NOVEL FUNCTIONAL FORMS AND PARAMETERIZATION METHODS FOR AB INITIO FORCE FIELD DEVELOPMENT

by

Mary J. Van Vleet

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

(Chemistry)

at the

UNIVERSITY OF WISCONSIN–MADISON

2017

Date of final oral examination: 08/15/17

The dissertation is approved by the following members of the Final Oral Committee:
J.R. Schmidt, Associate Professor, Chemistry
Clark R. Landis, Professor, Chemistry
Qiang Cui, Professor, Chemistry
Arun Yethiraj, Professor, Chemistry
Reid Van Lehn, Assistant Professor, Chemical and Biological Engineering

Soli Deo gloria.

ACKNOWLEDGMENTS

It is customary for authors of academic books to include in their prefaces statements such as this: "I am indebted to ... for their invaluable help; however, any errors which remain are my sole responsibility." Occasionally an author will go further. Rather than say that if there are any mistakes then he is responsible for them, he will say that there will inevitably be some mistakes and he is responsible for them....

Although the shouldering of all responsibility is usually a social ritual, the admission that errors exist is not — it is often a sincere avowal of belief. But this appears to present a living and everyday example of a situation which philosophers have commonly dismissed as absurd; that it is sometimes rational to hold logically incompatible beliefs.

— David C. Makinson (1965)

CONTENTS

LIST OF TABLES

LIST OF FIGURES

NOVEL FUNCTIONAL FORMS AND PARAMETERIZATION METHODS FOR AB INITIO FORCE FIELD DEVELOPMENT

Mary J. Van Vleet

Under the supervision of Professor J.R. Schmidt
At the University of Wisconsin-Madison

Molecular simulation is an essential tool for interpreting and predicting the structure, thermodynamics, and dynamics of chemical and biochemical systems. The fundamental inputs into these simulations are the intra- and intermolecular force fields, which provide simple and computationally efficient descriptions of molecular interactions. Consequently, the utility of molecular simulation ultimately depends on the fidelity of the force field to the underlying (exact) potential energy surface. This dissertation describes a number of novel advances designed to improve the accuracy and predictive power of (specifically ab initio) intermolecular force fields. By fitting ab initio force fields to first-principles-based functional forms and chemically-meaningful parameters, and by taking frequent advantage of the physically-motivated partitioning afforded by Symmetry-Adapted Perturbation Theory (SAPT) and Iterated Stockholder Atoms (ISA) approaches, we demonstrate how the resulting force fields can be applied to describe a broad range of molecular systems in different chemical and physical environments. Our newly-developed MASTIFF approach achieves quantitative accuracy with respect to both high-level electronic structure theory and experiment, and is thus well suited for use in 'next-generation' ab initio force field development and large-scale molecular simulation.

J.R. Schmidt

ABSTRACT

Molecular simulation is an essential tool for interpreting and predicting the structure, thermodynamics, and dynamics of chemical and biochemical systems. The fundamental inputs into these simulations are the intra- and intermolecular force fields, which provide simple and computationally efficient descriptions of molecular interactions. Consequently, the utility of molecular simulation ultimately depends on the fidelity of the force field to the underlying (exact) potential energy surface. This dissertation describes a number of novel advances designed to improve the accuracy and predictive power of (specifically ab initio) intermolecular force fields. By fitting ab initio force fields to first-principles-based functional forms and chemically-meaningful parameters, and by taking frequent advantage of the physically-motivated partitioning afforded by Symmetry-Adapted Perturbation Theory (SAPT) and Iterated Stockholder Atoms (ISA) approaches, we demonstrate how the resulting force fields can be applied to describe a broad range of molecular systems in different chemical and physical environments. Our newly-developed MASTIFF approach achieves quantitative accuracy with respect to both high-level electronic structure theory and experiment, and is thus well suited for use in 'next-generation' ab initio force field development and large-scale molecular simulation.

PUBLISHED WORK AND WORK IN PREPARATION

[95] Van Vleet, M. J.; Misquitta, A. J.; Stone, A. J.; Schmidt, J. R. *J. Chem. Theory Comput.* **2016**, *12*, 3851–3870.

[263] Van Vleet, M. J.; Misquitta, A. J.; Schmidt, J. R. *J. Chem. Theory Comput.* **2017**, submitted.

[3] Van Vleet, M.; Weng, T.; Schmidt, J. *Chem. Rev.* **2017**, invited, manuscript in preparation.

1 INTRODUCTION AND BACKGROUND

1.1 Molecular Simulation: History and Importance

What are the functions of proteins in the body? How can we identify new and better drugs for improved disease treatment, or optimal materials for designing efficient solar cells? What are the microscopic mechanisms by which chemicals interact, undergo phase transitions, or react to form entirely new species? Increasingly, these and other essential chemical questions can be addressed with the aid of computer simulation,[18–22] enabling us to, for example, peer into the detailed mechanisms of enzyme catalysis,[23] watch proteins fold,[24–27] virtually screen for novel drug candidates,[28] improve industrial materials,[29–31] and directly simulate hard-to-understand phase transformations at an atomistic level.[32,33] The question, of course, is: how?

To understand the manner in which computer simulation can be used to obtain obervable experimental properties of interest,[†] we first summarize several fundamental physical principles that help define the fundemand inputs and important techniques required for molecular simulation. The foundation for molecular simulation comes from the field of statistical mechanics, where by the mid 19-$^{\text{th}}$ century it was discovered that experimental observables, which we denote $\mathcal{O}$, depend entirely on a system's temperature, T, and the energies available to that system,

$$\langle \mathcal{O} \rangle = \frac{\int d\mathbf{p} d\mathbf{x}\, \mathcal{O}(\mathbf{p}, \mathbf{x}) \exp(-H(\mathbf{p}, \mathbf{q})/k_B T)}{\int d\mathbf{p} d\mathbf{x}\, \exp(-H(\mathbf{p}, \mathbf{x})/k_B T)} \tag{1.1}$$

Here the angle brackets denote that we're interested in some *average* value of the observable, $\mathbf{p}$ and $\mathbf{x}$ denote, respectiely, the momentum and positions of the particles in the system, H defines the classical Hamiltonian describing the potential and kinetic energy of that system, and k_B is Boltzmann's constant.[34] Technically, this expression only holds for classical systems at constant temperature, however conceptually-similar expressions can be derived outside of these assumptions. Even though atomic behavior is, in principle, quantum mechanical, subsequent research

over the years have shown that very satisfactory molecular properties, particularly for heavier atoms at higher temperatures,[35] can be obtained by treating molecular systems according to the above classical description. As a result of this highly important insight (indeed, Karplus, Levitt, and Warshel won the Nobel prize in 2013 for these ideas and related work),[35] it becomes possible to use Eq. (1.1) to extract information regarding macroscopic properties of interest given knowledge of the potential and kinetic energies available in a classical description of the molecular system.

In practice, most chemical systems contain large numbers of particles, and as the number of degress of freedom in the system increase, so does the dimensionality of the integral in Eq. (1.1). In most practical scenarios, analytical integration is not possible, and for large systems it is computationally prohibitive to solve for $\langle \mathcal{O} \rangle$ by standard numerical integration techniques.[34] In 1953, however a group of researchers[36] showed how Eq. (1.1) can be systematically estimated $\langle \mathcal{O} \rangle$ via appropriate random sampling of the integrand of Eq. (1.1) based on a probability distribution ρ:[34,37]

$$\langle \mathcal{O} \rangle = \langle \mathcal{O} \rangle_{\text{ens}} = \lim_{\tau_{\text{obs}} \to \infty} \frac{1}{\tau_{\text{obs}}} \sum_{\tau=1}^{\tau_{\text{obs}}} \mathcal{O}(\mathbf{\Gamma}(\tau)) \tag{1.2}$$

In contrast to Eq. (1.1), Eq. (1.2) is the average over a total number of sampled observations, τ_{obs}, taken of the system and its properties, and $\mathbf{\Gamma}(\tau)$ denotes the collective positions and momenta that define the state of the system at each sample point. The key technique that defines the resulting Metropolis Monte Carlo (MC) algorithm is known as 'importance sampling': provided we cleverly choose our probability distribution, ρ, to be identical to the Boltzmann distribution, $\rho \propto \exp(-H(\mathbf{p}, \mathbf{x})/k_B T)$, the right-hand side of Eq. (1.2) converges fairly rapidly as a function of τ_{obs}. The interested reader is directed to Allen and Tildesley[34] for detailed information on

[†] We've been intentionally vague about what these 'experimental properties of interest' might be, as the experimental properties one finds important vary considerably between applications. To offer some concrete examples, drug discovery studies are often interested in the binding free energies between proteins and prospective drug molecules,[28,31] and the optimization of putative solar cell materials often focuses on open circuit voltages and/or short-circuit currents.[31]

the exact techniques, algorithms, and practical concerns involved in importance sampling. In general, however, it is sufficient to know that MC is one of the main algorithms used to evaluate average molecular properties, and that ultimately the MC technique depends on the accuracy with which we can evaluate the potential and kinetic energies of a given system.

As an alternative strategy to MC, at the turn of the 20$^{\text{th}}$ century Ludwig Boltzmann proposed his now-famous 'ergodic' hypothesis.[38] This hypothesis states that, over sufficiently long time periods, the 'ensemble average' defined in Eq. (1.1) becomes identical to the 'time-average' that results from studying the system's dynamical behavior,

$$\langle \mathcal{O} \rangle = \langle \mathcal{O} \rangle_{\text{time}} = \langle \mathcal{O}(\mathbf{\Gamma}(t) \rangle_{\text{time}} = \lim_{t_{\text{obs}} \to \infty} \frac{1}{t_{\text{obs}}} \int_0^{t_{\text{obs}}} \mathcal{O}(\mathbf{\Gamma}(t)) dt, \tag{1.3}$$

where t_{obs} indicates the length of time over which we average the system's properties.[34] The right-hand side of Eq. (1.3) shows that, so long as we know and can solve for the equations of motion that govern a given system's behavior, we can simulate the time evolution of that system until the r.h.s. of Eq. (1.3) converges, thereby obtaining a prediction for $\langle \mathcal{O} \rangle$. Classical mechanics is governed simply by Newton's laws of motion,

$$-\frac{dE(\mathbf{x})}{d\mathbf{x}} \equiv \mathbf{F}(\mathbf{x}) = m\ddot{\mathbf{x}}, \tag{1.4}$$

with m a mass, and $\mathbf{F}$ the force acting on a particle. Thus as with MC, we need only know the potential energy of the system (and by extension its constituent forces) in order to solve for the time-dependent positions, momenta, and (ultimately) properties of a system. The resulting integration of Newton's equations of motion, a process known as molecular dynamics (MD), has made it possible to study both the kinetic and thermodynamic properties of a wide range of molecular systems, historically beginning with monatomic liquids in 1964, and soon leading to the first polymer and protein simulations in 1975 and 1977, respectively.[18] Ever since these

first studies, and as a complement to MC, MD simulation has increasingly become a preeminent tool in the investigation and prediction of chemical phenomena.

1.2 Molecular Simulation: Challenges and Unanswered Questions

Recent successes in both MC and MD have shown great promise for using molecular simulation in the understanding, interpretation, and even prediction of experimental results,[18] making simulation a powerful complement to traditional experimental tools.[19–21,29–31,39,40] Nevertheless, accurate and insightful molecular simulation depends on success in the following three critical aspects of any MD/MC simulation:[25]

1. We must be able to accurately and efficiently quantify the potential energy, E_n, of any state n of the system that might get sampled by the MD/MC simulation.[41–44] Henceforth we will collectively refer to these energies, given as a function of the system state Γ, as the potential energy surface (PES) of a system (Some visual examples of representative PESs are shown in Fig. 1.1).

2. We must be able to obtain a representative sample of all states of the system over a sufficiently long timescale (commensurate with the timescale(s) of the chemical phenomena of interest) so as to obtain converged property predictions.[45–47] This class of problems is often referred to simply as 'sampling issues'.

3. Especially when interested in *interpreting* chemical phenomena, we must be able to analyze the results of a simulation in such a way as to garner detailed, chemically-intuitive insight into the problem at hand.[48–50] While this task is relatively straightforward for homogeneous liquids and other 'simple' systems, it can become decidedly difficult for analyzing complex properties and mechanisms, such as with using simulation to investigate protein folding mechanisms.

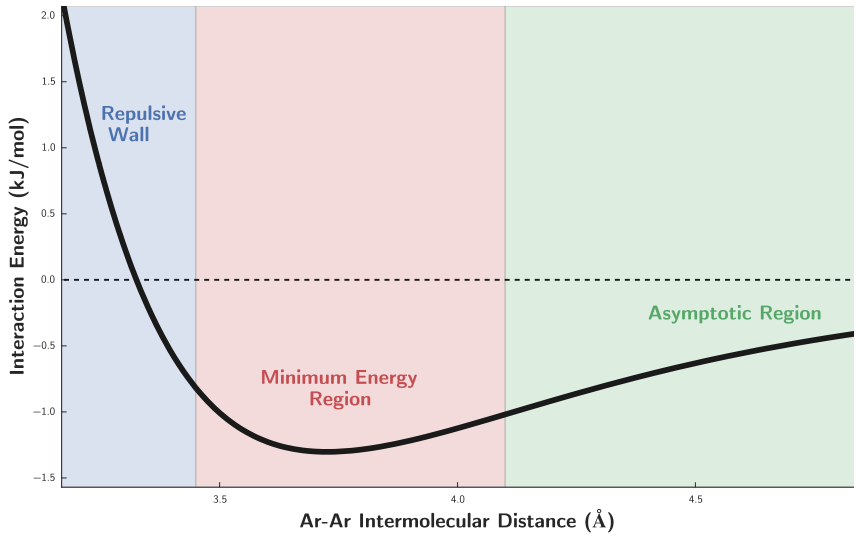
Though our focus in this dissertation will be on the first point (that of computing potential energies for molecular simulation), all three aspects of molecular simulation are challenging in their own right, and form highly active and important areas of research.[22] Moreover, there is significant interplay between these areas in terms of research development. As an example, improvements in sampling methods often lead to increased computational efficiency, thereby enabling use of more accurate (but more costly) representations of the PES. Conversely, the next section will discuss how the development of cost-effective potential energy functions is often necessary for running simulations over long enough time scales to ensure representative sampling and robust interpretation of the simulation results. Clearly, insofar as molecular simulation is concerned, both accuracy and computational efficiency are of paramount importance.

In the pursuit of increasingly accurate, insightful, and predictive molecular simulation, it is clear that we must be able to quantitatively represent the PES of any molecular system, and, furthermore, that our mathematical representation of this PES must be sufficiently accurate and cost-effective so as to enable simulation that is chemically insightful (given the type of simulation analysis required for a particular problem or application) and computationally affordable (in accordance with the length of molecular simulation that will need to be run in order to appropriately deal with any sampling issues). Bearing these stipulations in mind, we can now broadly state the guiding question for this dissertation: in the pursuit of accurate and insightful molecular simulation, how can we optimally obtain a mathematical description of the PES?

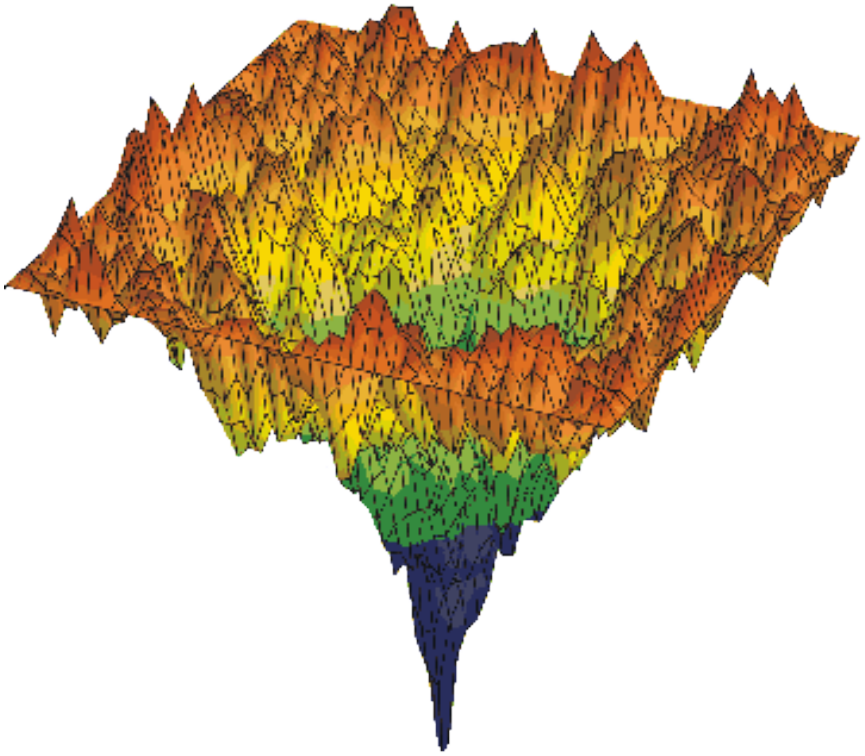
1.3 Molecular Simulation: Cost Considerations

Calculations of a PES ultimately rely on finding approximate solutions to the time-independent Schrödinger equation,

$$\hat{H}\Psi = E\Psi \tag{1.5}$$



(a) One-dimensional PES for the argon dimer.



(b) Three-dimensional representation of an N-dimensional PES for a protein. Copied under a CC license from Chaplin [51]

Figure 1.1: Simple and complex potential energy surfaces for molecular systems.

where $\hat{H}$ is a quantum-mechanical operator describing both the kinetic and potential energies of the nuclei and electrons, Ψ is a wavefunction, and E is the energy of the system as a function of the nuclear coordinates. A plethora of approximations exist for solving Eq. (1.5), each with their own disadvantages and advantages, and a full discussion of the accuracy of such electronic structure theories (ESTs) can be found in Cramer[52] and other texts. These methods differ greatly into terms of computational cost with respect to system size, and so we begin our discussion of ESTs in terms of the cost-efficiency with which various methods might be used in molecular simulation. At the high cost end of the spectrum, extremely accurate 'gold-standard' EST calculations can be run using a method known as CCSD(T), whose cost scales as N^7 with respect to the number of electrons in the system. More approximate methods, such as MP2 and HF, scale as N^5 and N^4, respectively, and Density Functional Theory (DFT) (frequenctly regarded as the 'computationally-affordable' workhorse of EST) scales even more modestly as N^3. To put these scalings in context, however, Fig. 1.2 shows the largest system sizes (given as a number of atoms) which a given EST is capable of computing using available computational resources. Though these estimates are taken from the early 2000s (since which the 'TeraFlops MPP' supercomputer has been superseded in 2016 by various PetaFlops supercomputers with 1000x the computer power), severel of the conclusions are still the same, namely that CCSD(T) and most other ESTs remain too expensive to be employed in large-scale molecular simulation.

Compounding the above scaling problem, molecular simulation requires that we compute, not just one snapshot of a molecular PES, but millions, billions, or (for protein folding simulations) even trillions of such energy snapshots. Thus depending on the lengths and timescales involved in the chemical processes under study, even the cheapest DFT ESTs are typically too expensive for routine molecular simulation. (For reference, in 2014 DFT-based simulations were reported on roughly 10^3-atom systems using state-of-the-art facilities,[41] whereas $10^5 - 10^6$ atoms can be required for running representative simulations of proteins.[21,25])

As computing power continues to increase, there is no doubt that advanced and accurate ESTs will eventually be used in larger-scale molecular simulation to

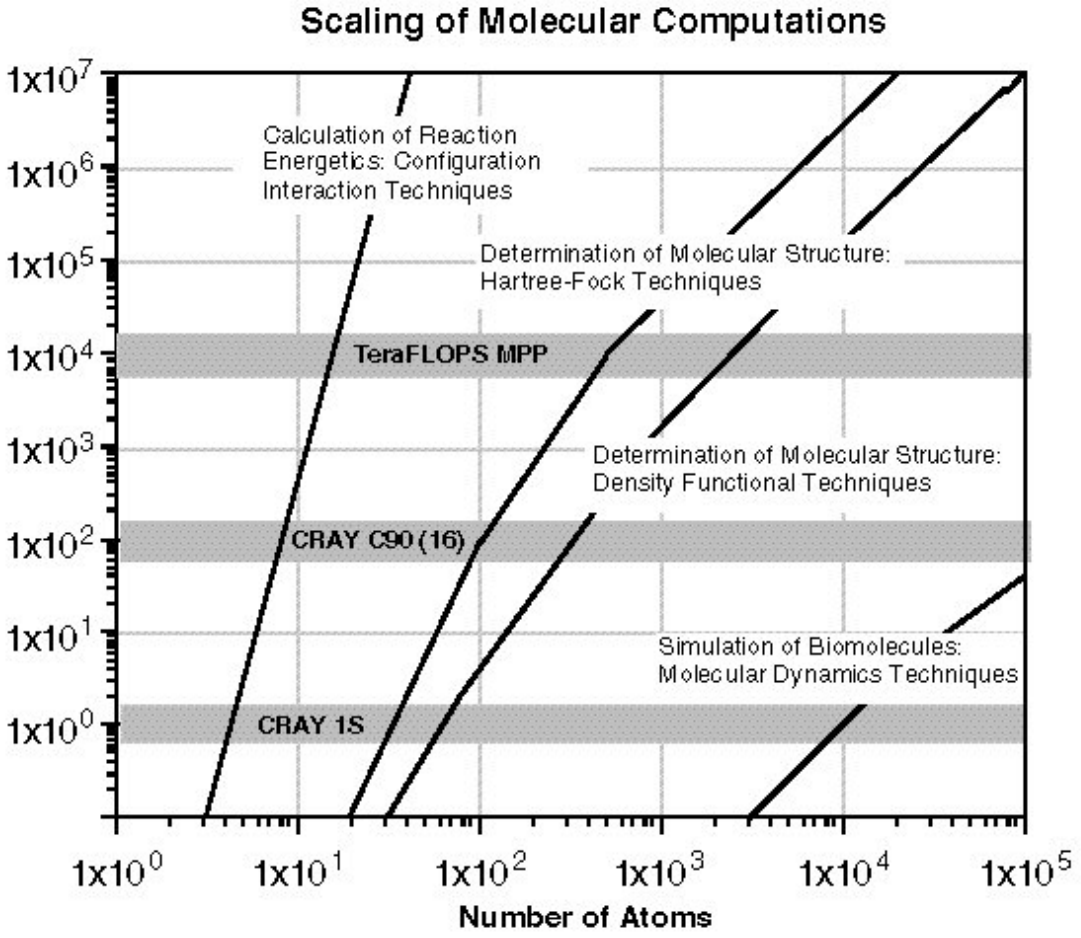


Figure 1.2: The relative computing power required for molecular computations at four levels of theory. In the absence of screening techniques, the formal scaling for configuration interaction, Hartree-Fock, density functional, and molecular dynamics is: N6, N4 , N3 and N2 , respectively. Reprinted from Ref. 8.

investigate problems of both chemical and biological import.[41] In the meantime, however, and for investigating problems on extremely large time- and length-scales, lower-cost methods for calculating the PES are required. In order to achieve this low-cost scaling, a popular strategy is to model the PES using simple mathematical expressions which depend only on the positions of the nuclei in the system. Such models are referred to as 'force fields', which are defined both by the choice of functional forms (that is, mathematical formulae) and parameters that go into them. The computational cost of force fields nominally scales as N^2 (and usually scales as $N \log N$ in practice) with respect to the number of *atoms* (rather than electrons) in the system, thus representing significnat computational savings compared to the ESTs described above. Fig. 1.2 shows the scaling of these methods in the N^2 scaling limit, from which it becomes clear that we can use force fields to study system sizes 2–3 *orders of magnitude* larger than what is possible with DFT. Indeed, in the early 1990s Chan and Dill[9] provided a useful estimation of the computing power needed to simulate a variety of important chemical and biological processes (Fig. 1.3), and showed that, using these cost-efficient force fields, we are not far off from the time when atomistic simulations of protein folding and/or aggregation can be achieved. Some of the first simulations of protein folding have already been reported, and using computationally-efficient force fields we can expect this trend to continue into the foreseeable future.[25]

1.4 Force Fields

Despite the advantages and opportunities afforded by their computational efficiency, molecular simulation can also be *limited* by force fields in the sense that, in the absence of fortuitous error cancellation, the predictive accuracy of molecular simulation is inextricably tied to the accuracy of the force field used to run the simulation. For this reason, one of the central challenges facing molecular simulation today is the development of new and more accurate force fields.[41,53]

To understand why the development of accurate force fields remains so challenging, it's worthwhile to briefly discuss the development process itself, both in

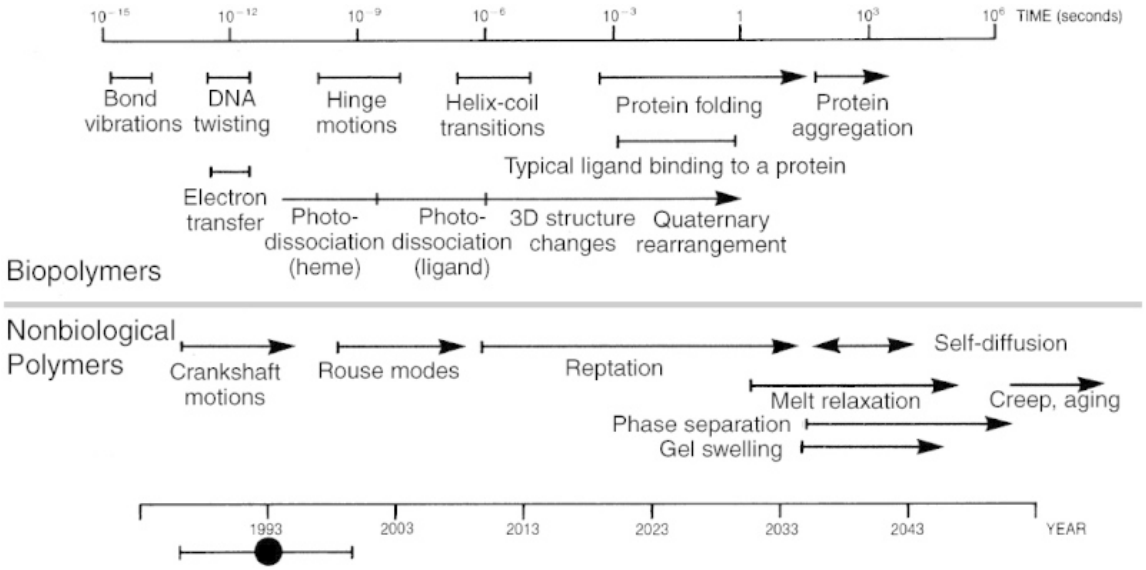


Figure 1.3: Time scales for various motions within biopolymers (upper) and nonbiological polymers (lower). The year scale at the bottom shows estimates of when each such process might be accessible to brute force molecular simulation on supercomputers, assuming that parallel processing capability on supercomputers increases by about a factor of 1,000 every 10 years (i.e., one order of magnitude more than Moore's law) and neglecting new approaches or breakthroughs. Reprinted with permission from H.S. Chan and K. A. Dill. Physics Today, 46, 2, 24, (1993).[9]

terms of the functional forms that get used in force fields as well as the manner in which these functional forms are parameterized. Traditionally, force fields have been crafted by an 'empirical' development proccess,[54] in which the force field functional form is parameterized so as to reproduce select experimental properties of interest. The obvious advantage of such a strategy is that, so long as one parameterizes and investigates a limited scope of chemical, physical, and/or structural conditions, there is a good chance that empirical force fields will be of good accuracy in providing a microscopic picture of the macroscopic experimental properties. Thus, for instance, empirical force fields can have remarkable success in simulating the behavior of folded proteins in biologically-relevant environments.[26,42,55]

Despite these successes, empirical force field development also faces significant

challenges. The first challenge is one of 'transferability': outside of the parameterization scope discussed above, there is little guarantee that empirical force fields will retain the good accuracy that can be expected within the original parameterization conditions.[53,56] To continue with our protein example, it has recently been shown how many empirical force fields, all of which generally provide similar predictions regarding the properties of folded proteins, differ widely when it comes to predicting a structurally-distinct class of partially unfolded, 'intrinsically-disorded' proteins.[26,57] Similarly, and despite much effort, it's still difficult to find an empirically-developed force field capable of correctly describing water across a wide range of physical and chemical environments.[58,59] More distressingly than the transferability problem, force field accuracy can sometimes be an issue even within the limited range of experimental conditions over which the force field was originally parameterized, and time-consuming re-parameterization methods must often be employed in order to correct for deficiencies in the original force field parameters.[53,56]

Arguably, the underlying reason why empirical force field development is limited (both in terms of accuracy and transferability) is that the force fields themselves are typical based on rather limited, or 'effective', physics.[60–62] Explicit many-body polarization, for instance, is not often accounted for in empirical force fields, despite the fact that it is known to be an important factor in many important chemical phenomenon.[53,58,63] Similarly, accurate multipolar expansions of electrostatic energies are often reduced to more approximate point charge models,[64] charge penetration effects are usually neglected,[60,61] and exchange effects are described by an overly-repulsive (but computationally convenient) $1/r_{ij}^{12}$ functional form.[60–62,65,66] In some cases, these modeling choices are justified by increased gains in computational efficiency; indeed, it is only within the past decade or so that explicit polarization and higher-order multipolar electrostatic treatments have become computationally-affordable.[63,64,67–70] In other cases, however, empirical force field development is limited by the significant complications involved in parameterization. With empirical force field development, each additional parameter must be optimized on the basis of costly molecular simulations. Moreover, and because experiment typically

probes only the average total energy of a given system, parameters in empirical force fields must be fit simultaneously. Models with many parameters are usually too time-consuming to optimize, and too prone to issues of overfitting,[71] to warrant the effort. For these reasons, it is likely that empirically-fit force fields will remain restricted to parametrically simple, physically-approximate, models.

To circumvent the practical limitations of empirical force field development, an alternate strategy is to fit force fields, not directly against experimental properties, but rather to benchmark calculations of the underlying PES itself.[54] The drawback of such a first-principles, or 'ab initio', methodology, is obvious: by not fitting to experimental quantities, the resulting force fields will not closely match experiment unless we accurately and systematically account for all the relevant physics for a given system. For this reason, comparisons between an ab initio force fields and experiment (Chapter 3) are often complicated by factors such as the accuracy of the underlying benchmark PES or the treatment of many-body and/or quantum effects.[72–74]

Nevertheless, ab initio force field development has several clear advantages over its empirical counterpart. First, and especially for systems where experimental data is lacking, ab inito force fields can be fit to calcuated data in order to make novel experimental predictions. Furthermore (and as discussed in Section 1.5.2), ab initio force fields can be fit, not merely to the total energy of a system, but also on a component-by-component basis to individually reproduce each physically-meaningful contribution to the PES. This, along with the simplicitly afforded by directly parameterizing the PES, means that ab initio force fields can be fit to more complicated and more physically-motivated functional forms, thus enabling the possibility of increased accuracy in molecular simulation. Furthermore, we show in Section 1.5.2 how advanced ab initio parameterization methods can lead to decreased reliance on error cancellation and minimize overfitting, thus augmenting both the accuracy and transferability of the resulting force fields. Finally, with ab initio force fields we can easily assess the fit quality compared to an underlying benchmark PES; as will be a theme of this dissertation, such an ability to directly compare between putative model PESs enables us quickly evaluate new and im-

proved functional forms and parameterization methods for ab initio force field development.

1.5 Ab initio force field development with SAPT and ISA

As implied throughout the preceding discussion, goals for ab initio force field development are as follows:

1. **Accuracy**: Ab initio force fields should ideally be able to reproduce a benchamrk PES (as calculated from high-quality EST) to within chemical accuracy or better, with the knowledge that accuracy compared to the PES will be well-correlated with accuracy compared to experiment

2. **Transferability**: The parameters and functional forms used ab inito force fields should be transferable between chemical and physical environments without loss of accuracy

3. **Cost-Efficiency**: The computational cost of ab initio force fields should ideally be comparable to that of empirically-derived models

4. **Physicality**: So as to minimize a reliance on error cancellation and promote accuracy and transferability, functional forms and parameters for ab initio force fields should be grounded in accurate and physically-meaningful first principles theories

5. **Simplicity**: When possible, and where the accuracy and physicality of the model is not compromised, the parameterization methodologies and functional forms used in ab initio force field development should be kept as simple as possible, particularly so as to avoid overfitting

A number of strategies for ab inito force field development are present in the literature,[41,54] however here we focus on the general approach used in our group[75]

to generate optimal ab initio force fields. Additionally, as the intramolecular portion of a force field is usually more straightforward to optimize, we limit our discussion to the functional forms and parameters used in developing the intermolecular part of the potential. In what follows, we describe three main strategies employed in our group to guide ab inito force field development: separation of the N-body potential into 2- and many-body contributions via the many-body expansion (MBE) (Section 1.5.1), decomposition and subsequent component-by-component parameterization of the total two-body interaction energy using Symmetry-Adapted Perturbation Theory (SAPT) (Section 1.5.2), and characterization of the atom-in-molecule contributions to each energy component via Iterated Stockholder Atoms (ISA) (Section 1.5.3).

1.5.1 The Many-Body Expansion

For an N-body system (here and throughout we use the terms 'body' and 'atom' synonymously), the molecular PES is given as a $3N-6$ dimensional function of particle positions,[4,54,63,76]

$$V_N(\vec{r}_1,\vec{r}_2,\ldots,\vec{r}_N) = \sum_i^N V_1(\vec{r}_i) + \sum_{i<j}^N \Delta V_2(\vec{r}_i,\vec{r}_j) + \sum_{i<j<k}^N \Delta V_3(\vec{r}_i,\vec{r}_j,\vec{r}_k) + \ldots \quad (1.6)$$

Here, and without loss of generality, we have expressed this roughly 3N-dimensional surface as a 'many-body' expansion of n-body cluster interactions. Thus V_1 describes one-body, or intramolecular, contributions to the overall PES. ΔV_2 is referred to as the 'pair potential', and represents the difference in interaction energies between a two-body cluster, or 'dimer', and the individual monomers themselves. In a similar fashion, ΔV_3 corresponds to the non-additive contributions (energy not accounted for in ΔV_2) to the interaction energies of trimers, and ΔV_4 and higher-order terms are defined analagously.

The utility of the MBE comes from the fact that, aside from many-body polarization, for which the complete N-body effects can readily be calculated,[54,77] the MBE is typically rapidly convergent, and often only ΔV_2 and ΔV_3 terms are

required to completely and accurately describe V_N.[54,78] In fact, the combination of ΔV_2 and N-body polarization often account for upwards of 90–95% of the total interaction energy,[4,78] such that the accuracy of a given ab initio force field depends primarily on the accuracy of the pair potential itself. When required, explicit terms for ΔV_3 can easily be added to an ab initio force field as an additive correction, and accurate models for ΔV_3 have been outline in previous work.[4] Nevertheless, we can usually restrict our focus to the development of accurate models for ΔV_2, with the knowledge that accuracy in describing ΔV_2 will have a direct effect on accuracy with respect to V_N and/or experiment.

1.5.2 SAPT

Having limited our attention to modeling the pair potential, ΔV_2, a second technique we can employ in the development of ab initio force fields is to fit our force field parameters on a component-by-component basis to a physically-meaningful Energy Decomposition Analysis (EDA) of dimer interaction energies. Force field fitting on a component-by-component basis enables the following:

1. By increasing the amount of ab initio data used in the force field fits, we reduce the possibility of overfitting the potential, which in turn enables transferability[75]

2. By enforcing a one-to-one correspondence between force field functional forms and benchmark ab initio energies, we reduce reliance on error cancellation and ensure that all fitted parameters describe the intended physical feature

3. By evaluating the resulting fits on a component-by-component basis, we can directly relate errors in the potential to errors in the individual energy comopnents, thus providing insight into how a current model might be improved

In most cases (with the work described in Chapter 4 being an exception), we use Symmetry-Adapted Perturbation Theory (SAPT) as our EDA of choice. SAPT, and

DFT-SAPT in particular (a variant of SAPT based on a DFT-based description of monomers, which scales reasonably as N^5 with respect to the number of electrons in the system), serves as an accurate yet affordable approximation to the gold-standard CCSD(T) calculations discussed earlier. Theories and formalisms for SAPT are reviewed in Ref. 79–81, and a variety of examples of SAPT-based ab initio force field development is given in Ref. 75, 82. Overall, ab inito force fields fit to DFT-SAPT energies have been shown to lead to good accuracy in experimental property predictions,[82,83] thus justifying our approach. Furthermore, and as is especially important in the development of *transferable* ab initio force fields, SAPT provides a natural and physically-meaningful decomposition into energy components of electrostatics, exchange, induction, and dispersion. By fitting these energy terms on a component-by-component basis, the SAPT energy decomposition can be fully taken advantage of to yield (as in Ref. 83) a library of accurate and transferable force field parameters with broad applicability to a range of chemical and physical environments.

1.5.3 ISA-Distributed Multipole Analysis (DMA)

As a complement to SAPT-based fitting approaches, we and others have shown how distributed molecular property calculations can be used to obtain select force field parameters in a transferable manner without recourse to direct fitting.[54,84,85] By distributed, we mean here that the results of the molecular calculations, which yield properties for individaul monomers, can be decomposed into meaningful atom-in-molecule contributions for use in the pair potential, ΔV_2. Historically, many of these distribution schemes have been based on a Distributed Multipole Analysis (DMA), and methods for obtaining distributed multipolar electrostatic,[86,87] polarization,[88] and dispersion[83,89,90] parameters from DMA are well-documented.[54,81]

More recently, Misquitta et al.[91] has, in conjunction with important contributions from Lillestolen and Wheatley,[92,93] built upon the existing class of Hirshfeld[94] atom-in-molecule charge partitioning schemes to develop a new distribution scheme based on Iterated Stockholder Atoms (ISA), termed ISA-DMA. In brief, ISA-DMA

operates by partitioning a monomer electron density into atomic contributions,

$$\rho_i(\mathbf{r}) = \rho_I(\mathbf{r}) \frac{w_i(\mathbf{r})}{\sum_{a \in I} w_a(\mathbf{r})}, \tag{1.7}$$

where $\rho(\mathbf{r})$ is an electron density, $w_a(\mathbf{r})$ is a spherically-symmetric weight function which is iteratively determined in the course of the ISA analysis, and lower- and upper-case subscripts represent, respectively, atomic or molecular quantities.[91] Using these atom-in-molecule charge densities, recent work has shown how new and/or improved parameters for multipolar electrostatics (Chapter 5 and Ref. 91), exchange-repulsion (Chapter 2 and Ref. 95), and dispersion (Chapter 5) are now possible. Notably, ours is not the only group to take advantage of the ISA-DMA or related methods,[96] and a number of ab intio force fields have been developed with the aid of such distribution schemes.[85,97–99] Regardless, the ISA-DMA parameters have shown good promise for the development of accurate and transferable force fields, and the manner in which we can include these parameters in force field development will be a main focus of this dissertation.

1.6 Outline

Having described the utility of molecular simulation and the important goals of accuracy and transferabilty in force field development, the purpose of this dissertation is to describe new and better methods for obtaining functional forms and parameters that will lead to improved ab initio force field development. To this end, Chapter 2 describes an approach whereby the ISA partitioning scheme can be used to develop new models for the SAPT exchange-repulsion energy and related short-range energy contributions. The methods in Chapter 2 neglect important effects due to the orientation dependence, or 'atomic-level anisotropy', of the ISA charge densities, and Chapter 3 extends the original ISA-based method to account for this atomic-level anisotropy in an accurate cost-efficient manner amenable to large-scale molecular simulation. As a third investigation of methods development

for ab inito force fields, Chapter 4 explores the ways in which additional EDA methods (aside from SAPT) can be used to benchmark and parameterize ab initio force fields in cases where SAPT itself is in error. Finally, in the course of our research we have developed many automated tools and best practices for ab initio force field development (SAPT-based or otherwise), and these practical considerations are the subject of Chapters 5 and 6. Overall conclusions and avenues for future research are the subject of Chapter 7.

Part I

Published Work

2 BEYOND BORN–MAYER: IMPROVED MODELS FOR SHORT-RANGE REPULSION IN AB INTIO FORCE FIELDS

2.1 Introduction

Molecular simulation is an essential tool for interpreting and predicting the structure, thermodynamics, and dynamics of chemical and biochemical systems. The fundamental inputs into these simulations are the intra- and intermolecular force fields, which provide simple and computationally efficient descriptions of molecular interactions. Consequently, the predictive and explanatory power of molecular simulations depends on the fidelity of the force field to the underlying (exact) potential energy surface.

In the case of intermolecular interactions, the dominant contributions for non-reactive systems can be decomposed into the following physically-meaningful energy components: electrostatic, exchange-repulsion, induction and dispersion.[54,78,100–102] At large intermolecular distances, where monomer electron overlap can be neglected, the physics of intermolecular interactions can be described entirely on the basis of monomer properties (e.g. multipole moments, polarizabilities), all of which can be calculated with high accuracy from first principles.[103] In conjunction with associated distribution schemes that decompose molecular monomer properties into atomic contributions,[54,78,87–89,91,104] these monomer properties lead to an accurate and computationally efficient model of 'long-range' intermolecular interactions as a sum of atom-atom terms, which can be straightforwardly included in common molecular simulation packages.

At shorter separations, where the molecular electron density overlap cannot be neglected, the asymptotic description of intermolecular interactions breaks down due to the influence of Pauli repulsion, charge penetration and charge transfer. These effects can be quantitatively described using modern electronic structure methods,[79,80,101,105,106] but are far more challenging to model accurately using computationally inexpensive force fields. For efficiency and ease of parameterization, most

simple force fields use a single 'repulsive' term to model the cumulative influence of (chemically distinct) short-range interactions. These simple models have seen comparatively little progress over the past eighty years, and the Lennard-Jones[107] (A/r^{12}) and Born-Mayer[108,109] ($A\exp(-Br)$) forms continue as popular descriptions of short-range effects in standard force fields despite some well-known limitations (*vide infra*).

Because the prediction of physical and chemical properties depends on the choice of short-range interaction model,[60–62,110–120] it is essential to develop sufficiently accurate short-range force fields. This is particularly true in the case of ab initio force field development. A principle goal of such a first-principles approach is the reproduction of a calculated potential energy surface (PES), thus (ideally) yielding accurate predictions of bulk properties.[75] Substantial deviations between a fitted and calculated PES lead to non-trivial challenges in the parameterization process, which in turn can often degrade the quality of property predictions. The challenge of reproducing an ab initio PES becomes particularly pronounced at short inter-molecular separations, where many common force field functional forms are insufficiently accurate. For example, the popular Lennard-Jones (A/r^{12}) functional form is well-known to be substantially too repulsive at short contacts as compared to the exact potential.[60–62,65,66] While the Born-Mayer ($A\exp(-Br)$) functional form is more physically-justified[109] and fares better in this regard,[65] substantial deviations often persist.[121] In addition, parameterization of the Born-Mayer form is complicated by the strong coupling of the pre-exponential (A) and exponent (B) parameters, hindering the transferability of the resulting force field. These considerations, along with the observed sensitivity of structural and dynamic properties to the treatment of short-range repulsion,[110] highlight the need for new approaches to model short-range repulsive interactions.

Our primary goal in this Chapter is to derive a simple and accurate description of short-range interactions in molecular systems that improves upon both the standard Lennard-Jones and Born-Mayer potentials in terms of accuracy, transferability, and ease of parameterization. Our focus is on ab initio force field development, and thus we will use the fidelity of a given force field with respect to an accurate

ab initio PES as a principle metric of force field quality. We note that other metrics may be more appropriate for the development of empirical potentials, where Lennard-Jones or Born-Mayer forms may yield highly accurate 'effective' potentials when parameterized against select bulk properties. Nonetheless, we anticipate that the models proposed in this Chapter may prove useful for empirical force field development in cases where a more physically-motivated functional form is necessary.[60–62]

The outline of this Chapter is thus as follows: first, we derive a new functional form capable of describing short-range repulsion from first principles, and show how the standard Born-Mayer form follows as an approximation to this more exact model. Our generalization of the Born-Mayer functional form allows for an improved description of a variety of short-range effects, namely electrostatic charge penetration, exchange-repulsion, and density overlap effects on induction and dispersion. Crucially, we also demonstrate how the associated atomic exponents can be extracted from first-principles monomer charge densities via an iterated stockholder atoms (ISA) density partitioning scheme, thereby reducing the number of required fitting parameters compared to the Born-Mayer model. Benchmarking this 'Slater-ISA' methodology (functional form and atomic exponents) against high-level ab initio calculations and experiment, we find that the approach exhibits increased accuracy, transferability, and robustness as compared to a typical Lennard-Jones or Born-Mayer potential. In addition, we show how the ISA-derived exponents can be adapted for use within the standard Born-Mayer form (Born-Mayer-sISA), while still retaining retaining many of the advantages of the Slater-ISA approach. As such, our methodology also offers an opportunity to dramatically simplify the development of both empirically-parameterized and ab initio simulation potentials based upon the standard Born-Mayer form.

2.2 Theory

We begin with a formal treatment of the overlap model for the exchange-repulsion between two isolated atoms, and then extend these results to develop a general-

ized model for the short-range interactions in both atomic and molecular systems. Finally, we show how the conventional Born-Mayer model can be derived as an approximation to this more rigorous treatment.

2.2.1 Models for the exchange-repulsion between isolated atoms

It is well known that the exchange-repulsion interaction between two closed-shell atoms i and j is proportional, or very nearly proportional, to the overlap of their respective charge densities:[122]

$$E_{ij}^{exch} \approx V_{ij}^{exch} = K_{ij}(S_{\rho}^{ij})^{\gamma} \tag{2.1}$$

$$S_{\rho}^{ij} = \int \rho_i(\mathbf{r})\rho_j(\mathbf{r})d^3\mathbf{r}. \tag{2.2}$$

Here and throughout, we use E to denote quantum mechanical energies, and V to denote the corresponding model/force field energies. Recently two of us have provided a theoretical justification for this repulsion hypothesis (or overlap model), and have shown that $\gamma = 1$ provided that asymptotically-correct densities are used to compute both the atomic densities and E_{ij}^{exch}.[54,123] As this is the case for the calculations in this work, we assume $\gamma = 1$ throughout the Chapter.

The overlap model has frequently been utilized in the literature and has been found to yield essentially quantitative accuracy for a wide variety of chemical systems.[122,124,125] Prior work exploiting the overlap model has generally followed one of two strategies. Striving for quantitative accuracy, several groups have developed approaches to evaluate Eq. (2.2) via either numerical integration or density fitting of ab-initio molecular electron densities, ρ_i (e.g. SIBFA, GEM, effective fragment potentials).[126? –134] These force fields, while often extremely accurate, lack the simple closed-form analytical expressions that define standard force fields (such as the Lennard-Jones or Born-Mayer models) and thus are often much more computationally expensive than conventional models.

In contrast, and similar to our objectives, the overlap model has also been used in the development of standard force fields. In this case, the molecular electron

density as well as the overlap itself is drastically simplified in order to yield a simple closed-form expression that can be used within a conventional molecular simulation package.[122,124,125] As we show below, the Born-Mayer model can be 'derived' via such an approach. At the expense of some accuracy, the resulting overlap-based force fields exhibit high computational efficiency and employ well-known functional forms.

Building on this prior work, our present goal is to derive rigorous analytical expressions and improved approximations for both ρ_i and Eq. (2.2), facilitating the construction of ab initio force fields that exhibit simplicity, high computational efficiency, fidelity to the underlying PES, and (with only trivial modifications) compatibility with standard simulation packages. We first start with the case of isolated atoms, where it is well-known that the atomic electron density decays asymptotically as

$$\rho_{r\to\infty}(r) \propto r^{2\beta} e^{-2\alpha r} \tag{2.3}$$

where the exponent $\alpha = \sqrt{2I}$ is fixed by the vertical ionization potential I, $\beta = -1 + \frac{Q}{\alpha}$, and $Q = Z - N + 1$ for an atom with nuclear charge $+Z$ and electronic charge $-N$.[123,135–137] The exponential term dominates the asymptotic form of the density, and the r-dependent prefactor may be neglected[91,124,125,138]. In this case, the density takes the even simpler form

$$\rho_{r\to\infty}(r) \approx D e^{-Br}, \tag{2.4}$$

where D is a constant that effectively absorbs the missing r-dependent pre-factor and B is an exponent that is now only approximately equal to 2α.

In the case of two identical atoms, substitution into Eq. (2.2) yields a simple

expression for the density overlap, S_ρ,[11,12]

$$\begin{aligned} S_\rho^{ii} &= \frac{\pi D^2}{B^3} P(Br_{ii}) \exp(-Br_{ii}) \\ P(Br_{ii}) &= \frac{1}{3}(Br_{ii})^2 + Br_{ii} + 1 \end{aligned} \tag{2.5}$$

as well as (via Eq. (2.1)) the exchange-repulsion energy[125,139]:

$$V_{ii}^{exch} = A_{ii}^{exch} P(Br_{ii}) \exp(-Br_{ii}). \tag{2.6}$$

Here, r_{ii} represents an interatomic distance, and A_{ii}^{exch} indicates a proportionality constant that is typically fit to calculated values of the exchange-repulsion energy. The only approximations thus far are the use of the overlap model and the simplified asymptotic form of the atomic charge density.

