_B^2 - \alpha_A^2)} e^{-\alpha_A r} - \frac{\alpha_A^2}{(\alpha_A^2 - \alpha_B^2)} e^{-\alpha_B r}\right) = \frac{q_A q_B}{r} \cdot f_{damp}^{overlap1}(r) \tag{2.10a}$$

where q_A and q_b are the total electron charges of atoms A and B, for the charge-charge portion of the interaction. Piquemal and co-workers take a two-center approach to approximating the integral,

$$U_4 = \frac{q_A q_B}{r}\left(1 - e^{-\beta_A r}\right)\left(1 - e^{-\beta_B r}\right) = \frac{q_A q_B}{r} \cdot f_{damp}^{overlap2}(r) \tag{2.10b}$$

where, as laid out in our previous work (Ref. 20), a second parameter is introduced to describe the overlap. While the derivations of these formulae are slightly different, mathematically these U_4 overlap damping functions constitute the only functional difference between the models of Gordon and co-workers and Piquemal and co-workers. For simplicity's sake, the approach of Eq. 2.10a will be referred to as model 1 and Eq. 2.10b as model 2. They can be implemented, however, in an identical manner. These overlap damping functions allow us to calculate the charge penetration corrected charge-charge electrostatic interaction between any two sites:

$$U_{electrostatic}^{charge-charge} = \frac{Z_A Z_B}{r} + \frac{Z_A q_B}{r} f_{damp}(r) + \frac{Z_B q_A}{r} f_{damp}(r) + \frac{q_A q_B}{r} f_{damp}^{overlap}(r). \tag{2.11}$$

The AMOEBA model, however, has more than just charges on every atom. It uses a multipole expansion representing the charge distribution at every site. The energy between two AMOEBA multipole sites, i and j, is given by,

$$U_{multipole} = \boldsymbol{M}_i^t \boldsymbol{T}_{ij}^{classical} \boldsymbol{M}_j \tag{2.12}$$

where $\boldsymbol{M}_i$ and $\boldsymbol{M}_j$ represent the multipole moments on atoms i and j respectively, and

$$\boldsymbol{T}_{ij}^{classical} = \begin{bmatrix} 1 & \frac{\partial}{\partial x_j} & \frac{\partial}{\partial y_j} & \frac{\partial}{\partial z_j} & \frac{\partial^2}{\partial x_j^2} & \cdots \\ \frac{\partial}{\partial x_i} & \frac{\partial^2}{\partial x_i \partial x_j} & \frac{\partial^2}{\partial x_i \partial y_j} & \frac{\partial^2}{\partial x_i \partial z_j} & \frac{\partial^3}{\partial x_i \partial x_j^2} & \cdots \\ \frac{\partial}{\partial y_i} & \frac{\partial^2}{\partial y_i \partial x_j} & \frac{\partial^2}{\partial y_i \partial y_j} & \frac{\partial^2}{\partial y_i \partial z_j} & \frac{\partial^3}{\partial y_i \partial x_j^2} & \cdots \\ \frac{\partial}{\partial z_i} & \frac{\partial^2}{\partial z_i \partial x_j} & \frac{\partial^2}{\partial z_i \partial y_j} & \frac{\partial^2}{\partial z_i \partial z_j} & \frac{\partial^3}{\partial z_i \partial x_j^2} & \cdots \\ \frac{\partial^2}{\partial x_i^2} & \frac{\partial^3}{\partial x_i^2 \partial x_j} & \frac{\partial^3}{\partial x_i^2 \partial y_j} & \frac{\partial^3}{\partial x_i^2 \partial z_j} & \frac{\partial^4}{\partial x_i^2 \partial x_j^2} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \left(\frac{1}{r_{ij}}\right) \tag{2.13}$$

is the classical point multipole interaction matrix. We can see in Eq. 2.13 that the interaction matrix, $\boldsymbol{T}_{ij}$, for AMOEBA without charge penetration is obtained simply by taking repeated derivatives of the classical coulomb potential, 1/r. To account for charge penetration, not just in charge-charge interactions, but in all multipole interactions up to arbitrary order, we simply insert the charge penetration damped potential in place of the classical potential. This yields the charge penetration corrected multipole interaction matrix,

$$\boldsymbol{T}_{ij} = \begin{bmatrix} 1 & \frac{\partial}{\partial x_j} & \frac{\partial}{\partial y_j} & \frac{\partial}{\partial z_j} & \frac{\partial^2}{\partial x_j^2} & \cdots \\ \frac{\partial}{\partial x_i} & \frac{\partial^2}{\partial x_i \partial x_j} & \frac{\partial^2}{\partial x_i \partial y_j} & \frac{\partial^2}{\partial x_i \partial z_j} & \frac{\partial^3}{\partial x_i \partial x_j^2} & \cdots \\ \frac{\partial}{\partial y_i} & \frac{\partial^2}{\partial y_i \partial x_j} & \frac{\partial^2}{\partial y_i \partial y_j} & \frac{\partial^2}{\partial y_i \partial z_j} & \frac{\partial^3}{\partial y_i \partial x_j^2} & \cdots \\ \frac{\partial}{\partial z_i} & \frac{\partial^2}{\partial z_i \partial x_j} & \frac{\partial^2}{\partial z_i \partial y_j} & \frac{\partial^2}{\partial z_i \partial z_j} & \frac{\partial^3}{\partial z_i \partial x_j^2} & \cdots \\ \frac{\partial^2}{\partial x_i^2} & \frac{\partial^3}{\partial x_i^2 \partial x_j} & \frac{\partial^3}{\partial x_i^2 \partial y_j} & \frac{\partial^3}{\partial x_i^2 \partial z_j} & \frac{\partial^4}{\partial x_i^2 \partial x_j^2} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \left(\frac{1}{r_{ij}}\right) f_{damp}(r), \tag{2.14}$$

where f_{damp} is either 1 (for nuclear-nuclear interactions), the damping function from Eq. 2.7 (for the second and third terms of the interaction energy), or the overlap damping function from Eqs. 10a or 10b (for the fourth term of the interaction energy). Using the charge penetration corrected multipole interaction matrices, we can express the new AMOEBA multipole interaction energy of any two sites as:

$$U_{electrostatic}^{CP} = \frac{Z_i Z_j}{r} + Z_i \boldsymbol{T}_{ij}^{damp} \boldsymbol{M}_j + Z_j \boldsymbol{T}_{ji}^{damp} \boldsymbol{M}_i + \boldsymbol{M}_i^t \boldsymbol{T}_{ij}^{overlap} \boldsymbol{M}_j. \quad (2.15)$$

Eq. 2.15 allows us to account for the effects of charge penetration up to arbitrary order multipole expansion. For AMOEBA, which has multipole interactions up to quadrupole-quadrupole, this means that the charge penetration model can be made fully consistent with the multipole model. See Appendix A for explicit damping functions for all AMOEBA multipole interaction components.

2.3 Parameterization

The goal of including charge penetration in the AMOEBA model is to more accurately reproduce the energies of electrostatic interactions between molecules at short range. Because both models 1 and 2 contain empirical parameters, we will seek to optimize them by fitting to a database of relevant intermolecular electrostatic energies. In our previous work, the S101 and S101x7 databases where constructed for this purpose.[23] The S101 database contains 101 unique pairs of both homodimers and heterodimers of common organic molecules. It contains the widely used S66 database[27] along with some additional relevant biomolecular interactions. The S101x7 database is constructed by placing each dimer pair from the S101 database at 0.70, 0.80, 0.90, 0.95, 1.00, 1.05 and 1.10 times their equilibrium intermolecular distance. A schematic representation of all the dimer pairs included in the S101 database is shown in figure 2.5.

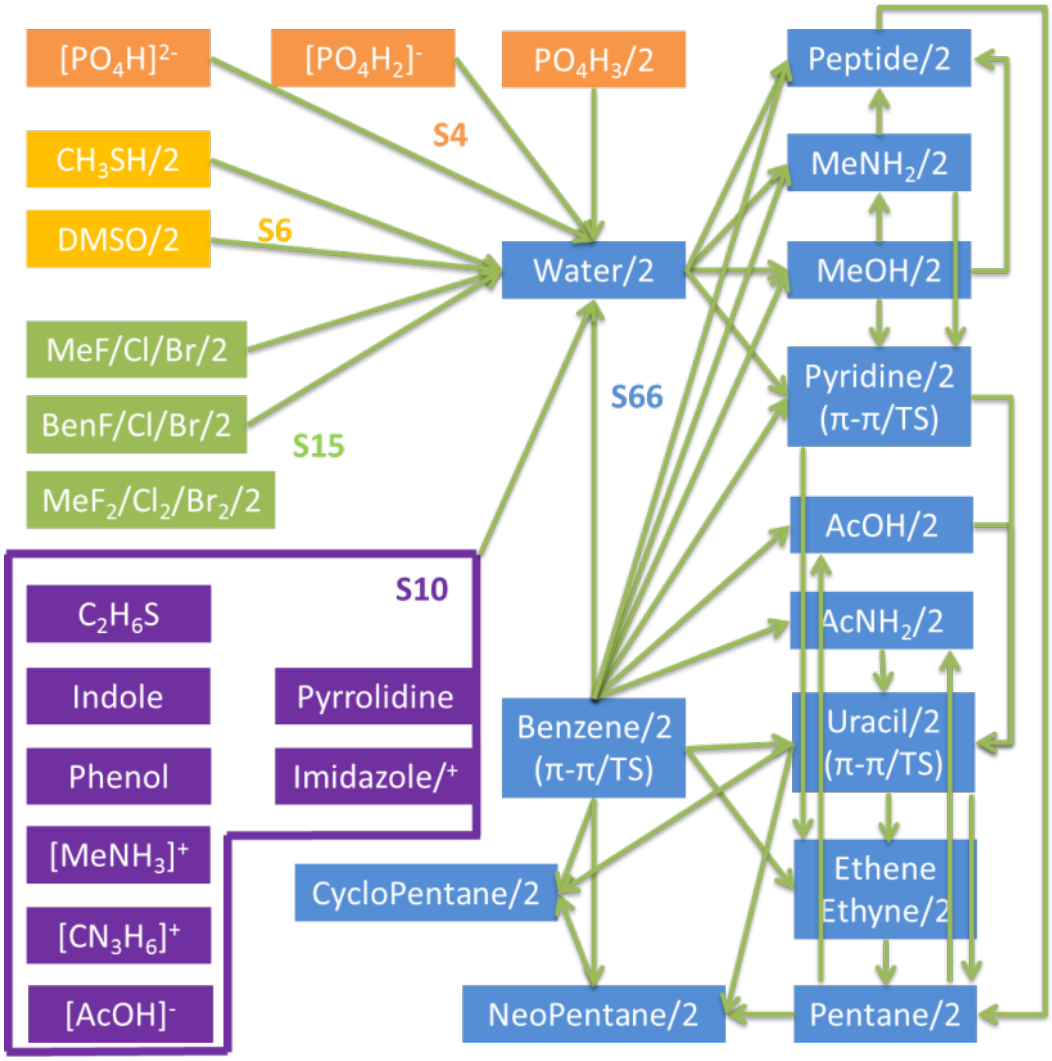


Figure 2.5 Dimer pairs in the S101 database.

Arrows connect monomers that form dimers. A "/2" designation indicates a homodimer. A "/+" designation indicates both neutral and positively charged forms. Reproduced from reference 20

In all of the parameterization that follows, the entire S101x7 database was used with the exception of interactions involving ethyne. The omission of ethyne allows direct comparison with the results from our previous work.

To parameterize the charge penetration models against the S101x7 database, accurate intermolecular electrostatic energies are needed for all dimer pairs. In the previous work,

Symmetry Adapted Perturbation Theory (SAPT)[28] calculations where performed to obtain these energies. SAPT calculations decompose intermolecular energies into physically meaningful components; the intermolecular energy between two monomers is broken down into electrostatic, induction, exchange-repulsion and dispersion energies. For the S101x7 database, SAPT2+ calculations[29,30], estimated at the complete basis set (CBS) limit as described in Ref. 22, were carried out to return the *ab initio* electrostatic interaction energy of each dimer pair.

The parameters of model 1 and model 2 were optimized by performing a nonlinear least squares fit to minimize the difference between the AMOEBA electrostatic energy (with charge penetration), $U_{electrostatic}^{AMOEBA}$, and the SAPT electrostatic energy, $U_{electrostatic}^{SAPT}$, for each dimer pair. For models 1 and 2, two methods of parameterizing are proposed. In the first method one parameter, α, is assigned per element. In the second, one α is assigned per charge penetration class. These classes, as listed in table 2.2, are simply chosen to allow for different descriptions of atoms of the same element but different physiochemical classifications.

Element	Charge-charge Damping				Charge Penetration Class	Charge-charge Damping			Higher-order Damping		
	Model 1 α (Å^{-1})	Model 2 α (Å^{-1})	Model 2 γ	Model 3 ζ (Å^{-1})		Model 1 α (Å^{-1})	Model 2 α (Å^{-1})	Model 2 γ	Model 1 α (Å^{-1})	Model 2 α (Å^{-1})	Model 2 γ
Hydrogen (H)	4.0026	10.000	0.8720	1.2976	non-polar (H-C)	3.4345	3.5474	0.8710	3.2484	3.2624	0.8823
					aromatic (H-C)	3.9419	4.0006		3.4437	3.4080	
					polar, water (H-X)	5.0049	10.000		3.2632	3.4317	
Carbon (C)	3.0957	2.9137		1.2100	sp^3, tetrahedral	3.3863	3.2136		3.5898	3.7576	
					sp^2, aromatic	3.1205	2.9268		3.2057	3.2569	
					sp^2, carbonyl, etc.	3.1702	2.9349		3.1286	3.1971	
Nitrogen (N)	3.7321	3.4066		1.4502	sp^3, tetrahedral	3.2519	3.2317		4.0135	3.9410	
					sp^2, aromatic	3.6979	3.4199		3.6358	3.7534	
					sp^2, other	3.4264	3.3110		3.7071	3.8244	
Oxygen (O)	4.1390	3.5677		1.4114	sp^3, hydroxyl, water	3.7975	3.7038		4.1615	4.2449	
					sp^2, aromatic	3.7770	3.6686		4.3778	5.0908	
					sp^2, carbonyl	3.4938	3.4509		3.7321	3.6146	
Phosphorous (P)	3.0661	2.5969		1.3369	phosphate	3.1539	2.6076		2.7476	3.0668	
Sulfur (S)	2.9570	2.5965		1.1156	sulfide, thiol	3.2046	2.7320		3.3112	3.3826	
					sulfur (IV)	3.3824	2.6353		2.6247	2.9057	
Fluorine (F)	4.4875	4.2333		1.5955	organofluoride	4.4314	4.2730		4.4675	10.000	
Chlorine (Cl)	3.5173	2.9092		1.2102	organochloride	3.5060	2.8887		3.4749	3.5035	
Bromine (Br)	3.7202	2.5924		1.2259	organobromide	3.7150	2.5820		3.6696	3.7146	

Table 2.2 Atom classes and fitted parameters for charge penetration models.

The choice of classes is based on the knowledge that the electronic structure of an sp2 hybridized carbon, for example, will be generally different than that of an aromatic carbon. While it is certainly true that differences in electron distribution exist even amongst atoms of the same charge penetration class (the electronic structure of every sp2 hybridized carbon is not exactly the same), the guiding principle is to include only the minimal level of atomic classification to allow the model to be easily transferable.

For model 2, the parameter, β, is fixed as a fraction of α, $\beta = \gamma \cdot \alpha$, where the parameter, γ, is taken to be universal to avoid over-fitting. Allowing β to float for every charge penetration class has the potential, of course, to improve the overall fit, but at the cost of losing physical meaningfulness. Recall from Eq. 2.10b that although the β parameter is specific to the overlap function in model 2, the two electron clouds that are overlapping are supposed to already be described by the parameter α. Allowing both α and β to float in the fit would allow two different parameters to describe essentially the same physics. Instead fitting one universal parameter γ simply describes how β should be generally related to α in approximating the overlap between molecules. It should be noted that the parameterization strategy here for model 2 differs slightly from previous work. It is chosen in this way to best fit the AMOEBA multipole model and provide for a direct comparison with model 1 on the same test set.

The results of fitting model 1 and model 2 are shown in table 2.2. Three fits were performed for each model. First the S101x7 database of intermolecular electrostatic energies was fit using only charge-charge damping with parameters assigned by element. Next, the same charge-charge damping fit was performed with parameters assigned by class. Then the database was fit using higher-order damping with damping of all AMOEBA multipole interactions (up to and including quadrupole-quadrupole).

In addition to parameterizing models 1 and 2, a third model, due to Wang and Truhlar[18-20] has been parameterized as well. This model, developed for application in QM/MM calculations, is included as a point of comparison. However, it is not developed any further than charge-charge damping using parameters assigned by element as it has several properties that make it unsuitable for implementation in AMOEBA. First, the model can be unstable with respect to the parameters of interacting atoms. If two closely interacting atoms have parameters that are close, but not identical, the overlap damping functions of the model breaks down. Second, expanding the model to include higher-order damping to make it fully consistent with the AMOEBA multipole model is computationally intractable with this model. The expressions that form the overlap damping functions, as seen in Eqs. 8 and 9 in Ref. 19 are much more complex functions of the radial distance between atoms, r. Taking the successive derivatives necessary for higher-order damping terms would produce expressions too expensive to calculate for our purposes. Third, even if such derivatives were deemed necessary, the model's framework is incompatible with higher-order damping. The damping functions used in Wang and Truhlar's model are meant to simulate the outer Slater-type orbitals of atoms. With this being the case, rather than treat all of an atom's electrons as damped, the model only treats a maximum of 2 as damped. This treatment is acceptable for charge-charge damping since charge is spherically symmetric and one simply treats the remaining electrons as part of the "core". This is, however, problematic for higher-order damping because there is no such simple partitioning of the electrons that make up an atom's dipole and quadrupole moment. It would be nonsensical to apply the model's damping terms meant for two electrons, to an atom's dipole and quadrupole interactions.

In the following section the fits produced by the parameterization of all three models is presented. The fits of each model to the S101x7 database will be used along with some important

validation tests and theoretical arguments to determine which model and which parameterization strategy to implement in AMOEBA.

2.4 Results

To understand how charge penetration improves the electrostatic model of AMOEBA, we must understand how the current AMOEBA model without a charge penetration correction performs. Figure 2.6 shows how AMOEBA's prediction of intermolecular electrostatic energies compares to the SAPT *ab initio* electrostatic energy values on the S101x7 database.

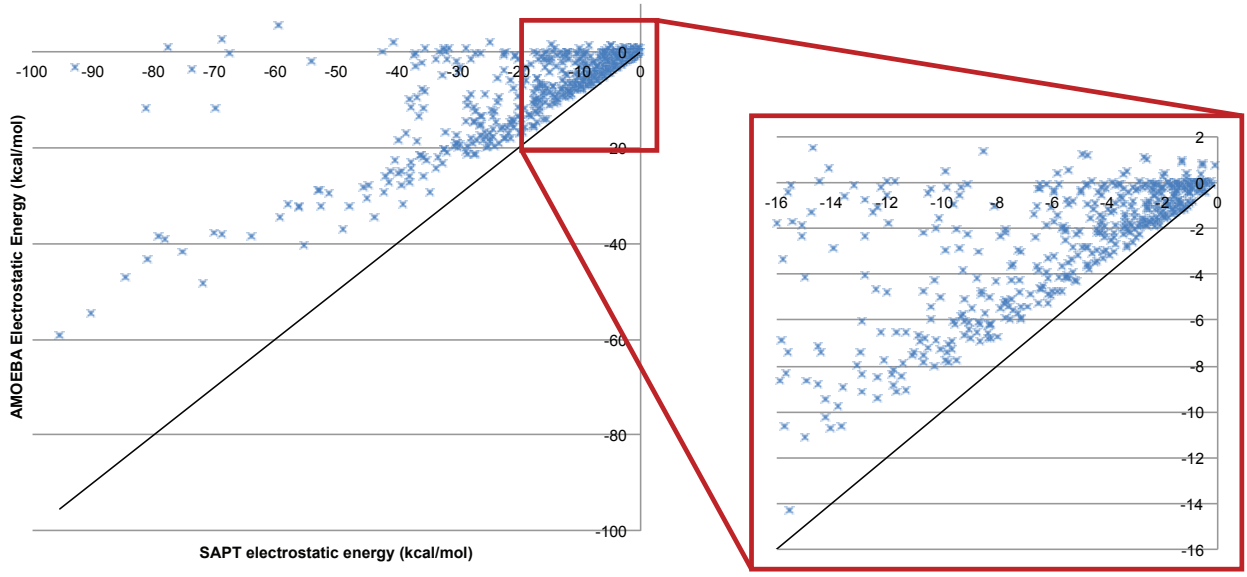


Figure 2.6 AMOEBA, multipole-only intermolecular electrostatic energy of dimers in S101x7 database. The multipole-only electrostatic energy for each dimer is plotted against the benchmark SAPT electrostatic energy. The diagonal, y=x line indicates what would be perfect agreement. Compared to the benchmark calculations, the multipole-only model systematically overestimates the electrostatic energy.

Figure 2.6 reveals that using only a multipole expansion to describe the electrostatic interactions between molecules systematically overestimates the electrostatic energy at short range. The

pervasive gap illustrated in figure 2.6 illustrates the need for including charge penetration in the electrostatic model of the AMOEBA force field.

The most naïve method of applying a charge penetration correction is to assign one parameter per element and damp only the charge-charge electrostatic interactions. As a first test of the theory, this strategy was implemented for models 1, 2 and 3. Each model was then parameterized by fitting to the S101x7 database. The overall results of assigning parameters by element and damping only the charge-charge electrostatic interactions are illustrated in the first cluster of columns in figure 2.7.

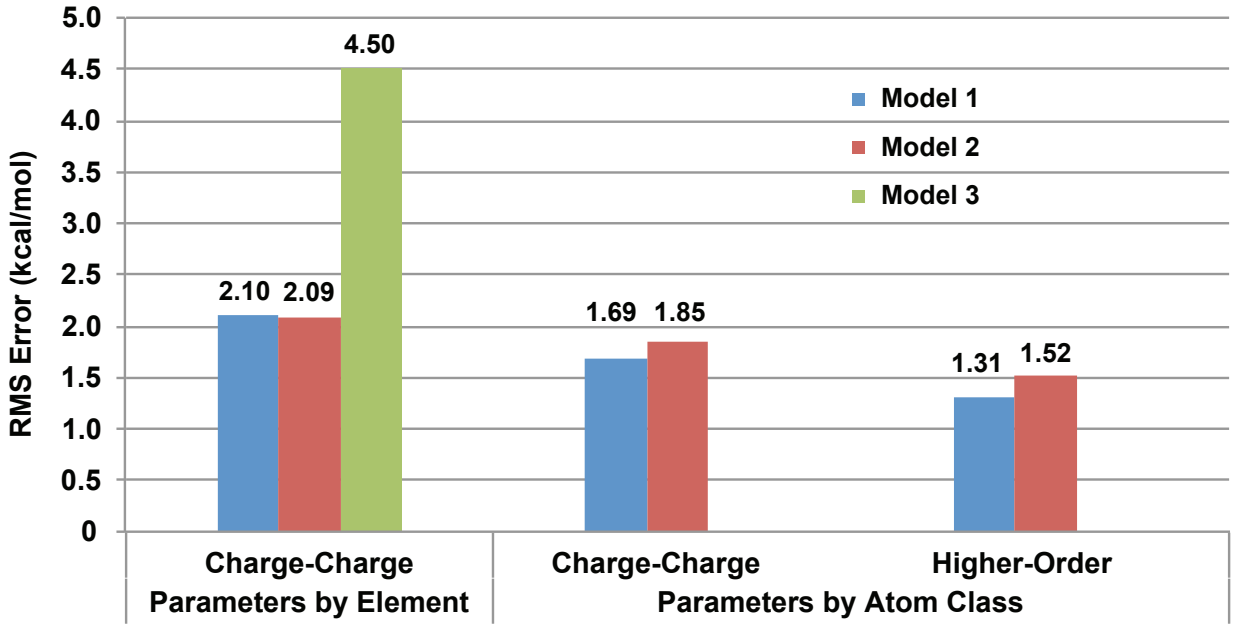


Figure 2.7 Root mean square error of AMOEBA electrostatic energy with charge penetration on S101x7 database.

Multiple charge penetration models were tested. The first cluster of columns represents the results of parameters fit by element with charge-charge damping only. The second cluster is the results of having parameters assigned by class and charge-charge damping. The third cluster is the results for including higher-order damping in addition to having parameters assigned by class. (RMS error of AMOEBA with multipoles-only is 13.4 kcal/mol)

It is clear that all three models perform much better than the current AMOEBA multipole only model. The RMS error of the multipole-only model for electrostatic energies on the S101x7 database is 13.4 kcal/mol. Models 1, 2 and 3 bring that error down to 2.1 kcal/mol, 2.1 kcal/mol and 4.5 kcal/mol respectively, showing that even a naïve damping strategy starts to capture the missing physics. It is also apparent that models 1 and 2 perform much better, even at this low level of implementation, than model 3. Additionally, note that despite having fewer parameters, model 1 performs nearly identically to model 2 for this implementation. Complete statistics for each of these fits, including a breakdown by intermolecular distance, are available in Appendix A.

While assigning parameters by element produces an improvement over the multipole-only AMOEBA model, it ignores some key physiochemical properties of elements in different bonding environments relevant to interpreting the α parameter. The α parameter with units, $Å^{-1}$, can be understood as the inverse of the physical extent of the electron cloud of an atom. From *ab initio* electronic structure calculations we know that in general this property can change substantially based on the bonding environment of an atom. For this reason, we fit models 1 and 2 with parameters assigned by class to the S101x7 as described in the preceding section. The overall results of assigning parameters by class and still damping only the charge-charge electrostatic interactions are illustrated in the second cluster of columns in figure 2.7. The first thing to note is the absence of a fit for model 3. Once the parameter set is expanded to include classes, model 3 becomes highly unstable. As noted, before this is due to numerical instability when parameters in the model become close. This is practically unavoidable for class-based parameters, so model 3 is excluded from this point forward. More importantly, however, we notice also that splitting out different parameter classes improves the overall fit to the S101x7 database for models 1 and 2. Assigning parameters by class improves the performance on the RMS error. Again, despite having

fewer parameters, model 1 outperforms model 2 in this case. This improvement is largely due to allowing different classes for the same element. For example, table 2.2 shows that for model 1 the parameter for hydrogen in the element based fit splits quite significantly when one allows different classes to vary. The element parameter, 4.0 $Å^{-1}$ splits into parameters of 3.4 $Å^{-1}$, 3.9 $Å^{-1}$ and 5.0 $Å^{-1}$ for non-polar, aromatic and polar hydrogen respectively. This extra flexibility in the parameterization, rooted in basic physiochemical properties improves our overall description of the electrostatics. Again, specific statistics for class-based fits can be found in the Appendix A.

Splitting out separate chemical classes for parameters improves the performance of our charge-charge damping charge penetration model, but it unfortunately does not meet the criteria of being fully consistent with the AMOEBA multipole electrostatic model. To test the fully integrated model we implemented charge penetration damping for all multipole interaction terms (up to and including quadrupole-quadrupole) for both models 1 and 2. We will refer to this model as "higher-order" damping. The overall results, illustrated in the third and final cluster of columns in figure 2.7, show the improvement that this model brings. Implementing a fully integrated higher-order damping model with class-based parameters brings the RMS error on the entire S101x7 database for models 1 and 2 down to 1.31 kcal/mol and 1.52 kcal/mol respectively. Full statistical analysis can be found in Appendix A. These numbers represent a dramatic improvement over the current AMOEBA multipole-only RMS error of 13.43 kcal/mol. More importantly they also improve on the errors from our charge-charge damping implementations. A significant portion of the improvement is due to improvement in the performance on the closest dimer pairs in the S101x7 database. Among dimers that are separated by 0.70 and 0.80 of their equilibrium distance, model 1 with higher-order damping reduced that error from 2.75 kcal/mol to 2.27 kcal/mol, and model 2 reduced it from 4.36 kcal/mol to 2.64 kcal/mol. Importantly, this improvement does not

sacrifice the fit at more accessible distances. For model 1 the RMS error on dimers with intermolecular separations of 0.90 to 1.10 times their equilibrium distance dropped to under 1 kcal/mol compared with an error of over 4 kcal/mol for the current multipole-only model. Lastly, these fits give a slight edge to the simpler model 1 over model 2. Model 1 performs 16% better than model 2 on overall RMS errors in the S101x7 database when higher-order damping is included. The absolute percent error of model 2 on the electrostatic energies of the S101x7 database is 10%, while model 1 gives 7%.

Figure 2.7 lays out the overall performance of each of the implementations described above. It is clear from this data that model 1 with higher-order damping and parameters assigned by class gives the best fit to the electrostatics of the S101x7 database. The improvement this model gives on each individual dimer pair is shown in figure 2.8.

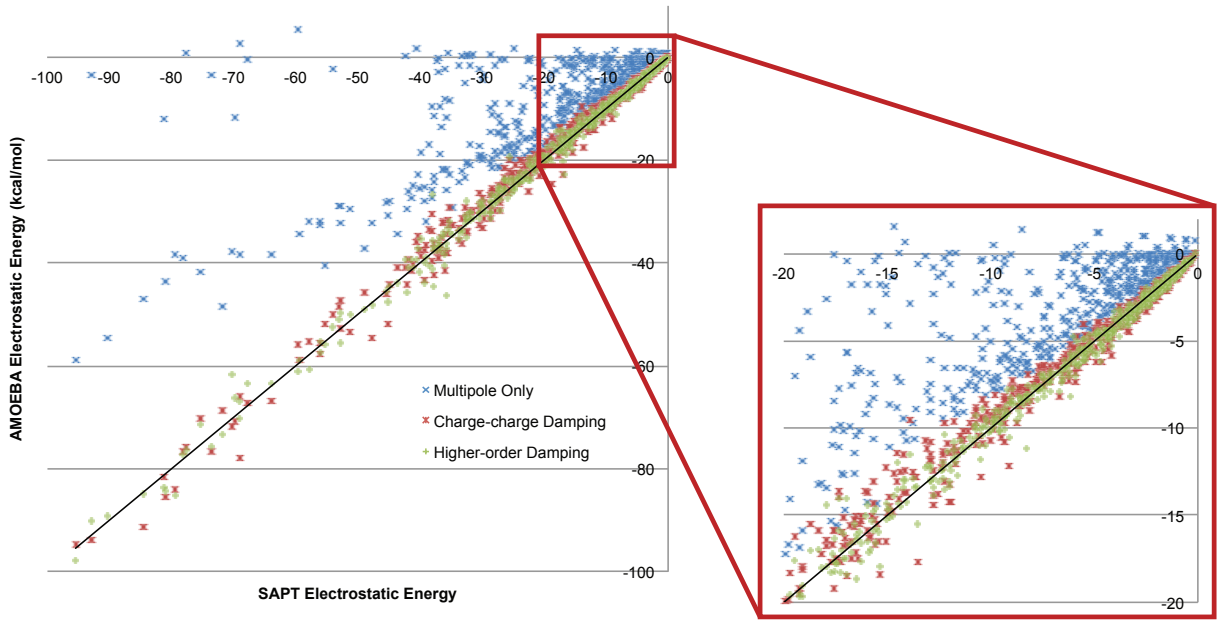


Figure 2.8 AMOEBA intermolecular electrostatic energy with and without charge penetration of S101x7 database dimers.

The AMOEBA electrostatic energy both without (multipole-only) and with (model 1 with charge-charge or higher-order damping) charge penetration is plotted against benchmark SAPT electrostatic energy calculations. The diagonal, y=x line indicates what would be perfect agreement. Including higher-order damping in the charge penetration model yields the best agreement with *ab initio* electrostatic energies.

Figure 2.8 shows that across the board model 1 with higher-order damping is superior to simple charge-charge damping and represents a dramatic improvement over the current multipole-only model. This is borne out in a handful of important and instructive examples. Figure 2.9 lays out the results for fitting the water dimer, figure 2.10 shows two important orientations of the benzene dimer and figure 2.11 shows the model's performance on phosphate ions.

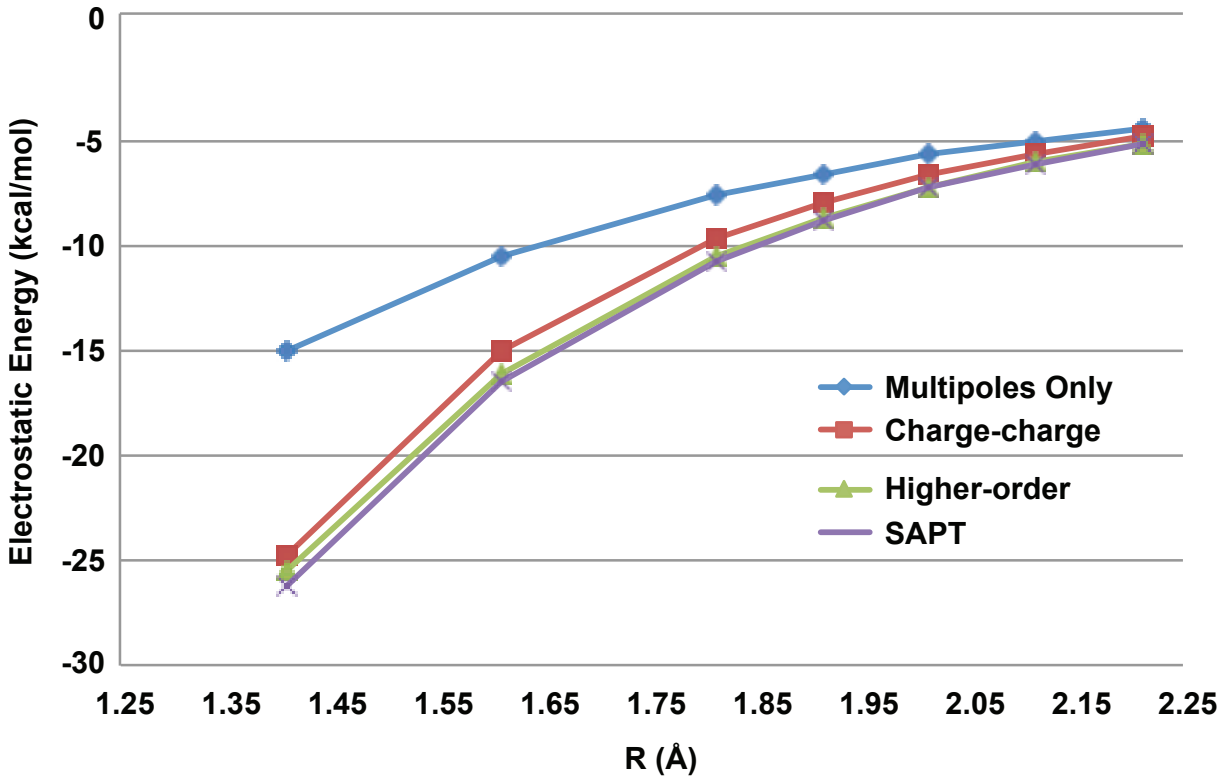
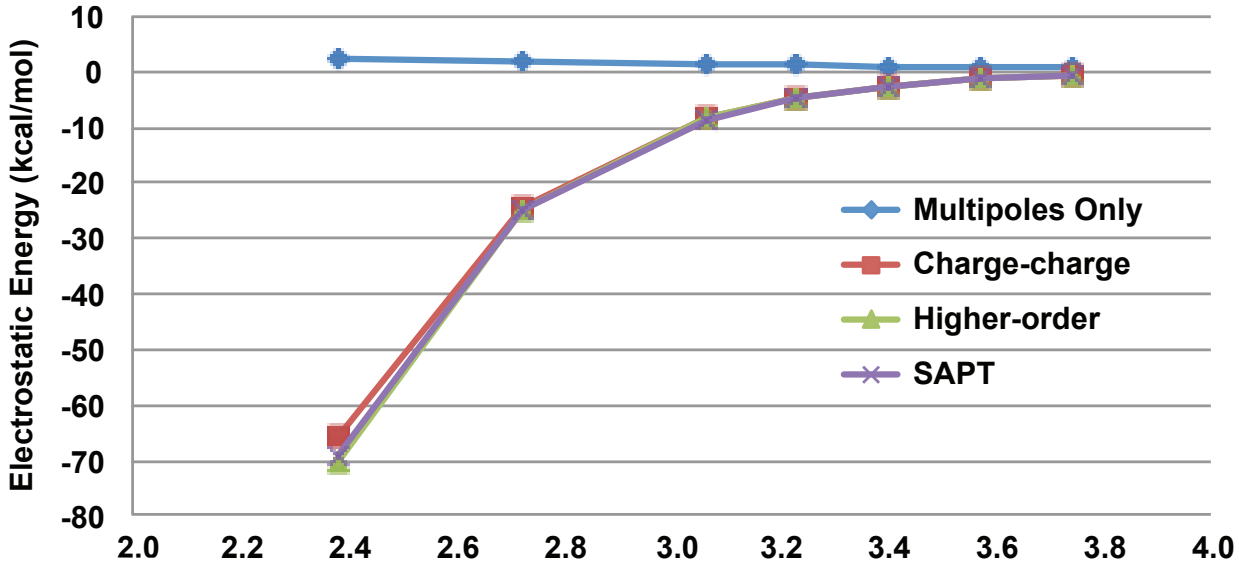


Figure 2.9 Water dimer electrostatics.

AMOEBA dimer electrostatic energies without (multipoles-only) and with (model 1 with charge-charge and higher-order damping) charge penetration are plotted against benchmark SAPT electrostatic energies.

a)



b)

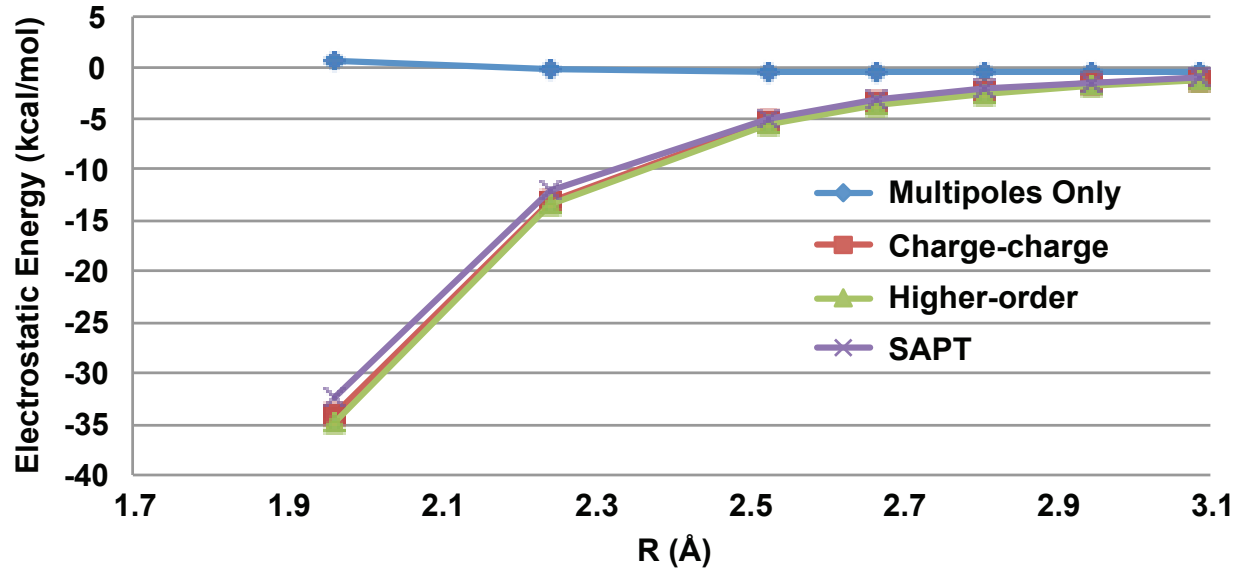
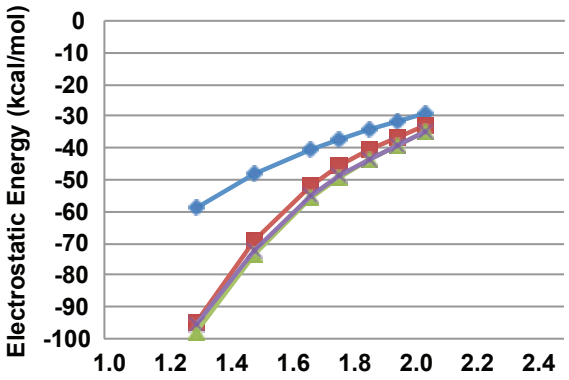


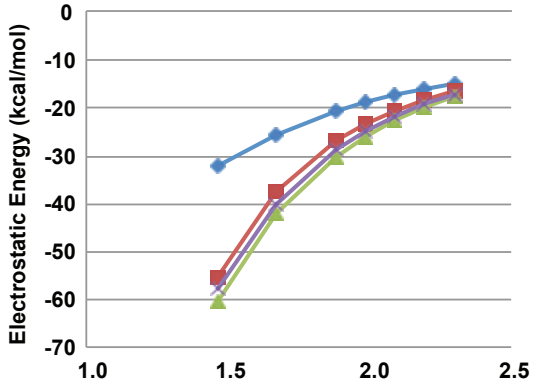
Figure 2.10 Benzene (a) Sandwich and (b) T-shape dimer electrostatics.

AMOEBA dimer electrostatic energies without (multipoles-only) and with (model 1 with charge-charge and higher-order damping) charge penetration are plotted against benchmark SAPT electrostatic energies.

a)



b)



c)

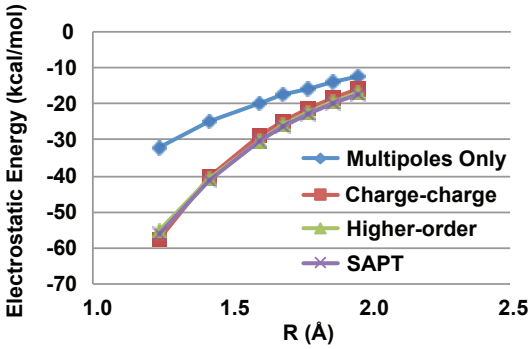


Figure 2.11 Phosphate-water dimer electrostatics.

AMOEBA dimer electrostatic energies without (multipoles-only) and with (model 1 with charge-charge and higher-order damping) charge penetration are plotted against benchmark SAPT electrostatic energies. Results are shown for PO4H (a), PO4H2 (b) and PO4H3 (c).

These three examples represent important relevant biomolecular interactions that the current multipole-only model fails to accurately capture. Moreover, all three also show that an integrated higher-order damping model is needed to achieve the highest level of agreement with SAPT electrostatic data. These examples show that not only does the model generally improve the quality of electrostatics across a wide dataset, but it also performs well on individual examples, such as the benzene sandwich dimer, that inspired our investigation of the charge penetration phenomenon.

2.5 Validation

The fit to the S101x7 database with model 1 higher-order damping is a welcome result. The model dramatically improves the quality of the electrostatic fit for those electrostatic interactions over AMOEBA's current multipole-only model and it outperforms all of the other relevant damping models proposed. There are, however, some considerations that need to be addressed to validate model 1 with higher-order damping as the best option for capturing the physics of charge penetration. First, we would like to show that in addition to giving the best fit, model 1 is also the most robust option. Second, we need to know to what extent this charge penetration model is independent of the AMOEBA multipole model. And most importantly, we must validate that this model is capturing a real physical phenomenon.

It is important our charge penetration model not only provides a good fit to *ab initio* electrostatic data, but also that the model is robust. To evaluate robustness, we must evaluate the sensitivity of the model to small changes in the parameters. Model 3 does not pass this parameter sensitivity requirement. Figure 2.12 shows the behavior of the oxygen–sulfur electrostatic

interaction in the DMSO–water dimer as the difference between oxygen and sulfur parameters gets smaller.

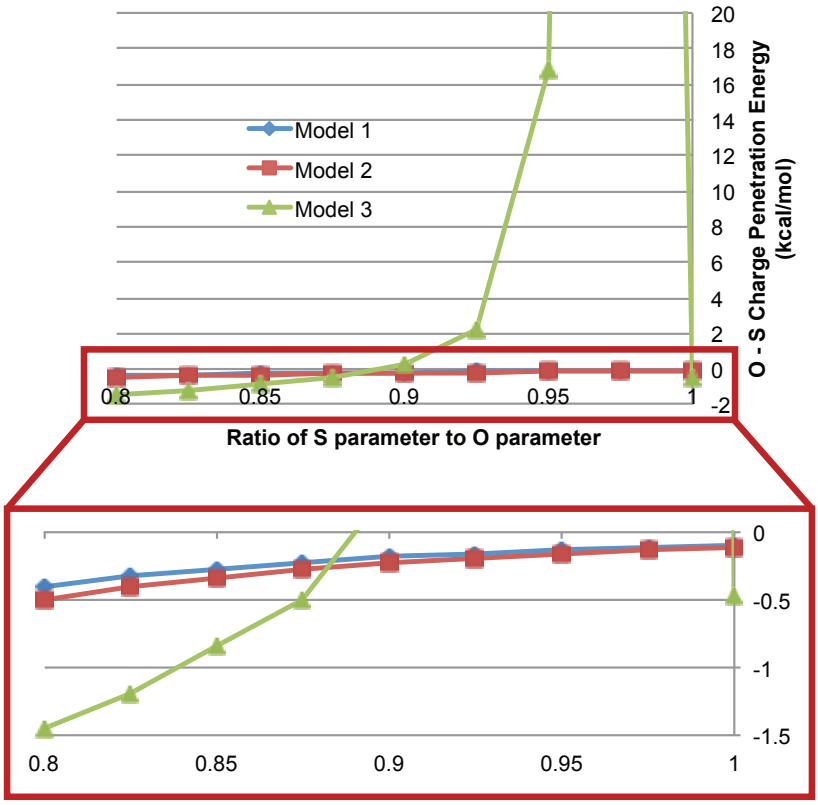


Figure 2.12 Charge penetration model stability.

The oxygen-sulfur electrostatic interaction energy for the water-DMSO dimer is plotted as a function of the difference between the oxygen and sulfur charge penetration parameters. As the ratio of the parameters approaches unity, model 3 becomes unstable.

Clearly model 3 breaks down as the two parameters get close to one another. Moreover, the problem is compounded as the intermolecular distance decreases. Since the zeta parameter multiplies the interatomic distance, r, everywhere in the damping function, the problem gets worse as monomers get closer together. Model 2 does not suffer from any such numerical instability, but

it is sensitive to the parameter, γ, that determines the overlap damping function. Table 2.3 shows that if the closest dimers are left out of our fit to the electrostatic data, γ changes from 0.88 to 0.90.

	Model 1	**Model 2**
Parameters from fit to full S101x7 database	1.31 kcal/mol	1.52 kcal/mol ($\gamma = 0.88$)
Parameters from fit to S101x7 database excluding the closest points (0.8 – 1.1)	1.40 kcal/mol	1.83 kcal/mol ($\gamma = 0.90$)

Table 2.3 Charge penetration model parameter sensitivity.

Models 1 and 2 were fit to the S101x7 database excluding the closest points (all dimers except those at 0.7 times the equilibrium distance). The parameters generated from that fit are then tested on the full database. Model 2, particularly the γ parameter, proves to be the more sensitive to this change.

Moreover, if we use the γ that comes out of the fit where we leave out the closest points, the RMS error for the full S101x7 database jumps from 1.52 kcal/mol to 1.83 kcal/mol. Model 1 on the other hand does not suffer from any such sensitivity. If we leave out the closest dimer pairs and fit parameters to our model, table 2.3 shows that those parameters do almost as well as the parameters fit to the full S101x7 database. The RMS error for model 1 in this case goes up by less than 0.1 kcal/mol. By these tests model 1 shows the strength with respect to numerical stability and parameter transferability we expect a robust charge penetration model to have.

In addition to being the most robust option, model 1 also shows good model independence from the AMOEBA multipole model. AMOEBA follows a defined protocol for determining charge, dipole and quadrupole parameters for each monomer[2] and we should expect that our model should, for the most part, be independent of that specific protocol. In other words, the multipole model and the charge penetration model should not depend on each other. To test this, we use the toy example, benzene. When determining the electrostatic parameters for benzene, multiple values for the opposing charges of the carbons and hydrogens will give nearly identical fits to the electrostatic potential on a grid of points around the molecule. Although the AMOEBA multipole protocol fixes those charge values semi-arbitrarily, we wanted to see if choosing otherwise would break our model 1 charge penetration model. Figure 2.13 demonstrates that model 1 accurately reproduces the electrostatic potential regardless of which potential-fitted charge-dipole-quadrupole model one chooses. This validates an important feature of the model: that it is independent of the specifics of potential fitting protocol for the AMOEBA multipole model.

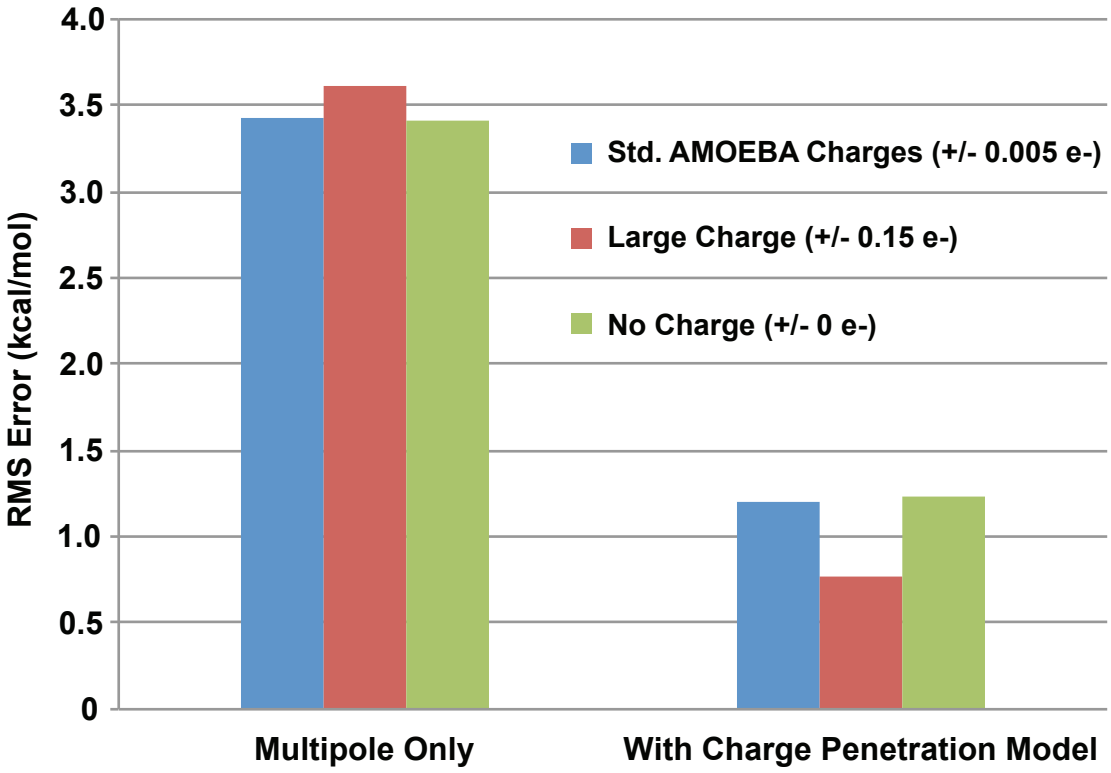


Figure 2.13 Charge penetration model independence.

Three different benzene multipole models were chosen with charges fixed at +/-0.005 e^-, +/- 0.15 e^-, and 0 e^- that give roughly equivalent electrostatic potential fits. The charge penetration model was then applied to all three models. RMS errors of the electrostatic potential on a grid of points around benzene for each model are plotted. The charge penetration significantly lowers the error regardless of multipole model.

Lastly, but most importantly, for our model to be valid, we must prove that it is capturing a real physical effect. At the heart of the charge penetration phenomenon is the fact that the electrostatic potential around an atom at short range cannot be reproduced by a simple point multipole approximation without accounting for the extent of the atom's charge density. To validate that the model is describing this physics we tested to see if our charge penetration model, model 1 with higher-order damping, could accurately reproduce the *ab initio* electrostatic potential

around a molecule at short range. Figure 2.14 shows that without exception the charge penetration model dramatically improves the electrostatic potential fit around every monomer in the S101 database. This is the validation we are looking for. Not only does our model correct the practical problem of bad intermolecular electrostatic energies at close range, but it does so by accurately capturing the physical reality of molecules' finite charge distributions.

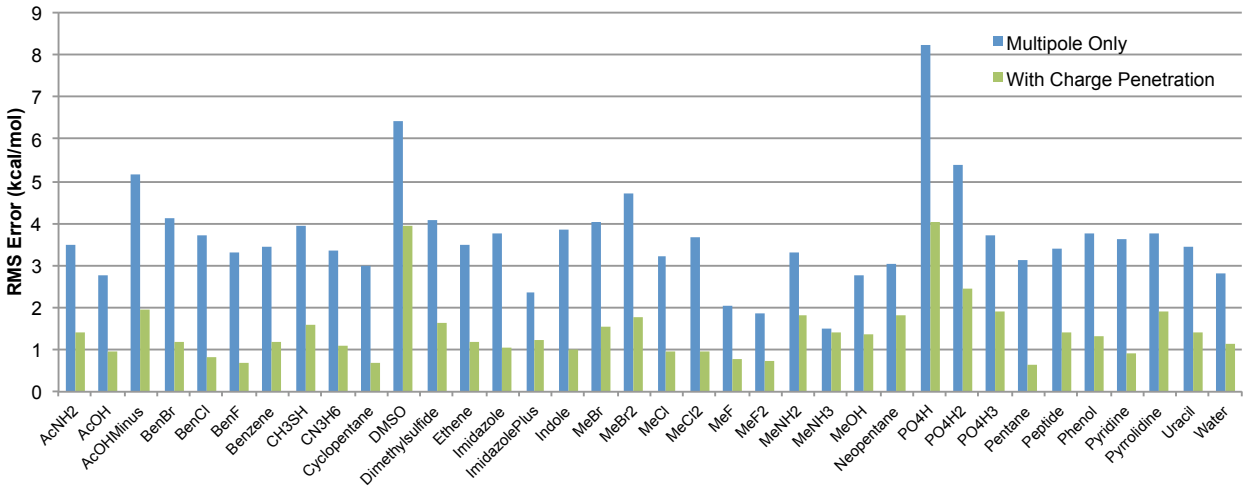


Figure 2.14 Charge penetration model performance on electrostatic potential of monomers in S101 database. The RMS error of the electrostatic potential on a grid of points around each monomer is plotted. Including charge penetration improves the fit to the electrostatic potential for every monomer.

2.6 Test Case: Nucleic Acid Base Stacking

As stated in the introduction, charge penetration effects are important in a broad range of close-contact biomolecular interactions. One essential example is the stacking interactions of nucleobases in DNA and RNA sequences. Parker and Sherrill recently showed that without an explicit accounting for charge penetration, force fields struggle to accurately reproduce the *ab initio* electrostatic energies of these interactions.[8] For instance in an AC:GT base step, the mean

absolute errors (MAE) of the AMBER[31,32] and CHARMM[33] force fields relative to the SAPT electrostatic energy were over 20 kcal/mol. Likewise, we find that AMOEBA without charge penetration gives an electrostatic energy MAE over 20 kcal/mol as well. However, when we apply our charge penetration function with parameters fixed to their values from the S101x7 fit, the MAE drops dramatically to nearly 2 kcal/mol. This improvement is not unique to the AC:GT base step. As shown in figure 2.15, the MAE of our AMOEBA model with charge penetration is significantly lower for every base step combination.

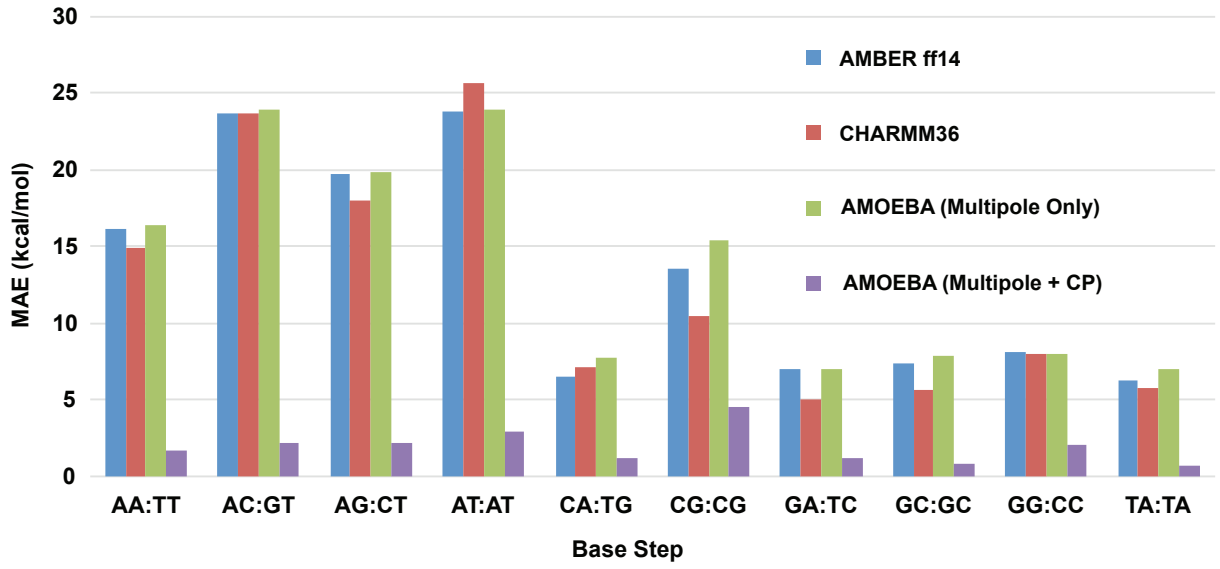


Figure 2.15 Mean absolute electrostatic interaction energy error relative to SAPT0 for ten stacked base steps. Including charge penetration lowers the MAE in the electrostatic interaction energy for every base step combination.

Moreover, this improvement in the electrostatic description of nucleobase stacking holds even for non-equilibrium stacking arrangements. Figure 2.16 shows that for the six structural

parameters that define the stacking interaction, [34] the AMOEBA + charge penetration model does far better than AMBER, CHARMM or the current AMOEBA force field.

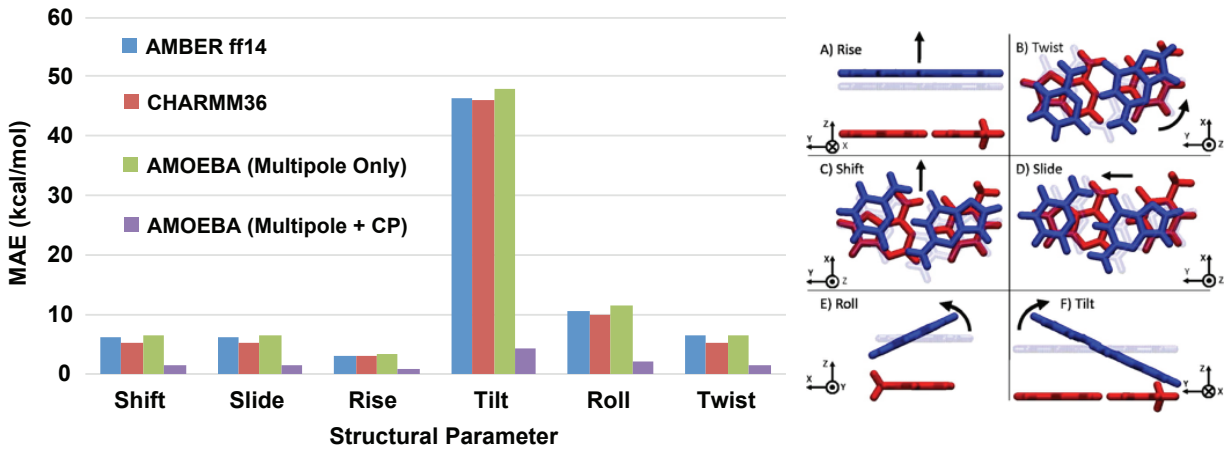


Figure 2.16 Mean absolute electrostatic interaction energy error relative to SAPT for six structural parameters.

Including charge penetration lowers the MAE for variation along every degree of freedom in the nucleobase stacking interaction. Inset reproduced from reference 7.

These data confirm, as asserted by Parker and Sherrill, that including charge penetration is an absolute necessity for a robust nucleic acid force field model. This imperative is highlighted in two standout cases of the TA:TA base step. Figure 2.17 shows the performance of force field models against SAPT electrostatics versus the nucleobase rise.

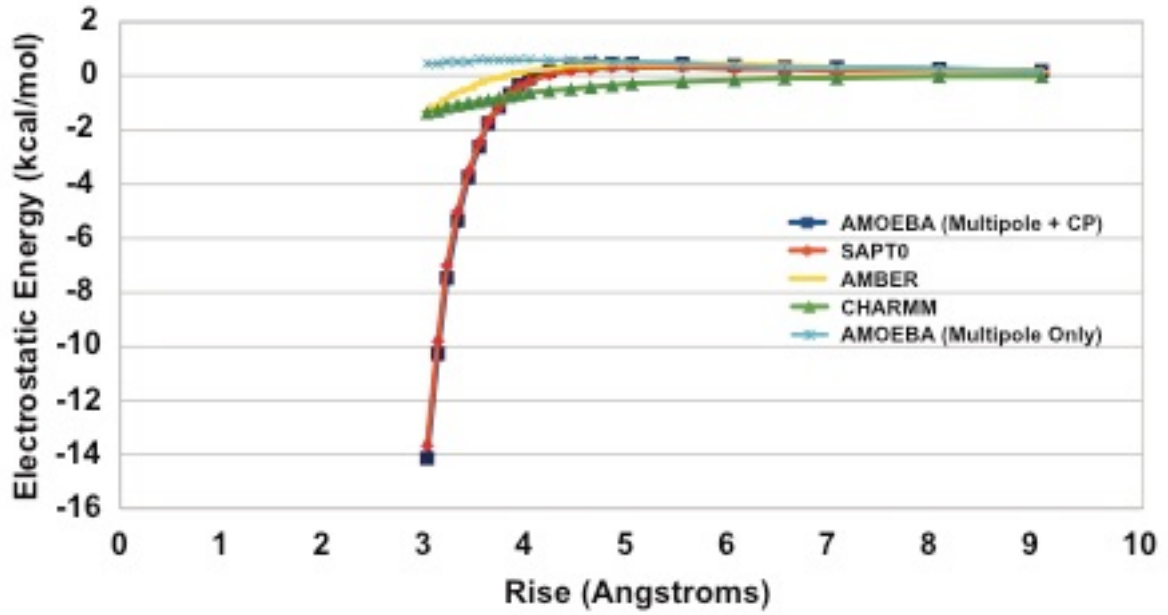


Figure 2.17 Electrostatic energy of a stacked TA:TA interaction *vs.* Rise.

Including charge penetration reproduces the *ab initio* SAPT electrostatic energy over the range of rise parameters. The behavior is consistent with that of the benzene dimer interaction (see figure 2.10).

It is immediately clear that the AMOEBA + charge penetration model put forward here is the only model that accurately reproduces the electrostatic nature of this interaction. The same is seen in figure 2.18 where we examine the electrostatic energy as a function of the tilt parameter.

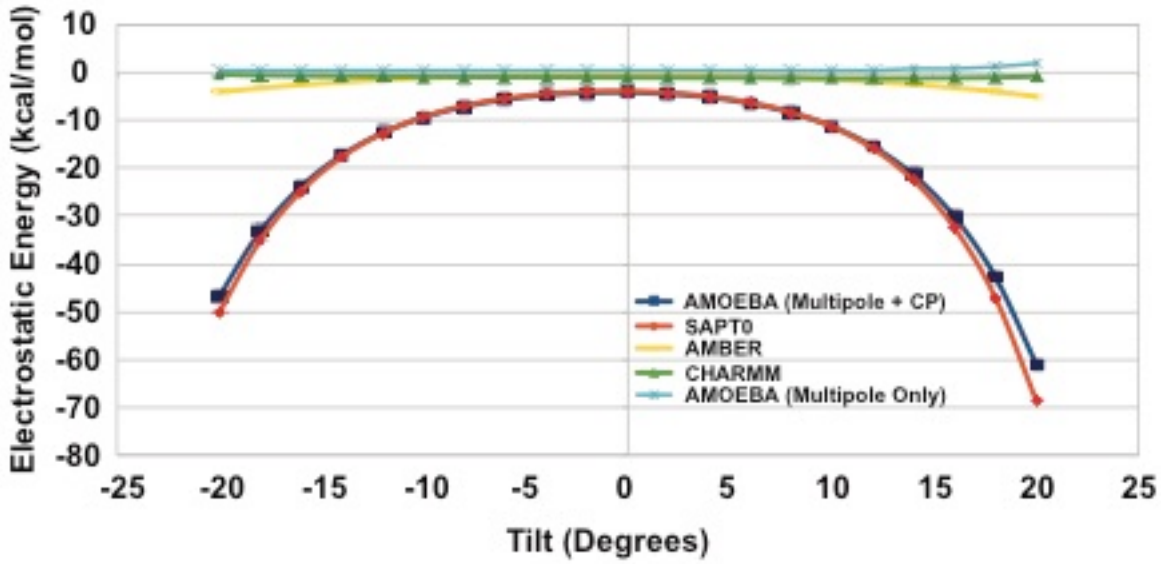


Figure 2.18 Electrostatic energy of a stacked TA:TA interaction vs. Tilt.

Including charge penetration reproduces the *ab initio* SAPT electrostatic energy over the range of tilt parameters. Tilt-like interactions are not part of the S101x7 database, so this behavior shows a level of transferability for the model.

Again, the model including charge penetration is the only model that agrees with the quantum mechanics. This same improvement persists across all structural parameters of the TA:TA base step. Figures for the other four parameters can be found in Appendix A. It is worth noting that not only is this an important test case because of its direct relation to biomolecular applications for the force field. It is also important because it shows that the model, parameterized against a particular test set (S101x7) performs well on interactions well outside of that set. These results give us confidence in the transferability of our charge penetration model.

2.7 Conclusions

The goal of the AMOEBA force field is to model the physics of biomolecular interactions using approximations that make calculations on large systems tractable. Our work here shows that to accurately capture the physics of short-range intermolecular interactions a charge penetration

term is absolutely necessary. Without accounting for charge penetration, even an advanced point multipole model cannot accurately reproduce electrostatic interactions at short range. These discrepancies in intermolecular interactions crucial to biomolecular systems are large enough that they cannot be ignored. Fortunately, we have also shown that charge penetration can be corrected for with the implementation of a simple set of damping functions. This is not necessarily a new conclusion. Previous work on AMOEBA as well other classical force field models have demonstrated the efficacy of using damping functions to capture charge penetration. We have demonstrated here that the higher-order damping functions we have developed for model 1 represent the best, most integrated method for implementing charge penetration in the AMOEBA force field.

There are some key reasons why using model 1 with higher-order damping makes the most sense for AMOEBA. The first reason is the most obvious. On an extensive test set of relevant molecular dimers, model 1 with higher-order damping produced the most accurate results. We have shown that including higher-order damping provides a substantial increase in model accuracy and model 1 performs well at this purpose. The practical purpose of including charge penetration in the force field is to accurately describe intermolecular interactions and by this direct measure model 1 with higher-order damping does the best.

The model does more than simply give good numbers, however. Model 1 is derived from the fundamental physics of atomic charge distributions. The damping function that describes the electrostatic potential around an atom in this model comes directly from the charge distribution of a hydrogen-like atom. The overlap damping function comes directly from an approximation of the overlap integral between two hydrogen-like charge densities. The model does contain empirical parameters, but those parameters are given physical meaning by the derived functions they sit in.

A natural question is why the similar model 2 with one extra parameter does not give better results than model 1. The simple answer is that it appears the two models are intrinsically aligned with different multipole models. AMOEBA takes a two-step approach to assigning multipole parameters. First distributed multipole analysis (DMA) is performed to obtain initial charge, dipole and quadrupole parameters. Then, those parameters are optimized by fitting to the electrostatic potential on a grid of points around the molecule. Because the overlap function in model 1 is constructed starting from a simple one-electron potential, model 1 seems to align nicely with the electrostatic potential fit method for determining AMOEBA multipoles. In contrast it seems that the two-center integral method used by model 2 might perform better with multipoles that are not potential-fitted. This theory is borne out by the results of figure 2.19. Figure 2.19 illustrates that model 2 with its extra free parameter, does perform better on the S101x7 database when simple DMA multipoles are used instead of potential fitted ones.

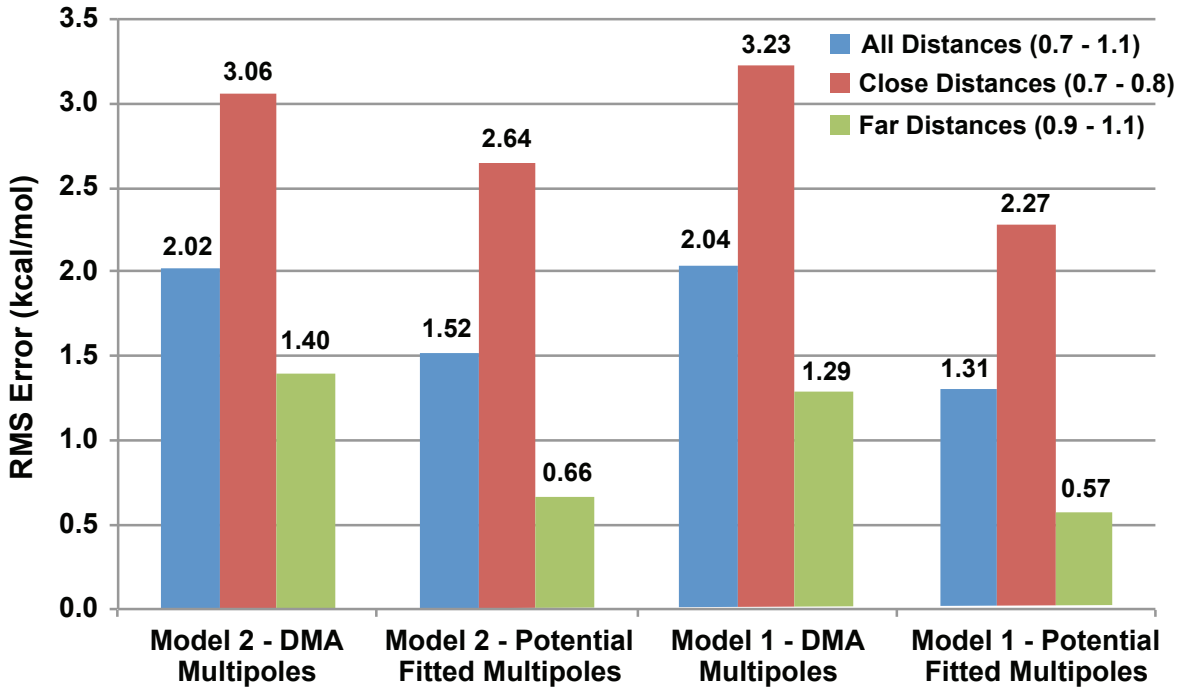


Figure 2.19 Charge penetration model agreement with AMOEBA potential-fit multipole model.

Models 1 and 2 are fit to the S101x7 database using either DMA or potential-fit multipoles. RMS electrostatic energy error is plotted. Model 2 performs slightly better when DMA multipoles are used, but model 1 with potential-fit multipoles gives the best overall fit.

Using the AMOEBA potential fitted multipoles however does better overall and much better when paired with model 1. The origin of this difference between models 1 and 2 in instructive. It shows that despite its relative simplicity, model 1 seems to provide a better intrinsic fit for the AMOEBA force field.

Not only is the model conceptually aligned with the AMOEBA multipole model, but it is fully integrated with it as well. Prior charge penetration models have damped charge-charge interactions or a handful of higher order interactions[13,14], but here we have derived damping functions for multipole interactions up to arbitrary order. This does two important things. First, it improves the overall accuracy of our intermolecular electrostatic energies. And second, it gives us a fully integrated multipole electrostatic–charge penetration model. The charge, dipole, quadrupole moments of a multipole expansion are all functions of the underlying charge density distribution. Thus, every interaction of these moments should be damped by the function that describes that charge density. Our higher-order charge penetration model satisfies this requirement and does so in a simple, straightforward way.

Importantly, the charge penetration model doesn't just fit one set of data. We have demonstrated that it passes multiple validation tests. First, the model proved to be robust. There is no numerical instability and the parameters are not overly sensitive. Second, the model is independent of the multipole model. This means that even if a slightly different set of multipole moments that fit the electrostatic potential are chosen for a given molecule, our charge penetration model will still give the same improvement in the fit. These validation tests indicate not only that

our model is viable, but that it is not beholden to the test set or the multipole model. In addition, we have shown that our charge penetration model has some measure of predictive power. On the biologically significant test of electrostatics in nucleic acid base stacking, our charge penetration model accurately predicted the electrostatic energies of base stacking over a wide range of non-equilibrium structural parameters. This result displays the promise this model shows in its application to simulations of real biological systems.

Finally, our higher-order charge penetration model captures a real physical effect. The charge penetration phenomenon is a direct result of the fact that atoms have charge distributions representing their electron densities. We have shown that our charge penetration function captures exactly this physics. When we use our model to fit the electrostatic potential on a grid of point surrounding a molecule, the error in the electrostatic fit from the simple point multipole approximation goes down for every tested case. This gives us the highest degree of certainty that we are doing more than just adding in another degree of freedom to our electrostatic function. The damping functions derived for our higher-order damping model accurately describe the electrostatic environment around molecules, and since the effect is necessarily short-range, the computational cost of accounting for charge penetration in this way is minimal. The damping terms can be implemented utilizing a short-range cutoff or can be computed for every pairwise interaction in the real-space portion of an Ewald summation approach. In either case, the additional cost beyond that of the standard AMOEBA electrostatic model is small. By describing this simple physics in a simple way, our model allows us to more accurately predict intermolecular interactions between biomolecules.

2.8 Further Work

The work in this chapter demonstrated that model 1 is an effective and efficient way to capture the electrostatic interactions between molecules at short range. For the HIPPO force field, however, one small change was made. Rather than use the damping function proposed in equation 2.7, we choose the similar function,

$$f_{damp}(r) = 1 - \left(1 + \frac{1}{2}\alpha r\right) e^{-\alpha r}, \tag{2.16}$$

where the only difference is the addition of the polynomial prefactor. This function meets all of the criteria required by section 2.2 and the remainder of the derivation in the text is identical.

There are two conceptual reasons why equation 2.16 was chosen over the original damping function. The first is that it gives a model for the electrostatic potential that is more closely aligned with the hydrogen-like atom. Inspection of equation 2.16 compared to equation 2.6 shows the clear similarity. This similarity is important because it affects the density which forms the bedrock for the rest of the model. While equation 2.7, *vis a vis* Poisson's equation, represents the potential due to the density,

$$\rho(r) = \frac{q\alpha^2}{4\pi r} e^{-\alpha r}\,, \tag{2.17}$$

equation 2.16 corresponds to the density

$$\rho(r) = \frac{q\alpha^3}{8\pi} e^{-\alpha r}. \tag{2.18}$$

As will be apparent in Chapter 5, for the HIPPO repulsion model we will need terms of the type,

$$\int \Phi_A \Phi_B dv \tag{2.19}$$

where Φ_A and Φ_B are related to the square roots of the densities on interacting atoms A and B respectively. It is a subtle difference, but the integrals of the type $\int e^{-\alpha_A r_A - \alpha_B r_B} dv$ that come

from the equation 2.18 definition are analytically solvable. Integrals of the type $\int \frac{1}{\sqrt{r_A}} e^{-\alpha_A r_A} \frac{1}{\sqrt{r_B}} e^{-\alpha_B r_B} dv$, corresponding to the equation 2.17 definition, however, are not. This makes equation 2.16 a more natural choice as the groundwork for the rest of the HIPPO model.

In fact, this alternate model was proposed previously by Slipchenko and Gordon. They showed that the differences between the two definitions are small, but the description the electrostatic potential around an atom is slightly better with the alternate model.[13] My work also shows that the differences in how well the model fits the SAPT electrostatic data are small. Using the equation 2.16 definition, the RMS error on the S101x7 database is 1.1 kcal/mol as compared to 1.3 kcal/mol for the definition in this published work. Because it was found that a larger number of atom classes were needed to accurately describe other energy components in the model, these classes were likewise used in this parameterization of electrostatics. The classes and parameters for the updated model are shown in table 2.4.

	Class	α (ang^{-1})		Class	α (ang^{-1})
1	H (nonpolar)	4.2097	15	N (sp2)	3.9413
2	H (nonpolar, Alkane)	4.3225	16	N (aromatic)	3.9434
3	H (polar, NH/N aromatic)	5.5155	17	O (sp3, hydroxyl, water)	4.7004
4	H (polar, OH)	4.7441	18	O (sp2, carbonyl)	4.2263
5	H (aromatic, CH)	4.953	19	O (O− in AcO−)	4.0355
6	H (polar, SH)	4.3952	20	O (O− in HPO42−)	4.4574
7	C (sp3)	4.2998	21	O (O− in H2PO4−)	4.5154
8	C (sp3, Alkane)	4.5439	22	O (O in H3PO4)	4.3312
9	C (sp2, Ethene)	3.5491	23	P (phosphate)	2.813
10	C (sp2, CO)	5.9682	24	S (sulfide, RSH)	3.362
11	C (sp)	1000	25	S (sulfur IV, DMSO)	2.7272
12	C (aromatic, CC)	3.8056	26	F (organofluorine)	5.508
13	C (aromatic, CX)	3.8066	27	Cl (organochloride)	3.6316
14	N (sp3)	3.9882	28	Br (organobromine)	3.2008

Table 2.4 Atom classes and parameters for the updated electrostatics (charge penetration) model.

Lastly, to avoid problems for larger atoms, the division of core vs. valence electrons was also changed. Rather than treat all electrons as part of the model charge density, only the valence electrons are included. The remainder are treated as part of the point positive core charge. This change is also reflected in the parameters listed in table 2.4.

2.8 References

1 Ren, P. & Ponder, J. W. Polarizable Atomic Multipole Water Model for Molecular Mechanics Simulation. *The Journal of Physical Chemistry B* **107**, 5933-5947, doi:10.1021/jp027815+ (2003).

2 Ren, P., Wu, C. & Ponder, J. W. Polarizable Atomic Multipole-Based Molecular Mechanics for Organic Molecules. *Journal of Chemical Theory and Computation* **7**, 3143-3161, doi:10.1021/ct200304d (2011).

3 Stone, A. J. & Alderton, M. Distributed multipole analysis. *Molecular Physics* **56**, 1047-1064, doi:10.1080/00268978500102891 (1985).

4 Ponder, J. W. *et al.* Current Status of the AMOEBA Polarizable Force Field. *The Journal of Physical Chemistry B* **114**, 2549-2564, doi:10.1021/jp910674d (2010).

5 Shi, Y. *et al.* Polarizable Atomic Multipole-Based AMOEBA Force Field for Proteins. *Journal of Chemical Theory and Computation* **9**, 4046-4063, doi:10.1021/ct4003702 (2013).

6 Sinnokrot, M. O. & Sherrill, C. D. Highly accurate coupled cluster potential energy curves for the benzene dimer: sandwich, T-shaped, and parallel-displaced configurations. *The Journal of Physical Chemistry A* **108**, 10200-10207 (2004).

7 Sherrill, C. D. *et al.* Assessment of Standard Force Field Models Against High-Quality ab Initio Potential Curves for Prototypes of Pi-Pi, CH/Pi, and SH/Pi Interactions. *J. Comput. Chem.* **30**, 2187-2193 (2009).

8 Parker, T. M. & Sherrill, C. D. Assessment of Empirical Models versus High-Accuracy Ab Initio Methods for Nucleobase Stacking: Evaluating the Importance of Charge Penetration. *Journal of Chemical Theory and Computation* **11**, 4197-4204 (2015).

9 Piquemal, J.-P., Cisneros, G. A., Reinhardt, P., Gresh, N. & Darden, T. A. Towards a Force Field Based on Density Fitting. *J. Chem. Phys.* **124**, 104101 (2006).

10 Cisneros, G. A. *et al.* Simple formulas for improved point-charge electrostatics in classical force fields and hybrid quantum mechanical/molecular mechanical embedding. *International Journal of Quantum Chemistry* **108**, 1905-1912, doi:Doi 10.1002/Qua.21675 (2008).

11 Freitag, M. A., Gordon, M. S., Jensen, J. H. & Stevens, W. J. Evaluation of charge penetration between distributed multipolar expansions. *Journal of Chemical Physics* **112**, 7300-7306, doi:Doi 10.1063/1.481370 (2000).

12 Piquemal, J. P., Gresh, N. & Giessner-Prettre, C. Improved formulas for the calculation of the electrostatic contribution to the intermolecular interaction energy from multipolar

expansion of the electronic distribution. *Journal of Physical Chemistry A* **107**, 10353-10359, doi:Doi 10.1021/Jp035748t (2003).
13 Slipchenko, L. V. & Gordon, M. S. Electrostatic energy in the effective fragment potential method: Theory and application to benzene dimer. *Journal of computational chemistry* **28**, 276-291 (2007).
14 Slipchenko, L. V. & Gordon, M. S. Damping functions in the effective fragment potential method. *Molecular Physics* **107**, 999-1016 (2009).
15 Spackman, M. A. The use of the promolecular charge density to approximate the penetration contribution to intermolecular electrostatic energies. *Chemical Physics Letters* **418**, 158-162, doi:Doi 10.1016/J.Cplett.2005.10.103 (2006).
16 Stone, A. J. Electrostatic Damping Functions and the Penetration Energy. *Journal of Physical Chemistry A* **115**, 7017-7027, doi:Doi 10.1021/Jp112251z (2011).
17 Tafipolsky, M. & Engels, B. Accurate Intermolecular Potentials with Physically Grounded Electrostatics. *Journal of Chemical Theory and Computation* **7**, 1791-1803, doi:Doi 10.1021/Ct200185h (2011).
18 Wang, B. & Truhlar, D. G. Including Charge Penetration Effects in Molecular Modeling. *Journal of Chemical Theory and Computation* **6**, 3330-3342, doi:Doi 10.1021/Ct1003862 (2010).
19 Wang, B. & Truhlar, D. G. Partial Atomic Charges and Screened Charge Models of the Electrostatic Potential. *Journal of Chemical Theory and Computation* **8**, 1989-1998, doi:Doi 10.1021/Ct2009285 (2012).
20 Wang, B. & Truhlar, D. G. Screened Electrostatic Interactions in Molecular Mechanics. *Journal of Chemical Theory and Computation* **10**, 4480-4487, doi:10.1021/ct5005142 (2014).
21 Cisneros, G. A., Piquemal, J.-P. & Darden, T. A. Generalization of the Gaussian Electrostatic Model: Extension to Arbitrary Angular Momentum, Distributed Multipoles and Speedup with Reciprocal Space Methods. *J. Chem. Phys.* **125**, 184101 (2006).
22 Duke, R. E., Starovoytox, O. N., Piquemal, J.-P. & Cisneros, G. A. GEM*: A Molecular Electronic Density-Based Force Field for Molecular Dynamics Simulations. *J. Chem. Theory Comput.* **10**, 1361-1365 (2014).
23 Wang, Q. *et al.* A General Model for Treating Short-Range Electrostatic Penetration in a Molecular Mechanics Force Field. *Journal of Chemical Theory and Computation* (2015).
24 Narth, C. *et al.* Scalable Improvement of SPME Multipolar Electrostatics in Anisotropic Polarizable Molecular Mechanics Using a General Short-Range Penetration Correction up to Quadrupoles. *J. Comput. Chem.* **37**, 494-506 (2016).
25 Stone, A. J. *In The Theory of Intermolecular Forces*. 94 (Oxford University Press Inc., 1996).
26 Coulson, C. in *Mathematical Proceedings of the Cambridge Philosophical Society.* 210-223 (Cambridge Univ Press).
27 Rezac, J. & Hobza, P. Extrapolation and Scaling of the DFT-SAPT Interaction Energies toward the Basis Set Limit. *Journal of Chemical Theory and Computation* **7**, 685-689, doi:Doi 10.1021/Ct200005p (2011).
28 Jeziorski, B., Moszynski, R. & Szalewicz, K. Perturbation Theory Approach to Intermolecular Potential Energy Surfaces of van der Waals Complexes. *Chemical Reviews* **94**, 1887-1930, doi:10.1021/cr00031a008 (1994).

29 Hohenstein, E. G. & Sherrill, C. D. Density fitting of intramonomer correlation effects in symmetry-adapted perturbation theory. *The Journal of Chemical Physics* **133**, 014101-, doi:doi:http://dx.doi.org/10.1063/1.3451077 (2010).
30 Parker, T. M., Burns, L. A., Parrish, R. M., Ryno, A. G. & Sherrill, C. D. Levels of symmetry adapted perturbation theory (SAPT). I. Efficiency and performance for interaction energies. *Journal of Chemical Physics* **140**, 094106 (2014).
31 Cornell, W. D. *et al.* A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **117**, 5179-5197 (1995).
32 Wang, J., Cieplak, P. & Kollman, P. A. How Well Does a Restrained Electrostatic Potential (RESP) Model Perform in Calculating Conformational Energies of Organic and Biological Molecules? *J. Comput. Chem.* **21**, 1049-1074 (2000).
33 Foloppe, N. & MacKerell Jr, A. D. All-atom empirical force field for nucleic acids: I. Parameter optimization based on small molecule and condensed phase macromolecular target data. *Journal of computational chemistry* **21**, 86-104 (2000).
34 El Hassan, M. A. & Calladine, C. R. The Assessment of the Geometry of Dinucleotide Steps in Double-Helical DNA: A New Local Calculation Scheme. *J. Mol. Biol.* **251**, 648-664 (1995).

Chapter 3: Induction (Polarization)

From the perspective of perturbation theory, polarization is the next logical step in building a classical force field. After the first order electrostatics, the dominant part of the second order in RS perturbation theory is the polarization component. SAPT uses the term “induction” for this term. I will explain the subtle, but important differences between the two in this chapter. The two, however, are conceptually similar. The basic idea is that the electronic structure of molecules should be able to respond to changes in their environment. In this chapter I will start by explaining why this is important and then proceed to lay out the HIPPO induction model.

3.1 Introduction

One defining characteristic of biomolecular systems is their heterogeneity. Proteins are composed of a wide range of elements in a variety of chemistries. Most important phenomena occur in water which has a unique set of molecular properties. Ions are known to play large parts in many biomolecular interactions. This diversity of molecules along with the fact that most biology occurs in the liquid phase means that often molecules in a given simulation are encountering a number of different environments. The way to capture the effect of this heterogeneity is by including polarization in the potential energy function. Polarization, simply put, allows atoms (and the molecules that they belong to) to respond to their environment. A simple, broad example illustrates how important including this feature is. The dielectric constant of water is 80, while the dielectric constant of the interior of a protein is 2-5. If one wishes to model the behavior of a molecule (a protein-binding drug molecule, for instance) equally well in both environments, one must include some function for polarization.

Despite the fact that polarization is not included in the standard point charge force field model, evidence is mounting that a large number of biomolecular simulation applications need

polarization in order to be accurate. Ion channel simulations, for instance, are thought to require polarization to accurately reproduce the flux of ions through a membrane.[1] Polarization is necessary to reproduce many properties of DNA and RNA.[2] The charged nature of lipid headgroups means that polarization is likely required to reproduce membrane - solvent/protein interfaces.[3] And polarization is known to be important for accurate simulations of charged species, such as ionic liquids.[4] This non-exhaustive list of applications underscores the importance of polarization in biomolecular simulations. In fact, these reasons lie at the heart of why the AMOEBA polarizable force field was first developed.

HIPPO, which is a direct descendant of AMOEBA, carries forward with this emphasis on polarization. As I will show in this chapter, it uses the same point inducible dipole formalism as its predecessor, but it uses its definition of a model charge density to produce an even more accurate model.

3.1.1 Overview of Existing Models

There are a variety of ways in which a given model can incorporate polarization. For a comprehensive overview of methods see references 3 and 5. Briefly, there are three different classes of methods of incorporating polarization. The first class is fluctuating charge methods. In these models, the charge of each atom in a simulation is changed slightly according to its environment. The total charge of the system remains constant, but the distribution of charges among atoms changes. Examples are the charge equilibration model, CHEQ, and charge-transfer type models.[6,7] The second class of methods is Drude oscillator-based approaches. In this method a Drude oscillator consisting of a negatively charged particle attached to a spring is affixed to each atom in the system. These Drude particles are free to move according to the electric field they experience. An example is the CHARMM (Chemistry at Harvard Molecular Mechanics)

Drude force field.[8] The final class of methods for including polarization is the point inducible dipole models. This can be thought of as the analytic equivalent of the Drude oscillator approach. Rather than a charge on a spring, each atom has point dipole that is induced by the electric field it experiences. Examples of this type of model for polarization are the Amber ff02, MPID (Multipole and Induced Dipole) and AMOEBA force fields.[9-11]

All three classes give many-body effects, the primary purpose of any polarization model. This means that the total energy of a system is no longer trivially equal to the sum of all pairs of interactions. Accurately reproducing these many-body effects is the standard by which polarizable models should be judged. Interestingly, recent work has shown that these effects can be more or less equally well captured by either fluctuating charge or induced dipole models.[12] Additionally, work on the MPID force field has shown that there exists a direct mapping from Drude models to point inducible dipole models.[11] Because there are potential numerical and practical problems with fluctuating charge and Drude models, and with these equivalencies in mind, I choose to build HIPPO with a point inducible dipole model.

3.1.2 Induction vs. Polarization

Up until this point, I have used the terms induction and polarization relatively interchangeably. As was stated before, this is because conceptually they are not that different; both describe a molecule or atom responding to its environment. However, there is a useful distinction to be drawn between these two terms. In SAPT, the induction energy of a dimer pair is the energy associated with the rearrangement of monomer A's electrons in the presence of the electric field of monomer B. Qualitatively, there are two effects that can happen in this interaction. The first is classical polarization. In this effect, the electrons of monomer A shift slightly in response to the influence of monomer B, but remain attached to monomer A. The

second is charge transfer. In this effect, some fraction of monomer A's electrons make the jump from monomer A's electron density over to monomer B's electron density. In this way these electrons (or often just a fraction of an electron) can now be qualitatively assigned to monomer B.

This distinction between the two effects defines the difference between polarization and induction. Induction is the entire interaction (polarization plus charge transfer), whereas polarization is only the first response component. As described here, this distinction is admittedly mathematically ill-defined. However, in section 3.2, I will describe quantitatively how this separation is made.

3.1.3 Overview of HIPPO Induction Model

The HIPPO induction model has two parts. The first is a many-body, point inducible dipole model for polarization and the second is a pairwise exponential function for charge transfer. Both parts contain novel elements for biomolecular force fields.

In the polarization component, HIPPO replaces the previous, empirical methods of the AMOEBA model with a more physics-based approach. AMOEBA utilized the method of Thole to damp dipole-dipole interactions at short range so as to avoid so-called "polarization catastrophe".[13,14] This model has been shown to be effective, but lacks a physical rationale for the damping function. HIPPO replaces this damping function with a damping function drawn directly from the electrostatic work presented in Chapter 2. I will show that in addition to being physically motivated, this model produces more accurate molecular polarizabilities and many-body energies than the previous Thole model.

The HIPPO induction model also includes a function to describe short-range charge transfer. As I will show in section 3.4, a simple exponential function is capable of capturing this effect quite accurately. Moreover, I will show why the pairwise approximation is a good one for this term. Taken together with the polarization function, the total induction model is capable of very close agreement with SAPT induction results.

3.2 Theory

The HIPPO induction model consists of two parts: polarization and charge transfer. Because these effects are qualitatively different, I will derive the functions for each independently. The first section will describe the polarization, which is responsible for the entirety of many-body effects in the HIPPO force field and the second will describe the simpler charge transfer function.

3.2.1 HIPPO Polarization Derivation

The polarization model of HIPPO is a point inducible dipole model. Each atom in a given system has an inducible dipole determined by:

$$\vec{\mu} = \alpha \vec{E} \tag{3.1}$$

where μ is the induced dipole, α is the polarizability, and E is the electric field at the site. The key to solving this equation to find the induced dipoles is in the electric field. The field at a given site is:

$$\vec{E}_{total} = \vec{E}_{perm} + \vec{E}_{induced} \,. \tag{3.2}$$

where the permanent and induced fields are separated for clarity.

The term E_{perm} in equation 3.2 represents the electric field due to the permanent moments on all the other sites,

$$\vec{E}_{perm}(R_i) = \sum_{j \neq i} Z_j \nabla \mathrm{T} + \vec{M}_\mathrm{j} \nabla \mathbf{T}^* \tag{3.3}$$

where Z represents the core charge and M represents the vector of the multipole components (charge, dipole and quadrupole),

$$\vec{M} = \left(Q, \left[\mu_x, \mu_y, \mu_z\right], \begin{bmatrix} \Theta_{\mathrm{xx}} & \Theta_{\mathrm{xy}} & \Theta_{\mathrm{xz}} \\ \Theta_{\mathrm{yx}} & \Theta_{\mathrm{yy}} & \Theta_{\mathrm{yz}} \\ \Theta_{\mathrm{zx}} & \Theta_{\mathrm{zy}} & \Theta_{\mathrm{zz}} \end{bmatrix} \right). \tag{3.4}$$

The interaction matrices, T and $\mathbf{T}^*$ are taken directly from the electrostatics function defined in Chapter 2.

$$\begin{aligned} T &= \frac{1}{r_{ij}} \\ \boldsymbol{T}^* &= \begin{bmatrix} 1 & \nabla & \nabla^2 \end{bmatrix} \left(\frac{1}{r_{ij}} f_{ij}^{damp}(r_{ij}) \right) \end{aligned} \tag{3.5}$$

The damping function in equation 3.5 is identical to equation 2.16, defining the electrostatic potential in Chapter 2,

$$f_{ij}^{damp}(r_{ij}) = 1 - \left(1 + \frac{1}{2}\alpha_j r_{ij}\right) e^{-\alpha_j r_{ij}} \tag{3.6}$$

where the α parameter here is not to be confused with the polarizability. The interpretation for this definition of the permanent electrostatic field is simple. It is simply the gradient of the electrostatic potential of a set of charge densities defined in Chapter 2. As we will show in section 3.4, the alpha parameters that come from the electrostatics model (defined in table 2.3) can be used directly for this calculation.

The second part of the electric field at any given point comes from the induced dipoles themselves. $E_{induced}$ in equation 3.2 is defined similarly to E_{perm} as,

$$\vec{E}_{induced}(R_i) = \sum_{j \neq i} \vec{\mu}_j^{ind} \nabla^2 \left(\frac{1}{r_{ij}} f_{ij}^{overlap}(r_{ij}) \right) \tag{3.7}$$

where μ^{ind} represents the induced dipole at each site. In this case, rather than the one-center damping function that defines the permanent electrostatic potential and field, the two-center, overlap function

$$f_{ij}^{overlap}(r_{ij}) == \begin{cases} 1-\left(1+\frac{11}{16}\alpha r_{ij}+\frac{3}{16}(\alpha r_{ij})^2+\frac{1}{48}(\alpha r_{ij})^3\right)e^{-\alpha r_{ij}}, & \alpha_i=\alpha_j \\ 1-A^2\left(1+2B+\frac{\alpha_i}{2}r_{ij}\right)e^{-\alpha_i r_{ij}}-B^2\left(1+2A+\frac{\alpha_j}{2}r_{ij}\right)e^{-\alpha_j r_{ij}}, & \alpha_i\neq\alpha_j \end{cases} \quad (3.8)$$

is used. This is because each induced dipole is represented as part of the same density that defines the rest of the atom. Thus, the induced dipole – induced dipole interaction is an interaction between two interacting densities. Equation 3.8 again is identical to the overlap function set by the electrostatics model in the Chapter 2. Again, the physical rationale is almost trivial. Each induced dipole is represented as a density and the size of that density is set by the electrostatics model parameterized in Chapter 2.

Taken together, the damping functions in equations 3.6 and 3.8 serve a practical purpose in addition to satisfying simple physics arguments. The previous AMOEBA induced dipole model used an empirical damping method due to Thole for the permanent and induced fields. These damping functions effectively replace the old method with physical rationale. The Thole approach was designed to prevent polarization catastrophe, and as we will show in section 3.4, the HIPPO model achieves the same end with even greater accuracy.

Because both sides of equation 3.1 contain the induced dipoles, one must solve a system of linear equations in order to find the dipoles. To make this more readily apparent, we can re-write equation 3.1 in tensor form as,

$$\mathbf{T}\boldsymbol{\mu} = \mathbf{E}_{\mathbf{perm}}, \quad (3.9)$$

where $\mathbf{E}_{\mathbf{perm}}$ represents the permanent electrostatic field and,

$$\mathbf{T} = \alpha^{-1} - \mathcal{T}, \quad (3.10)$$

where $\mathcal{T}$ is the induced dipole – induced dipole interaction tensor,

$$\mathcal{T} = \nabla^2 \left(\frac{1}{r_{ij}} f_{ij}^{overlap}(r_{ij}) \right). \tag{3.11}$$

Using this notation, the energy due to the induced dipoles can be written as,

$$U_{pol} = \frac{1}{2} \boldsymbol{\mu}^T \mathbf{T} \boldsymbol{\mu} - \mathbf{E}_{\boldsymbol{perm}}^{\mathbf{T}} \boldsymbol{\mu}. \tag{3.12}$$

There are two different options for how to solve this system of equations to get the polarization energy in the HIPPO force field. The first is a variational, iterative method. In this method we define a residual,

$$\mathcal{R} \equiv \left(\frac{\mathrm{d}U}{\mathrm{d}\boldsymbol{\mu}} \right)^T = \mathbf{T}\boldsymbol{\mu} - \mathbf{E}_{\mathbf{perm}} \tag{3.13}$$

and minimize this with respect to the induced dipoles, ultimately requiring this to be zero. In practice this is done using an iterative preconditioned conjugant gradient solver. Convergence is declared once the RMS change in induced dipoles from one iteration to the next becomes lower than some threshold (typically 1×10^{-5} or 1×10^{-6} Debye). Because the residual is assumed to be very close to zero in this method, the total energy can be re-written in the simpler form:

$$U = -\frac{1}{2} \mathbf{E}_{\boldsymbol{perm}}^{\mathbf{T}} \boldsymbol{\mu}. \tag{3.14}$$

The other method that can be used for the HIPPO polarization model is the OPT method of Andrew Simmonett and co-workers.[15,16] In this method, rather than require a zero residual as with variational approaches, the polarization energy and force is calculated through a perturbation theory expansion in the induced dipoles. For a full description of this method, see references 15 and 16. In short, however, the induced dipoles are defined in a power series expansion as:

$$\boldsymbol{\mu_n} = \boldsymbol{\mu_{(0)}} + \lambda \boldsymbol{\mu_{(1)}} + \lambda^2 \boldsymbol{\mu_{(2)}} + \cdots \lambda^n \boldsymbol{\mu_{(n)}} \tag{3.15}$$

with the various orders,

$$\begin{aligned} \boldsymbol{\mu}_{(0)} &= \boldsymbol{\alpha}\mathbf{E}_{\mathbf{perm}}, \\ \boldsymbol{\mu}_{(1)} &= \boldsymbol{\alpha}\boldsymbol{\mathcal{T}}\boldsymbol{\alpha}\mathbf{E}_{\mathbf{perm}}, \\ \boldsymbol{\mu}_{(2)} &= \boldsymbol{\alpha}\boldsymbol{\mathcal{T}}\boldsymbol{\alpha}\boldsymbol{\mathcal{T}}\boldsymbol{\alpha}\mathbf{E}_{\mathbf{perm}}, \\ &\vdots \\ \boldsymbol{\mu}_{(n)} &= (\boldsymbol{\alpha}\boldsymbol{\mathcal{T}})^{\mathbf{n}}\boldsymbol{\alpha}\mathbf{E}_{\mathbf{perm}}. \end{aligned} \tag{3.16}$$

With this definition, the energy is:

$$U = -\frac{1}{2}\mathbf{E}_{\boldsymbol{perm}}^{\mathbf{T}}\boldsymbol{\mu}_{\boldsymbol{n}}. \tag{3.17}$$

Typically, the expansion is carried out to either 3rd or 4th order to match the fully converged result. This approach is usually faster to compute than fully converged variational methods. Work by Simmonett and co-workers has shown, however, that the induced dipoles from 3rd and 4th order expansions match the variational result for many systems quite well.

As with any force field, HIPPO utilizes a set of exclusion rules to determine which interactions in a large molecule are treated with intramolecular *vs.* intermolecular energy terms. For polarization these rules merit a brief discussion. There are two sets of rules. One that applies just to induced dipole – induced dipole interactions, and one that applies to induced dipole interactions – permanent moment interactions.

The induced dipole – induced dipole exclusion rules are the simpler of the two. These are based on connectivity. For all atoms that are separated by one bond (so-called 1-2 interactions), the induced dipole – induced dipole interaction is scaled by 0.2. For all other interactions (1-3, 1-4, 1-5, etc.) the full interaction is used. This effectively means that for things that are bonded, HIPPO is counting on the bond energy term to pick up most of the change in energy as the distance between 1-2 atoms changes. Practically, this scaling must be done to avoid polarization catastrophe due to atoms that are very close polarizing each other.

The induced dipole – permanent multipole exclusion rules are slightly more nuanced. The AMOEBA polarization model introduced the concept of "polarization groups". These, as illustrated in figure 3.1 are groups of atoms that form a cohesive chemical group.

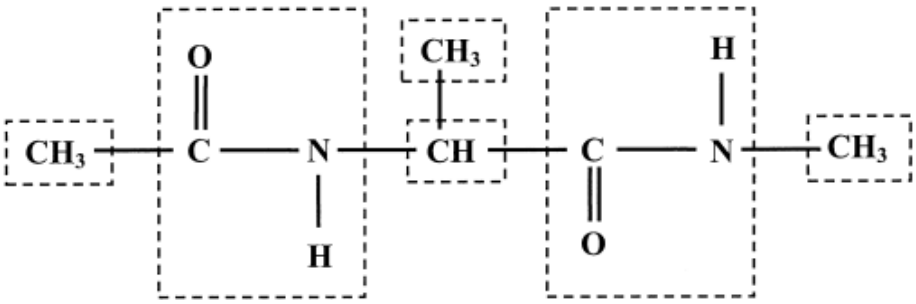


Figure 3.1 Example of polarization group scheme.

These groups represent the division of the AMOEBA force field polarization model for alanine dipeptide. Reproduced from reference 17.

Under this scheme, AMOEBA performed the calculations to obtain the induced dipoles with group-based exclusion rules. All intragroup induced dipole – permanent multipole interactions were scaled by 0 (completely excluded) and all intergroup induced dipole – permanent multipole interactions were scaled by 1 (completely included). (AMOEBA has no exclusion rules for induced dipole – induced dipole interactions; everything is completely included.) However, once the induced dipoles are obtained, a different set of induced dipole – permanent multipole scaling rules is used to compute the AMOEBA polarization energy. These rules are connectivity-based. Under these rules 1-2 interactions are scaled by 0, 1-3 interactions are scaled by 0 and 1-4 and longer interactions are scaled by 1. The one wrinkle is that in the case of a 1-4 interaction where both atoms happen to be in the same group, the scale factor is changed to 0.5 instead of 1. Having these dual sets of exclusion rules (one for calculating the dipoles, and one for calculating

the energy) leads to some complications that I have tried to avoid in the HIPPO polarization model.

The HIPPO polarization model uses a hybrid group-connectivity framework for the polarization exclusion rules. Pengyu Ren showed that using polarization groups can lead to good transferability of a polarization model.[17] The HIPPO model retains the polarization group mechanism, but discards the unwieldy elements of having two separate sets of exclusion rules. Under this single set of rules, we still define polarization groups and then use those groups to define a single set of more nuanced exclusion rules, summarized in table 3.1

	Intra-group	Inter-group
1-2	0.0	0.5
1-3	0.0	0.5
1-4	0.5	1.0
1-5 and beyond	1.0	1.0

Table 3.1 HIPPO polarization model exclusion rules.

Although these rules have not been thoroughly tested, they give a very close facsimile of the original, well-tested AMOEBA rules without incurring the costs of the previous dual-exclusion method.

3.2.2 HIPPO Charge Transfer Derivation

The second part of the total HIPPO induction model is the charge transfer. For this portion of the induction energy, I choose to use a simple pairwise exponential functional form. As we will show in section 3.4, this is a good approximation for most organic systems. The functional form is based on the assumption laid out in section 3.1.2 that aside from the polarization, the other dominant effect in SAPT induction is charge hopping from an atom in monomer A to a different atom in monomer B.

Qualitatively, this charge transfer effect can be shown to be the effect of electron tunneling into the nuclear potential well of an opposing monomer.[18] Alston Misquitta showed that the amount of charge transferred is proportional to an exponential. Because the remainder of the induction energy for most intermolecular interactions at distances of consequence in biomolecular simulations is quite small, I choose a simple empirical model that captures this qualitative observation. The HIPPO charge transfer model is:

$$U_{ct} = \sum_{i \neq j} -A_i e^{-\eta_j r_{ij}} - A_j e^{-\eta_i r_{ij}} . \tag{3.18}$$

Here, A_i and A_j represent the maximum magnitude of charge that can be transferred from atom i and j, respectively, and η_i and η_j represent the exponentials of the nuclear wells of i and j, respectively. For a given pair, the interpretation of equation 3.18 is that the first term gives the energy of charge transferring from i to j and the second term gives the transfer in the opposite direction.

3.3 Methods

The general strategy I employed for parameterizing the HIPPO induction model falls into two parts. The first part is meant to establish the many-body polarization portion of the function. This is done first because of the number of direct connections with experimental observables that is possible for polarization. The second part is to fill in the difference with charge transfer. Although this is certainly an approximation and generalization, the functional form is flexible enough to handle this.

To parameterize the polarization model, I fit atomic polarizabilities to molecular polarizability data. Full anisotropic molecular polarizabilities for the 36 molecules in the S101 database were calculated using the Psi4 quantum chemistry program using density functional

theory (DFT) with the WB97XD functional and the aug-cc-pVTZ basis set. Atomic polarizabilities were constrained to the 28 classes defined in table 2.3 of Chapter 2. I used a least squares fitting program written in the Tinker molecular mechanics software package to optimize these 28 atomic polarizabilities.[19] Additionally, I calculated 3- and 4-body energies for a variety of molecular clusters. These calculations were performed at the MP2 level of theory with an aug-cc-pVTZ basis set.

After determining the atomic polarizabilities, I proceeded to fit the charge transfer function. I performed a log-weighted least squares fit that minimized the residual,

$$U_{ind}(SAPT) - \left(U_{pol}(HIPPO) + U_{ct}(HIPPO)\right) \tag{3.19}$$

on each dimer in the S101x7 database. In equation 3.19, U_{pol}(HIPPO) is calculated using the polarizabilities determined in the previous step. The parameters optimized in this fit are only the A and η parameters from equation 3.18. This least squares fitting program was also implemented in Tinker.

3.4 Results

3.4.1 Molecular Polarizabilities

Because SAPT does not discriminate between polarization and charge transfer type excitations, another source of data is needed for the parameterization of the HIPPO polarization model. Fortunately, there is an experimental observable quantity available for polarization: molecular polarizability. The first step is to calculate molecular polarizabilities for a set of molecules. For organic molecules, calculating molecular polarizabilities with high-level *ab initio* methods is known to be nearly identical to experimental polarizabilities and gives the full anisotropic polarizability tensor, which is sometimes unavailable from experimental data. I

calculated the molecular polarizabilities of every molecule in the S101 database. The results are listed in table 3.2.

Molecules	XX polarizability (ang^3)	YY polarizability (ang^3)	ZZ polarizability (ang^3)
Water	1.3853	1.431	1.4918
MeOH	2.9491	3.022	3.4493
MeNH2	3.5788	3.675	4.1802
Peptide	5.8740	7.657	9.1497
Uracil	6.1162	11.01	13.7840
Pyridine	6.0438	10.76	11.3683
AcOH	3.9460	5.478	5.8124
Benzene	6.6435	11.96	11.9693
Ethene	3.3782	3.734	5.2637
Pentane	8.4985	9.107	11.2340
Neopentane	9.5372	9.539	9.5431
Cyclopentane	7.8998	9.101	9.1027
CH3SH	4.9665	5.012	6.3118
DMSO	6.7630	8.23	8.4210
PO4H	9.0964	9.256	9.6770
PO4H2	6.5318	6.671	7.1143
PO4H3	5.5003	5.596	5.9249
BenF	6.5209	11.87	12.1878
BenCl	7.7548	12.97	16.3675
BenBr	8.5962	13.71	18.3138

MeNH3	2.8210	2.853	3.2479
ImidazolePlus	4.0514	7.109	7.4532
CN3H6	3.5660	5.917	5.9879
AcOHMinus	5.1132	7.392	7.6815
DimethylSulfide	6.2777	7.202	8.3401
Imidazole	5.0451	8.125	8.5001
Pyrrolidine	7.6730	8.494	8.8057
Phenol	6.9451	12.39	13.6796
Indole	8.8472	15.71	20.5356
AcNH2	4.4862	6.278	6.6882
CH2F2	2.4791	2.616	2.7829
CH2Cl2	5.2380	5.827	8.0503
CH2Br2	6.9463	7.577	11.1370
MeF	2.4490	2.449	2.6632
MeCl	3.8911	3.893	5.3231
MeBr	4.8358	4.839	6.6588

Table 3.2 S101 monomer molecular polarizabilities.

The advantage of using *ab initio* molecular polarizabilities is that it gives the anisotropic polarizability tensor. For roughly spherical molecules, like water, this doesn't matter much, but for molecules like benzene, table 3.2 shows how dramatically the polarizability can vary depending on the orientation. This orientation-dependent information is crucial to developing a polarization model.

With these molecular polarizabilities, the next step is to fit the HIPPO polarization model to reproduce them. 28 chemical classes (identical to those laid out in the "Further Work" section of Chapter 2) are assigned and then fit with a nonlinear least-squares routine. The optimization minimized the residual on each component of the molecular polarizability simultaneously. Because the DFT functional is known to have errors for phosphates, the PO_4H, PO_4H_2, and PO_4H_3 compounds were excluded from the fit. The parameters determined from this fit are presented in table 3.3.

class	name	Polarizability (ang^3)
1	H (nonpolar)	0.373
2	H (nonpolar, Alkane)	0.504
3	H (polar, NH/N aromatic)	0.005
4	H (polar, OH)	0.3698
5	H (aromatic, CH)	0.1106
6	H (polar, SH)	0.2093
7	C (sp3)	0.755
8	C (sp3, Alkane)	0.9354
9	C (sp2, Ethene)	1.9384
10	C (sp2, CO)	0.6577
11	C (sp)	n/a
12	C (aromatic, CC)	1.5624
13	C (aromatic, CX)	1.2811
14	N (sp3)	1.4289
15	N (sp2)	1.4545
16	N (aromatic)	1.3037
17	O (sp3, hydroxyl, water)	0.6645
18	O (sp2, carbonyl)	1.4266
19	O (O− in AcO−)	1.8809
20	O (O− in HPO42−)	n/a
21	O (O− in H2PO4−)	n/a
22	O (O in H3PO4)	n/a
23	P (phosphate)	n/a
24	S (sulfide, RSH)	3.1967
25	S (sulfur IV, DMSO)	2.458

26	F (organofluorine)	0.4717
27	Cl (organochloride)	2.366
28	Br (organobromine)	3.4458

Table 3.3 HIPPO atomic polarizabilities from fit to S101 molecular polarizabilities.

The parameters show clear periodic trends and largely agree with chemical intuition regarding the relative "sizes" of atoms in various chemical bonding environments.

The quality of the fit to the S101 molecular polarizabilities of the parameters shown in table 3.3 is plotted in figure 3.2.

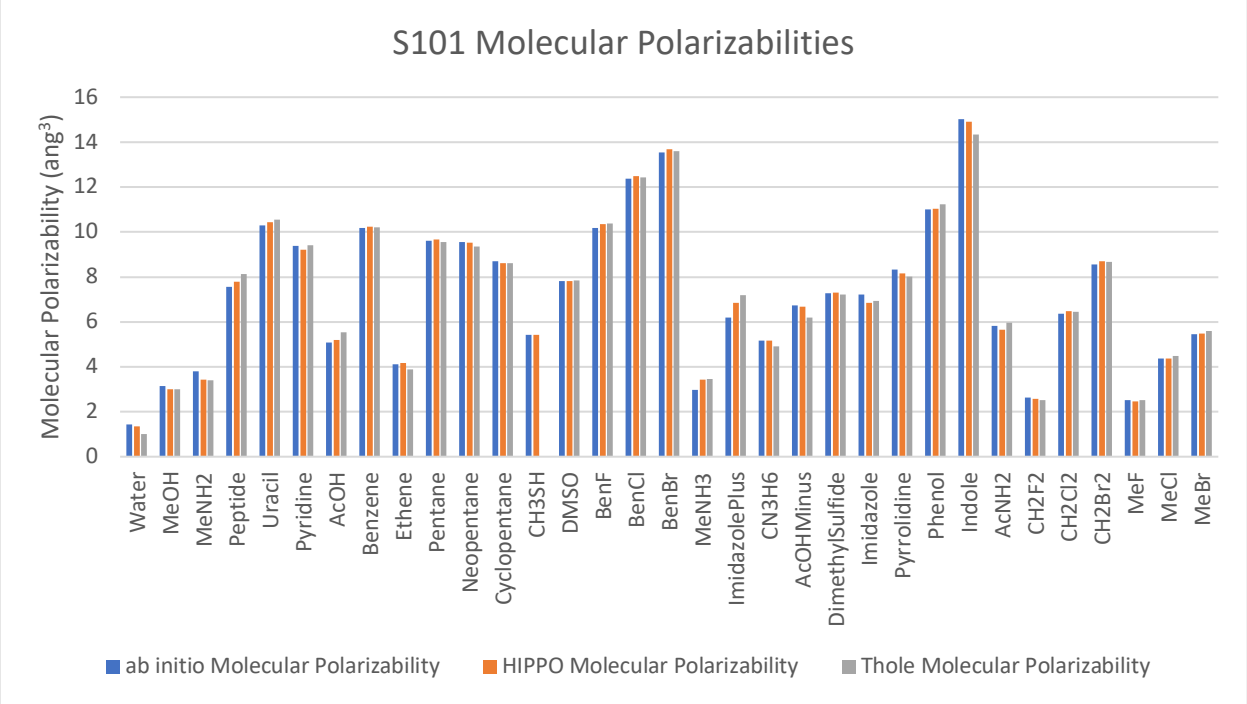


Figure 3.2 Molecular Polarizabilities of S101 Monomers.

Polarizabilities are the average of the three principal components of the molecular polarizability tensor.

The results in figure 3.2 show that the HIPPO model fits the molecular polarizabilities of a wide range of molecules very well. Figure 3.2 only shows the isotropic polarizability (average of the 3 principal components), but the RMS error for the dataset, presented in table 3.4 illustrates that

the fit is good across orientations as well. Figure 3.2 also includes the results of using the existing AMOEBA Thole model with the atomic polarizabilities allowed to vary across the same 28 classes as the HIPPO model. Clearly, the Thole model also fits the data well. However, as illustrated by the RMS error presented in table 3.4, the fit is slightly worse than the HIPPO model.

This gives us a high degree of confidence that using the physically motivated framework of the HIPPO model is not detrimental to the accuracy of the polarization component. In fact, it seems to perform slightly better.

	RMS Error (ang^3)
Thole	0.50
HIPPO	0.28

Table 3.4 RMS Error for Molecular Polarizabilities of S101 Monomers

One example will illustrate an important feature of the HIPPO polarization model. Plotted in figure 3.3 are the principal components of the molecular polarizability of benzene.

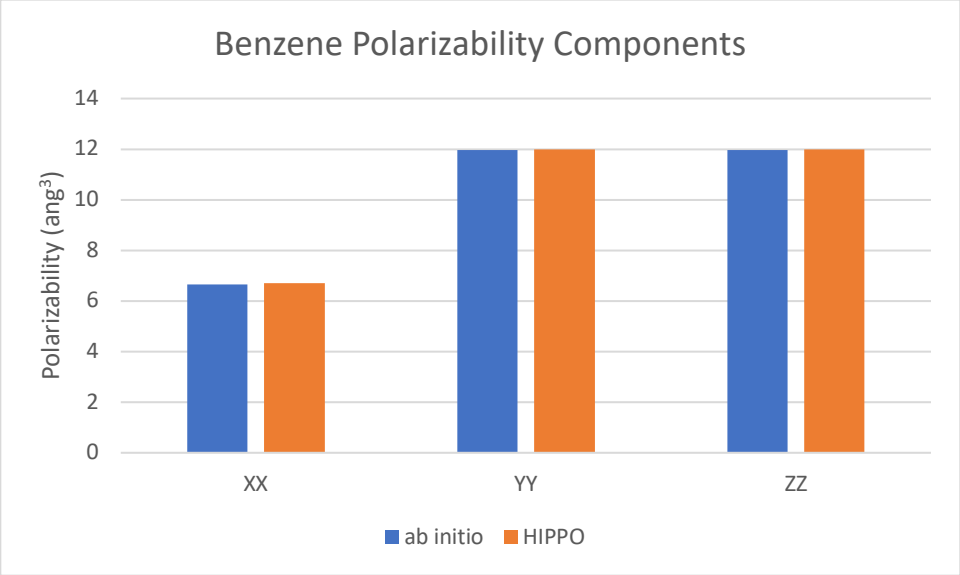


Figure 3.3 Benzene Polarizability Components.

The components in figure 3.3 shows not only how well the HIPPO model fits molecular polarizability, but that anisotropic atomic polarizabilities are not necessary for the model. Benzene has a large, almost 2-fold, difference between the in-plane (YY and ZZ) and out-of-plane (XX) components of the polarizability, but the simple isotropic polarizabilities of the HIPPO model are sufficient to capture this difference. This example holds for the anisotropic polarizabilities of other S101 molecules as well.

3.4.2 SAPT Induction vs. HIPPO Polarization

As was stated in the introduction, the HIPPO model draws a distinction between pure polarization, or the linear order response to an electric field, and the full induction energy. This assertion, however, that there is a meaningful difference between the two has not been backed up with data. To investigate this, I compared the polarization energies that come from the polarization model fit to molecular polarizabilities to SAPT induction energies. In this test, if the

two are close, a distinction between polarization and induction is unnecessary. If, however, they are not, it shows that there is a larger issue at play.

The start of this investigation is an example. I examined the polarization energy of the water dimer at the various distances included in the S101x7 database and compared these energies to the SAPT induction energies of the same structures. Shown in figure 3.4 is the result of that comparison.

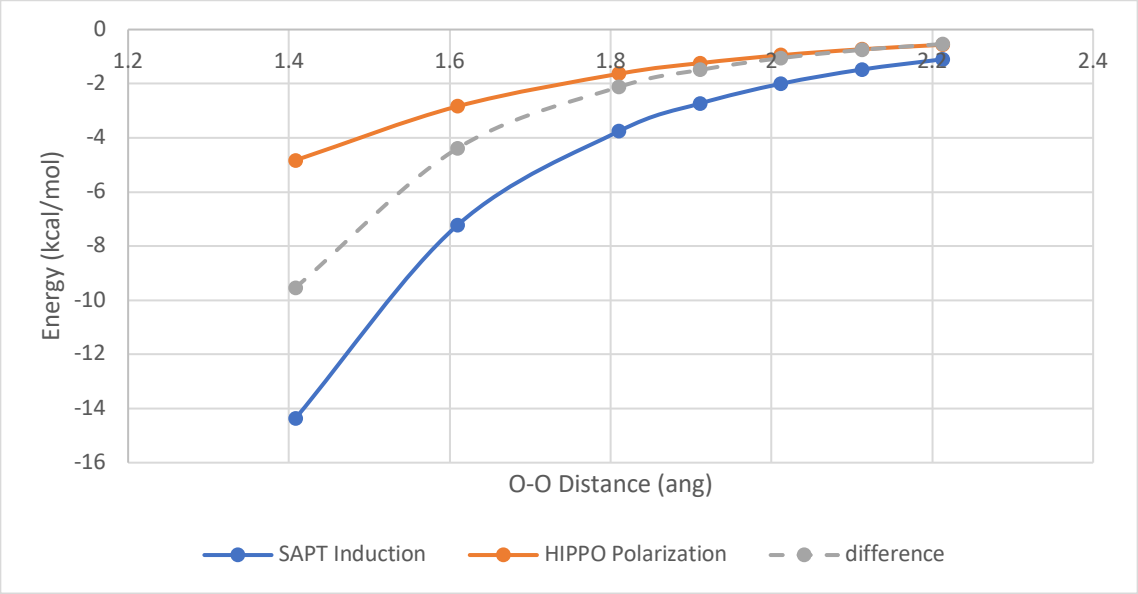


Figure 3.4 SAPT induction *vs.* HIPPO polarization energy for the water dimer dissociation curve.

The SAPT induction curve is clearly more attractive, especially at short range, than the HIPPO polarization curve. The dotted gray curve shows the difference (SAPT – HIPPO).

The HIPPO polarization curve in figure 3.4 clearly shows the canonical, roughly $1/r^3$,dipole – dipole interaction energy dependence (by construction). The SAPT curve, however, does not match the HIPPO curve. Moreover, the difference, also plotted in figure 3.4, reveals that it is not a simple scaling or additive factor between the two.

To rule out the possibility that the induced dipoles were wrong because the electric field is wrong, I plotted the electric field for the water dimer. Shown in figure 3.5 is the magnitude of the electric field due to the opposing monomer at each nuclear position.

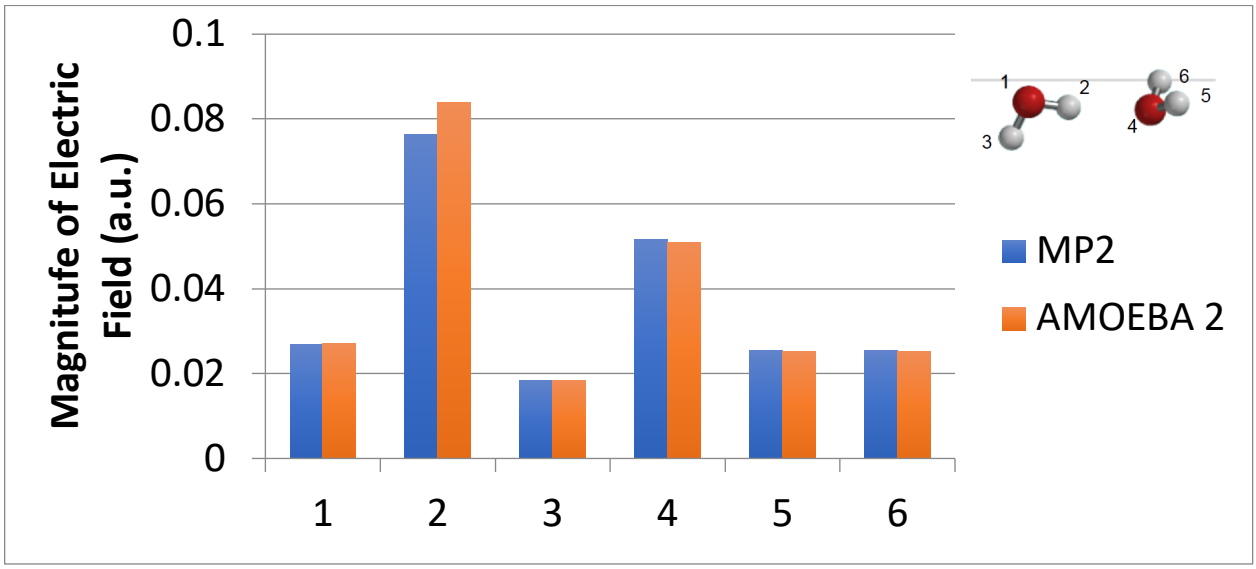


Figure 3.5 Electric field of the equilibrium water dimer.

"AMOEBA 2" references the working title of HIPPO. "MP2" references the level of *ab initio* theory used. The inset shows the numbering of nuclear coordinates. Electric fields are calculated from the static (unperturbed) electron density of opposing monomer.

Figure 3.5 shows that the electric field is not what causes the difference seen in figure 3.4. The electric field magnitudes (and directional components, not plotted) are very close to the *ab initio* result. This nearly conclusively shows that a simple, linear order polarization model is not capturing all of the physics included in the SAPT induction energy. The electric fields and polarizabilities are both accurate to within a few percent. There is nothing else that goes in to the polarization energy calculation. This must mean that the difference is coming from somewhere else.

3.4.3 Charge Transfer

In recent work Alston Misquitta has shown that the difference between the linear response polarization energy and the SAPT induction energy can be largely attributed to charge transfer.[18] According to this hypothesis the charge transfer energy should be an exponential function that described tunneling of electrons on one monomer into the nuclear wells of the other monomer. In order words, this describes electrons hopping from one molecule to another rather than simply shifting around on their parent molecule. To test this hypothesis, I attempted to fit the difference between the SAPT induction and HIPPO polarization energies with a simple exponential function (equation 3.18). The results for the water dimer are shown in figure 3.6.

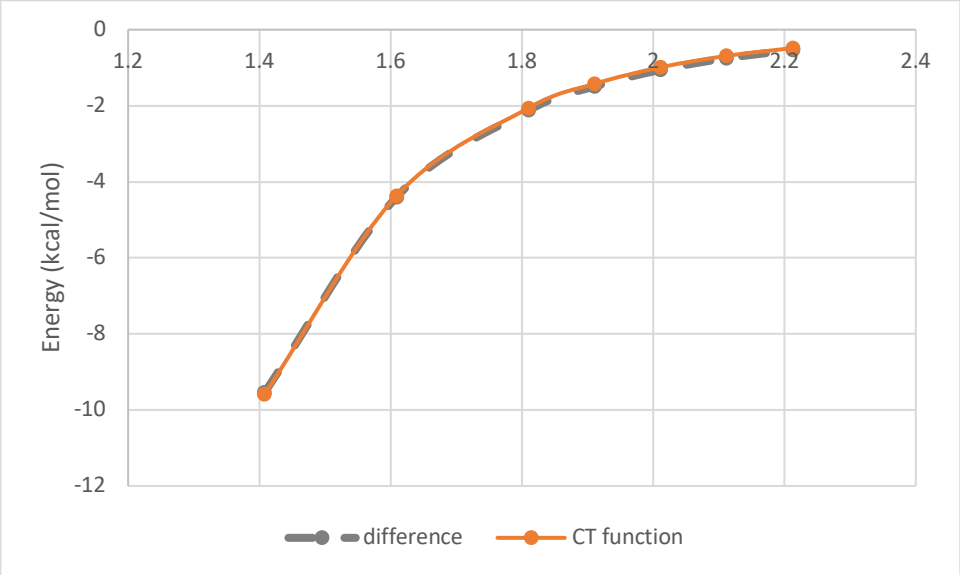


Figure 3.6 Charge transfer function fit to SAPT induction – HIPPO polarization difference.

The simple exponential function proposed matches the shape of the difference. CT function defined in equation 3.18.

Clearly, the difference between induction and polarization follows a straightforward exponential dependence. To assess if this relation holds more broadly, I performed a much more exhaustive fit to the entire S101x7 database. To do this, I calculated the polarization energy of each dimer

with the HIPPO polarization model and subtracted it from the SAPT induction energy. I then performed a log-weighted least squares fit, optimizing the 28 A and η parameters that define the charge transfer function. The results of this fit are shown in figure 3.7.

The trends in figure 3.7 confirm what was observed in the water dimer case study. The HIPPO polarization energy is uniformly underbound compared to the SAPT induction energy, and a simple pairwise exponential function can effectively fill the gap between the two.

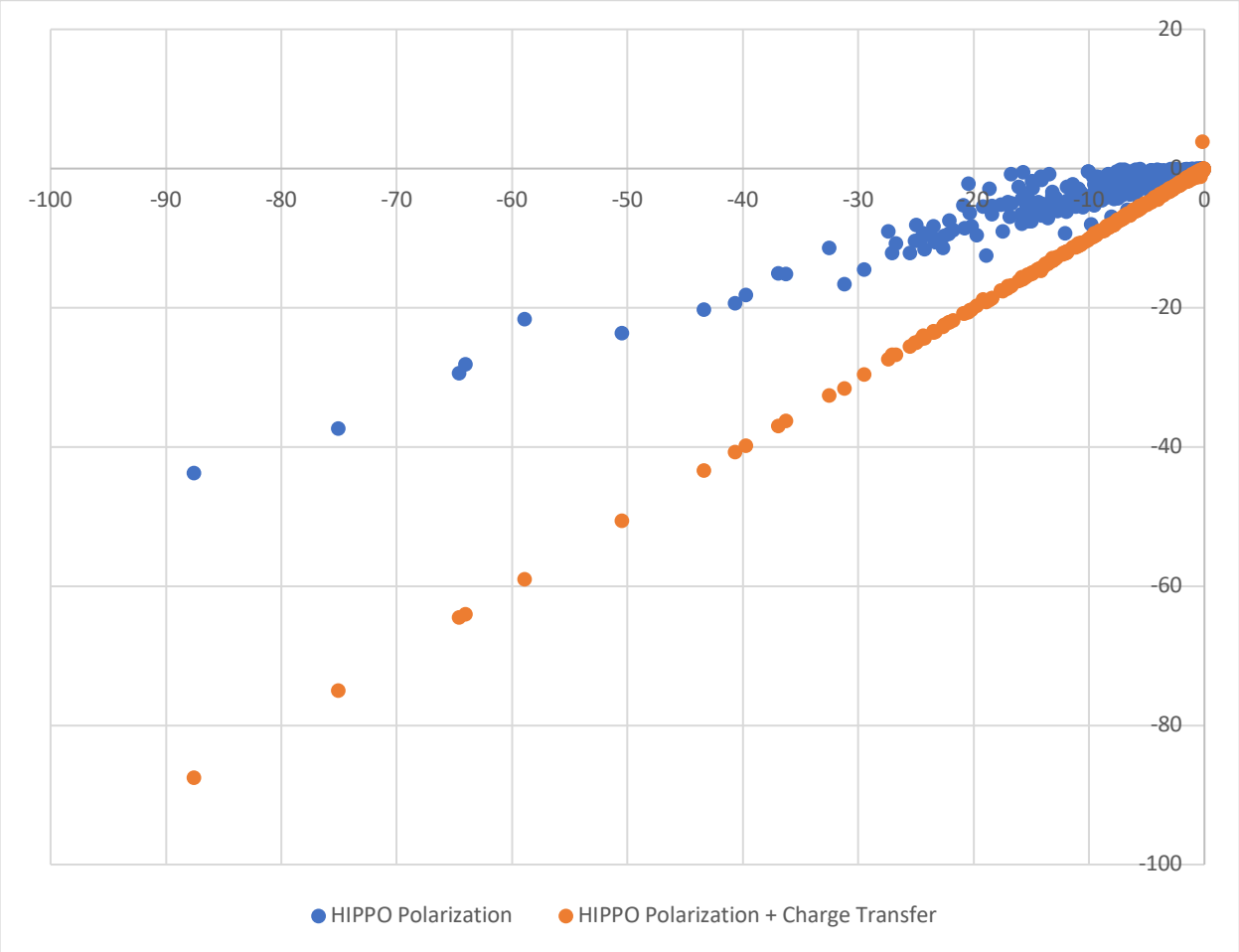


Figure 3.7 HIPPO Polarization + Charge Transfer on S101x7 Database.