For the general case of two hetero-atoms, substitution of Eq. (2.4) into Eq. (2.2) yields the more complicated expression[11,12]

$$\begin{aligned} S_\rho^{ij} =& \frac{16\pi D_i D_j \exp(-\{B_i + B_j\} r_{ij}/2)}{(B_i^2 - B_j^2)^3 r_{ij}} \times \\ & \Bigg[\left(\frac{B_i - B_j}{2}\right)^2 \left(\exp\left(\{B_i - B_j\}\frac{r_{ij}}{2}\right) - \exp\left(-\{B_i - B_j\}\frac{r_{ij}}{2}\right) \right) \\ & \quad \times \left(\left(\frac{B_i + B_j}{2}\right)^2 r_{ij}^2 + (B_i + B_j) r_{ij} + 2 \right) \\ & - \left(\frac{B_i + B_j}{2}\right)^2 \exp\left(\{B_i - B_j\}\frac{r_{ij}}{2}\right) \times \left(\left(\frac{B_i - B_j}{2}\right)^2 r_{ij}^2 - (B_i - B_j) r_{ij} + 2 \right) \\ & + \left(\frac{B_i + B_j}{2}\right)^2 \exp\left(-\{B_i - B_j\}\frac{r_{ij}}{2}\right) \times \left(\left(\frac{B_i - B_j}{2}\right)^2 r_{ij}^2 + (B_i - B_j) r_{ij} + 2 \right) \Bigg], \end{aligned} \tag{2.7}$$

which is too cumbersome to serve as a practical force field functional form. However, since the above expression reduces to Eq. (2.5) in the limit $B_i = B_j$, and because $|B_i -$

$B_j|$ is small for most atom pairs, we have found that Eq. (2.7) may be approximated using Eq. (2.5) with an *effective* atomic exponent B. An expansion of Eq. (2.7) about $B_i = B_j$ suggests that this effective exponent should be given by the arithmetic mean, $B_{ij} = \frac{1}{2}(B_i + B_j)$. However, a Waldman-Hagler style analysis[140] (Section 2.A) suggests instead that a more suitable exponent is given by the geometric mean combination rule,

$$B = B_{ij} \equiv \sqrt{B_i B_j}. \tag{2.8}$$

As shown in the Supporting Information of Ref. 95, this approximate overlap model (Eq. (2.5) and Eq. (2.8)) is of comparable accuracy to the exact overlap from Eq. (2.7). Thus the density overlap and (force field) exchange energies of arbitrary hetero-atoms take the simple forms

$$S_\rho^{ij} = D_{ij} P(B_{ij}, r_{ij}) \exp(-B_{ij} r_{ij}) \tag{2.9}$$

$$D_{ij} = \pi D_i D_j B_{ij}^{-3} \tag{2.10}$$

$$P(B_{ij}, r_{ij}) = \frac{1}{3}(B_{ij} r_{ij})^2 + B_{ij} r_{ij} + 1 \tag{2.11}$$

and

$$V_{ij}^{\mathrm{exch}} = A_{ij}^{exch} P(B_{ij} r_{ij}) \exp(-B_{ij} r_{ij}). \tag{2.12}$$

Due to the connection with the overlap between two s-type Slater orbitals, we refer to Eq. (2.12) as the Slater functional form. Note that this expression reduces to the standard Born-Mayer function by making the further approximation $P(B_{ij} r_{ij}) = 1$, although it is known[125,141] that this is a poor approximation with the B_{ij} as defined above. Instead, as we shall demonstrate in Section 2.4, the exponents B_{ij} need to be scaled for accurate use with a Born-Mayer functional form.

Variants of the polynomial pre-factor from Eq. (2.9) have previously been recognized and used in intermolecular interaction models.[109,139,141] Early work by Buckingham[109] hypothesized that the functional form of Eq. (2.12) would be more

accurate than the Born-Mayer form, though no attempt was made to provide a closed-form expression for P. More recent potentials have incorporated a low-order polynomial into the exchange repulsion term, either by direct parameterization[142–146] or indirectly by fitting the exchange to S_ρ/r^2 rather than to S_ρ itself.[124,125,147] Kita et al. have derived (but not tested) Eq. (2.6) for the homoatomic case.[147] Recently, and most similar to the spirit of the present Chapter, York and coworkers have derived a model based upon the overlap of Slater-type orbitals for use in QM/MM simulations, yielding an expression identical to Eq. (2.7).[148–150] Those authors treated D_i and D_j as empirical fitting parameters and estimated atomic exponents (B_i and B_j) via atomic-charge dependent functions. In contrast, we will demonstrate that utilization of the far simpler functional form from Eq. (2.12), in conjunction with exponents calculated from analysis of the first-principles molecular electron density, yields much higher computational efficiency and simplifies the parameterization process without significant loss of accuracy.

For an arbitrary pair of interacting atoms, A_{ij}^{exch} can be obtained by fitting to calculated exchange-repulsion energies. However, assuming that the overlap proportionality factor K_{ij} is a universal constant (or, alternatively, separable with $K_{ij} = K_i K_j$), then

$$A_{ij}^{exch} = \left(K_i \sqrt{\frac{\pi}{B_i^3}} D_i \right) \left(K_j \sqrt{\frac{\pi}{B_j^3}} D_j \right) \equiv A_i^{exch} A_j^{exch}, \tag{2.13}$$

thus providing a combination rule for heteroatomic interaction in terms of purely atomic quantities. The universality and separability of K_{ij} are, at present, empirically rather than theoretically justified.[54,151,152] The A_i^{exch} can then be obtained, for example, by a straightforward fitting of calculated ab initio *homoatomic* exchange-repulsion energies.

2.2.2 Models for other short-range interactions between isolated atoms

Beyond the exchange-repulsion, the density-overlap model may also be used to model other short-range interaction components, such as the electrostatic charge penetration energy and the short-range induction and dispersion energies (that is, the portion modulated by charge overlap). Indeed, it has been demonstrated that the electrostatic charge penetration energy is approximately proportional to the exchange-repulsion energy, and consequently to the charge density overlap,[54,91] which has provided a successful basis for modeling the electrostatic charge penetration energy.[83,153] While the relation between short-range induction and charge overlap is less clear, recent results have demonstrated that the charge-transfer energy, which is the dominant short-range component of the induction energy,[154] is approximately proportional to the first-order exchange energy,[123,155] and prior work has successfully used the overlap hypothesis to describe the short-range induction.[54,83,153] We therefore model the electrostatic charge penetration and short-range induction interactions as

$$V_{ij}^{pen} = A_{ij}^{pen} P(B_{ij}, r_{ij}) \exp(-B_{ij} r_{ij}) \tag{2.14}$$

and

$$V_{ij}^{ind,sr} = A_{ij}^{ind} P(B_{ij}, r_{ij}) \exp(-B_{ij} r_{ij}). \tag{2.15}$$

Aside from the pre-factors A_{ij}, these expressions are identical to that for the exchange-repulsion term.

The behavior of the dispersion interaction at short distances poses a special challenge. In order to model the short-range dispersion and to resolve the unphysical, mathematical divergence of the $1/r^n$ terms as $r \to 0$, Tang and Toennies have shown that the terms in the dispersion expansion should be damped using an

appropriate incomplete gamma function

$$f_n(x) = 1 - e^{-x} \sum_{k=0}^{n} \frac{(x)^k}{k!} \tag{2.16}$$

$$x = -\frac{d}{dr} \left[\ln V^{exch}(r) \right] \; r \tag{2.17}$$

that accounts for both exchange and charge penetration effects.[156,157] Note that the form of this damping factor depends on the model used for exchange repulsion. For the Slater functional form (Eq. (2.12)),

$$x_{Slater} = B_{ij} r_{ij} - \frac{2B_{ij}^2 r_{ij} + 3B_{ij}}{B_{ij}^2 r_{ij}^2 + 3B_{ij} r_{ij} + 3} r_{ij}. \tag{2.18}$$

Alternatively, if we replace the Slater functional form with the less accurate Born-Mayer expression, x simplifies to the result originally given by Tang and Toennies:

$$x_{Born\text{-}Mayer} = B_{ij} r_{ij}. \tag{2.19}$$

2.2.3 Models for short-range interactions between molecules

The overlap repulsion hypothesis can be extended to molecules[54,151,158–160] by writing the molecular density ρ_I as a superposition of atomic densities

$$\rho_I(\mathbf{r}) = \sum_{i \in I} \rho_i(\mathbf{r}) \tag{2.20}$$

where i represents an atom in molecule I. In this case,

$$V_{IJ}^{exch} = \sum_{i \in I} \sum_{j \in J} V_{ij}^{exch} \tag{2.21}$$

$$V_{ij}^{exch} = K_{ij} S_{\rho}^{ij} = \int \rho_i(\mathbf{r}) \rho_j(\mathbf{r}) d^3\mathbf{r}. \tag{2.22}$$

Note that the form of Eq. (2.22) is identical to the corresponding expression between isolated atoms, but requires partitioning of the molecular charge density into atom-in-molecule densities, ρ_i, each decaying according to an effective atom-in-molecule density decay exponent, B_i.

In principle, such atom-in-molecule exponents could be estimated from the ionization potentials of the corresponding isolated atoms,[83,139] but this approach neglects the influence of the molecular environment. A more appealing possibility is to directly evaluate the atom-in-molecule densities via partitioning of the calculated monomer densities. Density partitioning has not yet (to our knowledge) been applied in the context of the overlap model to solve for Eq. (2.22), however several successful efforts in force field development have recently relied on an atoms-in-molecule approach in order to obtain accurate scaling relationships for intermolecular force field parameters.[161–163] In particular, Cole et al. utilized a density-derived electrostatic and chemical (DDEC) partitioning scheme[164,165] to generate Lennard-Jones dispersion and short-range repulsion parameters, though the latter parameters were calculated implicitly by enforcing the coincidence of the potential minimum and the calculated atomic radius.

While no unique atom-in-molecule density partitioning scheme exists, an ideal approach should yield atom-in-molecule densities that strongly resemble those of isolated atoms, e.g. maximally spherical and asymptotically exponential.[91,166–168] The recently developed iterated stockholder partitioning of Lillestolen and Wheatley obeys this first important constraint of sphericity.[92,93] As a non-trivial extension of the original Hirshfeld method,[94] iterated stockholder atom (ISA) densities are defined as

$$\rho_i(\mathbf{r}) = \rho_I(\mathbf{r}) \frac{w_i(\mathbf{r})}{\sum_{a \in I} w_a(\mathbf{r})} \tag{2.23}$$

where the converged shape functions $w_i(\mathbf{r})$ are spherical averages of the atomic

densities $\rho_i(\mathbf{r})$:

$$w_i(\mathbf{r}) = \langle \rho_i(\mathbf{r}) \rangle_{\text{sph}}. \tag{2.24}$$

This formulation ensures, by construction, that the sum of atomic densities reproduces the overall molecular density. Furthermore, the maximally spherical nature of the atom-in-molecule densities naturally facilitates a description of short-range interactions via a simple isotropic site-site model.

Misquitta et al. have developed a rapidly convergent implementation of the ISA procedure (BS-ISA[91]) using a basis set expansion which, in addition to exhibiting good convergence with respect to basis set, also leads to asymptotically-exponential atomic densities. Consequently, the BS-ISA method is our preferred density partitioning scheme. As an example, the spherically-averaged atomic densities for acetone are shown in Fig. 2.1. For simplicity, and because a full treatment of the anisotropy is beyond the scope of this Chapter, we subsequently refer to the spherically-averaged atomic densities (i.e. the shape functions, $w_i(r)$) as atomic or atom-in-molecule densities.

From Fig. 2.1 we see that the ISA atomic shape functions (that is, the spherically-averaged ISA atoms-in-molecule density) decay exponentially outside the core region. However, note that the exponents governing the spherical density decay, B_i^{ISA}, differ from those of the free atoms. The ISA densities have been observed to account for electron movement in the molecule, and the consequent density changes brought about by this movement tend to be manifested in the region of the density tails.[91] The ISA exponents can be obtained by a weighted least-squares fit to the BS-ISA atomic density (see Section 2.3 for details), with the resulting fitted atomic densities shown in Fig. 2.1. Note that even a single exponential is remarkably successful in reproducing the entirety of the valence atomic density.

Given these fitted ISA exponents, we can now apply our short-range interaction

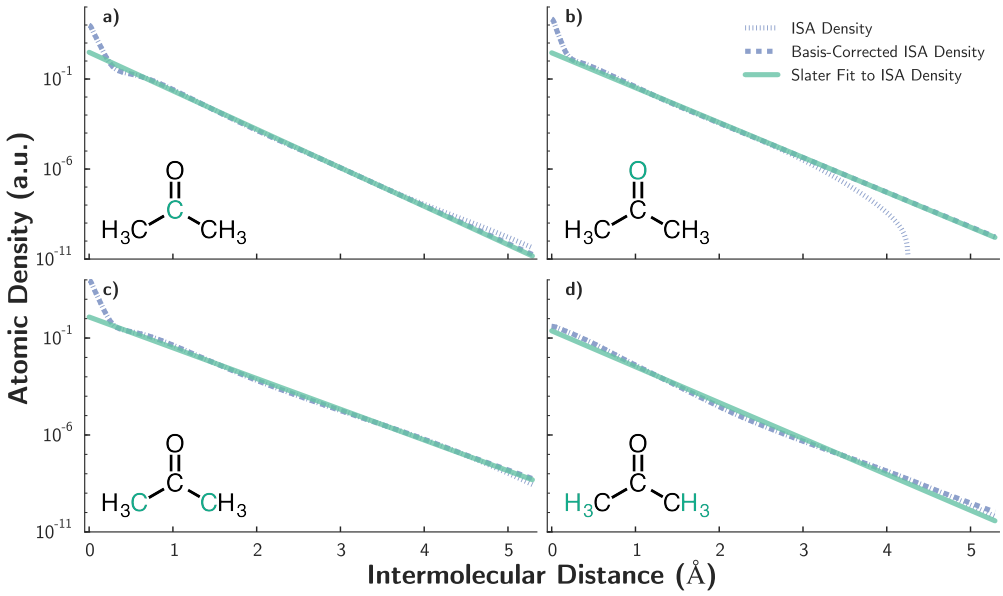


Figure 2.1: BS-ISA and fitted shape functions for each atom type in acetone: a) carbonyl carbon, b) oxygen, c) methyl carbon, d) hydrogen. BS-ISA shape functions (dotted line) for each atom type have been obtained at a PBE0/aug-cc-pVTZ level of theory. A modified BS-ISA shape function (dashed line) corrects the tail-region of the BS-ISA function to account for basis set deficiencies in the BS-ISA algorithm. A single Slater orbital of the form $D_i^{ISA} \exp(-B_i^{ISA} r)$ (solid line) is fit to the basis-corrected BS-ISA shape function, and the obtained B_i^{ISA} value is used as an atomic exponent in the functional form of Aniso-Iso FF. Results for acetone are typical of molecules studied in this Chapter.

formalism to polyatomics,

$$
\begin{aligned}
V^{sr} &= \sum_{ij} A_{ij}^{sr} P(B_{ij}, r_{ij}) \exp(-B_{ij} r_{ij}) \\
P(B_{ij}, r_{ij}) &= \frac{1}{3}(B_{ij} r_{ij})^2 + B_{ij} r_{ij} + 1 \\
A_{ij}^{sr} &= A_i^{sr} A_j^{sr} \\
B_{ij} &= \sqrt{B_i^{ISA} B_j^{ISA}}
\end{aligned}
\tag{2.25}
$$

where the molecular short-range energy is now a sum of atom-atom contributions.

In conjunction with appropriately damped atomic dispersion (Eqs. (2.16) and (2.18)), Eq. (2.25) completely defines our new short-range force field. We refer to this new functional form and set of atomic exponents as the Aniso-Iso FF.

2.3 Computational Methods

To evaluate the Slater-ISA FF against conventional Born-Mayer and/or Lennard-Jones models, we compare the ability of each resulting short-range force field to reproduce benchmark ab initio intermolecular interaction energies for a collection of representative dimers. Such a metric is directly relevant for ab initio force field development. Even for an empirically-parameterized force field, however, fidelity to an accurate ab initio potential should be well correlated with the highest level of accuracy and transferability achievable with a given short-range methodology.

We have developed the Slater-ISA FF, Born-Mayer, and Lennard-Jones force fields using benchmark energies calculated using the symmetry-adapted perturbation theory based on density-functional theory (DFT-SAPT or SAPT(DFT)[169–177]). DFT-SAPT provides interaction energies that are comparable in accuracy to those from CCSD(T) and which are rigorously free from basis set superposition error.[101,178] Additionally, at second-order, DFT-SAPT also provides an explicit interaction energy decomposition into physically-meaningful contributions: the electrostatic, exchange-repulsion, induction, and dispersion energies. This decomposition is vital to the development of models as it allows the development of separate terms for each type of short-range interaction. Terms of third and higher order are estimated using the δHF correction[179] which contains mainly higher-order induction contributions. Following prior work,[83,166] and for the purposes of fitting to the DFT-SAPT data, we keep the second-order induction term and the δHF term separate.

Since the Slater-ISA and Born-Mayer force fields describe only short-range interactions (i.e. those terms which are modulated by the overlap of the monomer electron densities), they must both be supplemented with additional long-range terms that describe the electrostatic, polarization, and dispersion interactions. Here

we have chosen a long-range potential of the form

$$V_{lr} = V_{multipole} + V_{dispersion} + V_{pol} \tag{2.26}$$

where

$$V_{multipole} = \sum_{ij} \sum_{tu} Q_t^i T_{tu} Q_u^j \tag{2.27}$$

includes distributed multipole contributions from each atom up to quadrupoles,

$$V_{dispersion} = -\sum_{ij} \sum_{n=3}^{6} \frac{C_{ij,2n}}{r_{ij}^{2n}} \tag{2.28}$$

describes isotropic dispersion, and V_{pol} is the polarization energy modeled by Drude oscillators[180,181] as in Ref. 83. The accuracy of each of these terms is expected to minimize errors in the long-range potential, simplifying the comparison between short-range force field functional forms. Nonetheless, we expect that our results will be qualitatively insensitive to the particular choice of long-range force field and acknowledge that simpler alternatives may be preferred for the development of highly efficient simulation potentials. In the case of the Lennard-Jones force field, we replace Eq. (2.28) with the simple $C_{ij,6}/r_{ij}^6$ dispersion term that is standard to the Lennard-Jones model.

We used a test set consisting of one atom (argon) and 12 small organic molecules (see Fig. 2.2) from which dimer potentials could be generated (we will use the term 'dimer' to mean two, potentially dissimilar, interacting molecules or atoms), yielding 91 dimer combinations (13 homo-monomeric, 78 hetero-monomeric). This wide range of systems allowed us to evaluate both the accuracy and transferability of the Slater-ISA model compared to conventional Born-Mayer and/or Lennard-Jones models.

A detailed description of this overall methodology is provided below.

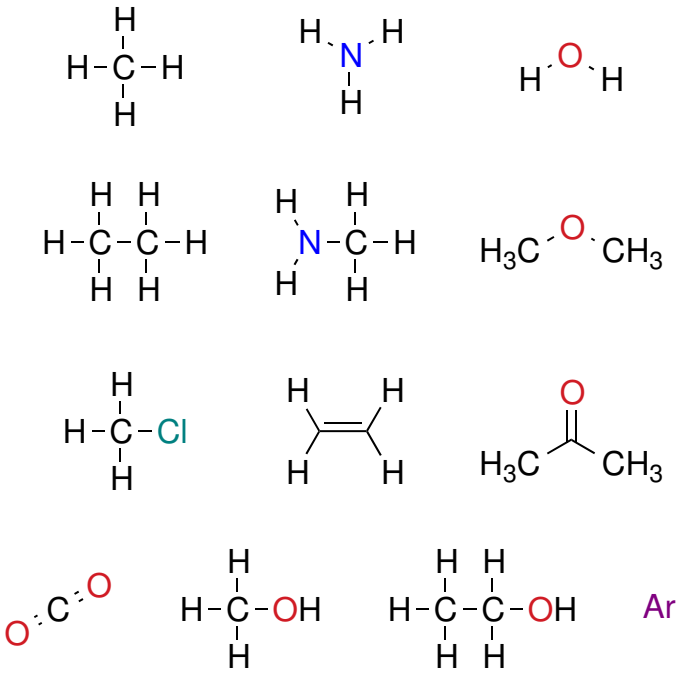


Figure 2.2: The 13 small molecules included in the 91 dimer (13 homomonomeric, 78 heteromonomeric) test set. Cartesian geometries for all of these molecules are given in Section A.1.

2.3.1 Construction of the 91 dimer test set

Monomer geometries for each of the 13 small molecules were taken from the experimental NIST [CCCBDB] database[182] and can be found in Section A.1. For acetone and methyl amine, experimental geometries were unavailable, and thus the computational NIST [CCCBDB] database was used to obtain geometries at a high level of theory (B3LYP/aVTZ for acetone, CCSD(T)/6-311G* for methyl amine). For each of the 91 dimers, a training set was constructed using DFT-SAPT (PBE0/AC) interaction energies calculated at 1000 quasi-random dimer configurations. These configurations were generated using Shoemake's algorithm,[183] subject to the constraint that the nearest atom pairs be separated by between 0.75 and 1.3 of the sum of their van der Waals radii. This ensured adequate sampling of the potential energy surface in the region of the repulsive wall. The DFT-SAPT interaction energies

were evaluated using an asymptotically corrected PBE0 functional (PBE0/AC) with monomer vertical (first) ionization potentials computed using the Δ-DFT approach at a PBE0/aVTZ level of theory. Unless otherwise noted, all DFT-SAPT calculations used an aVTZ basis set in the dimer-centered form with midbond functions (the so-called DC+ form), and were performed using the MOLPRO2009 software suite.[184] The midbond set consisted of a 5s3p1d1f even-tempered basis set with ratios of 2.5 and centered at $\zeta = 0.5, 0.5, 0.3, 0.3$ for the s,p,d, and f shells, respectively. This set was placed near the midpoint of the centers of mass of the two interacting monomers.

A small fraction of DFT-SAPT calculations exhibited unphysical energies, which were attributed to errors in generating the optimized effective potential used during the DFT-SAPT (PBE0/AC) calculations; these points were removed from the test set.

2.3.2 BS-ISA Calculations

BS-ISA atomic densities were obtained using CamCASP 5.8[84,185,186] following the procedure of Misquitta et al.[91] For the BS-ISA calculations, an auxiliary basis was constructed from an RI-MP2 aVTZ basis set with *s*-functions replaced by the ISA-set2 supplied with the CamCASP program; CamCASP's ISA-set2 basis was also used for the ISA basis set.[91] A custom ISA basis set for Ar was used (even tempered, $n_{min} = -2, n_{max} = 8$)[91] as no published basis was available. BS-ISA calculations were performed with the A+DF algorithm, which allows the ISA functional to be mixed with some fraction, ζ, of the density-fitting functional. Following the recommendations of Misquitta et al.[91], we have used $\zeta = 0.1$ for the multipole moment calculations, and $\zeta = 0.9$ for the density partitioning used to determine the B_{ij} coefficients.

2.3.3 Determination of $B_i^{\mathbf{ISA}}$

The BS-ISA-derived atomic exponents, B_i^{ISA}, were obtained from a weighted least-squares fit to the spherically averaged BS-ISA atomic densities (shape functions),

$w_i(\mathbf{r})$. In some cases, numerical instabilities and basis-set limitations of the BS-ISA procedure yielded densities that exhibited non-exponential asymptotic behavior.[91] To correct for these unphysical densities, we extrapolated the exponential decay of the valence region to describe the BS-ISA tails also. Details of this procedure can be found in Section 5.B. The ISA atom-in-molecule exponents were then derived via a log-weighted fit to the tail-corrected shape-functions $w^a(\mathbf{r})$ for densities within the cutoff $10^{-2} > w^a > 10^{-20}$ a.u. This region was chosen to reproduce the charge density most accurately in the valence regimes most likely to be relevant to intermolecular interactions.

2.3.4 Force Field Functional Forms and Parameterization

The general structure of the force fields V_{FF} for both the Slater-ISA FF and the Born-Mayer-type models are given by the following equations:

$$
\begin{aligned}
V_{FF} &= \sum_{ij} V_{ij}^{exch} + V_{ij}^{elst} + V_{ij}^{ind} + V_{ij}^{\delta^{HF}} + V_{ij}^{disp} \\
V_{ij}^{exch} &= A_{ij}^{exch} P(B_{ij}, r_{ij}) \exp(-B_{ij} r_{ij}) \\
V_{ij}^{elst} &= -A_{ij}^{elst} P(B_{ij}, r_{ij}) \exp(-B_{ij} r_{ij}) + \sum_{tu} Q_t^i T_{tu} Q_u^j \\
V_{ij}^{ind} &= -A_{ij}^{ind} P(B_{ij}, r_{ij}) \exp(-B_{ij} r_{ij}) + V_{pol}^{(2)} \\
V_{ij}^{\delta^{HF}} &= -A_{ij}^{\delta^{HF}} P(B_{ij}, r_{ij}) \exp(-B_{ij} r_{ij}) + V_{pol}^{(3-\infty)} \\
V_{ij}^{disp} &= -\sum_{n=3}^{6} f_{2n}(x) \frac{C_{ij,2n}}{r_{ij}^{2n}} \\
A_{ij} &= A_i A_j \\
C_{ij,2n} &= \sqrt{C_{i,n} C_{j,n}} \\
f_{2n}(x) &= 1 - e^{-x} \sum_{k=0}^{2n} \frac{(x)^k}{k!}
\end{aligned}
\tag{2.29}
$$

For the Slater-ISA FF:

$$\begin{aligned} B_i &= B_i^{\mathrm{ISA}} \\ B_{ij} &= \sqrt{B_i B_j} \\ P(B_{ij}, r_{ij}) &= \frac{1}{3}(B_{ij} r_{ij})^2 + B_{ij} r_{ij} + 1 \\ x &= B_{ij} r_{ij} - \frac{2B_{ij}^2 r_{ij} + 3B_{ij}}{B_{ij}^2 r_{ij}^2 + 3B_{ij} r_{ij} + 3} r_{ij} \end{aligned} \tag{2.30}$$

For all Born-Mayer type models:

$$\begin{aligned} P(B_{ij}, r_{ij}) &= 1 \\ x &= B_{ij} r_{ij} \end{aligned} \tag{2.31}$$

For the Born-Mayer-IP FF:

$$\begin{aligned} B_i &\equiv B_i^{\mathrm{IP}} = 2\sqrt{2I_i} \\ B_{ij} &= \frac{B_i B_j (B_i + B_j)}{B_i^2 + B_j^2} \end{aligned} \tag{2.32}$$

For the Born-Mayer-sISA FF:

$$\begin{aligned} B_i &= 0.84 B_i^{\mathrm{ISA}} \\ B_{ij} &= \sqrt{B_i B_j} \end{aligned} \tag{2.33}$$

Of the parameters in these force fields, only the coefficients A_i were fit to reproduce DFT-SAPT dimer energies (details below). All other force field parameters were derived from first-principles atom or atom-in-molecule properties. Exponents for the Slater-ISA FF and the Born-Mayer-sISA FF were derived from BS-ISA calculations, while exponents for the Born-Mayer-IP FF were determined from vertical ionization potentials of the isolated atoms. Dispersion coefficients ($C_{ij,2n}$) were either used directly from Ref. 83 or were parameterized using analogous methods

in the case of argon. Distributed multipoles Q_t^i for each system were obtained from the BS-ISA-based distributed multipoles scheme (ISA-DMA)[91], with the expansion truncated to rank 2 (quadrupole). Note that here, $t = 00, 10, \ldots, 22s$ denotes the rank of the multipole in the compact notation of Stone[78]. (In addition to rank 2 ISA-DMA multipoles, we also tested the use of DMA4 multipoles[87] as well as the use of rank 0 charges obtained from the rank truncation or transformation[187] of either ISA-DMA or DMA4 multipoles; the effect of including a Tang-Toennies damping factor[83,156] was studied in all cases. Each of these alternative long-range electrostatic models proved either comparably or less accurate for both the Slater-ISA FF and the Born-Mayer-IP FF in terms of their ability to reproduce the DFT-SAPT electrostatic energy, and are not discussed further.) Long-range polarization (V_{shell}) was modeled using Drude oscillators in a manner identical to Ref. 83. As in our prior work, during parameterization, the Drude energy was partitioned into 2nd ($V_{pol}^{(2)}$) and higher order ($V_{pol}^{(3-\infty)}$) contributions, where $V_{pol}^{(2)}$ is the Drude oscillator energy due to static charges (excluding intra-molecular contributions), and $V_{pol}^{(3-\infty)}$ is the difference between the fully converged Drude energy, V_{shell}, and $V_{pol}^{(2)}$. Force field parameters for all homo-monomeric systems are located in the Supporting Information of Ref. 95.

A weighted least-squares fitting procedure was used to fit A_i parameters to the benchmark DFT-SAPT (PBE0/AC) interaction energies on a component-by-component basis. That is, four separate optimizations[83] were performed to directly fit V^{exch}, V^{elst}, V^{ind}, and $V^{\delta^{HF}}$ to, respectively, the following DFT-SAPT quantities (notation as in Ref. 172):

$$\begin{aligned} E^{exch} &\equiv E_{exch}^{(1)} \\ E^{elst} &\equiv E_{pol}^{(1)} \\ E^{ind} &\equiv E_{ind}^{(2)} + E_{ind\text{-}exch}^{(2)} \\ E^{\delta^{HF}} &\equiv \delta(HF). \end{aligned} \tag{2.34}$$

For V^{disp}, no parameters were directly fit to the DFT-SAPT dispersion,

$$E^{disp} \equiv E^{(2)}_{disp} + E^{(2)}_{disp\text{-}exch}, \tag{2.35}$$

but were instead obtained solely from monomer properties as described above. Finally, note that no parameters were directly fit to the total DFT-SAPT energy,

$$E_{int} = E^{exch} + E^{elst} + E^{ind} + E^{\delta^{HF}} + E^{disp}, \tag{2.36}$$

for either the Slater-ISA FF or the Born-Mayer-IP FF. Rather, V_{FF} was calculated according to Eq. (2.29).

Data points for each fit were weighted using a Fermi-Dirac functional form given by

$$w_i = \frac{1}{\exp((-E_i - \mu_{eff})/kT) + 1}, \tag{2.37}$$

where E_i is the reference energy and μ_{eff} and kT were treated as adjustable parameters. The parameter kT, which sets the energy scale for the weighting function, was taken to be $kT = \lambda|E_{min}|$; here E_{min} is an estimate of the global minimum well depth. Unless otherwise stated, we have used $\lambda = 2.0$ and $\mu_{eff} = 0.0$. These defaults were chosen to minimize overall average attractive RMSE for all 91 dimer sets. Increases or decreases in the λ factor correspond to the weighting of more or fewer repulsive configurations, respectively.

In the case of Lennard-Jones, the standard Lennard-Jones functional form was used for the van der Waals terms, with Coulomb and polarization terms modeled exactly as for the Slater-ISA FF:

$$V^{LJ}_{FF} = \sum_{ij} \frac{A_{ij}}{r^{12}_{ij}} - \frac{C_{ij,6}}{r^{6}_{ij}} + V_{pol} + \sum_{tu} Q^{i}_{t} T_{tu} Q^{j}_{u} \tag{2.38}$$

Lorentz-Berthelot combination rules were used to obtain heteroatomic A_{ij} and C_{ij} parameters. Unlike with the Slater-ISA FF and Born-Mayer models, V^{LJ}_{FF} was fit to

the total DFT-SAPT (PBE0/AC) energy, with A_{ij} and $C_{ij,6}$ as fitting parameters. The weighting function from Eq. (2.37) was used in fitting.

2.3.5 Potential Energy Surface Scans

In order to visually assess fit quality, representative one-dimensional scans of the potential energy surface were calculated for several dimer pairs along low-energy dimer orientations. For each dimer pair, the minimum energy configuration of the 1000 random dimer points was selected as a starting configuration, and additional dimer configurations (not necessarily included in the original 1000 points) were generated by scanning along some bond vector. In the case of the ethane dimer, two carbon atoms (one on each monomer) were used; for acetone, the carbonyl carbon on each monomer defined the bond vector.

2.3.6 Molecular Simulations

All bulk simulations were run using OpenMM release version 7.0.[188] Enthalpies of vaporization were computed from

$$\Delta H_{vap} = (E_{pot}(g) + RT) - E_{pot}(l)$$

where $E_{pot}(g)$ and $E_{pot}(l)$ were determined from NVT simulations at the experimental gas and liquid densities, respectively. Calculated liquid densities were determined from NPT simulations. In all cases, the OPLS/AA force field was used for the intramolecular potential.[189] All simulations used a Langevin integrator with a 0.5 fs time step and a 1 ps^{-1} friction coefficient; NPT simulations used a Monte Carlo barostat with a trial volume step every 5^{th} move. Periodic boundary conditions, particle-mesh Ewald, and a non-bonding cutoff of 1.2nm with added long-range corrections were used to simulate a unit cell of 222 molecules. After an equilibration period of at least 600ps, simulation data was gathered from production runs lasting at least 200ns.

2.4 Results and Discussion

The Slater-ISA methodology for short-range intermolecular interactions has been derived from a simple but rigorous physical model of overlapping monomer electron densities. In practice, this approach differs from the conventional Born-Mayer approach in both the choice of the short-range functional form (with the latter omitting the polynomial pre-factor) and the source of the exponents (with the former derived from ISA analysis of the monomer density). Our principal goal is to examine the influence of these modifications on the accuracy and transferability of the resulting force fields.

We initially benchmark the Slater-ISA FF against a conventional Born-Mayer potential, Born-Mayer-IP FF. The latter approach has been used extensively in prior work,[75,83] and both approaches use identical numbers of fitted parameters. Following prior work, combination rules for the Born-Mayer-IP FF are as in Ref. 83. (We have tested the effect of using a geometric mean for the Born-Mayer-IP FF; results do not differ qualitatively from those presented below.) Owing to its popularity, we also compare the Slater-ISA FF to a Lennard-Jones functional form (LJ FF).

We first assess the accuracy of the Slater-ISA FF, Born-Mayer-IP FF, and LJ FF against benchmark ab initio intermolecular interaction energies and experimental 2^{nd} virial coefficients, enthalpies of vaporization, and liquid densities. We next examine parameter transferability, assessing the extent to which parameters from pure homo-monomeric systems can be re-used (without further optimization) to describe mixed interactions. To assess parameter robustness, we also study the sensitivity of each methodology to changes in the weighting function (Eq. (2.37)). Finally, we explore the application of BS-ISA-derived exponents within the Born-Mayer functional form as a straightforward method for simplifying the parameterization (and potentially increasing the accuracy) of a wide variety of standard ab initio and empirically-parameterized force fields.

2.4.1 Accuracy: Comparison with DFT-SAPT

For each of the 91 molecule pairs described in the Computational Methods section, parameters for the Slater-ISA FF, Born-Mayer-IP FF, and LJ FF were fit to reproduce DFT-SAPT (PBE0/AC) interaction energies calculated for a set of 1000 dimer configurations. These 91,000 total configurations and corresponding DFT-SAPT energies are collectively referred to as the '91 dimer test set'. As a primary indication of accuracy, root-mean-square errors (RMSE) and mean signed errors (MSE), both with respect to DFT-SAPT, were computed for each methodology and for each dimer pair. Because these RMSE and MSE are dominated by repulsive contributions, and owing to the thermodynamic importance of attractive configurations, so-called 'attractive RMSE/MSE' were also computed by excluding net repulsive configurations (as measured by the DFT-SAPT total energy). The overall RMSE/MSE for all 91 dimers were then averaged to produce one 'characteristic RMSE/MSE' for the entire test set. Since these errors varied considerably in magnitude depending on the dimer in question, this overall average was taken in the geometric mean sense. (Results with an arithmetic mean do not differ qualitatively). Note that when computing the characteristic MSE, only the magnitude of each MSE, ||MSE||, was considered.

Characteristic RMSE and ||MSE|| across the 91 dimer test set are shown in Fig. 2.3 and Table 2.1. Overall, the Slater-ISA FF exhibits smaller errors compared to the Born-Mayer-IP FF. On average, the characteristic total energy RMSE for the Slater-ISA FF decrease by 33% relative to the Born-Mayer-IP FF. Even excluding repulsive configurations (dominated by short-range interactions), errors for the Slater-ISA FF are lower by 11% compared to the Born-Mayer-IP FF, demonstrating modest gains in accuracy even over the most energetically-relevant regions of the potential. A more detailed analysis of each of the 91 pairs of molecules shows that in an overwhelming 93% of such cases, force fields derived from the Slater-ISA method have smaller RMSEs compared to their Born-Mayer-IP counterparts (70% if only attractive configurations are considered). Regardless of the metric used, the Slater-ISA FF produces force fields with higher fidelity to the underlying benchmark interaction energies.

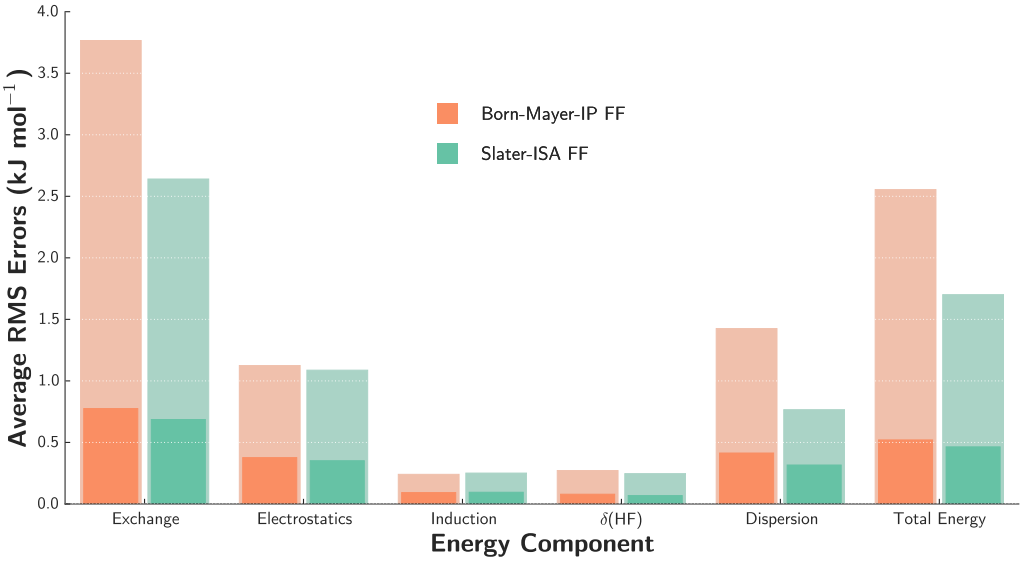


Figure 2.3: Characteristic RMSE (as described in the main text) for the Born-Mayer-IP FF (orange) and the Slater-ISA FF (green) over the 91 dimer test set. The translucent bars represent total RMSE for each energy component, while the smaller solid bars represent 'Attractive' RMSE, in which repulsive points have been excluded.

Component	Dimer-Specific Fits			Transferable Fits		
	Slater-ISA FF (kJ mol^{-1})	Born-Mayer-IP FF (kJ mol^{-1})	LJ FF (kJ mol^{-1})	Slater-ISA FF (kJ mol^{-1})	Born-Mayer-IP FF (kJ mol^{-1})	LJ FF (kJ mol^{-1})
Exchange	2.641 (0.686)	3.766 (0.775)	—	2.718 (0.720)	4.033 (0.836)	—
Electrostatics	1.087 (0.351)	1.126 (0.377)	—	1.134 (0.351)	1.231 (0.378)	—
Induction	0.251 (0.095)	0.241 (0.093)	—	0.278 (0.101)	0.265 (0.098)	—
δHF	0.246 (0.068)	0.272 (0.079)	—	0.274 (0.076)	0.304 (0.081)	—
Dispersion	0.766 (0.317)	1.425 (0.414)	—	0.766 (0.317)	1.425 (0.414)	—
Total Energy						
RMSE	1.701 (0.464)	2.554 (0.520)	1.984 (0.603)	1.650 (0.456)	2.698 (0.555)	2.054 (0.640)
$\|\|MSE\|\|$	0.216 (0.057)	0.539 (0.127)	0.322 (0.345)	0.175 (0.051)	0.569 (0.112)	0.311 (0.368)

Table 2.1: Comparison of characteristic RMSE (as described in the main text) over the 91 dimer test set for the Slater-ISA FF, Born-Mayer-IP FF and LJ FF. For the total energy, both characteristic RMSE and MSE have been shown, with only the magnitude of the MSE, $\|\text{MSE}\|$, displayed. 'Attractive' RMSE, representing the characteristic RMSE for the subset of points whose energies are net attractive ($E_{int} < 0$), are shown in parentheses to the right of the total RMS errors; 'attractive' $\|\text{MSE}\|$ are likewise displayed for the total energy. As discussed in Section 2.4.3, the 'Dimer-Specific Fits' refer to force fields whose parameters have been optimized for each of the 91 dimers separately, whereas the 'Transferable Fits' refer to force fields whose parameters have been optimized for the 13 homodimers and then applied (without further optimization) to the remaining 78 mixed systems. Unless otherwise stated, a default weighting function of $\lambda = 2.0$ (see Eq. (2.37)) has been used for all force fields in this Chapter.

It is also instructive to consider each energy component individually. As might be expected, improvements in the description of E^{exch} are pronounced, with the characteristic RMSE from the Slater-ISA FF being 30% smaller than that from the Born-Mayer-IP FF. Examining each dimer pair separately (see **??** for homomonomeric fits, representative of the entire test set) we also find that, in general, the Slater-ISA FF is far better at reproducing *trends* in the exchange energy compared to the Born-Mayer-IP FF. This qualitative result is also reflected in the smaller ||MSE|| values for the Slater-ISA FF as compared to the Born-Mayer-IP FF. Nevertheless, there remains a fair amount of scatter in the exchange energies for several dimer pairs, particularly for molecules with exposed lone pairs or delocalized π systems. We hypothesize that this scatter is due to a breakdown of the isotropic approximation made in the Theory section, a conclusion supported by observations on the pyridine dimer system recently made by some of us.[155] It it therefore quite possible that the observed 30% RMSE reduction underestimates the true error reduction that might be observed if such anisotropy were accounted for.

From Fig. 2.3, we see that the dispersion energy model from the Slater-ISA FF is also a substantial improvement; for dispersion, characteristic RMSE are 46% smaller for the Slater-ISA FF compared to the Born-Mayer model. This should not be a counter-intuitive result: while both potentials use identical dispersion coefficients, they differ in the damping model used. In the Born-Mayer-IP FF, the standard Tang–Toennies damping model is employed, and the damping parameters only depend on free atom ionization potentials; in the Slater-ISA FF, on the other hand, the damping parameters are obtained from the ISA shape functions, and thus take molecular environment effects into account. Even when only considering attractive dimer configurations (solid bar in Fig. 2.3), errors in the dispersion energy component are reduced by 23%, demonstrating the importance of the damping function across the potential surface. From these results, and in agreement with related literature studies,[190] we conclude that use of the standard Tang-Toennies damping function based on atomic ionization potentials[83,156,191–195] lacks quantitative predictive power compared to the Slater-ISA model. Note that neither the Slater-ISA FF nor the Born-Mayer-IP FF are directly fitted to the DFT-SAPT dispersion energies

(all parameters are determined from monomer properties), making this accuracy particularly striking. We hypothesize that the effect of the Slater-ISA approach is greater for dispersion than for first-order exchange because here (in contrast to the exchange energy) there are no fitted parameters to compensate for deficiencies in the exponents or functional form of the Born-Mayer-IP FF.

In contrast to the exchange and dispersion energies, the Slater-ISA FF and the Born-Mayer-IP FF show nearly identical errors for the electrostatic and the induction (2^{nd} order induction plus δHF) energies. In these cases, the two models differ only in the parameters and functional form used to represent the exponentially-dependent short-range terms of these energy components, namely the penetration component for the electrostatic term and the penetration/charge-transfer term for the induction. The lack of improvement between the Slater-ISA and Born-Mayer-IP models may imply that we are not able to capture the physics of these particular short-range interactions with either the Slater-functional of Born-Mayer functional forms. Alternatively, the assumption that the short-range components of the electrostatic and induction energies are proportional to the exchange-repulsion may need to be re-examined. As discussed in Section 2.2.2, this proportionality is known to be approximately valid, but as yet there does not seem to be a deeper theoretical understanding of these short-range terms that may lead to a better model. Nevertheless, absolute errors in the electrostatic and induction components are relatively small for both models. Thus overall, the Slater-ISA FF functional form is promising for treating a wide variety of short-range effects.

The comparison between the Slater-ISA FF and the LJ FF is slightly more complicated, owing to the differences in long-range potential and fitting methodology (see Section 2.3.4). As such, we compare the Slater-ISA FF to several versions of the LJ FF (for which characteristic RMSE and ||MSE|| are shown in Table 2.2). Using the same weighting function and constraining the Coulombic and polarization terms to be identical to the Slater-ISA FF, we see that the resulting Lennard-Jones force field (LJ FF, $\lambda = 2.0$) is significantly worse than the Slater-ISA FF, both in terms of total RMSE and attractive RMSE. Furthermore, by comparing the ||MSE|| of both force fields, we see that errors in LJ FF are much more *systematic* than in the

	LJ FF Dimer-Specific Fits		LJ FF Transferable Fits	
	$\lambda = 2.0$ (kJ mol^{-1})	$\lambda = 0.1$ (kJ mol^{-1})	$\lambda = 2.0$ (kJ mol^{-1})	$\lambda = 0.1$ (kJ mol^{-1})
RMSE	1.984 (0.603)	6.058 (0.413)	2.054 (0.640)	5.760 (0.457)
$\|\|MSE\|\|$	0.322 (0.345)	1.610 (0.041)	0.311 (0.368)	1.410 (0.060)

Table 2.2: Comparison of characteristic RMSE and ||MSE|| over the 91 dimer test set for the various Lennard-Jones models. The LJ models are not parameterized on a component-by-component basis, thus RMSE/||MSE|| values are only shown for the total FF energies. 'Attractive' errors, representing the characteristic RMSE/||MSE|| for the subset of points whose energies are net attractive ($E_{int} < 0$), are shown in parentheses to the right of the total errors. 'Dimer-Specific Fits' and 'Transferable Fits' are as in Table 2.1.

Slater-ISA FF: in order to reproduce the repulsive wall correctly, the Lennard-Jones potential generally underestimates the well-depth by a considerable fraction (see the Supporting Information of Ref. 95 for ethane as a typical example).

Given the failure of the LJ FF ($\lambda = 2.0$) force field to reproduce the energetically important region of the PES, we also compared the Slater-ISA FF to a 'best-case' scenario Lennard-Jones force field which correctly reproduces the minimum energy region at the expense of the repulsive wall. These LJ FF ($\lambda = 0.1$) fits have total RMSE errors nearly 4 times that of the Slater-ISA FF; indeed, the LJ FF ($\lambda = 0.1$) reproduces the repulsive wall only qualitatively. Insofar as the repulsive wall is concerned, the Slater-ISA FF is far superior to the Lennard-Jones short-range model. Nevertheless (and much more importantly for molecular simulation), the attractive region of the potential is reproduced surprisingly well by LJ FF. Characteristic attractive RMSE for the LJ FF ($\lambda = 0.1$) are slightly lower than those for Slater-ISA FF, although the former has one additional free parameter per atom type and is also fit directly to reproduce the total energy. Likewise, attractive ||MSE|| between the Slater-ISA FF and the LJ FF ($\lambda = 0.1$) are comparable. As we show in the Supporting Information of Ref. 95, however, and as is well known in the literature, weighting the Lennard-Jones potential in this manner does not necessarily capture important

information from the long-range attractive tail or repulsive wall of the PES, such that the LJ FF ($\lambda = 0.1$) is not always expected to yield good property predictions. This latter point will be demonstrated in Section 2.4.2.

In order to compare the performance of the Slater-ISA FF against popular standard force fields, we also developed a 'best case scenario' non-polarizable point charge Lennard-Jones model, results for which are shown in the Supporting Information of Ref. 95. Unsurprisingly, this force field is worse (in an RMSE and ||MSE|| sense) than all other force fields studied in this Chapter, thus demonstrating how important accurate models for long-range electrostatics and polarization are to the overall accuracy of ab initio force fields.

Argon Dimer

We now turn to several specific case studies. The Ar dimer provides an interesting test case to examine directly the impact of the polynomial pre-factor included in the Slater-ISA FF functional form. Since Ar is an atomic species, we should have $B_{Ar}^{ISA} = B_{Ar}^{IP}$. For numerical reasons, the Slater-ISA FF and Born-Mayer-IP FF exponents differ by 0.03 a.u.; however, this difference is insignificant, and the two FFs differ mainly in the polynomial pre-factor. Fig. 2.4 shows the potential energy surface (PES) for the argon dimer computed using the Slater-ISA FF and the Born-Mayer-IP FF. Here the default weighting scheme has been used so as to best reproduce the energetically attractive region. Note that, while both potentials reproduce the minimum energy configurations correctly, the Born-Mayer-IP FF considerably overestimates the exchange energy (and thus the total energy) along the repulsive wall. The Slater-ISA FF, on the other hand, maintains excellent accuracy in this region of the potential. This result is particularly notable because the repulsive wall is not heavily weighted in the fit. (A point 10 kJ mol^{-1} along the repulsive wall, for instance, is weighted only 3% as heavily as a point near the bottom of the well). A similar, though smaller, increase in accuracy is seen in the fit to the DFT-SAPT dispersion energies, where the Slater-ISA FF is better able to model the energies for shorter interatomic separations. This increased accuracy is entirely attributable to

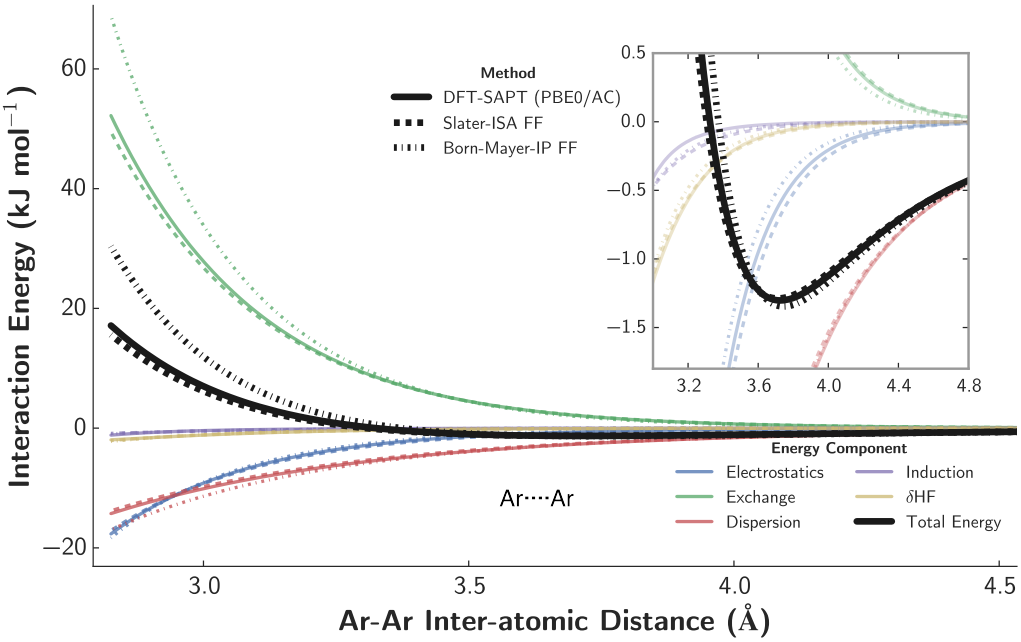


Figure 2.4: Potential energy surface for the argon dimer. Interaction energies for the Slater-ISA FF (dashed curves) and the Born-Mayer-IP FF (dash-dotted curves) are shown alongside benchmark DFT-SAPT (PBE0/AC) energies (solid curves). The energy decomposition for DFT-SAPT and for each force field is shown for reference.

the functional form employed, as the dispersion parameters are identical between the two FFs.