A simple pairwise exponential function is able to fit the difference between SAPT induction and HIPPO polarization.

Figure 3.8 shows that as a complete function the HIPPO Polarization + Charge Transfer model can accurately reproduce the SAPT induction energy. This fit has an RMS error of under 1 kcal/mol, meeting the chemical accuracy requirement.

While the success of a simple exponential at fitting the difference between induction and polarization is not conclusive evidence for the qualitative interpretation behind the charge transfer explanation, it is a strong suggestion. Charge transfer is a necessarily ambiguous terminology when it comes to intermolecular interactions. It requires that one partition electrons in an unphysical manner. However, all reasonable charge transfer functions proposed are exponential in form. The fact that the difference in this case is so purely exponential seems to say that qualitatively there is some electron transfer occurring.

3.4.4 Three and Four Body Energies

Beyond polarization's part in the SAPT expansion, the other major reason to include polarization in the HIPPO model is for its many-body effects. The many-body energy is the amount of energy missing from a system of particles if one only calculated the energy as the sum of all the pair energies. In the case of organic molecules this is known to be a large effect, but it is missing (by construction) in the standard pairwise biomolecular force fields. Because calculating this many-body energy comes with the large cost associated with solving the system of equations for induced dipoles, I set out to check that the many-body energies of the model were accurate.

As an initial test case I chose the water trimer. I calculated the three-body energy of the water trimer for a variety of different intermolecular distances using the MP2 *ab initio* method

and compared those energies to the HIPPO polarization three-body energy. The structures were generated starting from the configuration shown in figure 3.8 and varying d_1 and d_2. Figure 3.10 shows the resulting three-body energies for a range of intermolecular distances. The three-body energies were calculated according to the formula,

$$E_{3B} = E_{total} - \sum_i E_i - \sum_i \sum_j E_{ij} \tag{3.20}$$

where E_{total} is the total energy of the system, E_i are the monomer energies, and E_{ij} are the energies of each pair of monomers.

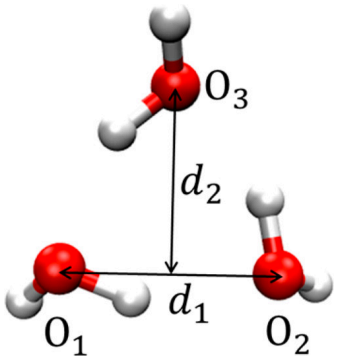


Figure 3.8 Water Trimer Geometry.

The distances d_1 and d_2 were varied systematically and three-body energies evaluated.

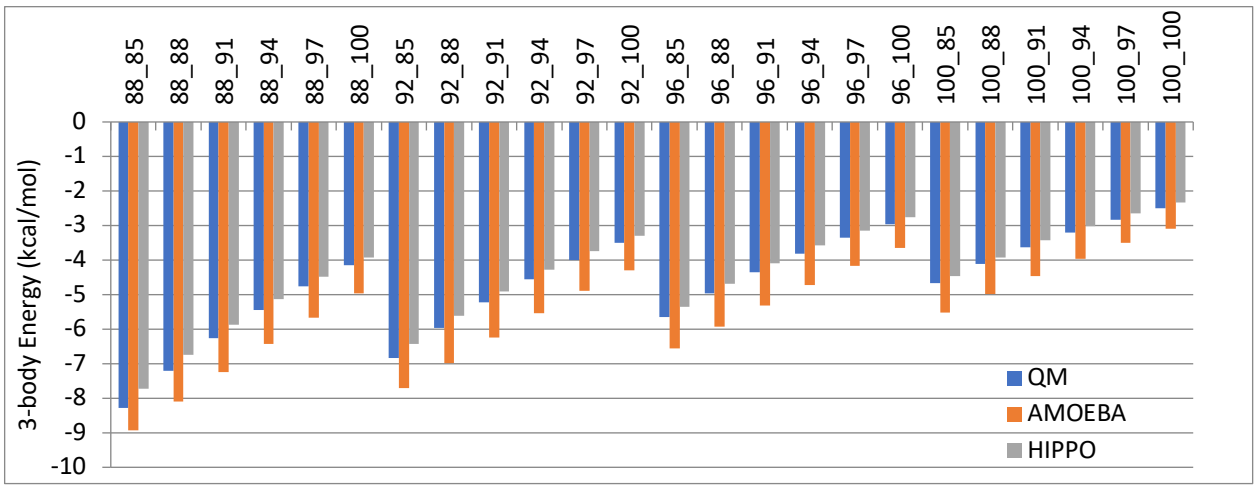


Figure 3.9 Water Trimer Three-Body Energies.

Each set of columns represents the three-body energy for a given configuration. In the x-axis notation for configurations, X_Y, X and Y refer to the percent of the equilibrium distances d_1 and d_2, respectively. "QM" refers to an MP2 calculation with aug-cc-pVTZ basis set and "AMOEBA" refers to the water03 polarization model.[10]

The results in figure 3.9 show good behavior of the HIPPO three-body energy across a range of intermolecular distances. In fact, the three-body energy is consistently closer to the MP2 result than the old AMOEBA model. This is not surprising, given that the AMOEBA water molecular polarizability is slightly larger and slightly more anisotropic than the experimental polarizability.

A subtle, but notable feature of figure 3.9 is that for every data point the HIPPO polarization model three-body energy is slightly lower in magnitude than the QM result. Given the nature of our model, this should be expected. Polarization is not the only many-body effect in intermolecular interactions. Dispersion, exchange-repulsion and even charge transfer are all known to also have, with the exception of exchange-repulsion, attractive many-body components. The QM three-body energy includes all of these phenomena without distinction, so we should expect a polarization-only many-body model to slightly underestimate the total three-body effect. Work from Demerdash and co-workers has shown that the many-body effect of

dispersion, exchange-repulsion and charge transfer is small for most organic molecules, especially water.[20] This, combined with the data observed in figure 3.9, leads to two conclusions. First, in agreement with Demerdash and co-workers' results, we do not need a many-body dispersion or exchange-repulsion function for HIPPO. And second, the many-body component of charge transfer, although certainly non-zero, is so small that we can chose to effectively ignore it for the purposes of an induction model whose goal is accuracy to within 1 kcal/mol. This validates the decision made to use a simple pairwise exponential to describe the charge transfer effect in this model. It should be noted that since, as stated earlier, the distinction between polarization and charge transfer is a matter of taste rather than hard fact, other force fields have reached the opposite conclusion. For example, the SIBFA (Sum of Interactions Between Fragments Ab Initio Computed) force field uses a computationally expensive many-body charge transfer term.[21,22] Given the HIPPO definition, however, of the polarization model as the linear response molecular polarizability and charge transfer as the remainder, the pairwise approximation seems to be a good one.

To verify that the many-body energy agreement observed in figure 3.9 is not unique to this particular configuration of water molecules, I tested the HIPPO polarization model on a range of other water clusters. For these larger clusters I also computed the four-body energy,

$$E_{4B} = E_{total} - \sum_i E_i - \sum_i \sum_j E_{ij} - \sum_i \sum_j \sum_k E_{ijk} \qquad (3.21)$$

where the final E_{ijk} term represents the sum of the energies of every set of trimers in the system. The results for a number of configurations of the water tetramer are shown in table 3.5.

	QM	HIPPO
	01 - Prism	
3-body	-9.0748127	-8.8611
4-body	-0.6128176	-0.9333

	02 - Cage	
3-body	-9.2219	-9.0533
4-body	-0.6517	-0.7587
	03 - Bag	
3-body	-10.4123	-10.3599
4-body	-1.3541	-1.4591
	04 - Cyclic Chair	
3-body	-11.5925	-11.9124
4-body	-2.0484	-2.253
	05 - Book 1	
3-body	-10.3471	-10.4967
4-body	-1.3316	-1.4523
	06 - Book 2	
3-body	-10.1075	-10.1501
4-body	-1.23	-1.3367
	07 - Cyclic Boat 1	
3-body	-11.2001	-11.4394
4-body	-1.8828	-2.0472
	08 - Cyclic Boat 2	
3-body	-11.1891	-11.4684
4-body	-1.8677	-2.0274

Table 3.5 Water Tetramer 3- and 4-Body Energies.

QM results are calculated at the MP2 level of theory. All energies are in kcal/mol.

These tetramer configurations are taken from reference 23 and represent a range of structures found in liquid water. The results show that the HIPPO agreement with QM many-body energies is accurate across conformations. The same holds true for larger clusters, as well. Shown in table 3.6 are three- and four-body energies for water clusters with up to eight molecules.

	8-mer	
	QM	HIPPO
3-body	-15.892902	-15.9016
4-body	-1.1263756	-1.5807

	6-mer	
	QM	HIPPO

3-body	-9.0748127	-8.8611
4-body	-0.6128176	-0.9333

5-mer

	QM	HIPPO
3-body	-9.1722077	-9.1677
4-body	-1.309086	-1.5763

4-mer

	QM	HIPPO
3-body	-6.3015546	-6.1326
4-body	-0.5776596	-0.8102

Table 3.6 Large Water Cluster Many-Body Energies.

QM results are calculated at the MP2 level of theory. All energies are in kcal/mol.

The structures listed in table 3.6 are also taken from reference 23. Again, the three- and four-body energies of the HIPPO polarization model match the QM results closely. I also tested the model on benzene trimers to see if how it performs on non-polar compounds. The results for a range of distances of the benzene trimer are plotted in figure 3.10.

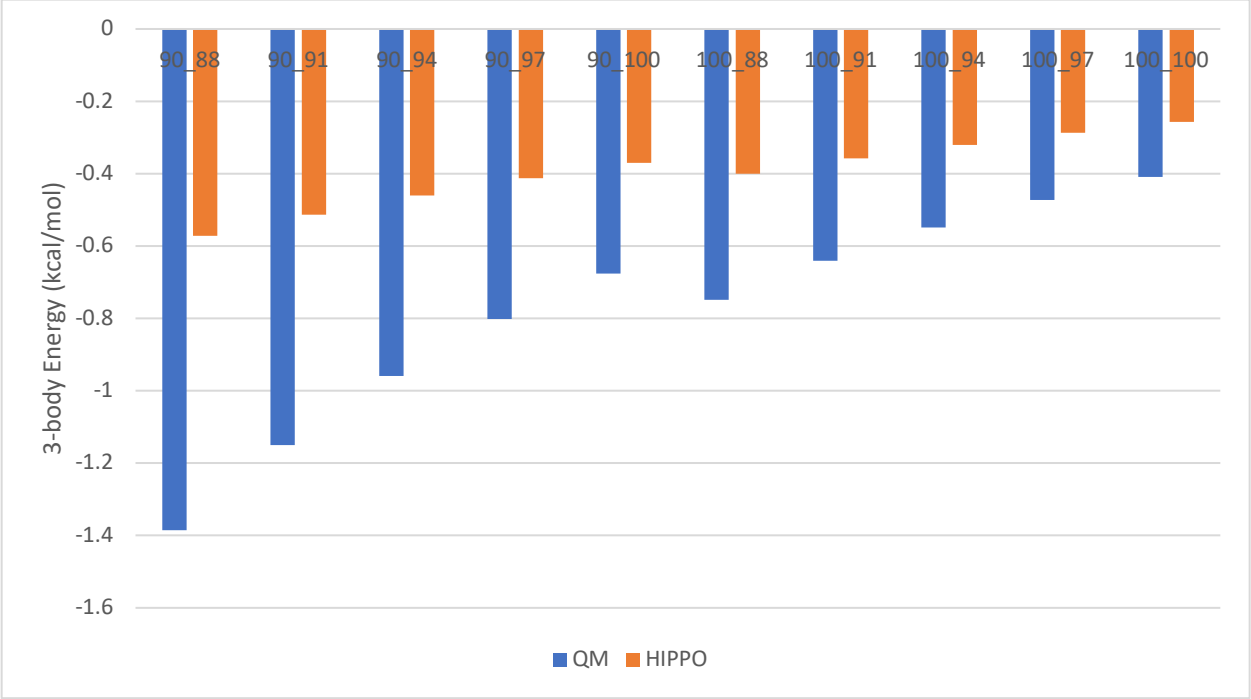


Figure 3.10 Benzene Trimer 3-body Energies.

Across the board, benzene trimers exhibit much less 3-body energy than water trimers. The absolute errors of the HIPPO polarization model are similar to those for water. Many-body dispersion likely makes up the remainder of the missing 3-body energy in this case. In the x-axis notation for configurations, X_Y, X and Y refer to the percent of the equilibrium distances d_1 and d_2, respectively. "QM" refers to an MP2 calculation with aug-cc-pVTZ basis set.

Although, compared to water, the percent error in the three-body energy is much larger, the absolute error is similar. This is because benzene trimers have much lower three-body energies in general due to the fact that they are nonpolar. The remaining three-body energy is likely primarily many-body dispersion. However, even in this case where many-body dispersion has maximum effect, the absolute error is still always less than 1 kcal/mol. This underscores the decision to forgo including other many-body effects in the force field.

3.5 Discussion and Conclusions

The summary of the HIPPO polarization + charge transfer model is fairly simple. It is a straightforward application of the atomic density model introduced in Chapter 2 to a polarizable induced dipole model. Because we are parameterizing against SAPT induction, this then has a charge transfer model layered on top of it. This simple approach yields a model that is physically motivated, has few free parameters and gives good results relative to both experiment and *ab initio* computations.

The physical interpretation of the polarization model is a big advance over the previous AMOEBA polarization model. The AMOEBA method used Thole-style damping to prevent polarization catastrophe, but these damping functions were not based in any particular physical model. In fact, the damping function used in AMOEBA is just one of the number of empirical possibilities suggested in Thole's original paper.[13] HIPPO achieves the same empirical end with a model that is rooted in physics. The damping functions, both permanent multipole – induced dipole and induced dipole – induced dipole, are derived directly from the electrostatic field generated by the atomic electron density model introduced in Chapter 2. In other words, the induced dipole is subject to the same electrostatic fields and is represented by the same density model as every other particle in the system. This gives some intuition for what the Thole model was approximating. Polarization catastrophe is an artifact of the point approximation. The HIPPO polarization model removes this artifact in a physically rational manner.

A side-effect of this physical motivation is that the HIPPO polarization model reduces the number of free parameters in the polarization model. Previously, a damped polarization model would need both an atomic polarizability for each atom, along with a damping coefficient. HIPPO removes the damping coefficient from the equation because it is set by the electrostatics

function determined in Chapter 2. This means that only the atomic polarizabilities need to be fit for the model.

Despite having this restriction imposed on the function, the HIPPO model performs quite well in fitting to experimental data. The polarization model is able to fit a broad range of molecular polarizabilities to a high degree of accuracy. An array of 32 molecules from the S101 database can be fit with just 28 atom classes. In validation tests, it also predicts the many-body energies of water and benzene clusters to within 1 kcal/mol. The validation tests suggest that the many-body energy, the primary reason to incur the cost of a polarization function in a biomolecular force field, is well-reproduced by the HIPPO polarization model.

Of course, there are some shortcomings of the model, due to its approximate nature, that merit consideration. The first, and most obvious is that it lacks any other many-body terms outside of polarization. Work by many groups has shown the importance of many-body dispersion and even exchange repulsion in some cases[24,25] Whether these matter to biomolecular simulations is a matter of much debate. The results here certainly don't settle the dispute, but they do provide an interesting data point. Comparing the water many-body data (where many body dispersion and repulsion effects are thought to be small) with the benzene many-body (where they are thought to be large), shows that indeed benzene shows a larger percent error in the many-body energy with the HIPPO model. This seems to validate the view that many-body dispersion is important. However, in absolute magnitude, the effect is still quite small, even in benzene. It suggests that for a threshold accuracy of 1 kcal/mol, even for non-polar substances, many-body dispersion may not be necessary.

Of course, more testing could be done to validate the HIPPO model on many-body energies. The present work only explored water and benzene as examples, but a fuller

examination would have merit. Because of the computational cost of constructing such a database of *ab initio* calculations, I did not pursue this in the current work. However, such a task is not computationally intractable. Further work in the area is highly encouraged.

The other lacking physics in the HIPPO polarization model is the truncation of the polarization expansion. HIPPO includes only linear order dipole polarizability. However, real atoms have higher order multipole (quadrupole, octupole, etc.) polarizabilities and each of these also have higher order polynomial (quadratic, cubic, etc.), sometimes called hyper-, polarizabilities. SAPT draws no such distinction. The SAPT induction energy includes all orders in both dimensions. In effect, the approximation being made by the HIPPO polarization model is illustrated graphically in figure 3.11.

	Linear	Quadratic	Cubic & beyond
Dipole	$\boldsymbol{\mu_{ind}} = \alpha \boldsymbol{E}$	$+\beta \boldsymbol{E}^2$	$\cdots$
Quadrupole	$\boldsymbol{\Theta_{ind}} = A \nabla \boldsymbol{E}$	$+B(\nabla \boldsymbol{E})^2$	$\cdots$
Octopole & beyond	$\vdots$	$\vdots$	$\ddots$

Figure 3.11 The HIPPO induction model in terms of the infinite-order polarizability expansion.

The green shaded area represents what is included in the HIPPO polarization model. The red shaded regain represents the entire rest of the expansion is covered by HIPPO's charge transfer model.

The green shaded region represents the portion of the polarization expansion that the HIPPO polarization function captures. Clearly, this is only the leading term in an infinitely long expansion. This leaves the rest of the expansion to be covered by something else. In the case of HIPPO, this is the charge transfer function. There is no doubt that this is an approximation. However, there are two pieces of evidence that suggest that it is not a bad one. First, as indicated by the many-body calculations here and elsewhere[20] the magnitude of these higher-order contributions to the many-body energy seems to be small. And second, the remainder of the SAPT induction energy is clearly exponential. (If the missing piece was, for instance, quadrupole

polarizability, we would expect the remainder to be polynomial.) This fact hints at a possible physical interpretation: the two main induction effects in intermolecular interactions may be linear order dipole polarization (a relatively small shift in electron density) and large scale intermonomer rearrangement (a large shift). In other words, it may be that the medium-order terms of the polarization expansion, don't matter all that much. How to test this hypothesis is not clear, but it is an interesting suggestion from the results shown in this work.

This leads to our final conclusion about the HIPPO total induction model: the charge transfer model empirically works. The remainder of the SAPT induction energy left over after linear-order polarization energy is computed is largely exponential and largely pairwise. This lends itself to a simple pairwise function, and this work shows that such a function works to within the required chemical accuracy for intermolecular interactions. The function is fast compute and fits the SAPT data well. This is clearly the least physically motivated term of the HIPPO force field, but, as I will show in Chapter 6, tuning to fit condensed phase properties is necessary regardless and this charge transfer term make a natural starting point for that tuning.

3.6 References

1 DeMarco, K. R., Bekker, S. & Vorobyov, I. Challenges and advances in atomistic simulations of potassium and sodium ion channel gating and permeation. *The Journal of physiology* **597**, 679-698 (2019).

2 Zhang, C. *et al.* AMOEBA polarizable atomic multipole force field for nucleic acids. *Journal of chemical theory and computation* **14**, 2084-2108 (2018).

3 Baker, C. M. Polarizable force fields for molecular dynamics simulations of biomolecules. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **5**, 241-254 (2015).

4 Salanne, M. Simulations of room temperature ionic liquids: from polarizable to coarse-grained force fields. *Physical Chemistry Chemical Physics* **17**, 14270-14279 (2015).

5 Jing, Z. *et al.* Polarizable force fields for biomolecular simulations: Recent advances and applications. *annual review of biophysics* (2017).

6 Bauer, B. A. & Patel, S. Recent applications and developments of charge equilibration force fields for modeling dynamical charges in classical molecular dynamics simulations. *Theoretical Chemistry Accounts* **131**, 1153 (2012).

7 Rick, S. W., Stuart, S. J. & Berne, B. J. Dynamical fluctuating charge force fields: Application to liquid water. *The Journal of chemical physics* **101**, 6141-6156 (1994).
8 Lemkul, J. A., Huang, J., Roux, B. & MacKerell Jr, A. D. An empirical polarizable force field based on the classical drude oscillator model: development history and recent applications. *Chemical reviews* **116**, 4983-5013 (2016).
9 Cieplak, P., Caldwell, J. & Kollman, P. Molecular mechanical models for organic and biological systems going beyond the atom centered two body additive approximation: aqueous solution free energies of methanol and N-methyl acetamide, nucleic acid base, and amide hydrogen bonding and chloroform/water partition coefficients of the nucleic acid bases. *Journal of Computational Chemistry* **22**, 1048-1057 (2001).
10 Ren, P. & Ponder, J. W. Polarizable Atomic Multipole Water Model for Molecular Mechanics Simulation. *The Journal of Physical Chemistry B* **107**, 5933-5947, doi:10.1021/jp027815+ (2003).
11 Huang, J., Simmonett, A. C., Pickard IV, F. C., MacKerell Jr, A. D. & Brooks, B. R. Mapping the Drude polarizable force field onto a multipole and induced dipole model. *The Journal of chemical physics* **147**, 161702 (2017).
12 Mei, Y. *et al.* Numerical study on the partitioning of the molecular polarizability into fluctuating charge and induced atomic dipole contributions. *The Journal of Physical Chemistry A* **119**, 5865-5882 (2015).
13 Thole, B. T. Molecular polarizabilities calculated with a modified dipole interaction. *Chemical Physics* **59**, 341-350 (1981).
14 Van Duijnen, P. T. & Swart, M. Molecular and atomic polarizabilities: Thole's model revisited. *The Journal of Physical Chemistry A* **102**, 2399-2407 (1998).
15 Simmonett, A. C., Pickard IV, F. C., Shao, Y., Cheatham III, T. E. & Brooks, B. R. Efficient treatment of induced dipoles. *The Journal of chemical physics* **143**, 074115 (2015).
16 Simmonett, A. C., Pickard IV, F. C., Ponder, J. W. & Brooks, B. R. An empirical extrapolation scheme for efficient treatment of induced dipoles. *The Journal of chemical physics* **145**, 164101 (2016).
17 Ren, P. & Ponder, J. W. Consistent treatment of inter-and intramolecular polarization in molecular mechanics calculations. *Journal of computational chemistry* **23**, 1497-1506 (2002).
18 Misquitta, A. J. Charge transfer from regularized symmetry-adapted perturbation theory. *Journal of chemical theory and computation* **9**, 5313-5326 (2013).
19 Rackers, J. A. *et al.* Tinker 8: software tools for molecular design. *Journal of chemical theory and computation* **14**, 5273-5289 (2018).
20 Demerdash, O., Mao, Y., Liu, T., Head-Gordon, M. & Head-Gordon, T. Assessing many-body contributions to intermolecular interactions of the AMOEBA force field using energy decomposition analysis of electronic structure calculations. *The Journal of chemical physics* **147**, 161721 (2017).
21 Gresh, N., Claverie, P. & Pullman, A. Intermolecular interactions: Elaboration on an additive procedure including an explicit charge-transfer contribution. *International journal of quantum chemistry* **29**, 101-118 (1986).
22 Gresh, N., Piquemal, J. P. & Krauss, M. Representation of Zn(II) complexes in polarizable molecular mechanics. Further refinements of the electrostatic and short-range

contributions. Comparisons with parallel ab initio computations. *Journal of Computational Chemistry* **26**, 1113-1130, doi:Doi 10.1002/Jcc.20244 (2005).
23 Liu, C., Qi, R., Wang, Q., Piquemal, J.-P. & Ren, P. Capturing many-body interactions with classical dipole induction models. *Journal of chemical theory and computation* **13**, 2751-2761 (2017).
24 Tkatchenko, A., DiStasio Jr, R. A., Car, R. & Scheffler, M. Accurate and efficient method for many-body van der Waals interactions. *Physical review letters* **108**, 236402 (2012).
25 Jeziorski, B., Bulski, M. & Piela, L. First-Order perturbation treatment of the short-range repulsion in a system of many closed-shell atoms or molecules. *International Journal of Quantum Chemistry* **10**, 281-297 (1976).

Chapter 4: Dispersion

At the time that the polarization model was finalized, the plan was not to replace the existing AMOEBA van der Waals model. The model up until this point was still called "AMOEBA 2", since the changes had been modifications of the existing functional form rather than wholescale changes. This changed for three reasons. First, I attempted to construct a water model by using the electrostatics and polarization functions as described above and reparametrizing the existing Buffered 14-7 van der Waals potential. The results were lackluster. The water model suffered from the same problems as both published AMOEBA water models: the first peak of the oxygen-oxygen radial distribution function was difficult to bring into agreement with experiment. Second, the dispersion and exchange-repulsion parts of the Buffered 14-7 van der Waals function fit the SAPT data very poorly. And third, upon taking a step back, I realized that I could actually use the density model I established to model dispersion (and subsequently, exchange-repulsion) in a manner that approximates the SAPT perturbation theory approach. This decision to overhaul the van der Waals function constitutes the break point between "AMOEBA 2" and HIPPO. The name had not been born yet, but, conceptually, the groundwork had been established.

The following is taken from a published paper in which I describe how the HIPPO density model can be used to construct a physically grounded model for dispersion. In particular, this model gives physical meaning to the known need for "dispersion damping". Literature going back decades has acknowledged that while the leading dispersion term is proportional to $\sim 1/r^6$ at long range, this term must be modified by an exponential damping function at short range to agree with *ab initio* calculations. What has been lacking, however, is a rationale for why this

damping is necessary and what form it should have. The HIPPO dispersion model provides this with simple density-based arguments.

4.1 Introduction

The range of possible problems for molecular mechanics models to solve is immense. For problems that are too large to solve with Schrodinger's equation but too small to be observed experimentally, we rely on classical models to make predictions and generate hypotheses. This ability has made molecular mechanics force fields integral to the study of problems from RNA folding[1] to new alloy characterization[2]. Because they are classical approximations to quantum mechanical reality, the success of these models is entirely dependent on how accurate that approximation is on a wide variety of systems. To achieve this, most force fields split the interaction energies of interacting atoms into physically meaningful components. Among the most significant of these components is the dispersion interaction that arises from the correlation of instantaneous induced dipoles.

No force field can provide fully accurate predictions for every component of the total energy of a system. In current models this has been typically handled by careful cancellation of errors between the various components (electrostatics, polarization, repulsion, dispersion, *etc*.). More recently, however, a new crop of next-generation force fields is emerging that aim to reduce this dependence on error cancellation by comparing directly to *ab initio* energy decomposition analysis data.[3-10] We are working on a model with this same objective. Previously we have shown that it is possible to accurately model electrostatics (to within 1 kcal/mol) in regions where previously error cancellation had long been relied upon, the so-called "charge penetration" error.[11] In this work we shall demonstrate that the same is possible for the

dispersion interaction component. While this does not represent a complete force field capable of condensed phase simulations, it is an important step toward such a full model.

Accurately modeling dispersion in classical force fields is known to be important, particularly for biological systems. On a phenomenological level, dispersion is what causes neutral atoms and molecules to be weakly attracted to each other. This makes it essential to modeling simple Lennard-Jones fluids such as liquid argon, but it is also critically important to more complex systems. Dispersion has been shown to be an essential component of modeling nucleic acid structure,[12] where it contributes to the so-called stacking energy of nucleic acid bases. It is known to play a part in halogen bonding, supporting, along with electrostatics, the stabilization energy of the interaction.[13,14] Additionally, long-range dispersion is widely recognized to be important for the simulation of lipid bilayers.[15] This broad spectrum of applications motivates the necessity of accurate dispersion models.

The history of dispersion models dates back to Fritz London, who first established the canonical $1/r^6$ dependence of the London Dispersion energy. This model has been enormously influential. The vast majority of biological force fields in use today still use this simple model (Amber[16], CHARMM[17], *etc*.), or derivatives thereof such as the attractive part of Halgren's buffered 14-7 potential[18] used in the AMOEBA[19] force field. It is well known, however, that the $1/r^n$ potential expansion breaks down for short-range interactions where charge distributions of interacting molecules overlap.[20] There is a long history of attempts to correct this divergence though the use of damping functions. An important early damped dispersion model was the empirical HFD (Hartree-Fock-Dispersion) scheme proposed by Scoles and coworkers.[21] Another notable attempt to describe this phenomenon was undertaken by Tang and Toennies who introduced a damping function parameterized to account for the overlap in charge distributions.[22]

A comprehensive review of dispersion damping functions is beyond the scope of this work, but the original Tang and Toennies report provides a thorough overview of dispersion damping functions up to that point. These types of formalisms have seen the widest use as dispersion corrections to DFT calculations.[23] DFT-D schemes have used the Tang-Toennies function, as well as various other damping functions proposed by Wu and Yang[24], Chai and Head-Gordon[25], and Johnson and Becke[26]. Despite wide use in the DFT community, damped dispersion functions have been taken up in decidedly fewer molecular mechanics models. Notably, the Effective Fragment Potential (EFP) model employs a dispersion model that utilizes an overlap-based, parameter-free modification of the Tang-Toennies damping function.[27,28] And recently, Verma et al. proposed using the dispersion part of the DFT-D3 formulation of Grimme[23] as a molecular mechanics model.[29] However, while it has been shown that previous damping functions can effectively account for the change in dispersion upon charge overlap, they do so largely empirically. In the case of the Tang-Toennies damping function for example, the form is based on a Born-Mayer potential described by an empirically fit width parameter.

In this paper we propose a damped dispersion function similar in spirit to that of Tang and Toennies but rooted in a physical model of charge distribution overlap. In previous work we have shown that a relatively simple model can capture the physical extent of atomic charge distributions that leads to the so-called charge penetration error in electrostatic interactions between molecules.[11] Here we will show this same model can be used directly and without modification to create a dispersion model that is elegantly unified with the electrostatic model. This unification is possible because both the electrostatic and dispersion terms depend on the density. The electrostatic term is simply the interaction between two static densities, while the dispersion term arises from the interaction of densities associated with instantaneous induced

dipoles. In this work we will show that the same rough description of the density can be used in both cases to great effect. This will be done in five parts. First, we elucidate the theory that starts from dipole-dipole interactions and gives rise to this new damped dispersion function. Second, we describe the methods of the study. Third, we evaluate the performance of this function against benchmark Symmetry Adapted Perturbation Theory (SAPT) calculations. Fourth, we will describe how the model has been implemented with dispersion particle mesh Ewald (DPME) to boost its efficiency. And lastly, we will discuss the implications of this work and some general conclusions.

4.2 Theory

To present our new damped dispersion model, we shall first revisit a simple derivation of the original London dispersion model. We do so first and foremost because it forms the basis for our damped model, but also because it is instructive. One of the defining characteristics of a damped dispersion model, as we shall argue later in the paper, is that it has a straightforward physical interpretation. Dispersion is correctly said to be a fundamentally non-classical phenomenon, but the model we use to describe it need not to be so bound. We will show that an interpretable model of dispersion can be constructed from physical models of atomic polarizability and charge density.

4.2.1 London Dispersion

For our description of canonical London Dispersion energy, we will follow that of Maitland, Rigby, Smith and Wakeham.[30] The dispersion energy between two atoms arises from the interaction between instantaneous dipoles of those atoms. To model this system, we consider a simplified one-dimensional Drude oscillator model as illustrated in Figure 4.1.

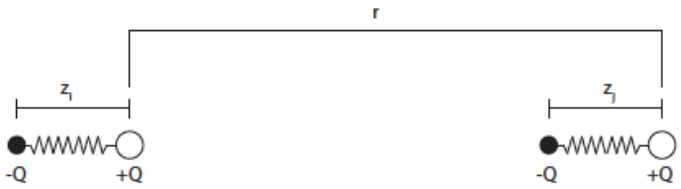


Figure 4.1. Classical Model of Dispersion

In this representation each atom is represented by a fixed charge +Q bound by a spring with spring constant, k, to and an equal and opposite charge, –Q with mass, M. This model is crude, but it captures the essential elements of the dispersion interaction. At any point in time each atom has a dipole moment, μ = Qz (dependent on the atomic polarizability determined by *k*) and those dipole moments are free to interact with each other.

When atom *i* and atom *j* are infinitely separated, the Schrödinger equation for each can be written as,

$$\frac{1}{M}\frac{\partial^2 \Psi_i}{\partial z_i^2} + \frac{2}{\hbar^2}\left(E_i - \frac{1}{2}kz_i^2\right)\Psi_i = 0 \quad , \tag{4.1}$$

where the potential energy term is merely the energy of a simple harmonic oscillator. The same can be written for atom *j*. The solutions to this equation can be found trivially, yielding ground state energies of,

$$E_i = \frac{1}{2}\hbar\omega_0 \quad \text{and} \quad E_j = \frac{1}{2}\hbar\omega_0 \quad , \tag{4.2}$$

where the frequency, ω_0 is:

$$\omega_0 = \sqrt{\frac{k}{M}} \quad . \tag{4.3}$$

In the complete non-interacting limit, the total energy of the system is:

$$E(r \to \infty) = E_i + E_j = \hbar\omega_0 \quad . \tag{4.4}$$

This limit in itself is not useful, but if we consider what happens when the two atoms get closer, we shall see that it sets a useful reference for our potential energy function. If we bring the two atoms closer so that they do interact, but not so close that their charge distributions overlap, our Schrödinger equation is not trivially longer separable. The wave equation for two interacting atoms now includes the electrostatic interaction between the two dipoles and becomes,

$$\frac{1}{M}\frac{\partial^2\Psi}{\partial z_i^2} + \frac{1}{M}\frac{\partial^2\Psi}{\partial z_j^2} + \frac{2}{\hbar^2}\left(E - \frac{1}{2}kz_i^2 - \frac{1}{2}kz_j^2 - U_{electrostatic}\right)\Psi = 0 \quad . \tag{4.5}$$

One can see that in addition to the simple harmonic oscillator terms, a new potential appears in equation 4.5. This is the potential energy at any given instant between the two interacting instantaneous multipole distributions. For the dipole-dipole interaction of the simple Drude model of figure 4.1, the form of this potential is easily obtained from simple electrostatics:

$$U_{electrostatic} = U_{dipole-dipole} = \nabla\nabla U_{chg-chg} = \nabla\nabla\left(\frac{q_i q_j}{r}\right) = \frac{1}{r^3}\left(\vec{\mu}_i \cdot \vec{\mu}_j - 3\frac{\left(\vec{\mu}_i \cdot \vec{r}_{ij}\right)\left(\vec{r}_{ij} \cdot \vec{\mu}_j\right)}{r^2}\right) . \tag{4.6}$$

If we plug in the Drude dipoles from figure 4.1, μ = Qz, equation 4.6 becomes,

$$U_{dipole-dipole} = \frac{1}{r^3}\left(\mu_i\mu_j - 3\frac{\left(\mu_i r\right)\left(\mu_j r\right)}{r^2}\right) = -\frac{2\mu_i\mu_j}{r^3} \quad . \tag{4.7}$$

This dipole-dipole energy is the source, as we shall show, of the canonical $1/r^6$ leading term dependence of the dispersion energy.

Combining equation 4.7 with equation 4.5 yields,

$$\frac{1}{M}\frac{\partial^2\Psi}{\partial z_i^2} + \frac{1}{M}\frac{\partial^2\Psi}{\partial z_j^2} + \frac{2}{\hbar^2}\left(E - \frac{1}{2}kz_i^2 - \frac{1}{2}kz_j^2 - \frac{2\mu_i\mu_j}{r^3}\right)\Psi = 0 \quad . \tag{4.8}$$

Following the transformation of variables of Maitland, Rigby, Smith and Wakeham, we define,

$$\lambda_1 = \frac{z_i + z_j}{\sqrt{2}}, \qquad \lambda_2 = \frac{z_i - z_j}{\sqrt{2}} \tag{4.9}$$

and rewrite equation 4.8 as,

$$\frac{1}{M}\frac{\partial^2 \Psi}{\partial z_i^2} + \frac{1}{M}\frac{\partial^2 \Psi}{\partial z_j^2} + \frac{2}{\hbar^2}\left(E - \frac{1}{2}k_1\lambda_i^2 - \frac{1}{2}k_2\lambda_j^2\right)\Psi = 0 \tag{4.10}$$

where,

$$k_1 = k + \frac{2Q^2}{r^3}, \qquad k_2 = k - \frac{2Q^2}{r^3}. \tag{4.11}$$

Equation 4.10 is simply a transformed version of the original problem of two independent harmonic oscillators. It can be solved in the same manner giving,

$$E(r) = \frac{1}{2}\hbar\left(\omega_1 + \omega_2\right), \tag{4.12}$$

where,

$$\omega_1 = \sqrt{\frac{k_1}{M}} = \omega_0\sqrt{1 - \frac{2Q^2}{r^3 k}}, \qquad \omega_2 = \sqrt{\frac{k_2}{M}} = \omega_0\sqrt{1 + \frac{2Q^2}{r^3 k}}. \tag{4.13}$$

One can see that as r becomes large, ω_1 and ω_2 converge to ω_0 where we recover the independent oscillator solution. For small perturbations ω_1 and ω_2 can be approximated with a binomial expansion,

$$\sqrt{1+x} = 1 + \tfrac{1}{2}x - \tfrac{1}{8}x + \ldots \tag{4.14}$$

so, the total energy becomes,

$$E(r) = \hbar\omega_0 - \frac{Q^4\hbar\omega_0}{2r^6k^2} + \ldots. \tag{4.15}$$

The final step is to subtract the energy of infinitely separated atoms. This gives the dispersion potential energy,

$$U_{dispersion} = E(r) - E(\infty) = -\frac{Q^4 \hbar \omega_0}{2k^2 r^6} + \ldots , \tag{4.16}$$

where the canonical r^6 dependence arises from the first non-zero term from the application of the binomial expansion.

It should be noted that while the dipole-dipole interaction is the dominant electrostatic term of equation 4.5, there are terms arising from higher-order multipole interactions as well. The dipole-quadrupole and quadrupole-quadrupole interactions giving rise to the $1/r^8$ and $1/r^{10}$ potentials are derived in Appendix A. There are a number of models that use these terms, including EFP, SIBFA, and Misquitta and Stone's model for small organic molecules.[27,31,32] For a perspective on the importance of these higher order terms for the case of the neon dimer, the reader is directed to the work of Bytautas and Ruedenberg.[33] The latter reference showed that even for this simple dimer, the $1/r^8$ and $1/r^{10}$ terms are nearly impossible to distinguish at reasonable separations. There are odd-power terms ($1/r^7$, $1/r^9$, *etc.*) that can be included in the expansion as well. These arise from mixing of the even order terms, are highly angularly dependent, and spherically average to zero at long range.[34] There has also been recent work on incorporating these terms into dispersion models,[35,36] where these higher-order terms give successively better approximations to the exact dispersion energy. As we will show, however, to reach the stated accuracy goal of < 1 kcal/mol, only the leading term will be necessary.

For most systems the perturbation of the dipole-dipole interaction energy is small compared to the energy holding the electrons to their respective atoms. This makes taking only the leading r^6 term of equation 4.16 a good approximation for most long-range intermolecular interactions. In practice this is done by introducing a parameter, *C*, to capture this dependence,

$$U^{dispersion} = -\frac{C_6^i C_6^j}{r^6} . \tag{4.17}$$

This model will be referred to throughout the remainder of the paper as the London Dispersion model. Unfortunately, this method of approximation starts to break down when the charge distributions of interacting atoms start to overlap. We will handle this situation through introduction of short-range damping, but rather than rely on empiricism for the damping function, we look to the underlying electrostatics to provide a consistent model.

4.2.2 Short-Range Electrostatics

A long-standing problem in the modeling of electrostatics for molecular mechanics models is the so-called charge penetration error. The error arises when charge distributions of interacting atoms overlap, causing the true electrostatic energy of the interacting densities to diverge from the point charge or point multipole approximation. We have shown in previously published studies[11,37,38] that a simple hydrogen-like approximation of the Coulomb potential does a remarkably good job at correcting this error.

Why is this germane to a study of dispersion? Dispersion, as shown above, can be modeled as arising from a dipole-dipole interaction. In the context of the multipolar AMOEBA force field we have shown that the hydrogen-like approximation to the Coulomb interaction can be extended to the interactions between higher-order multipole moments. In fact, including these corrections for charge-dipole, dipole-dipole, dipole-quadrupole, *etc*. interactions is essential to the transferability and accuracy of the model.[11] Here we show that the dipole-dipole interaction arising from this earlier model can be used directly to create a new damped dispersion model.

To illustrate where the dipole-dipole damping comes from, we follow a similar derivation to that of Ref. 11. The potential due to the electrons for this model is defined as,

$$V(r) = \frac{q_i}{r}\left(1 - e^{-\alpha_i r}\right) \tag{4.18}$$

where r is the distance from the center of the charge distribution and α is a parameter describing the width of the distribution. Application of Poisson's equation,

$$\nabla^2 V = \frac{\rho}{\varepsilon_0}, \tag{4.19}$$

yields the corresponding density,

$$\rho(r) = \frac{q_i \alpha_i^2 \varepsilon_0}{r} e^{-\alpha_i r}. \tag{4.20}$$

These two quantities can be used to approximate the Coulomb interaction energy between two charge distributions,

$$U_{electrostatic}^{chg-chg} = \int\int \frac{\rho_i(\mathbf{r}_i)\rho_j(\mathbf{r}_j)}{\mathbf{r}_{ij}} d\mathbf{r}_i\, d\mathbf{r}_j = \frac{1}{2}\left(\int \rho_i(\mathbf{r}_i) V_j(\mathbf{r}_j) d\mathbf{r}_i + \int \rho_j(\mathbf{r}_j) V_i(\mathbf{r}_i) d\mathbf{r}_j\right). \tag{4.21}$$

Application of the one-center integral method of Coulson[39] gives

$$U_{electrostatic}^{chg-chg} = \frac{q_i q_j}{r}\left(1 - \frac{\alpha_j^2}{\left(\alpha_j^2 - \alpha_i^2\right)} e^{-\alpha_i r} - \frac{\alpha_i^2}{\left(\alpha_i^2 - \alpha_j^2\right)} e^{-\alpha_j r}\right). \tag{4.22}$$

Equation 4.22 gives the charge-charge electrostatic energy. To get the dipole-dipole energy, recall that the full multipole energy of the *i-j* interaction can be written:

$$\begin{aligned} U_{electrostatic}^{total} &= U^{chg-chg} + U^{chg-dipole} + U^{dipole-chg} + U^{dipole-dipole} + \ldots \\ &= q_i T_{ij} q_j + q_i \nabla T_{ij} \mu_j - \mu_i \nabla T_{ij} q_j + \mu_i \nabla\nabla T_{ij} \mu_j + \ldots \end{aligned} \tag{4.23}$$

For a point-point interaction T_{ij} is simply 1/r, but for our model direct inspection of equation 4.22 yields

$$T_{ij} = \frac{1}{r}\left(1 - \frac{\alpha_j^2}{\left(\alpha_j^2 - \alpha_i^2\right)} e^{-\alpha_i r} - \frac{\alpha_i^2}{\left(\alpha_i^2 - \alpha_j^2\right)} e^{-\alpha_j r}\right) = \frac{1}{r} f_1^{damp}. \tag{4.24}$$

We can now apply this new relation for T_{ij} to the definition of the dipole-dipole energy from equation 4.23.

$$U_{damp}^{dipole-dipole} = \mu_i \nabla\nabla T_{ij} \mu_j = f_3^{damp} \frac{\vec{\mu}_i \cdot \vec{\mu}_j}{r^3} - f_5^{damp} \frac{3\left(\vec{\mu}_i \cdot \vec{r}_{ij}\right)\left(\vec{r}_{ij} \cdot \vec{\mu}_j\right)}{r^5} \tag{4.25}$$

where f_3 and f_5 are the damping terms that come from derivatives of the f_1^{damp} term of equation 4.24,

$$f_3^{damp} = 1 - \frac{\alpha_j^2}{\left(\alpha_j^2 - \alpha_i^2\right)}\left(1 + \alpha_i r\right)e^{-\alpha_i r} - \frac{\alpha_i^2}{\left(\alpha_i^2 - \alpha_j^2\right)}\left(1 + \alpha_j r\right)e^{-\alpha_j r}$$
$$f_5^{damp} = 1 - \frac{\alpha_j^2}{\left(\alpha_j^2 - \alpha_i^2\right)}\left(1 + \alpha_i r + \frac{1}{3}\left(\alpha_i r\right)^2\right)e^{-\alpha_i r} - \frac{\alpha_i^2}{\left(\alpha_i^2 - \alpha_j^2\right)}\left(1 + \alpha_j r + \frac{1}{3}\left(\alpha_j r\right)^2\right)e^{-\alpha_j r} \tag{4.26}$$

Now let us compare equation 4.24 and equation 4.6. Clearly the difference between the point dipole-dipole interaction and the new model's dipole-dipole interaction is the damping terms that arise from the hydrogen-like model of charge density. For large separations f_3 and f_5 approach one and we recover the point interaction. For small density overlaps, f_3 and f_5 represent a perturbation that damps the point dipole-dipole interaction.

4.2.3 Overlap Damped Dispersion

To derive our damped dispersion model, we start from the earlier derivation of London Dispersion. Equations 4.1-4.5 remain the same, but instead of inserting the point dipole-dipole interaction energy into equation 4.5, we now substitute our damped dipole-dipole interaction from equation 4.25. Following our simple one-dimensional Drude model we obtain

$$\frac{1}{M}\frac{\partial^2 \Psi}{\partial z_i^2} + \frac{1}{M}\frac{\partial^2 \Psi}{\partial z_j^2} + \frac{2}{\hbar^2}\left(E - \frac{1}{2}kz_i^2 - \frac{1}{2}kz_j^2 - U_{dipole-dipole}\right)\Psi = 0$$
$$U_{damp}^{dipole-dipole} = f_3^{damp} \frac{\vec{\mu}_i \cdot \vec{\mu}_j}{r^3} - f_5^{damp} \frac{3\left(\vec{\mu}_i \cdot \vec{r}_{ij}\right)\left(\vec{r}_{ij} \cdot \vec{\mu}_j\right)}{r^5} \tag{4.27}$$

where $U^{dipole\text{-}dipole}$ can be simplified to:

$$U_{dipole-dipole} = f_3^{damp}\frac{\mu_i\mu_j}{r^3} - f_5^{damp}\frac{3(\mu_i r)(\mu_j r)}{r^5} \tag{4.28}$$

Inserting this into the Schrödinger equation yields

$$\frac{1}{M}\frac{\partial^2\Psi}{\partial z_i^2} + \frac{1}{M}\frac{\partial^2\Psi}{\partial z_j^2} + \frac{2}{\hbar^2}\left(E - \frac{1}{2}kz_i^2 - \frac{1}{2}kz_j^2 - \left(3f_5^{damp} - f_3^{damp}\right)\frac{\mu_i\mu_j}{r^3}\right)\Psi = 0 \tag{4.29}$$

.

This can be solved by the same transformation as the non-damped case discussed earlier where,

$$k_1 = k - \frac{2Q^2}{r^3}f_{dispersion}^{damp}, \qquad k_2 = k + \frac{2Q^2}{r^3}f_{dispersion}^{damp}$$
$$f_{dispersion}^{damp} = 3f_5^{damp} - f_3^{damp} \tag{4.30}$$

.

This results in the solution

$$E = \frac{1}{2}\hbar(\omega_1 + \omega_2)$$
$$\omega_1 = \sqrt{\frac{k_1}{M}} = \omega_0\sqrt{1 - \frac{2Q^2}{r^3k}f_{dispersion}^{damp}}, \qquad \omega_2 = \sqrt{\frac{k_2}{M}} = \omega_0\sqrt{1 + \frac{2Q^2}{r^3k}f_{dispersion}^{damp}} \tag{4.31}$$

.

Applying the binomial expansion and subtracting the energy of infinitely separated atoms yields the damped dispersion energy:

$$U_{dispersion}^{damp} = -\frac{Q^4\hbar\omega_0}{2k^2r^6}\left(f_{dispersion}^{damp}\right)^2 + \ldots \tag{4.32}$$

.

Just as before, for small density overlaps the leading term of equation 4.32 dominates. To convert this into a parameterized molecular mechanics model we again introduce C_6 parameters, giving our final model energy:

$$U_{dispersion}^{damp} = -\frac{C_6^iC_6^j}{r^6}\left(f_{dispersion}^{damp}\right)_{ij}^2 \tag{4.33}$$

This model represents an elegant and simple unification of the electrostatics and dispersion models for molecular mechanics force fields. We will refer to this model throughout

the remainder of the paper as the "Overlap Damped Dispersion" model. It has some important features:

1. The model retains the canonical $1/r^6$ asymptotic behavior as f tends to unity at large separations.

2. The damping function has a straightforward physical interpretation: it is the integral of the overlap of between charge distributions on interacting atoms.

3. The damping function follows a similar exponential form as other previously proposed dispersion damping functions.

4. The damping function has no adjustable parameters. The parameters are fixed from the electrostatics charge penetration damping function.

As we will show in Section 4, this model, in addition to being theoretically compelling, produces good agreement with dispersion energies from *ab initio* energy decomposition analysis calculations.

4.3 Methods

The damped dispersion model we propose requires the fitting of C_6 parameters. To obtain these parameters, validate their robustness and assess the model's accuracy, we set out a four-step protocol. First, we assemble a database of representative molecular interactions. Second, we perform benchmark *ab initio* reference calculations on that database. Third, we fit the parameters of our model to the reference *ab initio* data. Fourth, we assess the robustness of the fit by validation of the model on systems outside of the database.

For the scope of this study we intend to parameterize our model for the chemical space of biomolecules. To this end we use the previously constructed S101x7 database [40] for fitting. This database consists of 101 distinct pairs of molecular dimers. For each of these dimers, seven

points along the dissociation curve are established at 0.7, 0.8, 0.9 0.95, 1.0 1.05 and 1.1 times the equilibrium intermolecular distance. Details on how the structures were generated are available in Ref. 37. The dimers in this set represent a cross section of typical interactions found in protein and nucleic acid systems. We note that the points in the dataset at 0.7x the equilibrium distance are important despite the fact that they are rarely sampled in condensed phase simulations for most systems. These points are included to ensure that the shape of the potential at the closest sampled points (often 0.8x the equilibrium distance) is accurately captured. A summary of all the pair interactions is presented in figure 4.2.

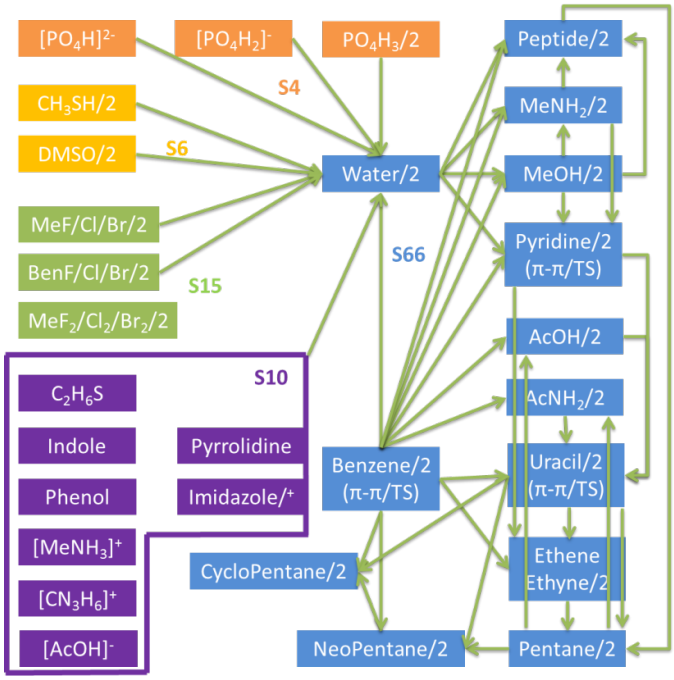


Figure 4.2. Dimer pairs in the S101 database.

Arrows indicate heterodimers, while "/2" indicates a homodimer. Reprinted with permission from Wang, Q. *et al.* General model for treating short-range electrostatic penetration in a molecular mechanics force field. *Journal of Chemical Theory and Computation* 11, 2609-2618. Copyright 2015 American Chemical Society.

In order to parameterize our model, a set of dispersion reference data is required. Because dispersion is not a physical observable, we must rely on an *ab initio* energy decomposition analysis (EDA) to generate our reference data. We have chosen Symmetry Adapted Perturbation Theory (SAPT)[41] for this purpose. SAPT has a number of features that make it a reasonable choice. First, SAPT is a perturbation theory approach that takes the electron density of monomers as its unperturbed state. This is an exact analogy to molecular mechanics models where distributed multipoles are calculated from monomer densities. Second, because of this correspondence, SAPT is the theory that was used to generate the parameterization of the electrostatic model referenced in Section 2. Using SAPT here as well ensures a straightforwardly unified model. Finally, SAPT is a well-established theory with a proliferation of studies analyzing its accuracy with respect to various orders and basis sets. We use the SAPT2+ level of theory as defined by Sherrill *et al.*[42] with Dunning correlation consistent basis sets[43,44] to estimate the complete basis set (CBS) limit[45] for the SAPT energy components. The SAPT2+ method with large, augmented basis sets has been previously shown to give errors relative to CCSD(T)/CBS of about 0.3 kcal/mol. This was chosen over the cheaper to compute SAPT0 method, which gives errors of around 0.5 kcal/mol. In order to minimize the difference between our SAPT calculations and gold-standard CCSD(T), we evaluated the residual,

$$R = \left| E_{total}^{CCSD(T)/CBS} - \left(E_{non-dispersion}^{SAPT\,2+/CBS} + cE_{dispersion}^{SAPT\,2+/CBS} \right) \right|, \tag{4.34}$$

where ($E_{non\text{-}dispersion}$ + $E_{dispersion}$) represents the total SAPT2+ energy, with a scale factor, c, introduced as a parameter. Minimizing this residual with respect to c yielded a scale factor of c = 0.89 that is used to scale all dispersion energies. For further details on the construction of the reference data for S101x7, please see Ref. 37. The Psi4 program was used to perform all SAPT calculations.[46-48] All structures and reference data are available at the S101x7 online repository.[40]

To obtain C_6 parameters, we performed a nonlinear least square fit of the SAPT dispersion reference data using a Levenberg-Marquardt algorithm implemented in the Tinker molecular mechanics software package. To test the robustness of this parameterization we leave out some of the data points and repeat the fit. The model with the new parameters is then evaluated on the excluded points. As a validation test case, we also evaluate the performance of the model on previously published nucleic acid interaction data.[49]

The last part of the study is evaluation of the dispersion particle mesh Ewald (DPME) method. DPME has been implemented in a locally modified version of Tinker and is available through the Tinker GitHub site.[50] To evaluate the efficiency of this implementation, the DPME method is tested on a 36 Å periodic cube containing 1,600 water molecules. The PME summation was performed with a ~1 Å grid, 5th order *B*-splines, and an Ewald coefficient of 0.4. Timings are computed on a 6-core, 2.66 GHz Intel Xeon processor for 100 energy evaluations using standard (non-Ewald) and DPME Overlap Damped Dispersion.

4.4 Results

4.4.1 Model Accuracy

The question that we are attempting to answer in this study is whether or not a damped dispersion model that is consistent with an underlying electrostatic model is demonstrably more accurate relative to *ab initio* data than simpler counterparts. To test this question, we employed a two-step approach. First, we compared the pure London Dispersion model with our new Overlap Damped Dispersion model on the S101x7 database. Then we compared these models to recently published work fitting the S101x7 database with a buffered 14-7 potential function.

To compare the damped and non-damped $1/r^6$ dispersion potentials we fit both to the SAPT2+ dispersion values from the S101x7 database. The results of these fits are presented in table 4.1 and figure 4.3.

	London Dispersion	**Overlap Damped Dispersion**
Total Root Mean Square Error (RMSE)	1.19	0.52
Short-range RMSE (0.7 – 0.8x equil dist)	1.52	0.65
Long-range RMSE (0.9 – 1.1x equil dist)	1.04	0.46

Table 4.1. Goodness of Fit on S101x7 Database (kcal/mol).

Clearly the Overlap Damped Dispersion potential performs better on this set of data, displaying a total root mean square error of 0.52 kcal/mol as opposed to 1.2 kcal/mol for the pure London Dispersion function. Figure 4.3 illustrates how the Overlap Damped Dispersion model consistently fits the SAPT dispersion data better than the non-damped model over a range of interactions energies.

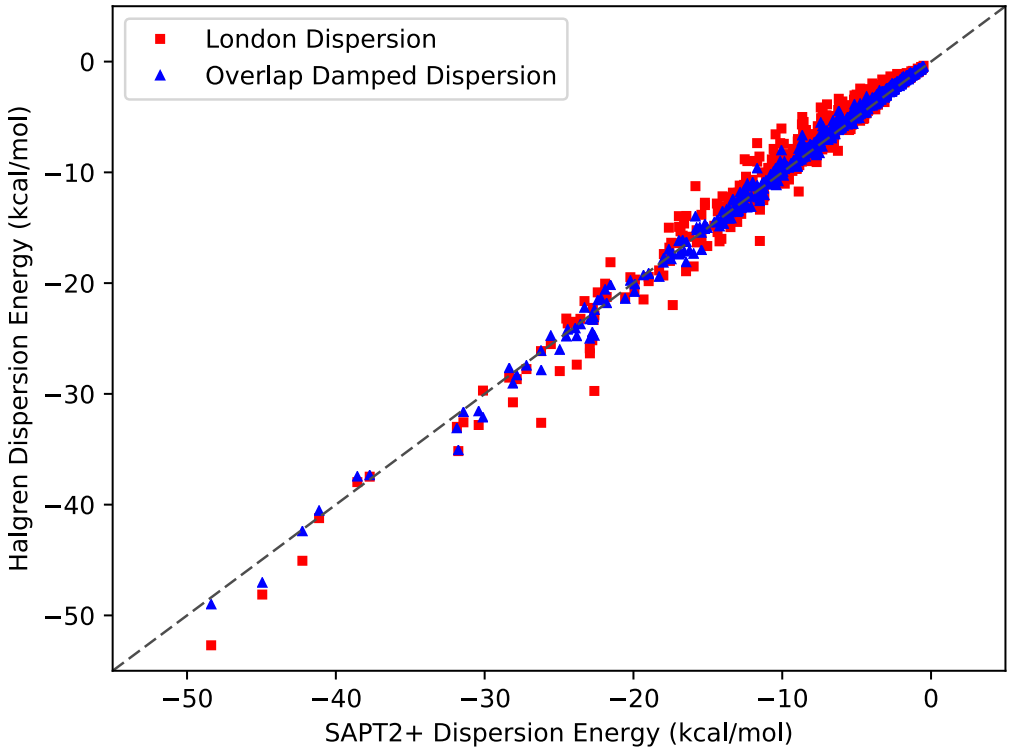


Figure 4.3. Damped and Undamped Dispersion Models against SAPT2+ Dispersion Energies.

The diagonal, y = x dashed line indicates perfect agreement. The Overlap Damped Dispersion model produces a significantly improved fit.

Moreover, the difference in fit quality between the short-range and long-range points shown in table 4.1 is much smaller for the Overlap Damped Dispersion model. This seems to indicate that the damping is having the short-range effect we hoped it might. This issue will be examined further in the robustness tests.

It is instructive to note exactly what is and is not being fit in these two models. For both models the only parameters being fit are one C_6 coefficient per atom class. (Atom class definitions can be found in table 4.2. They are identical to those defined in Ref. 11.) It bears emphasizing that for the Overlap Damped Dispersion model the damping parameters (α_i in

equation 4.26) are not allowed to vary; they are fixed at the values determined in Ref. 11. These values, recapitulated here in table 4.2,

Element	**Atom Class**	**α ($Å^{-1}$)**
	non-polar	3.2484
Hydrogen (H)	aromatic	3.4437
	polar, water	3.2632
	sp^3	3.5898
Carbon (C)	aromatic	3.2057
	sp^2	3.1286
	sp^3	4.0135
Nitrogen (N)	aromatic	3.6358
	sp^2	3.7071
Oxygen (O)	sp^3, hydroxyl, water	4.1615
	aromatic	4.3778
	sp^2, carbonyl	3.7321
Phosphorous (P)	phosphate	2.7476
Sulfur (S)	sulfide	3.3112
	sulfur IV	2.6247
Fluorine (F)	organofluoride	4.4675
Chlorine (Cl)	organochloride	3.4749
Bromine (Br)	organobromide	3.6696

Table 4.2. Fixed Electrostatic Damping Parameters.

describe the physical extent of an atom's electron distribution. They were fit to the SAPT electrostatic energies of the same S101x7 database in the previous study. A comparison of the C_6 parameters between the damped and non-damped models in table 4.3 shows that the Damped Dispersion model exhibits a smoother variation within classes.

Element	**Atom Class**	**London Dispersion C_6 (Å^6 kcal/mol)**	**Damped Dispersion C_6 (Å^6 kcal/mol)**
	non-polar	3.4118	6.3960
Hydrogen (H)	aromatic	4.7993	5.7678
	polar, water	0.9114	5.1133
	sp^3	28.5333	18.1732
Carbon (C)	aromatic	23.2125	23.3605
	sp^2	26.1301	23.0103
	sp^3	33.6562	21.4927
Nitrogen (N)	aromatic	18.2114	19.7421
	sp^2	30.6586	19.4543
	sp^3, hydroxyl, water	25.5861	15.1656
Oxygen (O)	aromatic	25.2794	14.8569
	sp^2, carbonyl	23.1181	18.4344
Phosphorous (P)	phosphate	46.4113	44.8658
Sulfur (S)	sulfide	62.1844	52.8970
	sulfur IV	39.0781	59.2558
Fluorine (F)	organofluoride	15.0568	13.6549

Chlorine (Cl)	organochloride	44.4420	45.7799
Bromine (Br)	organobromide	59.9587	62.0655

Table 4.3. Model C_6 Parameters

The fact that a similar set of parameters produces a damped dispersion model that yields a fit that is 0.5 kcal/mol better than the non-damped model, despite having the exact same number of fitting parameters, is instructive. It shows us that the quality is not due to any extra flexibility in the fitting procedure. This hints that our model may be seizing some of the same physical reality captured in the electrostatics model.

The London Dispersion model is widely used, but it is certainly not the only simple dispersion model used in molecular mechanics force fields. One alternative is Halgren's buffered 14-7 potential.[18] As discussed in section 2, the $1/r^6$ term is only the first term in the expansion of the dispersion energy. The buffered 14-7 potential,

$$U_{vdW} = \sum_{vdw} \sum_{i \neq j} \varepsilon_{ij} \left(\frac{1+\delta}{\rho_{ij}+\delta} \right)^7 \left(\frac{1+\gamma}{\rho_{ij}^7+\gamma} - 2 \right), \qquad \rho_{ij} = \frac{r_{ij}}{\sigma_{ij}}, \tag{4.35}$$

attempts to accommodate higher order terms by means of the buffered $1/r^7$ attractive term to describe dispersion. The buffered 14-7 van der Waals potential has been used in a number of force fields, including AMOEBA, for which a good amount of analysis involving the S101x7 database has already been done. In a recent study Qi, Wang and Ren fit the buffered 14-7 van der Waals potential to the sum of the exchange-repulsion and dispersion data from the S101x7 database yielding a model they call "vdw2016".[51] Given the quality of the total van der Waals energy reported, we set out to see how the corresponding dispersion energies compared to $1/r^6$ derived functions.

To assess the performance of the dispersion part of the vdw2016 model we performed calculations using only the attractive part of the buffered 14-7 potential defined in equation 4.35. The vdw2016 model differs slightly from the damped and non-damped $1/r^6$ dispersion models in its number of atom classes. Where we define just 18 atom classes for the molecules in S101, Qi, Wang and Ren find they need 28 to accurately model the van der Waals energy. For each class they allowed two parameters to vary: the well depth, ε, and radius, σ. Despite this greater flexibility in parameters, the vdw2016 model performs very poorly on predicting the dispersion part of the van der Waals energy. As is clearly seen in figure 4.4, it is not nearly attractive enough.

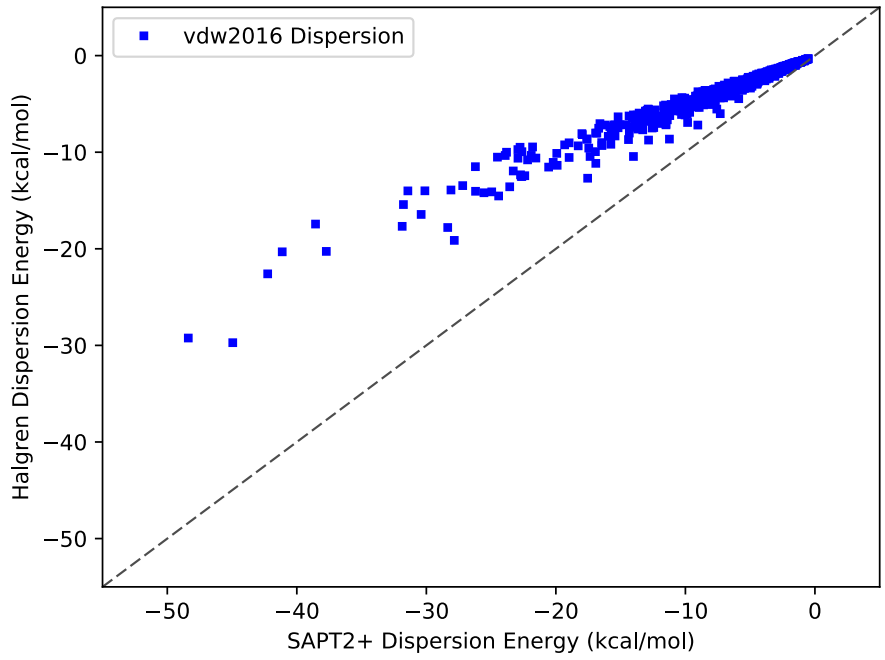


Figure 4.4. vdw2016 against SAPT2+ Dispersion.

The diagonal, y = x dashed line indicates perfect agreement. The vdw2016 model systematically underestimates the magnitude of the dispersion energy.

This is unsurprising given the nature of the fit that was performed. Since the target data was the sum of the exchange-repulsion and dispersion energies the fit is highly skewed by the exchange-repulsion energy. The exchange-repulsion can often be an order of magnitude large than dispersion, especially at short-range, and thus drives values obtained for the fit. This does not mean that vdw2016 does not make an adequate empirical total van der Waals model (indeed buffered 14-7 has almost always been used in its totality), but it does mean that this parameterization will not work as a stand-alone dispersion model if the goal is to reduce cancellation of errors.

While the vdw2016 model has been shown to yield good van der Waals energies, it does so to the detriment of having a separate and interpretable dispersion model. To attempt to remedy this, we performed a second fit of the buffered 14-7 van der Waals form to the S101x7 dataset, vdw2017, where the exchange-repulsion and dispersion components were fit independently. The results of this fit are shown in figure 4.5.

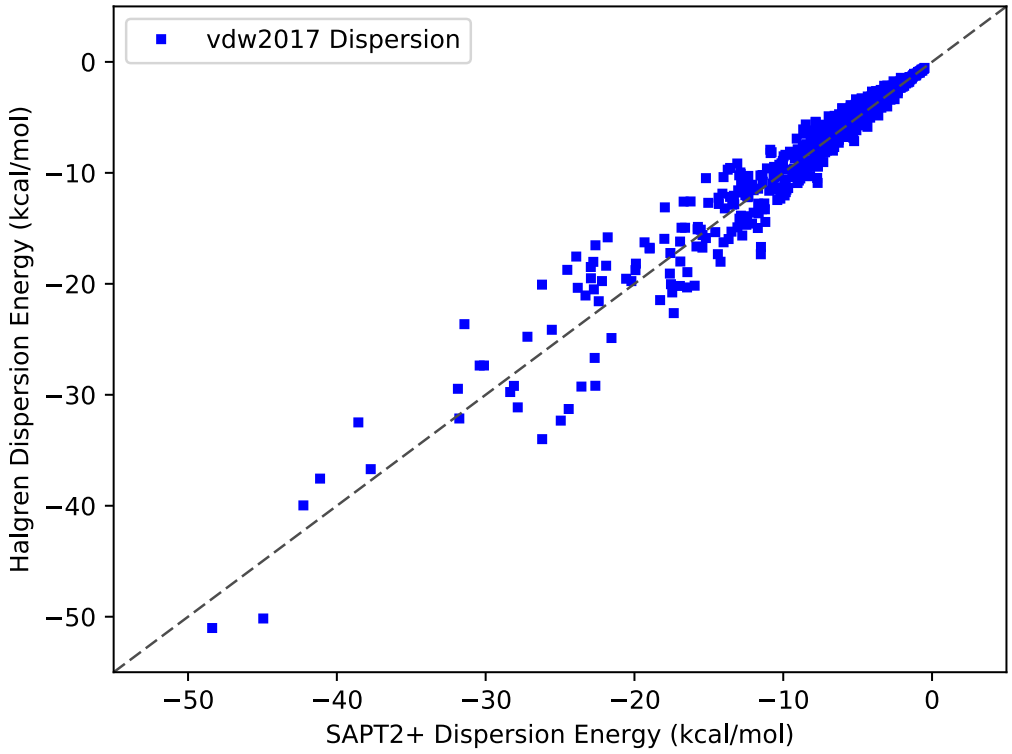


Figure 4.5. vdw2017 Dispersion against SAPT 2+ Dispersion Energies.

The diagonal y = x dashed line indicates perfect agreement. The vdw2017 model RMS error is 1.6 kcal/mol.

One can see that the systematic deviation in dispersion that plagues the vdw2016 model is largely alleviated in the new fit. However, the root mean square error for vdw2107 dispersion remains at 1.6 kcal/mol. This occurs despite preserving the extra flexibility of having 28 atom classes. This seems to show that while the buffered 14-7 may have a fortunate cancellation of errors for the total van der Waals energy, a $1/r^6$ asymptotic function is a more natural fit to the pure dispersion interaction.

Comparing the overall fits of the Halgren dispersion potentials to the (damped or non-damped) London Dispersion potentials it is clear that the latter produce a better fit to the S101x7

dataset. Since the empirical buffered 14-7 potential seems to offer no advantage in accuracy for dispersion, there is no reason to further pursue it as a viable, interpretable dispersion model for the purposes of this study. The next step is to assess whether the advantage in accuracy of the damped dispersion model is worth the extra complexity and computational effort.

4.4.2 Model Robustness

Although the Overlap Damped Dispersion model shows a better fit to the S101x7 dispersion dataset, we would like to be sure that this advantage over the simpler London Dispersion model is robust. To test this point, we employed two separate validation assessments. First, we interrogated the quality of the fit with regard to intermolecular distance. Here our aim was to ascertain which of the two functions is a more natural fit to the data. Second, we applied both models to cases outside of the S101 suite of dimers.

The S101x7 dataset contains sets of dimers arranged at seven different intermolecular distances (0.7, 0.8 0.9, 0.95, 1.0, 1.05 and 1.1 times the equilibrium distance). Because this data set includes a good amount of information about close contact points, we want to be sure that our models fit the short-range points well without sacrificing asymptotic behavior. To judge the long-range fit, we excluded all of the 0.7 and 0.8 times equilibrium data points and then reoptimized the parameters. The results, presented in the “Long-range” entries of table 4.4, show that for this near-equilibrium regime the London Dispersion and Overlap Damped Dispersion models give comparable fits.

The test of robustness is to then use the parameters that come out of the near-equilibrium fits and evaluate each model on the close-contact points that were left out of the fit. This shows how well the shape of the function matches the intrinsic shape of the dispersion dissociation

curve at short-range. As can be seen in table 4.4, there is a difference between the London Dispersion and Overlap Damped Dispersion models.

	London Dispersion	**Overlap Damped Dispersion**
Total Root Mean Square Error (RMSE)	3.12	0.67
Total Mean Signed Error (MSE)	0.91	-0.31
Short-range RMSE (0.7–0.8x equil dist)	5.55	0.84
Long-range RMSE (0.9–1.1x equil dist)	1.15	0.59
Short-range MSE (0.7–0.8x equil dist)	2.71	-0.12
Long-range MSE (0.9–1.1x equil dist)	0.20	-0.38

Table 4.4. Dispersion Model Robustness Test (kcal/mol)

The total RMS error of the Overlap Damped Dispersion model increases modestly when the close-contact points are included, as should be expected since these points were not included in the fit. The total RMS error of the non-damped London Dispersion model, however, rises dramatically. While the long-range quality of fit (those points that were included in the fit) is good for both models, the short-range quality (those points not included in the fit but included in the robustness test) is very different between the two models. The RMS error on the short-range test points with the Overlap Damped Dispersion model is less than 1 kcal/mol, but the RMS error of the London Dispersion model is over 5 kcal/mol. These errors are clearly caused by the inability of a simple $1/r^6$ function to adequately describe both the asymptotic and overlap regimes. Moreover, it is clear from table 4.4 that the Overlap Damped Dispersion model is not sacrificing accuracy in the asymptotic regime, where it is actually slightly better than the London

Dispersion model. A handful of illustrative examples show how the London Dispersion model fit to near-equilibrium points systematically predicts the dispersion energy to be too attractive. Figure 4.6 shows three examples where this effect is pronounced.

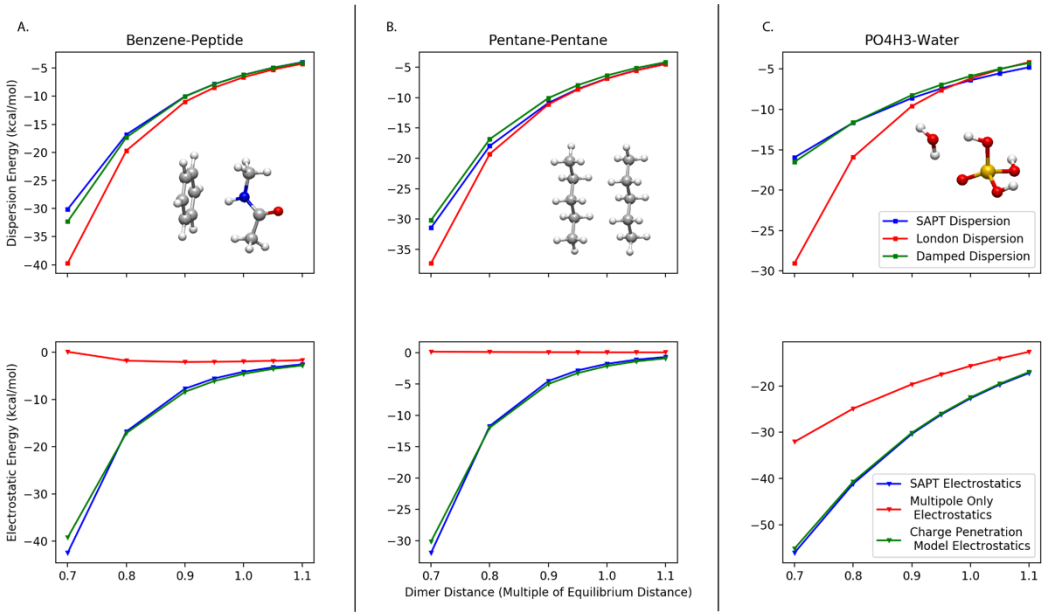


Figure 4.6. Examples of Dispersion (top row) and Electrostatic (bottom row) corrections for charge density overlap in (A) benzene-peptide, (B) pentane-pentane and (C) water-PO_4H_3 interactions.

The x-axis indicates dimer intermolecular distance as a fraction of each dimer's equilibrium separation. In all three examples the undamped, "classical" model diverges from the *ab initio* result at short-range, while the damped model follows the *ab initio* curve closely.

The pentane-pentane, benzene-peptide and water-PO_4H_3 interactions are all examples of important component interactions in biology. They also exhibit the importance of damping the dispersion energy at short range for an *ab initio*-based force field. Clearly, including the damping function from the electrostatic model improves the agreement with SAPT dispersion data at the closest points.

We suggest the effectiveness of this damping is fundamentally tied to the overlap in charge distributions. If we compare the non-damped London Dispersion curves with their corresponding non-damped electrostatic curves (no charge penetration correction) in figure 4.6, we see that the divergence of non-damped energies from their SAPT counterparts occurs at roughly the same separation. This suggests that deviation from the $1/r^6$ asymptotic behavior in the dispersion energy at short-range is also attributable to the overlap in charge distributions. We know the point multipole expansion model for electrostatic interactions is rigorously accurate until charge distributions begin to overlap. The fact the divergence in the point dipole derived dispersion energy occurs at a similar distance, suggests that the same effect is driving this phenomenon. Moreover, the fact that the exact same parameters can be used to accommodate the change from the asymptotic behavior for both electrostatics and dispersion indicates that these are separate manifestations of the same physical reality.

Although the London Dispersion model may be simpler and computationally less expensive than the Overlap Damped Dispersion model, it is clear from this robustness test that the latter provides a much better description of the dispersion interaction that spans both the close contact and asymptotic regimes. For the S101x7 dataset, generally, the 0.8x points represent the closest intermolecular distance for liquids at ambient conditions. The robustness test shows that a force field using the Overlap Damped Dispersion model will rely less on cancellation of errors in this area than an undamped model. Importantly, we note that the Overlap Damped Dispersion model retains the $1/r^6$ dependence at long-range as the damping factor quickly approaches unity when charge distributions no longer overlap. This gives us confidence that the shape of the function is well suited to the intrinsic shape of the dispersion dissociation curve.

4.4.3 Model Analysis and Validation

Having established the capability of the Overlap Damped Dispersion model for short-range interactions, we can ask how well this model performs on specific, important systems. Dispersion plays an important role in a range of biomolecular interactions and one should hope a good model would describe such interactions accurately. Two instructive examples are water-water interactions and benzene stacking interactions. Both also happen to be instances where charge density overlap plays a role in their short-range interactions.

The balance between water-water and water-biomolecule interactions is known to be important to accurate simulations of biomolecules. Recently, a study by Piana and co-workers demonstrated that simulations with a few commonly used water models overpredict the compactness of disordered and partially disordered proteins.[52] They suggest that this occurs because these typical water models underestimate water-water and water-protein dispersion interactions relative to *ab initio* dimer calculations. This conclusion may be overstated, since for the TIP3P and SPCE models discussed, this underestimation is largely handled through cancellation of errors within the rest of the force field. A goal of our work, however, is to reduce this reliance on such cancellation. The Overlap Damped Dispersion model directly addresses this problem through accurate prediction of the water dimer dispersion energy curve. As shown in figure 4.7, the damped model gives good overall agreement with the shape of the SAPT dispersion data.

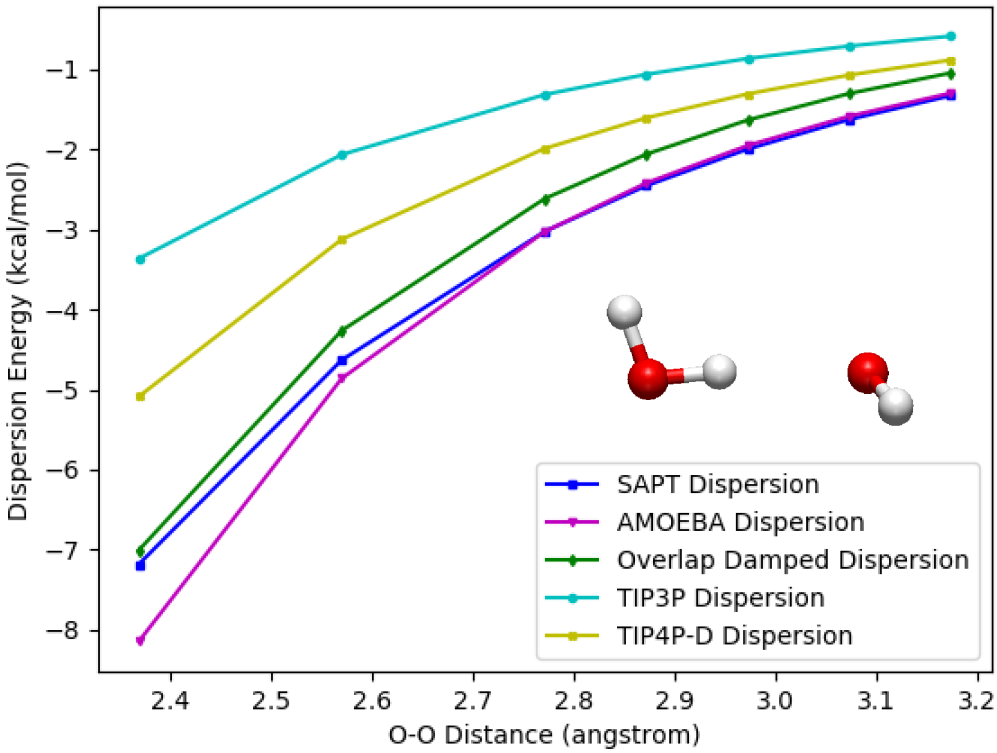


Figure 4.7. Performance of Various Water Dispersion Models against SAPT2+ Dispersion.

Model dispersion energies are compared to SAPT2+ dispersion energies for a range of intermolecular distance of the water dimer. TIP3P[53] and TIP4P-D[52] are undamped ~$1/r^6$ models, AMOEBA is the attractive, ~$1/r^7$, component of the buffered 14-7 potential with parameters from the water03 force field[19], and the Overlap Damped Dispersion model is from this work.

Also shown in figure 4.7 is the quality of the fit of the AMOEBA water03[19] model. Since this AMOEBA model is polarizable, one would expect the dispersion part of its van der Waals function should be close to the *ab initio* dispersion energy due to less reliance on cancellation of errors. Indeed, near equilibrium this model produces excellent agreement, but at short range the dispersion energy becomes too negative. While the absolute energy error may not be large for these close points, one can see that the error in the slope is much greater. At an O-O distance of

~2.6 Å, for example – well sampled in ambient water[54] – one can see that the water03 dispersion force is slightly too attractive. Recent work has suggested that cancellation of errors is responsible for the condensed phase behavior of AMOEBA water[55,56], but as these compensatory components are removed for the next generation of the model, the error in the dispersion becomes more important to address directly. It is not novel to suggest that modeling the short and long-range dispersion interactions simultaneously requires a damping function. What is shown here, however, is that a simple, rationally constructed and minimally parameterized model yields excellent agreement for this important interaction.

Another example interaction of importance in biomolecular modeling is the benzene "pi-stacking" interaction. In addition to being an important exemplar for nucleic acid structure and drug binding, this interaction falls into the qualitative "dispersion-bound" category[57], so accurately modeling it is imperative for a dispersion model. Figure 4.8 shows the performance of the Overlap Damped Dispersion model against SAPT.

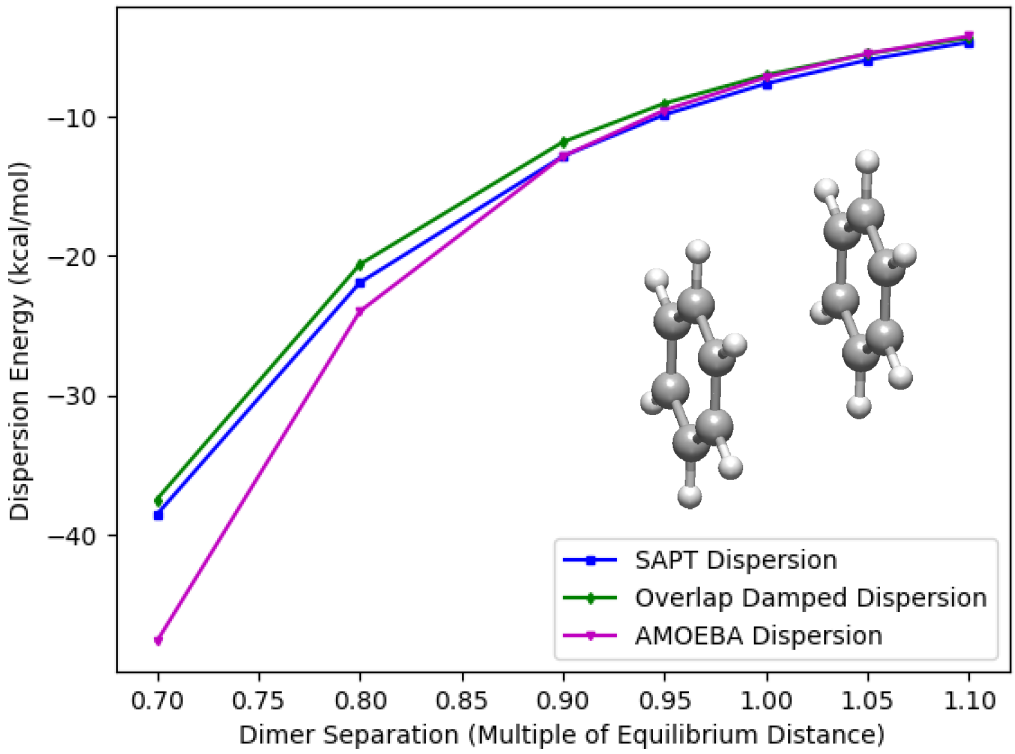


Figure 4.8. Benzene Dimer Dispersion.

Model dispersion energies are compared to SAPT2+ dispersion energies for a range of intermolecular distance of the benzene dimer. The AMOEBA model functional form is the same is in figure 4.7, with parameters taken from the AMOEBA09 force field.

One can see that the agreement of the Overlap Damped Dispersion model with the SAPT data is excellent across all benzene dimer separations. As was observed for the water dimer, the AMOEBA model produces good agreement near equilibrium, but characteristically deteriorates at short range. In particular the divergence begins at ~0.85 of the equilibrium separation or a ~3.2 Å C-C distance. This distance is a close contact for liquid benzene at room temperature and 1 atm – it falls near the start of the radial distribution function.[58] As a model system, it is also close to the stacking distance between bases in B-DNA, ~3.3 Å. Figure 4.8 shows that for small, but

relevant distances like this the shape of SAPT dispersion is more closely matched by the Overlap Damped Dispersion Model. Although less dramatic than with electrostatics, the deviation at short-range of the London Dispersion model is due to the same phenomenon that drives the divergence in the electrostatics of the benzene dimer. Figure 4.8 shows us that the same treatment can be applied to fix the errors in both classical models.

Finally, to check that the success at accurately fitting the S101x7 dataset is not the result of overfitting, we employ a validation test on a system outside of the training set. For this purpose, we chose to test the dispersion component of nucleic acid base stacking interactions. In previously published work, Parker and Sherrill performed SAPT energy decomposition analysis calculations on a set of nucleic acid structures to evaluate the performance of current force fields. In order to assess how well a given model reproduces the energy components of base stacking interactions, Parker and Sherrill performed SAPT calculations at equilibrium and near equilibrium geometries of all ten possible two base-pair steps of DNA: AATT, ACGT, AGCT, ATAT, CATG, CGCG, GATG, GCGC, GGCC and TATA. To generate trial geometries, Parker and Sherrill systematically varied the six geometrical degrees of freedom illustrated in figure 4.9 (shift, slide, rise, tilt, roll and twist) for each base-pair step. See reference 49 for structure generation specifics and calculation details.

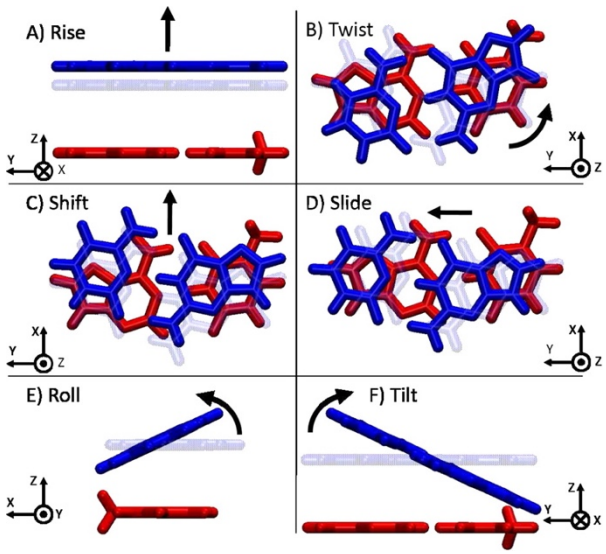


Figure 4.9. Illustration of the six degrees of freedom explored for nucleic acid structures.

The example shown is for the AC:GT base step. Reprinted with permission from Parker, T. M. & Sherrill, C. D. Assessment of Empirical Models versus High-Accuracy Ab Initio Methods for Nucleobase Stacking: Evaluating the Importance of Charge Penetration. *Journal of Chemical Theory and Computation* 11, 4197-4204. Copyright 2015 American Chemical Society.

To see how our model measures up, we compare the published nucleic acid SAPT dispersion energies with the dispersion energies predicted by our Overlap Damped Dispersion model using atom types as defined in table 4.3. The results for this test set are shown in figure 4.10.

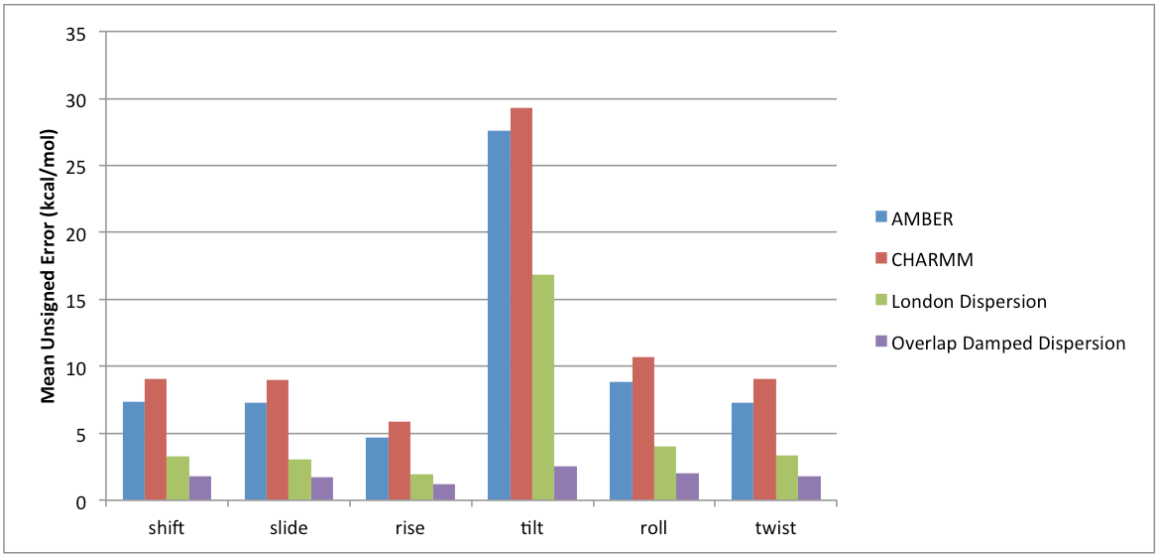


Figure 4.10. Mean Unsigned Error in Dispersion Energy for Nucleic Acid Structures.

Model error is relative to SAPT for each of the six structural parameters. The Overlap Damped Dispersion model reduces the error in the dispersion across all six degrees of freedom. Amber and CHARMM results from Ref. 49.

There are two important features to point out in the figure. First, one will notice that the London Dispersion model performs better than either the Amber or CHARMM nucleic acid dispersion models despite having an identical functional form. This, as noted by Parker and Sherrill, is primarily due to cancellation of errors in the partial charge models. These models do not explicitly include the effects of charge penetration, so the dispersion function is called upon to absorb some of the error in the electrostatics. What Parker and Sherrill find, however is that while this cancellation of errors strategy produces total energies within 1 kcal/mol relative to DW-CCSD(T**) for structures near B-form DNA, the error in the total energy across the range of potential energy surface scans is closer to 2 kcal/mol with some errors over 10 kcal/mol even for attractive points on the surface. One can see from figure 4.10 that parameterizing a $1/r^6$ (London Dispersion) model directly to SAPT reduces some of the need for cancellation of error, but not all. The second and more important feature one observes is the agreement throughout the

potential energy surface of the Overlap Damped Dispersion model. In addition to relieving itself of the cancellation of errors burden, one can see the damped model provides a minimum factor of two improvement in the mean unsigned error over the undamped London Dispersion model for every degree of freedom. This has little to do with the behavior of the dispersion energy at equilibrium; the divergence occurs primarily for structures where the electron densities of the two base-pairs start to overlap.

As an instructive example, take the change in dispersion energy with respect to tilt angle for the CATG base step shown in figure 4.11.

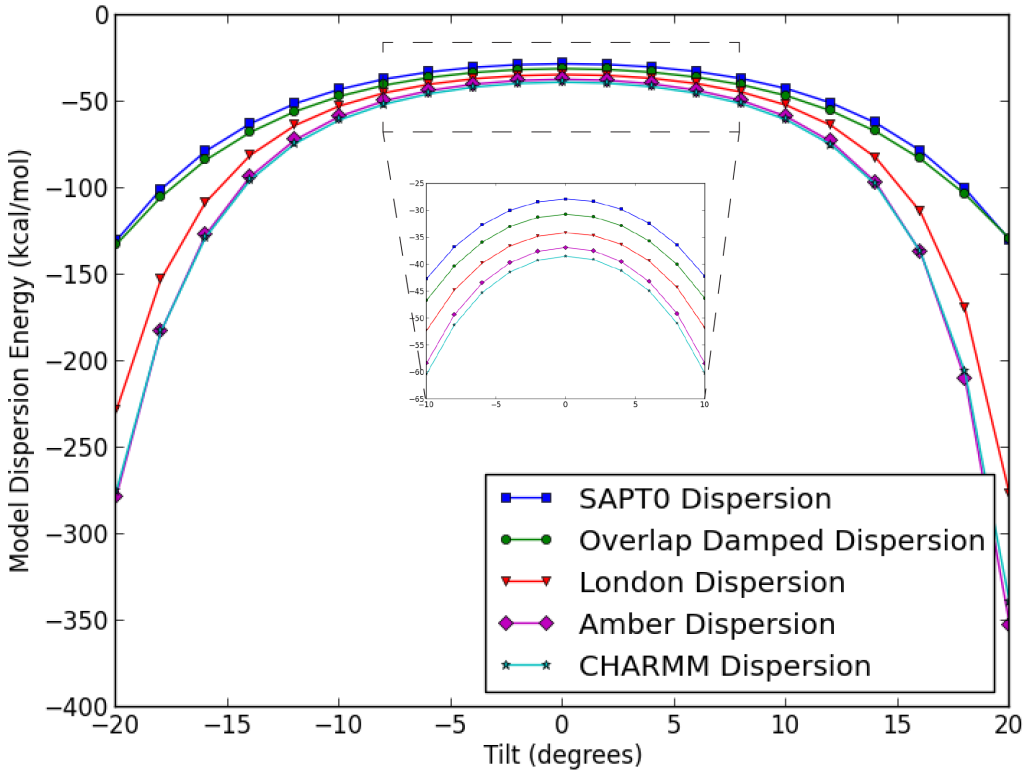


Figure 4.11. Dispersion Energy of CATG Interaction *vs.* Tilt.

The non-damped dispersion models uniformly overestimate the magnitude of the dispersion energy as the angle varies from equilibrium in either direction. The Overlap Damped Dispersion Model predicts the shape of the SAPT curve at both equilibrium and near-equilibrium geometries.

One can see that at equilibrium both the London and Overlap Damped Dispersion models predict the SAPT dispersion energy with good precision. However, as one changes the tilt angle in either direction, the dispersion energy of the undamped model diverges quickly from the SAPT while the Overlap Damped Dispersion model follows the shape of the SAPT curve with fidelity. This trend holds across all six degrees of freedom and all ten base pair steps. Plots like figure 4.11 for each combination are available in the supplementary information of reference 59. The divergence observed for non-damped models matters because it is not simply confined to high total energy areas of the DNA potential energy surface. In fact, Parker and Sherrill showed that for the stacked A-C pair (one half of the CATG base step) at a tilt angle of -15^{o} the total energy is -5 kcal/mol. This is only 0.5 kcal/mol above the minimum total energy of -5.5 kcal/mol. Figure 4.11 suggests that in order to accurately model this region of the potential energy surface without large cancellation of errors a damped dispersion model is necessary.

4.5 Dispersion Particle Mesh Ewald Summation

The accuracy of a molecular mechanics model is important, but so too is its efficiency. A good dispersion model must not only be accurate, but also fast to compute. While the accuracy of the Overlap Damped Dispersion Model has been solidly established in this paper, the exponentials required for its evaluation have the potential to slow potential energy calculations. To make the Overlap Damped Dispersion model computationally efficient and tractable for use in biomolecular simulations, we have implemented the model with particle mesh Ewald (PME) summation in the Tinker molecular mechanics software package. In this section we present a brief overview of the damped dispersion PME implementation and show how this

implementation provides a substantial speed and accuracy improvement over the standard cutoff-based van der Waals implementation.

Ewald summation is classically considered to be primarily a solution to the pairwise long-range electrostatics problem. The Σ1/r electrostatic potential is conditionally convergent which makes direct computation of the electrostatic energy of a periodic system difficult. To circumvent this problem, Ewald methods split the sum into short-range and long-range parts, with short-range part being computed directly and the long-range via Fourier transformation. This separation not only makes periodic calculations possible, but also increases the speed with which the energy and gradient can be evaluated.

The same method can be applied to the dispersion energy calculation. Here we note that the following derivation is by no means original. In fact, Essman and co-workers proposed the possibility of using particle mesh Ewald summation for dispersion in their 1995 paper describing the method of smooth particle mesh Ewald summation.[60] We present here a brief summary simply to show that the inclusion of a damping term in this case does not change the ability to use the method.

The total dispersion energy, as given by eq. 33 is:

$$U_{dispersion}^{damp} = -\sum_{i \neq j} \frac{C_6^i C_6^j}{r_{ij}^6} \left(f_{dispersion}^{damp} \right)_{ij}^2 . \tag{4.36}$$

This can be split into a short-range part, a long-range part and a "self" term,

$$U_{total}^{dispersion} = U_{short-range}^{dispersion} + U_{long-range}^{dispersion} + U_{self}^{dispersion} \tag{4.37}$$

with

$$U_{short-range}^{dispersion} = \sum_{i \neq j} \frac{C_6^i C_6^j}{r_{ij}^6} \left(f_{dispersion}^{damp} \right)_{ij}^2 \left(1 + \beta^2 r_{ij}^2 + \frac{1}{2} \beta^4 r_{ij}^4 \right) e^{-\beta^2 r_{ij}^2}$$

$$U_{long-range}^{dispersion} = \frac{2\pi^{9/2}}{3V} \sum_{\mathbf{m} \neq 0} |m|^3 \left[\frac{1}{2\left(\pi |m| / \beta\right)^3} \left(1 - 2\left(\pi |m| / \beta\right)^2 \right) e^{\left(-\left(\pi |m| / \beta\right)^2 + \sqrt{\pi} erfc\left(\pi |m| / \beta\right) \right)} \right] \hat{S}(\mathbf{m}) \hat{S}(-\mathbf{m})$$

$$U_{self}^{dispersion} = -\frac{\beta^6}{12} \sum_i C_i^2 + \frac{\beta^3 \pi^{3/2}}{6V} \left(\sum_i C_i \right)^2 \qquad (4.38a,b,c)$$

Equations 4.38a and 4.38b are commonly known as the direct space sum and reciprocal space sum, respectively. The variable, β, is the parameter determining the Gaussian width, **m** is defined by the reciprocal lattice vectors, **a**, as $\mathbf{m} = m_1\mathbf{a_1}^* + m_2\mathbf{a_2}^* + m_3\mathbf{a_3}^*$, and V is the volume of the unit cell. The structure factor, S, is defined for dispersion as,

$$\hat{S} = \sum_j C_j e^{i 2\pi (\mathbf{m} \cdot \mathbf{r}_j)} \quad . \qquad (4.39)$$

The summation in eq. 38b is handled in the same manner as the reciprocal space sum for electrostatics. Tinker uses the FFTW (Fastest Fourier Transform in the West) package to perform the needed Fourier transforms.[61] To speed the calculation and because the dispersion energy decreases quickly with distance, eq. 38a, the direct space sum, is truncated at a fixed distance.

For simple dispersion PME the choice of direct space cutoff matters very little; one simply chooses a cutoff that balances computational effort between direct space and reciprocal space. For Overlap Damped Dispersion PME, however, some care must be taken with the choice. This is because eqs. 38a, b and c as written, do not strictly sum to eq. 36. This imbalance is caused by the presence of the damping function in the direct space sum, without an equivalent component in the reciprocal space. In practice, however, this is easily overcome with a rational choice of cutoff distance. The function f_{damp} goes to unity very quickly with distance (much

faster than $1/r^6$ goes to zero), so reasonable cutoff distances are easy to obtain. Figure 4.12 shows dispersion energy as a function of cutoff distance.

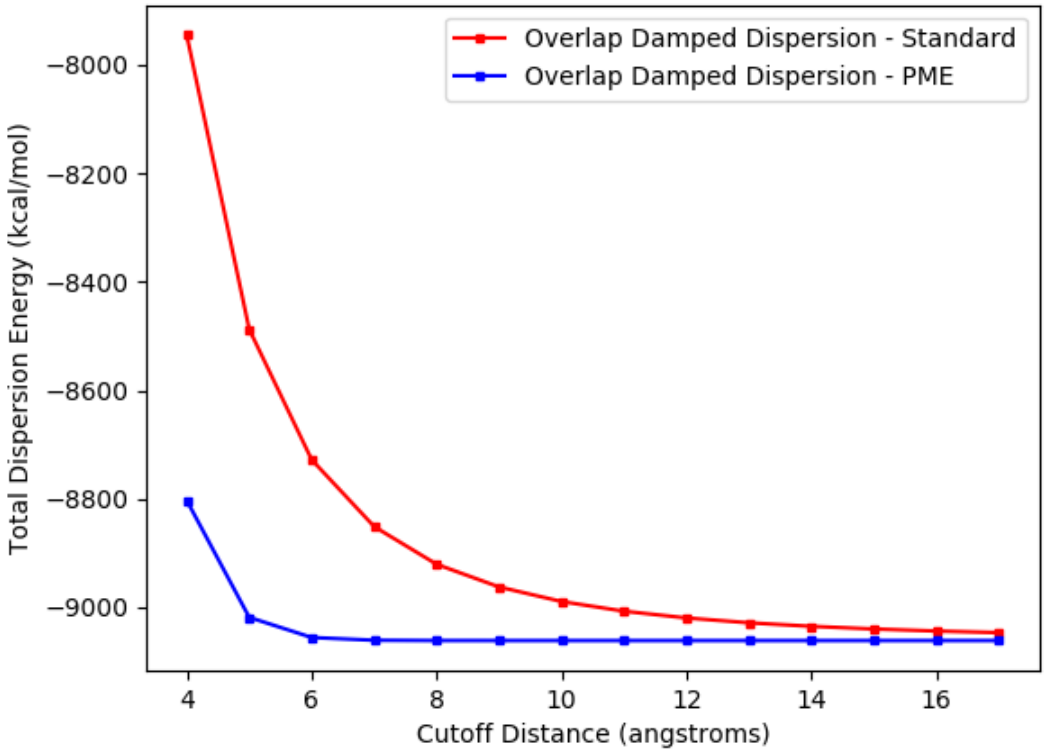


Figure 4.12. Cutoff Distance Convergence of the Overlap Damped Dispersion Model.

The total dispersion energy of a 36 Å water box is shown for the standard and particle mesh Ewald (PME) implementations of the Overlap Damped Dispersion model. The cutoff of the PME implementation refers to the cutoff of the real space summation. Computational details are enumerated in section 3.

One can see that for cutoffs longer than 6 Å, the energy of the PME implementation is constant, due to the fact that f_{damp} is effectively unity for all atom pairs outside of this radius. For our model we chose this cutoff of 6 Å to balance the direct and reciprocal space computational effort. Comparing the PME and non-PME curves in figure 4.12 illustrates an obvious advantage of using Ewald summation for dispersion interactions. While the non-PME curve converges to

the asymptotic total energy very slowly with cutoff distance, the Ewald sum is converged within the 6 Å cutoff distance. The slow convergence of the non-Ewald sum is the reason many molecular mechanics models use 12 to 16 Å cutoffs or van der Waals corrections term for their dispersion interactions.

Because our model is not forced to use a longer cutoff distance, it can be faster to compute than standard dispersion models. As a point of reference, in the AMOEBA model, the van der Waals calculations currently comprises 10-15% of the total calculation time. While this is certainly not the bottleneck for efficiency, it is important to keep this relative cost low. In figure 4.13 we compare computation times for our PME implementation of the model with the standard implementation for various cutoff distances.

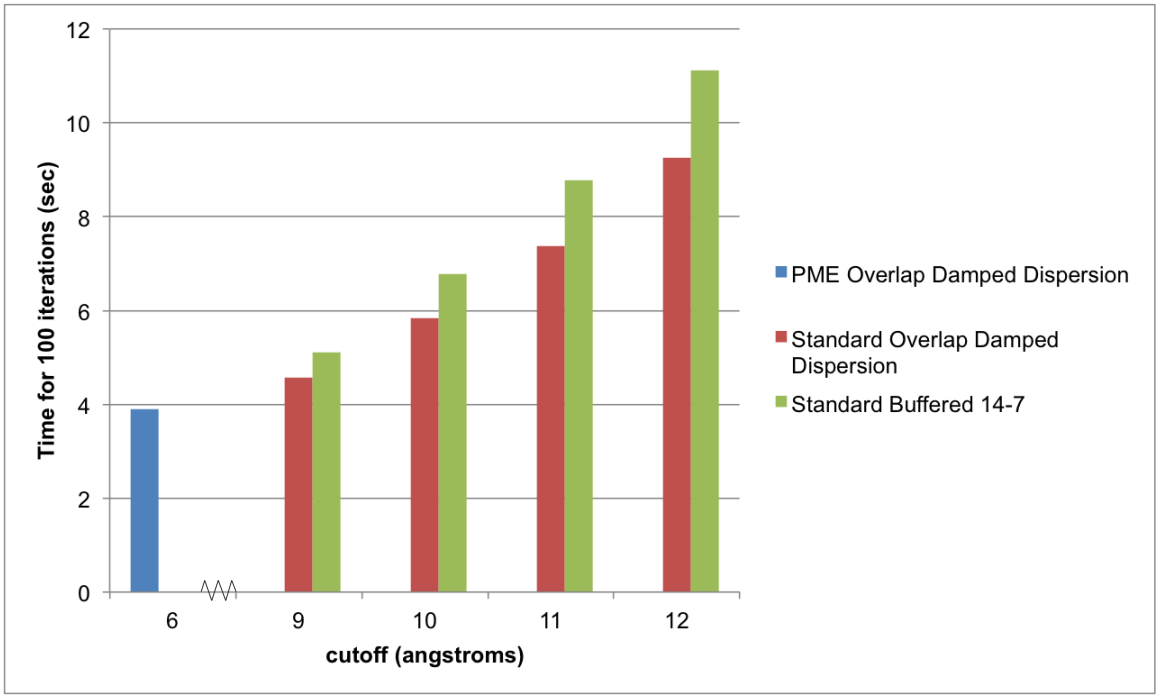


Figure 4.13. Computational Effort for Overlap Damped Dispersion.

The time to complete 100 iterations of Overlap Damped Dispersion and Buffered 14-7 is plotted as a function of cutoff distance.

Figure 4.13 shows that for standard implementation cutoff distances of greater than 9 Å, the PME implementation of the Overlap Damped Dispersion Model provides a performance boost relative to the non-PME implementation. As a point of reference, figure 4.13 also shows the speed of the AMOEBA buffered 14-7 van der Waals functional form with its suggested cutoff distance of 12 Å. Even compared to this model, which has no required exponential evaluation, the Overlap Damped Dispersion PME model provides a factor of 2.5 speed increase. The use of particle mesh Ewald summation minimizes the work needed in real space, thus enabling the use of our more complicated and accurate functional form without loss of computational efficiency.

Finally, our model benefits from utilization of simple combination rules. Using a multiplicative combination rule as indicated in equation 4.33 makes our DPME method exact. It is worth noting that many popular force fields, including all three mentioned in this paper (Amber, CHARMM and AMOEBA) use additive combining rules for van der Waals parameters. Particle Mesh Ewald methods can be used with additive combining rules in an approximate manner first proposed by Erik Lindahl and co-workers.[62,63] This method prevents what would otherwise be a discontinuity in the forces at the cutoff distance, but does introduce complexity when switching from direct to reciprocal space combining rules.[15] The Overlap Damped Dispersion PME model avoids this by explicitly parameterizing to a multiplicative combining rule. This implementation is certainly not unique or novel, but it is important to the future use of the Overlap Damped Dispersion model in a complete force field because it shows that the benefit of a more physics-based short-range model can be realized at no increase in cost.

4.6 Discussion and Conclusions

The universe of possible molecular mechanics models is immense. To discriminate between models and decide which models work best with each other, we must evaluate them based upon the goals they wish to achieve. The goal of our proposed model is use in biomolecular modeling, molecular dynamics simulation and free energy calculations. To this end it is important for the model to be accurate, transferable and interpretable. Accuracy has obvious importance for describing the interactions of biological molecules with fidelity, but transferability and interpretability are no less important. In this last section we summarize how the Overlap Damped Dispersion model measures on each of these attributes.

The data in table 4.1 show that when measured against Symmetry Adapted Perturbation Theory, damping is necessary to achieve 1 kcal/mol accuracy for the S101x7 data set. Interestingly, the Overlap Damped Dispersion model is also shown to be more accurate than the attractive part of the slightly more complex buffered 14-7 potential. The root of this behavior seems to lie with the behavior of the dispersion energy at short-range. Figure 4.6 shows that a damping function is necessary to fit both close-contact and large-separation dimer points. In most force fields this inaccuracy at short-range is handled through cancellation of error. Relying on such cancellation, however, will not work as force fields become more accurate and, more importantly, is not guaranteed to function favorably across the wide variety of intermolecular interactions that occur in biomolecular applications.

The transferability of the Overlap Damped Dispersion model is coupled to this idea of eliminating a reliance on cancellation of error. The most notable feature of the model, aside from its accuracy, is the fact that it has no additional adjustable parameters beyond the simple London Dispersion model. The damping function, as presented in section 2, is entirely determined by the electrostatic model presented in Ref. 11 with no additional fitting or parameterization. This

property of the model suggests two things. First, the Overlap Damped Dispersion model is easily transferable to a range of chemical space because of the limited number of parameters. Evidence of this is shown in figures 4.10 and 4.11 where the S101-fitted parameters were used to predict the dispersion energy of nucleic acid base stacking interactions. Second, it hints at a physical reality behind the model. The fact that parameters generated through fitting to intermolecular electrostatic interactions, where density overlap is the determining factor in short-range interactions, works well for our dispersion model is a strong indicator that the same phenomenon is driving short-range dispersion.

This physical picture of short-range dispersion is what makes the Overlap Damped Dispersion model interpretable. There are many damping functions that can be used to correct for the behavior of the dispersion energy at short-range. Several of these damping functions can likely be parameterized to yield results against Symmetry Adapted Perturbation theory that are as accurate as those presented here. What the current model offers over the alternatives is a physical interpretation. In this model the dispersion interaction is the result of the electrostatic interaction between the instantaneous induced dipoles of two distinct charge distributions. This characteristic of the model is valuable for two reasons. First, it gives us some intuition about the nature of intermolecular interactions. Second, the interpretability of the model makes it easier to extend the model to new areas of chemical space. We make no claim that the 18 atom classes used in this paper will accurately describe the all of the variety of chemistries in organic molecules. What is clear, however, is that the interpretation of the damping parameter as a measure of an atom's charge distribution gives a clear path to determining new parameters where necessary. In this way the Overlap Damped Dispersion model is systematically improvable. As advanced molecular mechanics models evolve this property will be important to their ongoing

development. As models grow to explicitly take into account the short-range interactions between molecules, this dispersion model fits neatly into that framework.

Accurately modeling the short-range interactions between molecules is important to making trustworthy predictions on a range of biomolecular problems. drug binding[64], intrinsically disordered protein behavior[65] and nucleic acid structure[66] are all areas where advanced force fields have been shown to be necessary for correct predictions. As models get more complex the tendency is to accumulate additional parameters and with them empiricism. In the case of dispersion, this paper shows that a simple physical model can be employed that adds no new parameters while reducing the need to rely on cancellation of errors. Moreover, combined with a dispersion particle mesh Ewald implementation, the evaluation of the necessary equations can be achieved as fast or faster than the standard implementation of simple non-damped models. This yields a simple, physically interpretable model ready for the next generation of advanced molecular mechanics models. We are currently working on incorporating this model, along with the previously published charge penetration function, into a complete force field.

4.7 Further Work

The model presented in this published work uses the first density model as published in the work from Chapter 2. However, as explained in the "Further Work" section of Chapter 2, the model density has changed slightly. The only part of the derivation that changes is equation 4.26. The damping factors are replaced with

$$f_3^{overlap} = \begin{cases} 1 - \left(1 + \alpha R + \frac{1}{2}(\alpha R)^2 + \frac{7}{48}(\alpha R)^3 + \frac{1}{48}(\alpha R)^4\right) e^{-\alpha R}, & \alpha_i = \alpha_j \\ 1 - A^2\left(1 + \alpha_i R + \frac{1}{2}(\alpha_i R)^2\right) e^{-\alpha_i R} - B^2\left(1 + \alpha_j R + \frac{1}{2}\left(\alpha_j R\right)^2\right) e^{-\alpha_j R} - \\ 2A^2 B(1 + \alpha_i R) e^{-\alpha_i R} - 2B^2 A\left(1 + \alpha_j R\right) e^{-\alpha_j R}, & \alpha_i \neq \alpha_j \end{cases} \tag{4.40}$$

and

$$f_5^{overlap} = \begin{cases} 1 - \left(1 + \alpha R + \frac{1}{2}(\alpha R)^2 + \frac{1}{6}(\alpha R)^3 + \frac{1}{24}(\alpha R)^4 + \frac{1}{144}(\alpha R)^5\right) e^{-\alpha R}, & \alpha_i = \alpha_j \\ 1 - A^2\left(1 + \alpha_i R + \frac{1}{2}(\alpha_i R)^2 + \frac{1}{6}(\alpha_i R)^3\right) e^{-\alpha_i R} - \\ B^2\left(1 + \alpha_j R + \frac{1}{2}\left(\alpha_j R\right)^2 + \frac{1}{6}\left(\alpha_j R\right)^3\right) e^{-\alpha_j R} - \\ 2A^2 B\left(1 + \alpha_i R + \frac{1}{3}(\alpha_i R)^2\right) e^{-\alpha_i R} - \\ 2B^2 A\left(1 + \alpha_j R + \frac{1}{3}\left(\alpha_j R\right)^2\right) e^{-\alpha_j R}, & \alpha_i \neq \alpha_j \end{cases} \tag{4.41}$$

with

$$B = \frac{\alpha_i^2}{\left(\alpha_i^2 - \alpha_j^2\right)^2}, \quad A = \frac{\alpha_j^2}{\left(\alpha_j^2 - \alpha_i^2\right)^2}. \tag{4.42}$$

The remainder of the derivation remains identical to that in section 4.2.

Just as with the electrostatics, the amount that this changes the quality of fit and validation examples is nearly negligible. The updated C_6 parameters are tabulated in table 4.5.

class	name	C_6
1	H (nonpolar)	6.4475
2	H (nonpolar, Alkane)	7.0153
3	H (polar, NH/N aromatic)	2.2238
4	H (polar, OH)	3.3513
5	H (aromatic, CH)	4.7752
6	H (polar, SH)	5.1227
7	C (sp3)	16.7895

8	C (sp3, Alkane)	18.1472
9	C (sp2, Ethene)	27.5956
10	C (sp2, CO)	4.1502
11	C (sp)	20
12	C (aromatic, CC)	25.1602
13	C (aromatic, CX)	18.8259
14	N (sp3)	33.4759
15	N (sp2)	34.2019
16	N (aromatic)	33.531
17	O (sp3, hydroxyl, water)	22.5286
18	O (sp2, carbonyl)	30.0411
19	O (O− in AcO−)	33.4521
20	O (O− in HPO42−)	33.76
21	O (O− in H2PO4−)	31.3032
22	O (O in H3PO4)	32.5028
23	P (phosphate)	3.8493
24	S (sulfide, RSH)	60.5601
25	S (sulfur IV, DMSO)	34.642
26	F (organofluorine)	13.4133
27	Cl (organochloride)	45.0491
28	Br (organobromine)	64.3023

Table 4.5. C_6 parameters for final HIPPO dispersion model.

These use the "new" density model described in the "Further Work" section of Chapter 2.

The difference in computational cost is likewise insignificant since the damping factors are already being computed for electrostatics. The update to the "new" density model, however, is important to the consistency of the model. With this update, all of the intermolecular potential energy terms in HIPPO are constructed from a singular density model per atom.

4.8 References

1 Bergonzo, C., Henriksen, N. M., Roe, D. R. & Cheatham, T. E. Highly Sampled Tetranucleotide and Tetraloop Motifs Enable Evaluation of Common RNA Force Fields. *RNA* **21**, 1578-1590 (2015).

2 Fan, F. *et al.* Mechanical Properties of Amorphous LixSi Alloys: A Reactive Force Field Study. *Model Simul Mater Sc* **21**, 074002 (2013).

3 McDaniel, J. G. & Schmidt, J. Next-Generation Force Fields from Symmetry-Adapted Perturbation Theory. *Annu Rev Phys Chem* **67**, 467-488 (2016).
4 Tafipolsky, M. & Ansorg, K. Toward a Pshycially Motivated Force Field: Hydrogen Bond Directionality from a Symmetry-Adapted Perturbation Theory Perspective. *J Chem Theory Comput* **12**, 1267-1279 (2016).
5 Schmidt, J., Yu, K. & McDaniel, J. G. Trensferable Next-Generation Force Fields from Simple Liquids to Complex Materials. *Acc Chem Res* **48**, 548-556 (2015).
6 Misquitta, A. J., Podeszwa, R., Jeziorski, B. & Szalewicz, K. Intermolecular Potentials based on Symmetry-Adapted Perturbation Theory with Dispersion Energies from Time-Dependent Density-Functional Calculations. *J Chem Phys* **123**, 214103 (2005).
7 Taylor, D. E., Rob, F., Rice, B. M., Podeszwa, R. & Szalewicz, K. A Molecular Dynamics Study of 1,1-Diamino-2,2-dinitroethylene (FOX-7) Crystal Using a Symmetry Adapted Perturbation Theory-based Intermolecular Force Field. *Phys Chem Chem Phys* **13**, 16629-16636 (2011).
8 Gordon, M. S., Slipchenko, L. V., Li, H. & Jensen, J. H. The Effective Fragment Potential: A General Method for Predicting Intermolecular Interactions. *Annu Rep Comput Chem* **3**, 177-193 (2007).
9 Cisneros, G. A. Application of Gaussian Electrostatic Model (GEM) Distributed Multipoles in the AMOEBA Force Field. *J Chem Theory Comput* **8**, 5072-5080 (2012).
10 El Khoury, L. *et al.* Importance of Explicit Smeared Lone-Pairs in Ansiotropic Polarizable Molecular Mechanics. Torture Track Angular Tests for Exchange-Repulsion and Charge Transfer Contributions. *J Comput Chem* **38**, 1897-1920 (2017).
11 Rackers, J. A. *et al.* An Optimized Charge Penetration Model for Use with the AMOEBA Force Field. *Phys Chem Chem Phys* **19**, 276-291 (2017).
12 Kolář, M., Kubař, T. & Hobza, P. On the Role of London Dispersion Forces in Biomolecular Structure Determination. *J Phys Chem B* **115**, 8038-8046 (2011).
13 Riley, K. E. & Hobza, P. Investigations into the Nature of Halogen Bonding Including Symmetry Adapted Perturbation Theory Analyses. *J Chem Theory Comput* **4**, 232-242 (2008).
14 Riley, K. E. & Hobza, P. The Relative Roles of Electrostatics and Dispersion in the Stabilization of Halogen Bonds. *Phys Chem Chem Phys* **15**, 17742-17751 (2013).
15 Leonard, A. N. *et al.* Comparison of Additive and Polarizable Models with Explicit Treatment of Long-Range Lennard-Jones Interactions using Alkane Simulations. *J Chem Theory Comput* (2017).
16 Wang, L.-P. *et al.* Building a More Predictive Protein Force Field: A Systematic and Reproducible Route to AMBER-FB15. *J Phys Chem B* **121**, 4023-4039 (2017).
17 Huang, J. *et al.* CHARMM36m: An Improved Force Field for Folded and Intrinsically Disordered Proteins. *Nat Methods* **14**, 71-73 (2017).
18 Halgren, T. A. The Representation of van der Waals (vdW) Interactions in Molecular Mechanics Force Fields: Potential Form, Combination Rules, and vdW Parameters. *J Am Chem Soc* **114**, 7827-7843 (1992).
19 Ren, P. & Ponder, J. W. Polarizable Atomic Multipole Water Model for Molecular Mechanics Simulation. *J Phys Chem B* **107**, 5933-5947 (2003).
20 Brooks, F. C. Convergence of Intermolecular Force Series. *Phys Rev* **86**, 92-97 (1952).
21 Ahlrichs, R., Penco, R. & Scoles, G. Intermolecular Forces in Simple Systems. *Chem Phys* **19**, 119-130 (1977).

22 Tang, K. T. & Toennies, J. P. An Improved Simple Model for the van der Waals Potential Based on Universal Damping Functions for the Dispersion Coefficients. *J Chem Phys* **80**, 3726-3741 (1984).
23 Grimme, S. Density Functional Theory with London Dispersion Corrections. *WIRES Comput Mol Sci* **1**, 211-228 (2011).
24 Wu, Q. & Yang, W. Empirical Correction to Density Functional Theory for van der Waals Interactions. *J Chem Phys* **116**, 515-524 (2002).
25 Chai, J.-D. & Head-Gordon, M. Long-Range Corrected Hybrid Density Functionals with Damped Atom–Atom Dispersion Corrections. *Phys Chem Chem Phys* **10**, 6615-6620 (2008).
26 Johnson, E. R. & Becke, A. D. A Post-Hartree-Fock Model of Intermolecular Interactions. *J Chem Phys* **123**, 024101 (2005).
27 Slipchenko, L. V. & Gordon, M. S. Damping Functions in the Effective Fragment Potential Method. *Mol Phys* **107**, 999-1016 (2009).
28 Adamovic, I. & Gordon, M. S. Dynamic Polarizability, Dispersion Coefficient C6 and Dispersion Energy in the Effective Fragment Potential Method. *Mol Phys* **103**, 379-387 (2005).
29 Verma, P., Wang, B., Fernandez, L. E. & Truhlar, D. G. Physical Molecular Mechanics Method for Damped Dispersion. *J Phys Chem A* **121**, 2855-2862 (2017).
30 Maitland, G. C., Rigby, M., Smith, E. B. & Wakeham, W. A. (Oxford University Press, 1981).
31 Misquitta, A. J. & Stone, A. J. Dispersion Energies for Small Organic Molecules: First Row Atoms. *Mol Phys* **106**, 1631-1643 (2008).
32 Gresh, N. *et al.* Complexes of a Zn-Metalloenzyme Binding Site with Hydroxamate-Containing Ligands. A Case for Detailed Benchmarkings of Polarizable Molecular Mechanics/Dynamics Potentials when the Experimental Binding Structure is Unknown. *J Comput Chem* **37**, 2770-2782 (2016).
33 Bytautas, L. & Ruedenberg, K. Correlation Energy and Dispersion Interaction in the ab Initio Potential Energy Curve of the Neon Dimer. *J Comput Chem* **37**, 2770-2782 (2008).
34 Buckingham, A. D. Theory of Long-Range Dispersion Forces. *Discuss Faraday Soc* **40**, 232-238 (1965).
35 Xu, P., Zahariev, F. & Gordon, M. S. The R-7 Dispersion Interaction in the General Effective Fragment Potential Method. *J Chem Theory Comput* **10**, 1576-1587 (2014).
36 Guidez, E. B., Xu, P. & Gordon, M. S. Derivation and Implementation of the Gradient of the R-7 Dispersion Interaction in the Effective Fragment Potential Method. *J Phys Chem A* **120**, 639-647 (2016).
37 Wang, Q. *et al.* General Model for Treating Short-Range Electrostatic Penetration in a Molecular Mechanics Force Field. *J Chem Theory Comput* **11**, 2609-2618 (2015).
38 Narth, C. *et al.* Scalable Improvement of SPME Multipolar Electrostatics in Anisotropic Polarizable Molecular Mechanics Using a General Short-Range Penetration Correction Up To Quadrupoles. *J Comput Chem* **37**, 494-506 (2016).
39 Coulson, C. Two-Centre Integrals Occurring in the Theory of Molecular Structure. *Math Proc Cambridge* **38**, 210-223 (1942).
40 Ren, P. *SAPT Database between Organic Molecules, Protein Side Chain Analogs*, <http://biomol.bme.utexas.edu/~ch38988/s101x7> (2015).

41 Jeziorski, B., Moszynski, R. & Szalewicz, K. Perturbation Theory Approach to Intermolecular Potential Energy Surfaces of van der Waals Complexes. *Chem Rev* **94**, 1887-1930 (1994).
42 Parker, T. M., Burns, L. A., Parrish, R. M., Ryno, A. G. & Sherrill, C. D. Levels of Symmetry Adapted Perturbation Theory (SAPT). I. Efficiency and Performance for Interaction Energies. *J Chem Phys* **140**, 094106 (2014).
43 Dunning Jr., T. H. Gaussian Basis Sets for Use in Correlated Molecular Calculations. I. The Atoms Boron through Neon and Hydrogen. *J Chem Phys* **90**, 1007-1023 (1989).
44 Kendall, R. A., Dunning Jr., T. H. & Harrison, R. J. Electron Affinities of the First-Row Atoms Revisited. Systematic Basis Sets and Wave Functions. *J Chem Phys* **96**, 6796-6806 (1992).
45 Halkier, A. *et al.* Basis-set Convergence in Correlated Calculations on Ne, N2, and H2O. *Chem Phys Lett* **286**, 243-252 (1998).
46 Hohenstein, E. G. & Sherrill, C. D. Density Fitting of Intramonomer Correlation Effects in Symmetry-Adapted Perturbation Theory. *J Chem Phys* **133**, 014101 (2010).
47 Turney, J. M. *et al.* Psi4: An Open-Source ab Initio Electronic Structure Program. *WIRES Comput Mol Sci* **2**, 556-565 (2012).
48 Parrish, R. M. *et al.* Psi4 1.1: An Open-Source Electronic Structure Program Emphasizing Automation, Advanced Libraries, and Interoperability. *J Chem Theory Comput* **13**, 3185-3197 (2017).
49 Parker, T. M. & Sherrill, C. D. Assessment of Empirical Models versus High-Accuracy Ab Initio Methods for Nucleobase Stacking: Evaluating the Importance of Charge Penetration. *J Chem Theory Comput* **11**, 4197-4204 (2015).
50 Rackers, J. A. *Tinker Github Repository, DPME Branch*, <https://github.com/JoshRackers/tinker/tree/dpme> (2017).
51 Qi, R., Wang, Q. & Ren, P. General van der Waals Potential for Common Organic Molecules. *Bioorgan Med Chem* **24**, 4911-4919 (2016).
52 Piana, S., Donchev, A. G., Robustelli, P. & Shaw, D. E. Water Dispersion Interactions Strongly Influence Simulated Structural Properties of Disordered Protein States. *J Phys Chem B* **119**, 5113-5123 (2015).
53 Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J Chem Phys* **79**, 926-935 (1983).
54 Brookes, D. H. & Head-Gordon, T. Family of Oxygen-Oxygen Radial Dirsribution Functions for Water. *J Phys Chem Lett* **6**, 2938-2943 (2015).
55 Mao, Y., Demerdash, O., Head-Gordon, M. & Head-Gordon, T. Assessing Ion-Water Interactions in the AMOEBA Force Field Using Energy Decomposition Analysis of Electronic Structure Calculations. *J Chem Theory Comput* **12**, 5422-5437 (2016).
56 Demerdash, O., Mao, Y., Liu, T., Head-Gordon, M. & Head-Gordon, T. Assessing Many-Body Contributions to Intermolecular Interactions of the AMOEBA Force Field Using Energy Decomposition Analysis of Electronic Structure Calculations. *J Chem Phys* **147**, 161721 (2017).
57 Rezác, J., Riley, K. E. & Hobza, P. S66: A Well-Balanced Database of Benchmark Interaction Energies Relevant to Biomolecular Structures. *J Chem Theory Comput* **7**, 2427-2438 (2011).