Consistent with prior literature,[125,141] these results suggest that neglect of the polynomial pre-factor P (as in standard Born-Mayer potentials) is *by itself* a poor approximation. However, as we show below, the Born-Mayer form can still be used as an accurate model in conjunction with appropriately scaled atomic exponents. Nonetheless, the more physically-motivated Slater form provides increased accuracy over a wider range of separations without recourse to empirical scaling.

Results for LJ FF are shown in the Supporting Information of Ref. 95; consistent with expectations for the Lennard-Jones model, the repulsive wall is overestimated by the $1/r_{ij}^{12}$ short-range functional form, and the magnitude of the attractive tail region is similarly overestimated by the effective $C_{ij,6}$ dispersion parameter. Note

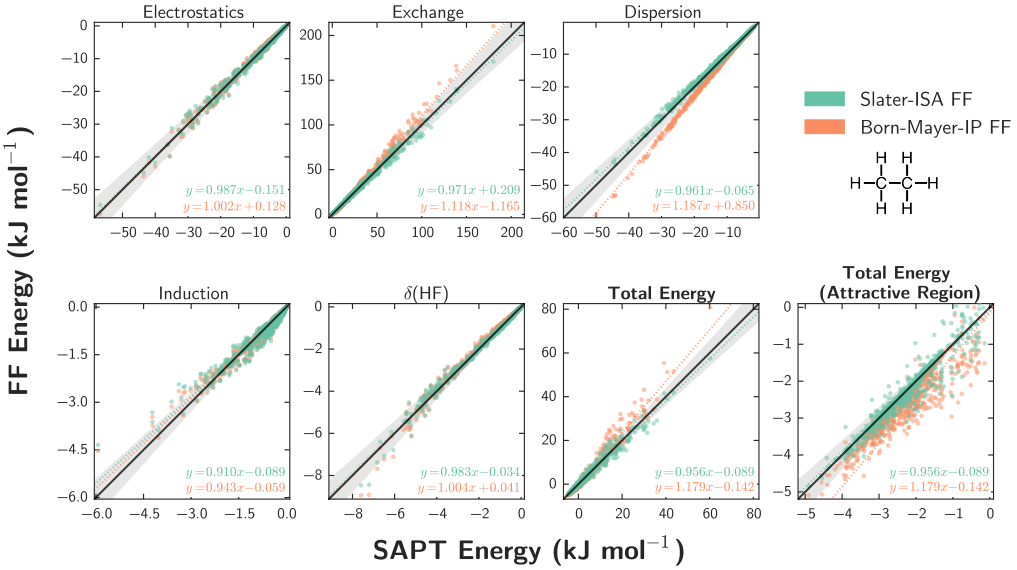


Figure 2.5: Force field fits for the ethane dimer using the Slater-ISA (green) and Born-Mayer-IP (orange) FFs. Fits for each energy component are displayed along with two views of the total interaction energy. The diagonal line (black) indicates perfect agreement between reference energies and each force field, while shaded grey areas represent points within ±10% agreement of the benchmark. To guide the eye, a line of best fit (dotted line) has been computed for each force field and for each energy component.

that this $C_{ij,6}$ coefficient has been fit to the total energy, and thus differs from the asymptotically-correct $C_{ij,6}$ parameter used for both the Slater-ISA FF and the Born-Mayer-IP FF. An alternative parameterization strategy would have been to use the asymptotically-correct $C_{ij,6}$ parameter in the LJ FF, but this would have worsened predictions along both the repulsive wall and the minimum energy configurations.

Ethane Dimer

We next discuss the ethane dimer and show both a scatter plot of the 1000 dimer interactions (Fig. 2.5) and a cut through the potential energy surface near the minimum (Fig. 2.6) as indications of force field quality.

As with the argon dimer, for the ethane dimer the Slater-ISA FF produces more

accurate exchange and dispersion energies compared to the Born-Mayer-IP FF. Here, the effects of the Slater-ISA FF for dispersion are even more pronounced, likely because the conventional damping of the Born-Mayer-IP FF is systematically in error due to differences in both the form of the damping function and exponents. As for the total interaction energy, we again find that the Born-Mayer-IP FF exhibits large errors for repulsive contributions, while the Slater-ISA FF naturally reproduces interactions for both attractive and strongly repulsive configurations. Even in the attractive regime, the Born-Mayer-IP FF is systematically too attractive. These systematic errors are the result of imperfect error cancellation between the exchange and dispersion components of the fit, and are discussed in more detail in Section 2.4.4.

Examining a specific cut across the ethane-ethane PES (Fig. 2.6) visually confirms these results. Both potentials do an excellent job of reproducing the benchmark DFT-SAPT energies in the minimum energy region, though the Born-Mayer-IP FF is slightly too attractive. (Other cuts of the PES would show the Born-Mayer-IP predictions to be significantly more in error, consistent with the scatter plots). Along the repulsive wall, however, the Born-Mayer-IP FF predictions worsen in comparison to those from the Slater-ISA FF. Finally, the PES shows an increased reliance on error cancellation between the various energy components for the Born-Mayer-IP FF compared to the Slater-ISA FF.

As shown in the Supporting Information of Ref. 95, the Lennard-Jones force field models are incapable of reproducing the entirety of the ethane PES; depending on the weighting function, either the repulsive wall or the attractive well can be reproduced, however no set of parameters can predict both regions simultaneously.

Acetone Dimer

The acetone dimer provides a final interesting example involving a moderately sized organic molecule. From both the scatter plots (Fig. 2.7) and the PES cross section (Fig. 2.8), it is evident that both the Slater-ISA and Born-Mayer-IP force fields do an excellent job of reproducing DFT-SAPT energies for the low energy

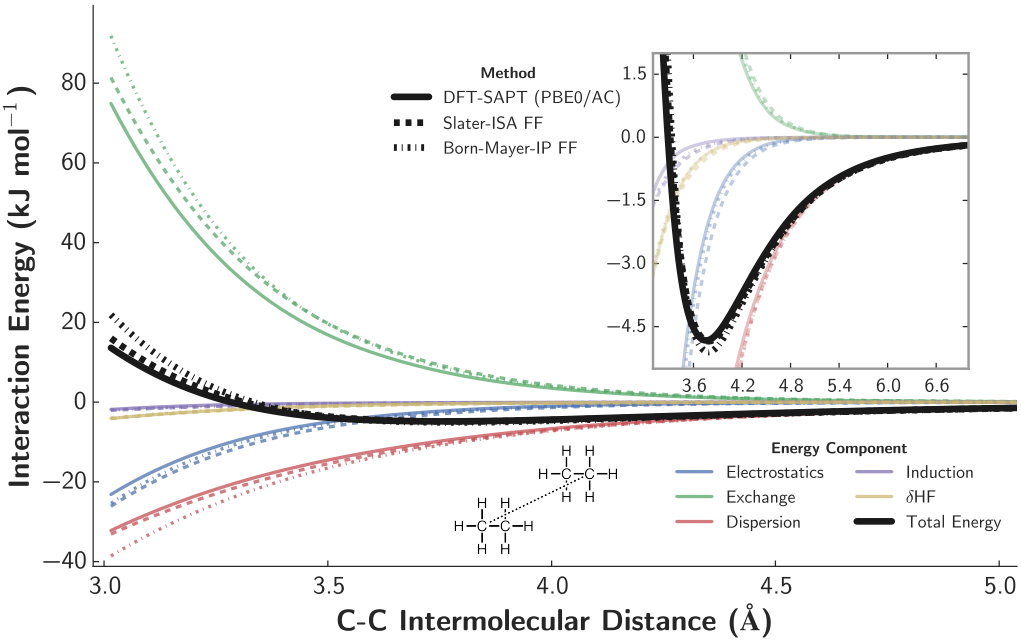


Figure 2.6: A representative potential energy scan near a local minimum for the ethane dimer. Interaction energies for the Slater-ISA FF (dashed curves) and the Born-Mayer-IP FF (dash-dotted curves) are shown alongside benchmark DFT-SAPT (PBE0/AC) energies (solid curves). The energy decomposition for DFT-SAPT and for each force field is shown for reference. The ethane dimer configuration in this scan corresponds to the most energetically attractive dimer included in the training set; other points along this scan are not included in the training set.

dimers. Along the repulsive wall, however, the Born-Mayer-IP FF shows larger systematic errors in each energy component, and seems to rely on error cancellation to achieve good agreement in the total energy. This reliance on error cancellation has two negative effects: Firstly, the additional scatter in the total energy of the Born-Mayer-IP FF fit, especially prominent for attractive configurations, indicates that this error cancellation is imperfect in certain cases. MSE for the Slater-ISA FF (-0.0115 kJ mol^{-1}) are an order of magnitude lower than for the Born-Mayer-IP FF (0.182 kJ mol^{-1}) in the attractive region of the potential. Secondly, as we shall later explore, reliance on error cancellation likely contributes to the somewhat decreased transferability of the Born-Mayer-IP FF as compared to the Slater-ISA FF.

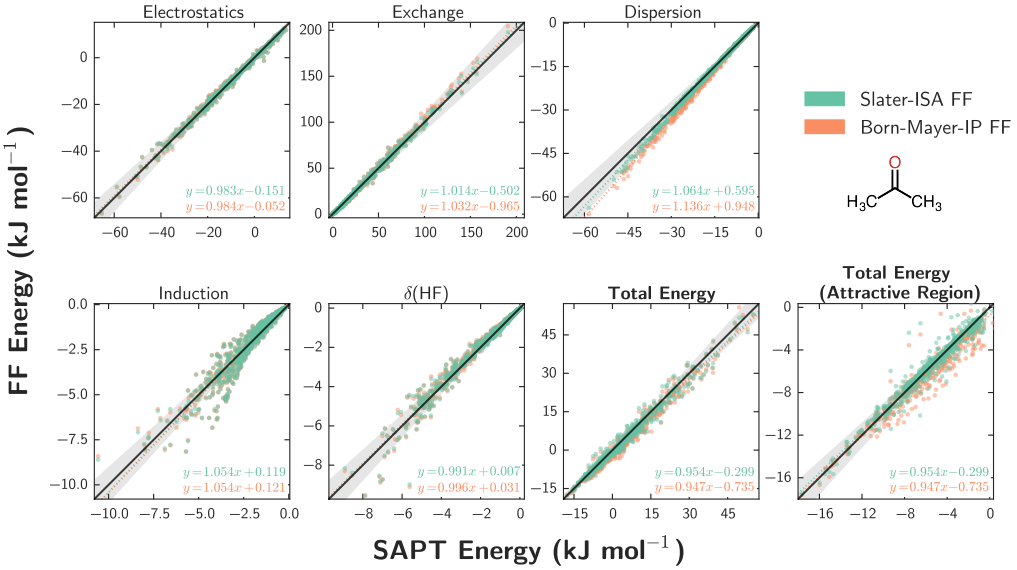


Figure 2.7: Force field fits for the acetone dimer using the Slater-ISA (green) and Born-Mayer-IP (orange) FFs, as in Fig. 2.5.

As shown in the Supporting Information of Ref. 95, the LJ FF predictions for acetone are reasonably good in both the tail and minimum energy regions of the potential, however the LJ FF grossly overpredicts the DFT-SAPT (PBE0/AC) energies along the repulsive wall.

2.4.2 Accuracy: Comparison with experiment

We have benchmarked the above force fields against experimental second virial coefficients and, in the case of ethane, enthalpies of vaporization and liquid densities. The classical 2nd virial coefficients were calculated for both argon and ethane using rigid monomer geometries, following the procedure described in Ref. 83. Enthalpies of vaporization and liquid densities were calculated using the OpenMM molecular simulation package[188] as described in Section 2.3. Higher-order multipole moments — which were negligible for these molecules — were neglected, and so only rank 0 terms were used in these calculations. Results are shown in Figures

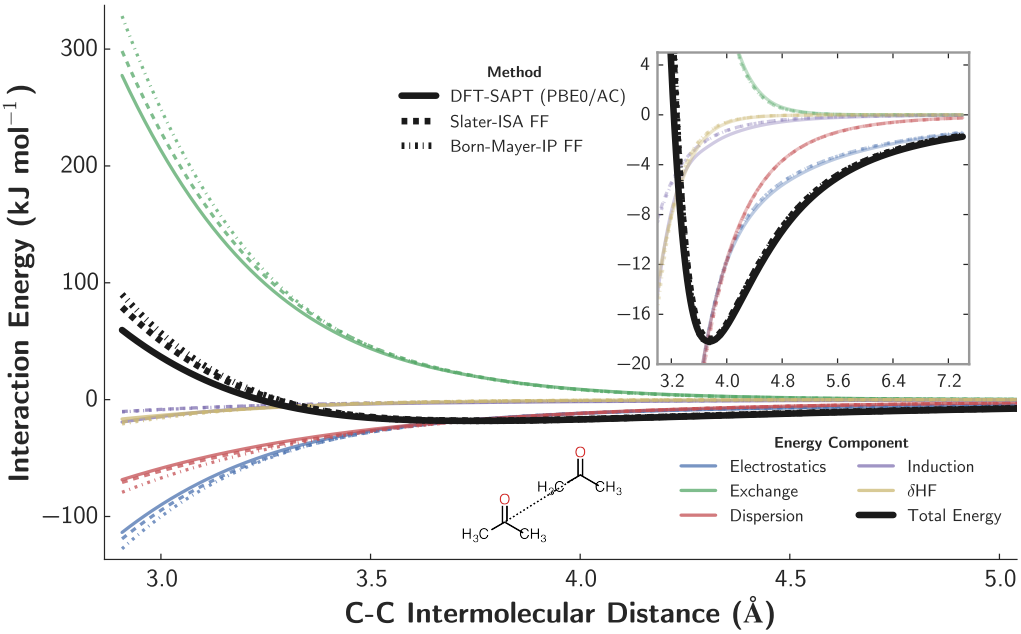


Figure 2.8: A representative potential energy scan near a local minimum for the acetone dimer. Interaction energies for the Slater-ISA FF (dashed curves) and the Born-Mayer-IP FF (dash-dotted curves) are shown alongside benchmark DFT-SAPT (PBE0/AC) energies (solid curves). The energy decomposition for DFT-SAPT and for each force field is shown for reference. The intermolecular distance is taken to be the internuclear distance between the two carbonyl carbons on each acetone monomer. The configuration in this scan corresponds to the most attractive dimer configuration included in the training set for the acetone dimer; other points along this scan have not explicitly been included in the training set.

2.9 and 2.10 as well as Table 2.4.

For argon, since both Slater-ISA FF and Born-Mayer-IP FF accurately reproduce the energetics of low-energy configurations, it is unsurprising that both force fields yield accurate virial coefficients over a wide range of temperatures. Errors in computed B_2 coefficients (for both potentials) are likely attributable to small errors in the DFT-SAPT (PBE0/AC) potential itself,[178] and, to a much lesser extent, the neglect of nuclear quantum effects at lower temperatures.[196] Despite the good (in an RMSE sense) fit quality of the LJ FF ($\lambda = 0.1$), this force field overpredicts the

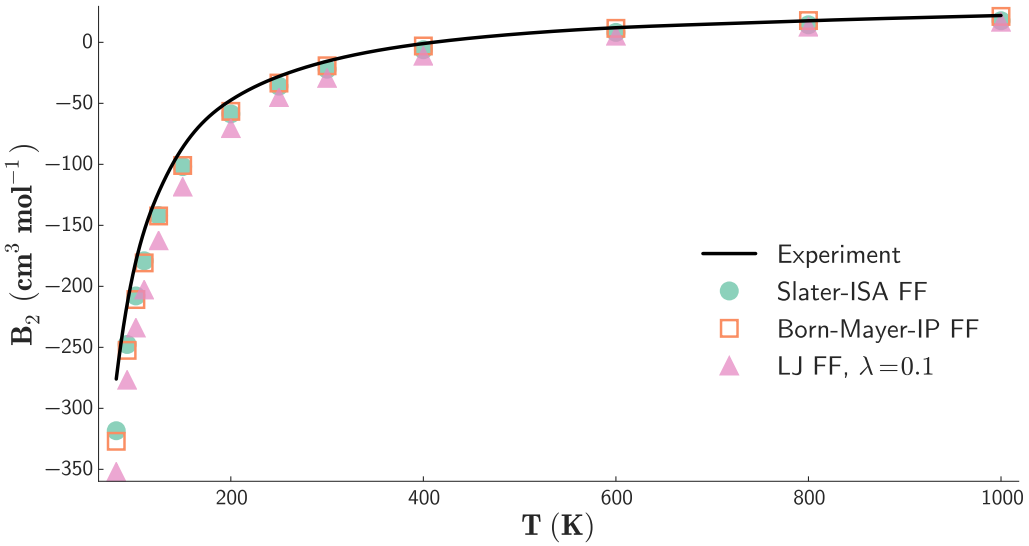


Figure 2.9: Second virial coefficients for argon. The Slater-ISA and the Born-Mayer-IP FFs are shown as green circles and orange squares, respectively; the black line corresponds to experiments from Ref. 10.

magnitude of the 2nd virial for argon, likely as a result of the effective dispersion coefficient, which overestimates the attraction in the tail region of the PES (see Supporting Information of Ref. 95). Although it is certainly possible to parameterize a Lennard-Jones model *empirically* for argon, such a force field would rely on a subtle cancellation of errors between the minimum energy- and tail-regions of the PES. As the proper balance is impossible to predict a priori, this result highlights one of the difficulties of using the less physical LJ model in the development of ab-initio force fields.

In the case of ethane, the Slater-ISA FF is in excellent agreement with experiment, whereas the Born-Mayer-IP FF underpredicts B_2 by as much as 20%. These results are indicative, not only of the more accurate functional form and parameterization of Slater-ISA FF, but also of the high accuracy of the underlying DFT-SAPT (PBE0/AC) benchmark energies. In this case, LJ FF also correctly predicts the virial. Using weighting functions for each model that are optimal for the 91 dimer test set as a whole ($\lambda = 2.0$ for the Slater-ISA FF and the Born-Mayer-IP FF, $\lambda = 0.1$ for the

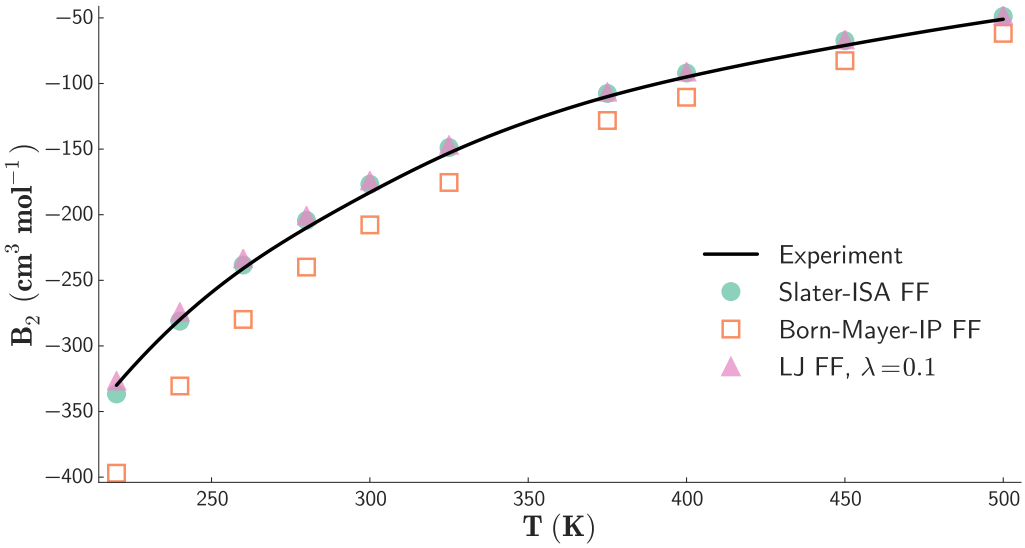


Figure 2.10: Second virial coefficients for ethane. The Slater-ISA and Born-Mayer-IP FFs are shown as green circles and orange squares, respectively; the black line corresponds to experiments from Ref. 10.

LJ FF), all force fields produce similar results for ΔH_{vap} and ρ (Table 2.4). These values are slightly overestimated by all force fields (especially in the case of the Born-Mayer-IP FF), which is to be expected given our neglect of many-body effects. McDaniel and Schmidt have calculated the 3-body correction for the Born-Mayer-IP FF; using this value as a global 3-body correction for all force fields, we see that both the Slater-ISA and the Lennard-Jones force fields compare very favorably to experiment, with the Slater-ISA FF perhaps slightly more accurate.

2.4.3 Transferability

The transferability of interaction potentials is a crucial aspect of practical molecular simulations. Here we examine 'parameter transferability', by which we mean the extent to which parameters from two homo-monomeric systems can combined to predict the intermolecular interactions of the resulting mixed hetero-monomeric system. As a measure of parameter transferability, we compared characteristic

RMSE and ||MSE|| relative to the benchmark data for two different parameterization schemes. For the 'Dimer-Specific Fits', A_{ij} parameters were obtained for each of the 91 dimer pairs individually; these results are identical to those discussed in the previous two subsections. In contrast, for the 'Transferable Fits', the A_{ij} parameters were fit to the 13 homonomeric dimer pairs and were re-used (without any further optimization) to calculate energies for the 78 mixed systems using the combination rules listed in Section 2.3.4. Results for each parameterization scheme are shown in Table 2.1. From the RMSE and ||MSE|| from the competing schemes, we see excellent parameter transferability for all force fields studied. For the Slater-ISA FF, characteristic RMSE and ||MSE|| for each component increase by a very small fraction upon constraining the fit; due to small error cancellation, errors in the total energy actually *decrease* somewhat with these constraints. (This is possible since the total energy is not directly fit.) The Born-Mayer-IP FF also displays a significant degree of transferability, though errors in the total energy increase slightly upon constraining the fit. As in prior work, the observed parameter transferability for both force fields can be attributed to our use of a term-by-term parameterization scheme (Section 2.3.4), which serves to minimize error cancellation between energy components and generate a more physically-meaningful (and thus transferable) set of parameters.[83,197] Finally, note that for four of the five interaction energy components the relative change in RMSE on constraining the fit is smaller for the Slater-ISA FF than the Born-Mayer-IP FF. The δHF term is the exception, but even here the relative change in errors from the two methods are comparable. This suggests that the Slater-ISA FF may be the more transferable of the force fields studied. Nevertheless, the Lennard-Jones model is surprisingly transferable, likely in part due to the same accurate and transferable 'long-range' electrostatics and polarization as the Slater-ISA FF. The non-polarizable, point-charge Lennard-Jones model (results for which are shown in the Supporting Information of Ref. 95) displays the least transferability (in both an RMSE and ||MSE|| sense) of all force fields studied.

Although we do not examine it here, we expect that the previously demonstrated success[4,83,141,197] of the Born-Mayer-IP FF with respect to 'environment transferabil-

ity' — the extent to which a single set of parameters can model a variety of phases and molecular environments — and 'atom type transferability' — the extent to which atoms in chemically similar environments can accurately be grouped together into 'types' and treated using one parameter set — would also apply to, or even be improved by, Slater-ISA FF. These issues are under investigation in our groups.

2.4.4 Robustness

One of the practical challenges of ab initio force field development is the robustness of the resulting force field quality with respect to the choice of an appropriate training set and/or weighting function. To this end, the default weighting function (Eq. (2.37), $\lambda = 2.0$) was varied to produce unconstrained fits that were skewed either towards attractive ($\lambda = 0.5$) or repulsive ($\lambda = 5.0$) configurations, and pairwise differences in force field total energies were computed between each weighting scheme. Characteristic root-mean-square pairwise differences (RMSD) between each weighting function are shown in Table 2.3; as before, 'attractive RMSD' were calculated by excluding repulsive points from consideration. Note that, on average, the default $\lambda = 2.0$ weighting scheme is optimal (in an RMSE sense) for both the Slater-ISA and Born-Mayer-IP FFs.

Overall, both the Born-Mayer-IP FF and the LJ FF display significant weighting function sensitivity. This sensitivity is not surprising; as both force fields are unable to reproduce the entirety of the potential energy surface, changing the weighting scheme (or equivalently, the balance of configurations in the training set) alters the parameters in the Born-Mayer-IP FF or the LJ FF models quite substantially. Even excluding repulsive configurations, RMSD of ~ 0.5 kJ mol^{-1} are typical for the Born-Mayer-IP FF. RMSD are somewhat smaller for the LJ FF (~ 0.3 kJ mol^{-1}), however qualitatively we see that differences in computed force field energies are systematic: smaller weighting functions capture the minimum energy region of the potential while overestimating the magnitudes of both the repulsive and tail regions of the potential, whereas larger weighting functions tend to underestimate the minimum energy region in order to correctly reproduce the repulsive wall.

Characteristic RMSD	$\lambda = 0.5$ vs 2.0 (kJ mol^{-1})	$\lambda = 0.5$ vs 5.0 (kJ mol^{-1})	$\lambda = 2.0$ vs 5.0 (kJ mol^{-1})
Slater-ISA FF	0.742 (0.207)	0.990 (0.273)	0.306 (0.086)
Born-Mayer-IP FF	1.866 (0.409)	2.632 (0.550)	0.797 (0.153)
LJ FF	1.301 (0.216)	1.605 (0.309)	0.324 (0.099)
Born-Mayer-sISA FF	0.611 (0.178)	0.810 (0.236)	0.293 (0.081)

Table 2.3: Characteristic RMS pairwise differences (RMSD) in force field total energies for different weighting functions with λ values as defined in Eq. (2.37); values shown are the (arithmetic mean, rather than geometric) RMSD across the 91 dimer test set. Characteristic 'Attractive' RMSD (as defined in Table 2.1) are shown in parentheses to the right of each overall RMSD.

Consequently, the Lennard-Jones model shows weighting-function sensitivity in a manner that is not entirely captured by the RMSD, but is instead reflected in the greater sensitivity of the LJ FF (as compared to the Born-Mayer-IP FF) in the prediction of experimental properties (*vide infra*).

Note that for practical force field development (as opposed to minimization of overall RMSE), the default weighting scheme for the Born-Mayer-IP FF and the LJ FF is suboptimal for many dimers in the test set. Because both the Born-Mayer-IP FF and the LJ FF must inherently compromise between accuracy near the minimum and along the repulsive wall, the weighting function requires system-specific fine-tuning in order to achieve proper balance. This empiricism creates significant challenges in the development of ab initio force fields.

By contrast, we find the Slater-ISA FF to be robust with respect to the choice of weighting function due to its more balanced treatment of repulsive and attractive regions of the potential energy surface. Average RMSD for the Slater-ISA FF are between two to three *times* smaller compared to the Born-Mayer-IP FF, and the Slater-ISA FF is relatively insensitive to the choice of weighting function. These conclusions hold for both attractive and overall RMSD. As a result, the Slater-ISA model largely eliminates the need for empirical fine-tuning of the weighting function, which in turn greatly simplifies the parameterization process and allows for a more robust prediction of chemical and physical properties.

Force Field	Weighting Function $\lambda = 0.1$	$\lambda = 0.5$	$\lambda = 2.0$	$\lambda = 5.0$	Experiment
	ΔH_{vap} (kJ mol^{-1}); ρ = 0.546 g L^{-1}, T = 184 K				
Slater-ISA FF	15.3 (14.7)	15.3 (14.6)	15.3 (14.7)	15.2 (14.6)	
Born-Mayer-IP FF	14.3 (13.7)	15.1 (14.5)	16.6 (15.9)	18.6 (18.0)	14.7
LJ FF	15.5 (14.9)	14.6 (13.9)	11.4 (10.7)	10.1 (9.5)	
	ρ (g L^{-1}); P = 1 atm, T = 184 K				
Slater-ISA FF	0.600 (0.566)	0.602 (0.568)	0.600 (0.566)	0.593 (0.559)	
Born-Mayer-IP FF	0.521 (0.487)	0.567 (0.533)	0.632 (0.598)	0.678 (0.644)	0.546
LJ FF	0.607 (0.573)	0.610 (0.576)	0.555 (0.521)	0.494 (0.460)	

Table 2.4: Enthalpies of vaporization and liquid densities for ethane as a function of force field and weighting function. Values in parentheses include an estimation of the 3-body correction (0.628 kJ mol^{-1} and 0.034 g mL^{-1} for the enthalpy of vaporization and liquid density, respectively) as computed in Ref. 4. Experimental data taken from Ref. 5 and Ref. 6.

For the ethane dimer, Fig. 2.11 shows overall force field energies for both the Slater-ISA and Born-Mayer-IP FFs for three weighting functions. Results for the Lennard-Jones models are shown in the SI, and are qualitatively similar to the Born-Mayer-IP FF results. The Born-Mayer-IP FF fits vary qualitatively with λ, leading to a relatively large uncertainty in calculated B_2 coefficients, enthalpies of vaporization, and liquid densities (see Table 2.4). By skewing the fits towards attractive configurations ($\lambda = 0.5$), the majority of attractive configurations are predicted without systematic error, though points along the repulsive wall (including those with net negative energies) are systematically too repulsive. Using a scheme which more heavily weights repulsive configurations, the Born-Mayer-IP FF regains semi-quantitative accuracy for repulsive configurations, albeit at the expense of a systematic increase in errors for the attractive dimer configurations. Finally, we reiterate that the optimal weighting function for the ethane dimer (here $\lambda = 0.5$ best reproduces the 2nd virial for the Born-Mayer-IP FF) is by no means universal for the molecules in the 91 dimer test set.

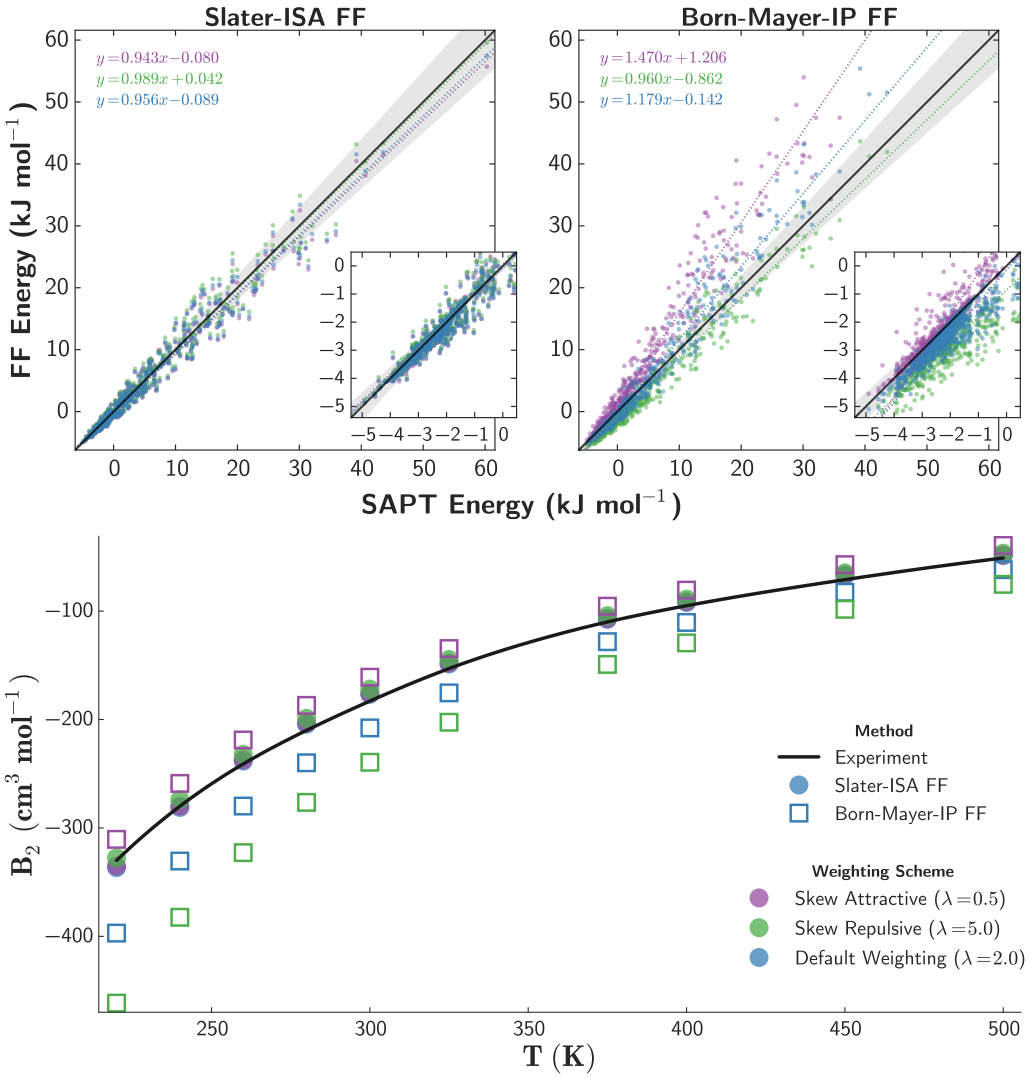


Figure 2.11: Comparison of the Slater-ISA FF and the Born-Mayer-IP FF in terms of sensitivity to the weighting function employed in parameter optimization for the ethane dimer. Three weighting functions, $\lambda = 0.5$ (purple), $\lambda = 2.0$ (blue), and $\lambda = 5.0$ (green) are shown, with higher λ values indicating more weighting of repulsive configurations.
(top) Total interaction energies for the Slater-ISA FF (left) and the Born-Mayer-IP FF (right) indicating the accuracy of each force field with respect to DFT-SAPT (PBE0/AC) benchmark energies. The diagonal line (black) indicates perfect agreement between reference energies and each force field, while shaded grey areas represent points within $\pm 10\%$ agreement of the benchmark. To guide the eye, a line of best fit (dotted line) has been computed for each force field and for each weighting function.
(bottom) Computed 2nd virial coefficients for ethane. Data for the Slater-ISA FF and the Born-Mayer-IP FF are depicted using shaded circles and open squares, respectively; colors for the different weighting functions are as above. Experimental data from Ref. 10 (black line) is also shown.

The Slater-ISA FF fits for the ethane dimer, on the other hand, are nearly completely insensitive to the weighting function, leading to little intrinsic uncertainty in the determination of parameters or in the computation of macroscopic properties. Some other dimers, particularly those where atomic anisotropy would be anticipated (e.g., water), exhibited slightly larger sensitivity to the weighting function. Nevertheless, the vast majority of dimers in the test set are qualitatively insensitive to the choice of weighting function, and can be optimized with the default $\lambda = 2.0$ weighting function without yielding undue systematic error in the attractive region of the potential, thus proving the enhanced robustness of the Slater-ISA FF model relative to conventional force fields.

2.4.5 Next-Generation Born-Mayer Models: Born-Mayer-sISA FF

We hypothesize that the increased accuracy, transferability, and robustness of the Slater-ISA FF is a direct result of its more physically-motivated functional form and its use of ISA-derived atomic exponents that directly account for the influence of the molecular environment. Nonetheless, we recognize that the standard Born-Mayer functional form remains extremely common, both in simulation software and in existing force fields. It is therefore fruitful to explore the extent to which the BS-ISA exponents themselves could be used in conjunction with a Born-Mayer functional form. These results are shown in Table 2.5.

Component	Dimer-Specific Fits			Transferable Fits		
	Slater-ISA FF (kJ mol^{-1})	Born-Mayer-ISA (kJ mol^{-1})	Born-Mayer-sISA (kJ mol^{-1})	Slater-ISA FF (kJ mol^{-1})	Born-Mayer-ISA (kJ mol^{-1})	Born-Mayer-sISA (kJ mol^{-1})
Exchange	2.641 (0.686)	7.030 (1.203)	2.677 (0.686)	2.718 (0.720)	6.968 (1.228)	2.764 (0.706)
Electrostatics	1.087 (0.351)	1.406 (0.589)	1.083 (0.352)	1.134 (0.351)	1.461 (0.598)	1.141 (0.352)
Induction	0.251 (0.095)	0.229 (0.097)	0.250 (0.096)	0.278 (0.101)	0.257 (0.101)	0.275 (0.101)
δHF	0.246 (0.068)	0.327 (0.120)	0.248 (0.068)	0.274 (0.076)	0.353 (0.122)	0.274 (0.076)
Dispersion	0.766 (0.317)	3.584 (0.890)	0.856 (0.336)	0.766 (0.317)	3.584 (0.890)	0.856 (0.336)
Total Energy						
RMSE	1.701 (0.464)	4.934 (1.054)	1.751 (0.453)	1.650 (0.456)	4.555 (1.035)	1.713 (0.446)
$\|MSE\|$	0.216 (0.057)	1.127 (0.505)	0.258 (0.063)	0.175 (0.051)	0.882 (0.516)	0.245 (0.057)

Table 2.5: Comparison of characteristic RMSE (as described in the main text) over the 91 dimer test set for the Born-Mayer-sISA approximation compared with other methods. For the total energy, both RMSE and absolute mean signed errors (MSE) have been shown. 'Attractive' RMSE, representing the characteristic RMSE for the subset of points whose energies are net attractive ($E_{int} < 0$), are shown in parentheses to the right of the total RMS errors; 'attractive' ||MSE|| are likewise displayed for the total energy. Slater-ISA FF, Born-Mayer-ISA, and Born-Mayer-sISA FF are as described in the main text, and the 'Dimer-Specific' and 'Transferable' fits are as described in Table 2.1.

As expected, direct insertion of the BS-ISA exponents into the Born-Mayer functional form (Born-Mayer-ISA) does not yield promising results. Indeed, the Born-Mayer-ISA FF has significantly worse RMSE and ||MSE|| than the Born-Mayer-IP FF. We reiterate that the $P = 1$ approximation from Eq. (2.25), yielding the conventional Born-Mayer form, is by itself a crude model. Rather, it becomes necessary to accompany this approximation by a corresponding exponent scale factor, ξ:

$$B_i = \xi B_i^{ISA}. \tag{2.39}$$

Following literature precedent,[125,141] we hypothesized that ξ could be treated as a universal constant. To test this conjecture, we computed reference density overlaps for a variety of isolated atom pairs (details in the Supporting Information of Ref. 95), and fitted each of these overlaps to a Born-Mayer function of the form $S_{ij} \approx K_{ij} \exp(-\xi B_{ij}^{ISA} r_{ij})$, where $K_{ij} = \frac{K}{B_{ij}^3}$ in line with Eq. (2.13). To very good approximation, both K and ξ can be treated as universal constants; that is, neither K nor ξ is sensitive to the value of B^{ISA}. However, fitted values of K and ξ do depend strongly on the range of r_{ij} values used in the optimization, yielding estimates ranging from 0.74 to 0.88.

As an alternative, we optimized ξ directly by minimizing RMSE against the 91 dimer test set. Results from various choices of ξ can be found in the Supporting Information of Ref. 95. In agreement with prior literature and our 'first-principles' analysis of overlaps, we find $\xi = 0.84$ to be optimal for minimizing characteristic overall and attractive RMSE, though in practice the errors are insensitive to $\xi \in [0.82, 0.86]$. We henceforth use $\xi = 0.84$ and refer to to this force field methodology (Born-Mayer functional form, ISA-derived exponents with scale factor $\xi = 0.84$) as the Born-Mayer-sISA FF. Parameters and homo-monomeric fits for the Born-Mayer-sISA FF can be found in the Supporting Information of Ref. 95 and in **??**.

From Table 2.5 we see that the Born-Mayer-sISA FF is comparable in quality to our original Slater-ISA FF methodology. For all attractive configurations, the Born-Mayer-sISA FF is equally accurate and transferable (Table 2.5). Furthermore,

as shown in Table 2.3, Born-Mayer-sISA FF displays similar parameter robustness to Slater-ISA FF. These results suggest that many of the advantages of the Slater-ISA FF procedure can be captured simply by using the (scaled) ISA exponents. Note, however, that the optimal scale factor likely exhibits some system dependence, and furthermore that the enhanced Slater functional form may be important where an accurate description of highly repulsive configurations is crucial.

We also examined the Slater-ISA FF and the Born-Mayer-sISA FF against force fields where B_i values were instead treated as soft constraints, rather than fixed parameters. Using entirely unconstrained exponents yields unphysical parameters and a severe degradation in force field transferability. Using exponents from the Slater-ISA FF and the Born-Mayer-sISA FF as Bayesian priors (in the sense used in Refs. 123, 155), we generated two new force fields with optimized exponents, denoted Slater-OPT and Born-Mayer-OPT, respectively. Characteristic RMSE and ||MSE|| for these force fields can be found in the Supporting Information of Ref. 95. We find that both methods yield only very minimal improvement, suggesting that the first-principles ISA exponents are already nearly optimal. Comparing the Born-Mayer-OPT exponents to those from Slater-ISA, we find a nearly identical average scale factor of $\gamma = 0.83 \pm 0.07$. Given that these optimal exponents can now be generated directly from first principles calculations of the molecular densities via the BS-ISA approach of Misquitta et al., we anticipate that the BS-ISA densities and resulting ISA exponents will be extremely useful in next-generation force field development in order to greatly simplify force field parameterization.

2.5 Conclusions and Recommendations

We have presented a new methodology for describing short-range intermolecular interactions based upon a simple model of atom-in-molecule electron density overlap. The resulting Slater-ISA FF is a simple extension of the conventional Born-Mayer functional form, supplemented with atomic exponents determined from an ISA analysis of the molecular electron density. In contrast to simple Born-Mayer or Lennard-Jones models, the Slater-ISA FF is capable of reproducing ab initio interac-

tion energies over a wide range of inter-atomic distances, and displays extremely low sensitivity to the details of parameterization. Furthermore, the Slater-ISA FF exhibits excellent parameter transferability. We thus recommend Slater-ISA FF for use in the development of future ab initio (and possibly empirically-parameterized) potentials, particularly where accuracy across wide regions of the potential surface is paramount.

More generally, we find that analysis of the ISA densities provides an excellent first-principles procedure for the determination of atomic-density decay exponents. This analysis improves upon existing approaches (which rely upon exponents derived from atomic radii or ionization potentials)[139,198–200] and explicitly incorporates the influence of the molecular environment. These exponents can be used within Slater-ISA FF without further parameterization. Alternatively, in conjunction with an appropriate scale factor, the exponents can be used to enhance the accuracy of standard Born-Mayer potentials and/or Tang-Toennies damping functions. The resulting Born-Mayer-sISA FF retains many of the advantages of Slater-ISA FF, but also maintains compatibility with existing force fields and simulations packages that do not support the Slater functional form. Given that the BS-ISA exponents appear to be essentially optimal with respect to additional empirical optimization, we strongly recommend use of these first-principles exponents in order to simplify (both ab initio and empirical) future force field development involving Born-Mayer or related functional forms.[112]

Overall, Slater-ISA FF enables a significantly increase in force field accuracy, particularly in describing short intermolecular contacts. Nevertheless, the neglect of atomic anisotropy remains, in some cases, a severe approximation.[201–203] Indeed, it has been shown by many authors[78,151,158,193] that quantitatively accurate A_{ij} parameters (and to a lesser extent, B_{ij} parameters) require incorporation of angular dependence for the generation of highly-accurate force fields. This anisotropy becomes crucial when describing systems containing lone pairs, hydrogen bonds, and/or π-interactions. Promisingly, BS-ISA densities naturally describe such anisotropy,[123,155,204] and a straightforward method for its inclusion (where essential) in ab initio force fields is the subject of Chapter 3.

2.A Waldman-Hagler Analysis of B_{ij} Combination Rule

The exact expressions for the overlap of two Slater densities $\rho_i = D_i \exp(-B_i r)$ and $\rho_j = D_j \exp(-B_j r)$ are shown here, first in the limiting case where the two exponents are equal ($B_i = B_j = B_{ij}$):

$$\begin{aligned} S^{ij}_{B_i=B_j} &= D_{ij} P(B_{ij}, r_{ij}) \exp(-B_{ij} r_{ij}) \\ D_{ij} &= \pi D_i D_j B_{ij}^{-3} \\ P(B_{ij}, r_{ij}) &= \frac{1}{3}(B_{ij} r_{ij})^2 + B_{ij} r_{ij} + 1, \end{aligned} \tag{2.40}$$

and second in the case where $B_i \neq B_j$:

$$\begin{aligned} S^{ij}_{B_i \neq B_j} =& \frac{16\pi D_i D_j \exp(-\{B_i + B_j\} r_{ij}/2)}{(B_i^2 - B_j^2)^3 r_{ij}} \times \\ & \left[\left(\frac{B_i - B_j}{2} \right)^2 \left(\exp\left(\{B_i - B_j\} \frac{r_{ij}}{2} \right) - \exp\left(-\{B_i - B_j\} \frac{r_{ij}}{2} \right) \right) \right. \\ & \quad \times \left(\left(\frac{B_i + B_j}{2} \right)^2 r_{ij}^2 + (B_i + B_j) r_{ij} + 2 \right) \\ & - \left(\frac{B_i + B_j}{2} \right)^2 \exp\left(\{B_i - B_j\} \frac{r_{ij}}{2} \right) \times \left(\left(\frac{B_i - B_j}{2} \right)^2 r_{ij}^2 - (B_i - B_j) r_{ij} + 2 \right) \\ & \left. + \left(\frac{B_i + B_j}{2} \right)^2 \exp\left(-\{B_i - B_j\} \frac{r_{ij}}{2} \right) \times \left(\left(\frac{B_i - B_j}{2} \right)^2 r_{ij}^2 + (B_i - B_j) r_{ij} + 2 \right) \right]. \end{aligned} \tag{2.41}$$

Each overlap formula has been given a subscript to indicate limits on B_i and B_j.

Our goal is to ascertain the extent to which $S^{ij}_{B_i \neq B_j}$ can be accurately modeled by the functional form and variables of $S^{ij}_{B_i = B_j}$. D_i and D_j are pre-factors appearing in both equations, and we set these variables to unity without loss of generality. To

find values of B_{ij} such that $S^{ij}_{B_i \neq B_j}(B_i, B_j, r_{ij}) \approx S^{ij}_{B_i = B_j}(B_{ij}, r_{ij})$, we first treat B_{ij} as a completely adjustable parameter, and later test for the existence of some simple combining function f such that $B_{ij} = f(B_i, B_j)$.

To optimize B_{ij}, we first require a training set of relevant $S^{ij}_{B_i \neq B_j}$ values. B_i, B_j, and r_{ij} are the only variables appearing in $S^{ij}_{B_i \neq B_j}$, and we could in principle fit B_{ij} values over a grid of B_i, B_j and r_{ij} combinations. However, we are only interested in the subset of points which are chemically relevant. Consequently, we developed a library of B_i values by deriving exponents from the ionization potentials of the first three rows of the periodic table (plus bromine and iodine). For each pair of elements, $B = 2\sqrt{2IP}$,[166] and a range of r_{ij} values corresponding to 0.8-1.2 times the sum of the van der Waals radii of the two atoms was selected. B_{ij} values in $S^{ij}_{B_i = B_j}$ were then optimized (in a least-squares sense) for each element pair separately; Mean absolute percent errors (MAPE) for fitted overlaps are shown in Fig. 2.12.

Relative errors for fitting are acceptably small for all element pairs. Excluding certain noble gases and alkali metals (He, Li, Ne, Na) from consideration, these being the elements with the most disparate B_i values compared to other elements, MAPE drops below 3% for all pairs, with the vast majority of MAPE below 1%. Our focus in this work is primarily on organic compounds where $|B_i - B_j|$ is small; empirically, these errors always translate to very small errors in the exchange energy itself. Use of an effective B_{ij} may require further testing in cases with extremely disparate B_i and B_j values.

We next tested whether the optimized B_{ij} could instead be modeled by a combination rule $B_{ij} = f(B_i, B_j)$. On the basis of symmetry and scaling considerations, Waldman and Hagler demonstrate that if a combination rule $f(B_i, B_j)$ exists, a plot of B_{ij}/B_i vs. B_j/B_i should lie on a single curve.[140] Remarkably (see 2.13), a geometric mean combination rule $B_{ij} = \sqrt{B_i B_j}$ models the fitted B_{ij} values near quantitatively. This result allows the computation of Slater overlaps using the much simpler form of $S^{ij}_{B_i = B_j}$ (2.40) from individual atoms-in-molecule exponents B_i and B_j.

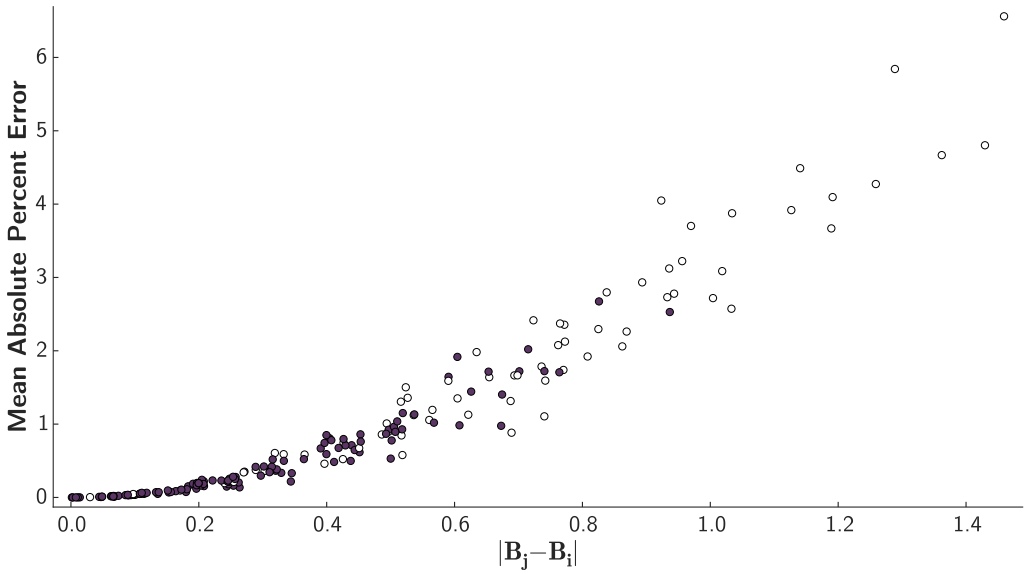


Figure 2.12: Mean absolute percent error of fitted overlap values as a function of the absolute difference between B_i and B_j values. Element pairs containing He, Li, Ne and/or Na are shown as empty circles. Deviations below 1% are seen for most element pairs, with noble gases and alkali metals posing a more significant challenge. Scatter in the plot is due to small variations in the absolute values of r_{ij} fit for each pair. As expected, $S^{ij}_{B_i \neq B_j}$ and $S^{ij}_{B_i = B_j}$ closely agree for $|B_i - B_j| \approx 0$.