58 Jorgensen, W. L. & Severance, D. L. Aromatic-Aromatic Interactions: Free Energy Profiles for the Banzene Dimer in Water, Chloroform, and Liquid Benzene. *J Am Chem Soc* **112**, 4768-4774 (1990).
59 Rackers, J. A., Liu, C., Ren, P. & Ponder, J. W. A physically grounded damped dispersion model with particle mesh Ewald summation. *The Journal of chemical physics* **149**, 084115 (2018).
60 Essmann, U. *et al.* A Smooth Particle Mesh Ewald Method. *J Chem Phys* **103**, 8577-8593 (1995).
61 Frigo, M. & Johnson, S. G. The Design and Implementation of FFTW3. *P IEEE* **93**, 216-231 (2005).
62 Wennberg, C. L., Murtola, T., Hess, B. & Lindahl, E. Lennard-Jones Lattice Summation in Bilayer Simulations has Critical Effects on Surface Tension and LIpid Properties. *J Chem Theory Comput* **9**, 3527-3537 (2013).
63 Wennberg, C. L. *et al.* Direct-Space Corrections Enable Fast and Accurate Lorentz-Berthelot Combination Rule Lennard-Jones Lattice Summation. *J Chem Theory Comput* **11**, 5737-5746 (2015).
64 Bell, D. R. *et al.* Calculating Binding Free Energies of Host–Guest Systems using the AMOEBA Polarizable Force Field. *Phys Chem Chem Phys* **18**, 30261-30269 (2016).
65 Huang, J. & MacKerell, A. D. Force Field Development and Simulations of Intrinsically Disordered Proteins. *Curr Opin Struc Biol* **48**, 40-48 (2018).
66 Parker, T. M., Hohenstein, E. G., Parrish, R. M., Hud, N. V. & Sherrill, C. D. Quantum-Mechanical Analysis of the Energetic Contributions to π Stacking in Nucleic Acids versus Rise, Twist, and Slide. *J Am Chem Soc* **135**, 1306-1316 (2013).

Chapter 5: Repulsion

The final piece of the HIPPO force field is the repulsion model. When the decision to abandon the empirical Buffered 14-7 potential was made, that created the need for a repulsion model that would match the SAPT exchange-repulsion component. Moreover, I wanted the repulsion model to not just fit the SAPT data, but to do so with a function that was derived from the physical meaning of exchange-repulsion in the SAPT approach. The result was the HIPPO Multipolar Pauli Repulsion model. Throughout this chapter and throughout this dissertation I refer to "Exchange-Repulsion" and "Pauli Repulsion" interchangeably. Unlike the distinction between induction and polarization, this is purely semantics. The two terms, in the way they are used in this dissertation, mean exactly the same thing: the effect on the intermolecular energy caused by imposing wavefunction antisymmetry. In simpler terms, molecules repel each other at short range because the Pauli exclusion rule prevents molecular densities from overlapping too much. This effect and how it can be accounted for in terms of a classical function will be explained in detail in this chapter.

The following is taken from a published paper in which I describe the Multipolar Pauli Repulsion model. It contains a derivation of the functional form as well as practical considerations such as agreement with SAPT and computational efficiency. In some ways this part of the model is the most novel element of HIPPO. It is the first anisotropic atomic repulsion function to ever be included in a biomolecular force field. This is a big leap from the Lennard-Jones-like van der Waals models that are prevalent in standard molecular mechanics force fields. Despite being novel to the field, however, the Multipolar Pauli Repulsion model is not unique in the context of HIPPO. The functional form is based on the exact same density model as the rest of the force field. This means that every term of the HIPPO functional form is based, in some

way, on that density model. So, while the model is notable in its own right, perhaps the most important thing about it is that it closes the loop on a simple, but radical interpretation of HIPPO as a new class of force field. For 30+ years, the standard force field has relied on the approximation that atoms can be represented as points. This assumption permeates both the electrostatics and van der Waals terms of these force fields. HIPPO takes one step further. It represents each atom as a point core plus an outer electron density. This new assumption permeates the entirety of the model in a consistent manner. This chapter sets in place the final piece of that model, but what remains to be tested, in the work presented in Chapter 6 and on-going work, is whether this density approximation will yield more predictive results for biomolecular simulations.

5.1 Introduction

The beauty of classical physics models is not *that* they work for describing most of our world, but rather *why* they work. While it is necessary for physics-based models to be accurate and predictive, these qualities alone are not sufficient. A true classical model must also be interpretable; that is to say, it must be a derivable approximation from first principles. Good classical models of everyday phenomenon are not just lucky; they are the true limiting behavior of fundamental physical laws. Nowhere is this principle more essential or more often forgotten than in molecular modeling. To solve difficult questions such as drug binding specificity or nanotube formation, fields from biology to materials science have come to rely on molecular mechanics models, or force fields, to generate hypotheses and make predictions. These predictions are only as good as the model used to make them, meaning that every part of the force field must contain a sufficient level of accuracy. In particular, one of the most important parts of any force field is the term responsible for intermolecular Pauli repulsion. This term,

which also goes by the names “steric” or “exchange” repulsion, is too often described as a mysterious “quantum mechanical” force. This could not be farther from the truth.[1-3] This paper intends to show that intermolecular Pauli repulsion is a simple consequence of Coulomb’s law and furthermore that this interpretation leads to an accurate classical model of Pauli repulsion.

The level of model accuracy needed is always a function of its intended use. For force fields this standard is referred to as “chemical accuracy”, which we define here as a fidelity in computed energies to within 1 kcal/mol. While this requirement is not universal, and higher accuracy may well be required for many applications in molecular interactions, a particular example will serve to rationalize its importance. A primary, current use for biomolecular force fields is prediction of drug binding affinities. Because of the relation $\Delta G = RT \log(K_D)$, at room temperature every order of magnitude in the binding affinity translates into 1.36 kcal/mol in the free energy of binding. This is of great practical importance since a factor of 10 variance in a drug’s binding affinity can be the difference between a medicine that hits a specific target *vs*. one that binds non-specifically. One of the most important factors to achieving this level of accuracy for atom-based force fields is anisotropy. Work with the AMOEBA (Atomic Multipole Optimized Energetics for Biomolecular Applications) force field has shown that adding atomic anisotropy *via* multipoles is an excellent way to make molecular mechanics models more accurate. Atomic multipoles are necessary to accurately predict the electrostatic potential around drug-like molecules[4] and they are needed to reproduce hydrogen bond geometries in water[5], proteins[6] and nucleic acids[7]. Recent work with AMOEBA, in the SAMPL6 challenge, showed that this more accurate model produces generally more accurate binding free predictions than its fixed charge counterparts.[8] All of this indicates that to reach the goal of “chemical accuracy”, the next generation of force fields will need to account for atomic and molecular anisotropy.

While the importance of anisotropy is broadly recognized for the electrostatic portions of molecular mechanics models, it is widely overlooked in the other terms, particularly Pauli repulsion. The repulsion term is acutely important because in most force fields it is the only consistent source of positive energy in the system. This means that in the delicate balance between attraction and repulsion that exists in all condensed phase systems, the repulsion term shoulders the burden for most of the second half of that equation. In the canonical, minimum energy water dimer for example, *ab initio* Symmetry Adapted Perturbation Theory (SAPT) energy decomposition calculations show that while the electrostatic, induction and dispersion contributions to the interaction are all negative, the exchange-repulsion is the only source of positive energy in the system.[9] Despite this, nearly all common biomolecular force fields including AMOEBA use relatively simple, isotropic repulsion schemes. Most commonly these are the $1/r^{12}$ repulsive Lennard-Jones potential, a Buckingham exponential form or, in the case of AMOEBA, Halgren's buffered 14-7 potential. Strong evidence is emerging that these isotropic Pauli repulsion functions may not be accurate enough to ensure "chemical accuracy" in biomolecular applications. Recent work by Anthony Stone has shown a strong angular dependence of the Pauli repulsion energy in halogen bonding interactions.[4] Furthermore, these sigma hole interactions are vitally important to the drug discovery process[10], and evidence has emerged that this angular dependence is a large source of the selective binding geometries of sigma hole associated drug candidates.[10-14] In order meet the standard needed for predictive biomolecular applications, we will need an accurate, physics-based, anisotropic Pauli repulsion model.

This work aims to present a classical Pauli repulsion model that is anisotropic and efficient to compute. In previous work we have shown that a rough model of atomic charge

density can dramatically improve the accuracy of electrostatic and dispersion models for short-range intermolecular interactions.[15] Here we will show that this same simple density formulation, coupled to an atomic multipole model yields a classical, physics-based model for Pauli repulsion. In addition to providing a qualitatively different level of accuracy compared to standard isotropic empirical models, this model also dispenses with the customary mysticism surrounding repulsion models. There is no attempt to write off another empirical model in terms of "quantum mechanical forces"; this model simply accounts for the loss in nuclear screening, relative to their isolated, unperturbed states, that molecules experience when their charge densities start to overlap. We will go about this in four stages. First, we will build out the theory underlying the model and its classical electrostatic interpretation. Second, we will describe the methods of the study. Third, we will demonstrate the accuracy of our Pauli repulsion model against benchmark SAPT data. And lastly, we will present pertinent discussion and conclusions.

5.1.1 A Brief History of Pauli Repulsion Models

Well before the advent of modern quantum mechanics scientists understood that molecules repel each other at short-range. Johannes van der Waals won a Nobel Prize in 1910 for "his work on the equation of state for gases and liquids", which postulated intermolecular interactions as the source of deviations from the ideal gas law.[16] It was not until the statement of the Pauli Exclusion Principle by Wolfgang Pauli in 1925[17] that physicists understood the explanation for the repulsive intermolecular interactions that keep molecules separated. The first model to approximate this Pauli repulsion phenomenon is due to Sir John Lennard-Jones, a man whose name is indelibly linked to the field of molecular modeling. Lennard-Jones first proposed a general polynomial form of the van der Waals potential in 1924[18] and suggested the now canonical 6-12 formulation in 1931.[19] While the $1/r^6$ attractive term was taken from London's

earlier work on dispersion[20], the $1/r^{12}$ repulsive term was chosen primarily out of convenience. Empirically, Lennard-Jones found that the $1/r^{12}$ term provided an adequate estimate of the repulsive forces between closed-shell atoms near equilibrium. It is worth noting that Lennard-Jones himself had no illusions about the limitations of such a functional form. In his 1931 lecture he acknowledges that for simple systems exchange energies fall off as $e^{-\alpha r}/r$. The Lennard-Jones potential, however, in its canonical form has been enormously influential. It was used in some of the very first molecular dynamics simulations of biological molecules[21] and continues to be the van der Waals function of choice for most popular biomolecular force fields including Amber, CHARMM, GROMOS and OPLS.

The next significant model function for intermolecular repulsion proposed was the simple exponential. Credited to Max Born and Joseph Mayer (1932)[22], and Richard Buckingham (1938)[23] this function has the form $A\ e^{-\alpha r}$. Both papers built on the work of John Slater, who in 1928 worked out the repulsive force between two helium atoms.[24] Slater found the repulsive force to be exponential of the form $P(r)\ e^{-\alpha r}$, where P(r) is a polynomial. Slater proposed, however, that to a reasonable approximation P(r) could be replaced with a constant. Born and Mayer tested this hypothesis on ionic cubic lattices and Buckingham extended their work to *ab initio* noble gas intermolecular forces. It is interesting to note that Lennard-Jones himself actually played a significant role in the development of these models. He, in fact, was responsible for communicating Buckingham's 1938 paper to the Royal Society. Despite its more substantial theoretical underpinnings, the Buckingham or Born-Mayer exponential functions have been less widely utilized in biomolecular force fields. Notably, the MM2, MM3 and MM4 force fields, designed for accurate conformational analysis and gas phase thermodynamics of hydrocarbons and simple organics, use an exponential repulsion function, however wide-spread adoption for

biomolecules was hampered in part due to the increased computational cost of exponential evaluation.[25-27]

There is one additional contribution to the class of simple, isotropic repulsion models that did not come until much later. In 1992 Thomas Halgren introduced the so-called "buffered 14-7" potential in an effort to raise the level of accuracy of van der Waals functions for organic and biomolecular force fields.[28] This function introduced the idea of buffering constants to fix the known systematic short-range over-repulsive nature of the canonical Lennard-Jones function. Halgren revisited the rare gas repulsion calculations that had guided Lennard-Jones's, Born and Mayer's, and Buckingham's model development with modern electronic structure methods and found that the buffered 14-7 form produced better fits than the 12-6 or exp-6 models he tested against. While the model yields good agreement with *ab initio* data, Halgren makes no attempt to justify it theoretically other than to present it as a perturbation on top of the accepted Lennard-Jones form. Henceforth it has commonly been accepted as a simple solution to the known problem of the excessive stiffness of the $1/r^{12}$ repulsive wall. Despite its lack of rigorous theoretical justification, this model has been used with some success in the Merck Molecular and the AMOEBA force fields.[29-31]

While these three models account for the vast majority of Pauli repulsion terms in biomolecular force fields, there are large number of "boutique" molecular mechanics repulsive functions tailored to modeling specific types of compounds. Often intended for specific, high-accuracy applications, these models are frequently anisotropic with a larger number of parameters. While we shall refrain from dissecting every known such potential function, a handful of models are notable. Anthony Stone proposed a water model with an atom-atom exponential repulsion that varied according to the relative local geometries of the interacting

water molecules.[32] Similarly, Misquitta and Stone developed a site-site anisotropic repulsive potential for pyridine that utilizes distributed densities to generate atomic repulsive "shape" parameters that enter the exponential.[33] In another example, the SAPT-5s water model uses an isotropic repulsive potential with a polynomial prefactor, but requires a number of off-atom sites for accuracy.[34] Lastly, an anisotropic short-range model that includes Pauli repulsion along with other short-range effects has been used to model polycyclic aromatic hydrocarbons.[35]

While the early models of Pauli repulsion acknowledged the general exponential nature of the term, it was not until the 1960s that more rigorous justification and specific functional dependence was provided. In 1961 Lionel Salem published the foundational paper for what would come to be known as the "orbital overlap" model of Pauli repulsion. Salem worked out the repulsive force experienced by two interacting helium atoms subject to the Hellman-Feynman Theorem and showed that the repulsion energy can be accurately modeled by S^2/R, where S is the overlap integral between the interacting orbitals and R is the distance between the atoms.[36] Not only did he derive this dependence from first principles, he showed numerically that such an approximation is quite good for the He dimer example. This work was followed by further validation by Musher and Salem[37], Murrell, Randic and Williams[38], and Murrell and Shaw[39]. The basis of this model is classical electrostatics. Salem showed unequivocally in the case of helium that the repulsive interaction experienced at close approach is caused by a depletion in electron density in the overlap region that de-screens the nuclei from each other, causing internuclear repulsion. Furthermore, Salem illustrated that the magnitude of this depletion is proportional to the square of the orbital overlap integral (S^2). Because of the essential problem of defining an orbital in a classical force field, this framework for repulsive models has been less widely used. Two notable exceptions are the SIBFA (Sum of Interactions

between Fragments) and EFP (Effective Fragment Potential) models. The SIBFA repulsive potential depends on the overlap between atom centers, bond centers and lone pairs, as well as a prefactor that accounts for the relative orientation of the interacting pairs.[40,41] EFP uses monomer LMOs (Localized Molecular Orbitals), which makes the potential transferable, but too expensive for large-scale biomolecular simulations.[42,43]

The final class of Pauli repulsion models that merits consideration are "density overlap" models. In 1976, Kita, Noda and Inouye performed molecular beam experiments with Cl^--X and Br^--X, where X = He, Ne, and Ar, and found the repulsive energy was proportional to Ω, the density overlap integral.[44] In 1981, Kim, Kim and Lee arrived at the same conclusion upon examination of experimental noble gas repulsion data.[45] This observation was first turned into a molecular mechanics model by Wheatley and Price, whose use of anisotropic atomic densities yielded an anisotropic repulsion model.[46] This model influenced Stone's original work on the water model mentioned above. It is also the basis for the Pauli repulsion term in the GEM (Gaussian Electrostatic Model) force field[47-50] and recent force fields developed by J.R. Schmidt and co-workers.[51,52] Although this model seems to match experimental data well, it has little in the way of theoretical grounding. Despite its seeming similarity to the orbital overlap model, the density overlap model has a fundamental units problem. This will be discussed in greater detail at the end of section 2. Theoretical considerations notwithstanding, the density overlap model has been parameterized to meet a range of modeling needs.[53-58] Section 2 will shed more light on the nature of this model's empirical success.

This history of Pauli repulsion models shows no clear consensus as to which functional form is best. The choice of model has posed two typical tradeoffs: first, between computational speed and model accuracy and second, between model transferability and number of parameters.

These tradeoffs, however, are not endemic to Pauli repulsion models. Here we present an orbital overlap model that sidesteps both of them. The result is a fast-to-compute, anisotropic, transferable Pauli repulsion function.

5.2 Theory

An explanation of any Pauli repulsion model must start with a basic understanding of the Pauli Exclusion Principle. The Exclusion principle is a consequence of two simple facts: electrons are fermions and they are indistinguishable. Let us consider a system of two electrons, with wave functions, $\phi_A(x_1)$ and $\phi_B(x_2)$. Since these electrons are indistinguishable, we must be able to swap labels and still end up with the same density. One can see that for the simple solution,

$$\phi(x_1, x_2) = \phi_A(x_1)\phi_B(x_2) \tag{1}$$

this condition is not met since it is not necessarily true that:

$$\phi_A(x_1)\phi_B(x_2) \neq \phi_A(x_2)\phi_B(x_1) \tag{2}$$

However, since both sides of equation 2 are solutions to the Schrodinger equation, we can use a linear combination to produce the total wave function:

$$\Phi(x_1, x_2) = \phi_A(x_1)\phi_B(x_2) \pm \phi_A(x_2)\phi_B(x_1) \tag{3}$$

The positive and negative versions of equation 3 correspond to symmetric and antisymmetric wave functions, respectively, and define the difference between bosons and fermions. Because electrons in nature are always observed to have antisymmetric wave functions, they are classified as fermions. The requirement that fermionic wave functions be antisymmetric is the essence of the Pauli Exclusion Principle. For an antisymmetric wavefunction if $\phi_A = \phi_B$, the total wavefunction, Φ, goes to zero, meaning that no two electrons may occupy the same state.

To illustrate how the Pauli Exclusion Principle leads to Pauli repulsion between molecules, let us consider the case of the helium dimer. As the simplest closed-shell dimer, He_2 provides a natural, general example for the repulsion between molecules. Our derivation will closely follow that of Salem's 1961 paper. All equations to follow are presented in atomic units, with $4\pi\varepsilon_0 = 1$.

Consider two helium atoms separated by a large distance, such that they do not interact. In this case the two atoms have distinct wave functions, ϕ_A and ϕ_B where both are real, spherically symmetric, exponentially decaying functions. If these two atoms are brought close enough to each other to interact, the total wave function, now a mix of ϕ_A and ϕ_B, must remain antisymmetric because of the Pauli Exclusion Rule. One way to do this is to construct orthonormal molecular orbitals from linear combinations of the atomic orbitals,

$$\psi_g = (2 + 2S)^{-1/2} (\phi_A + \phi_B)$$

$$\psi_u = (2 - 2S)^{-1/2} (\phi_A - \phi_B) \tag{4}$$

where S is the overlap integral,

$$S = \int \phi_A \phi_B dv \tag{5}$$

needed for normalization. These molecular orbitals fulfill our requirement that the total wavefunction,

$$\Psi = \psi_g(x_1)\psi_u(x_2) - \psi_u(x_1)\psi_g(x_2) = -\Psi \tag{6}$$

be antisymmetric.

After enforcing the Pauli Exclusion Rule, we can determine the total density that this antisymmetric wavefunction defines for the interacting helium dimer. It is worth noting that this is slightly different than the exact density because it lacks polarization effects. However, since

the effect of polarization is small for the helium dimer. The density is simply the square of the wave function,

$$\rho = \Psi^*\Psi = \psi_g^2 + \psi_u^2 \tag{7a}$$

$$= \frac{Z}{(1-S^2)}(\phi_A^2 + \phi_B^2 + 2S\phi_A\phi_B) \tag{7b}$$

where Z is the nuclear charge. When S, the overlap integral, is small we can approximate the prefactor in a binomial expansion,

$$\frac{1}{1-S^2} = 1 + S^2 + S^4 + \cdots \tag{8}$$

This gives us a good approximation to the total density,

$$\tilde{\rho} = \mathrm{Z}(\phi_A^2 + \phi_B^2) - 2ZS\phi_A\phi_B + ZS^2(\phi_A^2 + \phi_B^2) + \cdots \tag{9}$$

where terms of order S^3 or higher are dropped out. It is worth noting that this is slightly different than the exact density because it lacks polarization effects. However, since the effects of polarization are small for the helium dimer (as shown by Salem) this approximate density is accurate enough to ground a qualitative description of Pauli repulsion.

Equation 9 gives us an approximation of the true electron density of the helium dimer. Let us compare this result to the density that would have resulted if we had not imposed the Pauli Exclusion Principle,

$$\rho_0 = \mathrm{Z}(\phi_A^2 + \phi_B^2) \tag{10}$$

It is clear this reference density differs from the true density of equation 9. It is also clear from the preceding derivation that the difference between the two is entirely due to the imposition of the Pauli Exclusion Principle. We can calculate this difference by subtracting the true density from the reference,

$$\Delta\rho = \rho_0 - \tilde{\rho} = 2ZS\phi_A\phi_B - ZS^2(\phi_A^2 + \phi_B^2) \tag{11}$$

This gives us the change in density caused by the Pauli Exclusion Principle. Figure 5.1 shows how we can understand this change qualitatively.

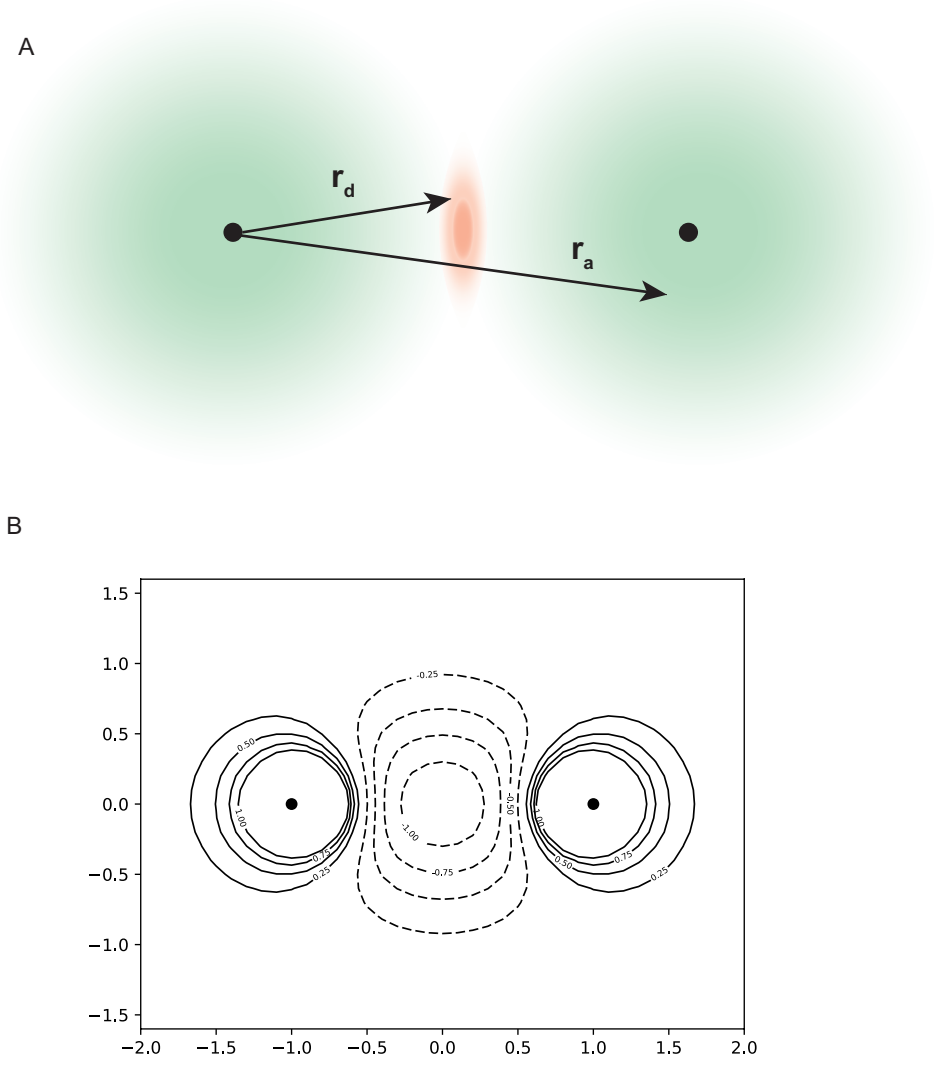


Figure 5.1. Change in electron density for interacting helium dimer.

(A) Representation of the density difference between the interacting and superimposed non-interacting densities. The green and red regions denote areas of electron accumulation and depletion, respectively upon enforcement of wavefunction antisymmetry. The distances r_a and r_d represent characteristic distances to each region from nucleus A. (B) Density difference from *ab initio* calculations. The difference between CCSD densities of the interacting and

non-interacting dimers at an internuclear distance of 1.8 angstroms is shown as a contour plot. Values shown are in thousandths of electrons/bohr^3.

There are two contributions to the change in density from the two terms in equation 11. The first term represents a depletion of electron density in the overlap region between nuclei, indicated by the red shaded region of figure 5.1A. And the second term represents an accumulation of electron density around the centers A and B, indicated by the green shaded regions of figure 5.1A. The validity of this spatial decomposition is shown in figure 5.1B. CCSD density difference calculations on the helium dimer show a clear pattern of depletion in the internuclear space accompanied by an accumulation near the nuclei. It is useful to note that since,

$$\int \Delta\rho dv = 0 \tag{12}$$

these two changes are exactly equal in magnitude, but opposite in sign. Of particular importance is the magnitude of the depleted charge in the overlap region. For the helium dimer shown in figure 5.1B with an internuclear separation of 1.8 angstroms (0.64 times the consensus He vdW diameter of 2.8 Å), the total depleted charge is only 0.01 e^-. This depletion, however, is the dominant contributor to the total SAPT repulsion energy of ~5 kcal/mol. In fact, if one simply computes Coulomb's law between the nuclei and the depleted "positive" charge located at the midpoint between the two nuclei, the result is ~7 kcal/mol, a slight overestimate of the repulsion energy.

In addition to the qualitative description of how the density changes upon imposition of the Pauli Repulsion Principle, we can quantitatively assess how this change affects the energy of the system. The change in energy for nucleus A is:

$$\Delta E_A = E_A(\rho_0) - E_A(\tilde{\rho}) = Z\int \frac{\Delta\rho}{r} dv = Z\left(\int \frac{\Delta\rho_d}{r} dv - \int \frac{\Delta\rho_a}{r} dv\right) \tag{13}$$

where $\Delta\rho_d$ and $\Delta\rho_a$ are the changes in density due to depletion and accumulation, respectively. Plugging in the two terms from equation 11 we find the change in energy,

$$\Delta E_A = Z\left(2ZS\int\frac{\phi_A\phi_B}{r}dv - ZS^2\int\frac{\phi_A}{r}dv - ZS^2\int\frac{\phi_B}{r}dv\right) \quad (14a)$$

$$= Z\left(2ZS\int\frac{\phi_A\phi_B}{r_d}dv - ZS^2\int\frac{\phi_B}{r_a}dv\right) \quad (14b)$$

where the middle term of 14a is identically zero by symmetry. Equation 14b introduces the notation for r illustrated in figure 5.1. The first integral in equation 14b, for small overlaps, is approximately zero everywhere except in that small overlap region. Similarly, the second integral of equation 14b is approximately zero for all of space except the local density of atom B. One can see from figure 5.1 that generally $r_d < r_a$, meaning that the depletion term of equation 14b outweighs the accumulation term, leaving

$$\Delta E_A \approx 2Z^2S\int\frac{\phi_A\phi_B}{r_d}dv \quad (15)$$

as a good approximation of the change in energy due to the Pauli Exclusion Principle. In his original paper, Salem confirmed the validity of this approximation for the helium dimer in the region of small overlap, showing that the energy of depletion term is over 10 times larger than the accumulation term at the van der Waals minimum. Since we are concerned only with small overlaps, a further reasonable approximation is to assume r_d to be constant,

$$r_d = \frac{R}{2} \quad (16)$$

where R is the internuclear distance. Plugging this into equation 15,

$$\Delta E_A \approx \frac{Z^2S}{R}\int\phi_A\phi_B dv \quad (17)$$

simplifies the integral to give our final result:

$$U_{Pauli} = \frac{Z^2}{R} S^2 \tag{18}$$

This is the simple energy difference caused by the imposition of the Pauli Repulsion Principle on our unperturbed reference state.

Equation 18 constitutes the "Orbital Overlap" model of Pauli repulsion and it is remarkable for three important reasons. First, it gives us a clear definition of Pauli repulsion. There is no mysterious "quantum force" that drives molecules apart – equation 18 reveals that the repulsion caused by enforcing the Pauli exclusion principle is electrostatic. The form of equation 18 bears a striking resemblance to Coulomb's law, $U = q_i q_j / R$, only with a factor of S^2 modulating the interaction. This similarity is not an accident. Figure 5.1 shows quite clearly that the main effect of requiring the wavefunction to be antisymmetric is a net loss of electron density, relative to the reference state, in the area between the two nuclei. This leads to an electrostatic repulsion between nuclei that is proportional to the overlap squared. Second, the form of equation 18 fits the asymptotic behavior we expect from a Pauli repulsion function. It is positive everywhere, making it indeed repulsive. Moreover, since S is proportional to an exponential, the repulsion energy goes to zero at long range and becomes large when molecules strongly overlap. Third, the orbital overlap model lends itself to molecular mechanics models because of how it was derived. The above derivation relies on a choice of reference state, in this case the unperturbed electron densities of the separated helium atoms. This is exactly analogous to the strategy of most molecular mechanics models. Partial charges, multipoles, polarizabilities, *etc*. are all assigned to molecules in force fields as gas-phase monomer properties. In other words, the unperturbed molecular electron density is also the "reference state" of molecular mechanics, making the orbital overlap model inherently consistent with the rest of the force field model.

This orbital overlap model of Pauli repulsion is by no means new, but it has not been widely taken up by molecular mechanics models. The reason for this is the challenge of determining the S^2 term for a classical model. Models like SIBFA and EFP have taken the strategy of explicitly calculating molecular orbital overlaps to directly obtain S^2. These models are certainly accurate and have the feature of giving realistic anisotropic repulsion, but they can be too slow for large-scale molecular dynamics simulation and pose problems for parameterization of biological macromolecules. An alternative is to use an empirical model of orbital overlap, but this leaves open the question of how to determine parameters defining the anisotropy. We present here a novel approach to this issue – an anisotropic model of Pauli repulsion can be faithfully constructed from the anisotropy encoded in an atom's multipole moments.

5.2.1 Multipole Overlap Pauli Repulsion

A model for orbital overlap requires some method for describing the electron distribution around a molecule. Previous work has shown that a simple, hydrogen-like model of charge density can be used to accurately predict the electrostatic interactions of dimers at short-range.[15] In this model the point Coulomb potential, $V = q / r$ in atomic units, was replaced by,

$$V(R) = \frac{q}{r}(1 - e^{-\alpha r}) \tag{19}$$

where α is a parameter introduced to describe the width of the electron density. For the present model, we modify this slightly, as suggested by Slipchenko and Gordon.[59] The potential generated by the electron of a hydrogen-like atom is:

$$V(r) = \frac{q}{r}[1 - (1 + Zr)e^{-2Zr}] \tag{20}$$

where Z is the nuclear charge. The form of equation 20 differs slightly from equation 19, suggesting a better approximation for the model,

$$V(r) = \frac{q}{r}\left[1 - \left(1 + \frac{1}{2}\alpha r\right)e^{-\alpha r}\right] \tag{20}$$

that more accurately captures the asymptotic behavior of the potential. From this potential we wish to build a model electron distribution. To do this we can apply Poisson's equation,

$$\nabla^2 V = -4\pi\rho, \tag{21}$$

to obtain the charge density that generates the potential of equation 19:

$$\rho(r) = \frac{q\alpha^3}{8\pi}e^{-\alpha r} \tag{22}$$

The density, however, does not directly give the information we need. In order to use the orbital overlap model of Pauli repulsion, we must have a model for orbitals. To get from a density model to an orbital model we apply,

$$\rho = \phi^*\phi. \tag{23}$$

If we impose the restriction that the orbitals be real, then we are left with model pseudo-orbitals:

$$\phi = \sqrt{\rho} = \sqrt{\frac{q\alpha^3}{8\pi}}e^{\frac{-\alpha r}{2}}. \tag{24}$$

It is important to be clear about the purpose of these model orbitals. What we are interested in for Pauli repulsion is the regime of small overlap. Correspondingly, these orbitals are simply meant to approximate the form of the outermost extent of an atom's electron distribution.

Within this pseudo-orbital model we can evaluate the overlap integral between interacting orbitals, A and B,

$$S = \int \phi_A\phi_B dv = \frac{\sqrt{q_A q_B \alpha_A^3 \alpha_B^3}}{8\pi}\int e^{\frac{-\alpha_A r_A - \alpha_B r_B}{2}} dv\,, \tag{25}$$

Performing the integration according to the method of Coulson[60] and refactoring yields the result,

$$S = \sqrt{\frac{q_A q_B \alpha_A^3 \alpha_B^3}{R}} f_{damp}(R), \tag{26}$$

where,

$$f_{damp}(R) = \begin{cases} \frac{\sqrt{R}}{\alpha^3}\left(1 + \frac{\alpha R}{2} + \frac{1}{3}\left(\frac{\alpha R}{2}\right)^2\right) e^{\frac{-\alpha R}{2}}, & \alpha_A = \alpha_B \\ \frac{1}{2X^3\sqrt{R}}\left[\alpha_A(RX - 2\alpha_B)e^{\frac{-\alpha_B R}{2}} + \alpha_B(RX + 2\alpha_A)e^{\frac{-\alpha_A R}{2}}\right], & \alpha_A \neq \alpha_B \end{cases}. \tag{27}$$

$$X = \left(\frac{\alpha_A}{2}\right)^2 - \left(\frac{\alpha_B}{2}\right)^2$$

There are two important features to note about S. First, it is asymptotically exponential, matching our general intuition about orbital overlap. Second, it depends on the respective atomic multipoles of A and B. We can elucidate this fact by writing S^2 in "Coulombic" form,

$$S^2 = q_A T_{pauli} q_B, \tag{28}$$

with,

$$T_{pauli} = \frac{\alpha_A^3 \alpha_B^3}{R} f_{damp}^2. \tag{29}$$

Equation 28 represents the charge-charge overlap term of our Pauli repulsion model. If we wish our model to be isotropic, we simply stop here and compute the repulsion energy according to equation 18. However, for multipolar force fields we are not bound to simply using the charge-charge component of the overlap. Following the example of Slipchenko and Gordon,[59] we can show how to compute the charge-dipole and higher-order multipole terms for orbital overlap as well.

Consider the overlap at distance, R, of a charge density, Q with a finite dipole, μ, where the dipole is represented by two equal and opposite charges, q^- and q^+, separated by a distance, d. For this interaction,

$$S^2_{charge-dipole} = S^2_{Q-q^-} + S^2_{Q-q^+} = QT_{pauli}\left(R - \frac{d}{2}\right)q^- + QT_{pauli}\left(R + \frac{d}{2}\right)q^+. \tag{30}$$

If we define the dipole moment,

$$\mu = qd, \tag{31}$$

then equation 30 becomes:

$$S^2_{charge-dipole} = Q\mu\frac{T_{pauli}\left(R + \frac{d}{2}\right) - T_{pauli}\left(R - \frac{d}{2}\right)}{d}. \tag{32}$$

If we take the limit as $d \rightarrow 0$, then

$$S^2_{charge-dipole} = Q\frac{\partial T_{pauli}}{\partial R}\mu. \tag{33}$$

Note that this is exactly analogous to the derivation of the electrostatic multipole interaction,

$$U_{electrostatic} = q_A T q_B + q_A \nabla T \mu_B - \mu_A \nabla T q_B + \mu_A \nabla\nabla T \mu_B + \cdots$$
$$\text{with } T = \frac{1}{R} \tag{34}$$

where the only difference is the kernel, T.

In the same way we can compute S^2 for arbitrarily order multipole orbital overlaps. In the current model we shall take this through quadrupole-quadrupole repulsion, yielding:

$$S^2_{total} = q_A T_{pauli} q_B + q_A \nabla T_{pauli} \mu_B - \mu_A \nabla T_{pauli} q_B + \mu_A \nabla^2 T_{pauli} \mu_B + q_A \nabla^3 T_{pauli} \Theta_B - \Theta_A \nabla^3 T_{pauli} q_B + \mu_A \nabla^4 T_{pauli} \Theta_B - \Theta_A \nabla^4 T_{pauli} \mu_B + \Theta_A \nabla^5 T_{pauli} \Theta_B. \tag{35}$$

This is the source of anisotropy in our model. Rather than introduce any new parameters, we simply use the shape of the atom encoded in the multipole moments to tell us about the anisotropy of repulsion. This result provides a total multipole overlap model of Pauli repulsion:

$$U_{pauli} = \frac{K_A K_B}{R} S_{total}^2. \quad (36)$$

The parameter, K, is introduced to set the relative sizes of different atom classes. This model will be referred to throughout the remainder of the paper as the "Multipolar Pauli Repulsion" model. A variant that uses only the first, charge-charge term of equation 35 will also be discussed. We refer to this as the "Isotropic Pauli Repulsion" model.

We intend to use this model on a broad array of complicated biomolecular intermolecular interactions. To determine the parameters, particularly K, that accurately describe these interactions, we have chosen to use *ab initio* SAPT Exchange Repulsion calculations to generate reference data. It should be noted that this is technically an approximation. While our derivation relies on a Hellman-Feynman theorem analysis of density differences, the SAPT Exchange Repulsion term has no associated density. This approximation, however, is necessary, accurate, and consistent with our model. The SAPT Exchange Repulsion energy is a direct approximation of the energy increase required to antisymmetrize the wavefunctions of two monomer reference states – exactly the quantity that our model is built to reproduce.

5.2.2 A Note on "Density Overlap" Models

The other major class of overlap models for Pauli repulsion utilizes density overlap. In these models the Pauli repulsion energy is modeled as,

$$U_{pauli} = K_{ij}\Omega, \quad (37)$$

where,

$$\Omega = \int \rho_i \rho_j dv\,. \quad (38)$$

This model is supported with some experimental evidence,[45,61] but it has serious flaws as an interpretable model for Pauli repulsion energy. This is can be simply illustrated with

straightforward dimensional analysis. Ω is the one-center integral of two densities which means that it has dimension (in atomic units),

$$\Omega = \frac{e}{a_0^3}\frac{e}{a_0^3}a_0^3 = \frac{e^2}{a_0^3}. \tag{39}$$

These are not the units of a Coulombic energy (e^2 / a_0 in atomic units). Although K_{ij} can be given units that yield an energy, this does not make the model Coulombic since the term is a constant that does not depend on distance. This is a problem for two reasons. First, it makes the model inconsistent with the Hellman-Feynman (Electrostatic) theorem. According to the Hellman-Feynman theorem, every intermolecular force is the result of applying Coulomb's law to a change in electron density. This dimensional analysis demonstrates that there exists no Hellman-Feynman-based justification for the density overlap model of Pauli repulsion since any such rationale must necessarily be Coulombic. It should be noted that there are other Pauli repulsion energy models that do not have Coulombic interpretations as well, the simple Lennard-Jones model among them. However, these models are, by and large, unitless which means that they do not suffer from the second and more important reason that the density overlap model is problematic – the density overlap model has the wrong distance dependence. Since the density overlap is proportional to charge squared over distance cubed, applying the electric constant no longer gives units of energy, but energy over distance squared. This is problematic because this formulation now explicitly depends on the unit chosen for distance. A consequence of this is that the radial dependence of the Pauli repulsion energy will be qualitatively incorrect. This is illustrated in figure 5.2 for the case of helium dimer repulsion. Both the S^2/R and density overlap models are governed largely by exponentials, as shown by the nearly straight lines on the semi-

log plot. However, the slope of the density overlap exponential function clearly differs from the SAPT Exchange Repulsion.

A simple model system explains this difference in radial dependence of the S^2/R and density overlap curves. If we assume that both atoms have an isolated electron density described by equation 22 (~ $e^{-\alpha r}$), the resulting Pauli repulsion energies of the density overlap and S^2/R models respectively will be:

$$U_{density\ overlap} = K_{ij} \int \rho_i \rho_j dv = K_{ij} \frac{q_i q_j}{(8\pi)^2} \frac{\pi}{\alpha^3} \left(1 + \alpha R + \frac{1}{3}(\alpha R)^2\right) e^{-\alpha R} \qquad (40)$$

and,

$$U_{S^2/R} = K_{ij} \frac{1}{R} \left(\int \phi_i \phi_j dv \right)^2 = K_{ij} \frac{1}{R} \frac{q_i q_j}{(8\pi)^2} \frac{(8\pi)^2}{\alpha^6} \left(1 + \frac{1}{2}\alpha R + \frac{1}{12}(\alpha R)^2\right)^2 e^{-\alpha R}. \qquad (41)$$

For this simple model the arguments of the exponentials of both models are identical, due to the orbital overlap, S, being squared. This explains the similarities in the curves in figure 5.2. However, the R-dependent, polynomial prefactor of equation 40 clearly differs from equation 41. It is this difference that causes the radial divergence illustrated in figure 5.2. We will further illustrate this phenomenon in Section 4 with additional noble gas and water dimer calculations.

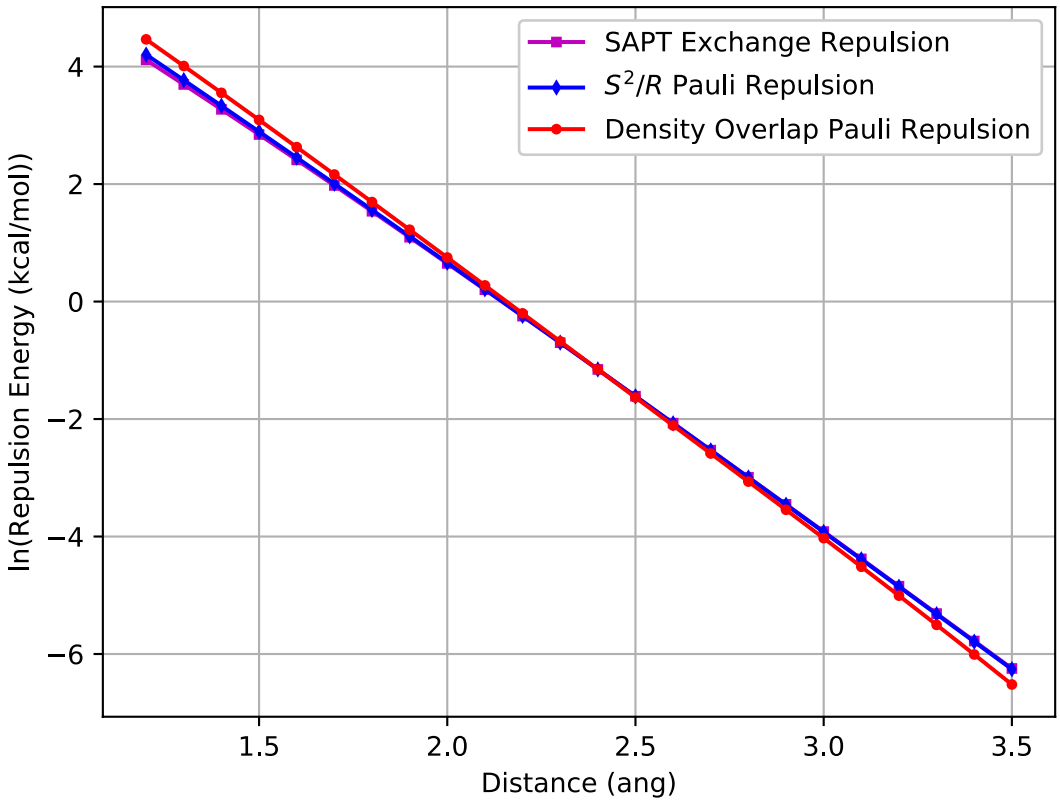


Figure 5.2. Radial dependence of the S^2/R (blue) and density overlap (red) models for the helium dimer. The repulsion energy of both models computed from monomer wavefunctions determined with the aug-cc-pVQZ basis set is compared to the SAPT2+ repulsion energy (magenta) and plotted on a semi-log scale. The proportionality constant of each model was fixed to reproduce the SAPT repulsion energy at 2.4 ang. The SAPT Exchange Repulsion curve is almost entirely obscured by the S^2/R curve. While the slope of the S^2/R curve matches that of the SAPT repulsion energy, the slope of the density overlap curve does not.

Despite this incorrect radial functional dependence, the density overlap model can still provide a serviceable empirical model. There are two ways that this has been done in prior efforts. The first is to choose the K_{ij} proportionality for a representative interaction distance. Since, like the orbital overlap model, the density overlap model is dominated by an exponential

term, if the K constant is chosen for a suitable interaction distance, the radial error may not become large within the range sampled in application. The other method is to include a distance dependent prefactor (1/R, $1/R^2$, *etc.*) in the function. This has been proposed by Neyland and Toennies, Andreev, and Soderhjelm and co-workers. [62-64] These models explicitly address the units of the function, albeit empirically, and have been shown to be valid over a wider range of interaction distances than the pure density overlap model.

As opposed to unitless models like Lennard-Jones or Buckingham functions which sidestep the question, the orbital overlap model explicitly satisfies the dimensional analysis test. If we take equation 35 and for simplicity, only consider the charge-charge term of S^2 (the same holds for the higher-order terms) this gives,

$$U_{pauli} = \frac{e^2}{a_0}, \tag{42}$$

the atomic units of a Coulomb energy. This is fully consistent with the Hellman-Feynman theorem interpretation of the Pauli repulsion energy. It is also what we should expect, given that the derivation of the orbital overlap model is built on a set of electrostatic interaction arguments. This makes the Multipole Overlap Pauli Repulsion model interpretable and, as we shall show in Section 4, this interpretability bestows on the model some measure of transferability.

5.3 Methods

5.3.1 Parameterization

To fit the Multipole Pauli Repulsion model, we utilized the previously published S101x7 database.[65] This database is meant to capture intermolecular interactions that are important for biomolecular applications and has been used to parameterize electrostatics and dispersion models.[15,66] The database contains 101 unique sets of molecular dimers with seven different points along the dissociation curve at 0.7, 0.8, 0.9, 0.95, 1.0, 1.05 and 1.1 times the equilibrium

distance. See reference 65 for a complete description of the generation of database geometries. We augmented the database with a set of methane and formaldehyde homodimers with geometries generated in the same manner. We used the publicly available SAPT2+ energy decomposition analysis calculations we previously published to extract the Exchange Repulsion (Pauli Repulsion) energy of each dimer pair in the database. The SAPT2+ Exchange Repulsion energies were computed using the so-called S^2 approximation which has been previously shown to be accurate for biomolecular fragment interactions.[67]

To fit the parameters of our model we defined 26 unique atom classes. These classes are assigned according to the qualitative chemical environment of each atom and are listed in table 5.1. They are adopted from the atom classes used in a previously published study of van der Waals energies of the S101x7 database.[68]

N	Pauli Repulsion Class	K	α	q
1	H (nonpolar)	2.25	4.63	1.00
2	H (nonpolar, alkane)	1.82	4.23	1.00
3	H (polar, N−H/N aromatic)	1.11	4.21	1.00
4	H (polar, O−H)	1.18	4.20	1.00
5	H (aromatic, C−H)	1.24	4.43	1.00
6	H (polar, S−H)	1.20	4.01	1.00
7	C (sp^3)	2.62	4.64	3.41
8	C (sp^2, alkene)	1.42	3.56	3.92
9	C (sp^2, C=O)	1.31	3.50	2.02
10	C (aromatic, C−C)	1.37	3.77	4.00
11	C (aromatic, C−X)	1.37	3.70	3.70

12	N (sp^3)	3.61	4.16	2.00
13	N (sp^2)	4.62	4.27	2.15
14	N (aromatic)	3.93	4.40	2.48
15	O (sp^3, hydroxyl, water)	3.57	4.74	3.00
16	O (sp^2, carbonyl)	1.43	4.14	6.00
17	O (O^- in AcO^-)	1.19	3.77	5.87
18	O (O^- in HPO_4^{-2})	1.25	3.73	5.82
19	O (O^- in $H_2PO_4^-$)	1.47	4.02	5.75
20	O (O in H_3PO_4)	1.63	4.15	5.78
21	P (phosphate)	1.74	4.40	4.98
22	S (sulfide, R−SH)	3.40	3.62	3.39
23	S (sulfur IV, DMSO)	1.52	3.33	6.00
24	F (organofluorine)	1.38	4.72	5.05
25	Cl (organochloride)	1.91	3.76	5.91
26	Br (organobromine)	2.02	3.52	6.63

Table 5.1. Atom classes and parameter values for the anisotropic Multipole Pauli Repulsion model.

Classes are taken from those in reference 68.

For each atom class, the model as defined by equation 35 requires three parameters: the size of the atom, K, the shape of the atom, α (as defined in equation 24), and the number of valence electrons, q (as defined in equation 35). The purpose of K and α is straightforward; together they set the strength and shape of the exponential repulsion between atoms. The purpose of q as a

parameter is slightly more nuanced. In point force fields (charge-only or multipolar) the atomic charges are a combination of the nuclear charge with the net electronic charge on that atom. This definition will not work for our model because only the electrons are involved in overlap. Furthermore, since we are only interested in the region of small overlaps, it does not make sense to use all of the electrons on each atom, as only the outermost part of the electron density is involved in overlap. Thus, the parameter, q, is best thought of as the maximum (not necessarily an integer) number of electrons that are involved in overlap for a particular atom. This turns out to be an important parameter because it sets how anisotropic the Pauli Repulsion of a specific atom will be. A large q will make the first, isotropic term of equation 35 large, while a small q will make the first term small relative to the higher-order anisotropic terms. To ward off overfitting, q for all hydrogens is set to be the negative of the total number of electrons (negative one plus the partial charge). Additionally, q for heavy atoms is constrained to lie between 2 and the number of valence electrons of the element. For the Isotropic Pauli Repulsion model (only using the first term of equation 35), q is set to be the number of valence electrons, since in the isotropic model q and K are redundant.

It is important to emphasize what is not being fit in this model. The dipole and quadrupole moments of each atom are taken directly from the Distributed Multipole Analysis (DMA) based procedure detailed in the appendix of reference 69. This does two important things. First, it makes the Multipolar Pauli Repulsion model consistent with the AMOEBA model and future AMOEBA-like models. In particular this means that the model can be used along with previously published AMOEBA-like electrostatics[15] and dispersion[66] models. Second, this insulates the Pauli Repulsion Model from the most common problem of anisotropic repulsion models: overfitting. The dipole moment of each atom has three independent

components and the traceless quadrupole has five. Empirically fitting these parameters would result in a massive overfitting problem. Past anisotropic models have either fit very specific models (*e.g.,* water) to large datasets[32,33] or used explicit atomic orbitals.[40,43] This model sidesteps the troubles associated with both of those approaches by using the multipoles that come from directly fitting the electrostatic potential around a molecule.

Because the Pauli Repulsion energy exhibits strong exponential character, we choose to perform a natural log fit to obtain parameters for the Multipolar and Isotropic Pauli Repulsion models. To do this we minimized the residual, log (SAPT Exchange) – log (model), for each dimer data point in the S101x7 dataset using a Levenberg-Marquardt least squares routine in the Tinker Molecular Mechanics package.[70] This prevents the closest, but rarely accessible, dimer points from biasing the fit. A third model, termed vdW2017 was also fit. This model uses the AMOEBA standard Buffered 14-7 functional form,

$$U_{vdW} = U_{replusion} + U_{dispersion} = \varepsilon_{ij}\left(\frac{1+\delta}{\rho_{ij}+\delta}\right)^{7}\left(\frac{1+\gamma}{\rho_{ij}^{7}+\gamma}-2\right), \qquad (43)$$

where δ and γ are global shape parameters, and ε and ρ are set by the Waldman-Hagler and arithmetic combining rules respectively as suggested in reference 68. All four parameters were allowed to vary in the fit with ε and ρ for each atom being set by the same atom classes presented in table 5.1. Additionally, a fixed "hydrogen-reduction factor" of 0.9 was applied to all hydrogens as described in reference 31. Equation 43 can be split into a positive and negative part which represent the contributions to repulsion and dispersion respectively. However, since the Buffered 14-7 parameters for the two terms are not independent, a (non-natural log weighted) least squares fit was carried out that simultaneously minimized (SAPT Exchange – vdW2017 Repulsion) and (SAPT Dispersion – vdW2017 Dispersion) for each dimer data point. In both fits,

S101 dimers including the triple-bonded ethyne molecule were excluded as was done previously, eliminating the S101 class for *sp*-hybridized carbon from the fitted parameters.

5.3.2 Computational Details

The Multipolar and Isotropic Pauli Repulsion models have been implemented in publicly available versions of the Tinker Molecular Mechanics package. [70-72] It is worth noting for future force field development that the additional overhead to compute the Multipolar Pauli Repulsion model on top of an existing Multipole Electrostatic calculation is small. As the similarity between equations 34 and 35 suggests, the intermediate quantities necessary for energy and forces are largely identical between the two models.

Lastly, we explored calculating the orbital overlap and density overlap directly from quantum mechanical calculations. Since the components of interacting dimers each have multiple occupied orbitals rather than single model pseudo-orbitals, we must define S^2 for this situation. Because we are working with orthogonal molecular orbitals in the LCAO (Linear Combination of Atomic Orbitals) convention, we use the sum-of-squares definition,

$$S^2 = \sum_i^A \sum_j^B \langle \psi_i | \psi_j \rangle^2 , \tag{44}$$

where the A and B represent the two monomers with sums over i and j, the occupied orbitals on A and B respectively.[62] We express the occupied orbitals in terms of atomic basis functions so that equation 44 is invariant under orthogonal transformations of the molecular orbitals. The density overlap is calculated on a grid according to equation 37. The QM orbital and density overlap calculations using SCF monomer orbitals with an aug-cc-pVDZ basis were performed using the Psi4NumPy program.[73]

The computational cost of the Multipolar Pauli Repulsion model was evaluated on a typical computer workstation. The time to complete 100 energy and force evaluations for different combinations of models was performed on a four core 3.4 GHz Intel Core i7 processor. The test was run on a 25 x 25 x 25 angstrom water box with 500 water molecules. The cutoff distance for Pauli repulsion is set to 5 angstroms and dispersion is handled via particle mesh Ewald summation.[66] For comparison, models including the Halgren buffered 14-7 potential were also included. For these calculations a van der Waals cutoff distance of 10 angstroms is used. To evaluate the cost in the context of a generalized AMOEBA-like model, timings are also presented that include polarization with the induced dipoles convergence criteria set to 10^{-5} Debye RMS.

5.4 Results

5.4.1 Noble Gas Dimers

To assess the validity of the orbital overlap model for Pauli repulsion we first considered the case of Pauli repulsion between noble gas dimers. Because they are neutral and have spherical symmetry, noble gas dimers are a natural first testing ground for a Pauli repulsion model. Specifically, we set out to test the underlying assumption that the Pauli repulsion energy should be proportional to S^2/R. The results, plotted for a range of distances of the neon and argon dimers, are plotted in figure 5.3.

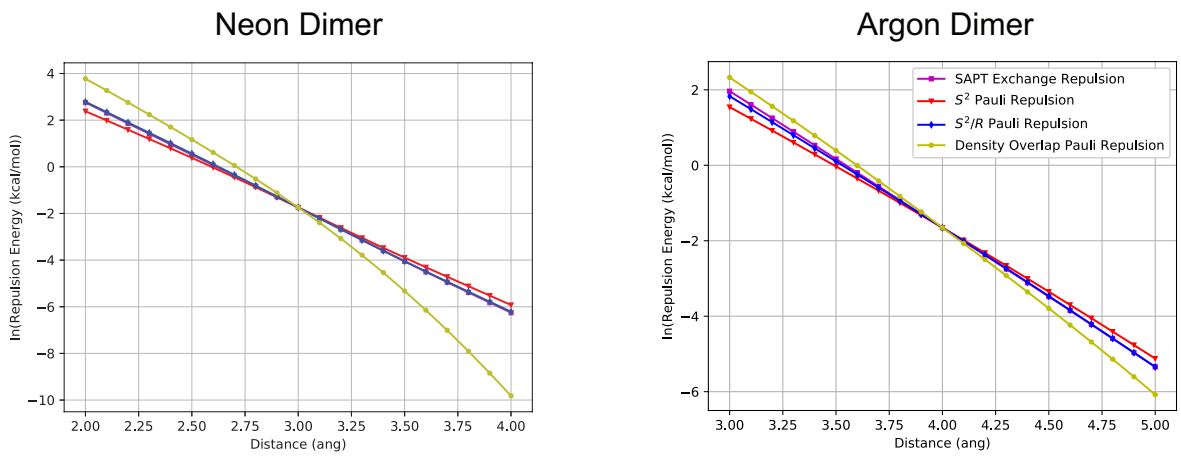


Figure 5.3. Comparison of QM monomer-based methods for estimating Pauli repulsion in noble gas dimers. The natural log of the repulsion energy is plotted against dimer separation distance to illustrate the exponential relationship. For a range of distances for Ne–Ne and Ar–Ar dimers, the S^2, S^2/R, and density overlap methods where tested against SAPT Exchange repulsion energy. The proportionality constant for each method was arbitrarily fixed to reproduce the middle value of the distance range (3.0 ang for Ne–Ne and 4.0 ang for Ar–Ar). In both cases the S^2/R method is virtually indistinguishable from the SAPT result.

There are several features worth noting in figure 5.3. The first is that for these simple systems the SAPT Exchange repulsion energy is clearly exponential. The plot reveals a near-linear relationship between the internuclear distance and the natural log of the SAPT Exchange repulsion energy. The second noteworthy feature is the quality of the S^2/R model energy calculated from SCF monomer orbitals. This is not necessarily surprising, given that the model was derived from the *ab initio* Pauli repulsion of the helium dimer, but agreement is remarkable, given that this estimate is obtained without the need for dimer calculations. The third important item of note is the poor quality of the density overlap approximation for the Pauli repulsion energy. The distance-dependence problem addressed in the preceding section is abundantly clear in this simple example. The density overlap is evidently not proportional to the Pauli repulsion energy for a meaningful range of distances for noble gas dimers. Although the rapid divergence

of the density overlap model at long range for the neon dimer is likely due use of the smaller aug-cc-pVDZ basis set, the same qualitatively different radial dependence is observed when the overlap is computed with a much larger (aug-cc-pV5Z) basis set.

The final feature to point out regarding the noble gas dimers is subtle, but important. Also plotted in figure 5.3 are the results for assuming that S^2, as opposed to S^2/R, is proportional to the Pauli repulsion energy. The results show that although the agreement might be close over a small range of distances, the overall slope is slightly too small. This shows the importance of the 1/R factor in the Pauli repulsion expression. There has been some discussion in the literature about what, if any, function of R should precede S^2 for Pauli repulsion.[41,62] These results clearly indicate that 1/R is the correct choice. As an empirical matter, of course, for more complicated molecules other choices can be made in the context of a total energy model. However, these results show the S^2/R model to be the most natural fit to the fundamental Pauli repulsion phenomenon.

5.4.2 S101x7 Dataset

Having established the validity of the S^2/R model for noble gas systems, we set about determining whether the model is appropriate for a more complicated dataset. The S101x7 database was chosen to represent a range of biomolecular dimer interactions and three models were fit: Multipolar Pauli Repulsion, Isotropic Pauli Repulsion and a Buffered 14-7 model. As stated in the Methods section, the first two repulsion-only models were fit with natural log weighted least squares, while the Buffered 14-7 model, termed vdW2017, was fit to unweighted SAPT dispersion and exchange repulsion data simultaneously. The results of these fits are given in table 5.2 and illustrated in figure 5.4.

	Total RMSE (kcal/mol)	Short-Range RMSE (0.7) (kcal/mol)	Intermediate RMSE (0.8 - 0.95) (kcal/mol)	Long-Range RMSE (1.0 – 1.1) (kcal/mol)
Multipolar Pauli Repulsion	1.71	4.14	0.99	0.37
Isotropic Pauli Repulsion	2.37	5.68	1.46	0.44
vdW2017 Repulsion	2.66	5.94	2.02	0.83

Table 5.2. Root mean square error on S101x7 dataset.

Shown are the errors relative to SAPT Exchange-repulsion. “Short-Range” indicates data points at 0.7 times the dimer equilibrium distance. “Intermediate” indicates data points 0.8 to 0.95 times the dimer equilibrium distance. “Long-Range” indicates data points at or beyond the dimer equilibrium distance. Note that all values are absolute errors, and not log-weighted.

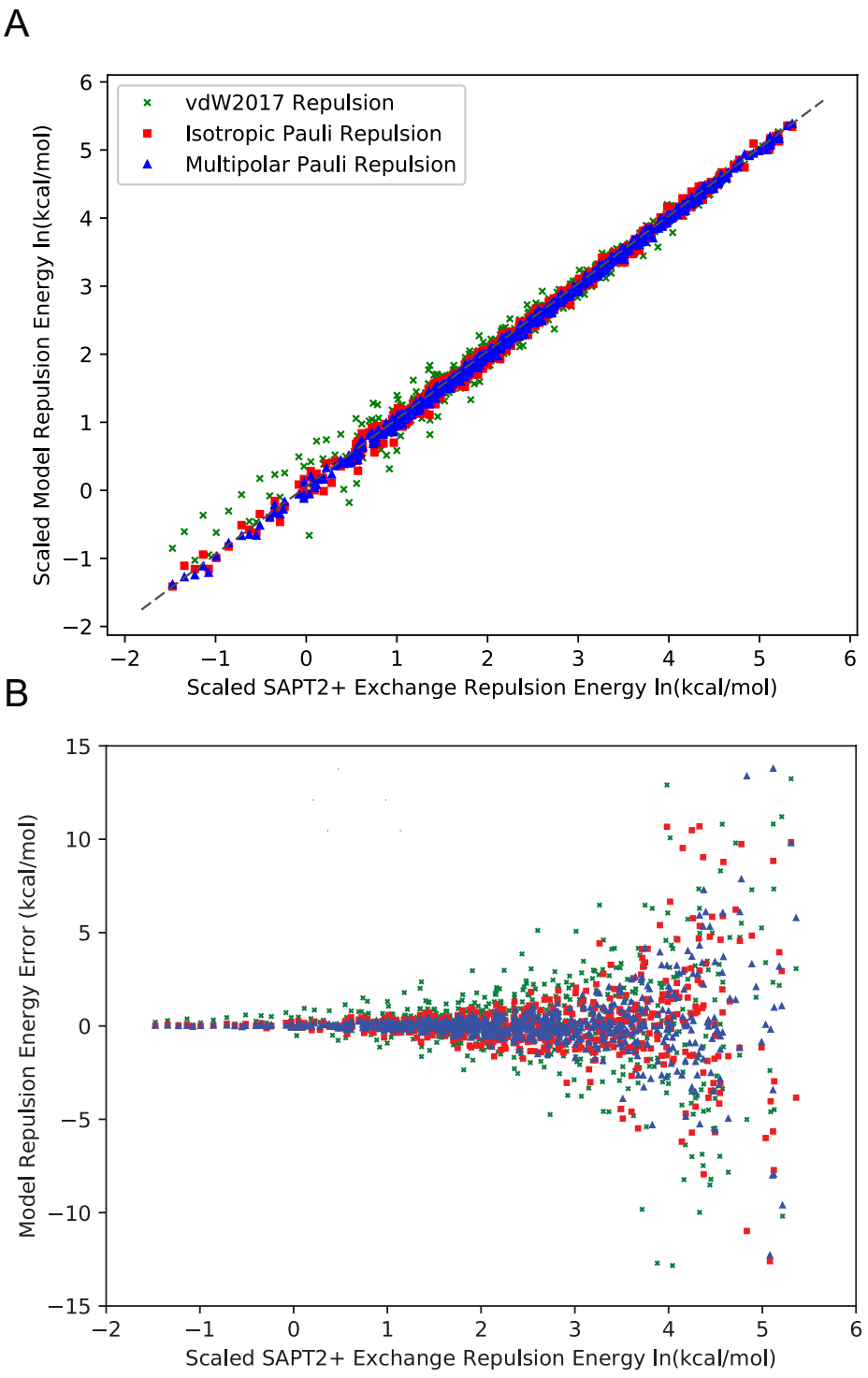


Figure 5.4. S101x7 Pauli repulsion energy.

(A) Model Pauli repulsion energy plotted against SAPT2+ Exchange repulsion energy for all dimers. Both model and SAPT data are plotted on natural log weighted access for clarity. Dashed line indicates perfect agreement. (B) Model error plotted against absolute (log-weighted) SAPT repulsion energy.

The results show the tradeoff in accuracy that is taken for using an isotropic model. The Isotropic Pauli Repulsion and vdW2017 models exhibit similarly large errors for the S101x7 dataset of 2.37 kcal/mol and 2.66 kcal/mol respectively. This error is driven by the closest contact points in the dataset but is still large for intermediate distances as well. Notably, the vdW2017 model has an error of close to 1 kcal/mol for points at equilibrium and beyond. To some extent, errors at this distance for the buffered 14-7 potential are compensating for the dispersion part of the function, but this comes at the detriment of having a separate and interpretable Pauli repulsion model. The Isotropic Pauli Repulsion function root mean square error, on the other hand, decays more rapidly with distance.

The quantitative benefit of using an anisotropic Pauli repulsion function is readily apparent from table 5.2 and figure 5.4. The Multipolar Pauli Repulsion model requires more terms to compute, but it fits the S101x7 nearly twice as well as its isotropic counterpart. The total RMSE is being driven almost entirely by the closest-range points (0.7x) in the dataset. For intermediate and near-equilibrium points, the Multipolar Pauli Repulsion model gives errors of well under 1 kcal/mol. Because the fitting was performed against log-transformed data, it is not surprising to see this behavior. Moreover, this behavior should be considered desirable since the 0.7x points of the dataset largely fall just outside of the realm accessed during molecular dynamics simulation under ambient conditions. Figure 5.4 illustrates the tighter fit to SAPT that is achieved with the anisotropic Multipolar Pauli Repulsion model. The results of these fits will be used to evaluate the models for the remainder of the paper. We will explore the factors

contributing to the superior fit of the Multipolar Pauli Repulsion model in the Discussion and Conclusions section.

As can be seen from figure 5.4A, there are some large errors in the fitted S101 dataset, particularly for the isotropic and vdW2017 models. Nearly all of these errors occur at the short-range 0.7x points of the database where the absolute repulsion energy is very high. This is apparent from the figure 5.4B, which shows the error of each point in the fits plotted against the log magnitude of the repulsion energy. All of the errors greater than 5 kcal/mol occur when the SAPT repulsion energy is greater than 50 kcal/mol. Additionally, many of the dimers where the anisotropic Multipole Pauli Repulsion model produces the greatest reduction in the error fit with intuition. Some of the largest decreases in error are for the DMSO-DMSO dimer, dimers involving phosphate, and pi-pi stacking interactions. These are all interactions with significant electrostatic anisotropy (large dipole and quadrupole moments) and the Multipole Pauli Repulsion model fits these data more precisely.

As stated in the introduction it is important that a repulsion model be interpretable in addition to being accurate. One simple measure of interpretability is the reasonableness of the fitted model parameters. To assess the sensibility of the parameters for the Multipolar Pauli Repulsion model we calculated an atomic "size" for each atom class defined in table 5.1. The metric for size, presented in table 5.3, is the atomic radius corresponding to an atom-atom homodimer internuclear distance at which the repulsive energy reaches 1.0 kcal/mol. Because the Multipolar Pauli Repulsion model is anisotropic, we only include the charge-charge portion of the energy to cleanly separate the size from the orientational dependence.

N	Pauli Repulsion Class	Radius (Å)
1	H (nonpolar)	1.12

2	H (nonpolar, alkane)	1.16
3	H (polar, N−H/N aromatic)	1.01
4	H (polar, O−H)	1.03
5	H (aromatic, C−H)	1.00
6	H (polar, S−H)	1.08
7	C (sp^3)	1.48
8	C (sp^2, alkene)	1.73
9	C (sp^2, C=O)	1.50
10	C (aromatic, C−C)	1.64
11	C (aromatic, C−X)	1.64
12	N (sp^3)	1.58
13	N (sp^2)	1.63
14	N (aromatic)	1.58
15	O (sp^3, hydroxyl, water)	1.50
16	O (sp^2, carbonyl)	1.64
17	O (O^- in AcO^-)	1.71
18	O (O^- in HPO_4^{-2})	1.75
19	O (O^- in $H_2PO_4^-$)	1.68
20	O (O in H_3PO_4)	1.66
21	P (phosphate)	1.55
22	S (sulfide, R−SH)	1.95
23	S (sulfur IV, DMSO)	2.02
24	F (organofluorine)	1.40

25	Cl (organochloride)	1.87
26	Br (organobromine)	2.05

Table 5.3. Atomic "size" for Multipolar Pauli Repulsion atom classes.

The radius is calculated as half the distance at which an atom-atom homodimer experiences 1.0 kcal/mol of repulsion energy. Only the charge-charge component of the repulsion is included. This is equivalent to the repulsion energy of the homodimer averaged over all possible dimer orientations at the standard distance.

Broadly, the sizes in table 5.3 show a chemically intuitive picture of atomic size. The sizes follow periodic trends and the differences across classes of the same element are reasonable. We note that although similar to the "size" (radius) parameter of the Lennard-Jones 12-6 or Halgren buffered 14-7 potentials, the size metric here should not be quantitatively compared. The size parameters in those van der Waals functions implicitly include the dispersion contribution in addition to repulsion.

The S101x7 database provides extensive coverage for biomolecular chemical space and the radial dependence of interactions. The results of the fit show that using an exponential-based function matches this radial dependence better than the buffered 14-7 potential. The errors of the Isotropic Pauli Repulsion and vdW2017 models are similar for the closest (0.7x) points of the dataset. However, at distances just past equilibrium the errors in the Isotropic Pauli Repulsion model become asymptotically smaller compared against the buffered 14-7 potential. This suggests radial scans of S101x7 are effective at determining the exponential parameter, α, and the prefactor, K. The S101x7 database, however, contains relatively less orientational information. This requires us to carefully consider the charge, q, parameters for heavy atoms that are largely responsible for handling the angular dependence of the Multipolar Pauli Repulsion

model. In the following test cases we explore a variety of systems that specifically target the angular degrees of freedom that are less sampled by the S101 dataset.

5.4.3 Water Dimers

Water is an important case in force field development, for the reason that it is the solvent in which interesting biomolecular phenomenon usually occur. In addition to being important for applications, water is also curious because of its anisotropic repulsive properties. To examine the performance of our model on this system, we shall consider three separate series of water dimers: one in which water dimer dissociation is considered, one in which the "flap angle" (defined in the inset of figure 5.6) of the water dimer is systematically varied, and one which consists of 10 well-studied independent stationary points on the water dimer surface.

Because water-water interactions are so important to the end goal of biomolecular simulations, the quality of the fit to the water dimer dissociation data of the S101x7 dataset is instructive. Figure 5.5 shows the performance of a number of repulsion models against reference SAPT2+ Exchange Repulsion energies plotted on a natural log scale.

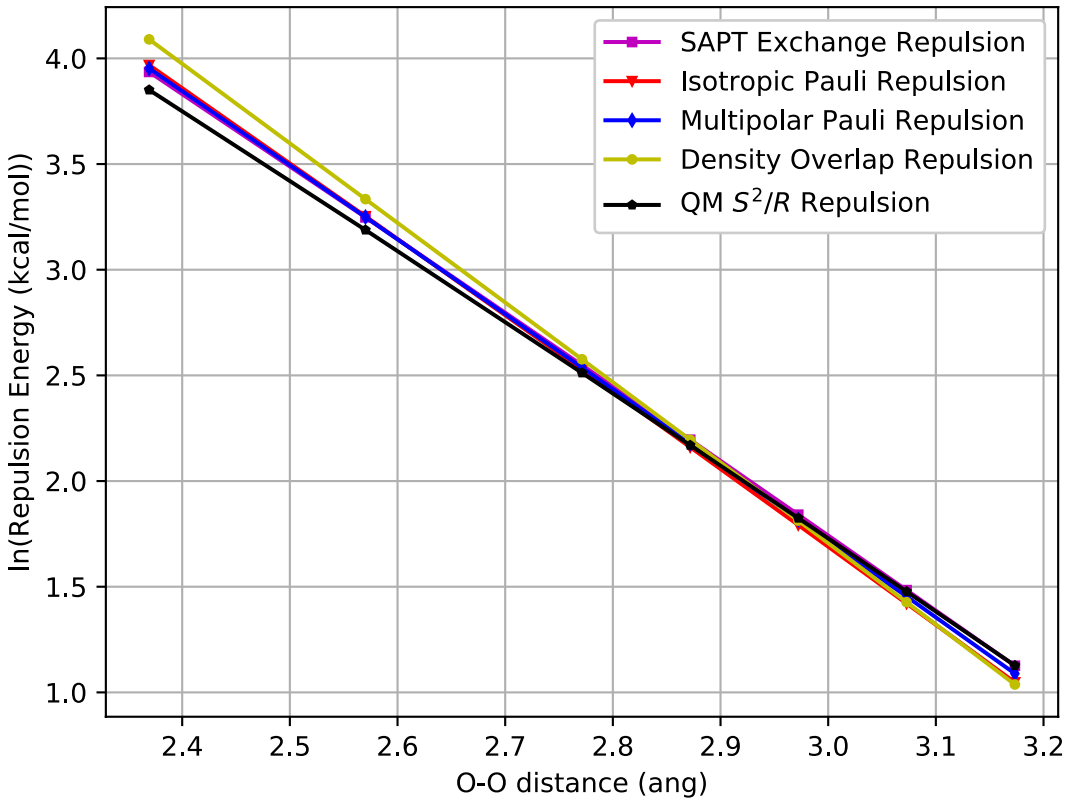


Figure 5.5. Water dimer dissociation Pauli repulsion.

Water monomer geometries were fixed at equilibrium and placed at distances from 0.7 to 1.1 times the equilibrium O–H distance. The SAPT Exchange Repulsion curve is almost entirely obscured by the Multipole Pauli Repulsion and Isotropic Pauli Repulsion curves. See text for definition of S^2/R and Density Overlap models.

The two Pauli repulsion models parameterized in this work compare quite well with the SAPT2+ result. Both the Isotropic and Multipolar Pauli Repulsion models capture magnitude of the interaction as well as the shape across the range of distances. This result is borne out by the S^2/R comparison also shown in figure 5.5. Although this is less straightforward to compute for polyatomic molecules, we defined R as the O-O distance and computed S^2 according to equation 44 with each monomer's MOs. This measure is also in very good agreement with the SAPT

results, with the divergence at short range likely due to our neglect of the hydrogens in the definition of R.

Also shown in figure 5.5 are the results from the QM density overlap calculation. The proportionality constant, K_{ij} for this model was chosen (as with the QM S^2/R model) to reproduce the SAPT repulsion energy at the equilibrium distance. One can see that the quantitative agreement is comparably good for this model. However, the density overlap model does show the same characteristic distance dependence problem illustrated for the noble gas dimers; it is slightly too repulsive at short range and not repulsive enough at long range. This erroneous distance dependence arises due to the unit consistency issue identified in Section 2. It is worth noting that for practical simulation purposes, this error in the radial dependence may be tolerable, given other sources of error in a force field.

Another important slice of the water potential energy surface is the "flap angle" energy dependence of the water dimer.[74,75] High level *ab initio* calculations predict this angle (θ in figure 5.6 inset) to be 57°.[76,77] Typical 3- and 4-site point charge force fields for water such as TIP3P, SPC, and TIP4P generally predict a flap angle to be too flat[78] (less than 57°) due to their inability to reproduce the molecular quadrupole moment of water. The opposite behavior was observed by Ren and Ponder when developing the original AMOEBA water model. Prior to their decision to scale down the quadrupole moments, the AMOEBA water model reproduced the molecular quadrupole moment very well, but predicted a flap angle of 70°.[31] A scaling of the quadruple moments by 70% served to correct the angle. Electrostatics, however, are only half of the story of the water dimer flap angle. Figure 5.6 shows that, in fact, *ab initio* electrostatics do strongly favor a flap angle of about 70°.

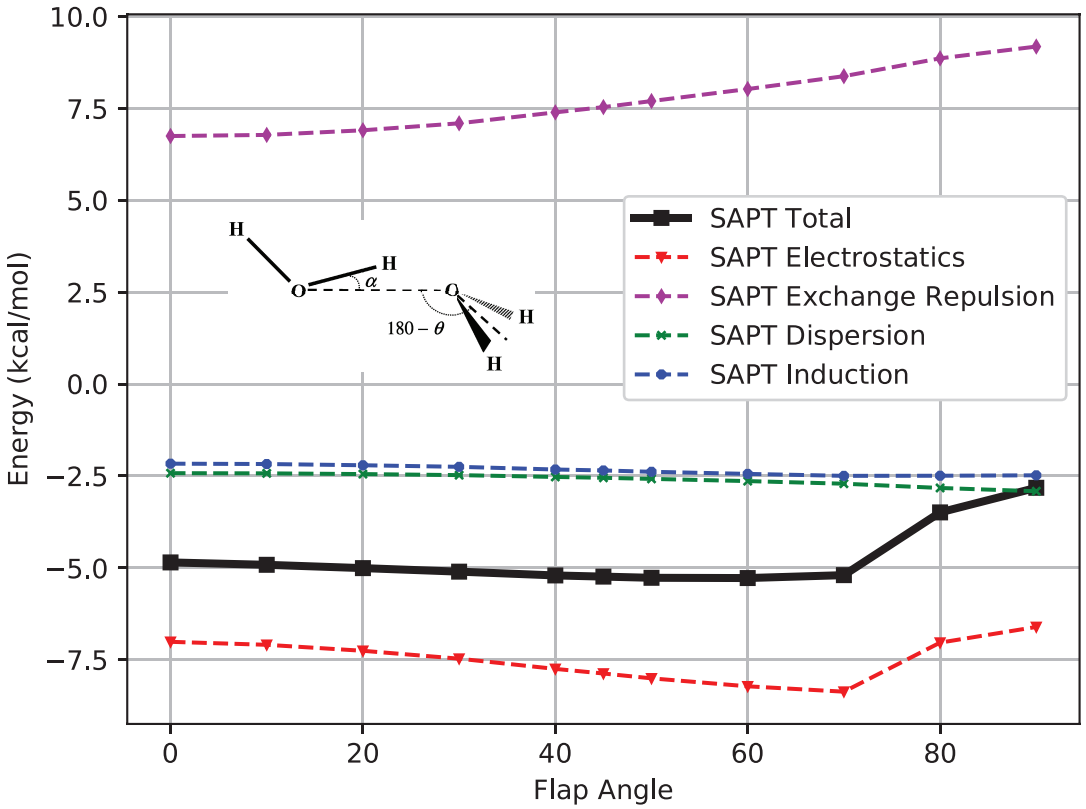


Figure 5.6. Water dimer "flap angle" SAPT energy decomposition analysis.

While the dispersion and induction components of the total energy are relatively flat across this slice of the potential energy landscape, the electrostatic and exchange repulsion change substantially and in opposite directions. The two trends largely cancel each other out in the total energy.

However, this does not correspond to the behavior of the total energy surface, which is basically flat for angles from 45° to 70°. To get this flat surface requires a compensating contribution from Pauli repulsion, and indeed figure 5.6 shows that as the flap angle is increased through this range, while the electrostatic energy consistently becomes more negative, the Pauli repulsion energy steadily trends more positive.

Which, if any, molecular mechanics models are capable of capturing this kind of phenomenon? Shown in figure 5.7 are several water dimer Pauli repulsion models evaluated for a range of flap angle values.

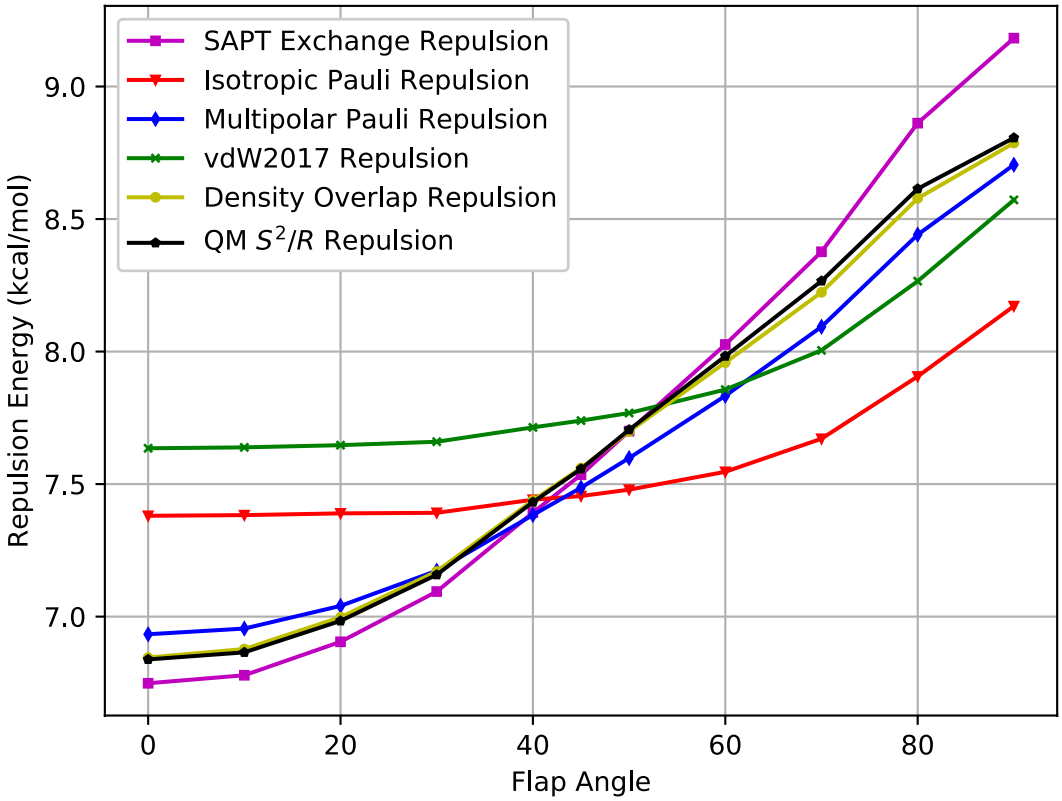


Figure 5.7. Water dimer "flap angle" Pauli repulsion.

The isotropic models (Isotropic Pauli Repulsion and vdW2017) clearly miss the sensitivity to angle change while the anisotropic models mirror the shape of the SAPT Exchange Repulsion.

The Multipolar Pauli Repulsion model as well as the QM-based S^2/R and density overlap methods all reproduce the shape of the angular dependence of the water dimer repulsion well. The Multipolar Pauli Repulsion Model, despite not being fit to any water angular dependence data (there are no angular scans in S101x7), reproduces the SAPT data quite well. This is due

entirely to the natural description of anisotropy that comes through the atomic multipoles. The fact that the electrostatic and exchange repulsion curves in figure 5.6 are nearly mirror images for a fixed distance is not a coincidence. The interpretation of this result is that while the quadrupole interactions become more attractive as the flap angle is increased, this same overlap causes the repulsion to increase as well. It is the same underlying change in overlap that is driving both trends. This is again borne out by the *ab initio* S^2/R calculation (evaluated in the same way as before) which also mirrors the SAPT result. Interestingly, the density overlap model also does very well in describing this angular dependence. This result makes sense because, in contrast to the dissociation case, the distance between atoms for this slice of the surface is largely unchanged throughout the scan. This means that the density overlap distance dependence problem is hidden, while the accounting of anisotropy (in this case implicitly through the density) gives the correct angular trend.

What is apparent from figure 5.7, however, is that isotropic Pauli repulsion models cannot capture the flap angle dependence of the water dimer. Neither the Isotropic Pauli Repulsion model nor the vdW2017 model experience any change in the repulsion energy until the flap angle becomes large enough that the hydrogen atoms of the acceptor molecule swing around to feel the repulsion of the donor oxygen. These models completely miss the quadrupole repulsion effect responsible for the shape of the flap angle repulsion curve.

The water dimer flap angle provides an excellent test case for cancellation of errors in advanced force fields. As force fields work to reproduce energy components individually, care must be taken to advance the physical models of each part in concert. If the electrostatic model is advanced to include anisotropy without including any such anisotropy in the repulsion, the electrostatics part of the force field will accurately capture that component of the energy.

However, for the flap angle degree of freedom, this will incur an error in the total energy of over 2 kcal/mol over an area in which the total energy should be essentially flat! Figure 5.7 shows that no cancellation of errors scheme for an isotropic repulsion model is sensitive to this degree of freedom; the only way to correct it is to include anisotropy in the repulsion as well. This is the reason why the original AMOEBA water model deviated from the physically-derived electrostatic model and scaled down the quadrupoles. Having a fully anisotropic Pauli repulsion function means that these components sit at the same level of theory, and this in turn allows us to regain sensitivity to cancellation of errors in the angular degrees of freedom.

The angular dependence of Pauli repulsion is not only apparent in the minimum energy water dimer. We also considered the ten water dimer structures introduced by van Tschumper, *et al.*[77] Figure 5.8 shows the error, relative to the SAPT2+ Exchange Repulsion energy for the Multipolar Pauli Repulsion, Isotropic Pauli Repulsion and vdW2017 models.

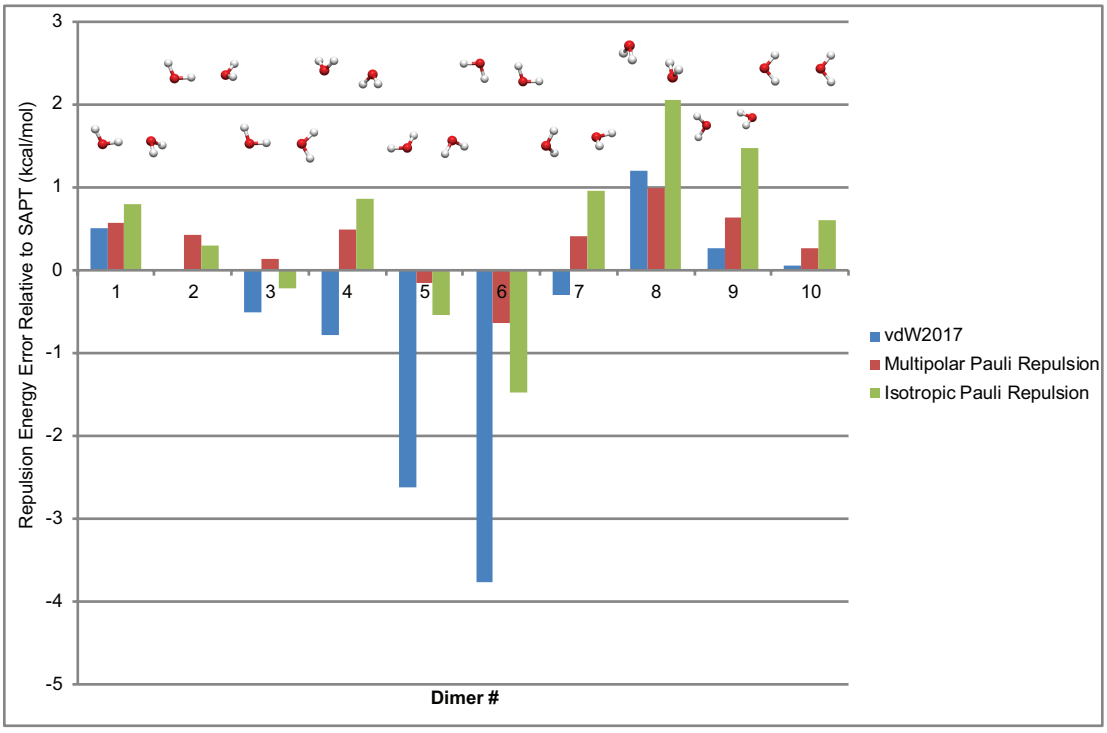


Figure 5.8. Pauli repulsion energy error for ten water dimers.

The error (SAPT minus model) is plotted for each configuration. The isotropic models exhibit errors of opposing sign for dimers 5 and 6 *vs.* 8 and 9. The Multipolar Pauli Repulsion model does not suffer from this constraint.

The Multipolar Pauli repulsion model displays errors of less than 1 kcal/mol for every dimer configuration. The two isotropic models, however, suffer from large, nonrandom errors on several of the dimers. Both the vdW2017 and Isotropic Pauli Repulsion models feature large and opposing errors on dimers 6 and 8. This indicates an angular repulsion dependence that an isotropic model is incapable of capturing. In fact, we attempted to fit the Isotropic Pauli Repulsion model directly to Exchange Repulsion energies of the ten water dimers and found the same opposing errors for dimers 6 and 8.

Taken as a whole, all of the data presented for water repulsion interactions tells a consistent story – that not including repulsion anisotropy on the water dimer potential energy surface will incur errors of 1-2 kcal/mol for accessible dimer configurations. It also shows the Multipolar Pauli Repulsion model is capable of bringing those errors down to ~ 0.5 kcal/mol. The ability of the model to predict angular dependence that is not in the fitting set, such as the flap angle and dimers 2-10, suggests the electrostatic multipole description is a natural fit for Pauli repulsion anisotropy.

5.4.4 The "Sigma Hole" Effect

As stated in the Introduction, halogen bonding *vis a vis* the so-called "sigma hole" effect is of particular interest to biomolecular force fields. Over 35% of drugs in clinical phase III trials contain at least one halogen atom.[10] The "sigma hole" terminology refers to the area of positive charge found at the distal tip of the halogen atom in a halogen-containing compound. It has long been accepted that this feature suggests the linear halogen B…X–Y bonding geometry characteristic of the "sigma hole" effect is driven by electrostatics.[79] Anthony Stone showed this

assumption is only partially correct. Using simple model systems, Stone showed that while electrostatics is indeed responsible for the overall attraction that causes halogen bonds to form, Pauli repulsion is largely responsible for the characteristic, often linear, geometry of these bonds.[4] Given the importance of halogen bond interactions we chose to consider a range of test systems to assess the quality of the Multipolar Pauli Repulsion model. We consider a pair of representative examples from Stone's work, a halobenzene example proposed by Nohad Gresh and co-workers,[13] an acetone-bromobenzene dimer suggested by Hobza and co-workers[80] and a drug-like dimer system from Alzate-Morales and co-workers.[12]

From Stone's work we consider two representative halogen bonding configurations: the "head-on" ammonia–ClF dimer and the "from the side" ethene–ClF dimer. Figures 5.9 and 5.10 show the results for the models on ammonia–ClF and ethene–ClF dimer respectively.

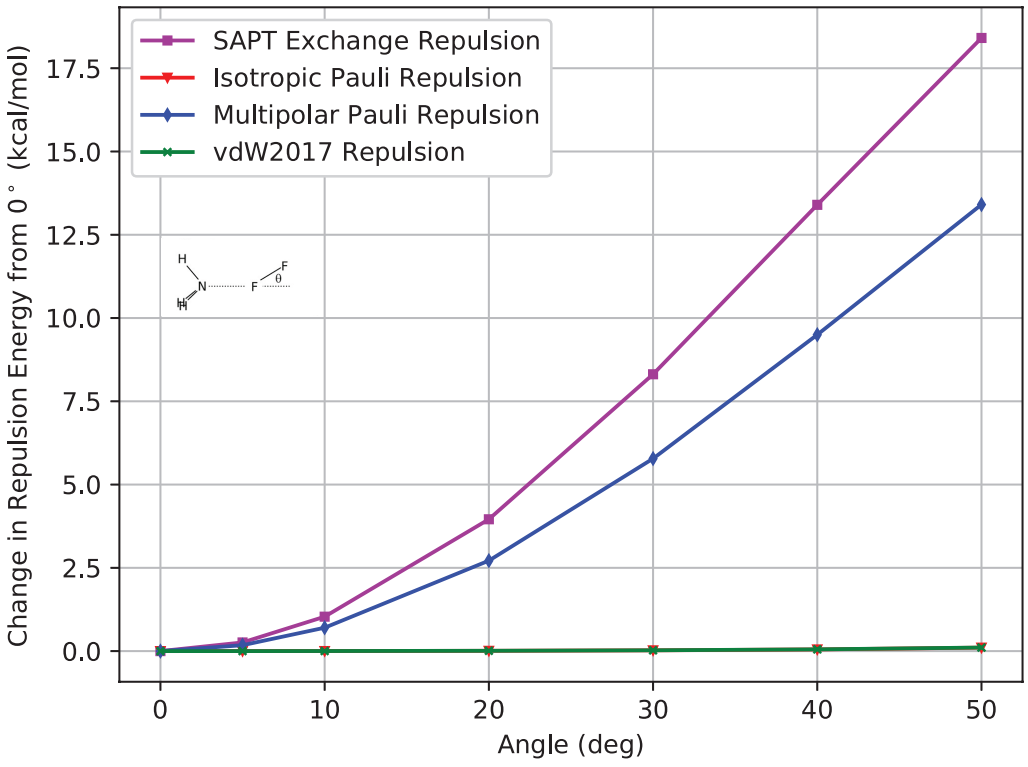


Figure 5.9. Variation of the Pauli repulsion energy with respect to tilt angle (B…X–Y) for the ammonia–ClF dimer.

The N to Cl distance is fixed at 2.376 Å. The Isotropic Pauli Repulsion and vdW2017 lines are indistinguishable.

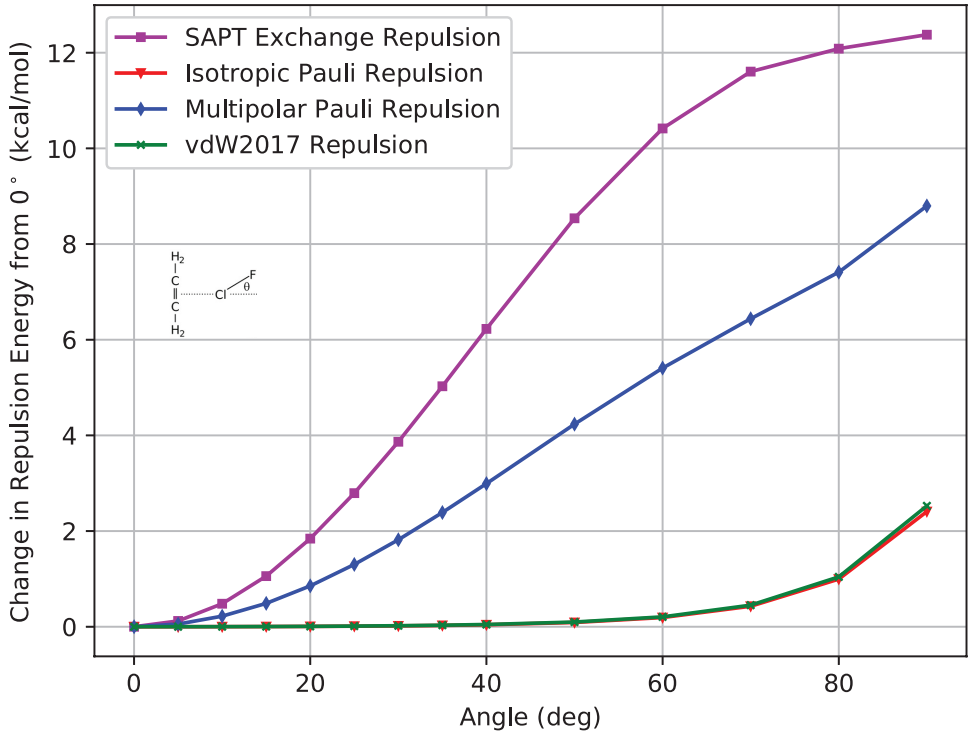


Figure 5.10. Variation of the Pauli repulsion energy with respect to tilt angle (B…X–Y) for the ethene–ClF dimer.

The C=C bond midpoint to Cl distance is fixed at 2.766 Å.

As the B…X–Y angle varies away from linear for both systems the *ab initio* Pauli repulsion rises sharply. The isotropic models miss this entirely. For ammonia–ClF, the repulsion energy of the Isotropic Pauli Repulsion and vdW2017 models is virtually flat throughout the scan and for ethene–ClF these models only start to vary once the fluorine swings around far enough to interact directly with the ethene molecule. This indicates that isotropic Pauli repulsion models

will miss the strong linear preference of these halogen-bonded complexes. The Multipolar Pauli Repulsion model, however, does not miss this angular effect. In both cases the anisotropic model correctly captures the immediate increase in Pauli repulsion that occurs as the halogen bond deviates from linearity. Although the Multipole Pauli repulsion model underestimates the anisotropy of repulsion in both examples, this is most likely a consequence of the DMA-based protocol used for determining multipole moments. For Cl-F the DMA Cl dipole and quadrupole moments differ from those of similar halogens in the S101 database. This is likely responsible for the difference, since the angular dependence of repulsion is driven by dipole and quadrupole interactions. However, the general qualitative agreement between SAPT and the Multipolar Pauli Repulsion model shows that a multipole-based description of electrostatics works well to describe the angular dependence of halogen bond repulsion in these cases.

The concept that force field anisotropy is necessary to accurately model halogen bonding is not new. Recent studies using the AMOEBA force field[81] as well as the SIBFA force field[13] have explored this idea. Both works stressed the importance of anisotropic electrostatics, but largely neglected a discussion about anisotropic repulsion. In particular the work of Gresh and co-workers studied the interactions of halobenzene–water complexes. To examine the repulsive contribution to these interactions, we chose the chlorobenzene–water dimer as a test system. Figure 5.11 shows the halogen bond angle scan for this system.

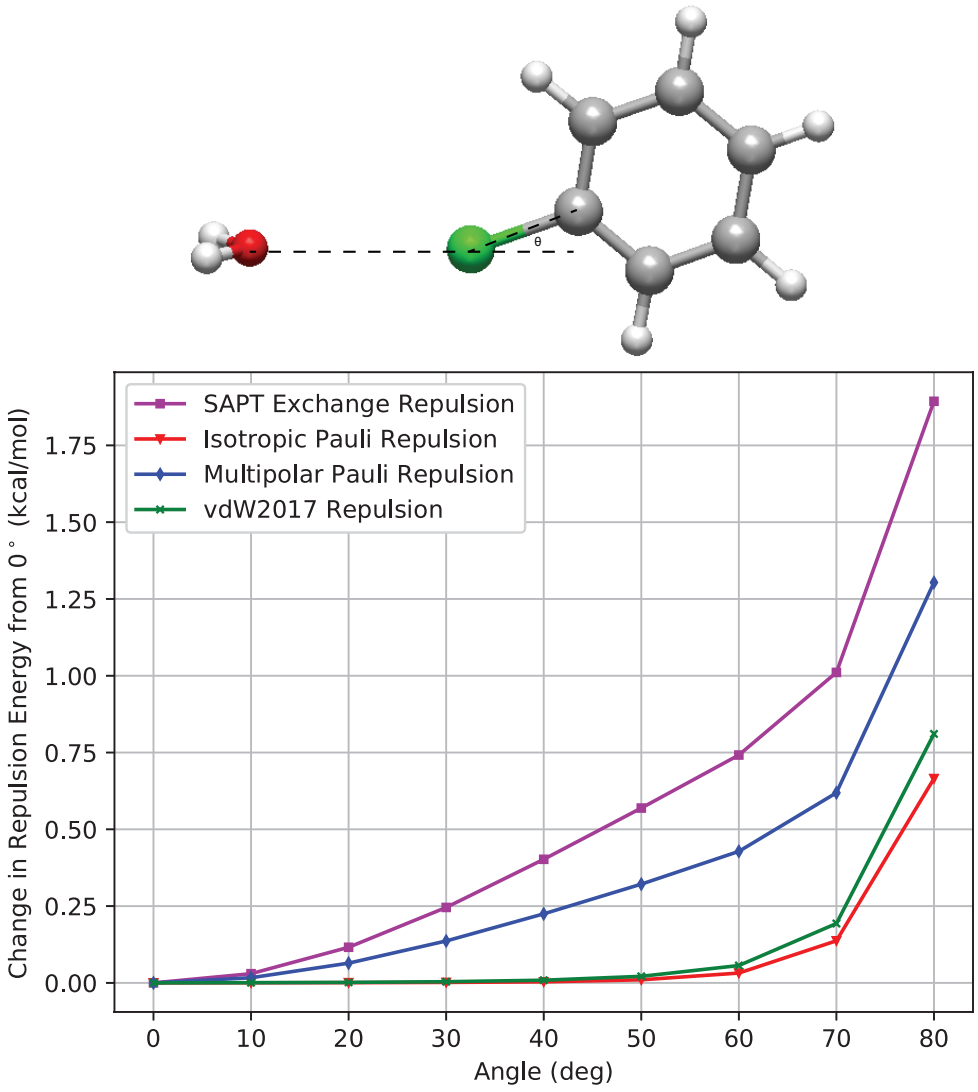


Figure 5.11. Variation of the Pauli repulsion energy with respect to tilt angle (B…X–Y) for the water-chlorobenzene dimer.

The O to Cl distance is fixed at 3.33 angstroms.

Although the energy variation is smaller for this system since the O to Cl contact distance is long, at 3.33 angstroms, the model trends are clear. The Multipolar Pauli Repulsion model mirrors the immediate change in repulsion energy that is felt when the complex deviates from

linear. The two isotropic models, however, do not sense this effect. The energy of these models is flat until a rotation of ~ 60 degrees, where steric repulsion begins.

Another useful test system for halogen bonding are bromobenzenes interacting with acetone. A crystallographic survey by Auffinger and co-workers showed that of the halogen bonded structures in the PDB, 70% involved a protein backbone carbonyl oxygen and that of those structures, 94% involved a halogen atom bonded to an aromatic or heterocyclic aromatic ring.[14] Hobza and co-workers proposed the bromobenzene–acetone complex as a simple probe for examining this kind of important halogen bonding. We used this probe to assess the quality of the carbonyl oxygen containing halogen bond behavior of the Multipolar Pauli Repulsion model. Shown in figure 5.12 is an angular scan of the acetone–bromobenzene Pauli repulsion energy surface.

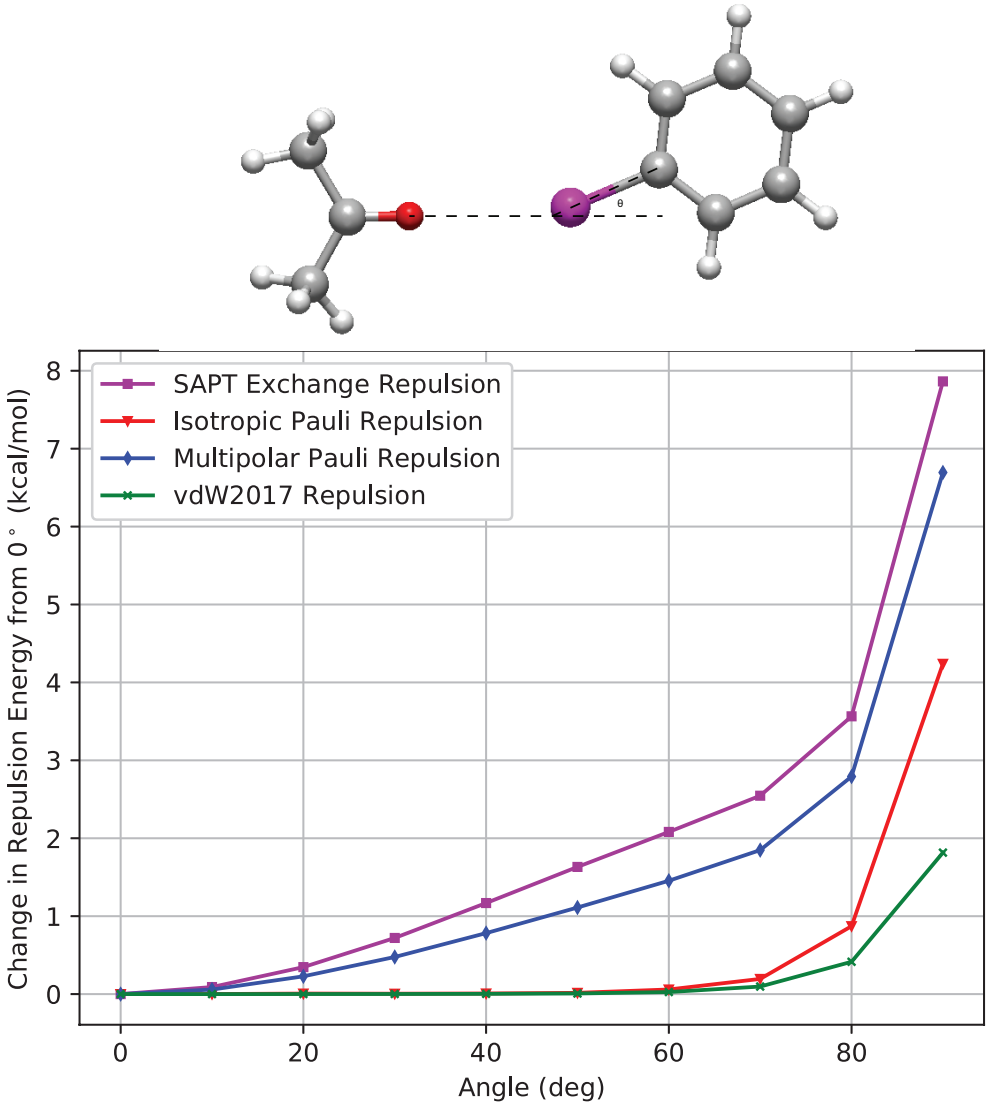


Figure 5.12. Variation of the Pauli repulsion energy with respect to tilt angle (B…X–Y) for the acetone–bromobenzene dimer.

The O to Br distance is fixed at 3.15 angstroms.

Clearly, while the Multipolar Repulsion Model is not in perfect agreement with SAPT, it is the only model that captures the angular dependence trend. The preference the acetone–bromobenzene system shows for the linear configuration is being driven in no small part by this angular dependence in Pauli repulsion. The Multipolar Pauli Repulsion model gets this

dependence qualitatively right. As expected, the isotropic models miss this variation in the rotational degrees of freedom.

The final example of halogen bonding we surveyed was a model "drug binding" system of *N*-methylacetamide (NMA) and chlorobenzene proposed by Alzate-Morales and co-workers.[12] This system was chosen to be a close approximation of a drug-like molecule interacting with a peptide backbone. Again, we performed SAPT calculations of the Exchange Repulsion energy at a range of interaction angles and compared these to the results from each model.

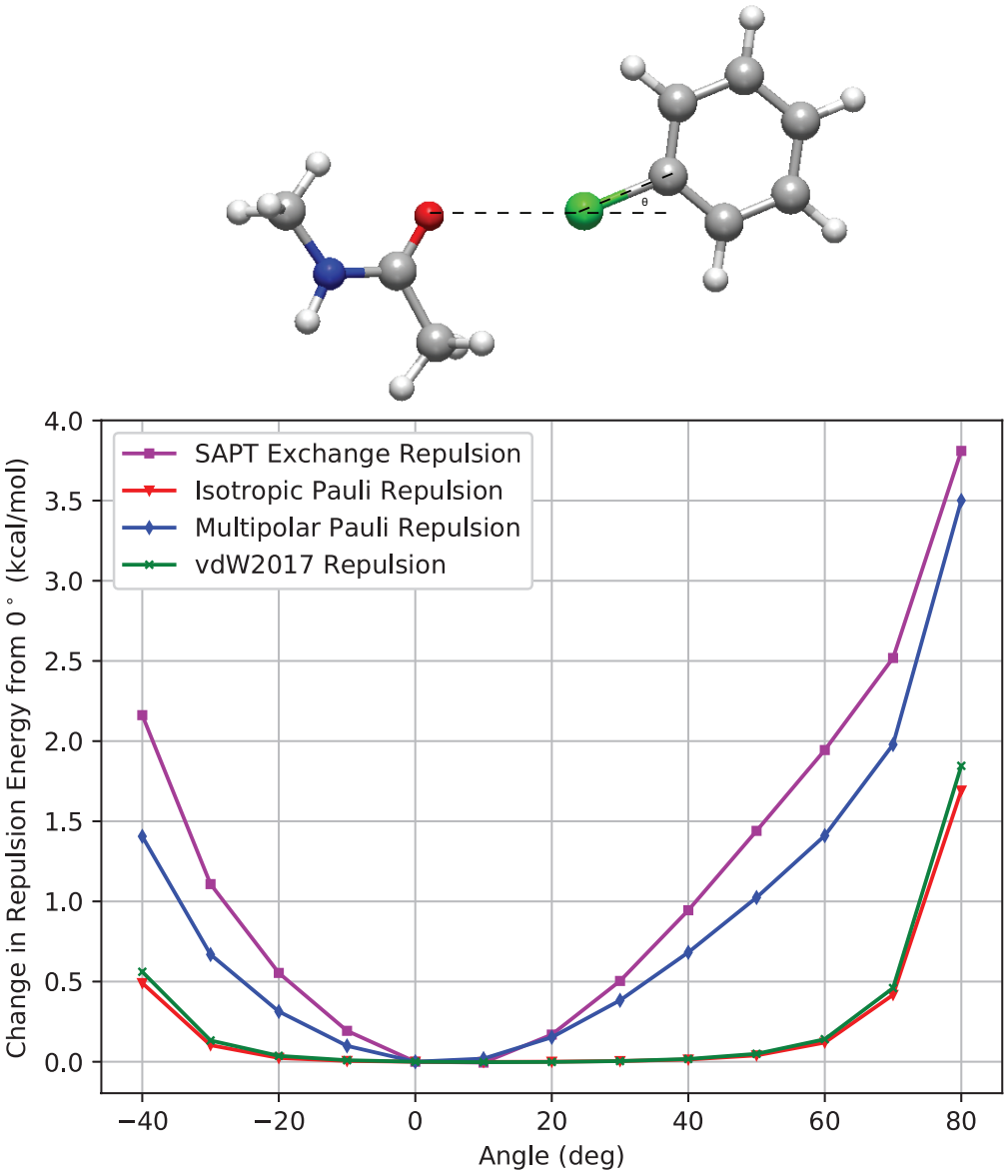


Figure 5.13. Variation of the Pauli repulsion energy with respect to tilt angle (B…X–Y) for the NMA-chlorobenzene dimer.

The O to Cl distance is fixed at 3.0 angstroms.

The results in figure 5.13 confirm the trend of the other test cases. The Multipolar Pauli Repulsion model correctly picks up the trends in both angular directions for the repulsion energy. Since NMA is an asymmetric molecule, we performed a scan from 80 to -40 degrees and found

that the Multipolar Pauli Repulsion model reproduces the increase in the repulsion energy on both sides of the well. The isotropic models both capture the beginning of steric repulsion that occurs at each end of the angular scan but are not sensitive to the anisotropic change in repulsion in the middle. The anisotropy in the Multipolar Pauli Repulsion model is picking up not only the increase in energy associated with rotating away from linear, but also the asymmetry about the O…Cl–C angle.

Much like hydrogen bonds, halogen bonds can be useful tools for molecular design because they are strong and exhibit a marked geometric preference. As has been shown in previous work, much of this strength and some of the geometry preference is expressed through the electrostatic, dispersion and polarization components of the intermolecular energy. However, the component most responsible for enforcing the generally linear geometry of halogen bonds is Pauli repulsion. The tests presented in this section show that no isotropic model (without employing off-atom repulsion sites) is capable of reproducing this effect. Moreover, this section shows that the electrostatic description of monomers via atomic multipole expansions is sufficient to capture the signature anisotropy of these interactions. It bears noting that despite the repetitive feel of the angular halogen bonding results shown, there are no anisotropic parameters being fit. The differences in anisotropy, including the asymmetric shape of the NMA–chlorobenzene well, are entirely determined by the *ab initio* derived atomic multipole moments of the molecules. The results here show that for “sigma hole” type interactions, the multipole moments can simultaneously provide a suitable description of both electrostatic and repulsion anisotropy.

5.4.5 Computational Cost

For any molecular mechanics model that aims to be useful for biomolecular simulation, the computational cost must be considered. In particular for the Multipolar Pauli Repulsion model, this is a matter of concern because multipole calculations are known to be computationally expensive. This model is intended to be used in tandem with a multipole electrostatics model; in particular it is parameterized against the AMOEBA multipole model, so we tested the computational efficiency in that context. The results in table 5.4 show that the additional cost for this model is minimal.

	Time for 100 energy and force evaluations (sec)
AMOEBA Electrostatics with Charge Penetration + Multipolar Pauli Repulsion + Dispersion	1.7
AMOEBA Electrostatics with Charge Penetration + vdW2017	1.4
AMOEBA Electrostatics with Charge Penetration + Polarization + Multipolar Pauli Repulsion + Dispersion	4.6
AMOEBA Electrostatics with Charge Penetration + Polarization + vdW2017	4.4

Table 5.4. Computational cost of the Multipolar Pauli Repulsion model.

Timings are for 100 energy and force evaluations in a standard Open-MP parallel implementation in Tinker.

When the Multipolar Pauli Repulsion Model is paired with our previously published damped dispersion model, the resulting calculations are around 20% slower than the current standard

AMOEBA buffered 14-7 van der Waals function. Furthermore, when this cost is put into the context of the entire AMOEBA energy function, including polarization, the extra cost become nearly negligible. The Multipolar Pauli Repulsion and Overlap Damped Dispersion combination yields a model that is 5% slower than its buffered 14-7 counterpart. Polarization is the costliest component of the AMOEBA force field, so adding a slightly more expensive Pauli repulsion function makes very little difference to the overall computational efficiency of the model.

We note here that this kind of computational efficiency is predicated upon two important factors. First, it relies upon using a multipole model for electrostatics. If one were to deploy the Multipolar Pauli Repulsion model as a stand-alone energy term, it would be an order of magnitude more expensive than a standard van der Waals function. When it is used with a multipole model and given that the multipole moments are constrained to be identical for the electrostatics and Pauli repulsion models, a large amount of the algebra to compute dipole and quadrupole forces is shared between the two models. Second, the speed of the model benefits greatly from employing cutoffs. The standard van der Waals cutoff for the AMOEBA force field is 9 to 12 Å. Because the Multipolar Pauli Repulsion model separates the repulsive and dispersive contributions to the van der Waals energy, it is free to use a much shorter cutoff. These tests were performed with a conservative truncated cutoff of 5 Å, but even better performance can be achieved by shortening this distance and employing a polynomial switching function.

5.5 Discussion and Conclusions

Pauli repulsion is one of the most important parts of any classical intermolecular potential energy model. In most energy decomposition analyses, it is the only component of the total energy that is always positive. This means that for condensed phase systems, total models rely

heavily on the Pauli repulsion term to reproduce bulk phase data. For this reason, it is important for a good Pauli repulsion model to be both accurate and physically interpretable. We have presented here the Multipolar Pauli Repulsion model as an option that fulfills both of these aims.

Despite the terms in which it is often discussed, there is nothing mystical about the phenomenon of Pauli repulsion. It is true the effect arises from the enforcement of the laws of quantum mechanics, but the result can be fully understood within a classical physics interpretation. The Pauli Exclusion Principle demands the total wavefunction of an electronic system be antisymmetric. If we take as our reference the unperturbed monomer wavefunctions of two interacting molecules, then upon overlap the enforcement of antisymmetry will lead to a loss in electron density in the overlap region. That loss of electron density, relative to the unperturbed reference state, causes a straightforward coulombic repulsion between the two nuclei. The Multipolar Pauli Repulsion model presented here follows this explanation and thus gives a classical physical interpretation of this quantum mechanical effect. The electrostatic nature of this effect is borne out by the success of the model in reproducing the anisotropy of repulsion using electrostatic multipole moments.

In addition to being physically interpretable, the Multipolar Pauli Repulsion model is shown to yield good quantitative fits to *ab initio* data. The model fits the S101x7 dataset to an accuracy of 1.7 kcal/mol and fits the near-equilibrium points of that dataset to an error of much less than 1 kcal/mol. This fit spans a large range of chemical space, from hydrogen bonds and halogen bonds to pi-pi interactions and charged species. The results are shown to be transferable to systems outside of the S101 set as well. Particularly, we have shown that the Multipolar Pauli Repulsion model captures the angular dependence of the repulsion energy associated with halogen bonding at a range of contact distances. These test systems show the transferability of

not only the exponential parameters that were fit, but also justify the claim to atomic multipole parameters as a description of anisotropic electron distribution overlap.

This work is certainly not the first to acknowledge the importance of the anisotropy of repulsion in intermolecular interactions and the Multipolar Pauli Repulsion model is not the first model to include such effects. What makes this model noteworthy is that it circumvents two obstacles that have traditionally stood in the way of adopting anisotropic models: parameter underdetermination and computational cost. Any atomic anisotropic model (repulsive or otherwise) requires a local frame and a set of parameters that obey the symmetries of that frame. In the absence of any richer set of data, these requirements mean that if one wishes to fit to intermolecular energies, there will be a large number of parameters to fit to a (usually small, depending on computational resources) set of scalar values. This is a recipe for overfitting and it is the reason that anisotropic repulsive models have largely been limited to specific systems for which large amounts of dimer data can be generated. The Multipolar Pauli Repulsion model evades this problem by constraining the parameters responsible for conferring anisotropy on the model to those that are derived from a much richer data source: the molecular density. The atomic multipolar parameters that are derived from *ab initio* monomer calculations not only constrain the parameter space of the model to avoid overfitting, but moreover, as shown in the Section 2, they do so through a series of theoretically justified approximations. The DMA multipoles come from the same monomer wavefunction that is required to calculate S^2, and this model uses that information to its advantage. This symbiosis not only stands the Multipolar Pauli Repulsion model on solid theoretical ground; it also ameliorates the concern of computational cost typically associated with anisotropic models. Because the local frames and atomic multipoles are identical between the electrostatics and repulsion models, there is very little

additional overhead incurred when using the Multipolar Pauli Repulsion model with a multipolar electrostatics model. Timings show that the cost of implementing a multipole-based repulsion model in an AMOEBA-like force field is minimal. By avoiding the overfitting and cost problems that have proved prohibitive, the Multipolar Pauli Repulsion model presented here provides a blueprint of one tractable way to include anisotropic repulsion in biomolecular force fields.

Not every force field needs the level of detail presented in this model. There are undoubtedly applications for which isotropic, point-based force fields are adequate for predicting quantities of interest. In fact, even for calculations that will require more advanced force fields, the Multipolar Pauli Repulsion model, as presented here is probably insufficient without additional tuning– several of the test cases presented, despite qualitative agreement with SAPT, fall short of accurately reproducing a truly *ab initio* potential energy surface. All force fields, advanced or not, will require some measure of error cancellation.

A final point of this paper is that in order to benefit from cancellation of errors the level of detail across different parts of the model must match. The water dimer data presented here shows this point nicely. If one uses an anisotropic multipolar description of electrostatics that, inevitably, has some error in the associated intermolecular angular degrees of freedom, the only way to cancel that error is by having the other components of the force field be sensitive to those same degrees of freedom. The water dimer example shows that for this important case (and likely many others) it is largely the Pauli Repulsion that provides the balancing force. This is specific evidence of the broader truism in molecular modeling that a theoretically "consistent" model is a good model. For a point charge force field it is possible that including an anisotropic repulsion model might make the model *worse* by introducing error that cannot be cancelled by other components of the force field. Likewise, for force fields based on multipolar electrostatics

models, we suggest that the Multipolar Pauli Repulsion model will not just be more accurate with respect to energy decomposition analysis, it will also confer the ability to achieve favorable cancellation of error across the total model.

While the theoretical framework presented here is applicable to any multipolar force field, the specifics of the parameterization and testing of the model are aimed at a particular goal. The development of the next generation of the AMOEBA force field is underway and the Multipolar Pauli Repulsion model has been constructed explicitly for that purpose. It is intended to be used with a multipolar description of electrostatics that includes our earlier work on charge penetration and in conjunction with our previously published Overlap Damped Dispersion model. Code that implements all of these components in the Tinker Molecular Mechanics software package is freely available on the web.[70] Work combining all of these components into a next-generation water model and full biomolecular force field will be reported in due course. The Multipolar Pauli Repulsion model provides a cheap, intuitive and interpretable way to put this important component of the future force field on an equal footing with its counterparts.

5.6 References

1 Bader, R. F. W. Pauli Repulsions Exist Only in the Eye of the Beholder. *Chem.-Eur. J.* **12**, 2896-2901 (2006).

2 Politzer, P., Murray, J. S. & Clark, T. Mathematical Modeling and Physical Reality in Noncovalent Interactions. *J. Mol. Model.* **21-31**, 52 (2015).

3 Deb, B. M. *Force concept in chemistry*. (Van Nostrand Reinhold, 1981).

4 Stone, A. J. Are Halogen Bonded Structures Electrostatically Driven? *J. Am. Chem. Soc.* **135**, 7005-7009 (2013).

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

6 Shi, Y. *et al.* Polarizable Atomic Multipole-Based AMOEBA Force Field for Proteins. *J. Chem. Theory Comput.* **9**, 4046-4063 (2013).

7 Zhang, C. *et al.* AMOEBA Polarizable Atomic Multipole Force Field for Nucleic Acids. *J. Chem. Theory Comput.* **14**, 2084-2108 (2018).

8 Laury, M. L., Wang, Z., Gordon, A. S. & Ponder, J. W. Absolute Binding Free Energies of the SAMPL6 Cucurbit[8]uril Host-Guest Challenge via the AMOEBA Polarizable Force Field. *J. Comput. Aid. Mol. Des.* **32**, 1087-1095 (2018).
9 Rybak, S., Jeziorski, B. & Szalewicz, K. Many-Body Symmetry-Adapted Perturbation Theory of Intermolecular Interactions. H2O and HF Dimers. *J. Chem. Phys.* **95**, 6576-6601 (1991).
10 Xu, Z. *et al.* Halogen Bond: Its Role beyond Drug–Target Binding Affinity for Drug Discovery and Development. *J. Chem. Inf. Model.* **54**, 69-78 (2014).
11 Grant Hill, J. & Legon, A. C. On the Directionality and Non-Linearity of Halogen and Hydrogen Bonds. *Phys. Chem. Chem. Phys.* **17**, 858-867 (2015).
12 Adasme-Carreño, F., Muñoz-Gutierrez, C. & Alzate-Morales, J. H. Halogen Bonding in Drug-Like Molecules: A Computational and Systematic Study of the Substituent Effect. *RSC Adv.* **6**, 61837-61847 (2016).
13 El Hage, K., Piquemal, J.-P., Hobaika, Z., Maroun, R. G. & Gresh, N. Could an Anisotropic Molecular Mechanics/Dynamics Potential Account for Sigma Hole Effects in the Complexes of Halogenated Compounds? *J. Comput. Chem.* **34**, 1125-1135 (2013).
14 Auffinger, P., Hays, F. A., Westhof, E. & Ho, P. S. Halogen Bonds in Biological Molecules. *Proc. Natl. Acad. Sci. USA* **101**, 16789-16794 (2004).
15 Rackers, J. A. *et al.* An Optimized Charge Penetration Model for Use with the AMOEBA Force Field. *Phys. Chem. Chem. Phys.* **19**, 276-291 (2017).
16 Van der Waals, J. D. *Over de Continuiteit van den Gas-en Vloeistoftoestand* Ph.D. thesis, Leiden, (1873).
17 Pauli, W. Über den Zusammenhang des Abschlusses der Elektronengruppen im Atom mit der Komplexstruktur der Spektren. *Z. Phys.* **31**, 765-783 (1925).
18 Jones, J. E. On the Determination of Molecular Fields.—II. From the Equation of State of a Gas. *Proc. R. Soc. Lon. Ser.-A* **106**, 463-477 (1924).
19 Lennard-Jones, J. E. Cohesion. *P. Phys. Soc.* **43**, 461-482 (1931).
20 London, F. Zur Theorie und Systematik der Molekularkräfte. *Z. Phys.* **63**, 245-279 (1930).
21 McCammon, J. A., Gelin, B. R. & Karplus, M. Dynamics of Folded Proteins. *Nature* **267**, 585-590 (1977).
22 Born, M. & Mayer, J. E. Zur Gittertheorie der Ionenkristalle. *Z. Phys.* **75**, 1-18 (1932).
23 Buckingham, R. A. The Classical Equation of State of Gaseous Helium, Neon and Argon. *Proc. R. Soc. Lon. Ser.-A* **168**, 264-283 (1938).
24 Slater, J. C. The Normal State of Helium. *Phys. Rev.* **32**, 349-360 (1928).
25 Allinger, N. L. Conformational Analysis. 130. MM2. A Hydrocarbon Force Field Utilizing V1 and V2 Torsional Terms. *J. Am. Chem. Soc.* **99**, 8127-8134 (1977).
26 Allinger, N. L., Yuh, Y. H. & Lii, J.-H. Molecular Mechanics. The MM3 Force Field for Hydrocarbons. 1. *J. Am. Chem. Soc.* **111**, 8551-8566 (1989).
27 Allinger, N. L., Chen, K. & Lii, J.-H. An Improved Force Field (MM4) for Saturated Hydrocarbons. *J. Comput. Chem.* **17**, 642-668 (1996).
28 Halgren, T. A. The Representation of van der Waals (vdW) Interactions in Molecular Mechanics Force Fields: Potential Form, Combination Rules, and vdW Parameters. *J. Am. Chem. Soc.* **114**, 7827-7843 (1992).
29 Halgren, T. A. Merck Molecular Force Field. I. Basis, Form, Scope, Parameterization, and Performance of MMFF94. *J. Comput. Chem.* **17**, 490-519 (1996).

30 Halgren, T. A. Merck Molecular Force Field. II. MMFF94 van der Waals and Electrostatic Parameters for Intermolecular Interactions. *J. Comput. Chem.* **17**, 520-552 (1996).
31 Ren, P. & Ponder, J. W. Polarizable Atomic Multipole Water Model for Molecular Mechanics Simulation. *J. Phys. Chem. B* **107**, 5933-5947 (2003).
32 Millot, C. & Stone, A. J. Towards an Accurate Intermolecular Potential for Water. *Mol. Phys.* **77**, 439-462 (1992).
33 Misquitta, A. J. & Stone, A. J. Ab Initio Atom–Atom Potentials Using CamCASP: Theory and Application to Many-Body Models for the Pyridine Dimer. *J. Chem. Theory Comput.* **12**, 4184-4208 (2016).
34 Mas, E. M. *et al.* Water Pair Potential of Near Spectroscopic Accuracy. I. Analysis of Potential Surface and Virial Coefficients. *J. Chem. Phys.* **113**, 6687-6701 (2000).
35 Totton, T. S., Misquitta, A. J. & Kraft, M. A First Principles Development of a General Anisotropic Potential for Polycyclic Aromatic Hydrocarbons. *J. Chem. Theory Comput.* **6**, 683-695 (2010).
36 Salem, L. The Forces Between Polyatomic Molecules. II. Short-Range Repulsive Forces. *Proc. R. Soc. Lon. Ser.-A* **264**, 379-391 (1961).
37 Musher, J. I. & Salem, L. Energy of Interaction between Two Molecules. *J. Chem. Phys.* **44**, 2943-2946 (1966).
38 Murrell, J. N., Randić, M. & Williams, D. R. The Theory of Intermolecular Forces in the Region of Small Orbital Overlap. *Proc. R. Soc. Lon. Ser.-A* **284**, 566-581 (1965).
39 Murrell, J. N. & Shaw, G. Intermolecular Forces in the Region of Small Orbital Overlap. *J. Chem. Phys.* **46**, 1768-1772 (1967).
40 Gresh, N. Energetics of Zn2+ Binding to a Series of Biologically Relevant Ligands: A Molecular Mechanics Investigation Grounded on ab Initio Self-Consistent Field Supermolecular Computations. *J. Comput. Chem.* **16**, 856-882 (1995).
41 Piquemal, J.-P., Chevreau, H. & Gresh, N. Toward a Separate Reproduction of the Contributions to the Hartree− Fock and DFT Intermolecular Interaction Energies by Polarizable Molecular Mechanics with the SIBFA Potential. *J. Chem. Theory Comput.* **3**, 824-837 (2007).
42 Jensen, J. H. & Gordon, M. S. An Approximate Formula for the Intermolecular Pauli Repulsion between Closed Shell Molecules. *Mol. Phys.* **89**, 1313-1325 (1996).
43 Jensen, J. H. & Gordon, M. S. An Approximate Formula for the Intermolecular Pauli Repulsion between Closed Shell Molecules. II. Application to the Effective Fragment Potential Method. *J. Chem. Phys.* **108**, 4772-4782 (1998).
44 Kita, S., Noda, K. & Inouye, H. Repulsive Potentials for Cl−R and Br−R (R= He, Ne, and Ar) Derived from Beam Experiments. *J. Chem. Phys.* **64**, 3446-3449 (1976).
45 Kim, Y. S., Kim, S. K. & Lee, W. D. Dependence of the Closed-Shell Repulsive Interaction on the Overlap of the Electron Densities. *Chem. Phys. Lett.* **80**, 574-575 (1981).
46 Wheatley, R. J. & Price, S. L. An Overlap Model for Estimating the Anisotropy of Repulsion. *Mol. Phys.* **69**, 507-533 (1990).
47 Piquemal, J.-P., Cisneros, G. A., Reinhardt, P., Gresh, N. & Darden, T. A. Towards a Force Field Based on Density Fitting. *J. Chem. Phys.* **124**, 104101 (2006).