2.B Force Field Fits for Homomonomeric Systems

Scatter plots are shown for each homomonomeric system as an indication of force field quality with respect to DFT-SAPT (PBE0/AC) benchmark energies (Fig. 3.7). As in the main text, fits for each energy component are displayed along with two views of the total interaction energy.

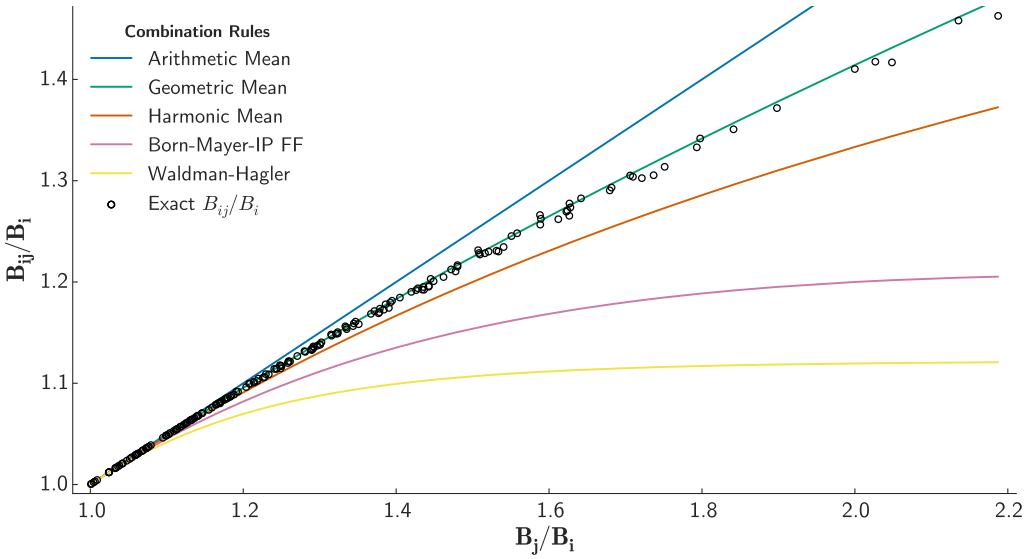
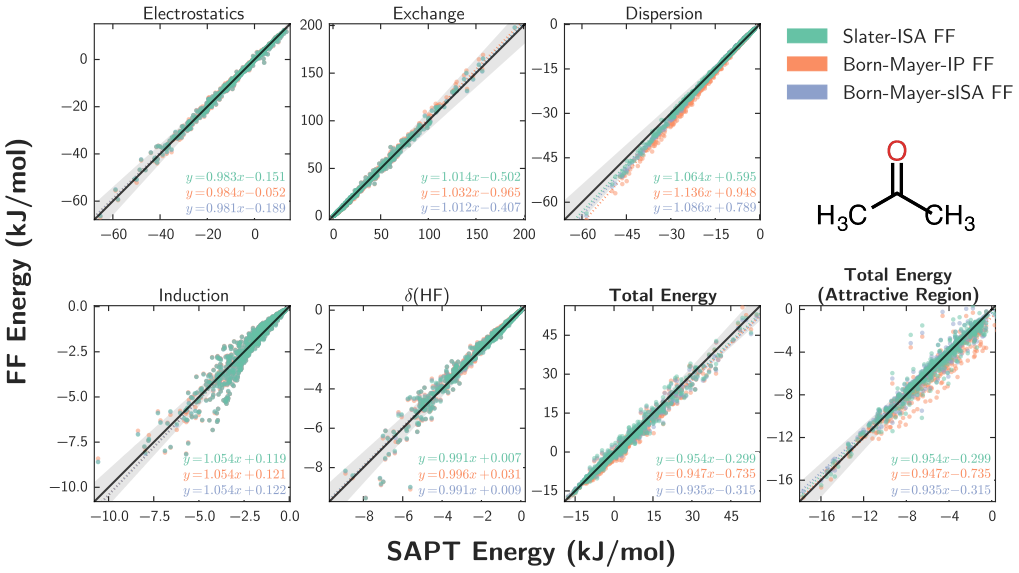
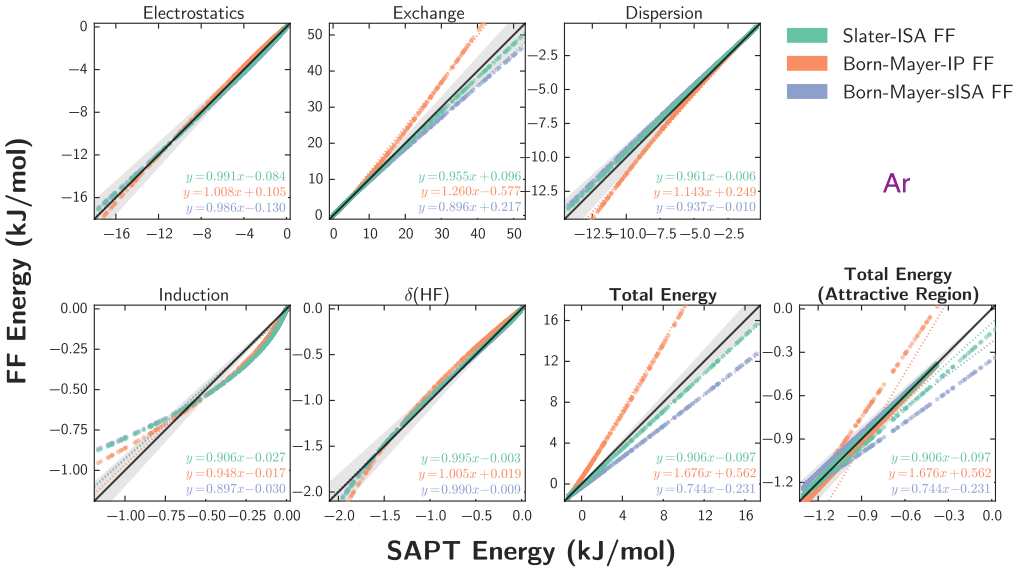


Figure 2.13: Waldman-Hagler-style analsysis of possible B_{ij} combination rules. Exact B_{ij} values are derived from fitting an approximate overlap density of the form $S_{ij} = A_{ij}K_2(r_{ij})\exp(-B_{ij}r_{ij})$ to the exact overlap density (as given by Rosen and by Tai[11,12]) of two distinct Slater orbitals whose exponents correspond to atomic exponents for the elements H-Ar, Cl, Br, and I. For each overlap pair, a range of r_{ij} values was used from 0.8 to 1.2 times the sum of the pair's van der Waals radii. The geometric mean combination rule $B_{ij} = \sqrt{B_i B_j}$ models the exact B_{ij} values with near-perfect agreement, justifying our choice of combination rule.

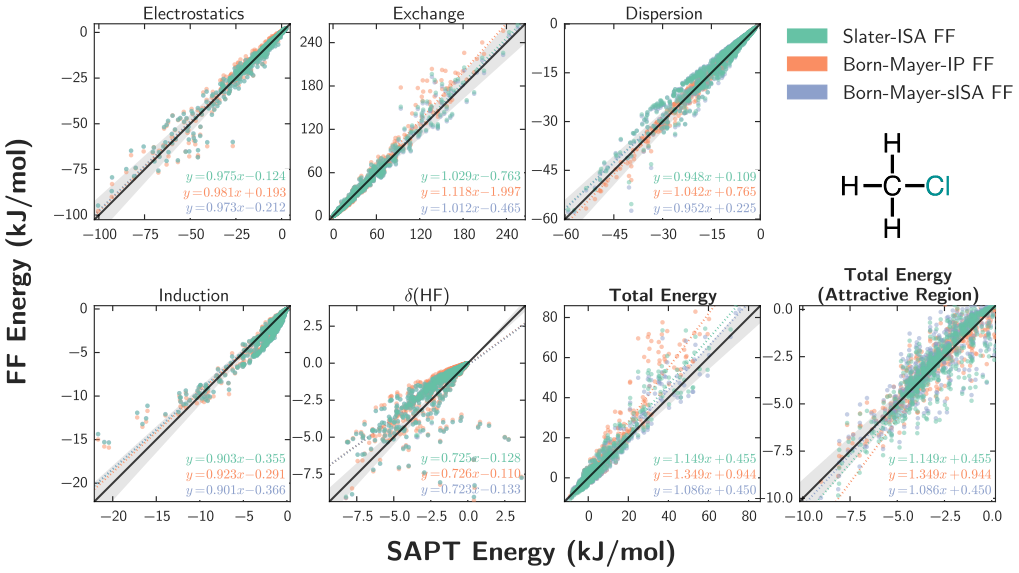
(a) Acetone Dimer



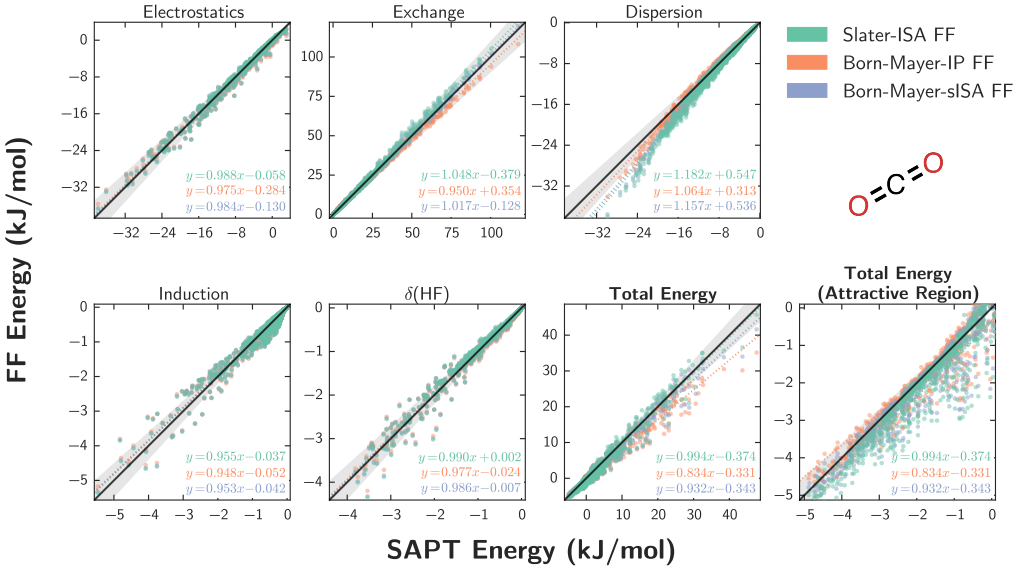
(b) Ar Dimer



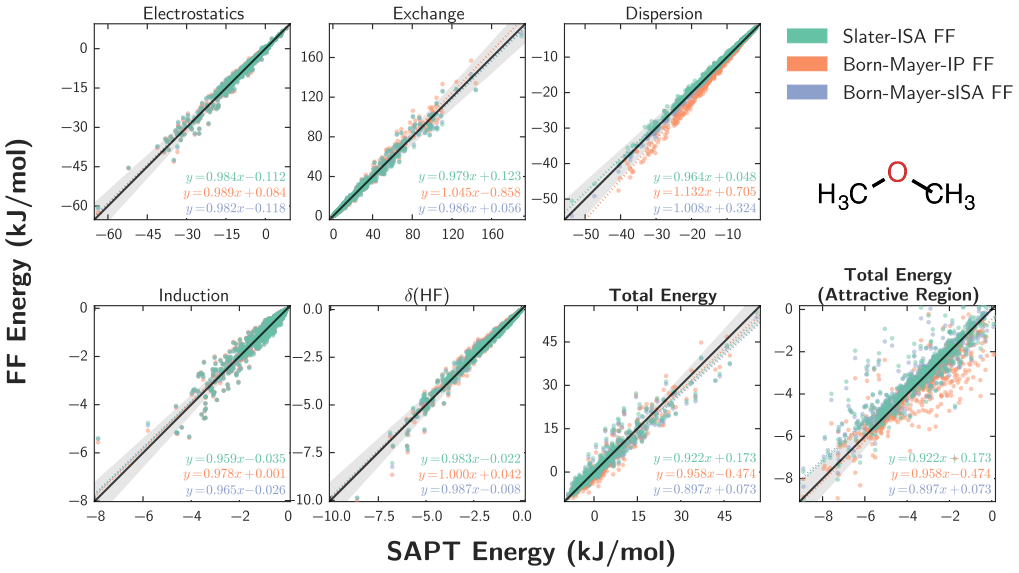
(c) Chloromethane Dimer



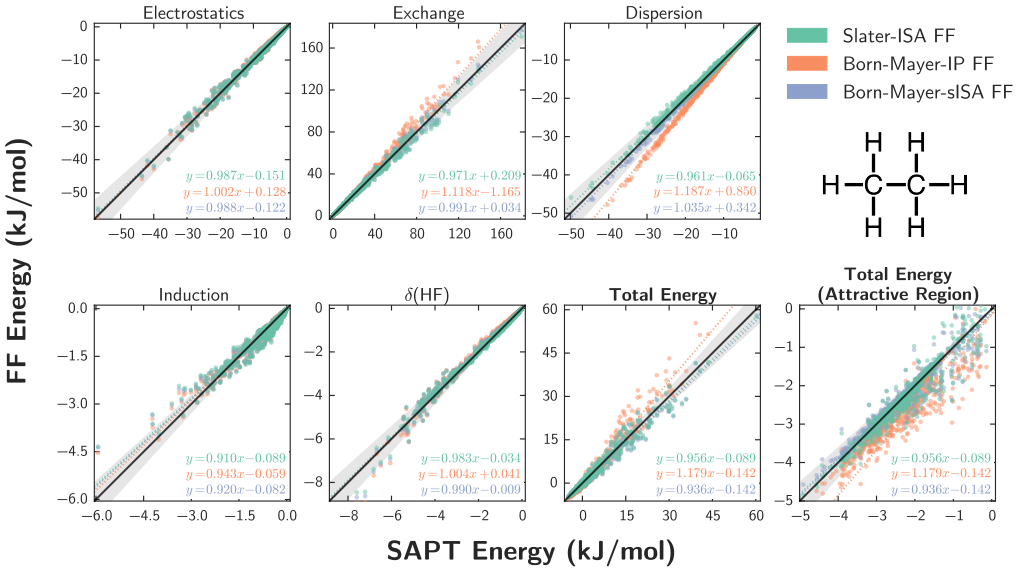
(d) CO_2 Dimer



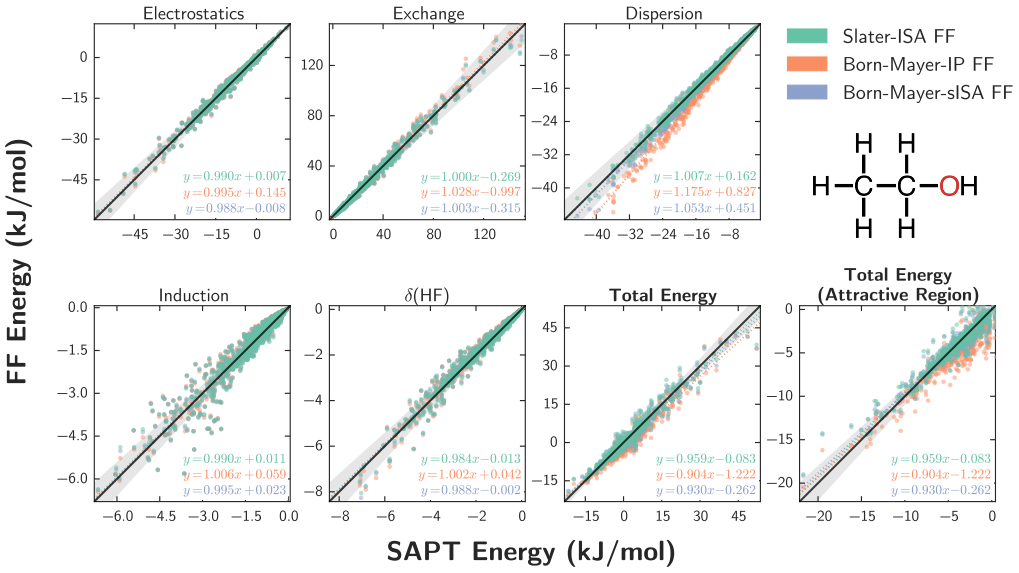
(e) Dimethyl Ether Dimer



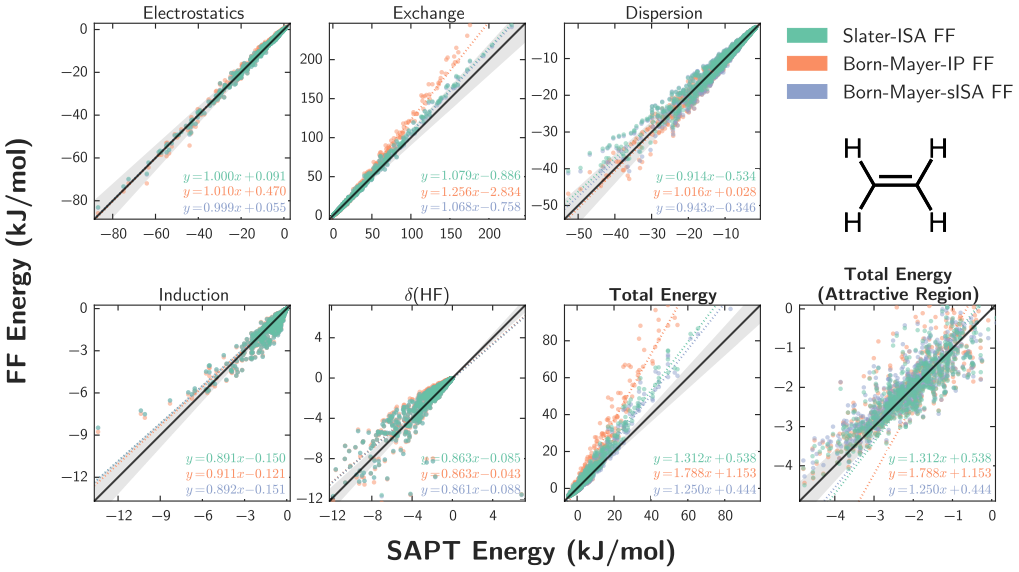
(f) Ethane Dimer



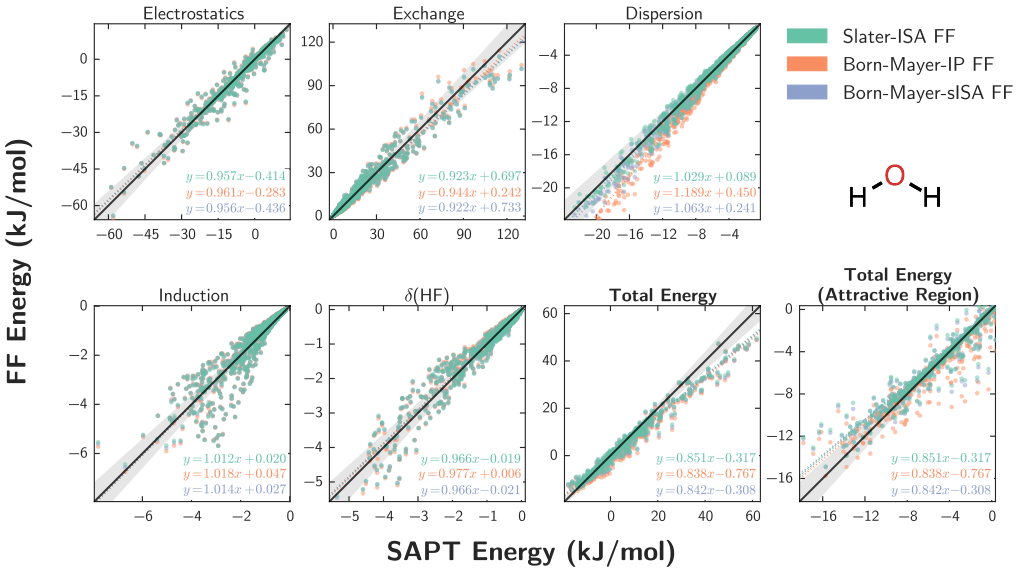
(g) Ethanol Dimer



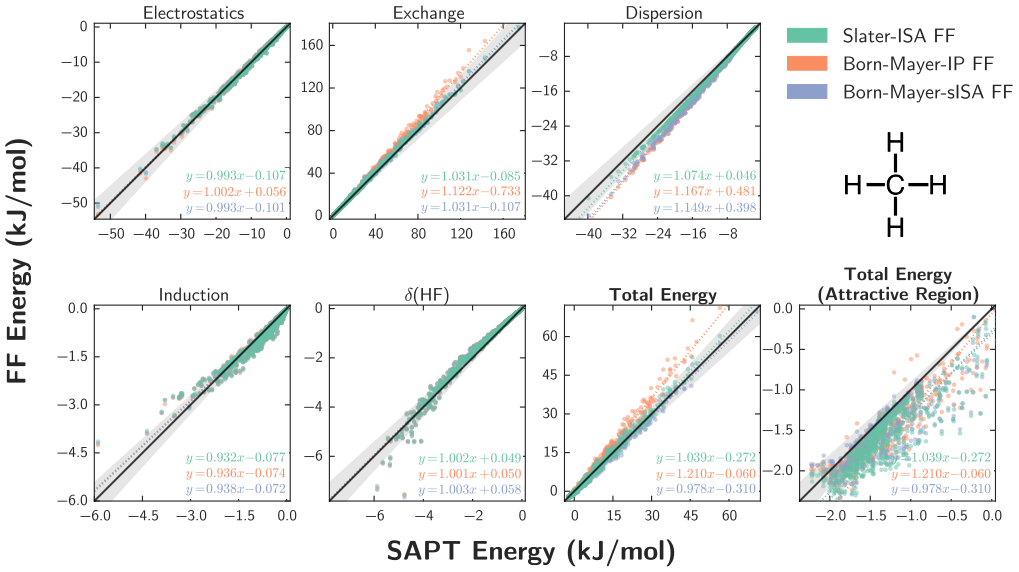
(h) Ethene Dimer



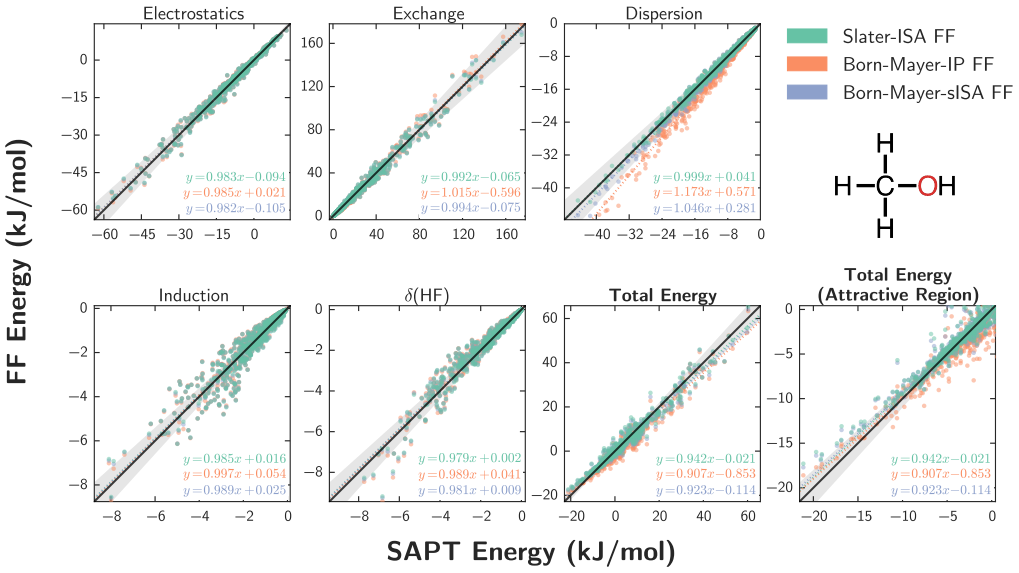
(i) H_2O Dimer



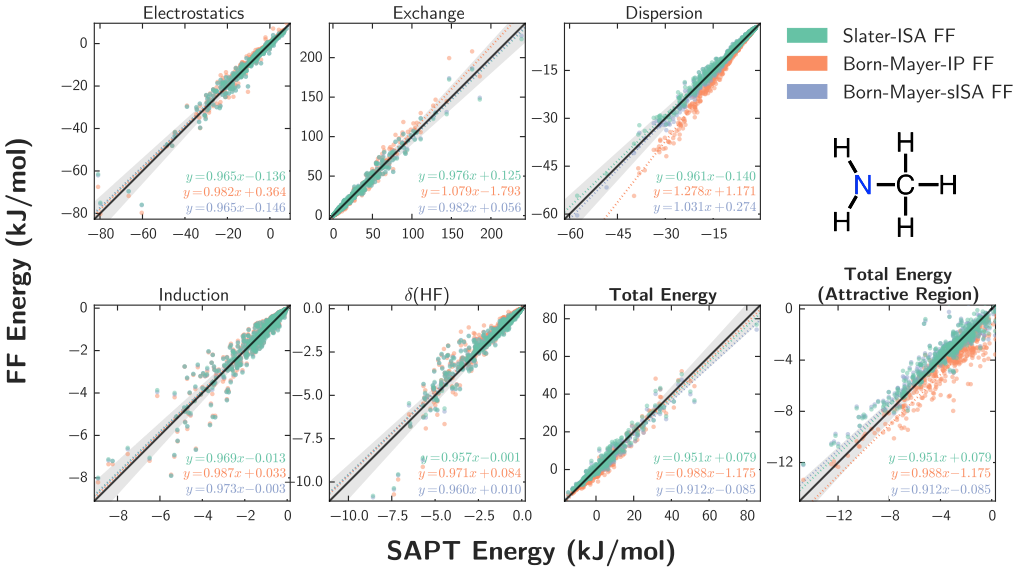
(j) Methane Dimer



(k) Methanol Dimer



(l) Methyl Amine Dimer



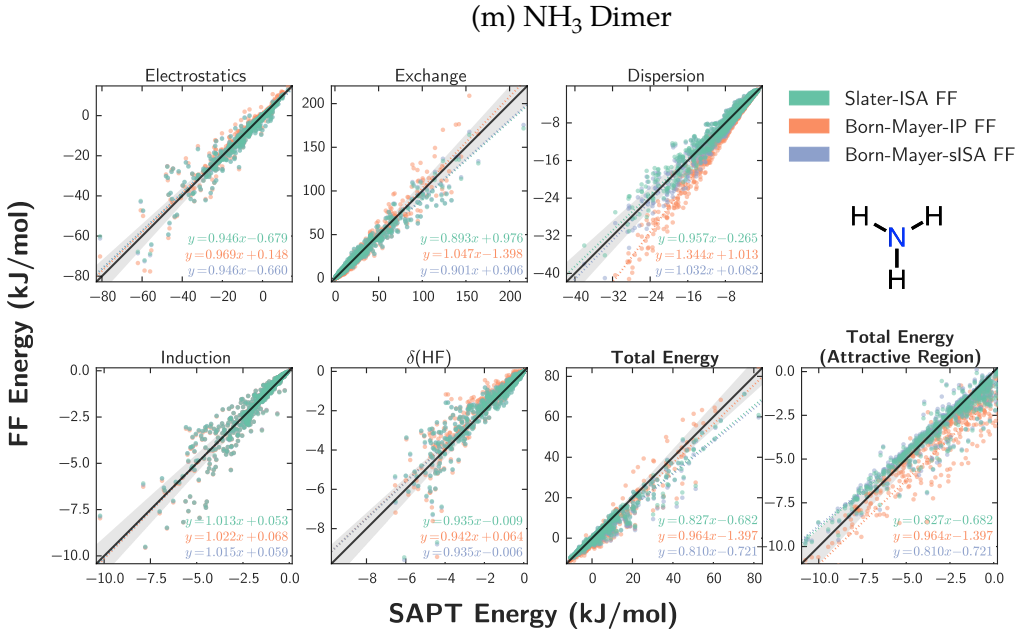


Figure 2.14: Force field fits for the homomoneric systems using the Slater-ISA FF (green), Born-Mayer-IP FF (orange) and Born-Mayer-sISA FF (blue). Fits for each energy component are displayed along with two views of the total interaction energy. The $y = x$ line (black) indicates perfect agreement between reference energies and each force field, while shaded grey areas represent points within $\pm 10\%$ agreement of the benchmark. To guide the eye, a line of best fit (dotted line) has been computed for each force field and for each energy component.

3 MASTIFF: A GENERAL APPROACH FOR INCORPORATING ATOMIC-LEVEL ANISOTROPY IN AB INITIO FORCE FIELDS

3.1 Introduction

Classical molecular simulation is a standard tool for interpreting and predicting the chemistry of an incredible host of systems ranging from simple liquids to complex materials and biomolecules. Such simulations always require, as input, a mathematical description of the system's potential energy surface (PES). In principle, the PES for most chemical systems can accurately be determined from one of several high-level electronic structure methods;[74,205,206] nevertheless, these calculations are currently too expensive to use in simulations of large systems and/or long timescales.[207] Consequently, most routine molecular simulation today is performed with the aid of force fields: computationally-inexpensive, parameterized mathematical expressions that approximate the exact PES. Because the accuracy and predictive capabilities of molecular simulation are directly tied to the underlying force field, one of the central challenges of molecular simulation is the development of highly accurate force fields. For ab initio force field development, this accuracy is principally defined by a force field's fidelity to the underlying exact PES.

As of now, several common shortcomings[62] inhibit the accuracy and predictive capabilities of standard ab initio force fields, and these limitations must be systematically addressed in order to generate improved, 'next-generation' force fields. One important shortcoming, which will be the focus of this Chapter, is the so-called 'sum-of-spheres' approximation,[208] in which it is assumed that the non-bonding interactions between molecules can be treated as a superposition of interactions between pairs of spherically-symmetric atoms. Put differently, the sum-of-spheres, or 'isotropic atom-atom', approximation assumes that the exact PES, E_{int} (which depends both on the center of mass distance R and relative orientation Ω between

molecules), can be modeled as

$$E_{int}(R, \Omega) \approx \sum_{ij} f(r_{ij}) \equiv V_{FF}, \tag{3.1}$$

where the above sum runs over all non-bonded pairs of atoms i and j with interatomic separation r_{ij}, and $f(r_{ij})$ is an arbitrary, distance-dependent function that defines the pairwise interaction. Here and throughout, we use E to denote the true PES, and V to denote the corresponding model/force field prediction. With some exceptions (vida infra), nearly all standard intermolecular force fields — ranging from the popular "Lennard-Jones plus point charges" model to more complex functional forms[75] — explicitly make use of the isotropic atom-atom model.

Notwithstanding the popularity of the model, there is good experimental and theoretical evidence to suggest that the sum-of-spheres approximation does not hold in practice.[78,208,209] Importantly, and as we argue in Section 3.5, models which include anisotropic (multipolar) electrostatics, but otherwise employ the sum-of-spheres approximation, are an improved but *still incomplete* model for describing the atomic-level anisotropy of intermolecular interactions. Experimentally, it has long been known that atom-in-molecule charge densities, as determined from x-ray diffraction, can exhibit significant non-spherical features, such as with lone pair or π electron densities.[210] Furthermore, statistical analyses of the Cambridge Structural Database have shown that the the van der Waals radii of atoms-in-molecules (as measured from interatomic closest contact distances) are not isotropically distributed, but rather show strong orientation dependencies, particularly for halogens and other heteroatoms.[201,211–215] These experimental studies are corroborated by a significant body of theoretical research on both the anisotropy of the atomic van der Waals radii as well as the non-spherical features of the atomic charge densities themselves.[203,204,215–218] These studies suggest that the sum-of-spheres approximation is an insufficiently flexible model for the subset of intermolecular interactions that arise from atomically non-spherical charge densities, and may help explain known difficulties in generating accurate isotropic atom-atom force fields for such important chemical interactions as π-interactions,[61,219,220] σ-bonding,[221–223] and hydrogen

bonding,[58] (see Ref. 64 and references therein).

Motivated by these observations, a small but important body of work has been devoted to directly addressing the limitations of the isotropic atom-atom model in the context of 'next-generation' force field development. As will be discussed in detail below (see Section 3.2), the general conclusion from these studies is that many components of intermolecular interactions (specifically electrostatics, exchange-repulsion, induction, and dispersion) can be more accurately modeled by functional forms that go beyond the sum-of-spheres approximation.[209,224,225] While few intermolecular potentials (and virtually no standard force fields amenable to routine molecular simulation) explicitly account for atomic-level anisotropy for each component of intermolecular interactions, several recent standard force fields have incorporated atomic-level anisotropy into their description of long-range electrostatics.[64] Some of these potentials (notably AMOEBA[225–227] and some water potentials[58,64]) are already employed in large-scale molecular simulation, often with very encouraging success.[64] Furthermore, others have shown that anisotropic potentials (some of which additionally model the anisotropy of exchange-repulsion and/or dispersion) lead to significant improvements in predicting molecular crystal structures.[64,99,151,191,228–230] These and other results strongly suggest that a complete incorporation of atomic anistropy into next-generation force fields will lead to increasingly accurate and predictive molecular simulations in a wider variety of chemical interactions.[224]

Given the importance of atomic-level anisotropy in defining intermolecular interactions, and the critical role that computationally-affordable standard force fields play in enabling molecular simulation, our present goal is to develop a general methodology for standard force field development that can both universally account for atomic-level anisotropy in all components of intermolecular interactions *and* that can be routinely employed in large-scale molecular simulation. Furthermore, and in line with our usual goals for force field development,[75] our aim is to develop a first-principles-based model that is as accurate and transferable as possible, all while maintaining a simple, computationally-tractable functional form that allows for robust parameterization and avoids over-/under-fitting. Thus, building on

prior work (both our own[54,75,91,95] and from other groups[209]), we present here a general ansatz for anisotropic force field development that, at minimal computational overhead, incorporates atomic-level anisotropy into all aspects of intermolecular interactions (electrostatics, exchange, induction, and dispersion) and that accounts for this anisotropy, not only in the asymptotic limit of large intermolecular separations, but also in the region of non-negligible electron density overlap. After motivating and establishing the functional forms used in our anisotropic force fields, we next demonstrate, using a large library of dimer interactions between organic molecules, the excellent accuracy and transferability of these new force fields with respect to the reproduction of high-quality ab initio potential energy surfaces. Lastly, we showcase how these new force fields can be used in molecular simulation, and benchmark the accuracy of our models with regards to a variety of experimental properties. The theory and results presented in this Chapter should be of general utility in improving the accuracy of (particularly ab initio generated) force fields, such that the complex, inherently anisotropic details of intermolecular interactions may eventually be routinely incorporated into increasingly rigorous and predictive molecular simulation.

3.2 Background

Before presenting our development methodology for atomically-anisotropic potentials, we provide the reader with a summary of prior approaches to ab initio force field development and to models going beyond the sum-of-spheres approximation. In discussing the effects of anisotropic charge distributions on intermolecular potentials, here and throughout we employ the fairly standard[231] decomposition of interaction energies into physically-meaningful components of electrostatics, exchange-repulsion, induction (which includes both polarization and charge-transfer), and dispersion. Many studies on atomically-anisotropic force field development have focused on incorporating anisotropy on a component-by-component basis, and so for clarity we discuss anisotropic modeling schemes for each energy component individually. As in Chapter 2,[95] we find it useful to separate and discuss in turn

the so-called 'long-range' effects (multipolar electrostatics, polarization, and dispersion) from those 'short-range' effects that arise only at shorter intermolecular separations due to the non-negligible overlap of monomer electron densities (e.g. charge penetration and exchange-repulsion). Finally, we take advantage of the many-body expansion[54,76] to separately consider into two- and many-body contributions, and primarily focus our discussion on improvements to the two-body interaction energies themselves.

3.2.1 Prior Models for Long-Range Interactions

The importance of atomic-level anisotropy in modeling long-range interactions, particularly as it pertains to electrostatics, is quite well studied. A number of groups have found that using atomic multipoles (rather than simple point charges) greatly improves both the electrostatic potential[216,232] and the resulting electrostatic interaction energies.[64,70,127,150,225,227,233,234] Though not without additional computational cost, atomic multipoles are now routinely employed in a number of popular force fields.[58,225,227] As an alternate, and often more computationally-affordable approach, other groups have used off-atom point charges to effectively account for anisotropic charge densities.[129,222,235,236] In line with chemical intuition, improvements from use of atomic multipoles/off-site charges are often particularly important in describing the electric fields generated by heteroatoms and carbons in multiple bonding environments.[237,238]

The induction and dispersion energies have also been shown to exhibit anisotropies that go beyond the sum-of-spheres model. For instance, it has been suggested that anisotropic polarizabilities (which effect both polarization and dispersion) are required to avoid an artificial over-stabilization of base stacking energies in biomolecules.[220] In order to more accurately treat polarization, several molecular mechanics potentials have made use of either off-site[239] or explicitly anisotropic polarizabilities.[236,240]. Similarly, the importance of anisotropic dispersion interactions has also been established,[54,89,90,241,242] particularly for π-stacking interactions,[62,220] and select potentials have incorporated directional dependence into the functional

form for dispersion by expanding dispersion coefficients in terms of spherical harmonics or, more generally, $\bar{S}$-functions (discussed in Section 3.A).[90,99]

3.2.2 Prior Models for Short-Range Interactions

At closer intermolecular separations, where overlapping electron densities between monomers leads to exchange-repulsion and charge-penetration effects, anisotropy is also important. Exchange-repulsion has known orientation dependencies which can play a quantitative role in halogen bonding[221,243] and other chemical interactions, and many authors have worked on developing different models for describing the anisotropy of exchange-repulsion. Some potentials (albeit not those that are amenable to large-scale molecular simulation) have numerically computed overlap integrals that can be used in conjunction with he density-overlap model popularised by Wheatley and Price[122,124,125,147,158] to quantify anisotropic exchange-repulsion, charge transfer, and/or charge penetration interactions.[127,233,234,244–246] Taking a more analytical approach, many other potentials have extended the Born–Mayer functional form[108] to allow for orientation-dependent pre-factors,[54,99,151,153,208,209,228,246,247] and model short-range effects using an anisotropic functional form originally proposed by Stone and Price:

$$V_{ij}^{exch} = G \exp[-\alpha_{ij}(R_{ij} - \rho_{ij}(\Omega_{ij}))]. \tag{3.2}$$

Here G is not a parameter, but rather an energy unit,[78] and α and ρ represent, respectively, the hardness and shape of the potential. In principle, one might also allow α to have orientation dependence, however this seems unnecessary in practice, as the hardness of the potential has been empirically found to behave more isotropically than its shape.[78] Similar to treatments of anisotropic electrostatics, this functional form typically expresses orientation dependence, Ω_{ij}, in terms of spherical harmonics and/or $\bar{S}$-functions.[78]

Finally, we note that, aside from exchange-repulsion, we are aware of relatively little research on the development of simple analytical expressions for the anisotropy

of other overlap effects, such as electrostatic/inductive charge penetration, charge-transfer, or short-range dispersion.

3.3 Theory and Motivation

Building on the extensive prior work that has led to a better understanding of atomic-level anisotropy and its effect(s) on intermolecular interactions, we now outline a methodology whereby atomic-level anisotropy can be incorporated into standard force fields amenable to large-scale molecular simulation. In particular, we aim to present a general methodology that optimally incorporates atomically-anisotropic effects given the following goals for ab initio force field development:

1. **Chemical accuracy with respect to ab initio benchmarks:** For systems that can be directly parameterized against high quality ab initio PES, the force field should exhibit chemical accuracy (average errors smaller than 1 kJ mol^{-1}) with respect to the ab initio benchmark; furthermore, any errors in the force field should be random rather than systematic

2. **Transferability across chemical environments:** Given force fields for two different pure systems, we should be able to accurately calculate (via simple combination rules and without additional parameterization) the PES of any system that is a mixture of the pure systems

3. **Simplicity:** The force field should be restricted to functional forms that are already compatible with, or could be easily implemented in, common molecular simulation packages

4. **Computational tractability:** The force field should be of minimal computational cost relative to existing polarizable multipolar force fields[227]

Given these goals, we now outline a detailed methodology for incorporating atomic-level anisotropy into each component (electrostatic, exchange-repulsion, induction, and dispersion) of intermolecular interactions, beginning with a new

model for short-range overlap effects and concluding with some new and/or revised theories for treating long-range interactions.

3.3.1 Anisotropic Models for Short-Range Interactions

Exchange-Repulsion

We begin by considering the exchange-repulsion, E_{ij}^{exch} that arises from the overlap of electron densities from two non-spherical atoms-in-molecules, i and j. Here and throughout, we closely follow the notation and theory used by Stone.[78] Without loss of generality, we can express the exchange repulsion between these two atoms as a function of their interatomic distance, r_{ij}, and relative orientation, Ω_{ij}. Furthermore, we can mathematically describe this relative orientation by assigning local coordinate axes to each i and j, such that the exchange energy is given by

$$E_{ij}^{exch}(r_{ij}, \Omega_{ij}) \equiv E_{ij}^{exch}(r_{ij}, \theta_i, \phi_i, \theta_j, \phi_j), \tag{3.3}$$

where θ_i and ϕ_i are the polar coordinates, expressed in the local coordinate system of atom i, that describe the position of atom j. Correspondingly, θ_j and ϕ_j define the position of i in terms of the local coordinate system of j. In principle the choice of these local coordinate frames is arbitrary. However, for the models introduced below, parameterization can be dramatically simplified by taking advantage of the local symmetry of an atom in its molecular environment and aligning the local coordinate frame with the principal axis of this local symmetry.[78] Some examples of these local axes are shown in Fig. 3.1.

We next make an ansatz that Eq. (3.3) is separable into radial- and angular-dependent contributions,

$$E_{ij}^{exch}(r_{ij}, \theta_i, \phi_i, \theta_j, \phi_j) \approx V_{ij}^{exch}(r_{ij}, \theta_i, \phi_i, \theta_j, \phi_j) = f(r_{ij})g(\theta_i, \phi_i, \theta_j, \phi_j) \tag{3.4}$$

thus subdividing the problem of finding a general functional form for E_{ij}^{exch} into two more tractable tasks. First, we must find an ideal sum-of-spheres model to describe

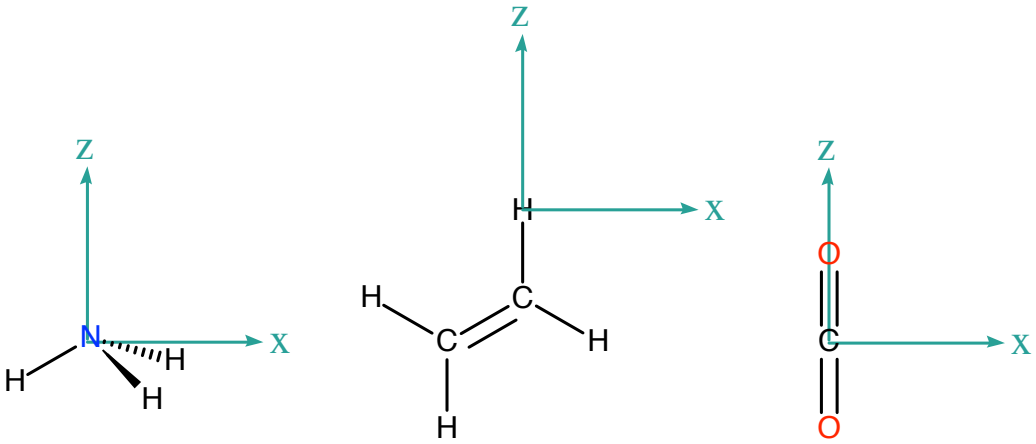


Figure 3.1: Local axis system, shown for select atoms in molecules.

the radial (isotropic) dependence of the force field, and second, we must find a way to model the orientation dependence as a multiplicative pre-factor to $f(r_{ij})$.

Given that the only requirement for $f(r_{ij})$ is that it be isotropic, how should a suitable model for $f(r_{ij})$ be chosen? Indeed, all standard isotropic force fields are of this general form, and thus might serve as a suitable starting point for anisotropic force field development. For reasons discussed below, in this Chapter we employ a simple and accurate model (Slater-ISA FF) from Chapter 2 for $f(r_{ij})$. This model can be derived from first-principles by approximating E_{ij}^{exch} as proportional to the overlap between spherically-symmetric atom-in-molecule (AIM) electron densities, each with density

$$\rho_i(r) = D_i \exp^{-B_i r}, \tag{3.5}$$

where D_i and B_i are both atom type-specific constants that can be parameterized from molecular electron densities and that represent, respectively, the shape and hardness of the AIM density. Using this approximation to the overlap model,[54,122,124,125,151,158–160]

the exchange energy between two atoms is then modeled by

$$\begin{aligned} E_{ij}^{exch} \approx V_{ij}^{exch} &\propto S_{\rho}^{ij} \\ &\approx A_{ij}^{exch} \left(\frac{(B_{ij} r_{ij})^2}{3} + B_{ij} r_{ij} + 1 \right) \exp(-B_{ij} r_{ij}) \end{aligned} \tag{3.6}$$

with combining rules

$$\begin{aligned} A_{ij}^{exch} &\equiv A_{i}^{exch} A_{j}^{exch}, \\ B_{ij} &\equiv \sqrt{B_i B_j}. \end{aligned} \tag{3.7}$$

S_{ρ}^{ij} is the electron density overlap between atoms and A_{ij} is a fitted proportionality constant.

Here and throughout we use Eq. (3.6) as our model for $f(r_{ij})$. This choice is primarily justified in Chapter 2 by the previously-demonstrated accuracy of the Slater-ISA formalism as compared to other sum-of-spheres models for repulsion.[95] Furthermore, and especially for simple test cases where one might expect the sum-of-spheres approximation to hold (such as with argon, methane, or ethane), we have shown (see Chapter 2) that the Slater-ISA FF correctly models intermolecular potential energy surfaces for a sizable library of intermolecular interactions over the asymptotic, attractive, and repulsive regions of the PES.

In addition to this empirical motivation for using the Slater-ISA formalism, there are good theoretical grounds to utilize it as a model for $f(r_{ij})$. Specifically, the AIM densities used to parameterize Slater-ISA FF are partitioned using an iterated stockholder atoms (ISA) procedure, and the resulting density profiles are guaranteed to be maximally spherical.[91–93] This condition of 'maximum sphericity' has two consequences. First, it suggests that the resulting Slater-ISA FF should be an optimal, or nearly optimal, isotropic atom-atom model. In other words, there is good reason to hope that our model for $f(r_{ij})$ completely accounts for the radial dependence of the potential, and consequently that models for $g(\theta_i, \phi_i, \theta_j, \phi_j)$ will truly represent the orientation dependence rather than simply over-fitting residual

errors from the radial functional form, thus retaining high transferability. Second, and relatedly, having maximally-spherical ISA densities suggests that anisotropic effects should be a minimal perturbation to the PES. This means that, to a first-order approximation, $g(\theta_i, \phi_i, \theta_j, \phi_j)$ is simply equal to 1. Furthermore, the non-spherical components of the ISA densities should provide us with guidance as to which atom types might require anisotropic treatment.

With the functional form for $f(r_{ij})$ determined, we now describe our model for $g(\theta_i, \phi_i, \theta_j, \phi_j)$. As motivated in Appendix 3.A, and under the ansatz of radial and angular separability, an approximate, transferable, and orientation-dependent expression for A_i^{exch} can be obtained by expanding A_i^{exch} in a basis of renormalized spherical harmonics,

$$C_{lm}(\theta, \phi) = \sqrt{\frac{4\pi}{2l+1}} Y_{lm}(\theta, \phi). \tag{3.8}$$

thus yielding

$$\begin{aligned} A_i^{exch}(\theta_i, \phi_i) &= A_{i,iso}^{exch}\big(1 + \xi^{exch}(\theta_i, \phi_i)\big), \\ \xi^{exch}(\theta_i, \phi_i) &\equiv \sum_{l>0,k} \alpha_{lk}^{exch} C_{lk}(\theta_i, \phi_i) \end{aligned} \tag{3.9}$$

for A_i^{exch} and subsequently

$$V_{ij}^{exch} = A_{ij}^{exch}(\Omega_{ij}) \left(\frac{(B_{ij} r_{ij})^2}{3} + B_{ij} r_{ij} + 1 \right) \exp(-B_{ij} r_{ij}) \tag{3.10}$$

with

$$A_{ij}^{exch}(\Omega_{ij}) = A_i^{exch}(\theta_i, \phi_i) A_j^{exch}(\theta_j, \phi_j) \tag{3.11}$$

for the exchange-repulsion potential. Note that, with the exception of the now orientation-dependent A_i^{exch}, the atomically-anisotropic model in Eq. (3.10) is identical to our previously-defined isotropic model (Eq. (3.6)).

In terms of parameterization for our newly-developed anisotropic model, note

that the a_{lk}^{exch} are free parameters which must be fit to ab initio data. Still, we and others have found the expansion in Eq. (3.9) to be very quickly convergent,[54,99,151,153,208,209,228,246,247] especially given a proper choice of coordinate system that eliminates many expansion terms via symmetry. In practice, only symmetry-allowed terms up to $l = 2$ seem to be required for heteroatoms, carbons in multiple bonding environments, and select hydrogens (see equations in Section 3.5), while many other atom types require no anisotropic parameters whatsoever. Encouragingly, isotropic atom types are easily modeled within this formalism simply by setting $\xi(\theta_i, \phi_i) = 0$.

Other Short-Range Effects

As in Chapter 2,[95] we have found that other short-range effects, namely charge penetration and short-range induction, can be modeled as proportional to exchange-repulsion. We take the same approach in the present Chapter, and the functional form for these two short-range effects is given by Eq. (3.10), with 'exch' superscripts replaced by the appropriate short-range energy term (see Section 3.4). Additionally, for induction, the long-range polarization must be damped, and for now this damping is modeled isotropically as in the AMOEBA force field.[227] Finally, to model short-range dispersion, we take the same Tang-Toennies[156,157] damping approach as in Chapter 2.[95] Because the argument to the damping function is given by

$$x = -\frac{d}{dr}\left[\ln V^{exch}(r)\right]\ r,$$

and because anisotropy only enters into the functional form as a multiplicative pre-factor, our functional form for damping remains unchanged compared to our previously-derived isotropic model.[95]

3.3.2 Anisotropic Models for Long-Range Interactions

Electrostatics

Theories for including anisotropy in long-range electrostatics are well established, and we refer the reader elsewhere for complete details on the required formalisms

for distributed multipole approaches.[54,78] In the present Chapter,

$$V_{ij}^{\text{multipole}} = \sum_{tu} Q_t^i T_{tu} Q_u^j$$

with multipolar interaction tensor T and parameterized moments Q for all multipole moments tu up to (in the present Chapter) rank 2.