48 Duke, R. E., Starovoytov, O. N., Piquemal, J.-P. & Cisneros, G. A. GEM*: A Molecular Electronic Density-Based Force Field for Molecular Dynamics Simulations. *J. Chem. Theory Comput.* **10**, 1361-1365 (2014).
49 Gokcan, H., Kratz, E. G., Darden, T. A., Piquemal, J.-P. & Cisneros, G. A. QM/MM Simulations with the Gaussian Electrostatic Model, A Density-Based Polarizable Potential. *The journal of physical chemistry letters* (2018).
50 Cisneros, G. A. Application of Gaussian Electrostatic Model (GEM) Distributed Multipoles in the AMOEBA Force Field. *J. Chem. Theory Comput.* **8**, 5072-5080 (2012).
51 Van Vleet, M. J., Misquitta, A. J. & Schmidt, J. R. New Angles on Standard Force Fields: Toward a General Approach for Treating Atomic-Level Anisotropy. *J. Chem. Theory Comput.* **14**, 739-758 (2018).
52 Van Vleet, M. J., Misquitta, A. J., Stone, A. J. & Schmidt, J. R. Beyond Born–Mayer: Improved Models for Short-Range Repulsion in ab Initio Force Fields. *J. Chem. Theory Comput.* **12**, 3851-3870 (2016).
53 Nobeli, I., Price, S. L. & Wheatley, R. J. Use of Molecular Overlap to Predict Intermolecular Repulsion in N⋯ H—O Hydrogen Bonds. *Mol. Phys.* **95**, 525-537 (1998).
54 Mitchell, J. B. O., Thornton, J. M., Singh, J. & Price, S. L. Towards an Understanding of the Arginine-Aspartate Interaction. *J. Mol. Biol.* **226**, 251-262 (1992).
55 Mitchell, J. B. O. & Price, S. L. The Nature of the N—H… O= C Hydrogen Bond: An Intermolecular Perturbation Theory Study of the Formamide/Formaldehyde Complex. *J. Comput. Chem.* **11**, 1217-1233 (1990).
56 Day, G. M. & Price, S. L. A Nonempirical Anisotropic Atom− Atom Model Potential for Chlorobenzene Crystals. *J. Am. Chem. Soc.* **125**, 16434-16443 (2003).
57 Tafipolsky, M. & Ansorg, K. Toward a Physically Motivated Force Field: Hydrogen Bond Directionality from a Symmetry-Adapted Perturbation Theory Perspective. *J. Chem. Theory Comput.* **12**, 1267-1279 (2016).
58 Domene, C., Fowler, P. W., Wilson, M., Madden, P. A. & Wheatley, R. J. Overlap-Model and ab Initio Cluster Calculations of Ion Properties in Distorted Environments. *Chem. Phys. Lett.* **333**, 403-412 (2001).
59 Slipchenko, L. V. & Gordon, M. S. Electrostatic Energy in the Effective Fragment Potential Method: Theory and Application to Benzene Dimer. *J. Comput. Chem.* **28**, 276-291 (2007).
60 Coulson, C. A. Two-Centre Integrals Occurring in the Theory of Molecular Structure. *Math. Proc. Cambridge* **38**, 210-223 (1942).
61 Inouye, H. & Kita, S. Experimental Determination of the Repulsive Potentials between K+ Ions and Rare-Gas Atoms. *J. Chem. Phys.* **56**, 4877-4882 (1972).
62 Söderhjelm, P., Karlström, G. & Ryde, U. Comparison of Overlap-Based Models for Approximating the Exchange-Repulsion Energy. *J. Chem. Phys.* **124**, 244101 (2006).
63 Nyeland, C. & Toennies, J. P. Modelling of Repulsive Potentials from Atom Charge Density Distributions: Interactions of Inert Gas Atoms. *Chem. Phys. Lett.* **127**, 172-177 (1986).
64 Andreev, E. On Asymptotic Calculation of the Exchange Interaction. *Theor. Chim. Acta* **28**, 235-239 (1973).
65 Wang, Q. *et al.* General Model for Treating Short-Range Electrostatic Penetration in a Molecular Mechanics Force Field. *J. Chem. Theory Comput.* **11**, 2609-2618 (2015).

66 Rackers, J. A., Liu, C., Ren, P. & Ponder, J. W. A Physically Grounded Damped Dispersion Model with Particle Mesh Ewald Summation. *J. Chem. Phys.* **149**, 084115 (2018).
67 Parker, T. M., Burns, L. A., Parrish, R. M., Ryno, A. G. & Sherrill, C. D. Levels of Symmetry Adapted Perturbation Theory (SAPT). I. Efficiency and Performance for Interaction Energies. *J. Chem. Phys.* **140**, 094106 (2014).
68 Qi, R., Wang, Q. & Ren, P. General van der Waals Potential for Common Organic Molecules. *Bioorgan. Med. Chem.* **24**, 4911-4919 (2016).
69 Ren, P., Wu, C. & Ponder, J. W. Polarizable Atomic Multipole-Based Molecular Mechanics for Organic Molecules. *J. Chem. Theory Comput.* **7**, 3143-3161 (2011).
70 Rackers, J. A. *et al.* Tinker 8: Software Tools for Molecular Design. *J. Chem. Theory Comput.* **14**, 5273-5289 (2018).
71 Rackers, J. A. *GitHub Branch for TinkerTools Development*, <https://github.com/JoshRackers/tinker/tree/amoeba2> (2018).
72 Ponder, J. W., Ren, P. & Piquemal, J.-P. *GitHub Site for TinkerTools*, <https://github.com/TinkerTools> (2018).
73 Smith, D. G. A. *et al.* Psi4NumPy: An Interactive Quantum Chemistry Programming Environment for Reference Implementations and Rapid Development. *J. Chem. Theory Comput.* **14**, 3504-3511 (2018).
74 Mahoney, M. W. & Jorgensen, W. L. A Five-Site Model for Liquid Water and the Reproduction of the Density Anomaly by Rigid, Nonpolarizable Potential Functions. *J. Chem. Phys.* **112**, 8910-8922 (2000).
75 Ren, P. & Ponder, J. W. Temperature and Pressure Dependence of the AMOEBA Water Model. *J. Phys. Chem. B* **108**, 13427-13437 (2004).
76 Klopper, W., van Duijneveldt-van de Rijdt, J. G. C. M. & van Duijneveldt, F. B. Computational Determination of Equilibrium Geometry and Dissociation Energy of the Water Dimer. *Phys. Chem. Chem. Phys.* **2**, 2227-2234 (2000).
77 Tschumper, G. S. *et al.* Anchoring the Water Dimer Potential Energy Surface with Explicitly Correlated Computations and Focal Point Analyses. *J. Chem. Phys.* **116**, 690-701 (2002).
78 Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **79**, 926-935 (1983).
79 Riley, K. E. *et al.* Halogen Bond Tunability II: The Varying Roles of Electrostatic and Dispersion Contributions to Attraction in Halogen Bonds. *J. Mol. Model.* **19**, 4651-4659 (2013).
80 Riley, K. E., Murray, J. S., Politzer, P., Concha, M. C. & Hobza, P. Br··· O Complexes as Probes of Factors Affecting Halogen Bonding: Interactions of Bromobenzenes and Bromopyrimidines with Acetone. *J. Chem. Theory Comput.* **5**, 155-163 (2008).
81 Mu, X. *et al.* Modeling Organochlorine Compounds and the σ-Hole Effect Using a Polarizable Multipole Force Field. *J. Phys. Chem. B* **118**, 6456-6465 (2014).

Chapter 6: Water

6.1 Introduction

For the most part, biology happens in water. That is to say the behavior of biological molecules occurs almost entirely within an aqueous environment. This simple fact makes water perhaps the most important molecule to model accurately in simulations of a biomolecular system. In all-atom molecular dynamics of proteins, for instance, it is common for >80% of the atoms in a simulation to belong to water. Despite its simple molecular structure, water is not a spectator to the goings-on of molecular biology. Water's specific properties and the nature of its interactions with biomolecules underpins an immense number of important biochemical phenomena, from the hydrophobic effect to screening effects to the concept of "buried pockets". For this reason, it is important for a water model to not only reproduce the properties of liquid water, but also for it to be compatible with the other parts of the biomolecular force field (protein, nucleic acid, small molecule, etc.). This need for both accuracy and compatibility has been a dilemma in the force field development community for decades. In this work, we introduce a new class of force field that presents a framework for satisfying both requirements. The HIPPO (Hydrogen-Like Intermolecular Polarizable Potential) model uses a groundwork of first-principles physics to reproduce the properties of pure water while remaining compatible with the broader biomolecular force field.

In a sense this work is testing a hypothesis. The proposition is that by grounding a classical force field in the first-principles physics of intermolecular interactions, we can produce a model that is accurate and efficient enough for use in biomolecular simulations. HIPPO tests this premise by introducing two novel features for a biomolecular force field. First, HIPPO includes a model atomic density on each atom. And second, HIPPO is derived and parameterized

to reproduce the *ab initio* electrostatic, polarization, dispersion and exchange-repulsion components from Symmetry Adapted Perturbation Theory (SAPT). As we will show here and have shown in previous works detailing the parts of the model, the former is necessary to achieve the latter. The density model makes HIPPO a new class of biomolecular force field, capable of accurately capturing the experimental condensed phase properties of water without sacrificing the quantum mechanical description of the molecule.

There are some good reasons to think that the above hypothesis may prove to be true, and that the project of constructing a density-based, first-principles-rooted water model may be worth the investment. Conceptually, the HIPPO model sits in a space that, as of yet, has been only partially explored by classical force fields for water. As illustrated in figure 6.1, there are two wings to the spectrum of current classical water models.

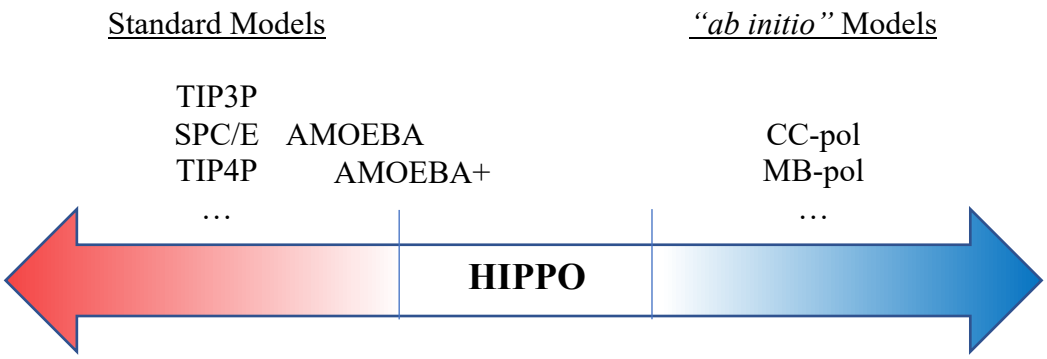


Figure 6.1. The Spectrum of Water Models

On one side is the point charge force field functional form. This is the most heavily explored and crowded part of the spectrum. Models like SCP/E, TIP3P, TIP4P and their numerous progeny all use the same point charge electrostatics + Lennard-Jones van der Waals energy function with

different parameters.[1-3] These models are parameterized empirically to reproduce the properties of liquid water. On the other side of the spectrum are the boutique classical water models. Models like CC-pol and MB-pol take much more complex functional forms that are not explicitly tied to a physical explanation.[4,5] Instead the high number of parameters are fit to a large number of data points on small clusters of water obtained with high-level quantum mechanics (QM) calculations. Both ends of this spectrum have problems with regard to having a water model that is both accurate for pure water and compatible with biomolecules. The empirical side relies on cancellation of errors. Although the cancellation may work in some cases, it is not guaranteed to work in all. On the other side, the "*ab initio*" models suffer from specificity. They work well for water but extending them to biomolecular simulations is intractable due to the combinatorial explosion of QM data needed to fully characterize complex systems. The middle of this spectrum, however, has been explored only sparingly. Toward the empirical end sits AMOEBA and the recently published AMOEBA+.[6,7] These models use a more physics-rich functional form than the point charge force fields but are still parameterized largely empirically. HIPPO aims to fill this open space in the spectrum, because it may be where an answer to the accuracy vs. compatibility dilemma lies.

In order to fill this space, there are several qualities that a force field will likely need. The following list is not prescriptive, but years of experience in water models has suggested four important guidelines of a good atomistic water force field.

1. The model should rely only sparingly on cancellation of errors. The proliferation of point charge water models and the commensurate lists of compatible biomolecular force fields is evidence to this point.

2. The model should have some method of constraining parameter space. All general biomolecular force fields will require some level of optimization. In order to be transferable, that available parameter space for optimization has to be small.
3. The model should be polarizable. The dielectric of various biomolecules can vary dramatically, so any water model should be able to respond to these changes in environment.
4. The model should be anisotropic. The potential energy surface of the smallest unit of water, the water dimer is highly anisotropic. The most accurate atomistic interatomic potentials have all required some level of atomic anisotropy.

Importantly, the absolute accuracy of the water model is not an element of this list. While there may yet come a day when a fully *ab initio* model is capable of accurately simulating water *and* the biomolecules it solvates, this is currently intractable. There are dozens of published water models with results more in agreement with experiment or QM data than what will be presented here. The goal of the HIPPO model is not to best those models in their intended purpose. (Although, as we will show, the agreement with experiment is, on the whole, very good.) The goal of the HIPPO model is to produce a model that gives satisfactory pure water results, while not straying from what makes it a groundwork for the larger project of a full, self-compatible biomolecular force field.

In this work, we will present a water model that satisfies the blueprint laid out for constructing a water model that may be able to solve the accuracy vs. generalizability dilemma. The HIPPO water model is based on a simple approximation: that every atom can be represented as a point core charge, surrounded by a model electron density. Every term of the force field, electrostatics, polarization, dispersion and repulsion, is derived from this density-based model

and parameterized against *ab initio* Symmetry Adapted Perturbation Theory. This ensures not only that we reduce the cancellation of errors relative to what is necessary in point charge force fields, but also that the parameter space for the model is tight and well-defined. The model includes polarization with a polarizable induced dipole model and has atomic anisotropy in the electrostatics, polarization and repulsion terms. The hypothesis is that these ingredients, with the parameter space restrictions they entail, will be capable of being both accurate for pure water and compatible with the rest of the biomolecular force field. This work tests the first part of that hypothesis, with the work on the second still in progress.

6.2 Theory

Part of what makes the HIPPO model unique is that it is not only parameterized against an *ab initio* perturbation theory method, but it is *derived* from a perturbation theory approach. This has been detailed for each individual term in our previous works, but a brief summary here bears repeating.[8-10] The groundwork for the model is the unperturbed monomer. In SAPT, this is the monomer electron density. In HIPPO this is represented by a multipole expansion of that density. Starting from that base, the first order of perturbation theory is the electrostatic interaction between monomer electron densities. HIPPO approximates this by introducing a model electron density and computing Coulomb's law between atomic density sites. The second order of perturbation theory consists of the polarization and dispersion. The uncorrelated polarization part is simply the response of one monomer's electron density based on the electric fields it experiences. HIPPO represents this naturally with the electric field generated by the first-order electrostatic model. Dispersion also involves the overlap of charge densities and HIPPO generates just such an approximation using its model densities. Lastly, the perturbation theory expansion must be corrected for the Pauli Exclusion Principle. To antisymmetrize the

wavefunction, an amount of electron density must be removed from the internuclear region and redistributed. The degree of this effect is known to be proportional to the overlap as well, and HIPPO uses its model density to approximate this overlap. This is a qualitative description, but it captures the essence of the rationale behind the model. What follows is a quantitative explanation of each term and how it is represented classically.

In the HIPPO force field every atom is represented by two components: a core point charge and a model valence electron density. The atomic electron density emulates that of a hydrogen-like atom,

$$\rho = \frac{Q\alpha^3}{8\pi} e^{-\alpha r} \tag{6.1}$$

where Q is the valence charge of the atom and α gives the shape of the density. This model density is used to derive all four intermolecular energy terms that compose the HIPPO force field,

$$U_{HIPPO} = U_{electrostatic} + U_{induction} + U_{dispersion} + U_{Pauli\ repulsion} \tag{6.2}$$

The general forms and derivations of these terms have been described in references 8, 9 and 10 which describe the piecewise development of the model. To provide a complete picture we offer here a comprehensive definition of each term.

6.2.1 Electrostatic Energy

Like its progenitor, AMOEBA, the HIPPO electrostatic term is anisotropic with multipole moments through quadrupole. Because each atom in the HIPPO force field is represented by a core and a density, the pairwise coulomb interaction has four components. The HIPPO electrostatic energy is defined as,

$$U_{electrostatic}^{HIPPO} = \sum_{i>j} Z_i T_{ij} Z_j + Z_i \boldsymbol{T}_{\boldsymbol{ij}}^{*} \vec{M}_j + Z_j \boldsymbol{T}_{\boldsymbol{ji}}^{*} \vec{M}_i + \vec{M}_i \boldsymbol{T}_{\boldsymbol{ij}}^{\boldsymbol{overlap}} \vec{M}_j \tag{6.3}$$

$$\vec{M} = \left(Q, \left[\mu_x, \mu_y, \mu_z\right], \begin{bmatrix} \Theta_{xx} & \Theta_{xy} & \Theta_{xz} \\ \Theta_{yx} & \Theta_{yy} & \Theta_{yz} \\ \Theta_{zx} & \Theta_{zy} & \Theta_{zz} \end{bmatrix} \right) \tag{6.4}$$

$$T_{ij} = \frac{1}{r_{ij}} \tag{6.5}$$

$$\boldsymbol{T}_{\boldsymbol{ij}}^{*} = \begin{bmatrix} 1 & \nabla & \nabla^2 \end{bmatrix} \left(\frac{1}{r_{ij}} f_{ij}^{damp}\left(r_{ij}\right) \right) \tag{6.6}$$

$$\boldsymbol{T}_{\boldsymbol{ij}}^{\boldsymbol{overlap}} = \begin{bmatrix} 1 & \nabla & \nabla^2 \\ \nabla & \nabla^2 & \nabla^3 \\ \nabla^2 & \nabla^3 & \nabla^4 \end{bmatrix} \left(\frac{1}{r_{ij}} f_{ij}^{overlap}\left(r_{ij}\right) \right) \tag{6.7}$$

where the first term represents the core – core repulsion, the second and third terms represent the core – density attractions and the fourth term represents the density – density repulsion. The M vector contains the multipole moments (charge, dipole and traceless quadrupole) and Q and Z represent the core and density charges constrained to satisfy the relation for the total partial charge, $q_i = Z_i + Q_i$. The f^{damp} and $f^{overlap}$ terms in equations 6.6 and 6.7 are of critical importance. They are a direct result of the electrostatic potential generated by the model density,

$$V(r) = \frac{Q}{r}\left[1 - \left(1 + \frac{1}{2}\alpha r\right) e^{-\alpha r}\right] \tag{6.8}$$

This gives the core – density attractions,

$$U_{core-density} = Z_i V_j\left(r_{ij}\right) = Z_i \left(\frac{1}{r_{ij}} f_{ij}^{damp}\left(r_{ij}\right) \right) q_j \tag{6.9}$$

$$f_{ij}^{damp}\left(r_{ij}\right) = 1 - \left(1 + \frac{1}{2}\alpha_j r_{ij}\right) e^{-\alpha_j r_{ij}} \tag{6.10}$$

yielding the "one-center" damping factor that goes into $\mathbf{T}^{*}$. The density – density repulsion is given by

$$U_{density-density} = \frac{1}{2}\left[\int \rho_i(\boldsymbol{r}) V_j(\boldsymbol{r}) dv + \int \rho_j(\boldsymbol{r}) V_i(\boldsymbol{r}) dv\right] = q_i \left(\frac{1}{r_{ij}} f_{ij}^{overlap}\left(r_{ij}\right) \right) q_j \tag{6.11}$$

$$f_{ij}^{overlap}(r_{ij}) = \begin{cases} 1 - \left(1 + \frac{11}{16}\alpha r_{ij} + \frac{3}{16}(\alpha r_{ij})^2 + \frac{1}{48}(\alpha r_{ij})^3\right) e^{-\alpha r_{ij}}, & \alpha_i = \alpha_j \\ 1 - A^2\left(1 + 2B + \frac{\alpha_i}{2} r_{ij}\right) e^{-\alpha_i r_{ij}} - B^2\left(1 + 2A + \frac{\alpha_j}{2} r_{ij}\right) e^{-\alpha_j r_{ij}}, & \alpha_i \neq \alpha_j \end{cases}$$

$$B = \frac{\alpha_i^2}{\left(\alpha_i^2 - \alpha_j^2\right)^2}, \quad A = \frac{\alpha_j^2}{\left(\alpha_j^2 - \alpha_i^2\right)^2} \tag{6.12}$$

where the integrals are evaluated according to the method of Coulson. The $f^{overlap}$ term is the "two-center" damping factor necessary to compute the fourth term of the HIPPO electrostatic model. The higher order terms necessary for higher multipole interactions are obtained by successive gradient operations applied to each of the damping factors as specified in equation 6.7. In the interest of clarity, the explicit equations for all orders of the multipole interaction energy are enumerated in Appendix C. In the limit of large α, both damping factors tend to unity and the point multipole interaction energy is recovered. In practice, the use of finite densities remedies the so-called charge penetration problem of electrostatics. In total, the electrostatic model has five parameters per atom: a core charge, Z, a valence charge, Q, a dipole moment, μ, a quadrupole moment, Θ, and a damping parameter, α.

6.2.2 Induction (Polarization)

In addition to the permanent core charge and density-based multipoles, HIPPO includes a point inducible dipole at every atomic site. The induction energy of the model is defined as

$$U_{induction}^{HIPPO} = \sum_i \frac{1}{2} \vec{\mu}_i^{ind} \vec{F}_i^{perm} - \sum_{i>j} \left(\varepsilon_i e^{-\eta_j r_{ij}} + \varepsilon_j e^{-\eta_i r_{ij}}\right) \tag{6.13}$$

where the first term represents the polarization energy of the induced dipoles interacting with the permanent electric field and the second term represents a small pairwise exponential charge transfer term. The polarization energy is the source of many-body energy in the force field. The induced dipoles are determined by solving the system of linear equations,

$$\boldsymbol{\mu} = \boldsymbol{\alpha}\left(\boldsymbol{F}^{\boldsymbol{perm}} + \boldsymbol{F}^{\boldsymbol{ind}}\right) \tag{6.14}$$

where the vectors are defined as $\boldsymbol{\mu} = [\mu_1, \mu_2, \mu_3, \ldots, \mu_n]$ and similarly for $\mathbf{F}^{\mathbf{perm}}$ (the permanent field), $\mathbf{F}^{\mathbf{ind}}$ (the induced field) and $\boldsymbol{\alpha}$ (the atomic polarizabilities). This system of equations can be solved either with a variational method, like preconditioned conjugate gradient (which is used in this work) or an analytical method such as the recently developed OPT or TCG schemes.[11-13] The permanent and induced electric fields are calculated in exactly the same manner, with the same parameters, as described in the electrostatic section. For completeness, full equations are detailed in Appendix D. The only additional parameter necessary for polarization is the polarizability, α, of each atom. The charge transfer function requires two parameters per atom: a prefactor, ε, and a damping factor, η.

6.2.3 Dispersion

The dispersion interaction between atoms arises from the interaction energy of correlated, instantaneous induced dipole moments. In the point approximation this gives the canonical $1/r^6$ dependence of London dispersion. Because our model, however, represents atoms' valence electrons as densities, the functional dependence is slightly modified. The dispersion energy between two atoms with instantaneous induced dipoles, μ_i and μ_j, is found by solving Schrödinger's equation,

$$\frac{1}{M}\frac{\delta^2\Psi}{\delta z_i^2} + \frac{1}{M}\frac{\delta^2\Psi}{\delta z_j^2} + \frac{2}{\hbar}\left(E - \frac{1}{2}kz_i^2 - \frac{1}{2}kz_j^2 - U_{dipole-dipole}\right)\Psi = 0 \tag{6.15}$$

where, for the case of correlated, parallel dipoles,

$$\begin{aligned} U_{dipole-dipole} &= \mu_i\left(\nabla^2\left(\frac{1}{r_{ij}}f_{ij}^{overlap}\left(r_{ij}\right)\right)\right)\mu_j = \frac{\mu_i\mu_j}{r^3}\lambda_3^{overlap} - \frac{3(\mu_i r)(\mu_j r)}{r^5}\lambda_5^{overlap} \\ &= \frac{\mu_i\mu_j}{r^3}\left(3\lambda_5^{overlap} - \lambda_3^{overlap}\right) = \frac{\mu_i\mu_j}{r^3}f_{damp}^{dispersion} \end{aligned} \tag{6.16}$$

The damping factors, λ_3 and λ_5, that define f_{damp} for dispersion are derived from the action of the gradient operator and are identical to those stated in the dipole-dipole interaction energy defined in Appendix C. Solving the Schrödinger equation defined in eq 6.15 yields,

$$E = \frac{1}{2}\hbar(\omega_1 + \omega_2) \tag{6.17}$$

$$\omega_1 = \omega_0\sqrt{1 - \frac{2Q^2}{r^3k}f_{damp}^{dispersion}}, \quad \omega_2 = \omega_0\sqrt{1 + \frac{2Q^2}{r^3k}f_{damp}^{dispersion}} \tag{6.18}$$

This energy expression can be effectively approximated with a binomial expansion,

$$\sqrt{1+x} = 1 + \frac{1}{2}\mathrm{x} - \frac{1}{8}\mathrm{x}^2 + \cdots \tag{6.19}$$

so the total energy becomes,

$$E = \hbar\omega_0 - \frac{Q^4\hbar\omega_0}{2r^6k^2} + \cdots \tag{6.20}$$

Subtracting the energy of two infinitely separated dipoles ($\hbar\omega_0$) and substituting the parameter C_6 for $\frac{Q^2\sqrt{\hbar\omega_0}}{\sqrt{2}k}$ gives the pairwise dispersion energy,

$$U_{dispersion}^{HIPPO} = -\sum_{i<j}\frac{C_6^i C_6^j}{r^6}\left(f_{damp}^{dispersion}\right)_{\mathrm{ij}}^2 \tag{6.21}$$

It is well known that accurate modelling of the dispersion energy at short range requires the use of a damping function. This model provides a non-empirical damping function derived from the dipole density – dipole density interaction. The model requires only one C_6 parameter per atom since the parameters for the damping function are fixed to their electrostatic values.

6.2.4 Repulsion

The final element of the HIPPO model is a density-based, multipolar model for Pauli Repulsion. Pauli repulsion is a consequence of the rearrangement of electron density that occurs

when the Pauli exclusion principle is applied to the electron densities of two unperturbed interacting molecules. We have shown that the primary change in electron density upon enforcement of antisymmetry, relative to the unperturbed reference state, is an evacuation of electron density from the internuclear region. We have also shown that the energy associated with this change is proportional to

$$U_{Pauli\ repulsion} \propto \frac{S^2}{r} \tag{6.22}$$

$$S = \int \phi_i \phi_j dv \tag{6.23}$$

where S is the overlap integral between the atomic orbitals on i and j, and r is the internuclear distance. To obtain suitable atomic orbitals to implement this model, we use the rule

$$\rho = \phi^* \phi \tag{6.24}$$

to define real, atomic pseudo-orbitals as:

$$\phi = \sqrt{\rho} = \sqrt{\frac{Q\alpha^2}{8\pi}} e^{\frac{-\alpha r}{2}} \tag{6.25}$$

We can then use these pseudo-orbitals to define the charge-charge portion of the overlap integral,

$$S = \int \phi_i \phi_j dv = \frac{\sqrt{Q_i Q_j}}{\sqrt{r}} \alpha_i \alpha_j f_{exp}^{repulsion}(R) \tag{6.26}$$

where, for $\alpha_i = \alpha_j$,

$$f_{exp}^{repulsion}(R) = \frac{\sqrt{r}}{\alpha^3}\left(1 + \frac{\alpha r}{2} + \frac{1}{3}\left(\frac{\alpha r}{2}\right)^2\right) e^{\frac{-\alpha r}{2}} \tag{6.27}$$

and, for $\alpha_i \neq \alpha_j$,

$$f_{exp}^{repulsion}(R) = \frac{1}{2X^3\sqrt{r}}\left[\alpha_i\left(Xr - 2\alpha_j\right)e^{\frac{-\alpha_j r}{2}} + \alpha_j(Xr + 2\alpha_i)e^{\frac{-\alpha_i r}{2}}\right]$$
$$X = \left(\frac{\alpha_i}{2}\right)^2 - \left(\frac{\alpha_j}{2}\right)^2 \tag{6.28}$$

This allows us to write S^2 in the familiar coulombic form,

$$S^2_{charge-charge} = Q_i T_{pauli} Q_j \tag{6.29}$$

with,

$$T_{pauli} = \frac{\alpha_i^2 \alpha_j^2}{r} \left(f_{exp}^{repulsion}\right)^2 \tag{6.30}$$

where T_{pauli} (and, in turn, S^2) is dominated at short range by the exponential $f^{repulsion}$ term.

The anisotropy of the HIPPO repulsion model is obtained through the multipole moments. Because S^2 has a clearly coulombic form, we can include higher order terms in the same manner as electrostatics,

$$S^2_{total} = \sum_{i>j} \vec{M}_i \boldsymbol{T}_{\boldsymbol{ij}}^{\boldsymbol{repulsion}} \vec{M}_j \tag{6.31}$$

$$\vec{M} = \left(Q, \left[\mu_x, \mu_y, \mu_z\right], \begin{bmatrix} \Theta_{xx} & \Theta_{xy} & \Theta_{xz} \\ \Theta_{yx} & \Theta_{yy} & \Theta_{yz} \\ \Theta_{zx} & \Theta_{zy} & \Theta_{zz} \end{bmatrix} \right) \tag{6.32}$$

$$\boldsymbol{T}_{\boldsymbol{ij}}^{\boldsymbol{repulsion}} = \begin{bmatrix} 1 & \nabla & \nabla^2 \\ \nabla & \nabla^2 & \nabla^3 \\ \nabla^2 & \nabla^3 & \nabla^4 \end{bmatrix} \left(T_{pauli}\right) \tag{6.33}$$

where the multipole moments are identical to those used in the electrostatics calculation. This definition of S^2 allows us to establish an anisotropic repulsion model we call the Multipolar Pauli Repulsion model,

$$U^{HIPPO}_{Pauli\ repulsion} = \sum_{i<j} \frac{K_i K_j}{r_{ij}} S^2_{total} \tag{6.34}$$

For a complete derivation of this model see our recently published work, reference 10. Full equations defining the model, with higher-order terms included, are presented in Appendix E. The HIPPO repulsion model introduces three additional parameters per atom: The proportionality constant, K, the exponential parameter, α, and the valence charge, Q. Note that

although similar in spirit to the electrostatics derivation, the parameters α and Q are allowed to differ from their values in the electrostatic energy term.

6.2.5 Rigid vs. Flexible

A choice must be made when building a water model of whether to allow the monomer to be flexible or not. Previously, the AMOEBA water model allowed monomer flexibility, but this required setting some unphysical parameter values. Specifically, the HOH ideal angle had to be set to 108.5°, wider than the known gas phase angle of 104.5° or the estimated liquid phase angle of ~106°. There are two reasons why employing this kind of empirical fix is problematic. First, in the case of water, in order to have the correct broadening of the HOH angle when transferring from gas phase to bulk Xantheas and co-workers have shown that coupling between the electrostatics and intramolecular energy terms is necessary.[14] This kind of detail is beyond the scope of the HIPPO model and difficult to generalize. The second reason is less practical and more theoretical. The intramolecular vibrational frequencies of the water molecular lie in the range 1000-4000 cm^{-1}. However, the value of k_BT at room temperature corresponds to a frequency of 200 cm^{-1}. This means that not only are these intramolecular vibrations quantized, but they are nearly always in their ground state. In other words, it should be impossible for the bulk environment of water to store any energy in the intramolecular degrees of freedom. This was pointed out most clearly and succinctly by van Gunsteren and co-workers.[15] Given these two considerations, it is likely that the best approximation to the intramolecular degrees of freedom that doesn't incur a large computational cost is to treat the molecule as rigid. This is what HIPPO does. The water monomer is held rigid in the average liquid phase geometry.[16]

6.3 Methods

6.3.1 Parameterization

In this section we will describe how the parameters for the HIPPO water model were obtained. The general strategy follows the perturbation theory framework laid out at the top of section 6.2. First, we set the reference monomer state by computing an *ab initio* electron density for the water molecule and performing distributed multipole analysis. Second, we fit the electron density model and related parameters to reproduce SAPT electrostatic, dispersion and repulsion results. Third, we determine the atomic polarizabilities for the water model. These steps give us a highly constrained set of starting parameters from which to start condensed phase optimization. This optimization was carried out in two stages: an initial global optimization, followed by a gradient-based fine-tuning of the model. This parameterization strategy allows us to satisfy the first-principles grounding of the model while still yielding a highly accurate condensed phase model.

The first step of parameterization is to determine the multipole moments of the model. HIPPO uses the proven AMOEBA multipole parameterization protocol for this.[17] We start with an MP2 monomer density calculation, perform Stone's Distributed Multipole Analysis (DMA) and then tune the dipole and quadrupole moments to fit the potential on a grid of points around the molecule.[18] The final density calculation is performed with an aug-cc-pVTZ basis set and the grid for the potential fit is constructed with shells of points starting one angstrom beyond the van der Waals surface of the molecule.

With the multipole moments established, the next step in parameterization is to set the intermolecular energy term parameters. In order to ensure compatibility with the biomolecular force fields we are developing in concert with this water model, I fit these terms based on the

large S101x7 database of intermolecular interactions, which includes water interacting with itself in addition to other biofragment-like molecules. I first fit the electrostatic damping parameters (work described in reference 8) and then fit the dispersion C_6 parameters (work described in reference 9). Finally, I fit the repulsion parameters (work described in reference 10)

This left the induction (polarization + charge transfer) as the final portion of the model to be determined. SAPT provides no clear metric for separating these two, so I had to choose some other data to fit the polarization model to before fitting the sum to the SAPT induction energy. The atomic polarizabilities were chosen to match the molecular polarizability of the water molecule. I also fit atomic polarizabilities for the other molecules in the S101 database. This set the polarization model. I then fit the charge transfer model to fill in the gap between the energy of this polarization model and the total SAPT induction energy for the S101x7 database (work described in Chapter 3 of this dissertation).

This procedure gave a valuable set of initial parameters. As mentioned in the introduction, HIPPO is not intended to be an "*ab initio"* force field, but it does have a relatively high number of parameters. This set of initial parameters ensured two things. First, it reduced the area of parameter space we had to search in optimization. And second, because the space is highly limited, it guaranteed that even with optimization the water model would remain compatible with the rest of the force field.

The first step of the optimization was to produce good agreement on fundamental properties of water at room temperature. Using the SciPy differential evolution optimizer, we scanned for models with the best agreement for enthalpy of vaporization and water dimer energies. In this procedure, only the dispersion, repulsion and charge transfer parameters were allowed to vary. The dispersion and repulsion parameters were constrained to be within 5% of

the initial values. The charge transfer parameters were allowed to vary broadly. We then took the top five sets of parameters and evaluated these with short liquid simulations. Of these, we selected the set that had the best overall agreement on the density, radial distribution function and dielectric constant at room temperature as well as agreement with SAPT and CCSD(T) data on the ten canonical water dimers.[19,20]

The final step of the optimization was to fine-tune the parameters to reproduce the properties of liquid water across a range of temperatures. For this, we used the ForceBalance program of Lee-Ping Wang.[21] We used data at temperatures of 249.15, 277.15, 298.15 (2x weight), and 373.15 K. The objective function was heavily weighted to reproduce the density and enthalpy of vaporization across those temperatures. Because we have guaranteed that we are close to a minimum, we found that despite these heavy weights on condensed phase data, the quality of the agreement with *ab initio* SAPT and CCSD(T) data degraded negligibly. As of the writing of this dissertation, this ForceBalance optimization is not yet complete. For the final version of this model, due to be published shortly, we have included data at many more temperatures and run longer simulations to converge properties fully. The parameters shown here, however, represent a point very near to what will be the final, fully-tuned model.

6.3.2 Software Implementation

A large chunk of the work of making this optimization procedure possible relied on efficient software implementations of HIPPO. I implemented the first version of HIPPO in the Tinker Molecular Mechanics software package.[22] This version, despite its more complicated functional form, was only 10% slower than the AMOEBA model. This implementation has been cleaned up and is now available in the publicly distributed release version of Tinker on GitHub. However, performing ForceBalance optimization requires a large number of long simulations

and we quickly found that the CPU-based Tinker implementation was too slow to do this effectively. To make the optimization process tractable, I, with assistance from Zhi Wang and Roseane Silva, implemented the HIPPO model in OpenMM.[23] We did this in a local version of our Tinker-OpenMM branch of OpenMM in order to maintain computability with the current Tinker-based work flow.[24] It can be found on my personal GitHub page at https://github.com/JoshRackers/Tinker-OpenMM. This version is over 10x faster than the Tinker CPU version and is allowing the current massive ForceBalance parameter fine-tuning. Work on an implementation of HIPPO in the main release of OpenMM from Stanford has just been completed.

6.4 Results

6.4.1 HIPPO Water Model Parameters

Listed in table 6.1 are the parameters that define the HIPPO water model. The parameters shaded in green are determined entirely by fits to *ab initio* data (MP2 and SAPT). The parameters shaded in red have been optimized according to the procedure laid out above in section 6.3.

parameter	**units**	**value**
O–H bond length	Å	0.97
H–O–H angle value	degree	106.1
O monopole	e	-0.3828
O dipole *Z*	e bohr	0.05477
O quadrupole *XX*	e bohr^2	0.69866
O quadrupole *YY*	e bohr^2	-0.60471
O quadrupole *ZZ*	e bohr^2	-0.09395
H monopole	e	0.1914
H dipole *Z*	e bohr	-0.20097
H quadrupole *XX*	e bohr^2	0.03881
H quadrupole *YY*	e bohr^2	0.02214

H quadrupole *ZZ*	e bohr^2	-0.06095
O polarizability	Å^3	0.795
H polarizability	Å^3	0.341
O Electrostatic Alpha	1/ang	4.7075
H Electrostatic Alpha	1/ang	4.7909
O Repulsion Size	kcal/mol	2.7502
H Repulsion Size	kcal/mol	1.9337
O Repulsion Alpha	1/ang	4.5673
H Repulsion Alpha	1/ang	4.8214
O Repulsion Charge	e	-3.2219
H Repulsion Charge	e	-0.81
O Dispersion C6	sqrt(ang^6*kcal/mol)	16.8783
H Dispersion C6	sqrt(ang^6*kcal/mol)	4.1580
O Charge Transfer Size	kcal/mol	1200.14
H Charge Transfer Alpha	1/ang	3.3837

Table 6.1. Parameters of the HIPPO water model.

Parameters shaded green are fit exclusively to *ab initio* data. Parameters shaded red are optimized for condensed phase results.

The remainder of the results shown in this section are generated from this set of parameters. As indicated in section 6.3, the parameters in the red shaded section of table 6.1 have changed by no more than 5% from their original values.

6.4.2 Gas Phase

Because the HIPPO model is rooted in quantum mechanics, we should expect the model to produce good agreement with gas phase properties of water monomers and dimers. The data shown in table 6.2 shows the properties of the HIPPO water monomer against the experimental gas phase results.

	AMOEBA03	HIPPO	experiment
dipole dz (Debye)	1.771	1.842	1.855

Quadrupole (Buckinghams)			
Qxx	2.502	2.592	2.63
Qyy	-2.168	-2.453	-2.5
Qzz	-0.334	-0.138	-0.13
Polarizability (ang^3)			
αxx	1.672	1.539	1.528
αyy	1.225	1.413	1.412
αzz	1.328	1.458	1.468

Table 6.2. Gas phase monomer water molecule properties.

Across the board, HIPPO is in better agreement with experiment for the monomer. This is largely because the original AMOEBA multipole parameters were not directly taken from *ab initio* calculations. The quadrupole moments in particular were scaled down by 70% to better fit the total water dimer potential energy surface. The other factor at play is the presence of a new polarization model in HIPPO. While AMOEBA used a Thole-type model for polarization, HIPPO uses a model that is wholly consistent with the electrostatics term of the force field. Table 6.2 shows that this yields better agreement with the experimental molecular polarizability of water.

The first step toward condensed phase from the monomer description is the water dimer. Although the water molecule itself is deceptively simple, the water dimer potential energy surface is quite complex. The global minimum of this surface is the well-known, canonical hydrogen-bonded configuration, but there are a wealth of features outside of this particular structure with importance to modeling water. We have examined this potential energy surface in a variety of slices. First, we have examined the dissociation curve as one pulls apart two monomers from the minimum energy configuration. Second, we have analyzed the angular dependency of hydrogen bonding by an angle scan around the minimum energy configuration.

And lastly, we have studied the behavior of the model on a set of ten representative low-lying potential energy minima structures on the water dimer potential energy surface.

The dissociation curve of the water dimer is particularly important because it tells us how water behaves as we compress or expand it. Shown in figure 6.2 are the HIPPO energy components plotted against their respective SAPT components for a set of seven points along the water dimer dissociation curve.

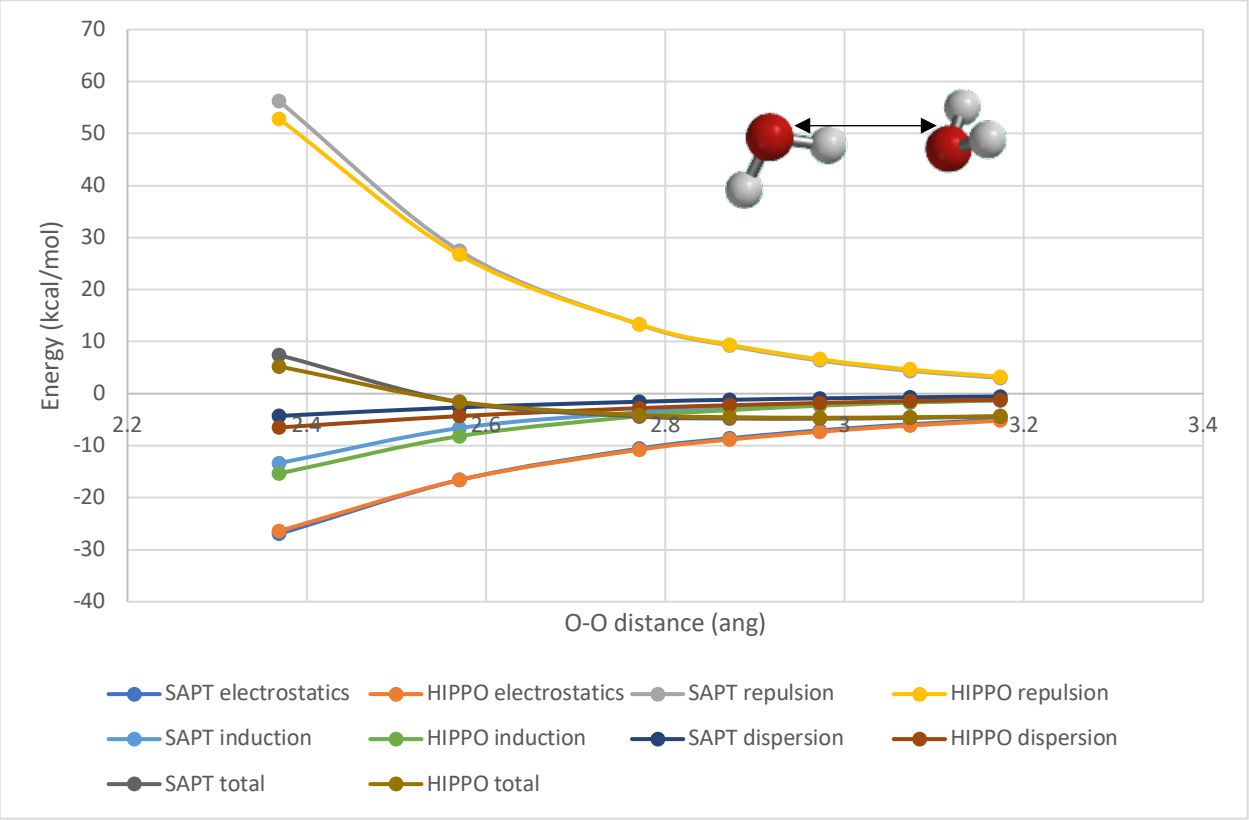


Figure 6.2. Water dimer dissociation energy components.

The energy components of HIPPO and SAPT are plotted against O-O distance (equilibrium ~2.9 ang).

Inspection of figure 6.2 shows that not only is the total energy of the HIPPO model in good agreement with SAPT across the dissociation curve, the components are as well. This a testament to the parameterization procedure. Because we established a set of initial parameters whose only

requirement was fitting their respective energy components, it appears that the final result has strayed very little.

The next slice of the dimer potential energy surface we examined is the angular dependence of the minimum energy water dimer. To make this slice, we took varying values of the water dimer "flap angle", defined in the inset of figure 6.3. For each angle we computed SAPT components and compared them against the HIPPO results.

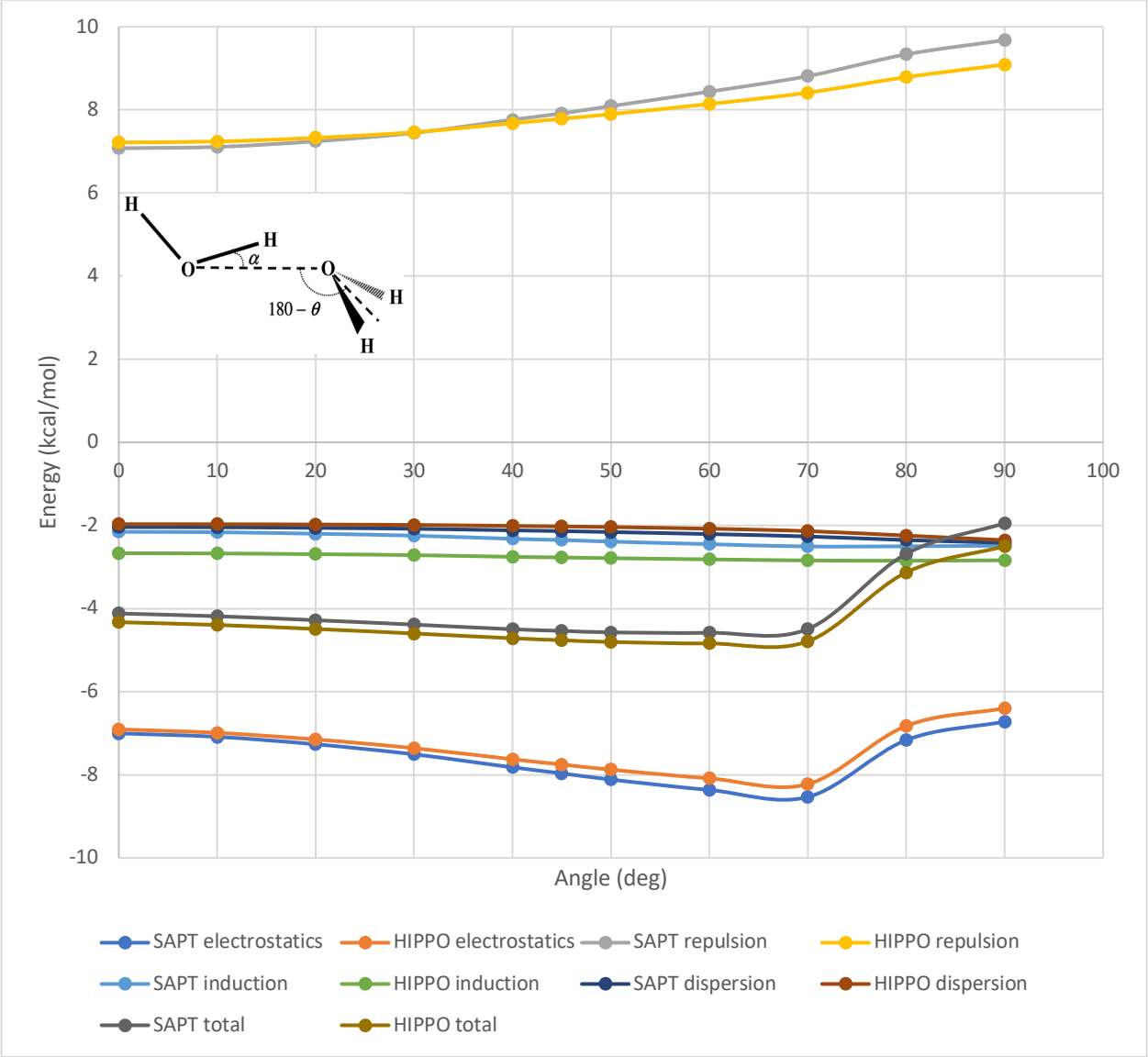


Figure 6.3. Water dimer flap angle energy components.

The water dimer is held in its equilibrium geometry and the angle θ is varied from 0 to 90°.

The data in figure 6.3 tell a clear story. The preference of the water dimer to sit at a flap angle of near 57° is the consequence of a tradeoff. The dispersion and induction components of the energy are nearly flat across this angular scan, but the electrostatics and repulsion trend in opposite directions. The electrostatics gets more attractive as the flap angle is increased. If this were the only anisotropic component of the force field, it would push the equilibrium flap angle to 70° or more. This is precisely the reason why the original AMOEBA force field reduced the quadrupole moments of both oxygen and hydrogen to artificially set the flap angle to 57°. In order to model this correctly without scaling, one must include the anisotropic effect of repulsion. We have shown in previous work, and figure 6.3 reiterates, that the Multipolar Pauli Repulsion model of HIPPO is capable of reproducing this angular dependence where typical van der Waals functions are not. This underscores point #4 of our list in the introduction. Anisotropy, and not just in the electrostatics, is crucial to correctly modeling the force field energy components in a way that prevents the need for large cancellation of errors.

The final part of the potential energy surface we assessed is a set of ten well-studied, stationary point water dimers. These dimers, illustrated in figure 6.4, represent a diversity of important water-water interactions away from the minimum energy configuration. Shown in figure 6.5 is the breakdown, by component, of the HIPPO water model against the SAPT calculations on these structures.

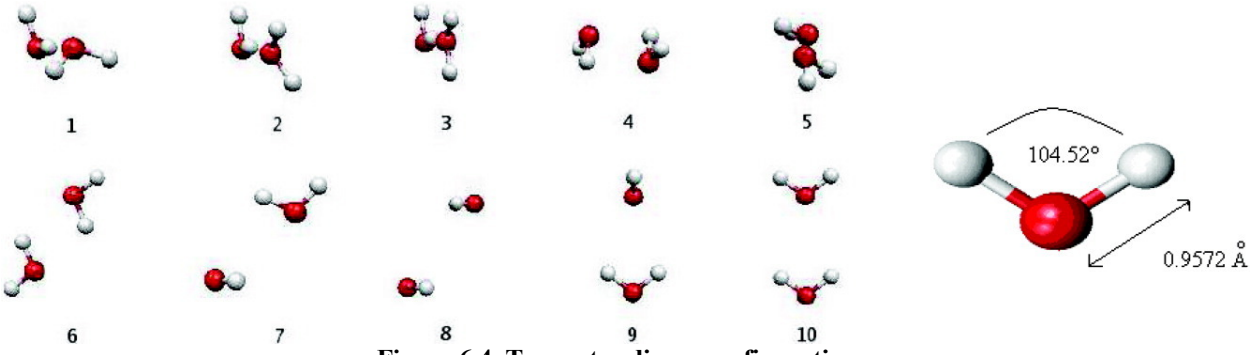


Figure 6.4. Ten water dimer configurations.

(Reproduced from reference 20)

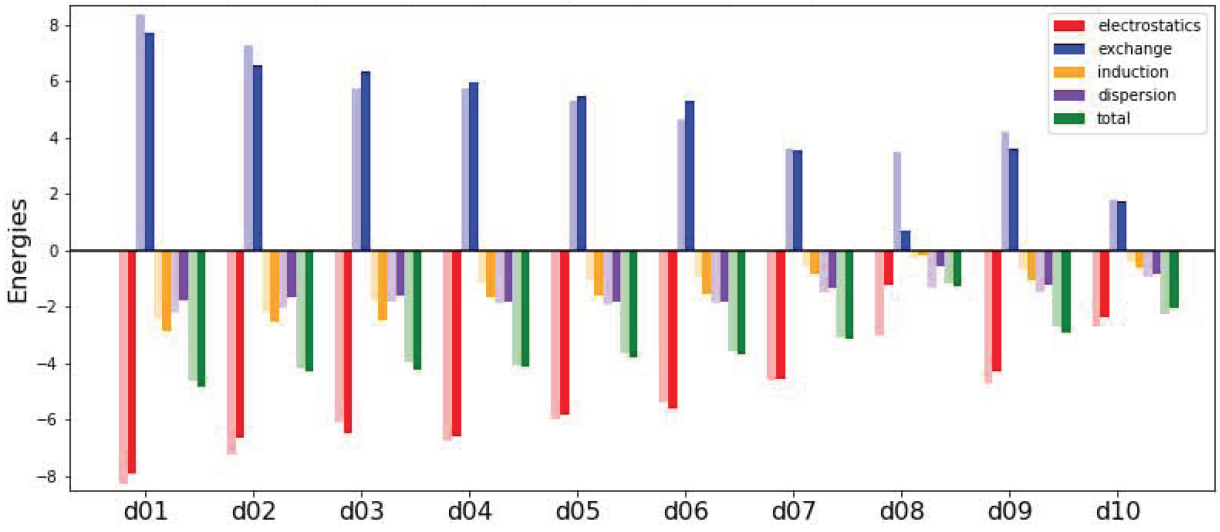


Figure 6.5 Energy component analysis of the ten water dimer configurations.

Light shaded bars represent SAPT and dark shaded bars represent HIPPO.

The results show that the HIPPO model is able to reproduce the SAPT energy components across a wide variety of intermolecular contacts. Although the only water-water data in the original fit was near the minimum energy conformation, the model is clearly able to describe equally well dimers with different contacts. Moreover, these are parameters that have been optimized for liquid phase properties. The data in figure 6.5 is showing that because we have limited the

parameter space to search, the quality of the agreement with SAPT in these dimer configurations is essentially unchanged.

In addition to the SAPT calculations on the water dimers, Piquemal and co-workers have done gold-standard CCSD(T) calculations on these structures. These total energies vary slightly from the SAPT results, so we wanted to check how HIPPO compared to these high-level calculations. The results are shown in figure 6.6.

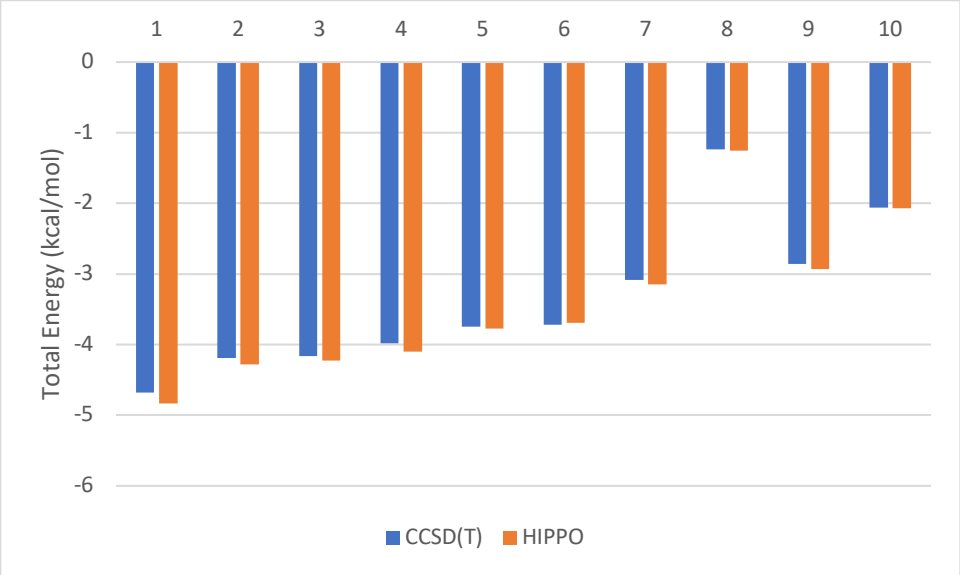


Figure 6.6. Total energy of ten water dimer structures.

Dimer numbers as illustrated in figure 6.4 are displayed across the x-axis.

These data show that although there are some small differences between SAPT and CCSD(T) total energies, the magnitudes are small. Furthermore, HIPPO does not appear to exhibit any sort of systematic bias relative to the gold-standard calculations.

The final element of gas phase data that is relevant to the HIPPO model is calculation of the many-body energy of small clusters. This has already been examined at length in Chapter 3, section 3.4.4., so I will not recapitulate that data here. It is worth noting that although the HIPPO

model has been optimized to liquid phase data, the electrostatics and polarization parameters were not included in that protocol. Therefore, all of the data in section 3.4.4. is exactly correct for the HIPPO water model. In short, the many-body energies of the HIPPO water model match the *ab initio* results for clusters of 3-8 molecules very accurately. In fact, they are more predictive than the original AMOEBA model.

6.4.3 Condensed Phase

The ultimate objective of the HIPPO water model is to use it in liquid phase simulations of biological molecules. To this end, the water model should be able to reproduce the fundamental properties of pure liquid water. Because most molecular dynamics simulations of biological molecules are performed at room temperature, we first analyzed the properties of the HIPPO water model at 298 K. Shown in table 6.3 are the relevant condensed phase properties of HIPPO water.

	HIPPO	**Experiment**
Density (g/cm^3)	0.994	0.997
Enthalpy of Vaporization (kcal/mol)	10.7	10.5
Self-Diffusion Coefficient (10^{-5} cm^2/s)	2.0	2.3
Dielectric	91	78

Table 6.3. Liquid phase properties of water at 298 K.

Clearly the model is in excellent agreement with fundamental water properties. In particular, the dielectric constant, which varies dramatically across various types of water models, is well

reproduced by the HIPPO model. We also compared the water radial distribution functions with those from experiment. These results are plotted in figure 6.7.

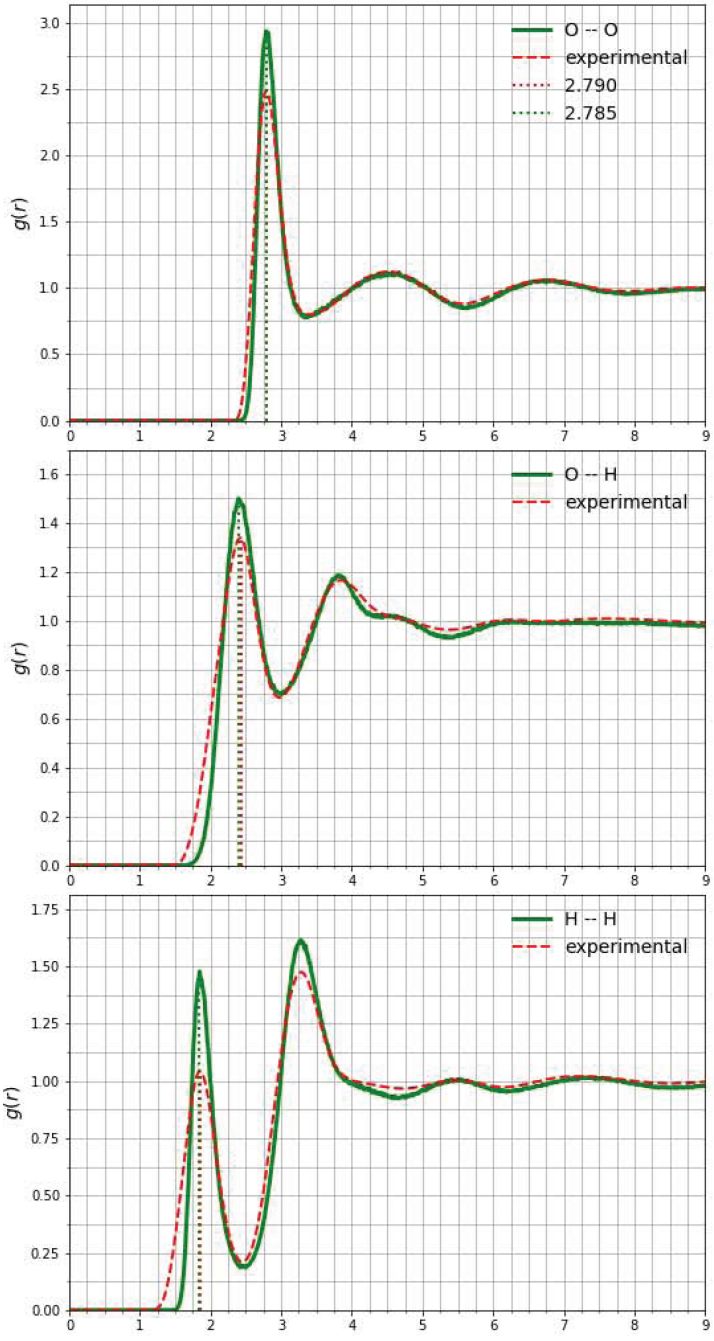


Figure 6.7. Radial distribution function of water.

Plots are shown for O-O, O-H, and H-H. X-axis units are angstroms.

The HIPPO model reproduces the radial distribution functions at room temperature as well or better than any classical force field. Most classical water models (including AMOEBA) predict the first peak of the O-O g(r) to be too far to the right and too high. The HIPPO O-O g(r), however, has a first peak whose distance and height are quite accurate. Additionally, where other models struggle to get relative heights of the first two peaks of the H-H g(r) correct, HIPPO correctly predicts the first to be shorter than the second.

Lastly, we used ForceBalance to optimize the performance of the HIPPO model across a range of temperatures. Shown in figure 6.8 is the temperature dependence of a variety of properties of the HIPPO model.

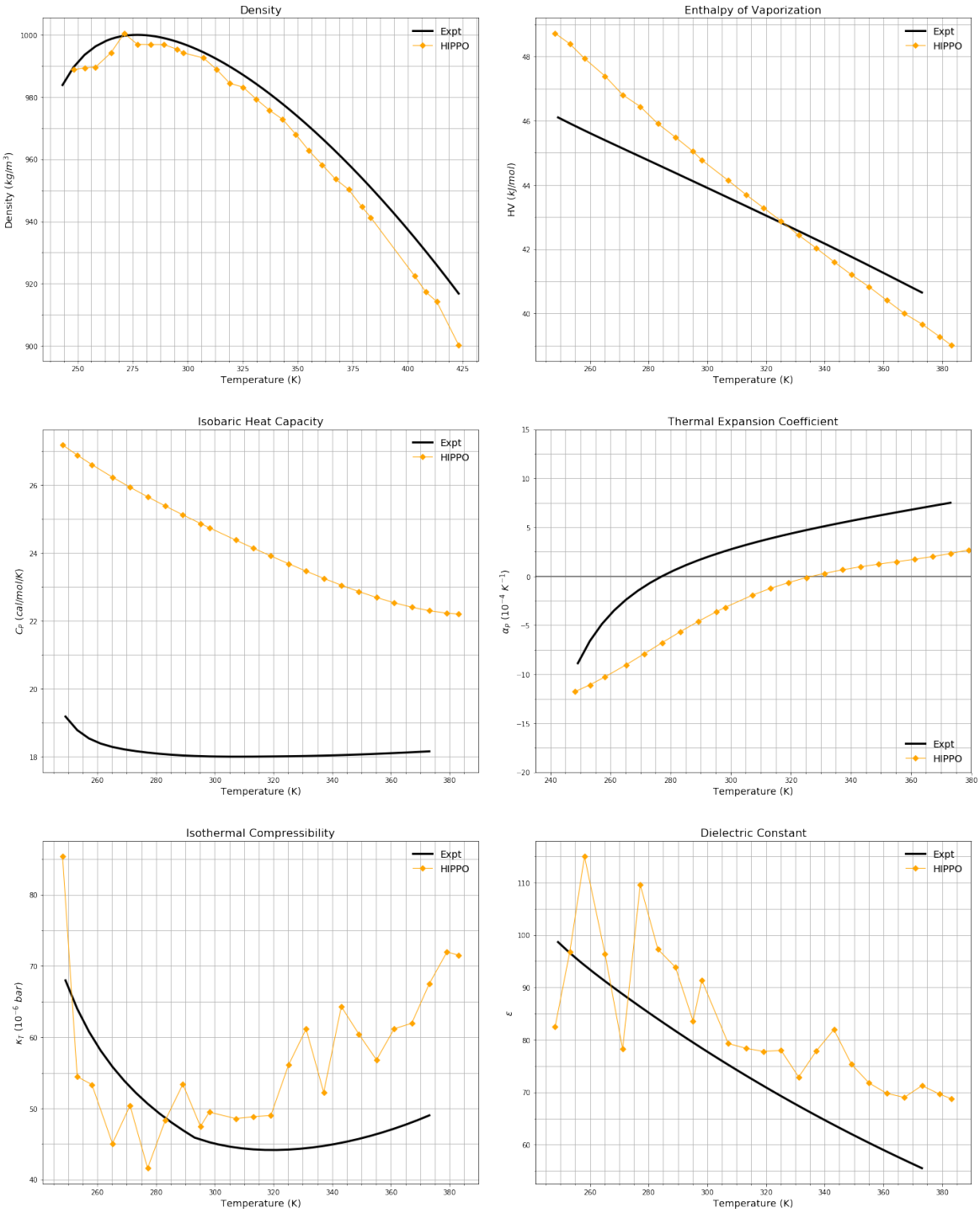


Figure 6.8. Temperature dependence of the HIPPO water model.