On the grounds of increased accuracy and ease of parameterization, here we have chosen to use a multipolar approach to describe the anisotropy of long-range electrostatics, However, for increased computational efficiency, off-site point charge models[64] could also be utilized.

Induction

Just as with electrostatics, long-range induction should properly be described by a distributed multipole expansion of interacting atomic polarizabilities.[54,99] Indeed, it has been shown that inclusion of higher-order and/or anisotropic polarizabilities greatly reduces errors in the two-body induction potential relative to commonly-used isotropic dipole polarizability models.[75,227,248–250] Because the model for the two-body induction also determines the many-body polarization energy, the proper treatment of induced multipoles becomes especially important in condensed phase simulation.[75,78,227]

Owing to the increased computational cost of these higher-order and anisotropic polarizability models, and because such functional forms are (as of now, and to our knowledge) not fully implemented in common molecular simulation packages, we neglect both higher-order and anisotropic contributions to the long-range induction in the present Chapter. As we shall show, however, errors in the induction potential limit the overall accuracy of our force fields for extremely polar molecules (notably water), and further improvements will likely require us to generate improved models for long-range induction.

Dispersion

Past research[78] has motivated an anisotropic atom-atom model for dispersion of the form

$$V_{ij}^{disp} = -\sum_{n=6} \frac{C_{ij,n}(\Omega_{ij})}{r_{ij}^{n}} \tag{3.12}$$

Note that, in this equation, both odd and even powers of n are allowed in the dispersion expansion. In order to make this model both computationally efficient and maximally compatible with our previous isotropic model for dispersion, we choose (as an ansatz) to model the dispersion anisotropy as an orientation-dependent prefactor that effects all isotropic $C_6 - C_{12}$ dispersion coefficients equally:

$$V_{ij}^{disp} = -A_i^{disp} A_j^{disp} \sum_{n=3}^{6} \frac{C_{ij,2n}}{r_{ij}^{2n}} \tag{3.13}$$

with

$$A_i^{disp} = 1 + \xi^{disp}(\theta_i, \phi_i) \tag{3.14}$$

and $\xi^{disp}(\theta_i, \phi_i)$ as in Eq. (3.9). Once again, Eq. (3.13) reduces to the isotropic case by setting $\xi^{disp}(\theta_i, \phi_i) = 0$. We must note that, though the functional form in Eq. (3.13) bears many similarities to Eq. (3.12), (unphysically) no odd powers of r show up in our proposed model for dispersion. Furthermore, the model utilizes the same anisotropic expansion for each dispersion coefficient. Nonetheless, we will show in Section 3.5 that this model yields significant accuracy gains in the dispersion energy with only minimal additional parameterization and model expense.

3.4 Technical Details

3.4.1 The 91 Dimer Test Set

Our benchmarking procedures are the same as in Chapter 2,[95] and we briefly summarize the relevant technical details. A full discussion of results and example calculations are presented in Section 3.5.

We have previously developed a large library of benchmark energies for interactions between the following 13 atomic and organic species: acetone, argon, ammonia, carbon dioxide, chloromethane, dimethyl ether, ethane, ethanol, ethene, methane, methanol, methyl amine, and water. Using these 13 monomers, we have generated a library of dimer interaction energies for each of the 91 possible unique dimer combinations (13 homomonomeric, 78 heteromonomeric). For each of these dimer combinations, interaction energies were computed at a DFT-SAPT[169–177] level of theory for 1000 quasi-randomly chosen dimer configurations, representing 91,000 benchmark interaction energies in total. As described below, parameters for a given force field methodology are then fit on a component-by-component basis to reproduce the benchmark DFT-SAPT energies.

3.4.2 Parameter Determination

We will present three types of force field fitting methodologies in this Chapter, termed Iso-Iso FF, Aniso-Iso FF, and Aniso-Aniso FF (alternately referred to as MASTIFF, as discussed below). The nomenclature of each name refers to, first, the isotropic/anisotropic treatment of multipolar electrostatics and, second, the isotropic/anisotropic treatment of dispersion and short-range effects. All studied force fields use the following general functional form:

$$V_{FF} = \sum_{ij} V_{ij}^{exch} + V_{ij}^{elst} + V_{ij}^{ind} + V_{ij}^{\delta^{HF}} + V_{ij}^{disp} \tag{3.15}$$

where

$$
\begin{aligned}
V_{ij}^{exch} &= A_{ij}^{exch} P(B_{ij}, r_{ij}) \exp(-B_{ij} r_{ij}) \\
V_{ij}^{elst} &= -A_{ij}^{elst} P(B_{ij}, r_{ij}) \exp(-B_{ij} r_{ij}) + \sum_{tu} Q_t^i T_{tu} Q_u^j \\
V_{ij}^{ind} &= -A_{ij}^{ind} P(B_{ij}, r_{ij}) \exp(-B_{ij} r_{ij}) + V_{pol}^{(2)} \\
V_{ij}^{\delta^{HF}} &= -A_{ij}^{\delta^{HF}} P(B_{ij}, r_{ij}) \exp(-B_{ij} r_{ij}) + V_{pol}^{(3-\infty)} \\
V_{ij}^{disp} &= -A_{ij}^{disp} \sum_{n=3}^{6} f_{2n}(x) \frac{C_{ij,2n}}{r_{ij}^{2n}} \\
P(B_{ij}, r_{ij}) &= \frac{1}{3}(B_{ij} r_{ij})^2 + B_{ij} r_{ij} + 1 \\
A_{ij} &= A_i A_j \\
B_{ij} &= \sqrt{B_i B_j} \\
C_{ij,2n} &= \sqrt{C_{i,2n} C_{j,2n}} \\
f_{2n}(x) &= 1 - e^{-x} \sum_{k=0}^{2n} \frac{(x)^k}{k!} \\
x &= B_{ij} r_{ij} - \frac{2B_{ij}^2 r_{ij} + 3B_{ij}}{B_{ij}^2 r_{ij}^2 + 3B_{ij} r_{ij} + 3} r_{ij}
\end{aligned}
\tag{3.16}
$$

For both Iso-Iso FF and Aniso-Iso FF, A_i is a fit parameter, and $A_{ij}^{disp} = 1$. For Iso-Iso FF (our completely isotropic model), the multipole expansion $\sum_{tu} Q_t^i T_{tu} Q_u^j$ is truncated to point charges, whereas Aniso-Iso FF and MASTIFF both use a multipole expansion up to quadrupoles. Finally, for our anisotropic model, MASTIFF, each A_i is treated as an orientation-dependent function, and is represented by the spherical harmonic expansion

$$
\begin{aligned}
A_i(\theta_i, \phi_i) &= A_{i,iso}\big(1 + \xi(\theta_i, \phi_i)\big), \\
\xi(\theta_i, \phi_i) &\equiv \sum_{l>0,k} a_{i,lk} C_{lk}(\theta_i, \phi_i)
\end{aligned}
\tag{3.17}
$$

where $A_{i,iso}$ and $a_{i,lk}$ are fitted parameters with the exception that $A_{i,iso}^{disp} = 1$.

Because DFT-SAPT provides a physically-meaningful energy decomposition into electrostatic, exchange-repulsion, induction, and dispersion terms, parameters for each term in Eq. (3.15) are directly fit to model the corresponding DFT-SAPT energy (see Ref. 95 and references therein for details on the DFT-SAPT terminology):

$$\begin{aligned} V^{exch} &\approx E^{exch} \equiv E_{exch}^{(1)} \\ V^{elst} &\approx E^{elst} \equiv E_{pol}^{(1)} \\ V^{ind} &\approx E^{ind} \equiv E_{ind}^{(2)} + E_{ind\text{-}exch}^{(2)} \\ V^{\delta^{HF}} &\approx E^{\delta^{HF}} \equiv \delta(HF) \\ V^{disp} &\approx E^{disp} \equiv E_{disp}^{(2)} + E_{disp\text{-}exch}^{(2)}. \end{aligned} \tag{3.18}$$

Fitting parameters on a component-by-component basis helps ensure parameter transferability and minimizes reliance on error cancellation. Note that no parameters are fit to reproduce the total energy and that, because the DFT-SAPT energy decomposition is only calculated to second-order, third- and higher-order terms (mostly consisting of higher-order induction) are estimated by $E^{\delta^{HF}}$.

Parameters Calculated from Monomer Properties

Of the parameters listed in Eq. (3.16), most do not need to be fit to the DFT-SAPT energies, but can instead be calculated directly on the basis of monomer electron densities. In particular, all multipolar coefficients, Q, polarizabilities (involved in the calculation of V_{pol}), dispersion coefficients C, and atom-in-molecule exponents, B^{ISA}, are calculated in a manner nearly identical to Ref. 95. Note that, for our atom-in-molecule exponents, we tested the effects of treating B^{ISA} both as a hard- and as a soft-constraint in the final force field fit. While the conclusions from this study are rather insensitive to this choice of constraint methodology, we have found that the overall force field quality is somewhat improved by relaxing the B^{ISA} coefficients in the presence of a harmonic penalty function (technical details of which can be found in the Supporting Information of Ref. 95). The optimized B coefficients

in this Chapter are always within 5–10% of the calculated B^{ISA} coefficients from Chapter 2, demonstrating the good accuracy of the B^{ISA} calculations themselves.

As a second distinction from our prior work, and for reasons of compatibility with the OpenMM[188] software we use for all molecular dynamics simulations, here our molecular simulations use an induced dipole model to describe polarization effects. Numerical differences between this model and the drude model used previously are very minor. Additionally, the Thole-damping functions used in this Chapter follow the same functional form used in the AMOEBA model,[225] with a damping parameter of 0.39.

Parameters Fit to Dimer Properties

In addition to the soft-constrained B parameters, all other free parameters (A and a parameters from Eq. (3.15) and Eq. (3.17)) are fit to reproduce DFT-SAPT energies from the 91 dimer test set described above. For each dimer pair, 4-5 separate optimizations (for exchange, electrostatics, induction, δHF, and, for MASTIFF, dispersion) were carried out to minimize a weighted least-squares error, with the weighting function given by a Fermi-Dirac functional form,

$$w_i = \frac{1}{\exp(-E_i/kT) + 1}, \tag{3.19}$$

where E_i is the reference energy and the parameter kT, which sets the energy scale for the weighting function, is calculated from an estimate of the global minimum well depth, E_{min}, such that $kT = 5.0|E_{min}|$.

Local Axis Determination

Identically to AMOEBA and other force fields that incorporate some degree of atomic-level anisotropy,[151,153,225] we use a z-then-x convention to describe the relative orientation of atomic species. By design, the z-axis is chosen to lie parallel to the principal symmetry axis (or approximate local symmetry axis) of an atom in its molecular environment, and the xz-plane is similarly chosen to correspond to a

secondary symmetry axis or plane. Based on the assigned symmetry of the local reference frame, many terms in the spherical expansion of Eq. (3.9) can then be set to zero, minimizing the number of free parameters that need to be fit to a given atom type. Representative local reference frames are shown for a few atom types in Fig. 3.1, and a complete listing of anisotropic atom types (along with their respective local reference frames and non-zero spherical harmonic expansion terms) are given in the Section 3.B.

CCSD(T) Force Fields

DFT-SAPT is known to systematically underestimate the interaction energies of hydrogen-bonding compounds, and can also exhibit small but important errors for dispersion-dominated compounds.[251] Consequently, for simulations involving CO_2, $CHCl_3$, NH_3, and H_2O, we refit our SAPT-based force fields to reproduce benchmark supermolecular, counterpoise-corrected CCSD(T)-F12a/aVTZ calculations on the respective dimers. All calculations were performed using the Molpro 2012 software.[252] Fits were still performed on a component-by-component basis, with the energy of most components matching the DFT-SAPT calculations used in Chapter 2.[95] However, so that the total benchmark energy corresponded to the total interaction energy calculated by CCSD(T)-F12a/aVTZ, the difference between coupled-cluster and SAPT energies was added to the SAPT dispersion energy,

$$\begin{aligned} V^{\text{exch}} &\approx E^{\text{exch}} \equiv E^{(1)}_{\text{exch}} \\ V^{\text{elst}} &\approx E^{\text{elst}} \equiv E^{(1)}_{\text{pol}} \\ V^{\text{ind}} &\approx E^{\text{ind}} \equiv E^{(2)}_{\text{ind}} + E^{(2)}_{\text{ind-exch}} \\ V^{\delta^{\text{HF}}} &\approx E^{\delta^{\text{HF}}} \equiv \delta(\text{HF}) \\ V^{\text{disp}} &\approx E^{\text{disp}} \equiv E^{(2)}_{\text{disp}} + E^{(2)}_{\text{disp-exch}} + \delta(\text{CC}), \end{aligned} \tag{3.20}$$

where $\delta(\text{CC}) \equiv E_{\text{int}}^{\text{CCSD(T)-F12a}} - E_{\text{int}}^{\text{DFT-SAPT}}$.

In fitting these CCSD(T)-f12a-based force fields, and to account for small errors in the original SAPT dispersion energy, we somewhat relaxed the constraint that

$A^{disp} = 1$ for all atom types, and instead let $0.7 \leqslant A^{disp} \leqslant 1.3$. This constraint relaxation led, in some cases, to modest improvements in the fitted potential.

CO_2 3-body potential

For the CO_2 dimer, we developed a three-body model to account for three-body dispersion effects. This three-body model is based on the three-body dispersion Axilrod-Teller-Muto (ATM) type model developed by Oakley and Wheatley[253]. These authors fit the ATM term with the constraint that the total molecular C_9 coefficient be 1970 a.u. Based on our own calculations using a CCSD/AVTZ level of theory,[254] we have obtained an isotropic molecular C_9 coefficient of 2246 a.u.; consequently, a 1.13 universal scale factor was introduced to the Oakley potential so as to obtain dispersion energies in line with this new dispersion coefficient.

3.4.3 Simulation Protocols

ΔH_{sub} for CO_2

For CO_2, the molar enthalpy of sublimation was determined according to

$$\begin{aligned}\Delta H_{sub} &= H_g - H_{crys} \\ &= (U_g + PV_g) - (U_{el,crystal,0K} + \Delta U_{el,crystal,0K\rightarrow T_{sub}} + PV_{crys} + E_{vib,crystal}) \\ &\approx (RT) - \left(U_{el,crystal,0K} + \int_{0K}^{T_{sub}} C_p\, dT \quad + E_{vib,crystal}\right)\end{aligned} \tag{3.21}$$

which assumes ideal gas behavior and $PV_g >> PV_{crys}$. For the crystal, an experimental measure of C_p was obtained from Ref. 255 and numerically integrated to obtain a value $\Delta U_{el,crystal,0K\rightarrow T_{sub}} = 6.70 \text{kJ mol}^{-1}$. Theoretical measures of $E_{vib,crystal} \approx 2.24 - 2.6 \text{kJ mol}^{-1}$ were obtained from (respectively) Ref. 256 and Ref. 257, and $U_{el,crystal,0K}$ was determined from the intermolecular force field using a unit cell geometry taken from experiment.[258]

Other CO_2 Simulations

To determine the densities and enthalpies of vaporization used in this Chapter, simulations were run in OpenMM using NPT and NVT ensembles, respectively. After an equilibration period of at least 100ps, data was collected for a minimum of 500ps, and uncertainties were calculated using the block averaging method. Average densities were obtained directly from simulation, and the molar enthalpy of vaporization for CO_2 was determined from the following formula:

$$\begin{aligned} \Delta H_{vap} &= H_g - H_{liq} \\ &= U_g - U_{liq} + P(V_g - V_{liq}) \end{aligned} \tag{3.22}$$

Note that, at the state points studied, the ideal gas approximation is insufficiently accurate, and thus simulations were run for both the gas and liquid phases at experimentally-determined densities and pressures.[7]

2nd Virial Calculations

Classical second virial coefficients were calculated for NH_3, H_2O, CO_2, and $CHCl_3$ using rigid monomer geometries and following the procedure described in Ref. 83.

3.5 Results and Discussion

3.5.1 Overview

We now turn to a discussion of the methods whereby we can compare our newly developed anisotropic force field methodology to various sum-of-spheres models. As is standard in ab initio force field development, we use a straightforward metric to evaluate force field quality: the accuracy with which a given force field functional form can reproduce high-quality ab initio benchmark energies. Furthermore, and because the functional forms introduced in Section 3.3 directly affect only the pairwise-additive portion of the intermolecular potential, we concentrate our efforts on assessing force field with respect to benchmark calculations of dimer

interaction energies, which directly measure the two-body portion of a system's total intermolecular interaction energy. (When required, and as discussed in Section 3.5.5, many-body effects can be accounted for separately and systematically using known methods).[4,259] In addition to this primary metric for force field quality, we also evaluate our force fields for their ability to reproduce select experimental properties. Importantly, however, experimental predictions from an ab initio force field significantly depend, not only on the fit quality of the pair potential, but also on the choice of benchmark electronic structure theory, treatment of many-body and/or quantum effects, etc. Because these factors complicate comparisons to experiment, here we treat experimental accuracy as an important, but secondary, metric for evaluating force field accuracy.

So as to systematically evaluate the effects of anisotropy on the development of intermolecular potentials, we compare three types of models. The first model, which we call Iso-Iso FF, uses a completely isotropic description of all energy components. Our second model, Aniso-Iso FF, accounts for long-range electrostatic anisotropy by including multipolar contributions (up to quadrupoles), but uses an isotropic model for all other terms in the intermolecular force field. Note that this model is virtually identical to the Slater-ISA FF model developed in Chapter 2, and that this manner of partially treating anisotropy is very similar in spirit to the popular AMOEBA[225,227] methodology. Finally, we develop Aniso-Aniso FF, which selectively incorporates anisotropy into all energy components (aside from long-range polarization) of the intermolecular potential. This model, which we also refer to with the moniker MASTIFF (a **M**ultipolar, **A**nisotropic, **S**later-**T**ype **I**ntermolecular **F**orce **F**ield), treats all electrostatic interactions via a multipole expansion with up to quadrupolar contributions, and includes anisotropic parameters for other terms of the force field (short-range interactions plus dispersion) for heteroatoms, atoms in multiple bonding environments, and associated hydrogens. A complete list of anisotropic atom types is given in the Section 3.B.

3.5.2 Accuracy: Comparison with DFT-SAPT

For each of the 91 dimer combinations described in Section 3.4, parameters were fit to reproduce Symmetry-Adapted Perturbation Theory (SAPT) energies calculated for 1000 different relative orientations of the constituent monomers. From these 'dimer-specific' fits, and as in described in Chapter 2,[95] we then averaged the root-mean-squared (RMSE) and mean signed errors (||MSE||) from each of the 91 fits to produce so-called 'characteristic RMSE/||MSE||', metrics representative of the errors associated with a given force field methodology. Typically, because the absolute magnitudes of the various energy components become large in the repulsive portion of the potential, these characteristic errors are dominated by repulsive configurations. As such, we have also calculated 'attractive RMSE/||MSE||' (aRMSE/a||MSE||), defined as the characteristic errors for the subset of configurations with net attractive total interaction energies. All computed characteristic RMSE are shown in Fig. 3.2. Unless otherwise stated, results in this section refer exclusively to the 'Dimer-specific' fits in Fig. 3.2, with an explanation and full discussion of so-called 'Transferable' fits given in Section 3.5.3.

Based on the characteristic RMSE shown in Fig. 3.2, both Aniso-Iso FF and MASTIFF offer substantial improvements over the completely isotropic model Iso-Iso FF. Though unsurprising, given the well-studied importance of higher-order electrostatic multipole moments, Aniso-Iso FF shows reduced RMSE/aRMSE that are (depending on the exact error metric used) roughly 30% smaller than Iso-Iso FF. Both RMSE and aRMSE measures showing similar gains in accuracy, indicating that inclusion of higher-order multipoles (henceforth 'multipolar electrostatic anisotropy') is important in both attractive and repulsive regions of the potential. Crucially, inclusion of additional 'short-range anisotropies' (anisotropic interactions arising from overlap of monomer electron densities, namely exchange-repulsion and electrostatic/inductive charge penetration) and long-range 'dispersion anisotropy' yields a *further* 40% reduction in RMSE/aRMSE for MASTIFF as compared to the Aniso-Iso FF. This latter result is highly important, as it suggests that, for the generation of highly accurate ab initio potentials, the combination of short-range and

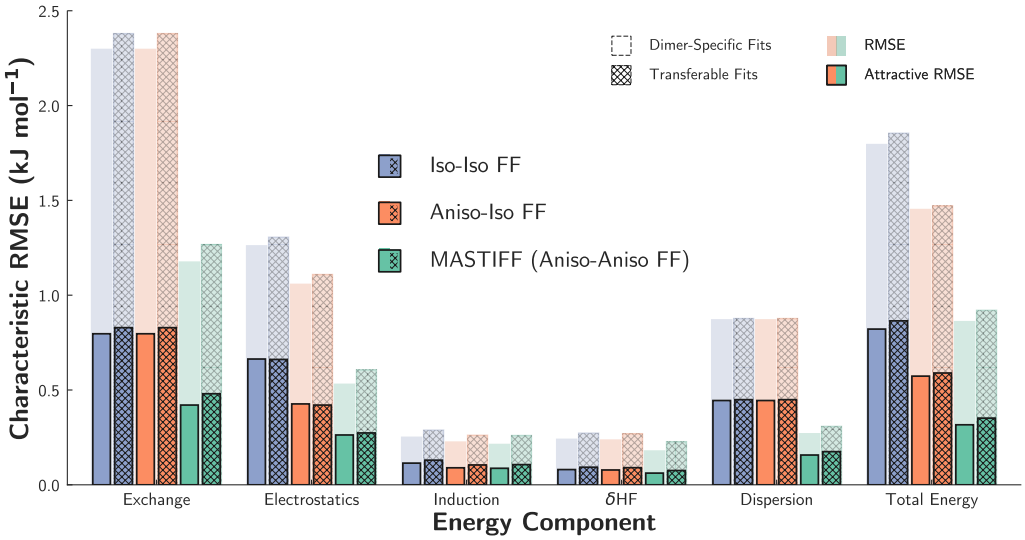


Figure 3.2: Characteristic RMSE (as described in the main text) for the Iso-Iso FF (purple), Aniso-Iso FF (orange), and MASTIFF (green) over the 91 dimer test set. The semi-transparent bars represent total RMSE for each energy component, while the smaller solid bars represent 'Attractive' RMSE, in which repulsive points have been excluded. For each force field, two types of fits, dimer-specific (solid) and transferable (hashed lines), are displayed; see Section 3.5.3 for details. Finally, note that, for Iso-Iso FF and Aniso-Iso FF, only the electrostatic and total energy RMSE's differ.

dispersion anisotropies are just as important to include as multipolar electrostatic anisotropy. Indeed, this substantial increase in force field accuracy, which arises from a full treatment of anisotropic effects, and is independent of improvements from multipolar electrostatic anisotropy, is one of the most important findings in the present Chapter. In summary, and encouragingly, the combination of multipolar electrostatic, short-range, and dispersion anisotropies result in an overall 60% reduction in RMSE/aRMSE when comparing Iso-Iso FF to MASTIFF.

To see exactly how an inclusion of anisotropy impacts each component of the potential, Fig. 3.2 also displays characteristic RMSE/aRMSE for each term in the force field description as compared to DFT-SAPT. Immediately, one can see that (aside from induction, discussed below), an inclusion of atomic-level anisotropy

greatly improves the description of each energy component. Unless otherwise stated, here we report results for aRMSE and dimer-specific fits, though similar values are obtained for overall RMSE and for transferable fits. Compared to Iso-Iso FF, exchange errors in MASTIFF are reduced by 47%. Electrostatic errors are reduced by an even larger 60%. By evaluating the ratio of electrostatic errors between different models, we find that $^{\text{aRMSE Aniso-Iso FF}}/_{\text{aRMSE Iso-Iso FF}} = 0.64$ and $^{\text{aRMSE MASTIFF}}/_{\text{aRMSE Aniso-Iso FF}} = 0.62$, suggesting that *both* higher-order multipoles and anisotropic charge penetration terms are necessarily to obtain an accurate description of the DFT-SAPT electrostatic energy. Finally, via an inclusion of dispersion anisotropy, aRMSE for dispersion are reduced by a significant 65%.

Though the trends for exchange, electrostatics, and dispersion universally suggest the importance of including atomic-level anisotropy, trends for terms describing the physics of polarization and charge-transfer (represented in DFT-SAPT by induction and δHF) are less encouraging. On the one hand, including higher-order multipoles substantially lowers RMSE for induction, with $^{\text{RMSE Aniso-Iso FF}}/_{\text{RMSE Iso-Iso FF}} = 0.70$. Because both Iso-Iso FF and Aniso-Iso FF use isotropic polarizabilities, and because the induction energy fundamentally depends only on the polarizabilities and the static electric field, this result is clearly due to an improved treatment of the static electric field via anisotropy of the multipolar electrostatics. Once again, this suggests that an anisotropic treatment of long-range electrostatics is crucial for accurate force field development. On the other hand, our functional form for anisotropic short-range induction (Eq. (3.15) and Eq. (3.17)) leads to no improvement in the induction RMSE, with $^{\text{RMSE Aniso-Iso FF}}/_{\text{RMSE Iso-Iso FF}} = 0.97$. This observed lack of improvement is likely due to a combination of factors. First, and perhaps most importantly, we have chosen in this Chapter to use isotropically-averaged dipole polarizabilities, but as with electrostatics, anisotropy and higher-order terms have been shown to be important in in the multipole expansion of atomic dipole polarizabilities.[54,99,236,248,260] Second, and though probably a smaller source of error, it is also unclear how to optimally model the distance dependence of the induction energy at short intermolecular separations, where penetration and charge-transfer effects become important and the long-range polarization terms must be damped.[95,154,261,262]

Given that the more elaborate short-range form of the MASTIFF induction model does not result in a tangible improvement, it is quite possible that alternative formulations are required for an accurate treatment of highly anisotropic induction.

To further analyze the effects of anisotropy on a molecule-by-molecule basis, we have calculated 'improvement ratios', defined as $^{\text{aRMSE Iso-Iso FF}}/_{\text{aRMSE MASTIFF}}$, for each energy component and for each homomonomeric species in the test set, results for which are shown in Table 3.1. (Improvement ratios for heteromnomeric species are given in the Supporting Information of Ref. 263, and we additionally provide scatter plots of each homomonomeric force field fit in Section 3.C.)

The most striking observation from the data presented in Table 3.1 is that the improvement ratios vary considerably with molecule. For example, with water the aRMSE is improved by an order of magnitude when anisotropy is included. On the other hand, no improvement is seen for hydrocarbons such as ethane and methane (also see the Section 3.C). Consequently, anisotropy in the short-range expansions may be necessary for only some atoms types (see Section 3.6). For the molecules studied in our test set, and in line with chemical intuition, we have found anisotropy to be particularly important for heteroatoms, π-bonded atoms, and all hydrogens bonded to anisotropic heavy atoms. Appealingly, this distinction between anisotropic and isotropic atom types simplifies force field parameterization and can enable more efficient molecular simulation (via a more cost-effective treatment of multipolar electrostatics) without sacrificing force field accuracy. Note that the current empirically-determined definitions of anisotropic atom types match both chemical intuition and the more quantitative measures of atomic anisotropy proposed by other groups.[204,216]

In general, the ordering of improvement ratios for exchange, electrostatics, dispersion, and the total energies are reasonably correlated. (As stated above, our model for anisotropic induction interactions is rather poor, and hence the improvement ratios for induction and δHF are relatively uncorrelated with the other components. A good model for anisotropic induction and δHF might easily change this result). Physically speaking, all atomically-anisotropic interactions arise from the same source (atomically-anisotropic electron densities), and so the

	Exchange	Electrostatics	Induction	δHF	Dispersion	Total Energy
H_2O (O,H)	4.96	13.12	1.69	1.88	8.20	11.54
CO_2 (C,O)	3.83	9.13	0.99	0.64	4.91	8.62
NH_3 (N,H)	3.15	5.36	0.90	2.86	2.45	5.78
Ethene (C,H)	1.44	1.46	1.00	1.00	7.59	4.16
Chloromethane (Cl)	3.17	4.03	1.36	1.04	4.20	4.08
Methyl Amine (N,H)	1.70	2.93	1.05	2.22	2.95	2.37
Methanol (O,H)	1.81	3.05	1.11	2.03	1.00	2.36
Dimethyl Ether (O)	1.30	2.07	1.38	1.19	1.85	2.30
Ethanol (O,H)	1.29	3.10	1.04	1.45	1.79	2.14
Acetone (O)	1.58	1.98	1.03	1.34	1.51	1.08
Ethane ()	1.00	1.26	1.05	1.01	1.00	1.08
Ar ()	1.00	1.00	1.00	1.00	1.00	1.00
Methane ()	1.00	0.93	0.99	1.01	1.00	0.94

Table 3.1: 'Improvement Ratios' for each homomonomeric species in the 91 dimer test set. For each dimer and energy component, the improvement ratio is calculated as the ratio of aRMSE between Iso-Iso FF and MASTIFF; values greater than 1 indicate decreased errors in the anisotropic model. Entries have been ordered according to the improvement ratio for the total energy.

observed correlation might have been expected. Nevertheless, there are some exceptions to this trend, such as with ethene and acetone. For ethene, relatively modest improvement ratios (roughly 1.4) are seen for exchange and electrostatics, whereas dispersion shows a much greater improvement ratio of 7.6. Since ethene homomonomeric interactions are dispersion-dominated, the improvement ratio for the total energy then roughly corresponds to that of dispersion. A larger test set (particularly one which includes more non-polar aromatic species) would be necessary to assess the generality of this result. For acetone, there is good correlation between the improvement ratios for exchange, electrostatics, and dispersion, which might lead one to suspect that the total energy improvement ratio would also be around 1.5-2.0. Nevertheless, for this molecule, the isotropic model benefits from error cancellation between energy components, and the total energy aRMSE between isotropic and anisotropic models are rather similar.

Crucially, electrostatics is most definitely not the only intermolecular interaction for which atomic-level anisotropy improves model quality. Indeed, for molecules like ethene, multipolar anisotropy in the electrostatic model is relatively unimportant, whereas dispersion anisotropy is essential for accurately modeling the π

interactions. Thus, for a given system, multipolar electrostatic, dispersion, and/or short-range anisotropies may all be important, and all relevant anisotropies must be accounted for in order to obtain good intermolecular models.

3.5.3 Transferability: Comparison to DFT-SAPT

From the above results it is clear that, when explicitly parameterized, an inclusion of anisotropy can greatly enhance the accuracy of an intermolecular potential. Nevertheless, for standard force field development, force field parameters must be *transferable* in order to be useful in the accurate prediction of intermolecular interactions in new chemical and/or physical environments. Indeed, in comparing simpler models to ones that introduce additional complexity, there is an ever-present danger that any accuracy gains from the more complex functional form are simply due to over-fitting or error cancellation,[71] ultimately resulting in an overly-complex model with poor predictive ability and limited transferability.

We have previously shown how, with models similar to Iso-Iso FF[75,83] or Aniso-Iso FF,[95] it is possible to generate transferable potentials with applicability to a broad range of chemical and physical environments.[75] This transferability has been attributed to a combination of the physically-meaningful energy decomposition of DFT-SAPT, our choice to parameterize on a component-by-component basis (rather than to the total energy), our use of physically-motivated functional forms, and our recourse to parameters calculated on the basis of monomer properties.[75,83,95]

MASTIFF largely shares this philosophy of force field development, and so we might also expect it to be transferable to heteromonomeric dimers. However, this transferability cannot be taken for granted because of the specific way in which we have included the anisotropy. First, we have relied on several separability ansatzes (Eq. (3.4) and Eq. (3.7)), and second, in doing so we have implicitly neglected potentially important interaction functions that depend on the relative orientation between monomers. Both of these assumptions may affect the transferability of the resulting force field.

To assess the transferability of the MASTIFF model, we analyze the extent

to which parameters developed for the homomonomeric systems can be used, without modification, to describe the interactions of the mixed dimers. Such an out-of-sample prediction, which is easily accomplished with out test set, is a direct measure of the extent to which our pair potentials can be applied to new chemical environments. For these transferable fits, parameters were fit to the 13 homomonomeric systems, and the combination rules shown in Eq. (3.15) were used to generate force fields for the remaining heteromonomeric systems. Thus, with these transferable fits we have essentially generated 78,000 predictions from fits to 13,000 data points. RMSE and aRMSE for these fits are shown in Fig. 3.2, and we treat relative differences between these quantities for the 'dimer-specific' and 'transferable' fits as a measure of the extent of transferability for each force field methodology.

Remarkably, all three force fields — Iso-Iso, Aniso-Iso, and MASTIFF — perform similarly for the dimer-specific and transferable fits, both for the individual interaction energy components and for the total interaction energy. The degree of transferability of the MASTIFF model is very encouraging, and indicates that the manner in which we have chosen to include the anisotropy is meaningful and does not lead to overfitting, but rather increases the accuracy of the intermolecular potentials for both in-sample and out-of-sample systems.

3.5.4 Comparison to Experiment: Second Virial Coefficients

In addition to comparisons with DFT-SAPT, we have also benchmarked our force fields against experimental second virial coefficients, which offer a direct experimental measure of the pair potential without the complication of many-body effects. Still, such comparisons to experiment depend, not only on the quality of the force field, but also on the accuracy of the benchmark electronic structure theory used to fit the force field. As compared to gold-standard CCSD(T)/CBS calculations, small (< 1 kJ mol^{-1}) but systematic inaccuracies can be present in DFT-SAPT/aVTZ+m[95] calculations, and so in this section we refit our potentials to a CCSD(T)-F12a/aVTZ+m benchmark, which serves as a computationally af-

fordable yet accurate prediction of the CCSD(T)/CBS limit.[17,264] We refer to these coupled cluster-based models with a -CC suffix, e.g. MASTIFF-CC, and details of the refitting procedure (which minimally affect the dispersion energies) can be found earlier in Section 3.4. Thus, aside from quantum effects (which are negligible for CO_2[265] and well-benchmarked for H_2O[266]), our second virial predictions should offer a fairly clean comparison between different models and experiment.

Using the -CC potentials, we have calculated second virial coefficients for each Iso-Iso FF-CC, Aniso-Iso FF-CC, and MASTIFF-CC and for the following systems: H_2O (Fig. 3.3), NH_3 (Fig. 3.4), $CHCl_3$ (Fig. 3.5), and CO_2 (Fig. 3.6). First, we find that the MASTIFF-CC methodology predicts virial coefficients which closely corresponds to experimental data. Given the range of systems tested (CO_2 dimer interactions are dispersion dominated, while $CHCl_3$, NH_3, and H_2O have relatively larger electrostatic and polarization contributions), this suggests that, when benchmarked against high-quality electronic structure theory, our anisotropic methodology offers a general strategy for quantitatively accurate pairwise potential development. Second, we note that the Iso-Iso FF-CC predictions are much worse than their MASTIFF-CC or Aniso-Iso FF counterparts, suggesting that an accurate treatment of long-range electrostatics is essential to obtain accurate virial coefficients. Finally, though Aniso-Iso FF-CC gives equally good predictions for some systems (notably $CHCl_3$) compared to the MASTIFF-CC method, virial coefficients for other systems (especially H_2O) are less accurate, suggesting that dispersion and short-range anisotropies are also important in many systems for the accurate prediction of virial coefficients. Consequently, and in summary, the minimal additional computational overhead (compared to Aniso-Iso FF-CC) and excellent accuracy of MASTIFF-CC permits us to recommend this fully anisotropic MASTIFF methodology for the prediction of dimer interaction energies and second virial coefficients.

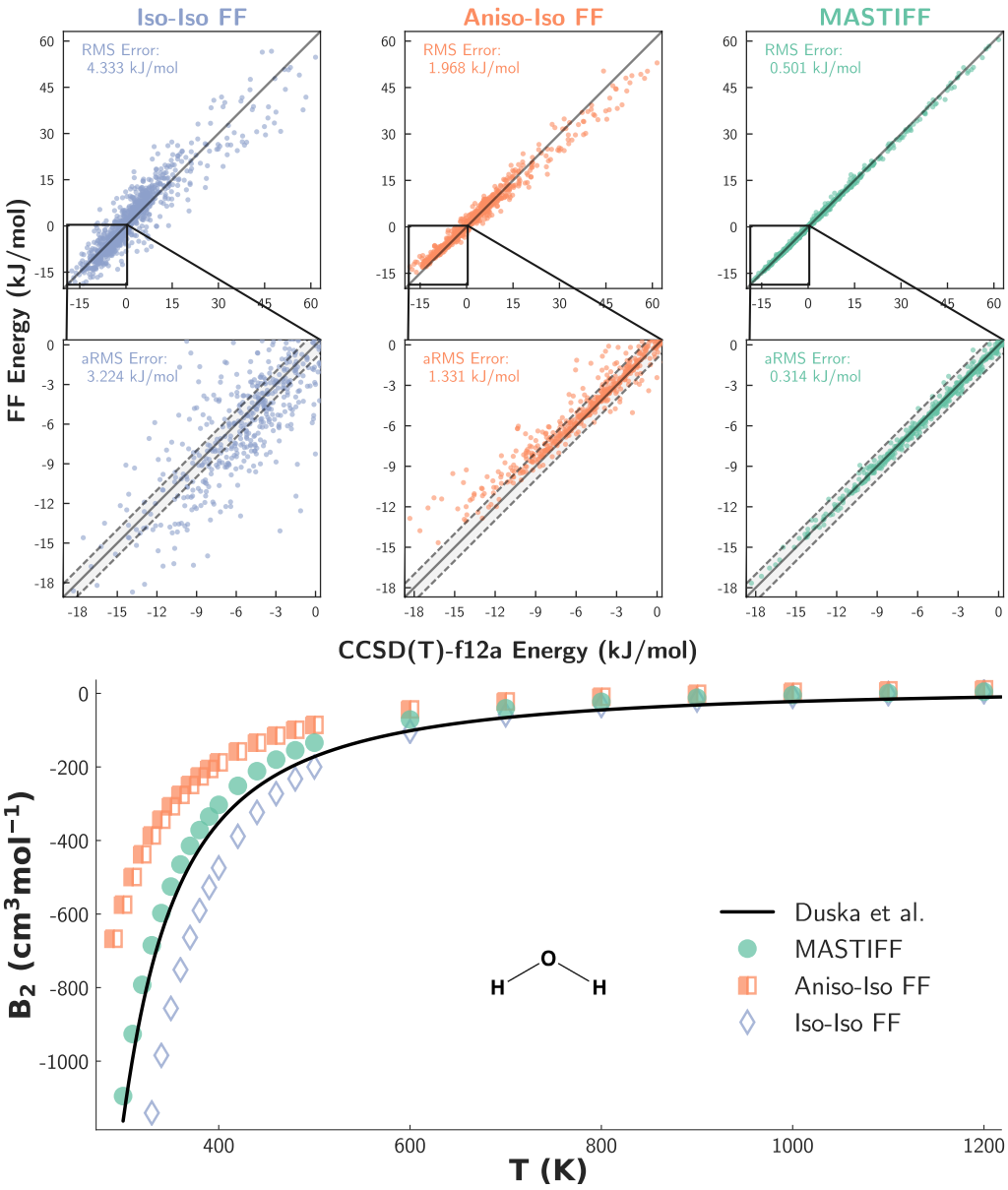


Figure 3.3: Classical second virial for water. Experimental data from Ref. 13. Note that some data points from Iso-Iso FF extend below the plot area.

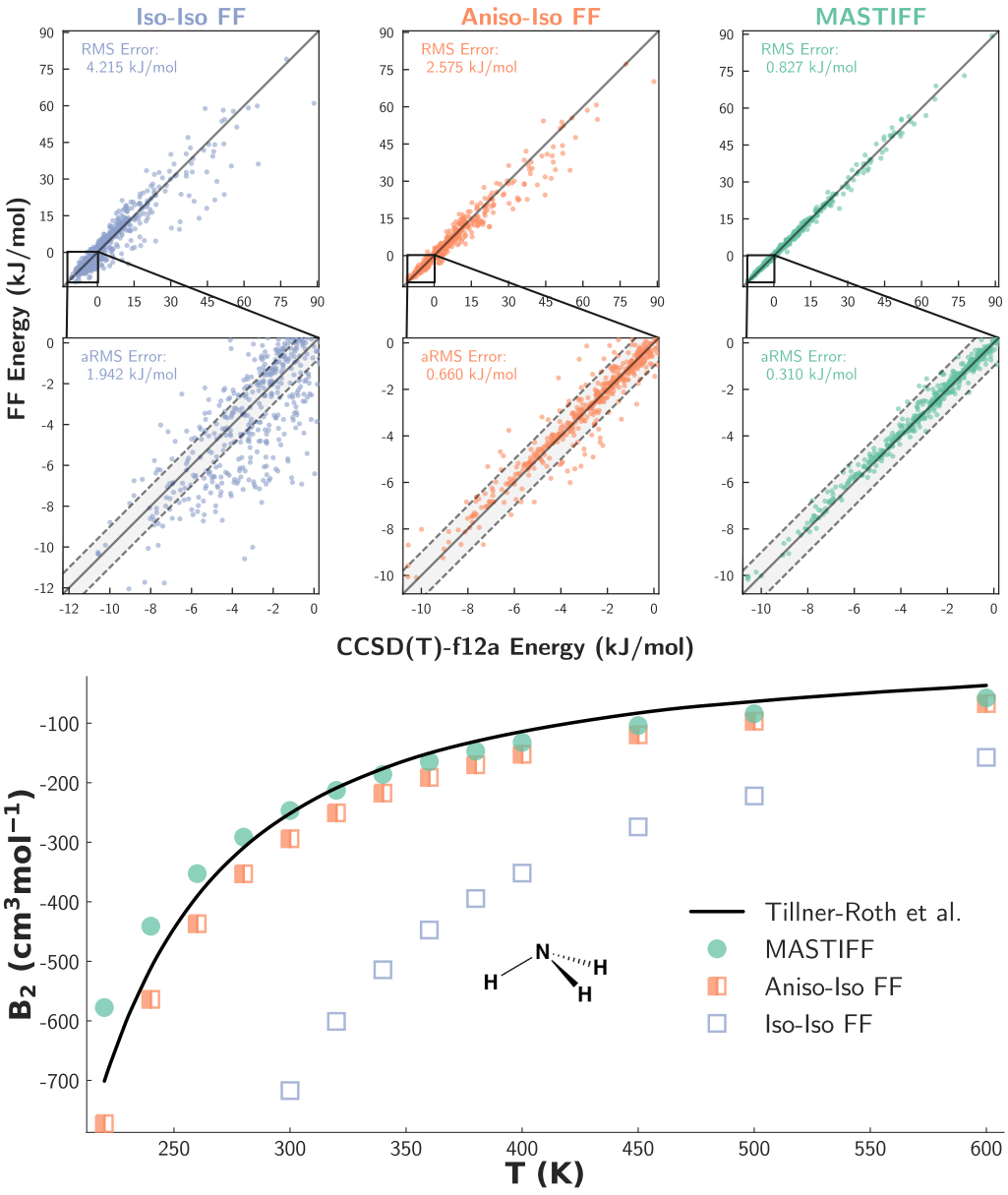


Figure 3.4: Classical second virial for ammonia. Experimental data from Ref. 14. Note that some data points from Iso-Iso FF extend below the plot area.

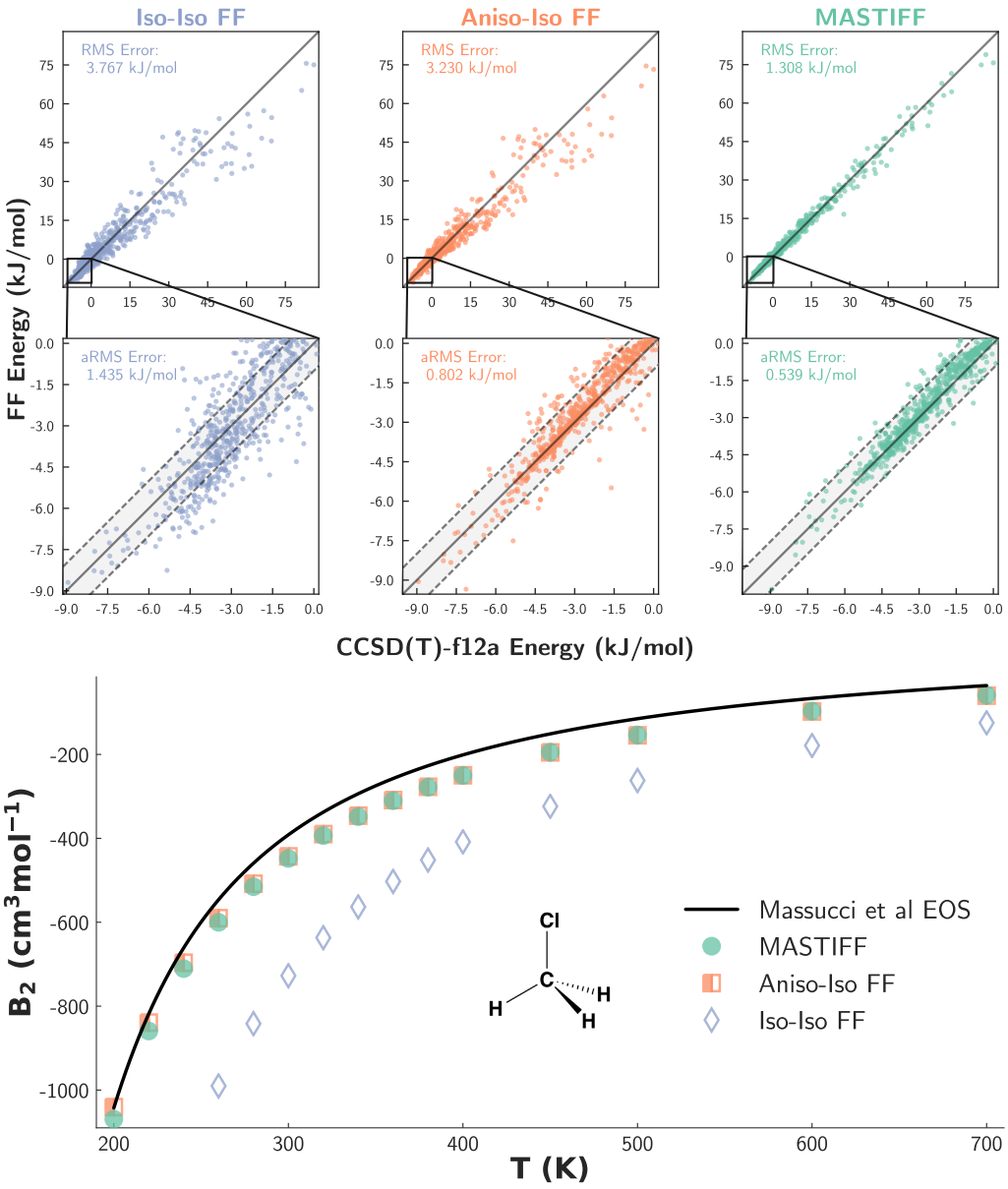


Figure 3.5: Classical second virial for chloromethane. Experimental equation of state (EOS) from Ref. 15. Note that some data points from Iso-Iso FF extend below the plot area.

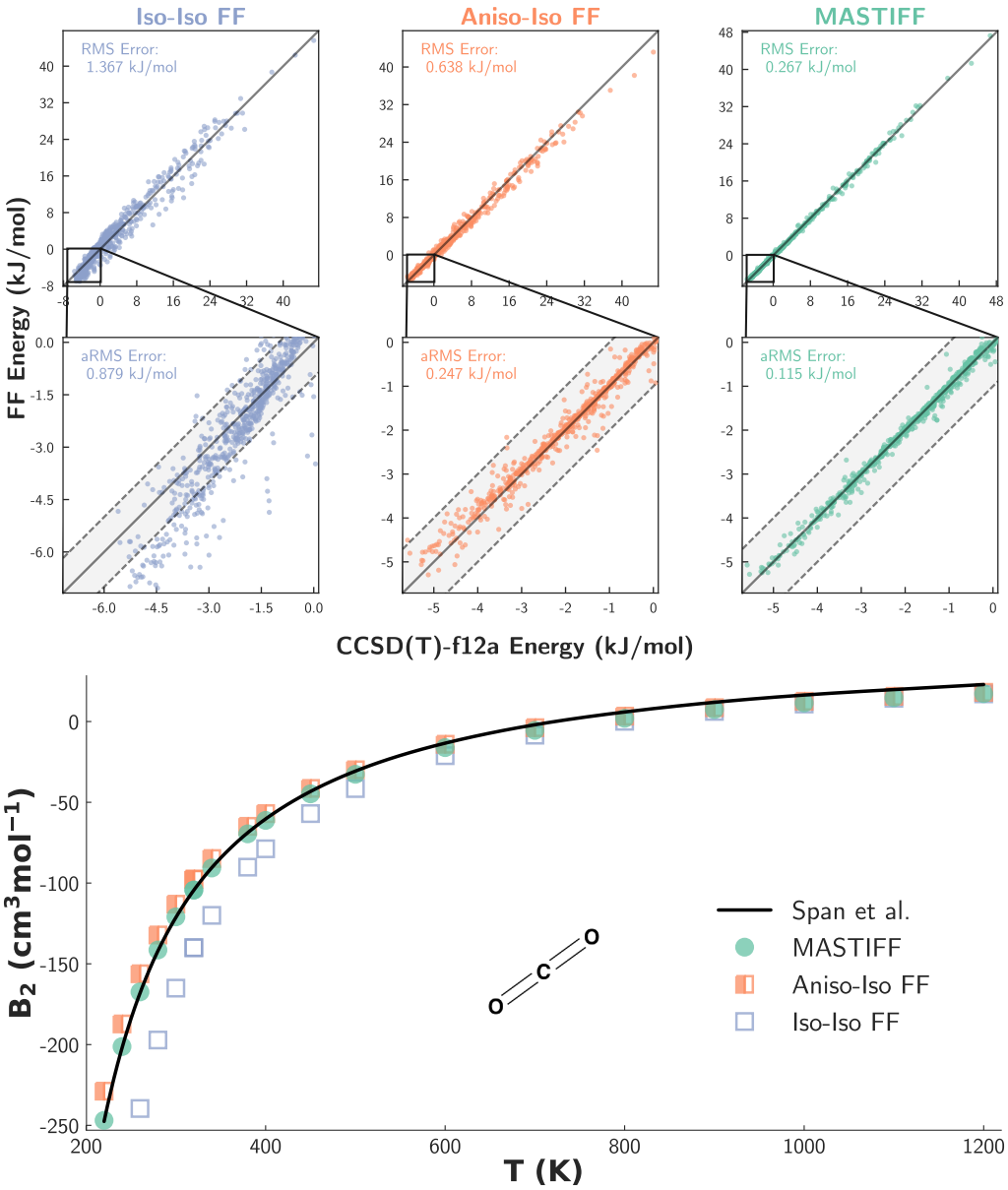


Figure 3.6: Classical second virial for CO_2. Experimental data from Ref. 7

3.5.5 Comparison to Experiment: Condensed Phase Properties of CO_2

To demonstrate the applicability of the MASTIFF methodology in condensed phase simulation, we have developed a complete many-body potential for CO_2, and have run bulk simulations involving a variety of vapor, liquid, supercritical, and solid phase points for preliminary comparisons to experiment. As above, we use the MASTIFF-CC potential to describe the pairwise potential and the many-body induction. As for other many-body effects, it is well-known[16,253,259] that three-body dispersion, and to a lesser extent, three-body exchange, are also important.[267] Thus, we model three-body dispersion via a modified version of the three-body dispersion potential developed by Oakley and Wheatley (see Section 3.4). Three-body exchange effects are not accounted for in our model, however prior work shows they are very small under the conditions studied here.[259]

Density predictions for the vapor, liquid, and supercritical phases of CO_2 are shown in Table 3.2, and enthalpies of sublimation and vaporization are shown in Table 3.3. We find it notable and highly encouraging that MASTIFF-CC reproduces *all* studied experimental properties to within a few percent. Of particular note is our excellent reproduction of the sublimation enthalpy, which critically depends on the lattice energy of the solid phase. Unlike with liquid or supercritical CO_2, where many dimer configurations are sampled, the solid consists of only four symmetry-unique configurations. Consequently, whereas an isotropic potential (which is in error for particular dimer configurations, but can take advantage of error cancellation to be accurate in an average sense) might yield good property predictions for the liquid phase, it would not be expected to correctly predict the solid phase (where beneficial error cancellation is unlikely). Indeed, most theories (including our previously developed SYM-3B model,[259] nearly all popular empirically-developed CO_2 models,[268] AMOEBA,[257] and many electronic structure theories[257]) struggle to correctly predict the solid phase properties of CO_2! For this reason, the enthalpy of sublimation is an extremely stringent test of force field quality,[268] and our accurate reproduction of this quantity is evidence for both the

excellent quality of the MASTIFF-CC potential in specific and of the importance of atomic-level anisotropy in general. Though more testing is needed to confirm the accuracy of our anisotropic force field for other phase points, our results suggest that, crucially, the MASTIFF-CC potential is transferable across the entire phase diagram of molecular CO_2, and is capable of describing the gas, liquid, supercritical, and solid phases.