The density, enthalpy of vaporization, heat capacity and thermal expansion coefficient are all converged. The isothermal compressibility and dielectric constant show erratic behavior, especially at low temperature, because the simulations were not run long enough to converge.

The results show qualitative agreement with the temperature dependence of water properties. Most notably the water model produces a temperature of maximum density that is roughly in agreement with experiment. Most water models in use for biomolecular simulation today do not exhibit this fundamental behavior. The trend of the enthalpy of vaporization for the HIPPO model is correct, but the slope is too steep. This produces a heat capacity that is too high – a problem is not unique to HIPPO. In fact, Paesani and co-workers showed that any classical model of water will suffer from this problem.[25] Even the *ab initio* MB-pol model suffers from this issue. They show that trend of the heat capacity of water is directly related nuclear quantum effects. Without performing path integral (PIMD) simulations, classical force fields will always overpredict the heat capacity. In a related manner, HIPPO predicts the thermal expansion coefficient to be uniformly too low. The isothermal compressibility data requires long simulations to be well converged, so it is difficult to draw any conclusions from the data shown here. The dielectric constant appears to be near convergence for high temperatures, but not for low temperatures. For the high temperature points it appears that the trend for the model is correct. The dielectric constant is notoriously sensitive, so reproducing this within +/- 10 near room temperature is noteworthy.

Unfortunately, the data presented here is an incomplete story. The ForceBalance procedure is very computationally expensive. Therefore, the optimization is not yet complete. The results presented here are only a point along the optimization path, it is not a minimum. It is difficult to predict how much better the model will become. There is certainly room for improvement with the model's agreement with temperature dependent properties. These current

results, however, represent a lower bound on how well the model will reproduce the properties of liquid water.

6.5 Discussion and Conclusions

The model presented here represents the first step in a larger project of constructing a full, self-compatible, first-principles-based force field. The HIPPO model is based on component parts that are each individually derived and parameterized to reproduce the physics of intermolecular interactions. However, this model is not an "*ab initio"* force field. It is not entirely fit to *ab initio* data, and it does not predict the properties of liquid water. Rather it is optimized to reproduce the properties of liquid water. The reason HIPPO is not an "*ab initio"* force field comes from a series of approximations. Each one stacked on top of the next.

The first approximation of the model is the Born-Oppenheimer (BO) approximation. Under BO, the nuclei are treated as fixed classical positive point charges relative to the electrons. The Schrödinger equation for the electronic wave function is solved under the static electric potential of the "clamped" nuclei. This introduces a not insignificant error already at the first level of approximation. So-called "nuclear quantum effects" (NQEs) are known to play a large part in the behavior of liquid water. They play such a large role, in fact, that the properties of D_2O at room temperature vary significantly from those of H_2O. The magnitude of these effects are known to be generally small, but not negligible.

The second level of approximation comes with how we practically solve Schrödinger's equation for molecules. The so-called "gold standard" of quantum chemistry, CCSD(T), is not exact and thus incurs some error on top of the error already present in the BO approximation. Furthermore, while the errors of CCSD(T) are small, SAPT is not in full agreement with CCSD(T). We expect that the SAPT total energy will be close to the CCSD(T) result, but there is

an error that is incurred here as well. The level of SAPT that was used to parameterize the HIPPO model has been shown to have a mean absolute error across a range of databases of ~0.3 kcal/mol relative to CCSD(T).[26] Since the SAPT calculations used here include close contact points on the dissociation curve, the error relative to CCSD(T) is likely even larger.

Finally, HIPPO is a classical approximation of SAPT. Our previous work has shown that the HIPPO model is capable to producing agreement with SAPT to within 1 kcal/mol for each component. This incurs an error on top of the approximations already made. Furthermore, an error of <1 kcal/mol in each component does not guarantee a total error of <1 kcal/mol, due to simple propagation of errors. There are a variety of reasons for this magnitude of error. Although the HIPPO model is derived to closely mirror the SAPT energy decomposition, there are certain elements that will be impossible to capture classically. Particularly in the components of repulsion and dispersion where the fermionic nature of electrons cannot be ignored, some amount of error will always be incurred by imposing a classical function. HIPPO is no exception.

If condensed phase chemistry were not so delicate, this sequence of approximations might be more forgiving. However, the properties of liquids are highly sensitive to small changes in intermolecular energies. Changes of fractions of a kcal/mol in the repulsive wall of a molecule's dimer potential energy surface can produce changes of several percent in the density of pure liquid. Water models such as MB-pol or CC-pol do not reproduce the properties of liquid water by nailing a series of identifiable approximations. These models work by setting out a functional form flexible enough, and set of high-level QM data large enough, that they get accurate properties *via* interpolation. HIPPO is not designed to do this. Despite the algebraic complexity of the functional form, it is a concrete, fast-to-compute physical model. HIPPO has a

clear, traceable lineage back to first-principles physics, even if that lineage prevents it from being truly "*ab initio"*.

There are three reasons why this is a sensible strategy for a water model intended for biomolecular simulation. The first is empirical accuracy. Even the "*ab initio"* models of water have shortcomings. Because they do not include nuclear quantum effects, and because CCSD(T) is not exact, and because interpolation is sometimes in error, there are water properties that these models predict incorrectly. By fitting directly to condensed phase data, HIPPO ensures that accuracy is obtained where it is needed: for simulations, where thermodynamics matter. HIPPO accounts for NQEs and the other sources of error through parameterization.

Of course, parameterization against condensed phase data is exactly how standard point charge force fields have been dealing with this problem for decades. If HIPPO is doing similarly, what makes it different? The distinction is that optimization for HIPPO is occurring in a much more limited parameter space. Optimization of point charge force fields suffers from being a massively underdetermined problem. The amount of uncorrelated experimental data available for fitting is on the same order of magnitude as the number of adjustable parameters. While it may seem counterintuitive, even though HIPPO has a larger number of parameters, the space it has to search is much smaller. This is because the initial step of fitting to SAPT gives a strong first guess and a set of hard guardrails for optimization. It turns what was previously an intractable global search into a local optimization problem.

The final reason to trust the HIPPO strategy is the capacity for compatibility with a more general force field. The protocol for determining the parameters for the HIPPO water model is not unique to water. The functional form is unchanged for every other compound that will make up the biomolecular part of the HIPPO model. The initial parameters for each of those parts have

been determined in the same manner as water: fitting to a database of SAPT intermolecular interaction energy components. We cannot show conclusively yet that the HIPPO water model will be compatible in condensed phase with these other parts, but there are strong suggestions that this will be the case. Conceptually, since all parts of the model are being derived from identical sets of approximations, it is not unreasonable to think that these parts will end up being compatible with each other. Practically, the liquid optimization procedure for water appeared to change the behavior of the individual components only very slightly. Figures provided in the results section attest to the fidelity of this optimized model's continued agreement with SAPT. The fact that there is no large-scale shift in the parameters or behavior suggests that the underlying compatibility may, in fact, be preserved.

As was stated in the introduction, this model is testing a hypothesis. The hypothesis is that a model that is built off of an identifiable series of approximations can yield an energy function that is both accurate for liquid water and general enough to be transferable to the complicated mixtures that biomolecular simulations entail. This work shows the first part of this proposition is true. HIPPO represents a "natural" potential energy function for representing liquid water. Optimization is certainly required to produce satisfactory agreement with experimentally measured condensed phase properties, but it is exactly that: optimization. The function is tied closely enough to the underlying quantum mechanics that determines the true potential energy surface that the initial parameters put us in the near neighborhood of a minimum that gives excellent agreement with experiment. This is by no means guaranteed to be the case. In general, given a function and arbitrary set of initial parameters, the probability the local minimum of that neighborhood will be a satisfactory condensed phase water model is vanishingly small. This suggests that the HIPPO functional form, and particularly its density

model on which the parts of the model are based, is a natural description for intermolecular interactions.

The evidence presented in this work makes a strong case to prove this distinction as a natural energy function for water. The second part of the hypothesis, however, remains unproven. What this work does give, however, is a suggestion that this behavior may not be unique to water. The protocol for determining the starting parameter values of the water model is based on the S101 dimer database, of which water-water interactions are only a small part. The rest of the database is made up of interactions between other bioorganic compounds required to build a full biomolecular force field. The fact that this function and protocol has produced a natural model for water suggests that the same may be possible for the other molecules in the S101 database. Moreover, since interactions between water and these other molecules is a significant part of the database, it is not unreasonable to postulate that these parts my end up being compatible as well.

6.6 References

1 Berendsen, H., Grigera, J. & Straatsma, T. The missing term in effective pair potentials. *Journal of Physical Chemistry* **91**, 6269-6271 (1987).

2 Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *The Journal of chemical physics* **79**, 926-935 (1983).

3 Jorgensen, W. L., Madura, J. D. & Swenson, C. J. Optimized intermolecular potential functions for liquid hydrocarbons. *Journal of the American Chemical Society* **106**, 6638-6646 (1984).

4 Bukowski, R., Szalewicz, K., Groenenboom, G. C. & Van der Avoird, A. Predictions of the properties of water from first principles. *Science* **315**, 1249-1252 (2007).

5 Babin, V., Leforestier, C. & Paesani, F. Development of a “first principles” water potential with flexible monomers: Dimer potential energy surface, VRT spectrum, and second virial coefficient. *Journal of chemical theory and computation* **9**, 5395-5403 (2013).

6 Ren, P. & Ponder, J. W. Polarizable atomic multipole water model for molecular mechanics simulation. *The Journal of Physical Chemistry B* **107**, 5933-5947 (2003).

7 Laury, M. L., Wang, L.-P., Pande, V. S., Head-Gordon, T. & Ponder, J. W. Revised parameters for the AMOEBA polarizable atomic multipole water model. *The Journal of Physical Chemistry B* **119**, 9423-9437 (2015).
8 Rackers, J. A. *et al.* An optimized charge penetration model for use with the AMOEBA force field. *Physical Chemistry Chemical Physics* **19**, 276-291 (2017).
9 Rackers, J. A., Liu, C., Ren, P. & Ponder, J. W. A physically grounded damped dispersion model with particle mesh Ewald summation. *The Journal of chemical physics* **149**, 084115 (2018).
10 Rackers, J. A. & Ponder, J. W. Classical Pauli repulsion: An anisotropic, atomic multipole model. *The Journal of chemical physics* **150**, 084104 (2019).
11 Aviat, F. *et al.* Truncated conjugate gradient: an optimal strategy for the analytical evaluation of the many-body polarization energy and forces in molecular simulations. *Journal of chemical theory and computation* **13**, 180-190 (2016).
12 Simmonett, A. C., Pickard IV, F. C., Ponder, J. W. & Brooks, B. R. An empirical extrapolation scheme for efficient treatment of induced dipoles. *The Journal of chemical physics* **145**, 164101 (2016).
13 Simmonett, A. C., Pickard IV, F. C., Shao, Y., Cheatham III, T. E. & Brooks, B. R. Efficient treatment of induced dipoles. *The Journal of chemical physics* **143**, 074115 (2015).
14 Fanourgakis, G. S. & Xantheas, S. S. Development of transferable interaction potentials for water. V. Extension of the flexible, polarizable, Thole-type model potential (TTM3-F, v. 3.0) to describe the vibrational spectra of water clusters and liquid water. *The Journal of chemical physics* **128**, 074506 (2008).
15 Tironi, I. G., Brunne, R. M. & van Gunsteren, W. F. On the relative merits of flexible versus rigid models for use in computer simulations of molecular liquids. *Chemical physics letters* **250**, 19-24 (1996).
16 Sprik, M., Hutter, J. & Parrinello, M. Ab initio molecular dynamics simulation of liquid water: Comparison of three gradient-corrected density functionals. *The Journal of chemical physics* **105**, 1142-1152 (1996).
17 Ren, P., Wu, C. & Ponder, J. W. Polarizable Atomic Multipole-Based Molecular Mechanics for Organic Molecules. *Journal of Chemical Theory and Computation* **7**, 3143-3161, doi:10.1021/ct200304d (2011).
18 Stone, A. Distributed multipole analysis, or how to describe a molecular charge distribution. *Chemical Physics Letters* **83**, 233-239 (1981).
19 Tschumper, G. S. *et al.* Anchoring the water dimer potential energy surface with explicitly correlated computations and focal point analyses. *The Journal of chemical physics* **116**, 690-701 (2002).
20 Reinhardt, P. & Piquemal, J. P. New intermolecular benchmark calculations on the water dimer: SAPT and supermolecular post-Hartree–Fock approaches. *International Journal of Quantum Chemistry* **109**, 3259-3267 (2009).
21 Wang, L.-P., Chen, J. & Van Voorhis, T. Systematic parametrization of polarizable force fields from quantum chemistry data. *Journal of chemical theory and computation* **9**, 452-460 (2012).
22 Rackers, J. A. *et al.* Tinker 8: software tools for molecular design. *Journal of chemical theory and computation* **14**, 5273-5289 (2018).

23 Eastman, P. *et al.* OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS computational biology* **13**, e1005659 (2017).
24 Harger, M. *et al.* Tinker-OpenMM: Absolute and relative alchemical free energies using AMOEBA on GPUs. *Journal of Computational Chemistry* (2017).
25 Reddy, S. K. *et al.* On the accuracy of the MB-pol many-body potential for water: Interaction energies, vibrational frequencies, and classical thermodynamic and dynamical properties from clusters to liquid water and ice. *The Journal of chemical physics* **145**, 194504 (2016).
26 Parker, T. M., Burns, L. A., Parrish, R. M., Ryno, A. G. & Sherrill, C. D. Levels of symmetry adapted perturbation theory (SAPT). I. Efficiency and performance for interaction energies. *The Journal of chemical physics* **140**, 094106 (2014).

Appendix A

Supplementary Information for Chapter 2

1. Damping Functions for Higher-order Damping

A. One-site damping functions

$$\lambda_1 = 1 - e^{-\alpha r}$$

$$\lambda_3 = 1 - (1 + \alpha r)e^{-\alpha r}$$

$$\lambda_5 = 1 - \left(1 + \alpha r + \left(\frac{1}{3}\right)(\alpha r)^2\right)e^{-\alpha r}$$

B. Two site damping functions, $\alpha_i \neq \alpha_k$

$$A = \frac{\alpha_k^2}{\alpha_k^2 - \alpha_i^2} \quad , \qquad B = \frac{\alpha_i^2}{\alpha_i^2 - \alpha_k^2}$$

$$\lambda_1 = 1 - Ae^{-\alpha_i r} - Be^{-\alpha_k r}$$

$$\lambda_3 = 1 - (1 + \alpha_i r)Ae^{-\alpha_i r} - (1 + \alpha_k r)Be^{-\alpha_k r}$$

$$\lambda_5 = 1 - \left(1 + \alpha_i r + \left(\frac{1}{3}\right)(\alpha_i r)^2\right)Ae^{-\alpha_i r} - \left(1 + \alpha_k r + \left(\frac{1}{3}\right)(\alpha_k r)^2\right)Be^{-\alpha_k r}$$

$$\lambda_7 = 1 - \left(1 + \alpha_i r + \left(\frac{2}{5}\right)(\alpha_i r)^2 + \left(\frac{1}{15}\right)(\alpha_i r)^3\right)Ae^{-\alpha_i r} - \left(1 + \alpha_k r + \left(\frac{2}{5}\right)(\alpha_k r)^2 + \left(\frac{1}{15}\right)(\alpha_k r)^3\right)Be^{-\alpha_k r}$$

$$\lambda_9 = 1 - \left(1 + \alpha_i r + \left(\frac{3}{7}\right)(\alpha_i r)^2 + \left(\frac{2}{21}\right)(\alpha_i r)^3 + \left(\frac{1}{105}\right)(\alpha_i r)^4\right)Ae^{-\alpha_i r} - \left(1 + \alpha_k r + \left(\frac{3}{7}\right)(\alpha_k r)^2 + \left(\frac{2}{21}\right)(\alpha_k r)^3 + \left(\frac{1}{105}\right)(\alpha_k r)^4\right)Be^{-\alpha_k r}$$

C. Two site damping functions, $\alpha_i = \alpha_k$

$$\lambda_1 = 1 - \left(1 + \left(\frac{1}{2}\right)\alpha r\right) e^{-\alpha r}$$

$$\lambda_3 = 1 - \left(1 + \alpha r + \left(\frac{1}{2}\right)(\alpha_i r)^2\right) e^{-\alpha r}$$

$$\lambda_5 = 1 - \left(1 + \alpha r + \left(\frac{1}{2}\right)(\alpha_i r)^2 + \left(\frac{1}{6}\right)(\alpha_i r)^3\right) e^{-\alpha r}$$

$$\lambda_7 = 1 - \left(1 + \alpha r + \left(\frac{1}{2}\right)(\alpha_i r)^2 + \left(\frac{1}{6}\right)(\alpha_i r)^3 + \left(\frac{1}{30}\right)(\alpha_i r)^4\right) e^{-\alpha r}$$

$$\lambda_9 = 1 - \left(1 + \alpha r + \left(\frac{1}{2}\right)(\alpha_i r)^2 + \left(\frac{1}{6}\right)(\alpha_i r)^3 + \left(\frac{4}{105}\right)(\alpha_i r)^4 + \left(\frac{1}{210}\right)(\alpha_i r)^5\right) e^{-\alpha r}$$

2. S101x7 SAPT Electrostatics Fit Statistics

Multipole Only

RMSE (Root Mean Square Error)	0.7 – 0.8	24.16611265
	0.9 – 1.1	4.350677453
	Total	13.43047692
MSE (Mean Signed Error)	0.7 – 0.8	19.15611117
	0.9 – 1.1	3.160407234
	Total	7.730608359
MUE (Mean Unsigned Error)	0.7 – 0.8	19.15611117
	0.9 – 1.1	3.160407234
	Total	7.730608359
Mean Percent Error	0.7 – 0.8	0.692902042
	0.9 – 1.1	1.330447394
	Total	57.41457894
Mean Absolute Percent Error	0.7 – 0.8	0.692902042
	0.9 – 1.1	1.330447394
	Total	57.41457894

Model 1 – Charge-charge – Element-based Parameters

RMSE (Root Mean Square Error)	0.7 – 0.8	3.530462178
	0.9 – 1.1	1.079421339
	Total	2.096053384
MSE (Mean Signed Error)	0.7 – 0.8	0.070609574
	0.9 – 1.1	0.443082553
	Total	0.336661702
MUE	0.7 – 0.8	2.483826596

(Mean Unsigned Error)	0.9 – 1.1	0.699765532
	Total	1.209497264
	0.7 – 0.8	0.014649782
Mean Percent Error	0.9 – 1.1	0.083136474
	Total	3.178442377
	0.7 – 0.8	0.100161271
Mean Absolute Percent Error	0.9 – 1.1	0.247685764
	Total	10.27679545

Model 2 – Charge-charge – Element-based Parameters

	0.7 – 0.8	3.416100288
RMSE (Root Mean Square Error)	0.9 – 1.1	1.196352255
	Total	2.087232479
	0.7 – 0.8	0.078644681
MSE (Mean Signed Error)	0.9 – 1.1	-0.51620766
	Total	-0.346249848
	0.7 – 0.8	2.524932979
MUE (Mean Unsigned Error)	0.9 – 1.1	0.885186383
	Total	1.35368541
	0.7 – 0.8	-0.024265813
Mean Percent Error	0.9 – 1.1	-0.456742535
	Total	-16.65888785
	0.7 – 0.8	0.110418585
Mean Absolute Percent Error	0.9 – 1.1	0.537220758
	Total	20.76386398

Model 3 – Charge-charge – Element-based Parameters

	0.7 – 0.8	8.245733619
RMSE (Root Mean Square Error)	0.9 – 1.1	1.070226765
	Total	4.499383648
	0.7 – 0.8	0.988135106
MSE (Mean Signed Error)	0.9 – 1.1	0.034202979
	Total	0.306755015
	0.7 – 0.8	3.662237234
MUE (Mean Unsigned Error)	0.9 – 1.1	0.671327234
	Total	1.525872948
	0.7 – 0.8	0.036913509
Mean Percent Error	0.9 – 1.1	0.046321109
	Total	2.181661167
	0.7 – 0.8	0.166713556
Mean Absolute Percent Error	0.9 – 1.1	0.39171247
	Total	16.37135331

Model 1 – Charge-charge – Class-based Parameters

RMSE (Root Mean Square Error)	0.7 – 0.8	2.746166031
	0.9 – 1.1	0.989597309
	Total	1.689436504
MSE (Mean Signed Error)	0.7 – 0.8	0.087677128
	0.9 – 1.1	0.396219574
	Total	0.30806459
MUE (Mean Unsigned Error)	0.7 – 0.8	1.970433511
	0.9 – 1.1	0.633159574
	Total	1.015237842
Mean Percent Error	0.7 – 0.8	0.00351084
	0.9 – 1.1	0.034173836
	Total	1.270649006
Mean Absolute Percent Error	0.7 – 0.8	0.081690656
	0.9 – 1.1	0.225204723
	Total	9.210035191

Model 2 – Charge-charge – Class-based Parameters

RMSE (Root Mean Square Error)	0.7 – 0.8	3.038892269
	0.9 – 1.1	1.043989404
	Total	1.848524579
MSE (Mean Signed Error)	0.7 – 0.8	0.03548617
	0.9 – 1.1	-0.415489574
	Total	-0.286639362
MUE (Mean Unsigned Error)	0.7 – 0.8	2.32576383
	0.9 – 1.1	0.759034255
	Total	1.206671277
Mean Percent Error	0.7 – 0.8	-0.024980534
	0.9 – 1.1	-0.379144054
	Total	-13.89772384
Mean Absolute Percent Error	0.7 – 0.8	0.104954218
	0.9 – 1.1	0.463356585
	Total	18.04779542

Model 1 – Higher-order – Class-based Parameters

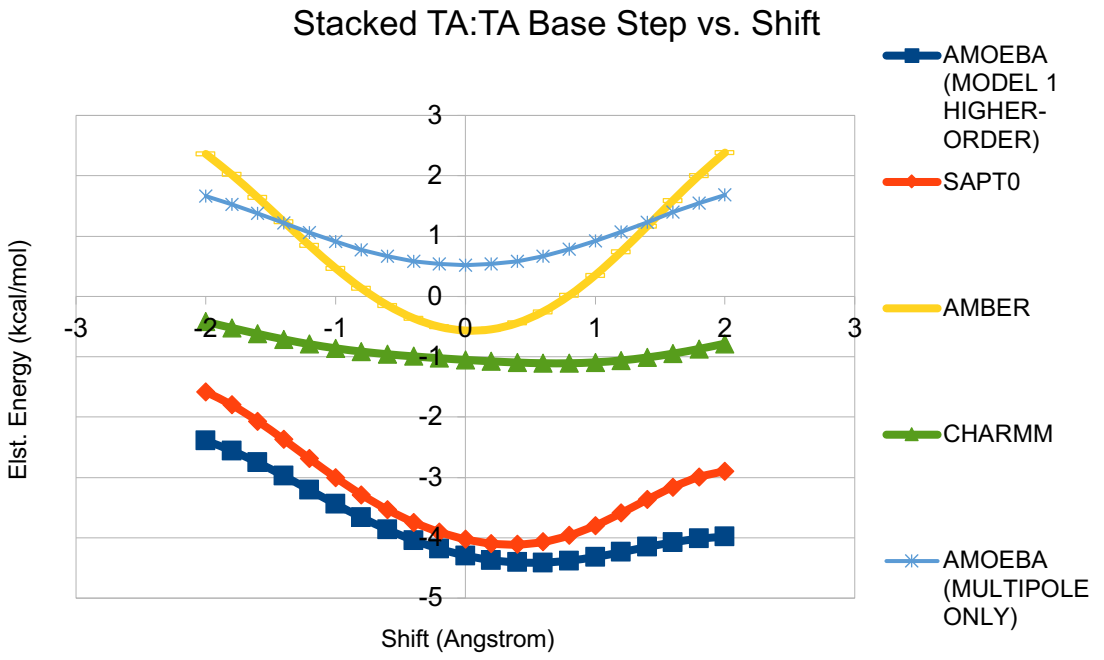
RMSE (Root Mean Square Error)	0.7 – 0.8	2.274214573
	0.9 – 1.1	0.566506743
	Total	1.306508613
MSE (Mean Signed Error)	0.7 – 0.8	0.047035638
	0.9 – 1.1	0.013938085
	Total	0.023394529
MUE (Mean Unsigned Error)	0.7 – 0.8	1.529532447
	0.9 – 1.1	0.376271277

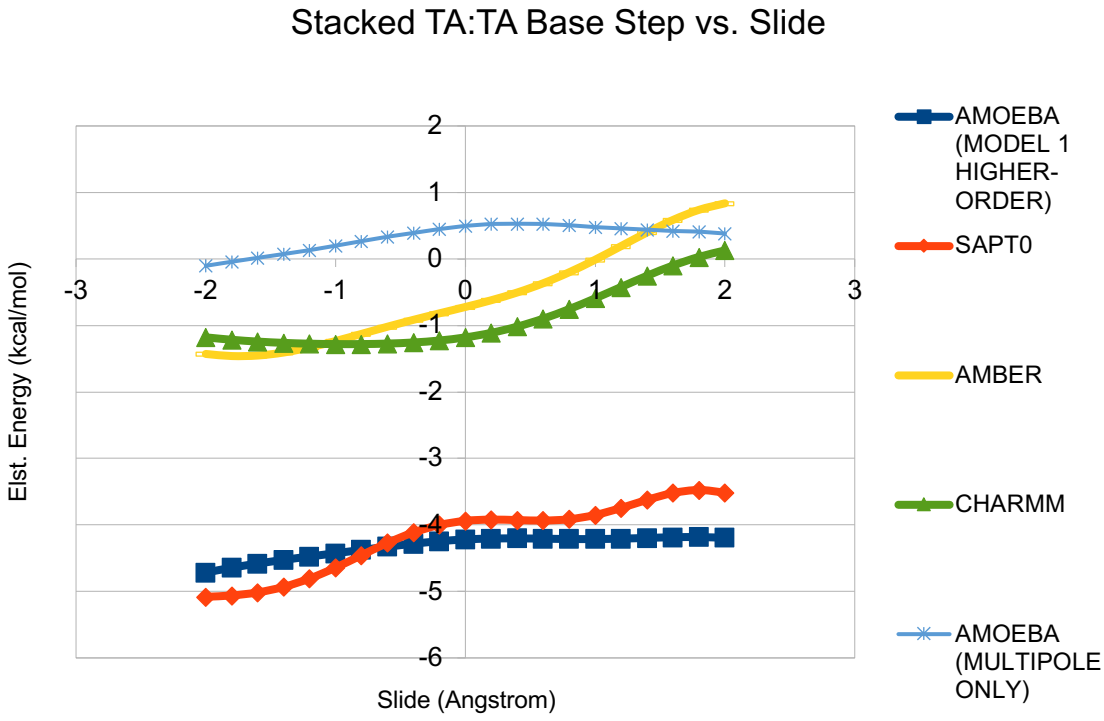
	Total	0.705774468
Mean Percent Error	0.7 – 0.8	-0.003770317
	0.9 – 1.1	-0.044059741
	Total	-1.627423843
Mean Absolute Percent Error	0.7 – 0.8	0.061419216
	0.9 – 1.1	0.18531193
	Total	7.495700609

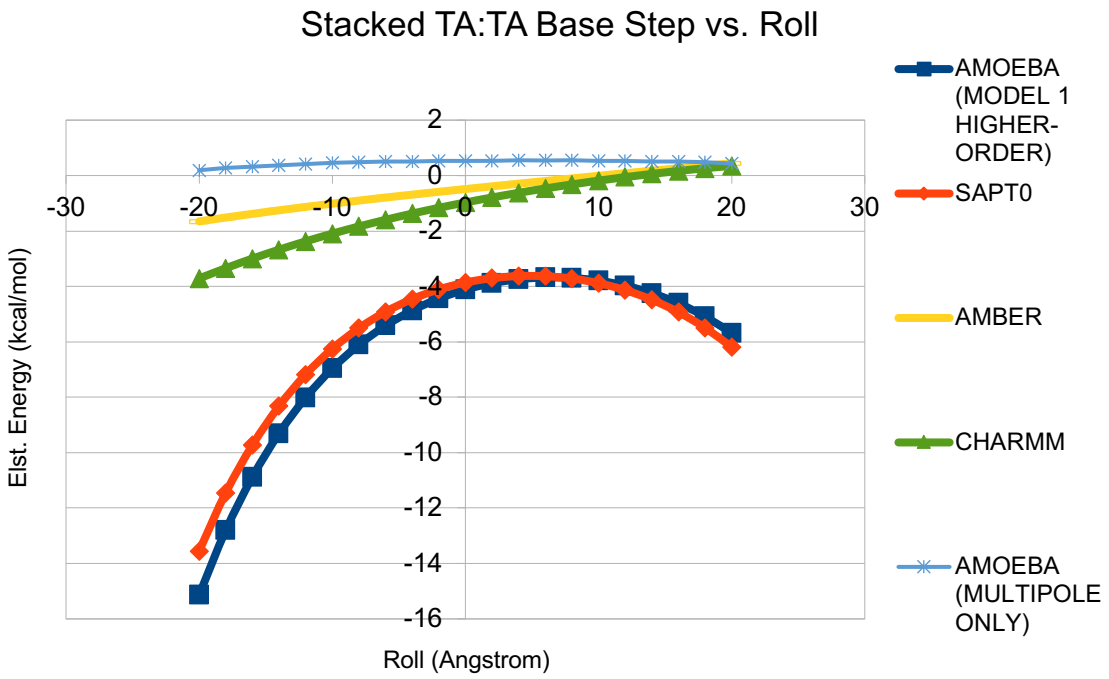
Model 2 – Higher-order – Class-based Parameters

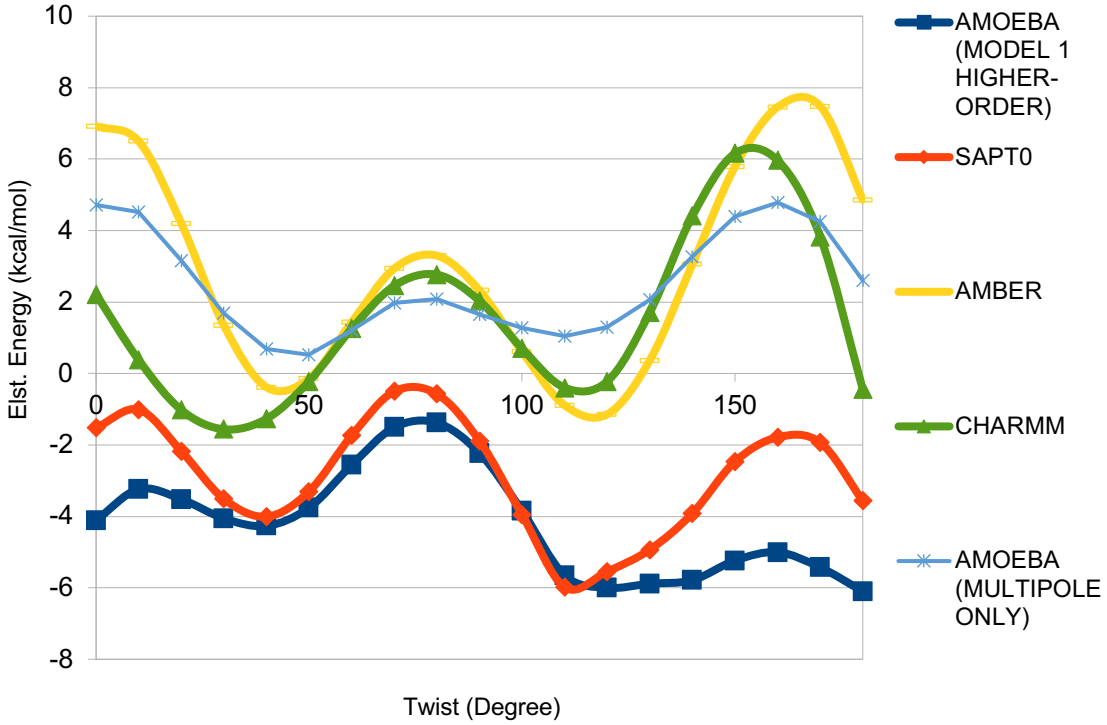
RMSE (Root Mean Square Error)	0.7 – 0.8	2.639920385
	0.9 – 1.1	0.661324567
	Total	1.517757283
MSE (Mean Signed Error)	0.7 – 0.8	0.009317553
	0.9 – 1.1	-0.115076596
	Total	-0.07953541
MUE (Mean Unsigned Error)	0.7 – 0.8	1.817476064
	0.9 – 1.1	0.466214043
	Total	0.852288906
Mean Percent Error	0.7 – 0.8	-0.013435382
	0.9 – 1.1	-0.151155074
	Total	-5.590329517
Mean Absolute Percent Error	0.7 – 0.8	0.074795428
	0.9 – 1.1	0.248929303
	Total	9.958838368

3. TA:TA Step Plots for Nucleic Acid Base Structural Parameters


Stacked TA:TA Base Step vs. Shift
Elst. Energy (kcal/mol)
Shift (Angstrom)
AMOEBA (MODEL 1 HIGHER-ORDER)
SAPT0
AMBER
CHARMM
AMOEBA (MULTIPOLE ONLY)


Stacked TA:TA Base Step vs. Slide
Elst. Energy (kcal/mol)
Slide (Angstrom)
AMOEBA (MODEL 1 HIGHER-ORDER)
SAPT0
AMBER
CHARMM
AMOEBA (MULTIPOLE ONLY)


Stacked TA:TA Base Step vs. Roll
Elst. Energy (kcal/mol)
2
0
-2
-4
-6
-8
-10
-12
-14
-16
-30
-20
-10
0
10
20
30
Roll (Angstrom)
AMOEBA (MODEL 1 HIGHER-ORDER)
SAPT0
AMBER
CHARMM
AMOEBA (MULTIPOLE ONLY)


Stacked TA:TA Base Step vs. Twist
Elst. Energy (kcal/mol)
10
8
6
4
2
0
-2
-4
-6
-8
0
50
100
150
Twist (Degree)
AMOEBA (MODEL 1 HIGHER-ORDER)
SAPT0
AMBER
CHARMM
AMOEBA (MULTIPOLE ONLY)

Appendix B

Supplementary Information for Chapter 4

As indicated in the text, the full dispersion interaction between two real atoms also includes higher-order components that give rise to $1/r^8$, $1/r^{10}$, *etc*. terms. These terms come from instantaneous higher-order multipole interactions between atoms. Similarly to the $1/r^6$ term, these can be derived from a simple Drude oscillator model of atomic polarizability. As the derivation of the origin of these terms is not readily available in the literature, we present here a derivation that continues the series started in the text.

Dipole-Quadrupole Dispersion

The derivation of the dipole-quadrupole dispersion energy starts from equation 4.5 in the text, where, instead of the dipole-dipole energy, the dipole-quadrupole interaction energy now enters into the Schrodinger equation,

$$\frac{1}{M}\frac{\partial^2\Psi}{\partial z_i^2}+\frac{1}{M}\frac{\partial^2\Psi}{\partial z_j^2}+\frac{2}{\hbar^2}\left(E-\frac{1}{2}kz_i^2-\frac{1}{2}kz_j^2-U_{dipole-quadrupole}\right)\Psi=0 \quad (1)$$

For a Drude oscillator dipole, interacting with a linear, Drude oscillator quadrupole, the energy of the interaction is given by:

$$U_{dipole-quadrupole}=\nabla\nabla\nabla U_{chg-chg}=-\frac{3\left(\mu_i\Theta_j+\mu_j\Theta_i\right)}{r^4} \quad (2)$$

If we assume the magnitude of the dipole and quadrupole moments on i and j to be identical, combining equations 1 and 2 yields,

$$\frac{1}{M}\frac{\partial^2\Psi}{\partial z_i^2}+\frac{1}{M}\frac{\partial^2\Psi}{\partial z_j^2}+\frac{2}{\hbar^2}\left(E-\frac{1}{2}kz_i^2-\frac{1}{2}kz_j^2-\frac{6\mu\Theta}{r^4}\right)\Psi=0 \quad (3)$$

We now follow a similar transformation of variables as in the text and define,

(4) $$\lambda_1 = \frac{z_i + z_j}{\sqrt{2}}, \qquad \lambda_2 = \frac{z_i - z_j}{\sqrt{2}}$$

and rewrite equation 4.8 as,

(5) $$\frac{1}{M}\frac{\partial^2 \Psi}{\partial z_i^2} + \frac{1}{M}\frac{\partial^2 \Psi}{\partial z_j^2} + \frac{2}{\hbar^2}\left(E - \frac{1}{2}k_1\lambda_i^2 - \frac{1}{2}k_2\lambda_j^2\right)\Psi = 0$$

where,

(6) $$k_1 = k + \frac{6Q^2 z_j}{r^4}, \qquad k_2 = k - \frac{6Q^2 z_j}{r^4}$$

Equation 5 is again a transformed version of the independent harmonic oscillator problem. It can be solved in the same manner giving,

(7) $$E(r) = \frac{1}{2}\hbar\left(\omega_1 + \omega_2\right),$$

where,

(8) $$\omega_1 = \sqrt{\frac{k_1}{M}} = \omega_0\sqrt{1 - \frac{6Qz_j}{r^4 k}}, \qquad \omega_2 = \sqrt{\frac{k_2}{M}} = \omega_0\sqrt{1 + \frac{6Qz_j}{r^4 k}}.$$

Applying the binomial expansion,

(9) $$\sqrt{1+x} = 1 + \tfrac{1}{2}x - \tfrac{1}{8}x + \ldots$$

the total energy becomes,

(10) $$E(r) = \hbar\omega_0 - \frac{36\hbar\omega_0}{8r^8 k^2} + \ldots.$$

We then subtract the energy of infinitely separated atoms. This gives the dipole-quadrupole term of the dispersion potential energy,

(11) $$E(r) = -\frac{36\hbar\omega_0}{8r^8 k^2} + \ldots.$$

It should be noted here that the spring constants, k, and frequencies, ω, are not the same as those for the dipole-dipole interaction. Therefore, just as with the dipole-dipole term, parameters are introduced to give the dipole-quadrupole dispersion model energy:

$$U^{dispersion} = -\frac{C_8^i C_8^j}{r^8} \quad . \tag{12}$$

Quadrupole-Quadrupole Dispersion

At the risk of repetition, the quadrupole-quadrupole derivation follows almost exactly the formulation above. The Schrodinger equation now reads,

$$\frac{1}{M}\frac{\partial^2 \Psi}{\partial z_i^2} + \frac{1}{M}\frac{\partial^2 \Psi}{\partial z_j^2} + \frac{2}{\hbar^2}\left(E - \frac{1}{2}kz_i^2 - \frac{1}{2}kz_j^2 - U_{dipole-quadrupole}\right)\Psi = 0 \quad . \tag{13}$$

In this case we have two linear, Drude oscillator quadrupoles, interacting with each other. The energy of the interaction is given by:

$$U_{quadrupole-quadrupole} = \nabla\nabla\nabla\nabla U_{chg-chg} = -\frac{6\Theta_i\Theta_j}{r^5} \quad . \tag{14}$$

Combining equations 13 and 14 yields,

$$\frac{1}{M}\frac{\partial^2 \Psi}{\partial z_i^2} + \frac{1}{M}\frac{\partial^2 \Psi}{\partial z_j^2} + \frac{2}{\hbar^2}\left(E - \frac{1}{2}kz_i^2 - \frac{1}{2}kz_j^2 - \frac{6\Theta_i\Theta_j}{r^5}\right)\Psi = 0 \quad . \tag{15}$$

The transformation of variables is identical to the dipole-quadrupole case,

$$\lambda_1 = \frac{z_i + z_j}{\sqrt{2}}, \qquad \lambda_2 = \frac{z_i - z_j}{\sqrt{2}} \tag{16}$$

which gives,

$$\frac{1}{M}\frac{\partial^2 \Psi}{\partial z_i^2} + \frac{1}{M}\frac{\partial^2 \Psi}{\partial z_j^2} + \frac{2}{\hbar^2}\left(E - \frac{1}{2}k_1\lambda_i^2 - \frac{1}{2}k_2\lambda_j^2\right)\Psi = 0 \tag{17}$$

where,

(18) $$k_1 = k + \frac{6Q^2 z_i z_j}{r^5}, \qquad k_2 = k - \frac{6Q^2 z_i z_j}{r^5}$$

Equation 17 again gives us the independent harmonic oscillator problem with the solution,

(19) $$E(r) = \frac{1}{2}\hbar\left(\omega_1 + \omega_2\right),$$

where,

(20) $$\omega_1 = \sqrt{\frac{k_1}{M}} = \omega_0\sqrt{1 - \frac{6Q^2 z_i z_j}{r^5 k}}, \qquad \omega_2 = \sqrt{\frac{k_2}{M}} = \omega_0\sqrt{1 + \frac{6Q^2 z_i z_j}{r^5 k}}.$$

Applying the binomial expansion as before, the total energy becomes,

(21) $$E(r) = \hbar\omega_0 - \frac{36Q^4 z_i^2 z_j^2 \hbar\omega_0}{8r^{10}k^2} + \ldots.$$

Subtracting the energy of infinitely separated atoms gives the quadrupole-quadrupole term of the dispersion potential energy,

(22) $$E(r) = -\frac{36Q^4 z_i^2 z_j^2 \hbar\omega_0}{8k^2 r^{10}} + \ldots.$$

Again, the spring constants, k, and frequencies, ω, are placeholders specific to this quadrupole-quadrupole interaction. To generalize, parameters are introduced to give,

(23) $$U^{dispersion}_{(quadrupole-quadrupole)} = -\frac{C^i_{10} C^j_{10}}{r^{10}},$$

the quadrupole-quadrupole dispersion energy.

As is apparent from the above sequence of derivations, this pattern of even power dispersion coefficients continues indefinitely for as many higher-order multipole moments as one wishes to include. (We should note that at $1/r^{10}$ terms and higher, multiple multipole interactions start to be included in terms. The $1/r^{10}$ term, for example, involves a dipole-octopole component

as well as quadrupole-quadrupole.) This pattern allows us to extrapolate to the full, parameterized dispersion expansion:

$$U_{ij}^{dispersion} = -\sum_{k\geq 6}^{even} \frac{C_k^i C_k^j}{r^k} \quad (24)$$

.

Appendix C

Electrostatic Energy of the HIPPO water model.

Core-Core:

$$U_{core-core} = Z_i T_{ij} Z_j$$
$$T_{ij} = \frac{1}{R}$$

Core-Density:

$$U_{core-density} = Z_i \boldsymbol{T}^*_{\boldsymbol{ij}} \vec{M}_j$$
$$\boldsymbol{T}^*_{\boldsymbol{ij}} = [1 \quad \nabla \quad \nabla^2] T^*$$
$$T^* = \frac{1}{R} f_1^{damp}$$
$$\nabla T^* = -f_3^{damp} \frac{R_\alpha}{R^3}$$
$$\nabla^2 T^* = f_5^{damp} \frac{3R_\alpha R_\beta}{R^5} - f_3^{damp} \frac{\delta_{\alpha\beta}}{R^3}$$
$$f_1^{damp} = 1 - \left(1 + \frac{1}{2}\alpha_j R\right) e^{-\alpha_j R}$$
$$f_3^{damp} = 1 - \left(1 + \alpha_j R + \frac{1}{2}(\alpha_j R)^2\right) e^{-\alpha_j R}$$
$$f_5^{damp} = 1 - \left(1 + \alpha_j R + \frac{1}{2}(\alpha_j R)^2 + \frac{1}{6}(\alpha_j R)^3\right) e^{-\alpha_j R}$$

Density-Density:

$$U_{density-density} = \vec{M}_i \boldsymbol{T}^{\boldsymbol{overlap}}_{\boldsymbol{ij}} \vec{M}_j$$
$$\boldsymbol{T}^{\boldsymbol{overlap}}_{\boldsymbol{ij}} = \begin{bmatrix} 1 & \nabla & \nabla^2 \\ \nabla & \nabla^2 & \nabla^3 \\ \nabla^2 & \nabla^3 & \nabla^4 \end{bmatrix} (T^{overlap})$$
$$T^{overlap} = \frac{1}{R} f_1^{overlap}$$
$$\nabla T^{overlap} = -f_3^{overlap} \frac{R_\alpha}{R^3}$$
$$\nabla^2 T^{overlap} = f_5^{overlap} \frac{3R_\alpha R_\beta}{R^5} - f_3^{overlap} \frac{\delta_{\alpha\beta}}{R^3}$$

$$\nabla^3 T^{overlap} = -f_7^{overlap}\frac{15R_\alpha R_\beta R_\gamma}{R^7} + f_5^{overlap}\frac{3\left(R_\alpha\delta_{\beta\gamma}+R_\beta\delta_{\alpha\gamma}+R_\gamma\delta_{\alpha\beta}\right)}{R^5}$$

$$\nabla^4 T^{overlap}$$
$$= f_9^{overlap}\frac{105R_\alpha R_\beta R_\gamma R_\eta}{R^9}$$
$$- f_7^{overlap}\frac{15\left(R_\alpha R_\beta\delta_{\gamma\eta}+R_\alpha R_\gamma\delta_{\beta\eta}+R_\alpha R_\eta\delta_{\beta\gamma}+R_\beta R_\gamma\delta_{\alpha\eta}+R_\beta R_\eta\delta_{\alpha\gamma}+R_\gamma R_\eta\delta_{\alpha\beta}\right)}{R^7}$$
$$+ f_5^{overlap}\frac{3\left(\delta_{\alpha\beta}\delta_{\gamma\eta}+\delta_{\alpha\gamma}\delta_{\beta\eta}+\delta_{\alpha\eta}\delta_{\beta\gamma}\right)}{R^5}$$

$$f_1^{overlap} = \begin{cases} 1-\left(1+\frac{11}{16}\alpha R+\frac{3}{16}(\alpha R)^2+\frac{1}{48}(\alpha R)^3\right)e^{-\alpha R}, & \alpha_i=\alpha_j \\ 1-A^2\left(1+2B+\frac{\alpha_i}{2}R\right)e^{-\alpha_i R}-B^2\left(1+2A+\frac{\alpha_j}{2}R\right)e^{-\alpha_j R}, & \alpha_i\neq\alpha_j \end{cases}$$

$$f_3^{overlap} = \begin{cases} 1-\left(1+\alpha R+\frac{1}{2}(\alpha R)^2+\frac{7}{48}(\alpha R)^3+\frac{1}{48}(\alpha R)^4\right)e^{-\alpha R}, & \alpha_i=\alpha_j \\ 1-A^2\left(1+\alpha_i R+\frac{1}{2}(\alpha_i R)^2\right)e^{-\alpha_i R}-B^2\left(1+\alpha_j R+\frac{1}{2}\left(\alpha_j R\right)^2\right)e^{-\alpha_j R}- & \\ \quad 2A^2B(1+\alpha_i R)e^{-\alpha_i R}-2B^2A\left(1+\alpha_j R\right)e^{-\alpha_j R}, & \alpha_i\neq\alpha_j \end{cases}$$

$$f_5^{overlap} = \begin{cases} 1-\left(1+\alpha R+\frac{1}{2}(\alpha R)^2+\frac{1}{6}(\alpha R)^3+\frac{1}{24}(\alpha R)^4+\frac{1}{144}(\alpha R)^5\right)e^{-\alpha R}, & \alpha_i=\alpha_j \\ 1-A^2\left(1+\alpha_i R+\frac{1}{2}(\alpha_i R)^2+\frac{1}{6}(\alpha_i R)^3\right)e^{-\alpha_i R}- & \\ \quad B^2\left(1+\alpha_j R+\frac{1}{2}\left(\alpha_j R\right)^2+\frac{1}{6}\left(\alpha_j R\right)^3\right)e^{-\alpha_j R}- & \\ \quad 2A^2B\left(1+\alpha_i R+\frac{1}{3}(\alpha_i R)^2\right)e^{-\alpha_i R}- & \\ \quad 2B^2A\left(1+\alpha_j R+\frac{1}{3}\left(\alpha_j R\right)^2\right)e^{-\alpha_j R}, & \alpha_i\neq\alpha_j \end{cases}$$

$$f_7^{overlap} = \begin{cases} 1-\left(1+\alpha R+\frac{1}{2}(\alpha R)^2+\frac{1}{6}(\alpha R)^3+\frac{1}{24}(\alpha R)^4+\frac{1}{120}(\alpha R)^5+\frac{1}{720}(\alpha R)^6\right)e^{-\alpha R}, & \alpha_i=\alpha_j \\ 1-A^2\left(1+\alpha_i R+\frac{1}{2}(\alpha_i R)^2+\frac{1}{6}(\alpha_i R)^3+\frac{1}{30}(\alpha_i R)^4\right)e^{-\alpha_i R}- & \\ \quad B^2\left(1+\alpha_j R+\frac{1}{2}\left(\alpha_j R\right)^2+\frac{1}{6}\left(\alpha_j R\right)^3+\frac{1}{30}\left(\alpha_j R\right)^4\right)e^{-\alpha_j R}- & \\ \quad 2A^2B\left(1+\alpha_i R+\frac{2}{5}(\alpha_i R)^2+\frac{1}{15}(\alpha_i R)^3\right)e^{-\alpha_i R}- & \\ \quad 2B^2A\left(1+\alpha_j R+\frac{2}{5}\left(\alpha_j R\right)^2+\frac{1}{15}\left(\alpha_j R\right)^3\right)e^{-\alpha_j R}, & \alpha_i\neq\alpha_j \end{cases}$$

$$f_9^{overlap} = \begin{cases} 1 - \left(1 + \alpha R + \frac{1}{2}(\alpha R)^2 + \frac{1}{6}(\alpha R)^3 + \frac{1}{24}(\alpha R)^4 \right. \\ \left. + \frac{1}{120}(\alpha R)^5 + \frac{1}{720}(\alpha R)^6 + \frac{1}{5040}(\alpha R)^7\right) e^{-\alpha R}, \quad \alpha_i = \alpha_j \\ 1 - A^2\left(1 + \alpha_i R + \frac{1}{2}(\alpha_i R)^2 + \frac{1}{6}(\alpha_i R)^3 + \frac{4}{105}(\alpha_i R)^4 + \frac{1}{210}(\alpha_i R)^5\right) e^{-\alpha_i R} - \\ B^2\left(1 + \alpha_j R + \frac{1}{2}\left(\alpha_j R\right)^2 + \frac{1}{6}\left(\alpha_j R\right)^3 + \frac{4}{105}\left(\alpha_j R\right)^4 + \frac{1}{210}\left(\alpha_j R\right)^5\right) e^{-\alpha_j R} - \\ 2A^2 B\left(1 + \alpha_i R + \frac{3}{7}(\alpha_i R)^2 + \frac{2}{21}(\alpha_i R)^3 + \frac{1}{105}(\alpha_i R)^4\right) e^{-\alpha_i R} - \\ 2B^2 A\left(1 + \alpha_j R + \frac{3}{7}\left(\alpha_j R\right)^2 + \frac{2}{21}\left(\alpha_j R\right)^3 + \frac{1}{105}\left(\alpha_j R\right)^4\right) e^{-\alpha_j R}, \qquad \alpha_i \neq \alpha_j \end{cases}$$

$$B = \frac{\alpha_i^2}{\left(\alpha_i^2 - \alpha_j^2\right)^2}, \quad A = \frac{\alpha_j^2}{\left(\alpha_j^2 - \alpha_i^2\right)^2}$$

Appendix D

Permanent electrostatic field of the HIPPO water model.

(field at induced dipole i, due to permanent moments of atom j)

$$\boldsymbol{F_i^{perm}}(R) = Z_j \nabla\left(\frac{1}{R}\right) + \mathrm{Q_j}\nabla\left(\frac{1}{R} f^{damp}(R)\right) + \boldsymbol{\mu_j} \cdot \nabla^2\left(\frac{1}{R} f^{damp}(R)\right) + \boldsymbol{\Theta_j} : \nabla^3\left(\frac{1}{R} f^{damp}(R)\right)$$

$$\nabla\left(\frac{1}{R} f^{damp}(R)\right) = -f_3^{damp}\frac{R_\alpha}{R^3}$$

$$\nabla^2\left(\frac{1}{R} f^{damp}(R)\right) = f_5^{damp}\frac{3R_\alpha R_\beta}{R^5} - f_3^{damp}\frac{\delta_{\alpha\beta}}{R^3}$$

$$\nabla^3\left(\frac{1}{R} f^{damp}(R)\right) = -f_7^{damp}\frac{15R_\alpha R_\beta R_\gamma}{R^7} + f_5^{damp}\frac{3\left(R_\alpha\delta_{\beta\gamma} + R_\beta\delta_{\alpha\gamma} + R_\gamma\delta_{\alpha\beta}\right)}{R^5}$$

$$f_3^{damp} = 1 - \left(1 + \alpha_j R + \frac{1}{2}\left(\alpha_j R\right)^2\right)e^{-\alpha_j R}$$

$$f_5^{damp} = 1 - \left(1 + \alpha_j R + \frac{1}{2}\left(\alpha_j R\right)^2 + \frac{1}{6}\left(\alpha_j R\right)^3\right)e^{-\alpha_j R}$$

$$f_7^{damp} = 1 - \left(1 + \alpha_j R + \frac{1}{2}\left(\alpha_j R\right)^2 + \frac{1}{6}\left(\alpha_j R\right)^3 + \frac{1}{30}\left(\alpha_j R\right)^4\right)e^{-\alpha_j R}$$

Induced dipole electrostatic field of the HIPPO water model.

(field at induced dipole i, due to induced dipole j)

$$\boldsymbol{F_i^{ind}}(R) = \boldsymbol{\mu_j^{ind}} \cdot \nabla^2\left(\frac{1}{R} f^{overlap}(R)\right)$$

$$\nabla^2\left(\frac{1}{R} f^{overlap}(R)\right) = f_5^{overlap}\frac{3R_\alpha R_\beta}{R^5} - f_3^{overlap}\frac{\delta_{\alpha\beta}}{R^3}$$

$$f_3^{overlap} = \begin{cases} 1 - \left(1 + \alpha R + \frac{1}{2}(\alpha R)^2 + \frac{7}{48}(\alpha R)^3 + \frac{1}{48}(\alpha R)^4\right)e^{-\alpha R}, & \alpha_i = \alpha_j \\ 1 - A^2\left(1 + \alpha_i R + \frac{1}{2}(\alpha_i R)^2\right)e^{-\alpha_i R} - B^2\left(1 + \alpha_j R + \frac{1}{2}\left(\alpha_j R\right)^2\right)e^{-\alpha_j R} - \\ \qquad 2A^2 B(1 + \alpha_i R)e^{-\alpha_i R} - 2B^2 A\left(1 + \alpha_j R\right)e^{-\alpha_j R}, & \alpha_i \neq \alpha_j \end{cases}$$

$$f_5^{overlap} = \begin{cases} 1 - \left(1 + \alpha R + \frac{1}{2}(\alpha R)^2 + \frac{1}{6}(\alpha R)^3 + \frac{1}{24}(\alpha R)^4 + \frac{1}{144}(\alpha R)^5\right) e^{-\alpha R}, & \alpha_i = \alpha_j \\ 1 - A^2\left(1 + \alpha_i R + \frac{1}{2}(\alpha_i R)^2 + \frac{1}{6}(\alpha_i R)^3\right) e^{-\alpha_i R} - \\ B^2\left(1 + \alpha_j R + \frac{1}{2}\left(\alpha_j R\right)^2 + \frac{1}{6}\left(\alpha_j R\right)^3\right) e^{-\alpha_j R} - \\ 2A^2 B\left(1 + \alpha_i R + \frac{1}{3}(\alpha_i R)^2\right) e^{-\alpha_i R} - \\ 2B^2 A\left(1 + \alpha_j R + \frac{1}{3}\left(\alpha_j R\right)^2\right) e^{-\alpha_j R}, & \alpha_i \neq \alpha_j \end{cases}$$

$$B = \frac{\alpha_i^2}{\left(\alpha_i^2 - \alpha_j^2\right)^2}, \quad A = \frac{\alpha_j^2}{\left(\alpha_j^2 - \alpha_i^2\right)^2}$$

Appendix E

Pauli Repulsion equations for the HIPPO water model.

$$U_{ij} = \frac{K_i K_j}{R} S_{total}^2$$

$$S_{total}^2 = \vec{M}_i \boldsymbol{T}_{\boldsymbol{ij}}^{\boldsymbol{repulsion}} \vec{M}_j$$

$$\boldsymbol{T}_{\boldsymbol{ij}}^{\boldsymbol{repulsion}} = \begin{bmatrix} 1 & \nabla & \nabla^2 \\ \nabla & \nabla^2 & \nabla^3 \\ \nabla^2 & \nabla^3 & \nabla^4 \end{bmatrix} \left(\alpha_i^3 \alpha_j^3 T^{pauli}\right)$$

$$T^{pauli} = \frac{1}{R} f_1^{rep}$$

$$\nabla T^{pauli} = -f_3^{rep} \frac{R_\alpha}{R^3}$$

$$\nabla^2 T^{pauli} = f_5^{rep} \frac{3R_\alpha R_\beta}{R^5} - f_3^{rep} \frac{\delta_{\alpha\beta}}{R^3}$$

$$\nabla^3 T^{pauli} = -f_7^{rep} \frac{15R_\alpha R_\beta R_\gamma}{R^7} + f_5^{rep} \frac{3\left(R_\alpha \delta_{\beta\gamma} + R_\beta \delta_{\alpha\gamma} + R_\gamma \delta_{\alpha\beta}\right)}{R^5}$$

$$\nabla^4 T^{pauli}$$

$$= f_9^{rep} \frac{105R_\alpha R_\beta R_\gamma R_\eta}{R^9}$$

$$- f_7^{rep} \frac{15\left(R_\alpha R_\beta \delta_{\gamma\eta} + R_\alpha R_\gamma \delta_{\beta\eta} + R_\alpha R_\eta \delta_{\beta\gamma} + R_\beta R_\gamma \delta_{\alpha\eta} + R_\beta R_\eta \delta_{\alpha\gamma} + R_\gamma R_\eta \delta_{\alpha\beta}\right)}{R^7}$$

$$+ f_5^{rep} \frac{3\left(\delta_{\alpha\beta}\delta_{\gamma\eta} + \delta_{\alpha\gamma}\delta_{\beta\eta} + \delta_{\alpha\eta}\delta_{\beta\gamma}\right)}{R^5}$$

$$f_1^{rep} = \left(f_{exp}\right)^2$$

$$f_{exp} = \begin{cases} \frac{1}{\alpha^3}\left(1 + \frac{\alpha R}{2} + \frac{1}{3}\left(\frac{\alpha R}{2}\right)^2\right) e^{\frac{-\alpha R}{2}}, & \alpha_i = \alpha_j \\ \frac{1}{2X^3 R}\left[\alpha_i (RX - 2\alpha_j) e^{\frac{-\alpha_j R}{2}} + \alpha_j (RX + 2\alpha_i) e^{\frac{-\alpha_i R}{2}}\right], & \alpha_i \neq \alpha_j \end{cases}$$

$$f_3^{rep} = 2 f_{exp} f'_{exp}$$

$$f'_{exp}$$

$$= \begin{cases} \frac{1}{\alpha^3}\frac{1}{3}\left(\frac{\alpha}{2}\right)^2 \left(1 + \frac{\alpha R}{2}\right) e^{\frac{-\alpha R}{2}}, & \alpha_i = \alpha_j \\ \frac{1}{2X^3 R}\left[\left(\frac{1}{2}\alpha_i \alpha_j X - \frac{\alpha_i \alpha_j^2}{R} - \frac{2\alpha_i \alpha_j}{R^2}\right) e^{\frac{-\alpha_j R}{2}} + \left(\frac{1}{2}\alpha_i \alpha_j X + \frac{\alpha_j \alpha_i^2}{R} + \frac{2\alpha_i \alpha_j}{R^2}\right) e^{\frac{-\alpha_i R}{2}}\right], & \alpha_i \neq \alpha_j \end{cases}$$

$$f_5^{rep} = 2\left(f_{exp} f''_{exp} + f'_{exp} f'_{exp}\right)$$

$$f''_{exp} = \begin{cases} \frac{1}{\alpha^3}\frac{1}{9}\left(\frac{\alpha}{2}\right)^4 e^{\frac{-\alpha R}{2}}, & \alpha_i = \alpha_j \\ \frac{1}{2X^3R^2}\left[\begin{matrix}\left(\frac{1}{4}\alpha_i\alpha_j^2X - \frac{\alpha_i\alpha_j^3}{2R} + \frac{\alpha_i\alpha_jX}{2R} - \frac{3\alpha_i\alpha_j^2}{R^2} - \frac{6\alpha_i\alpha_j}{R^5}\right)e^{\frac{-\alpha_jR}{2}} + \\ \left(\frac{1}{4}\alpha_j\alpha_i^2X + \frac{\alpha_j\alpha_i^3}{2R} + \frac{\alpha_j\alpha_iX}{2R} + \frac{3\alpha_j\alpha_i^2}{R^2} + \frac{6\alpha_j\alpha_i}{R^5}\right)e^{\frac{-\alpha_iR}{2}}\end{matrix}\right], & \alpha_i \neq \alpha_j \end{cases}$$

$$f_7^{rep} = 2\left(f_{exp}f'''_{exp} + 3f''_{exp}f'_{exp}\right)$$

$$f'''_{exp} = \begin{cases} \frac{1}{\alpha^3}\frac{1}{45}\left(\frac{\alpha}{2}\right)^5\frac{1}{R}e^{\frac{-\alpha R}{2}}, & \alpha_i = \alpha_j \\ \frac{1}{2X^3R^3}\left[\begin{matrix}\left(\frac{1}{8}\alpha_i\alpha_j^3X + \frac{3}{4}\frac{\alpha_i\alpha_j^2X}{R} + \frac{3}{2}\frac{\alpha_i\alpha_jX}{R^2} - \frac{1}{4}\frac{\alpha_i\alpha_j^4}{R} - \frac{3\alpha_i\alpha_j^3}{R^2} - \frac{15\alpha_i\alpha_j^2}{R^3} - \frac{30\alpha_i\alpha_j}{R^4}\right)e^{\frac{-\alpha_jR}{2}} + \\ \left(\frac{1}{8}\alpha_j\alpha_i^3X + \frac{3}{4}\frac{\alpha_j\alpha_i^2X}{R} + \frac{3}{2}\frac{\alpha_j\alpha_iX}{R^2} + \frac{1}{4}\frac{\alpha_j\alpha_i^4}{R} + \frac{3\alpha_j\alpha_i^3}{R^2} + \frac{15\alpha_j\alpha_i^2}{R^3} + \frac{30\alpha_j\alpha_i}{R^4}\right)e^{\frac{-\alpha_iR}{2}}\end{matrix}\right], & \alpha_i \neq \alpha_j \end{cases}$$

$$f_9^{rep} = 2\left(f_{exp}f''''_{exp} + 4f'''_{exp}f'_{exp} + 3f''_{exp}f''_{exp}\right)$$

$$f''''_{exp} = \begin{cases} \frac{1}{\alpha^3}\frac{1}{315}\left(\frac{\alpha}{2}\right)^5\frac{1}{R^3}\left(1+\frac{\alpha R}{2}\right)e^{\frac{-\alpha R}{2}}, & \alpha_i = \alpha_j \\ \frac{1}{2X^3R^4}\left[\begin{matrix}\left(\begin{matrix}\frac{1}{16}\alpha_i\alpha_j^4X + \frac{3}{4}\frac{\alpha_i\alpha_j^3X}{R} + \frac{15}{4}\frac{\alpha_i\alpha_j^2X}{R^2} + \frac{15}{2}\frac{\alpha_i\alpha_jX}{R^3} - \\ \frac{1}{8}\frac{\alpha_i\alpha_j^5}{R} - \frac{5}{2}\frac{\alpha_i\alpha_j^4}{R^2} - \frac{45}{2}\frac{\alpha_i\alpha_j^3}{R^3} - \frac{105\alpha_i\alpha_j^2}{R^4} - \frac{210\alpha_i\alpha_j}{R^5}\end{matrix}\right)e^{\frac{-\alpha_jR}{2}} + \\ \left(\begin{matrix}\frac{1}{16}\alpha_j\alpha_i^4X + \frac{3}{4}\frac{\alpha_j\alpha_i^3X}{R} + \frac{15}{4}\frac{\alpha_j\alpha_i^2X}{R^2} + \frac{15}{2}\frac{\alpha_j\alpha_iX}{R^3} + \\ \frac{1}{8}\frac{\alpha_j\alpha_i^5}{R} + \frac{5}{2}\frac{\alpha_j\alpha_i^4}{R^2} + \frac{45}{2}\frac{\alpha_j\alpha_i^3}{R^3} + \frac{105\alpha_j\alpha_i^2}{R^4} + \frac{210\alpha_j\alpha_i}{R^5}\end{matrix}\right)e^{\frac{-\alpha_iR}{2}}\end{matrix}\right], & \alpha_i \neq \alpha_j \end{cases}$$