Despite the excellent success of the MASTIFF-CC model, it is also worthwhile to address and understand its minor shortcomings. In particular, we have studied representative two- and three-body energies taken from a snapshot of the liquid at 273.15 K and 100 bar. For the two-body energies, we have compared against the extremely accurate Kalugina et al.[17] potential, while for three-body energies we have benchmarked against the Hellmann[16] PES. From these results (Section 3.D), it is clear that our pairwise MASTIFF potential is highly accurate for all configurations present in the liquid, with very small RMSE and no systematic error in the potential, such that the total two-body energy is accurate to within 0.05% compared to the Kalugina et al. PES. Once again, this result argues strongly for the accuracy and transferability of the MASTIFF methodology, and suggests that an inclusion of anisotropy is essential, not only for gas-phase clusters, but also for simulations of the bulk. By contrast, our three-body potential is systematically in error compared to the Hellmann PES. Though some of this error may be due to inaccuracies in the benchmark potential itself as compared to coupled-cluster,[16] most of this error is likely due to inaccuracies in our model for many-body CO_2 interactions. The atomically-isotropic treatment of three-body dispersion, neglect of higher-order dispersion terms , and neglect of explicit three-body exchange may all contribute to this error, and an improved model for many-body CO_2 interactions will be the subject of future research. Indeed, it is well-known that the density can be extremely sensitive to the treatment of many-body effects,[267] and it is highly probable that an improved many-body model would reduce the already small errors observed in our MASTIFF-CC predictions. Regardless, for now we conclude that, despite some small residual errors arising from the simplified treatment of many-body effects, our MASTIFF-CC methodology yields for an extremely accurate force field for CO_2

Phase	T (K)	P (bar)	Density (g/ml)	Exp.	% Error
Gas	300	50	0.131	0.128	2.34
Supercritical	320	140	0.728	0.703	3.56
Liquid	300	100	0.825	0.802	2.87
Liquid	273.15	100	1.000	0.974	2.67

Table 3.2: Select densities for CO_2 across a range of experimental conditions. Experimental data taken from the EOS of Ref. 7. Entries ordered by increasing experimental density.

Phases	T (K)	ΔH (kJ mol^{-1})	Exp.	% Error
s $\rightarrow$ g	194.76	25.0 ± 0.15	25.2	-0.8
l $\rightarrow$ g	288	7.92	7.80	-1.4

Table 3.3: Enthalpies of vaporization/sublimation for CO_2 at several temperatures. Experimental data taken from the EOS of Ref. 7. The uncertainty in the enthalpy of sublimation is due to ambiguity in the theoretical zero-point energy for CO_2 (see Section 3.4.

with applicability across a range of experimentally-important phases.

3.6 Conclusions and Recommendations

We have developed a comprehensive methodology for modeling atomic-level anisotropy in standard intermolecular force fields. By treating this anisotropy through a simple extension of standard isotropic force fields,[95] we have successfully demonstrated how this computationally-efficient treatment of atomic-level anisotropy leads to significant improvements in models for intermolecular interactions. Critically, and in contrast to popular assumption, we have shown how the accurate treatment of multipolar electrostatics does not *by itself* account for all energetically-important effects of atomic-level anisotropy. Rather, our results indicate that anisotropy may need to be included in the each electrostatic, exchange and

dispersion terms in order to obtain intermolecular force fields of the highest quality. In the present study, and in agreement with the more quantitative metrics proposed by others,[204,216] we have found a comprehensive model of atomic-level anisotropy to be particularly important for obtaining sub- kJ mol^{-1} accuracy for describing molecules with heteroatoms (particularly ones with exposed lone pairs), carbons in multiple bonding environments, and hydrogens bound to anisotropic heavy atoms. Our new intermolecular 'MASTIFF' force fields show great promise, not only with respect to high-quality electronic structure benchmark energies, but also with respect to experimental property predictions. Importantly, MASTIFF maintains high efficiency and transferability. and can easily be implemented in common software packages such as OpenMM for use in condensed phase simulations.[188]

Despite the advances presented in this Chapter, several aspects of our force field methodology require further improvement, and will be the subject of ongoing research. In particular, an improved description of induction effects will become essential for accurate bulk simulations of highly polarizable molecules such as water. We are now actively working to develop improved models that can describe both long-range anisotropic polarization and short-range polarization damping, as these aspects of the force field critically affect both the two- and many-body induction energies and can account for a sizable fraction of the total interaction energy in condensed phases. We anticipate that these improved models for induction will, in combination with an accurate description of three-body dispersion and exchange, yield a general approach to force field development that captures both the two- and many-body features of intermolecular interactions, in turn enabling highly accurate, 'next-generation' force field development capable of simulating a wide array of phases and chemical environments.

3.A Motivation for $g(\theta_i, \phi_i, \theta_j, \phi_j)$

As shown elsewhere,[103,269] an exact (under the ansatz of radial and angular separability) model for $g(\theta_i, \phi_i, \theta_j, \phi_j)$ is given by Stone's $\bar{S}$-functions, which form a complete basis set for describing any scalar function which depends on the relative

orientation between molecules, and are given (following Stone's notation[78]) by the formula

$$\bar{S}^{k_1k_2}_{l_1l_2j} = i^{l_1-l_2-j} \begin{pmatrix} l_1 & l_2 & j \\ 0 & 0 & 0 \end{pmatrix}^{-1} \sum_{m_1m_2m} [D^{l_1}_{m_1k_1}(\Omega_1)]^* [D^{l_2}_{m_2k_2}(\Omega_2)]^* C_{lm}(\theta,\phi) \begin{pmatrix} l_1 & l_2 & j \\ m_1 & m_2 & m \end{pmatrix}. \tag{3.23}$$

The general form of these $\bar{S}$-functions can be quite complicated, and involve both the Wigner D rotation matrices and Wigner 3j-symbols (quantities in parentheses) as well as the degree (l_1, l_2, and j) and order (m_1, m_2, and m for the global coordinate system, k_1 and k_2 for the various local coordinate systems) of the spherical harmonic tensors. Here subscripts reference either molecule 1 or molecule 2, and subscriptless quantities refer to the dimer as a whole.

In order to obtain a functional form for the exchange-repulsion that is amenable to simple combination rules (a necessary prerequisite for transferable potentials), we must somehow be able to separate $g(\theta_i,\phi_i,\theta_j,\phi_j)$ into monomer contributions. Unfortunately, many of the $\bar{S}$-functions depend on the relative orientation of the dimer itself, and thus must be excluded in the development of *transferable* potentials. Thus as a second ansatz (empirically validated by us in Section 3.5 and by others[270]) we neglect all contributions from $\bar{S}$-functions that depend on both local coordinate systems. This leaves us with two sets of $\bar{S}$-functions, namely

$$\bar{S}^{k0}_{l0l} = C_{lk}(\theta_i,\phi_i) \tag{3.24}$$

and

$$\bar{S}^{0k}_{0ll} = C_{lk}(\theta_j,\phi_j) \tag{3.25}$$

which are simply the renormalized spherical harmonics (Eq. (3.8)) expressed in each of the two local coordinate systems.

Given our truncated expressions for the $\bar{S}$-functions, we now need only extend our functional form for $f(r_{ij})$ to incorporate these anisotropic contributions. We

choose, in a manner analogous to literature precedent,[54,99,151,153,208,209,228,246,247] to expand the A_i^{exch} and A_j^{exch} parameters of Eq. (3.7) in terms of a truncated expansion of $\bar{S}$-functions. (In principle, we could also account for anisotropy in the B_{ij} parameters of our model for $f(r_{ij})$. However, previous literature suggests that in practice this 'hardness' parameter can often be treated as constant, and we also neglect its possible anisotropy in this Chapter.) Consequently, all short-range anisotropies are modeled in this Chapter by the expressions given in Eq. (3.9) and Eq. (3.10).

3.B Local Axis Definitions

For each molecule in the 91 dimer test set, listed below are any atom types which have been treated anisotropically. For each anisotropic atom type, the approximate symmetry and all terms included in the spherical harmonic expansion are listed to the right of the atom type. Additionally, the local axis reference frame for each anisotropic atom type is defined in the Axes subsection using the z-then-x convention employed by AMOEBA and other potentials. The first column of the axes subsection denotes the index of the anisotropic atom (atom ordering as in Section A.1), and the second column denotes whether the z or x axis is being defined. For certain local symmetries, the choice of x-axis is unimportant, and so not every anisotropic atom type has a defined x-axis. The remaining columns define the direction vector for the axis in terms of atomic indices. The first index (often the anisotropic atom itself) lists the start of the vector, and the endpoint of the vector is defined as the midpoint of all subsequently listed atoms.

To use water as an example, the oxygen atom is treated anisotropically using a spherical harmonic expansion that includes y10, y20, and y22c terms (notation as in Ref. 78). The z-axis points from the oxygen to the midpoint between the two hydrogens, and the xz plane (and subsequently the x-axis) is defined by one of the O–H bonds.

3.B.1 Acetone

```
OC  c2v     y10 y20 y22c

Axes
ATOM#   AXIS (z or x)   Atomic Indices defining vector (either 2 or more
integers)
1 z 1 0
1 x 0 2
```

3.B.2 Ar

```
Ar

Axes
ATOM#   AXIS (z or x)   Atomic Indices defining vector (either 2 or more
integers)
```

3.B.3 Chloromethane

```
Chloromethane
Cl c3v     y10 y20

Axes
ATOM#   AXIS (z or x)   Atomic Indices defining vector (either 2 or more
integers)
1 z 1 0
```

3.B.4 Carbon Dioxide

```
CO2
OCO cinfv y10 y20
CCO2 dinfh y20
```

```
Axes
ATOM#   AXIS (z or x)   Atomic Indices defining vector (either 2 or more
integers)
0 z 0 1
1 z 1 0
2 z 2 0
```

3.B.5 Dimethyl Ether

```
Dimethyl Ether
O c2v y10 y20 y22c

Axes
ATOM#   AXIS (z or x)   Atomic Indices defining vector (either 2 or more
integers)
0 z 0 1 2
0 x 0 1
```

3.B.6 Ethane

```
Ethane

Axes
ATOM#   AXIS (z or x)   Atomic Indices defining vector (either 2 or more
integers)
```

3.B.7 Ethanol

```
Ethanol
OH c2v y10 y20 y22c
HO cinfv y10 y20
```

```
Axes
ATOM#   AXIS (z or x)   Atomic Indices defining vector (either 2 or more
integers)
2 z 2 1 3
2 x 2 1
3 z 3 2
```

3.B.8 Ethene

```
Ethene
CM c2v y22c
HM cinfv y10 y20

Axes
ATOM#   AXIS (z or x)   Atomic Indices defining vector (either 2 or more
integers)
0 z 0 1
0 x 0 2
1 z 1 0
1 x 1 4
2 z 2 0
3 z 3 0
4 z 4 1
5 z 5 1
```

3.B.9 Water

```
H2O
OH2 c2v y10 y20 y22c
H2O cinfv y10 y20
```

```
Axes
ATOM#   AXIS (z or x)   Atomic Indices defining vector (either 2 or more
integers)
0 z 0 1 2
0 x 0 1
1 z 1 0
2 z 2 0
```

3.B.10 Methane

```
Methane

Axes
ATOM#   AXIS (z or x)   Atomic Indices defining vector (either 2 or more
integers)
```

3.B.11 Methanol

```
Methanol
OH1 c2v y10 y20 y22c
HO1 cinfv y10 y20

Axes
ATOM#   AXIS (z or x)   Atomic Indices defining vector (either 2 or more
integers)
1 z 1 0 5
1 x 1 0
5 z 5 1
```

3.B.12 Methyl Amine

```
Methyl Amine
```

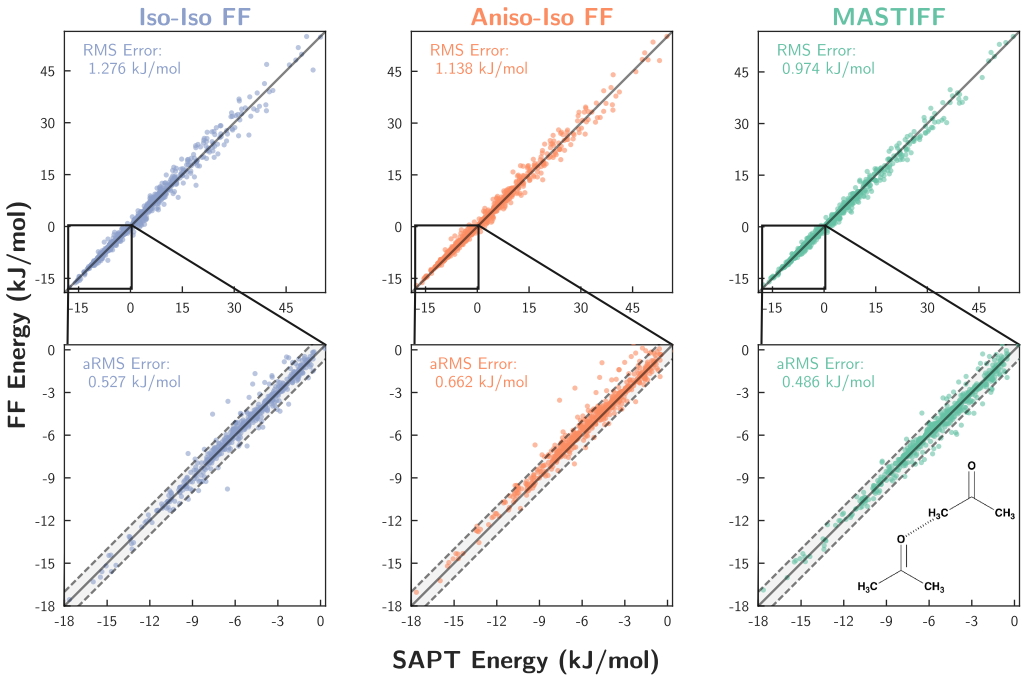
```
N1 c2v y10 y20 y22c
HN1 cinfv y10 y20

Axes
ATOM#   AXIS (z or x)   Atomic Indices defining vector (either 2 or more
integers)
1 z 1 0 5 6
1 x 1 0
5 z 5 1
6 z 6 1
```

3.B.13 Ammonia

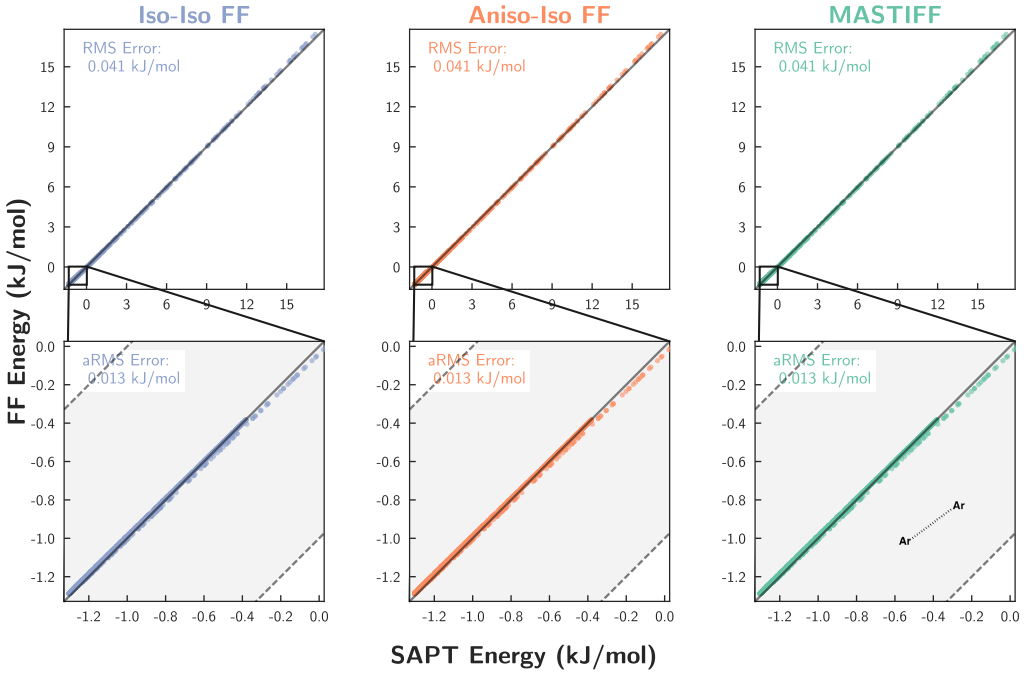
```
Ammonia
N c3v y10 y20
HN cinfv y10 y20

Axes
ATOM#   AXIS (z or x)   Atomic Indices defining vector (either 2 or more
integers)
0 z 0 1 2 3
0 x 0 1
1 z 1 0
2 z 2 0
3 z 3 0
```

3.C Homodimer Fits

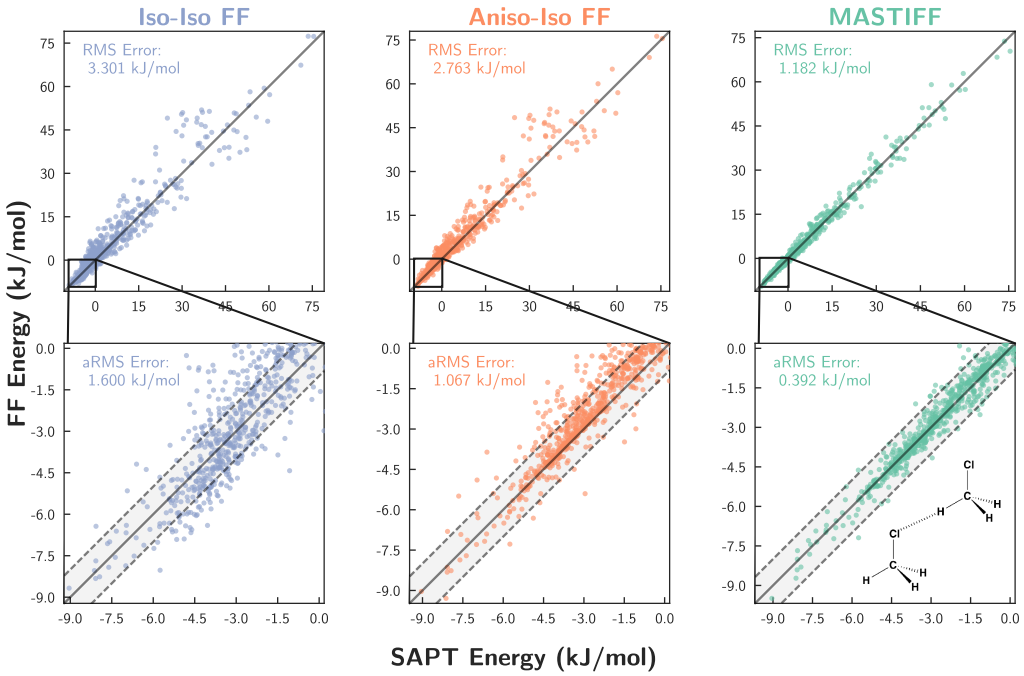
(a) Acetone Dimer



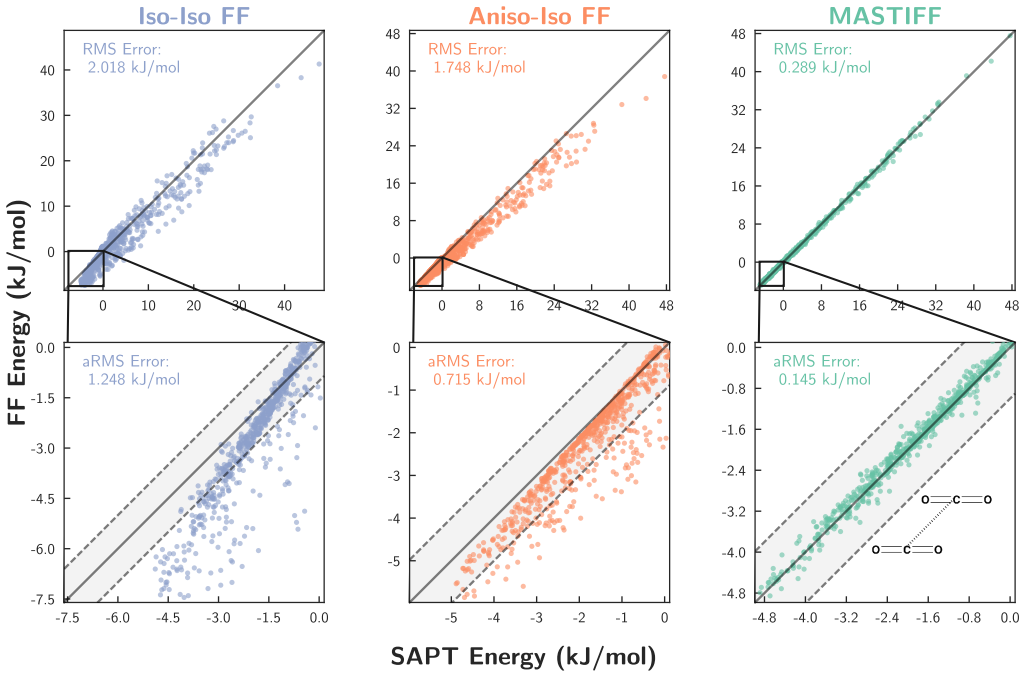
(b) Ar Dimer



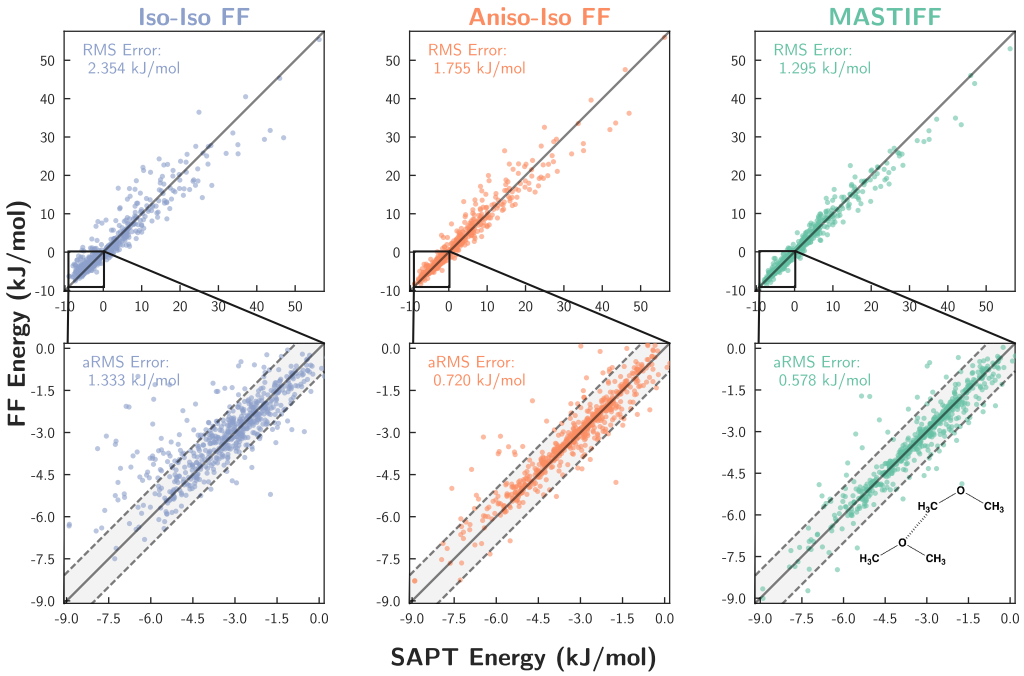
(c) Chloromethane Dimer



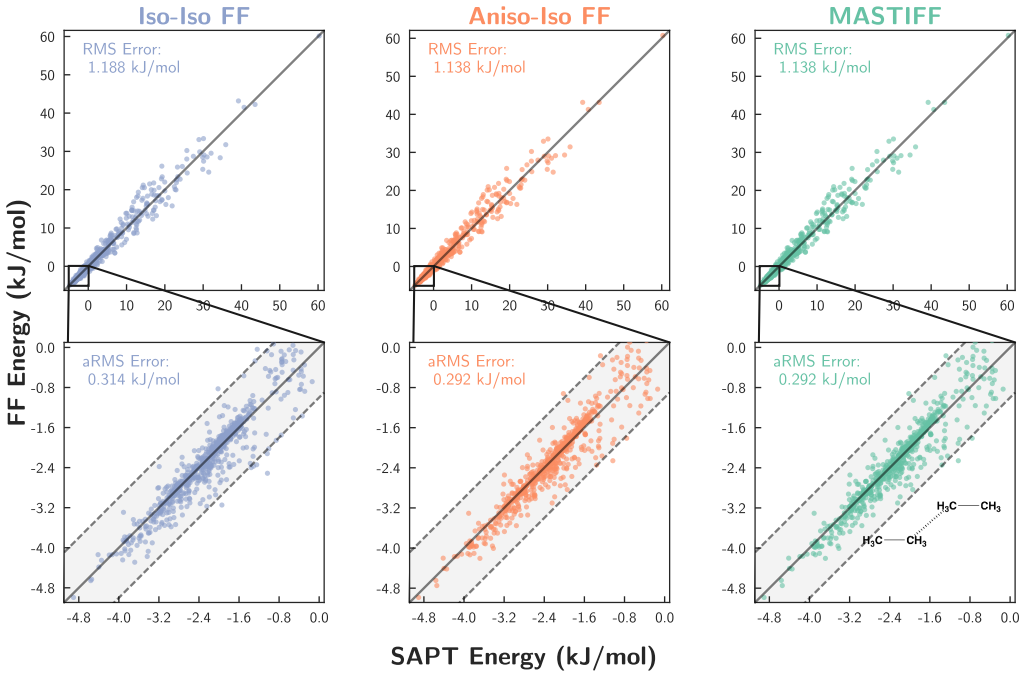
(d) CO_2 Dimer



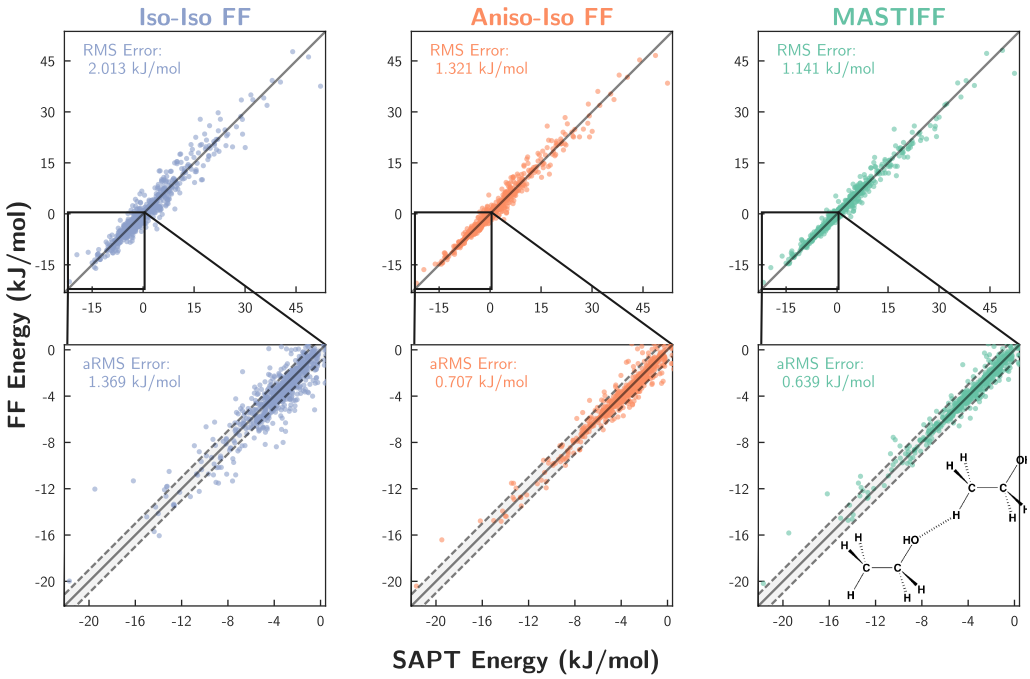
(e) Dimethyl Ether Dimer



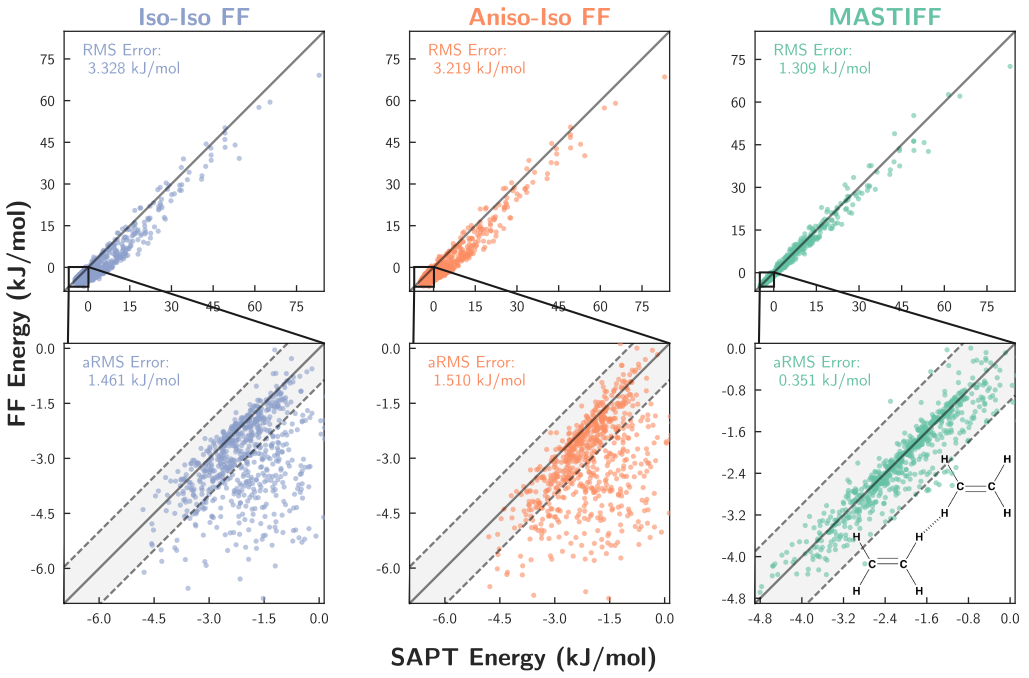
(f) Ethane Dimer



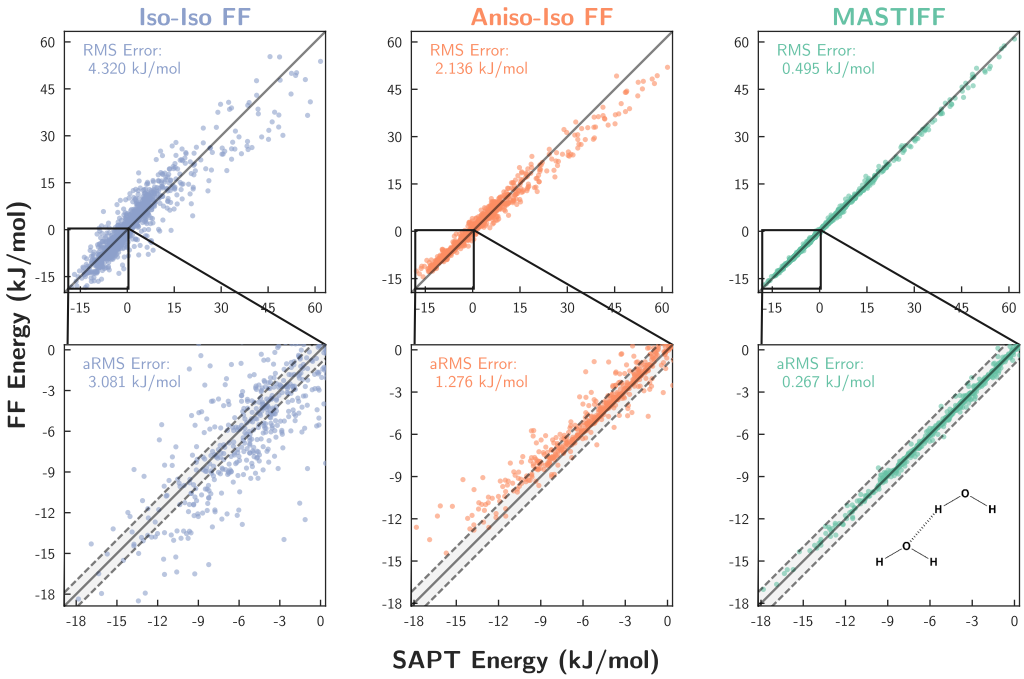
(g) Ethanol Dimer



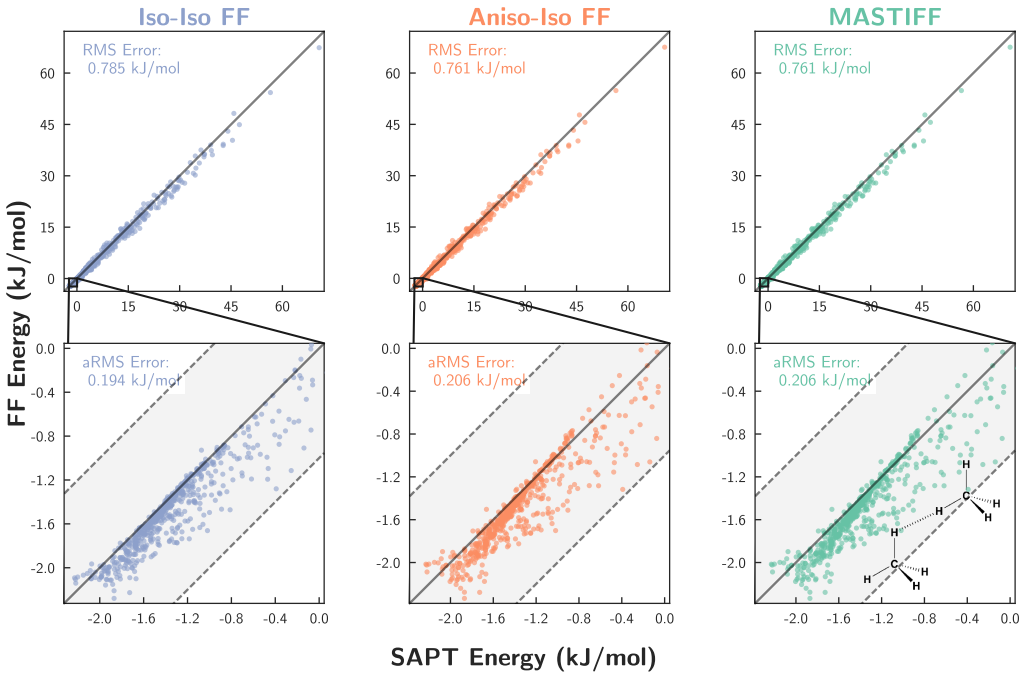
(h) Ethene Dimer



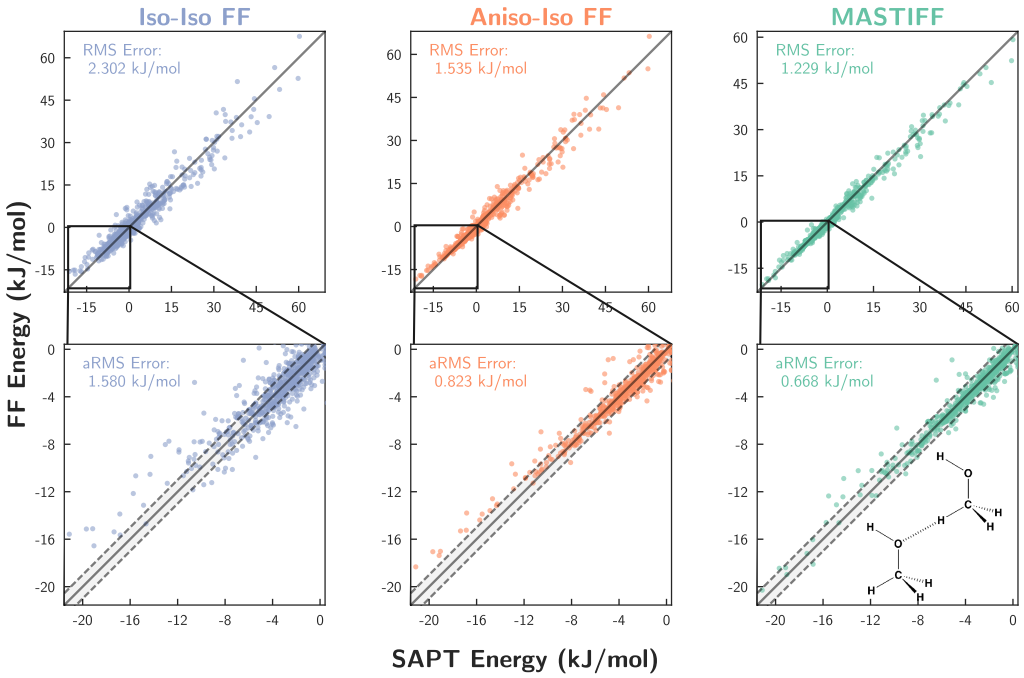
(i) H_2O Dimer



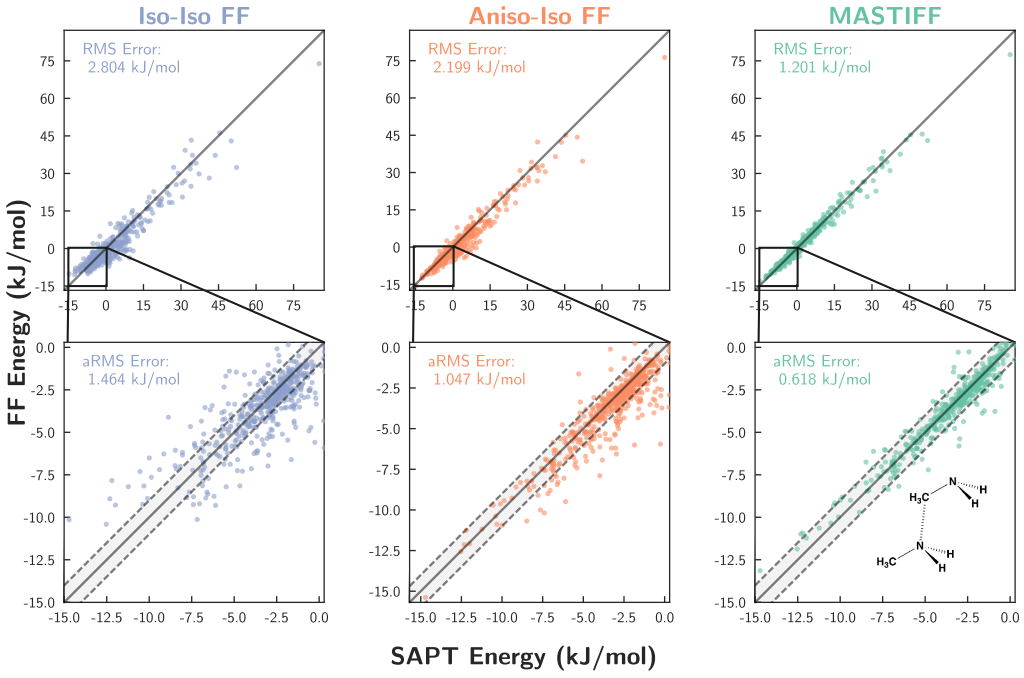
(j) Methane Dimer



(k) Methanol Dimer



(l) Methyl Amine Dimer



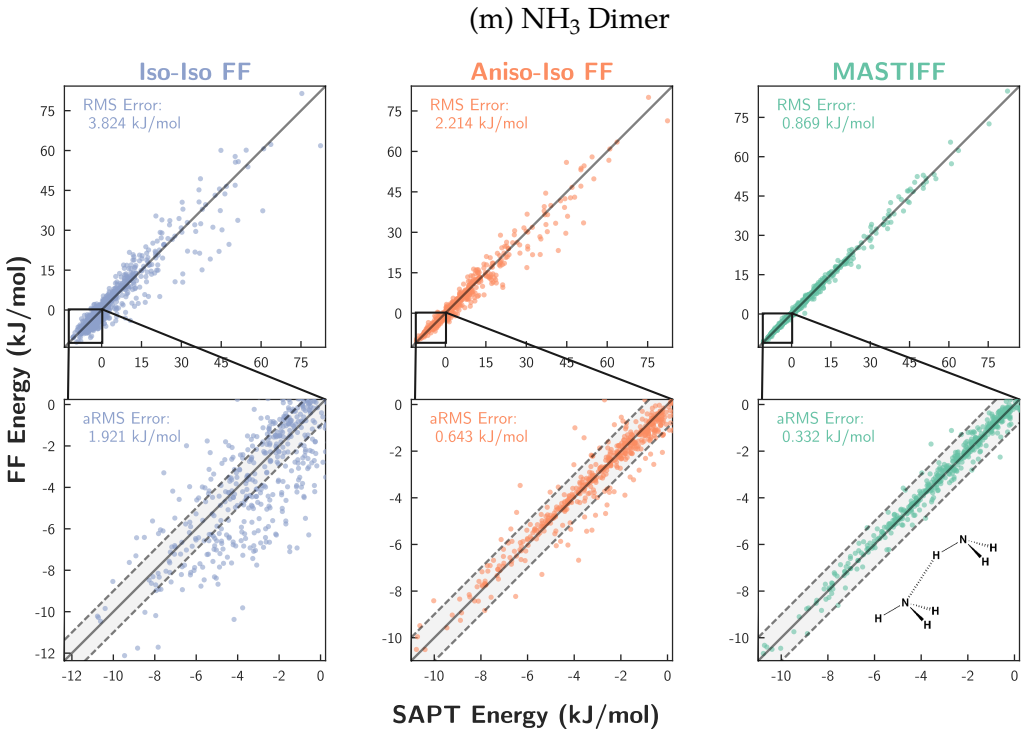


Figure 3.7: Force field fits for each homomoneric systems using the Iso-Iso FF (purple), Aniso-Iso FF (orange), and MASTIFF (purple). Two views of the fit to the total energy are displayed along with corresponding RMSE (aRMSE for the inset showing attractive configurations). The $y = x$ line (black) indicates perfect agreement between reference energies and each force field, while shaded grey areas represent points within ± 1 kJ/mol agreement of the benchmark.

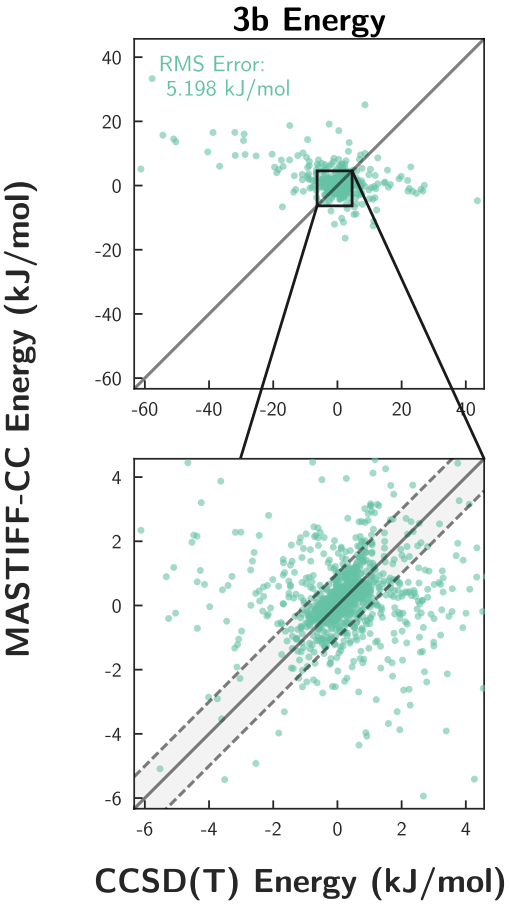


Figure 3.8: Three-body interaction energies for CO_2 as compared to the Hellmann[16] database of 9401 reference trimer configurations computed at a CCSD(T) level of theory.

3.D 2- and 3-body MASTIFF-CC CO_2 energies

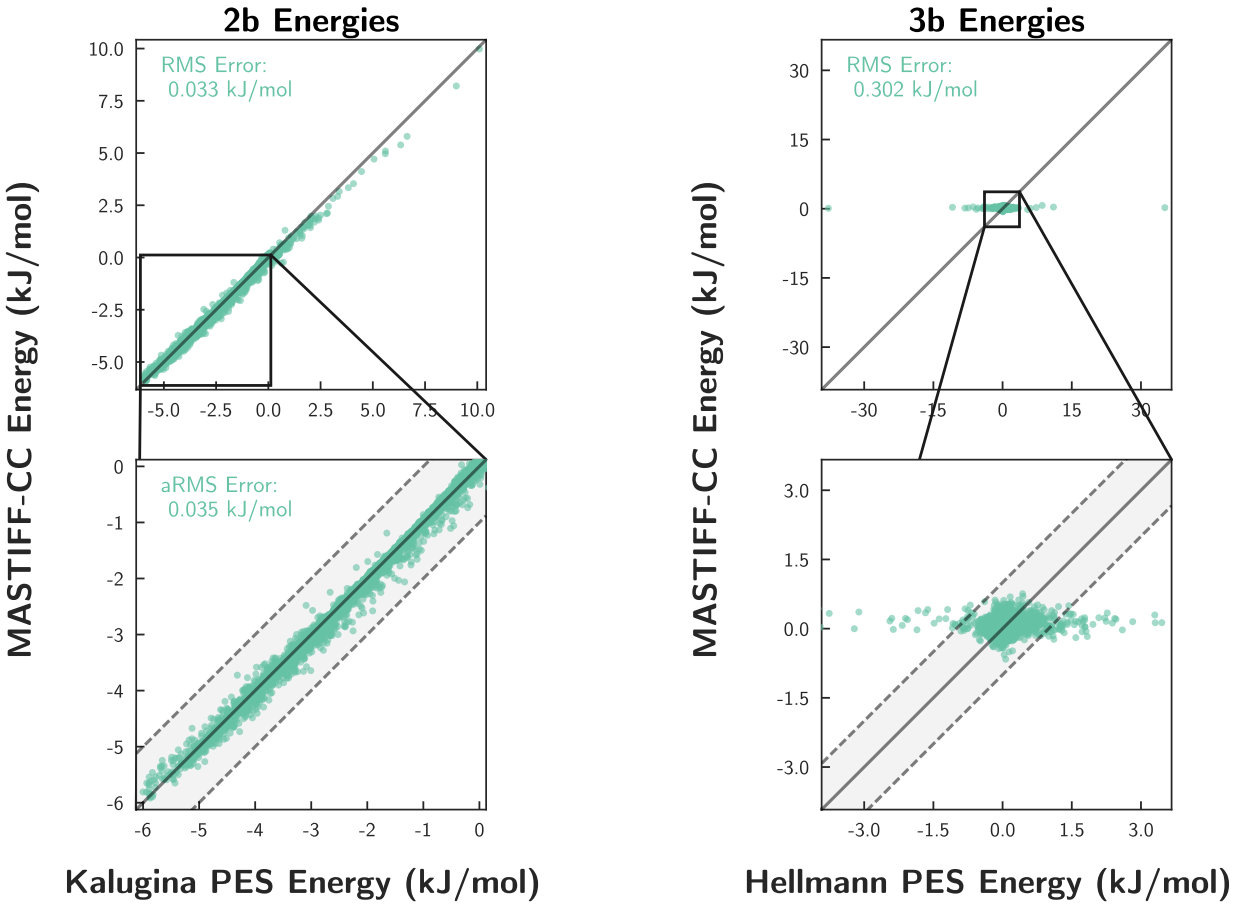


Figure 3.9: Force field quality for MASTIFF-CC in reproducing (left) two-body and (right) three-body CO_2 interaction energies. The $y = x$ line (solid) and ± 1 kJ/mol boundaries (dashed lines) are shown for reference. All dimer/trimer configurations were taken from a snapshot of CO_2 liquid simulated using MASTIFF-CC at 273.15 K and 100 bar. Reference energies are taken from the Kalugina et al.[17] PES for the two-body energies, and the Hellmann[16] PES for the three-body energies. A total of 62,583 dimer configurations and 43,784 trimer configurations are represented in the two plots.

Part II

Unpublished Work

4 AB INITIO FORCE FIELDS USING LMO-EDA

4.1 Preface

The preceding sections have been devoted to a development of various methodologies for ab initio intermolecular force field development, all generally assuming that Symmetry-Adapted Perturbation Theory (SAPT) can be used as a benchmark electronic structure theory. Critically, and especially given the developments discussed in Chapter 3, we can now usually expect our model force field energies to be within ~1 kJ/mol of the SAPT reference values! In spite of this success, this high precision between the model and SAPT energies can only lead to experimentally-accurate molecular simulation in the event that the SAPT energies themselves are accurate with respect to the exact underlying potential energy surface (PES). practice) Indeed, for systems where SAPT and CCSD(T) (a gold-standard electronic structure theory that closely matches the exact PES) disagree by several kJ/mol, there is little advantage in developing SAPT-based force fields with sub- kJ/mol precision. This limitation raises to two fundamentally important questions. First, for what types of systems might we expect SAPT to be inaccurate? Second, for the systems where SAPT and the exact PES are in disagreement, how might we best modify our typical methodology for ab-initio force field development?

The purpose of this chapter is to partially address these two questions, all within the specific context of force field development for Coordinatively-Unsaturated (CUS) Metal-Organic Frameworks (MOFs). Importantly, the results presented here were gathered from 2012–2014, so some important advances (namely those presented in Chapters 2 and 3) have yet to be incorporated into the force fields presented here, probably to the determinent of the accuracy and transferability that might be possible with the LMO-EDA-based methodology. Should this project be picked up in the future, it will likely prove advantageous to refit the LMO-EDA-based force fields described herein to the functional forms and monomer-based parameters discussed in Chapter 3.

4.2 Introduction

Metal-Organic Frameworks (MOFs) are an increasingly important class of compounds, and are defined as porous materials containing inorganic nodes connected by organic linkers. Within this general motif, more than 20,000 compounds have been reported and studied,[271] and this vast diversity of MOF materials shows great promise for chemical customization and optimization. Within the past two decades, a huge body of research has been devoted to the design and study of MOFs, and current applications range from gas separation and storage to catalysis and biomedical imaging.[271]

Somewhat recently, it has been discovered that so-called CUS MOFs can be created by activation of solvent-coordinated inorganic nodes to yield exposed (or 'open') metal sites.[272–274] These CUS-MOFs have been shown to exhibit exhibit excellent uptakes and selectivities in a number of gas separation and storage problems,[272,273,275] making this family of compounds an excellent target for future investigation and materials design. Owing to the vast scope of hypothetical CUS-MOF materials, however, and the number of chemically-distinct targets for gas separation/storage, it is unlikely that experiment alone can be used to screen for new and promising CUS-MOF materials.[276] Rather, a combination of experiment and computational modeling will be required to identify (or possibly even rationally design) optimal CUS-MOFs.[275–277]

Despite the utility of computational studies, it remains challenging to develop molecular models for CUS-MOFs.[274,276,277] Because the strong binding between metal and adsorbate leads to chemical environments substantially different from typical coordinatively-saturated MOFs, many standard force fields (such as UFF and DREIDING) that yield good predictions for these CS-MOFs can frequently (and substantially!) underpredict adsorption in CUS-MOFs.[276–278] Importantly, these underpredictions are especially prominent at low pressures, where metal-adsorbate interactions dominate.[276–278] While CUS-MOFs can sometimes be studied using quantum mechanical means,[277,279,280] clearly new and improved force fields will be required to perform in-depth simulations and large-scale screenings of these

materials, and such studies are already being undertaken.[281–284]

The goal of the present chapter is to present a general methodology for developing accurate and transferable force fields for CUS-MOFs. The current study is limited to a discussion of the MOF-74 series (a prototypical and well-studied CUS-MOF), however it is expected that the methods presented herein might also be applicable to other systems. After outlining this methodology (Sections 4.3 and 4.4), we next show how our force fields can be applied to accurately predict CO_2 adsorption isotherms in Mg-MOF-74. At the present time, we do not have results for other compounds in the M-MOF-74 series (M = Co, Cr, Cu, Fe, Mn, Ni, Ti, V, and Zn), largely as a result of techical challenges in the force field parameterization itself. We discuss these technical challenges in some detail, and conclude with our perspective on the challenges and opportunities associated with developing transferable force fields for the M-MOF-74 series and other similar CUS-MOF systems.

4.3 Background and Motivation

Prior work in our group has shown how, at least for coordinatively-saturated MOFs, accurate and transferable force fields can be generated for a wide variety of systems by fitting force field parameters on a component-by-component basis to reproduce an ab initio SAPT energy decomposition.[141,197] While full details for this force field development methodology can be found in Refs. 197, 285, a short workflow is given here for Mg-MOF-74:

1. Generate a representative cluster model from which interaction parameters can be determined for each pairwise interaction. An example cluster, used to parameterize Mg–CO_2 interactions in Mg-MOF-74, is shown in Fig. 4.1.

2. Using DFT-SAPT (a variant of SAPT with monomer densities given by Density Functional Theory (DFT)), compute a series of representative dimer interaction energies for the model cluster. For the cluster model in Fig. 4.1, representative dimers were generated by varying the position of CO_2 with respect to the

MOF cluster, and the corresponding DFT-SAPT total interaction energies are shown for a subset of representative points.

3. To determine partial charges for the system, generate representative clusters (as described in Section 4.5) for each the organic ligand and the inorganic node, and perform a Distributed Multipole Analysis (DMA) analysis on each cluster to determine partial charges for the overall system.

4. For each component of the DFT-SAPT interaction energy, parameterize the relevant functional forms (as detailed in Ref. 197 and Section 4.4) to reproduce the DFT-SAPT component energy.

Once parameterized, these SAPT-based MOF force fields can be used for calculating individual adsorption isotherms or even for high-throughput screening.[285]

In the generation of force fields for CUS-MOFs, we expect that many of the advantages of the development methodology for coordinatively-saturated MOFs (such as the component-by-component based parameterization and protocol for partial charge determination) should also translate well to CUS-MOF materials. Nevertheless, there are two reasons why a SAPT-based methodology is non-ideal for generating CUS-MOF force fields. First, and as shown in Fig. 4.1 for a representative Mg-MOF-74 cluster model, by comparing to benchmark CCSD(T)-f12 calculations we have discovered SAPT to be in error for CUS-MOF-like systems. DFT-SAPT is known to struggle with highly ionic systems (relative to CCSD(T) or DFT methods),[286,287] and so this error is perhaps not surprising. (Possible sources of the discrepancy between SAPT and CCSD(T)-f12 will be discussed in Section 4.8.) Nevertheless, and in the absence of fortuitous error cancellation, predictions from an ab initio force field can only be as good as the level of theory that they are parameterized against. Consequently, because SAPT underbinds CO_2 by a full 6 kJ/mol compared to CCSD(T)-f12, we would not expect to see good predictions for the CO_2 adsorption isotherm with a SAPT-based methodology. For CUS-MOFs and other similar systems, a new strategy for force field development is required.

As a second barrier to using a SAPT-based methodology, many of the compounds in the M-MOF-74 series are open-shell. Though this poses no fundamental

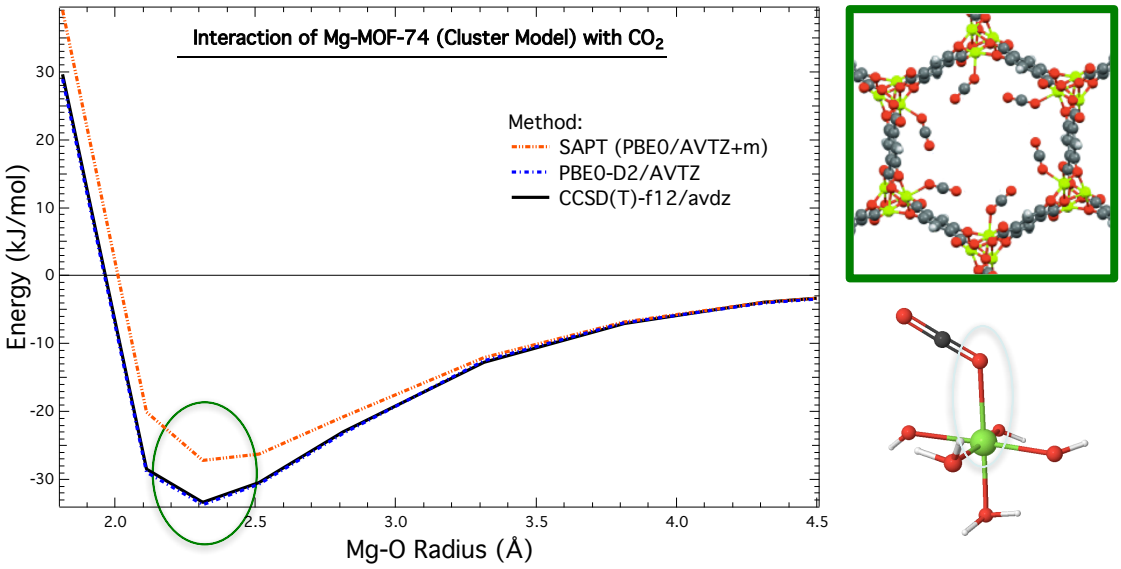


Figure 4.1: Model PES for interactions between CO_2 and Mg-MOF-74. (Left) Interaction energies between CO_2 and a cluster model of Mg-MOF-74 (shown bottom right), computed at a CCSD(T)-f12 (black), SAPT (orange), and/or PBE0-D2 (blue) level of theory. Discrepancies between SAPT and CCSD(T)-f12 in the minimum-energy region of the potential have been highlighted. (Top right) The structure of CO_2-bound Mg-MOF-74. (Bottom right) The structure of the cluster model used for Mg-MOF-74, where the circled atom pair indicates the relevant Mg-O radius from the x-axis in the leftmost figure.

issue, in practice most implementations of SAPT (aside from the seldom-used SAPT 2012 package developed in Krzysztof Szalewicz's group at Delaware) do not allow for computations of open-shell systems, and indeed SAPT-based studies of open-shell compounds are very rare.[288] For these reasons, a new, open-shell-compatible electronic structure benchmark is highly preferable.

4.4 Parameterizing CUS-MOF force fields with LMO-EDA

Based on the results for Mg-MOF-74, it is clear that, at least for CUS-MOFs, a new methodology is required which simultaneously keeps the important advantages of the old development strategy (especially the component-by-component based parameterization, which is essential for generating transferable force fields) while overcoming the limitations of SAPT itself. Put differently, for CUS-MOFs we should seek a new electronic structure theory benchmark and associated Energy Decomposition Analysis (EDA) with the following qualities:

1. High accuracy with respect to to CCSD(T)-f12 benchmark energies
2. Physically-meaningful energy decomposition into (at least) electrostatics, exchange, induction, and dispersion
3. Quantitative correspondence between the energy decompositions of SAPT and the new method for systems where total energies from SAPT, CCSD(T)-f12, and the new method agree

Assuming these three qualites are met, we expect to be able to generate force fields for CUS-MOFs that are both highly accurate and maximally-compatible with previous force fields developed for coordinatively-saturated MOF systems.

A substantial number of EDAs exist in the literature, and the interested reader is referred to Ref. 287 for a review and comparison of various popular methods. Aside from SAPT, which is a perturbative method, most EDAs are 'variational', meaning that the various energy components are calculated in stages from a series of constrained relaxations of the monomer wavefunctions into the optimized dimer wavefunction. For this reason, all variational EDAs are guaranteed to have total energies that match the result from a supermolecular interaction energy calculation. Furthermore, these EDAs are often implemented for wavefunction and DFT methods, thus allowing for significant flexibility (compared to the SAPT EDA) in terms of finding an EDA whose total energy closely matches CCSD(T)-f12. Indeed,

and as shown in Fig. 4.1, PBE0-D2 shows excellent agreement with CCSD(T)-f12 for a Mg-MOF-74 cluster model, and so any DFT-compatible EDA should meet our first criteria from above.

Although all variational EDAs yield the same total interaction energy for a given level of theory, many EDAs can differ substantially in terms of how this total energy is decomposed into chemically-meaningful components. At the time this research was completed, only a handful of variational EDAs distinguished each electrostatics, exchange, induction, and dispersion. (Notably, the recent second-generation ALMO-EDA[289] now separates their 'frozen' energy term into electrostatic, exchange, and dispersion components, and thus might be worth future investigation.) Of the popular EDA methods available in 2014, we found that LMO-EDA,[290,291] GKS-EDA,[292] and PIEDA[293] decompose the total interaction in a manner philosophically similar to SAPT, and include each electrostatic, exchange, induction, and dispersion terms. These three methods thus meet our second criteria for an optimal energy decomposition scheme for CUS-MOFs, and complete formalisms and details for the methods can be found in Refs. 290–293.

As for the last criterion, that of maximum correspondence between SAPT and a variational EDA, we have performed component-by-component analyses to compare SAPT to both LMO-EDA and GKS-EDA. PIEDA is known to overestimate the relative magnitude of the polarization energy, compared to SAPT, and thus was not considered in detail.[287] As for LMO-EDA and GKS-EDA (both of which are based on very similar theories, and tend to yield similar energy decompositions), we have in general found semi-quantitative to quantitative agreement with the SAPT energy decomposition, particularly for the electrostatic and exchange energies. Comparisons between LMO-EDA and SAPT are shown for the CO_2 dimer (Fig. 4.2) and for CO_2 interacting with a model Mg-MOF-74 compound (Fig. 4.3). GKS-EDA results are not shown, as the LMO-EDA and GKS-EDA results tend to be very similar, with the GKS-EDA results in slightly worse agreement with SAPT. For this reason, and because LMO-EDA does the best job of meeting our three criteria above, we choose in this work to use LMO-EDA as our new benchmark EDA for fitting CUS-MOF force fields.

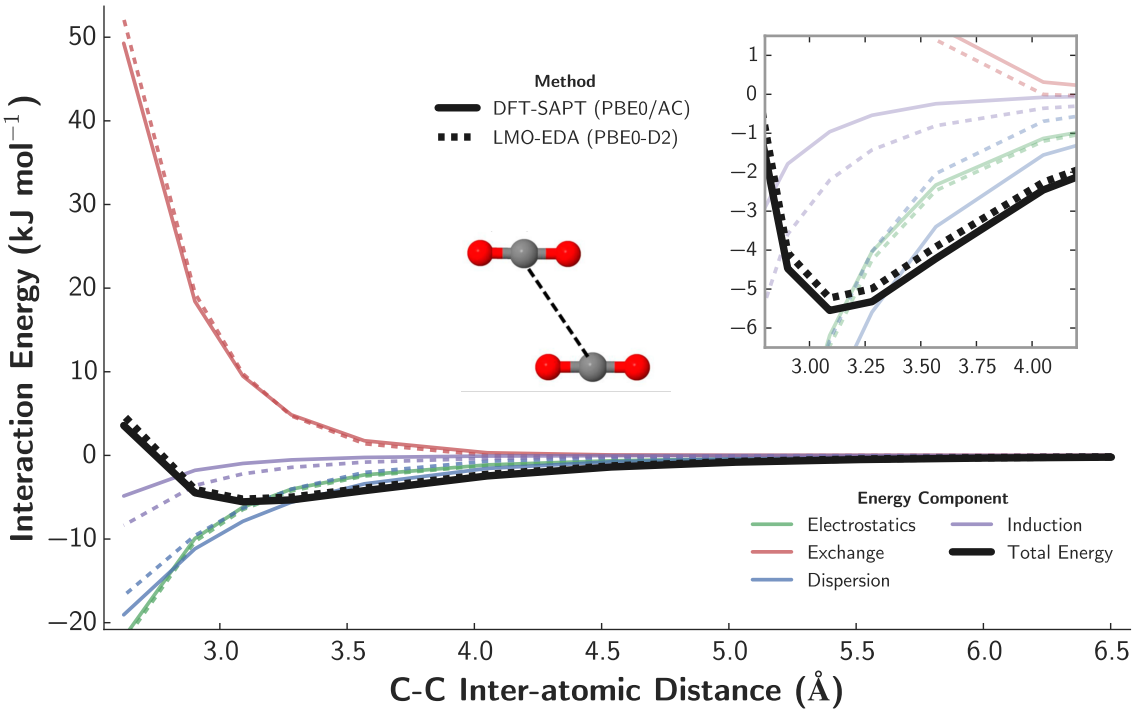


Figure 4.2: PES and associated energy decomposition for the slipped parallel geometry of the CO_2 dimer as a function of the C–C interatomic distance. The PES has been computed by both DFT-SAPT (PBE0) (solid lines) and LMO-EDA-PBE0-D2 (dashed lines), and each electrostatics (green), exchange (red), dispersion (blue), induction (purple), and total energy (black) components are displayed. Note that, for the DFT-SAPT energies, the δHF contribution has been incorporated into the induction energy.

In addition to describing the advantages of the LMO-EDA method, it is worthwhile to overview some of its relevant shortcomings and limitations. As with most variational EDA methods,[287] and especially for DFT-based methods, it becomes difficult to precisely assign and separate out the true 'dispersion' energy for a system. This limitation is also true of LMO-EDA, where the dispersion energy is defined as the difference in correlation energy between the monomer and dimer wavefunctions. (For density functionals employing Grimme's –D dispersion correction, this correction is also added to the LMO-EDA dispersion energy.) For functionals that

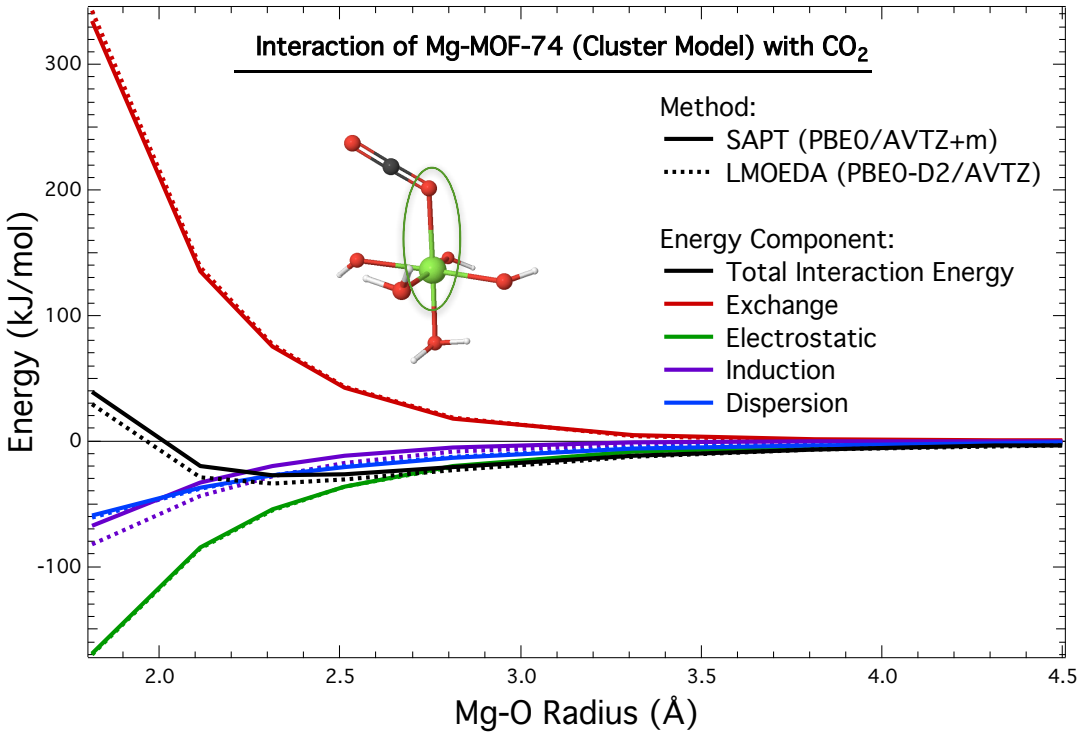


Figure 4.3: PES and associated energy decomposition for a CO_2 + MgO_5 cluster, as a function of the highlighted Mg–O interatomic distance. The PES has been computed by both DFT-SAPT (PBE0) (solid lines) and LMO-EDA-PBE0-D2 (dashed lines), and colors and labels for the energy decomposition are as in Fig. 4.2.

have a well-defined and theoretically-grounded distinction between the exchange and correlation functionals, the LMO-EDA energies tend to agree well with SAPT, and we have found good agreement (for instance) between SAPT and LMO-EDA-PBE0-D2. With other functionals, such as with our tests using the M06 functional, there is no separation between the exchange and correlation functionals, and LMO-EDA gives unphysical values for both the exchange and dispersion energies in this case. (Notably, GKS-EDA attempts to rectify this issue by changing the LMO-EDA formalism for dispersion. While this leads to qualitative agreement between SAPT and GKS-EDA for a wider variety of functionals, the quantitative agreement for the PBE0-D2 functional is somewhat worsened for the systems studied herein, and

we instead use LMO-EDA-PBE0-D2 for all results in this work.)

A second, and purely practical, limitation of LMO-EDA is its memory-intensive implementation in GAMESS. As will be discussed in detail later, calculations on a large (43 heavy atom) cluster model of Mg-MOF-74 were infeasible for us (using the Phoenix cluster in 2014) in all but the smallest VDZ basis set, and calculations on an identical cluster model of Co-MOF-74 could not be completed at all. For this reason, the LMO-EDA method is practically restricted to studies of smaller systems and/or basis sets.

4.5 Computational Methods

4.5.1 Partial Charge Determination

Partial charges for Mg-MOF-74 were determined in a manner analagous to Ref. 285 using the Q_{SBU} method. Two cluster models, one a hydrogen-capped DOBDC ligand environment, and one a capped MgO_5 inorganic chain, were constructed and analysed using a Distributed Multipole Analysis (DMA). The resulting DMA charges were then used to obtain charge paramters for the ligand and inorganic SBU, respectively. See Section 4.A for final charge parameters.

4.5.2 Force Field Fitting

Two types of force field functional forms were considered in this work. The first, a 'single-exponential' functional form, exactly matches that used in Ref. 83, with the exception that δHF parameters were not fit to the Mg atomtype. This fitting choice was due to the fact that LMO-EDA only provides a total induction term (rather than splitting into 2^{nd}- and higher-order induction energies, as with SAPT).

For the 'double-exponential' functional form used to fit the Mg-MOF-74-Yu cluster model, the same functional form was used as in the single-exponential case, with the exception that two sets of short-range interaction parameters (labeled Mg and Du in Section 4.A) were assigned to the Mg atomic center. This effectively

meant that Mg was described by two separate exponential decays, thus enabling additional parameterization flexibility for the force fields discussed in Section 4.6.

In all cases, force fields were fit using the Fortran code described in the Appendix of Ref. 81.

4.6 Results

4.6.1 Initial Force Field and Cluster Model Analysis

Originally, we attempted to fit Mg parameters on the basis of a small, 6 heavy atom cluster ('Mg-MOF-74-small', see Fig. 4.4 for chemical structure), which we felt would be representative of the Mg environment in Mg-MOF-74. Using the functional forms discussed in Section 4.5, force field parameters were fit to reproduce LMO-EDA-PBE0-D2 energies for a variety of CO_2/Mg-MOF-74-small interactions, with results shown in Fig. 4.4. Though select interaction energies disagree by several kJ/mol between LMO-EDA-PBE0-D2 and the force field energies, overall the agreement is reasonable, and the force field correctly reproduces trends in the interaction energies without significant systematic error.

Based on the agreement between PBE0-D2 and the force field, as well as between PBE0-D2 and CCSD(T)-f12, we expected to obtain good CO_2 adsorption isotherm predictions for the Mg-MOF-74 system itself. By contrast, our computed isotherm substantially underpredicts the experimental adsorption at low pressures, where Mg–CO_2 interactions are known to dominate. This underprediction strongly suggests that we had originally underestimated the magnitude of the Mg–CO_2 binding, a result which we were then able to attribute to our choice of cluster model (vida infra).

Cluster models for the M-MOF-74 series have been investigated by several groups, and it has been found in general that computed binding energies are sensitive both to the size of the cluster model as well as the treatment of geometry relaxation effects.[295,296] Consequently, we calculated the CO_2 binding energies and geometries of both our original Mg-MOF-74-small cluster as well as for two larger

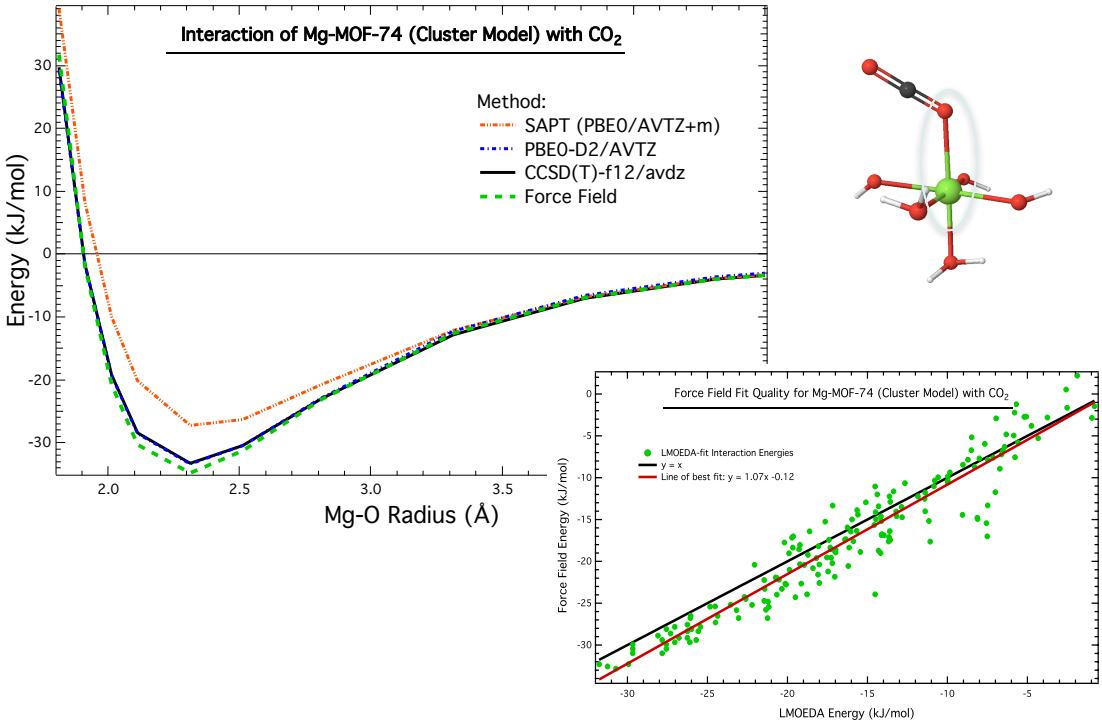
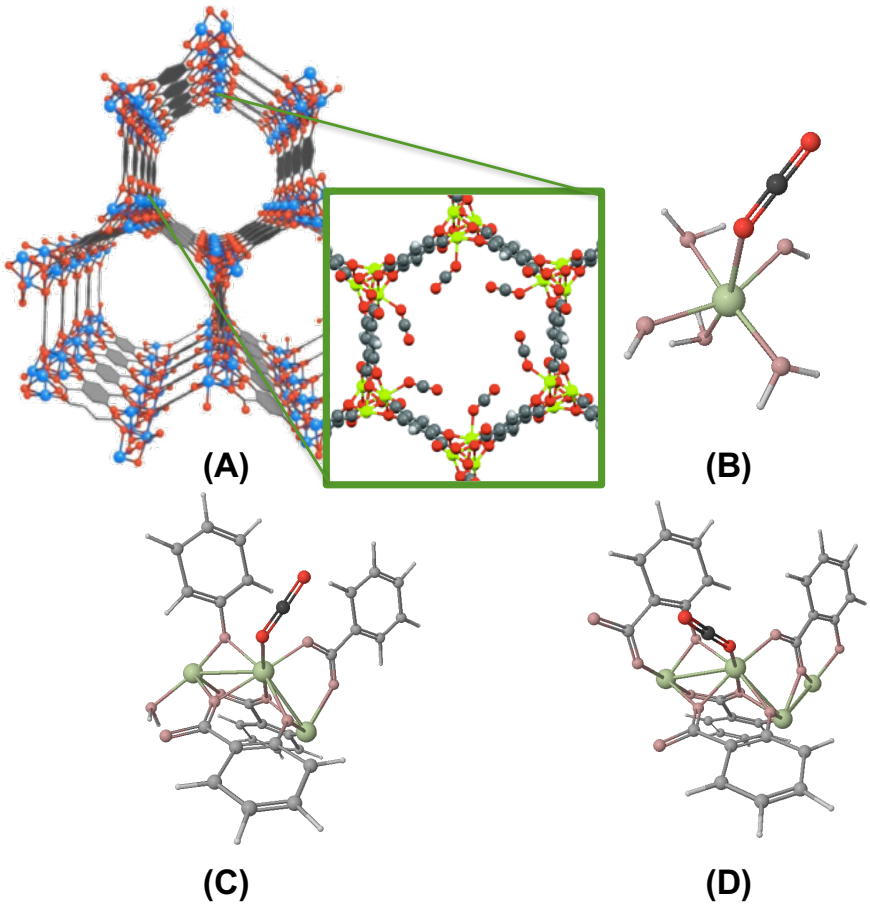


Figure 4.4: Force field fitting quality for the Mg-MOF-74-small cluster. (Top left) Various electronic structure benchmarks for Mg-MOF-74-small along with the classical potential. Each DFT-SAPT (orange dot-dashed), PBE0-D2 (blue dot-dashed), CCSD(T)-f12 (black solid), and the LMO-EDA-based force field (green dashed) are shown as a function of the Mg–O interatomic distance (non-bonding pair highlighted at top right). (Bottom right) Force field fit quality, as benchmarked against LMO-EDA-PBE0-D2, for a semi-random set of dimer configurations of the Mg-MOF-74-small cluster model interacting with CO_2. The black line establishes the $y = x$ benchmark, and the red line represents the line of best fit.



Model	CO_2 Binding Energy (kJ/mol)	Mg–O Interatomic Distance (Å)	Mg–O–C Tilt Angle (°)
A[279]	**-41.5**	**2.31**	**129**
B	-23.3	2.31	122
C	-31.4	2.28	123
D	-41.7	2.20	149

Figure 4.5: Various structures and cluster models for Mg-MOF-74 interacting with CO_2. (A) Full periodic Mg-MOF-74 structure with inset showing adsorbed CO_2 positions. (B) Mg-MOF-74-small cluster, containing 6 heavy atoms (not including CO_2). (C) Yu et al. cluster model for Mg-MOF-74, denoted in text as Mg-MOF-74-Yu. (D) Dzubak et al. cluster model for Mg-MOF-74, denoted in text as Mg-MOF-74-Dzubak. All cluster models as shown with optimized CO_2 positions, and bond lengths and angles for adsorbed CO_2 are given in the bottom table. Data for (A) was taken from Valenzano et al. using a B3LYP-D level of theory,[279] whereas data for (B-D) was computed in this work using PBE0-D2. Finally, note that the binding energy for (A) includes framework geometry relaxation effects, whereas (B-D) were computed using semi-rigid cluster geometries and only optimizing the CO_2 position and exposed MgO_5 pocket.

clusters developed in Refs. 274, 294. These latter two clusters, respectively denoted Mg-MOF-74-Yu and Mg-MOF-74-Dzubak, are the same size (each with 60 atoms), but have distinct stoichiometries and geometries. To test the influence of model cluster on the CO_2 binding energy/geometry, we performed two sets of optimizations of the Mg-MOF-74-Yu and Mg-MOF-74-Dzubak clusters: one in which only the CO_2 position was optimized, and one in which the exposed MgO_5 pocket was additionally relaxed. Binding geometries were relatively insensitive to the geometry relaxation, though binding energies varied by 2-5 kJ/mol, in agreement with other studies that have tested geometry relaxation effects.[295] Results for the $CO_2 + MgO_5$ relaxation are shown in Fig. 4.5.

Of the three studied cluster models, both Mg-MOF-74-small and Mg-MOF-74-Yu correctly reproduce the Mg–O interatomic distance and Mg–O–C tilt angle. These geometrical parameters arise primarily from electrostatic interactions between CO_2 and the MgO_5 pocket,[279] suggesting that both of these models capture such important interaction features. By contrast, the Mg-MOF-74-Dzubak model predicts a substantially shorter binding distance and increased tilt angle, both in contrast to results from the periodic system. In part, these deficiencies can be attributed to spurious CO_2 interactions with the exposed carbonyl capping groups in the Mg-MOF-74-Dzubak model, as these exposed carbonyls are not present in the periodic system or the other two cluster models. As a second distinction, a Mulliken charge analysis of the Mg-MOF-74-Dzubak cluster yields larger partial charges for the surrounding Mg atoms as compared to the Mg-MOF-74-Yu model, which may help explain the increased binding and shortened Mg–O contact in the Mg-MOF-74-Dzubak model.

There are also substantial differences in binding energies between the various cluster models. Importantly, Mg-MOF-74-small severely underbinds CO_2 compared to all other tested systems. These results for the Mg-MOF-74-small cluster indicate the inadequacy of such a small model, and likely explain the underprediction of the CO_2 adsorption isotherm from above. The Mg-MOF-74-Dzubak model shows best energetic agreement with the periodic system. Nevertheless, some of the Mg-MOF-74-Dzubak binding energy arises from truncation effects (as described above), and

the energetic agreement is thus due (at least in part) to error cancellation. Indeed, some of the binding energy in the periodic system arises from (attractive) long-range interactions, and thus we should expect to see a cluster model somewhat underpredict the binding energy. Primarily for its good agreement in binding geometries, and reasonable agreement in binding energy, we opt to use the Mg-MOF-74-Yu cluster model for the remainder of this work.

4.6.2 Final Mg-MOF-74 CO_2 Adsorption Isotherm

Using our new Mg-MOF-74-Yu cluster model, we next attempted to refit force field paramters for Mg. As discussed earlier, and because of the size of this new cluster (60 atoms), LMO-EDA-PBE0-D2 calculations became cost prohibitive in all but the smallest VDZ basis set, and thus could only be carried out for a limited set of points. Starting from the minimum energy configuration shown in Fig. 4.5, we fit Mg parameters to a 12-point scan along the Mg-O bond vector, with fit results shown in Fig. 4.6. Interestingly, though the functional form used in this fit was sufficient to accurately paramterize the interaction energies in the Mg-MOF-74-small cluster, the same force field methodology proved unsuccessful in paramterizing Mg-MOF-74-Yu–CO_2 interactions. We knew at the time that this inaccuracy was probably a consequence of uncertainties in correctly parameterizing the Mg short-range exponent. (See Chapter 2 for a full discussion of new methods for parameterizing the short-range potential.) Nevertheless, because the Slater-ISA methodology for short-range interactions had not yet been developed, we opted instead to fit the Mg interactions to a double exponential functional form, with each exponent corresponding to the ionization potential for either Mg^+ or Mg^{2+} (the two atomic environments most likely to correctly represent the Mg cation). As shown in Fig. 4.6, this form could excellently reproduce the Mg-MOF-74-Yu model PES.

Using the double exponential functional form from above, we recomputed the adsorption isotherm of CO_2 in Mg-MOF-74. Before comparing to experiment, and as recommended by others,[297] we scaled the experimental isotherm in order to account for the pore blocking effects that are common in the M-MOF-74 series.

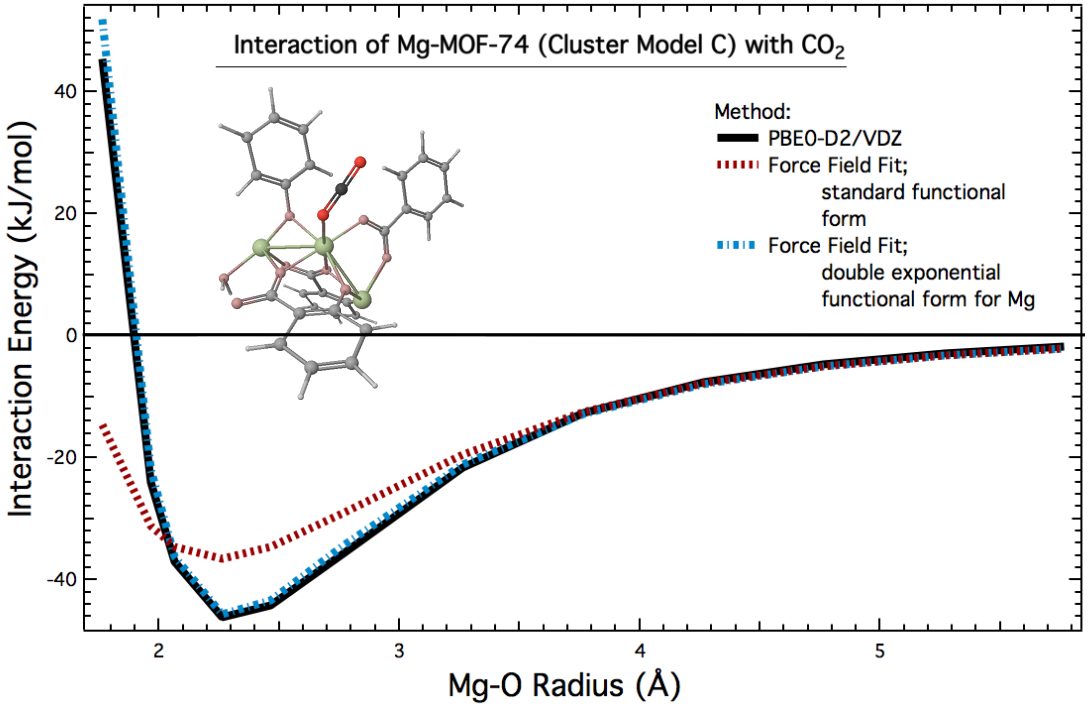


Figure 4.6: Force field fitting quality for the Mg-MOF-74-Yu cluster. A PBE0-D2 benchmark (black solid) is displayed along with two force field fits: single-exponential (red dashed) and double-exponential for Mg (blue dash-dotted). In either case, a cut of the PES is shown along the interatomic distance between the central Mg atom and the closest-contact oxygen atom in CO_2.

Using this scaled isotherm, we then obtain excellent agreement between our model potential and experiment (Fig. 4.7). Crucially, this accuracy is seen both at low- and high-pressure ranges, indicating the accuracy of the force field in modeling both the strong Mg–CO_2 binding as well as the weaker physisorption regime.

4.6.3 Transferability to Other Adsorption Isotherms

In addition to using our Mg paramters to compute the CO_2 adsorption isotherm, we also used our Mg force field in conjunction with the N_2 parameters developed by Yu et al.[166] to predict the N_2 adsorption isotherm. These predictions were generally

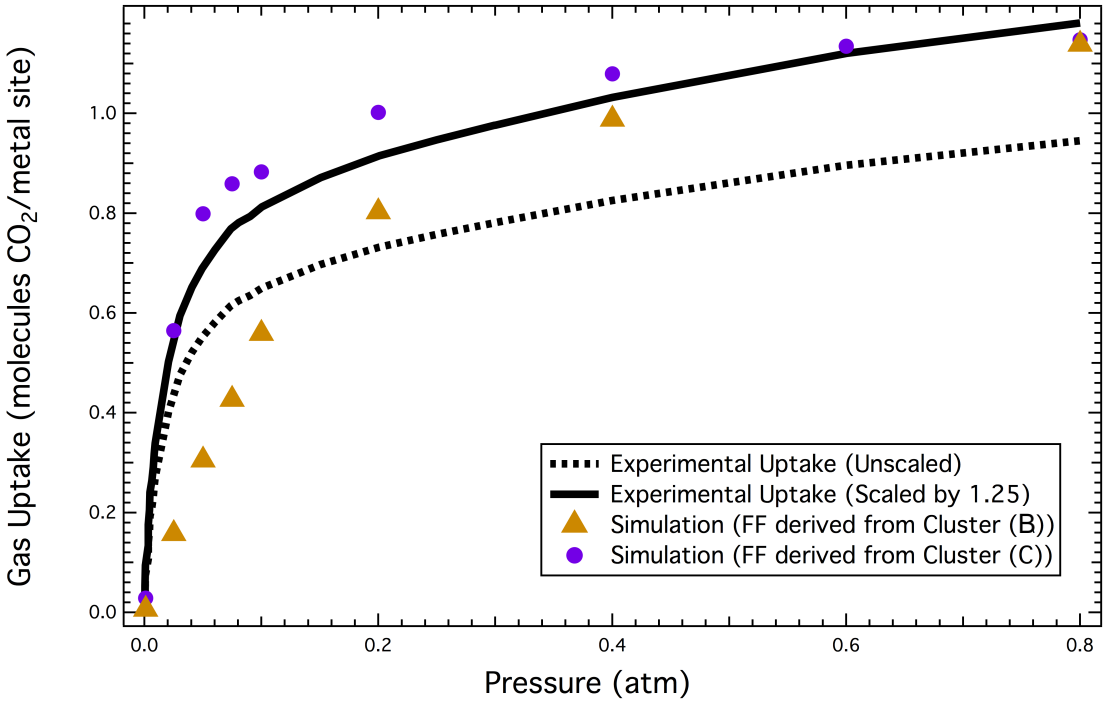


Figure 4.7: Predicted CO_2 adsorption isotherm for Mg-MOF-74. Two experimental isotherms are shown, one as directly measured by experiment (dashed line) and one scaled to account for pore block effects (solid line). Predictions from two force fields are also shown, where Mg parameters for each force field were fit either to the Mg-MOF-74-small cluster (gold triangles) or to the Mg-MOF-74-Yu cluster (purple circles). Cluster geometries are given in **??**.

poor, and results are not shown. Nevertheless, the poor N_2 results suggest a lack of transferability of our Mg parameters, possibly (and as discussed in the section on Future Work) due to the unphysical double-exponential functional form used to parameterize Mg.

4.6.4 Transferability to Other M-MOF-74 systems

As a second test of transferability, we also attempted to develop force fields for other compounds in the M-MOF-74 series, starting with Co-MOF-74. Unfortunately, the open-shell nature and increased electron count of Co-MOF-74 made LMO-EDA

calculations computationally prohibitive for any reasonable basis set, and these systems were not investigated further.

4.7 Conclusions

In this unpublished work, we have determined a new methodology for fitting force fields to CUS-MOFs. While largely following our previous methodology for MOF force field development, we have shown how an LMO-EDA Energy Decomposition Analysis can be used in leiu of SAPT to generate accurate ab initio benchmark energies in cases where SAPT itself is inaccurate. Using this new methodology, we have successfully modeled CO_2 interactions in Mg-MOF-74, and have simulated the adsorption isotherm for CO_2 in Mg-MOF-74 with good accuracy compared to experiment.

Ultimately, the methods presented herein suffered from a number of practical and fundamental issues (vida infra). Until these limitations can be fully addressed, we do not anticipate that the LMO-EDA method can be broadly used to develop transferable force fields for CUS-MOFs or other large systems where SAPT is in error.

4.8 Future Work

Throughout this Chapter, we have attempted to highlight some of the key limitations of our force field development methodology for CUS-MOFs. In summary, the following issues would need to be resolved in order to expand the scope and utility of the present research:

1. **Memory Limitations in GAMESS**: As evidenced in this work, relatively large (60+ atom) cluster models are required to correctly parameterize force fields for the M-MOF-74 system. While these cluster sizes do not present difficulties for standard DFT calculations with reasonable basis sets, the corresponding LMO-EDA calculations were, as implemented in the GAMESS software pack-

age, infeasible due to memory requirements. Some time was spent attempting to address these memory issues, particularly for the memory-intensive Edmiston-Ruedenberg localization subroutine that is the source of the problem. However, due to our lack of familiarity with the GAMESS software and the LMO-EDA source code, this pursuit was eventually dropped.

2. **Fundamental Issues with LMO-EDA**: As discussed in Section 4.4, the LMO-EDA method has several theoretical limitations. In particular, and especially for functionals with no defined separation between exchange and correlation functionals, LMO-EDA does not offer a clean separation between the exchange and dispersion energies. Furthermore, and unlike some recent EDA methods,[154,289] LMO-EDA cannot separate induction into charge transfer and polarization components.

3. **Transferability of the Force Field Functional Form**: While our final force field for studying CO_2 interactions in Mg-MOF-74 is highly accurate (both with respect to ab initio theory and with respect to experiment), it does not appear that this accuracy extends to models for the adsorption of other small molecules, such as N_2. This transferability limitation is almost certainly due to the chosen double-exponential functional form and/or the parameterization process used to obtain Mg parameters, and improvements to this methodology will be essential to make our work on the CO_2–Mg-MOF-74 system applicable to general force field development for CUS-MOFs. In particular, future work will require a better force field for describing short-range interactions, as the functional forms and paramters used in this work struggled to both accurately and transferably model the Mg-MOF-74 exchange energies.

While several of these issues (particularly practical limitations with the LMO-EDA implementation) have yet to be addressed in a meaningful fashion, several recent theoretical advances may pave the way for continued work on this project. Thus for CUS-MOFs and other systems where DFT-SAPT might be in error, we offer the following recommendations:

1. **Improved SAPT energies**: Recently, it has been proposed that the commonly used single-exchange ('S^2') approximation can lead to errors in the description of the induction energy, particularly for ionic systems.[286,298] While it is difficult to attribute errors in SAPT to a particular energy component, it may well be that SAPT poorly describes Mg-MOF-74 due to the S^2 treatment of the induction energy. In this case, eliminating the S^2 approximation might improve the DFT-SAPT total interaction energies, thus enabling SAPT to be used for (at least) closed-shell CUS-MOFs.

2. **New SAPT Correction Schemes**: As discussed in Chapter 3, deviations between SAPT and CCSD(T) can be rectified by adding a δCCSD(T) correction to the total SAPT energy. As discussed in Chapter 3, we have empirically had good success modeling this δCCSD(T) correction as part of the dispersion energy. Though this partitioning choice may require adjustment for treating Mg-MOF-74, the results in Chapter 3 indicate that simply correcting (rather than entirely ignoring) the DFT-SAPT energies is a promising strategy for transferable force field development.

3. **New EDA schemes**: Since this work was completed, a second-generation ALMO-EDA scheme has been implemented in the Q-Chem software package.[289] Crucially, and unlike its predecessor, this ALMO-EDA scheme now breaks up the interaction energy into electrostatic, exchange, polarization, charge-transfer, and dispersion components. While there is no guarantee that such an EDA could serve as the basis for CUS-MOF force field development (see Section 4.4), these and other recently developed EDAs may be worth investigation, and could eventually replace the (practically problematic) LMO-EDA method.

4. **Improved Force Field Functional Forms**: Since 2014, we have made significant progress in developing more accurate and transferable intermolecular force fields (see Chapters 2 and 3), and many of these advances particularly improve the description of the short-range potential itself. There is a good

chance that either the Slater-ISA FF or MASTIFF methodologies would yield high-quality force fields for Mg-MOF-74 and other systems. In this case, continued work on this project might be an exciting avenue for showcasing the MASTIFF methodology in the context of accurate inorganic/organometallic force field development.

4.A Force Field Parameters for CO_2 and Mg-MOF-74

Final force field parameters, fit using the double-exponential functional form above and the Mg-MOF-74-Yu cluster model, for CO_2 and Mg-MOF-74. These parameters should be read in as input into our group's lattice simulation code, see `http://schmidt.chem.wisc.edu/montecarlosimulationcodes` for details.

Listing 4.1: co2_mof74.pmt

```
lennard_jones_type      1       "1 for buckingham, 2 for lennard jones"

" parameters are listed as charge, A,B,C, polarizability
" units are A:kJ/mol,  B:A^-1,  C: KJ/mol*A^6   epsilon:KJ/mol,   sigma:A,
    polarizability: A^3 "
" polarizability is defined as q^2/k, spring constant is set to .1*1.8897^3 e^2/A^3"


solute_species
atom_type_parameters  ( q, Aexch, Aelec, Aind, Adhf, Adisp, C6, C8, C10, alpha )
2
C0      0.6573800       95510.43        -27846.98       -13425.1        2065.044
        0.0     6.891E02        0.0     0.0     1.1926153
O0      -0.328690       521902.066      -163908.84      -4475.8095      -26042.04
        0.0     1.8341E03       0.0     0.0     0.9009290

solute parameters for framework cross terms  ( Aexch, Aelec, Aind, Adhf, Adisp, B, C6, C8
    , C10, C12 )
C0      74376.65        24130.18        12513.37            795.36
        0.0     3.4384      1147.41867      6329.41038      29659.50100     183546.714
O0      354373.13       108208.5        2544.89         -18178.7
        0.0     3.7795          867.27598       4266.54582      28761.10636     132581.301

solute dhf cross terms (check code for input format if more than one cross term)
-6124.0
```

```
solute-solute exponents ( Bii , Bij , Bjj)
3.5105206        3.6993494        3.9288490


framework_species
atom_type_parameters    ( q, Aexch, Aelec, Aind, Adhf, Adisp, B, C6, C8, C10, C12, alpha )
9
C1   -0.1639350   612892.611   229850.679    -6511.529   -60036.930
         0.000         3.438     1628.820002      6821.530007     44464.989999    193602.980000
                 0.0
H1    0.2637500     8538.651     1678.771      -612.739     -502.639
         0.000         3.778      129.439978       679.640001      4995.299998         0.000000
                 0.0
C2   -0.3191850   612892.611   229850.679    -6511.529   -60036.930
         0.000         3.438     1628.820002      6821.530007     44464.989999    193602.980000
                 0.0
C3    0.4964850   612892.611   229850.679    -6511.529   -60036.930
         0.000         3.438     1628.820002      6821.530007     44464.989999    193602.980000
                 0.0
O3   -1.0339500     3398.424     1965.168      -182.412     -178.025
         0.000         2.457     2237.635879     29956.090890    561056.184030   7451461.601923
                 0.0
C4    0.9468300   263600.161   112896.479       -11.681    -2837.170
         0.000         3.438      772.870024      2349.180008     27539.189998    102366.260000
                 0.0
O4   -0.8903225   656757.170   174054.351    -45410.640   -33954.271
         0.000         3.779     1799.560008     11576.089993     50164.639999         0.000000
                 0.0
Mg    1.5906500      917.037     2417.463    -12542.799         0.000
         0.000         2.834      630.723467         0.000000         0.000000         0.000000
                 0.0
Du    0.0000000    29176.333        0.000   142260.481         0.000
         0.000         3.973        0.000000         0.000000         0.000000         0.000000
                 0.0
```

4.B Simulation Parameters CO_2 Adsorption in Mg-MOF-74

Lattice simulation parameters for CO_2 adsorption in Mg-MOF-74. Of particular importance is the 'orientation_try' keyword, which is necessary to sample the specific binding geometries CO_2 adopts when binding to the open-metal site. These simulation parameters should be read in as input into our group's lattice simulation code, see `http://schmidt.chem.wisc.edu/montecarlosimulationcodes` for details.

Listing 4.2: simulation_parameters.pmt

```
Simulation Methodology
energy_decomposition            yes        ! yes for our force fields, no for UFF LJ, etc.
solute_cross_parameter_set      yes        ! this should be set to yes if using different solute parameters
          ! for solute-solute and solute-framework interactions as in our force fields, no otherwise
C8_10_dispersion_terms          yes        ! set to yes if using C8, C10 dispersion terms as in our force fields
C12_dispersion                  yes
electrostatic_type              pme        ! either "pme" for particle-mesh ewald, "cutoff", or "none"
lj_comb_rule                    ZIFFF      ! "opls" or "standard" for lj, "standard" or "ZIFFF" for bkghm

Simulation Parameters
temperature                     296.0          ! temperature in Kelvin
too_close                       1.8            ! reject move if molecules are within this separation in Angstroms.
          ! helpful to avoid unnecessary energy calculations and to prevent drude oscillator catastrophes
lj_bkghm                        1              ! 1 for bkghm force field, 2 for lj
screen_type                     1              ! screening for coulomb potential: 0 = no screening, 1 = Tang-Toennies type screening for our force fields
springconstant                  0.1            ! spring constant for drude oscillators (au). set to 0.1 for our CO2/N2 force fields
thole                           2.0            ! thole parameter for intra-molecular drude oscillator screening.  Set to 2.0 for our CO2/N2 force fields.
drude_simulation                1              ! set to 1 if drude-oscillators are being used, 0 otherwise
pme_grid                        100            ! size of the pme grid
alpha_sqrt                      0.6            ! alpha sqrt for the electrostatic
```

```
    interactions
lj_asqrt                    0.6             ! alpha sqrt for the pme dispersion
lj_cutoff                   7.5             ! cutoff for long range LJ or C6,C8,C10
    dispersion interactions
ewald_cutoff                5.0             ! cutoff for real space pme
cav_grid_a                  30
cav_grid_b                  30
cav_grid_c                  30
na_nslist                   30              ! neighbour list searching grid
nb_nslist                   30              ! neighbour list searching grid
nc_nslist                   30              ! neighbour list searching grid
orientation_try             2000            ! max number of orientation samplings
REL_THRSH                   0.05            ! sampling threshold
ABS_THRSH                   3.0
BZ_CUTOFF                   100.0
```

Part III

Practical Matters

5 APPLIED FORCE FIELD DEVELOPMENT: ELECTRONIC STRUCTURE BENCHMARKS AND MONOMER PROPERTY CALCULATIONS

Due in part to the improvements in Chapters 2 and 3, the development protocol for SAPT-based, ab initio force fields is now fairly robust with respect to many parameterization details. Consequently, much of the workflow is now automated and requires little user input. The following sections are designed to give future users familiarity this workflow, not only as a "blackbox" tool, but also as a starting point for more complex and/or system-specific force field development. To this end, we first provide an overview of the workflow itself, and then describe the theoretical and practical details of each step in subsequent sections.

In order to gain expertise in practical force field development, new force field developers are encouraged to read through (in order) **??** and Chapters 5 and 6 to obtain a conceptual understanding of the force field development process, after which they should work on developng their own force field using the semi-automated workflow (Chapter 5) and the **P**arameter **O**ptimizer for **Inter**-molecular Force Fields (POInter) software (Chapter 6). Developing a force field for water makes for an excellent teaching example, however any interesting (and preferably small!) molecule will suffice.

5.1 Overview

As discussed in **??**, our SAPT-based force field methodology principally relies on modeling two-body interactions for a given system of interest. These two-body (i.e. dimer) interactions are completely defined by the positions and relative orientation of the two constituent monomers, and in practice we parameterize the two-body model based on benchmark SAPT energies for a series of gas-phase

dimer configurations.* We are usually interested in obtaining transferable parameters for a new molecule or atomtype, in which case it is often easiest to model the interactions between two identical monomers (a so-called homo-monomeric dimer interaction).†Still, there are reasons why it can be advantageous to instead study hetero-monomeric dimer interactions, and the workflow described herein applies equally to studying both homo-monomeric and hetero-monomeric dimer interactions.

Regardless of the chosen dimer of study, modeling a two-body PES involves two major steps. First, we must obtain benchmark two-body energies for a series of well-chosen dimer configurations. Second, we must calculate and/or fit all force field parameters so as to completely develop a force field for the two-body interaction energies. For the SAPT-based force fields described in Chapters 2 and 3, these two overarching steps lead to the following workflow:

* At first, it may seem counter-intuitive to focus so heavily on modeling the energetics of gas-phase dimers. After all, aren't we interested in simulating a wider variety of chemically-relevant systems, including homogeneous and heterogeneous liquids, solids, and super-critical phases? This apparent discrepancy can be resolved by looking at the many-body expansion (MBE) described in **??**. From this expansion, we see that *any* system can be modeled as a sum of two- and many-body interactions, with the two-body interactions plus N-body polarization (an energy term which we obtain automatically in Section 5.5.5) accounting for upwards of 90—95% of the total N-body energy.[4] Consequently, and regardless of whether we are ultimately interested in studying a homogeneous liquid or a heterogeneous supercritical phase, for ab initio force field development it's critical that we develop and parameterize accurate models for all two-body interactions. Thus in practice, our focus is often on developing new and improved force fields for gas-phase dimer interactions, always with the goal of using the MBE to run simulations on any N-body system of interest.

† In general, force field development based on homo-monomeric interactions involves the fewest atomtypes (and thus the fewest number of free parameters!), and is to be preferred. On the other hand, hetero-monomeric-based force field development can yield the best accuracy for studying specific systems where either transferability is difficult (see Chapter 4 for an example) or where computational expense is an issue. (Running large-basis-set SAPT calculations on the napthalene dimer, as an example, is currently infeasible, whereas benchmark calculations on napthalene-Ar interactions are affordable.)

I) Generate benchmark two-body energies

   1) Generate a series of well-chosen dimer configurations (see Section 5.2)
   2) Calculate DFT-SAPT benchmark energies for all dimer configurations from the previous step (see Section 5.3)
   3) Optionally (depending on system size and the accuracy of DFT-SAPT for the chosen system), calculate CCSD(T) or CCSD(T)-f12 benchmark energies in order to correct the DFT-SAPT energies above (see Section 5.4)

II) Parameterize the two-body PES

   1) For each unique monomer, obtain the following monomer-specific parameters:
      i. Multipole moments, Q (see Section 5.5.2)
      ii. ISA Exponents, B (see Section 5.5.3)
      iii. Dispersion Coefficients, C_n (see Section 5.5.4)
      iv. Induced Dipole Polarizabilities, α (see Section 5.5.5)
   2) Obtain all remaining force field parameters by fitting a chosen force field functional form to the two-body benchmark energies from Step I) (see Chapter 6)

The entire force field development process has been made reasonably 'black-box', and can be carried out via a handful of input files and easy-to-use run scripts. This semi-automated workflow for SAPT-based force field development is available for download at `https://github.com/mvanvleet/workflow-for-force-fields`, and should be sufficient for most routine force field development. Installation and usage instructions are included on the website, and are also reprinted in Fig. A.1 for conveninece. The remainder of this Chapter is designed to give new users a sense of the theory and practice involved in using the workflow, and we now turn to an in-depth discussion of each step.

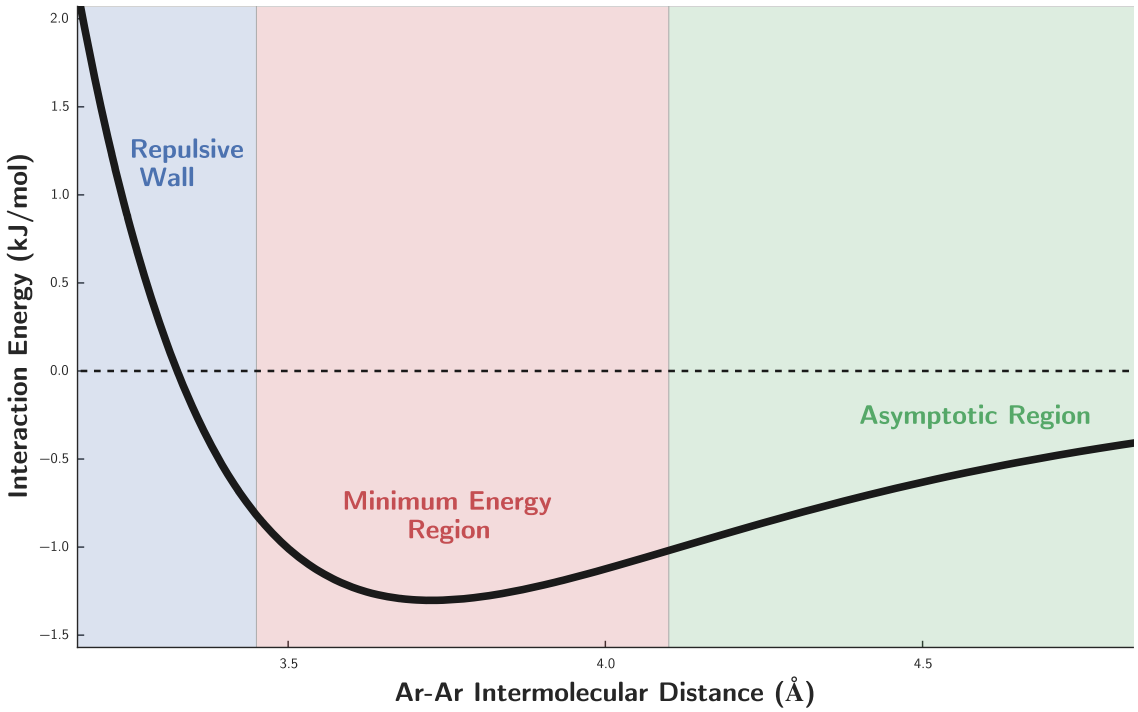


Figure 5.1: Generalized form of a PES showing the repulsive wall, minimum energy, and asymptotic regions of the argon dimer. Cutoffs between the different regions should be taken qualitatively.

5.2 Geometry Generation

5.2.1 Guiding Principles

For any given monomer(s) of interest, the first step in the force field development process is to choose a series of optimal dimer configurations. This 'optimal' set is highly dependent on the type of force field that is being fit, and indeed the recommendations offered below are specific to the SAPT-based force fields described in Chapters 2 and 3.

In general, and as shown in Fig. 5.1, a given PES will have (qualitatively) three different regions: a repulsive wall, a minimum energy region, and an asymptotic region. (Energies in the asymptotic region are usually attractive, but are sometimes

repulsive due to unfavorable electrostatic interactions). Based on the principles of statistical mechanics and the Boltzmann distribution, we know that (for a system at constant temperature T) the probability P_i of observing a system in state i is exponentially-dependent on the energy of that state:[299]

$$P_i \propto exp(-E_i/k_B T) \tag{5.1}$$

where k_B is the Boltzmann constant.

Due to the exponential relationship between the energetic stability of a given state and the probability of experimentally observing said state, *routine molecular simulation will most frequently sample dimer configurations near the minimum energy and asymptotic regions of the potential*.* Consequently, these two portions of the PES are the most important to accurately model with a force field. Nevertheless, and as discussed in Section 5.5, the asymptotic region of the PES primarily depends on functionals forms whose parameters are calculated from monomer properties, making the dimer-based fits described in Chapter 6 relatively insensitive to inclusion of configurations in this region.† By contrast, the force field parameters we directly fit to the dimer PES are primarily sensitive to the shape and location of the repulsive wall and (to a lesser extent) the minimum energy regions. Consequently, and based on the combination of their observation probability in molecular simulation as well as their importance in the improving the force field fitting process, *dimer configurations along the repulsive wall and (even more importantly) in the minimum energy region*

* Although it's straightforward to see why the minimum energy region gets sampled in molecular simulation, the importance of the asymptotic region may be hard to understand simply by looking at Eq. (5.1) and the 2-body PES shown in Fig. 5.1. We should recognize, however, that the two-body energy of an N-body system is determined both by nearest neighbor interactions (whose configurations are typically in the minimum energy region of the 2-body PES) and by the more distant, non-nearest neighbor interactions (whose configurations lie in the asymptotic region). The number of non-nearest neighbor interactions outweigh the closer-contact interactions, which in turn makes *both* the minimum energy and asymptotic regions of the potential important to correctly model.

†By contrast, other functional forms (e.g. Lennard-Jones) *do* have parameters that effect the asymptotic region, and for these force fields it would be important to include this region in the parameterization process.

*are the most important to parameterize in order to develop highly-accurate force fields.**

In practice, a standard procedure for optimally sampling across the PES has been established for the Slater-ISA FF and MASTIFF force fields. Though use of different functional forms might require a different relative sampling of the dimer PES, the next sections completely outline the theory and practical calculations that are involved in generating dimer configurations for the development of Slater-ISA FF and MASTIFF force fields.

5.2.2 Theory

Assuming rigid monomer geometries, a dimer configuration can be completely determined (without loss of generalization) by fixing the center of the first monomer at the origin and by placing the second monomer according to six variables. r, θ, and ϕ determine the position of the center of the second monomer, and the three-dimensional variable Ω determines the relative orientation of this second monomer about its center. In practice, Ω is most easily described by a quaternion, and the interested reader is referred to Ref. 300 for details.

For both the Slater-ISA FF and MASTIFF fits, dimer configurations are sampled psuedo-randomly using Shoemake's algorithm,[183] which ensures even sampling of the dimer configurations. Additionally, and in order to achieve a proper balance of sampling between the repulsive wall and minimum energy regions, the following dimer configurations are excluded from sampling:

1. Configurations with any atom-atom contact distance $r_{ij} \leqslant 0.8 \times (r_i^{vdw} + r_j^{vdw})$, where r_{ij} is the contact distance and r^{vdw} is the tabulated van der Waals radius for a given element

2. Configurations with all atom-atom contact distance $r_{ij} \geqslant 1.3 \times (r_i^{vdw} + r_j^{vdw})$

*Historical note: For force field functional forms which poorly model the repulsive wall (e.g. Lennard-Jones force fields or the Born-Mayer-IP FF described in Chapter 2), the force field fit quality strongly depends on the relative representation of repulsive and attractive dimer configurations, and including either too few or too many repulsive configurations can be problematic. Only with the advent of Slater-ISA FF and MASTIFF is the fit quality strictly improved by including repulsive configurations.

A working code for this sampling algorithm has been developed and is described below.

5.2.3 Practicals

In practice, generation of the dimer configurations is fairly straightforward. The three required input files – `input/dimer_info.dat`, `input/generate_grid_settings.inp`, and `input/<mona>_<monb>.inp` – are listed in Listings 5.1 to 5.3 and using the pyridine dimer as an example. (Here and throughout, we use angle brackets to indicate required arguments.) Each input file may need to be modified for the specific dimer under consideration, and comments within these input files explain any necessary system-specific changes.

Once all required input files have been created/modified, the geometry generation process can be carried out very simply from the main workflow directory by executing the command

```
./scripts/make_geometries.sh
```

5.3 SAPT Benchmarks

After geometry generation, the next step in the Workflow is to run benchmark DFT-SAPT calculations on all dimer configurations. For a detailed analysis of SAPT, and DFT-SAPT in particular, the reader is referred to Refs. 79–81. DFT-SAPT calculations can be performed in a fairly black-box manner using the Molpro software, though the following points are worth note:

1. For best accuracy, and as described in Ref. 166, monomer DFT calculations need to be asymptotically-corrected (AC) in order to achieve best accuracy. This asymptotic correction is computed as the difference between the HOMO and the vertical ionization potential for each monomer, and can be calculated automatically by running the command

   ```
   ./scripts/submit_ip_calcs.py
   ```

(The calculation takes only a few minutes for small molecules, but may take longer for larger systems.) Importantly, the HOMO calculation should be computed using the same basis set as the DFT-SAPT calculations themselves.

2. Accurate SAPT dispersion energies generally require use of midbond functions, as described in Ref. 166. Locations for the midbond functions can be specified in the `dimer_info.dat` file. For most small molecules (such as those described in Chapter 2), it is often sufficient to place a single midbond at the midpoint between each monomer's center of mass. For larger molecules, additional midbonds (especially ones near close-contact interaction sites) may be required.

3. The included workflow assumes an aVTZ+m basis set (where the +m represents the midbond functions). This is generally of sufficient accuracy for most systems, though an aVQZ+m basis set should be used when possible to ensure convergence of the DFT-SAPT dispersion energies.

Once the appropriate midbond functions have been added to the `dimer_info.dat` input file, and the AC calculations have finished, the DFT-SAPT input files can be generated by executing the command

```
./scripts/make_sapt_ifiles.py
```

The resulting input files can then be run using the Molpro software, either in serial or in parallel. *Care should be taken to ensure that multiple calculations do not end up on the same compute note, as this can often result in i/o caching issues and reduced computational efficiency.*

5.4 CCSD(T) Calculations

When affordable, CCSD(T) calculations should be run on (at least a subset of) the dimer configurations, both in order to benchmark the DFT-SAPT energies and to provide a δCCSD(T) correction for later fitting of the SAPT potential. Recently, an explicitly-correlated CCSD(T)-f12 method has been proposed, which for practical

purposes is identical to CCSD(T) but with faster basis set convergence.[264] Usually CCSD(T)-f12a/aVTZ+m is an excellent approximation of the CCSD(T)/CBS limit. The input files for CCSD(T)-f12/aVTZ+m calculations can be set up by executing the command

```
./scripts/make_ccsdt_ifiles.py
```

and by running each input file using the Molpro software package.

5.5 Monomer-Based Parameterization

While/after running the DFT-SAPT calculations, the next step in the Workflow is to compute various force field parameters which only depend on the identities of the individual monomers themselves. The following subsections describe the calculations of multipole moments (Section 5.5.2), short-range exponents (Section 5.5.3), dispersion coefficients (Section 5.5.4), and induced dipole polarizabilities (Section 5.5.5). First, however, we outline the scope and useful features of the CamCASP software used to perform these monomer property calculations.

5.5.1 Distributed Property Calculations using CamCASP

CamCASP is a collection of scripts and programs useful for (among other things) the calculation of distributed multipoles and polarizabilities.[84] Of particular importance is the choice of distribution method, as this determines how the various molecular properties of interest should be mapped onto corresponding atom-in-molecule properties. Currently, two main distribution (or 'charge partitioning') schemes are available in CamCASP: DMA[87] and Iterated Stockholder Atoms (ISA).[91] The theory behind the ISA procedure has already been detailed in **??**, and monomer property calculations using DMA are described in Ref. 81, 87, 88. In general, and where available, ISA-based properties are to be preferred, and we recommend an ISA-based parameterization scheme for obtaining multipoles and atom-in-molecule exponents. A DMA-based method is currently required for obtaining dispersion coefficients and static polarizabilities, though ISA-based strategies for these properties

are under active development and (in the case of dispersion) are discussed in Section 5.5.4. A complete overview of available property calculations and distribution schemes, along with relevant references, is given in Table 5.1.

Property	Parameterization Scheme	
	ISA	DMA
Multipoles	Section 5.5.2 Ref. 91	– Ref. 81, 87
Exponents	Section 5.5.3 Ref. 95	–
Dispersion Coefficients	Section 5.5.4 –	Section 5.5.4 Ref. 83
Dipole Polarizabilities	–	Section 5.5.5 Ref. 83

Table 5.1: Overview of ISA- and DMA-based methods for obtaining distributed monomer properties. Details for each monomer parameterization are given in the listed section and/or reference.

5.5.2 Multipoles

Practicals

ISA-based multipoles are described in detail in Ref. 91, and can be calculated using the CamCASP software. To set-up the ISA calculations, execute the command

```
./scripts/make_isa_files.py
```

which creates the necessary ISA files for calculating both distributed multipoles and exponents (see Section 5.5.3). After running these calculations (a process which may require several hours, depending on the molecule), the multipoles can be extracted simply by running

```
./scripts/workup_isa_charges.py
```

This work-up script produces several output files,

```
<monomer>_ISA_L4.mom
<monomer>_ISA_L2.mom
<monomer>_ISA_L0.mom
```

which correspond to multipole moments for various long-range electrostatic models. Using Stone's notation,[78] the Lx suffix refers to the highest order of multipole moments (L0 = point charges, L1 = dipoles, L2 = quadrupoles, etc.) included in the model. The L4 model is output by the CamCASP software package, and the L2 and L0 models are generated by rank-truncation (that is, zeroing out) of the higher-order multipole moments. For most routine force field development, the L2 model is to be preferred for its balance of accuracy and computational expense. Next, however, we discuss situations in which different electrostatic models may be desirable.

Advanced Multipole Parameterization Options

As stated above, for the purposes of obtaining sub- kJ/mol accuracy force fields it is often important to model the long-range electrostatics using ISA-based multipoles truncated to no farther than quadrupolar (i.e. 'rank 2' or L2)[78] contributions. Due to computational and/or software limitations, however, there exist practical cases where it becomes advantageous to exclude all higher-order multipole moments.[64] In such cases, two different types of long-range electrostatic models are useful. First, for reasonably isotropic molecules a good option is to rank-truncate the ISA multipoles to the L0 point charge contributions, thus yielding a so-called 'atom-centered point charge model'. On the other hand, for more anisotropic functional groups such as those described in Ref. 216, an atom-centered point charge model can be insufficiently accurate, making it necessary to model the long-range electrostatics by including additional 'off-center/off-site' point charges. Given a well-chosen set of off-site charges, an off-center point charge model usually can reasonably reproduce the effects of the neglected higher-order multipole moments.[235] In the

past, locations for the off-center charges have usually been manually tuned or optimized in a system-specific manner, though recent work suggests the possibility of switching to non-empirical methods in order to more easily calculate/optimize positions for the extra-atom sites.[129,301]

For atom-centered point charge models, the output of the `workup_isa_charges.py` script automatically provides the required rank-truncated multipole file (listed as `<monomer>_ISA_L0.mom` in the `isa/` sub-directory). Note that, because the `<monomer>_ISA_L0.mom` file is given as a simple rank-truncation of the more complete `<monomer>_ISA_L2.mom` multipoles, the L0 moments (that is, point charges) are identical between the two files.

For developing rank-transformed point charge models, Ferenczy et al. has programmed a method for calculating electrostatic potential-fitted charges, which can be thought of as a 'rank transformation' procedure. The author's MULFIT program can be downloaded online at `http://www-stone.ch.cam.ac.uk/pub/gdma/index.php`, and documentation for the program is found in the `documentation/` sub-directory of the Workflow. Assuming the `mulfit` executable is in your `$PATH`, a basic rank transformation can be performed using the following steps:

```
cp templates/mulfit.inp isa/<monomer>/OUT/
cd isa/<monomer>/OUT/
mulfit < mulfit.inp
```

Here the default `mulfit.inp` file is set to take in the L4 rank multipoles and rank-transform them to an L0 model. In this case, note that the L0 moments between the rank-transformed and rank-truncated moments will *not* be identical, and testing is required to ascertain which moments yield optimal force field parameters.

The MULFIT program can additionally be used to develop off-site point charge models. In this case, the input multipole file (default `ISA_L4.mom`) should be edited to include the additional sites, and an example of the required syntax is given in `documentation/examples/ISA_L4_offsites.mom` for a 4-site water model. Importantly, the MULFIT program does not help optimize the position(s) of the off-site charge(s), and thus the task of choosing the number and position(s) of the off-site(s) is left to the user.

After fitting multipole parameters with the MULFIT program, the program output gives two indications of fit quality. First, the agreement between the total reference and fitted multipoles moments is listed, and this should be taken as a primary indication of multipole quality. Second, the program gives a 'Goodness of fit' parameter, expressed as an energy. While difficult to interpret in an absolute sense, in comparing different rank-transformed models we have generally found that models with lower 'Goodness of fit' parameters yield better force field fits.

5.5.3 ISA Exponents

As described in **??** and Chapter 2, the ISA procedure produces a set of distributed atom-in-molecule (AIM) electron densities. The orientational average of each of these AIM densities, or 'shape-functions', are spherically-symmetric quantities that describe the radial decay of the AIM density.[91] As described in Chapter 2, and using the algorithm detailed in Section 5.B, the shape-functions can be fit to a Slater-type function in order to yield an isotropic, exponentially-decaying model for the ISA densities. Importantly, the Slater-exponents in this density model directly yield the exponents necessary to describe short-range effects (such as exchange-repulsion and charge penetration) in the two-body force field (see Chapter 2 for details).

Assuming the ISA calculations have already been run to obtain multipole moments (see previous section), the ISA exponents can be obtained very simply by running the command

```
./scripts/workup_isa_exponents.py
```

The resulting exponents are given in the file `isa/<monomer>.exp`, which uses a file format recognized by the POInter pre-prossessing scripts (see Chapter 6).

5.5.4 Dispersion Coefficients

Theory

Dispersion coefficients can also be determined from distributed molecular (that is, AIM) property calculations, using either an ISA- or DMA-based approach. The

method for obtaining distributed dispersion coefficients has been described in detail elsewhere for an assortment of DMA-based approaches,[78,81,83,89,90,141] and Ref. 81 in particular provides a useful summary of the different equations and molecular properties that are needed to derive the types of dispersion models used in Chapters 2 and 3. In brief, AIM dispersion energies can be obtained by integrating over distributed frequency-dependent polarizabilities for each monomer, and the interested reader is referred to Chapter 9 of Ref. 78 for complete details. Under the simplifying assumption that we can treat these frequency-dependent polarizabilities as isotropic, the dispersion energy expression is given by

$$E_{\text{disp}}^{ab} \approx -\frac{C_6^{ab}}{r_{ab}^6} - \frac{C_8^{ab}}{r_{ab}^8} - \dots \tag{5.2}$$

for each atom pair, where

$$C_6^{ab} = \frac{3}{\pi}\int_0^\infty \bar{\alpha}_1^a(i\omega)\bar{\alpha}_1^b(i\omega)d\omega, \tag{5.3}$$

$$C_8^{ab} = \frac{15}{2\pi}\int_0^\infty \bar{\alpha}_1^a(i\omega)\bar{\alpha}_2^b(i\omega) + \bar{\alpha}_2^a(i\omega)\bar{\alpha}_1^b(i\omega)d\omega, \tag{5.4}$$

and higher order terms are defined analagously. Here C_n^{ab} are the atom-atom dispersion coefficients, and $\bar{\alpha}_l^a$ are the rank l, isotropic, AIM frequency-dependent polarizabilities. The formalisms involved in evaluating Eqs. (5.3) and (5.4) can be somewhat involved, but for our purposes the important take-away is the understanding that the dispersion coefficients can be entirely determined by calculating the frequency-dependent polarizabilities for each atom in its molecular environment.

Although it is straightforward to calculate *molecular* frequency-dependent polarizabilities, a central difficulty in obtaining transferable dispersion coefficients is that, in order to evaluate Eqs. (5.3) and (5.4), we must have some physically-meaningful method for calculating *atom-in-molecule* polarizabilities. Many distribution strate-

gies exist in the literature, and here we focus on two such techniques. First, and as we have used in Chapters 2 and 3, one can utilize a DMA-based approach to partition the polarizabilities into AIM contributions. In this case, and due to deficiencies in the DMA partitioning scheme, the resulting atomic polarizabilities are not always positive-definite and monotonically-decaying, and this unphysical behavior can lead to a breakdown in transferable parameterization.[89] To correct for this undesirable behavior, McDaniel and Schmidt have proposed a constrained fitting procedure whereby atomic polarizabilities can be optimized in an iterative fashion, thereby generating transferable atomic polarizabilities at the expense of requiring a fairly large training set for each unparameterized atomtype (see Section 5.5.4 for details).

As an alternative to the above iterative polarization partitioning scheme, recently Misquitta has developed an ISA-based partitioning scheme to extract the atomic frequency-dependent polarizabilities. While this approach requires further testing, and is not yet published, the resulting 'ISA-pol' method appears to lead to a more physically-meaningful partitioning of the molecular polarizabilities. For practical purposes, this more physical partitioning enables us to determine transferable dispersion coefficients without resorting to large training sets. Formalisms and technical details related to ISA-pol are the subject of Section 5.5.4, and a comparison between the two methods for obtaining dispersion coefficients is given in Section 5.5.4. Finally, each method for obtaining dispersion coefficients requires a small amount of post-processing, and this is discussed in Section 5.5.4.

Iterative-DMA-pol

Theory As described in Ref. 81, the iterative-DMA-pol (iDMA-pol) method of McDaniel and Schmidt performs a constrained optimization of atomtype-specific frequency dependent polarizabilities by fitting all polarizabilities to reproduce the so-called 'point-to-point response', α_{PQ}. This point-to-point response is a molecular quantity that describes the change in electrostatic potential at point P due to an induced change in the electron density of a molecule caused by a point charge

perturbation q_Q at point Q. For an isotropic polarizability model,

$$\alpha_{PQ} = -q_Q \sum_{a,lm} T^{Pa}_{0,lm} \bar{\alpha}^a_l T^{aQ}_{lm,0} \tag{5.5}$$

where the T are the spherical harmonic interaction functions described above and in Ref. 78. Aside from the isotropic polarizabilities $\bar{\alpha}^a_l$, all quantities in Eq. (5.5) are directly calculated in CamCASP, enabling us to fit the isotropic polarizabilities on the basis of CamCASP property calculations (see Appendix A of Ref. 81 for details).

Practicals Using the iDMA-pol method in the Workflow has two software dependencies:

1. The iDMA-pol fiting program itself, which can be downloaded at `https://github.com/mvanvleet/p2p-fitting`. Three executables (`main_dispersion`, `main_drude`, and `localize.sh`) need to be added to your bash `$PATH` for the scripts listed in this section to work properly.

2. CamCASP, which can be downloaded from `http://www-stone.ch.cam.ac.uk/programs/camcasp.html`. CamCASP also requires several environment variables to be added to your bash `$PATH`, and some of these environment variables are also used by the iDMA-pol fitting program.

and requires two additional input files:

1. `input/<monomer>.atomtypes`: The iDMA-pol fitting program performs a constrained optimization whereby the $\bar{\alpha}^a_l$ are set to be identical for atoms with the same atomtype. Consequently, the `<monomer>.atomtypes` input file is required to specify the atomtypes in each monomer. This .atomtypes file has the same format as an .xyz file, with the exception that the element names for each atom are replaced with a user-defined atomtype. See Listing 5.4 for an example with pyridine.

2. `templates/dispersion_base_constraints.index`: As described below, with iDMA-pol it is usually advisable to only fit one or two atomtype polarizabilities at a time, with the remaining atomtype polarizabilities read in as hard constraints. The `dispersion_base_constraints.index` file lists these hard constraints in a block format,

```
CT
1
 7.14483224 7.11095841 6.87452508 6.19718464 4.87589777
 3.17818610 1.56461102 0.51670933 0.09175313 0.00367230
2
 20.26394042 20.00584110 17.66562710 14.33668329 12.03179893
 11.49156262 7.86254302 3.10936998 0.53746459 0.01774391
3
 77.37303638 73.13014787 24.68682297 −13.48390193 0.40172836
 29.76747226 34.31668916 17.88515654 3.13260459 0.10137127
```

   which lists each constrained atomtype along with 10 frequency-dependent polarizabilities for each polarizability rank (1-3). (CamCASP uses numerical integration to solve Eq. (5.3), and the 10 polarizabilities per rank correspond to the frequencies CamCASP needs to perform the numerical quadrature. See the CamCASP user manual for details.) Each polarizability block should be separated by a blank line, and the atomtypes listed in the .index file *must* match those in the .atomtypes file for any hard constraints to be successfully applied. Previously-fit atomtype polarizabilities from Ref. 83 are already included in `dispersion_base_constraints.index` so as to minimize the number of hard constraints that the user will need to add manually, and these hard constraints should be used whenever possible.

Once all required input files have been created, and assuming the IP calculations from Section 5.3 have already been performed, the CamCASP calculations necessary to run the iDMA-pol program can be performed by executing the command

```
./scripts/make_dmapol_files.py
```

and running the resulting input files through the CamCASP software (a process which can take several hours). Once the CamCASP calculations finish, dispersion coefficients can be obtained by running the following work-up script:

```
./scripts/workup_dispersion_files.sh
```

The resulting dispersion coefficients will be listed in the `dispersion/<monomer>.cncoeffs` output file.

When generating dispersion coefficients using iDMA-pol, the following sanity-checks should always be performed:

1. The `<monomer>_fit_dispersion.out` file lists the number and names of unconstrained atomtypes. Ensure that the number and type of unconstrained atomtypes match your expectations, and that the number of fit atomtypes is kept relatively small (1-2 max). If you need to fit multiple atomtypes simultaneously, or you obtain unphysical disperion coefficients (see next point), you'll likely need to utilize the iterative fitting algorithm outlined in Ref. 83 or obtain dispersion coefficients from an ISA-based scheme (Section 5.5.4).*

2. Dispersion coefficients should always be positive. Any negative dispersion coefficients are likely a sign of unphysical atomic polarizabilities (see next point).

3. Phsyically-speaking, the atomic polarizabilities at each rank should be positive definite, and monotomically-decreasing.[78,89] Unphysical behavior (especially at rank 3) is sometimes unavoidable, but often indicates poor fit quality and can lead to inaccurate and/or non-transferable dispersion coefficients. Always check the output `.casimir` files for the physicality (positive-definiteness and monatomic-decrease) of the frequency-dependent polarizabilities for each atomtype and each rank.

Finally, given a set of physical atomic polarizabilities and dispersion coefficients, dispersion coefficients from the iDMA-pol method can be worked-up using the post-processing scripts described in Section 5.5.4.

*Scripts to perform the iterative iDMA-pol fitting algorithm can be made available upon request.

ISA-pol

Theory Rather than iteratively fitting polarizabilities to reproduce the point-to-point response, with ISA it is possible to compute the atomic polarizabilities directly. First, note that the frequency-dependent, molecular polarizabilities are given by the following formula:

$$\alpha_{lm,l'm'}(\omega) = \int\int \hat{Q}_{lm}(\mathbf{r})\alpha(\mathbf{r},\mathbf{r'}|\omega)\hat{Q}_{l'm'}(\mathbf{r'})d\mathbf{r}d\mathbf{r'} \tag{5.6}$$

Here $\hat{Q}$ are the regular spherical harmonic operators (defined in Appendix A of Ref. 78) of rank l and order m, and $\alpha(\mathbf{r},\mathbf{r'}|\omega)$ is the frequency-dependent density susceptibility (FDDS), or charge density susceptibility, which measures the change in charge density at $\mathbf{r'}$ that results from a delta-function change in the electric potential at point $\mathbf{r}$. From **??**, we have that

$$1 = \sum_a \left(\frac{\bar{w}^a(\mathbf{r})}{\sum_m \bar{w}^m(\mathbf{r})} \right) = \sum_a \bar{p}_a(\mathbf{r}), \tag{5.7}$$

where the bars indicate that we have normalized the atom-in-molecule densities and weight functions. Substituting this equation into Eq. (5.6), we arrive at an ISA-based definition of the AIM polarizabilities:

$$\begin{aligned} \alpha_{lm,l'm'}(\omega) &= \int\int \hat{Q}_{lm}(\mathbf{r})\alpha(\mathbf{r},\mathbf{r'}|\omega)\hat{Q}_{l'm'}(\mathbf{r'})d\mathbf{r}d\mathbf{r'} \\ &= \sum_a \sum_b \int\int \hat{Q}_{lm}(\mathbf{r})p_a(\mathbf{r})\alpha(\mathbf{r},\mathbf{r'}|\omega)p_b(\mathbf{r'})\hat{Q}_{l'm'}(\mathbf{r'})d\mathbf{r}d\mathbf{r'} \\ &\equiv \sum_a \sum_b \alpha^{ab}_{lm,l'm'}(\omega) \end{aligned} \tag{5.8}$$

While this formula bears similarity to DMA-based polarization approaches,[89,90] the advantage of Eq. (5.8) is that the AIM polarizabilities are defined in a physically-meaningful and transferable manner. Consequently, with little refinement these ISA-based polarizabilities (ISA-pol) can be used to directly obtain transferable

dispersion coefficients for individual atom-in-molecule, all without recourse to the iterative fitting process required in Section 5.5.4.

Practicals The ISA-pol method has been completely implemented as of CamCASP-6.0, though the input scripts are (as of this writing) still in beta. Consult the CamCASP user manual or contact Alston Misquitta for up-to-date details and required input files.

Comparison between iDMA-pol and ISA-pol

Preliminary results for the ISA-pol method, tested on the 91 dimer test set from Chapter 2, appear to be of similar accuracy compared to the iDMA-pol method, though both methods appear to have their own strengths and weaknesses when it comes to obtaining dispersion coefficients for different atomtypes. A comparison between the two different methods is given in Table 5.2. Overall, ISA-pol appears to give more physically-meaningful atomic polarizabilities, whereas an isotropic iDMA-pol description is (for anisotropic systems) sometimes a better 'effectively anisotropic' model.*

Dispersion Coefficient Post-processing

Regardless of which distribution method is used, some post-processing is needed to transform the ISA-pol/iDMA-pol coefficients into optimal dispersion force field parameters. In particular, while the DFT-SAPT energies from Molpro and CamCASP should agree, in practice the different software packages use different kernels (ALDA+LHF and ALDA+CHF, respectively) to calculate the linear response func-

* A main difference between the iDMA-pol and ISA-pol coefficients is that iDMA-pol fits more strongly to the point-to-point (p2p) response function, whereas ISA-pol coefficients are set to the values calculated as in Section 5.5.4. Consequently, iDMA-pol is able to perform better as an 'effectively anisotropic' model. In principle, changing the defaults in CamCASP to use weight type 4 (which uses dipole-dipole terms as anchors, but completely fits higher ranking terms and thus fits the p2p better) or 3 (uses all terms as anchors) and a weight coefficient of 1e-5 (rather than 1e-3) should yield dispersion coefficients more similar to iDMA-pol, though this idea requires further testing.

iDMA-pol	ISA-pol
Ease of Parameterization	
• For systems with a single (or possibly two) unparameterized atomtype(s), straightforward to parameterize new atomtypes • For systems requiring dispersion coefficients for several unparameterized atomtypes, requires a library of systems containing these atomtypes, and an iterative procedure to fit the new atomtypes	• Straightforward for all molecules, regardless of number of unparameterized atomtypes
Physicality of the Distributed Polarizabilities	
• Polarizabilities tend to be positive-definite and monotonically-decaying at low rank, but not always for rank 3 • Physicality is highly-dependent on the quality of previously parameterized atomtypes	• With few exceptions, polarizabilities are positive-definite and monotonically-decaying at all ranks
Accuracy of the Dispersion Coefficients	
• Good to excellent accuracy for atomtypes which have been fit to a reproduce large library of molecular systems • Fair accuracy for certain atomtypes (such as chlorine or bromine) not parameterized to an extensive library • For anisotropic systems (such as CO_2), tends to give a better isotropic description than ISA-pol – we hypothesize that this is a result of directly fitting the point-to-point response, leading to an 'effectively-anisotropic' model	• Good to very good accuracy for all tested systems, regardless of what atomtypes are represented • Isotropic dispersion coefficients tend to give worse accuracy for anisotropic systems compared to iDMA-pol, whereas anisotropic dispersion models (see Chapter 3 based on ISA-pol are of similar accuracy to the iDMA-pol method

Table 5.2: Comparison between the iDMA-pol and ISA-pol methods.

tions. Consequently, this means that the dispersion coefficients calculated in CamCASP are intended to reproduce the CamCASP-calculated DFT-SAPT dispersion energies, but may only be approximately accurate for Molpro-calculated DFT-SAPT dispersion energies.* In practice, the CamCASP-calculated dispersion coefficients slightly underestimate the Molpro dispersion energies, and the coefficients need to be scaled (usually by a factor of 1.03-1.10, depending on the atomtype) to reproduce the Molpro energies. This scaling can be carried out by executing the command

```
./scripts/get_scaled_dispersion.py <scale_factor>
```

where `<scale_factor>` is chosen to reproduce the asymptotic Molpro DFT-SAPT energies (see Chapter 6 for details). This choice may require some testing, but 1.10 is usually a good default. The above script outputs files `dispersion/<monomer>.disp`, which can be used as input to the POInter program discussed in Chapter 6.

5.5.5 Polarization Charges

Theory In addition to frequency-dependent polarizabilities, some of the same techniques described in Section 5.5.4 can be applied to obtain the static polarizabilities that get used in modeling the SAPT induction energy. Though in principle ISA-based polarizabilities could be used, this technique has not yet been developed. Instead, an iDMA-pol-type procedure can be used to extract the necessary polarization parameters. The algorithms used to perform this procedure are described in Appendix A of Ref. 81. Due to the reduced number of coefficients that need to be fit, this optimization is generally more robust, and leads to more transferable dispersion parameters, than do the algorithms described in Section 5.5.4.

* Additional reasons for discrepancies between CamCASP and MOLPRO dispersion coefficients include the following:

1. For PBE calculations, CamCASP uses ALDA with PW91c correlation, whereas Molpro uses VWN
2. CamCASP writes kernels completely in the auxiliary basis set, whereas Molpro writes the kernel in a variety of basis sets

Practicals The drude oscillator fitting code has the same dependencies and input files as iDMA-pol, with the exception that the `dispersion_base_constraints.index` file is replaced with the following constraint file:

1. `drude_base_constraints.index`: As with iDMA-pol, it is usually advisable to only fit a few atomtype static polarizabilities at a time, with the remaining atomtype polarizabilities read in as hard constraints. The `drude_base_constraints.index` file lists these hard constraints in a block format,

```
C
1
 0.0

N
1
 -11.7529643

H
1
 -1.254
```

   which lists each constrained atomtype along with the rank 1 static polarizabilities. Each block should be separated by a blank line. Unlike with the generation of dispersion coefficients, an initial guess must be given for *all* atomtypes in the `<monomer>.atomtypes` input file. The format for the `drude_base_constraints.index` is such that positive polarizabilities correspond to these initial guesses, whereas zero or negative entries for the polarizabilities indicate that the atomtype should be treated as a hard constraint. Previously-fit atomtype polarizabilities from Ref. 83 are already included in `drude_base_constraints.index` so as to minimize the number of hard constraints that the user will need to add manually, and these hard constraints should be used whenever possible.

Assuming that the iDMA-pol calculations have already been run in CamCASP, the drude oscillator coefficients can be obtained simply by executing

```
./scripts/workup_drude_files.sh
```

As with the dispersion coefficients, care should be taken to ensure that the resulting drude oscillator charges are physically-meaningful (i.e. negative).

5.6 Dimer-Based Parameterization

After obtaining monomer parameters for a given system of interest, the final remaining task is to fit the remaining force field parameters to reproduce the DFT-SAPT calculations performed earlier in the Workflow. Dimer-based parameterization is carried out by the POInter program, which will be the subject of the next chapter. The Workflow is useful for preparing input files for this dimer-based parameterization, as follows:

```
./scripts/workup_sapt_energies.py
./scripts/gather_pointer_input_files.py
```

For modeling off-site point charges, the following additional steps are required (assuming the offites .xyz file has already been added to the input directory):

```
cd geometries
mkdir xyz
mv * xyz
cd ../
./scripts/get_sapt_file_wi_offsites.py
```

The output of these scripts will generate a .sapt file (containing results from the DFT-SAPT calculations, with atomtype labels taken from each `input/<monomer>.atomtypes` file) and a new directory, `ff_fitting`, which automatically sets up all of the input files and monomer parameters needed to easily run the POInter fitting code. The theory and practice of the POInter fitting code is the subject of the next chapter, however in practice the software can be run very simply by modifying the required input files (see Section 6.3 for details) and running

```
cd ff_fitting
(modify input scripts)
./run_pointer.py
```

5.A Input Scripts

In total, the workflow for force field development requires four input files, as follows:

Listing 5.1: generate_grid_settings.inp

```
# Generate Grid Settings file. Version 04.28.15
#
# General Scan Parameters:
n_points          1000  # Number of grid points (.xyz files) to output
geometry_file     pyridine_pyridine.inp #name of geometry file
output_name       pyridine_pyridine         #output file base name

# Hard Sphere cutoff parameters:
#
# Parameters below are used to define minimum and maximum acceptable distances
# for neighbor-neighbor interactions.  'cutoff_type' can either be set to
# 'absolute' or 'vdw'. In the former case, the hard sphere cutoff will be set
# to the absolute distances (in Angstroms) given by cutoff_min and cutoff_max,
# respectively. In the latter case, the hard sphere cutoff will be set to a
# fraction of the Van der Waals distance between two atoms.
cutoff_type       vdw   # either vdw or absolute
cutoff_min          0.8    # a positive float (ex. 0.8 for vdw or 2.0 for absolute)
cutoff_max          1.3   # a positive float (ex. 1.2 for vdw or 6.0 for absolute)

# The following are parameters defining the centers of monomer's a and b as well as the
    scan
# vector.
#
# The 'center' of each monomer is defined by default to be each
# monomer's center of mass, but can also be set to be either the center of an
# atom or a point in 3-space (relative to monomer coordinates given in input
# geometry file).
mona_origin_type  1   # choose 0 for center of mass (COM), 1 for atom#, and 2 for a
    specific point
mona_origin       6     # (either 'COM', point x,y,z , or atom# in monomer (indexing
    starts at 1), depending on choice of mona_origin_type above)
monb_origin_type  1   # choose 0 for COM, 1 for atom#, and 2 for a specific point
monb_origin       6    # (either 'COM', point x,y,z , or atom# in monomer (indexing
    starts at 1), depending on choice of mona_origin_type above)

# The scan vector should be a vector (given relative to the coordinates in
# monomer a) that defines the direction of internuclear seperation between the
# two monomers. It can either be given as a 3-membered list or by listing two
# monomer indices (scan vector will point from atom1 to atom2, indexing starts at 1).
```

```
scan_vector_type   0  # choose 0 for monomer indices, 1 for a specific point
scan_vector     9,6   # Give either as a 2 (if scan_vector_type==0) or a 3 (if
    scan_vector_type==1) membered,
                      # comma seperated list without spaces, i.e. '1.0,2.7,4.2' (no
                          quotes)

# Set bounds on moving the center of monomer b relative to the center of
# monomer a. min/max_r refers to the distance between the centers, while theta
# and phi correspond to the azimuthal and polar angles, respectively, of
# rotation about the vector scan_vector (given above).
#
# Give min/max angles as either integers/floats in terms of pi (i.e. setting
# 'max_theta 2' (no quotes) will yield  max_theta=2pi).
min_r               2.0
max_r               8.0
min_theta           0
max_theta           2
min_phi             0
max_phi             1
```

Listing 5.2: dimer_info.dat

```
############################
# DIMER INFORMATION FILE #
############################

# String names for monomers A and B:
# ----------------------------------
MonA_Name          pyridine
MonB_Name          pyridine

# Charges for monomers A and B:
# -----------------------------
MonA_Charge            0
MonB_Charge            0

# Midbond position(s); two integers indicating atom indices (indexed from 1)
# on monomers A and B, respectively, between which to place the midbond site.
# In lieu of an integer, COM can also be used to indicate the center of mass
# of the monomer.
# Multiple arguments can be given to produce multiple midbond functions.
# -----------------------------------------------------------------------------
midbond      com     com
```

Listing 5.3: pyridine_pyridine.inp

```
Pyridine Dimer; Optimzed with PBE0/cc-pVTZ Gaussian03 by AJ Misquitta
11
 H          -2.050322     1.274414     0.000000
 H          -2.147113    -1.203259     0.000000
 H           0.000000    -2.487558     0.000000
 H           2.147113    -1.203259     0.000000
 H           2.050322     1.274414     0.000000
 N           0.000000     1.382844     0.000000
 C          -1.134410     0.690452     0.000000
 C          -1.190513    -0.695795     0.000000
 C           0.000000    -1.403912     0.000000
 C           1.190513    -0.695795     0.000000
 C           1.134410     0.690452     0.000000
11
 H          -2.050322     1.274414     0.000000
 H          -2.147113    -1.203259     0.000000
 H           0.000000    -2.487558     0.000000
 H           2.147113    -1.203259     0.000000
 H           2.050322     1.274414     0.000000
 N           0.000000     1.382844     0.000000
 C          -1.134410     0.690452     0.000000
 C          -1.190513    -0.695795     0.000000
 C           0.000000    -1.403912     0.000000
 C           1.190513    -0.695795     0.000000
 C           1.134410     0.690452     0.000000
```

Listing 5.4: pyridine.atomtypes

```
11
pyridine; global coordinates
HM      -2.05032200      -0.00000000      -0.10843000
HM      -2.14711300      -0.00000000      -2.58610300
HM       0.00000000      -0.00000000      -3.87040200
HM       2.14711300      -0.00000000      -2.58610300
HM       2.05032200      -0.00000000      -0.10843000
N        0.00000000       0.00000000       0.00000000
CM      -1.13441000      -0.00000000      -0.69239200
CM      -1.19051300      -0.00000000      -2.07863900
CM       0.00000000      -0.00000000      -2.78675600
CM       1.19051300      -0.00000000      -2.07863900
CM       1.13441000      -0.00000000      -0.69239200
```

5.B Algorithm for Obtaining ISA Exponents

Unphysical asymptotic charge density decays occasionally arise in the ISA procedure due to basis set incompleteness and numerical instabilities. These unphysical decays can skew optimization of ISA-based exponents, B_i^{ISA}, and need to be corrected. Generally speaking, there exists some range of distances in the valence region that *does* exhibit the expected exponential decay; we extrapolate the decay from this intermediate region to describe the asymptotic region using the following algorithm:

1. Take the log of each atomic density (henceforth logdens) to linearize the asymptotic density.
2. Compute the 2^{nd} derivative of logdens. This can be done analytically, as the ISA procedure outputs an analytical expression (in terms of Gaussian basis functions) for the atomic density.
3. Determine the 'intermediate region' of exponential decay by locating the largest range where the 2^{nd} derivative of logdens is zero to within a fixed tolerance. Here we utilize a tolerance of 0.3 a.u. (absolute cutoff) or 190% of the smallest exponent in the Gaussian basis set (relative cutoff), whichever is smaller. The latter cutoff accounts for the eventual asymptotic Gaussian-type decay dictated by the smallest ζ in the ISA basis. The endpoints of this intermediate region are denoted $r1$ and $r2$, respectively.
4. Calculate the slope m and intercept b for the line defined by $r1$, $r2$, and their respective values of logdens.
5. Replace all values of logdens after $r2$ with $mr + b$. The resulting atomic density is labeled in the main text as 'Asymptotically-corrected ISA densities'.

A visual of these steps is shown in 5.2.

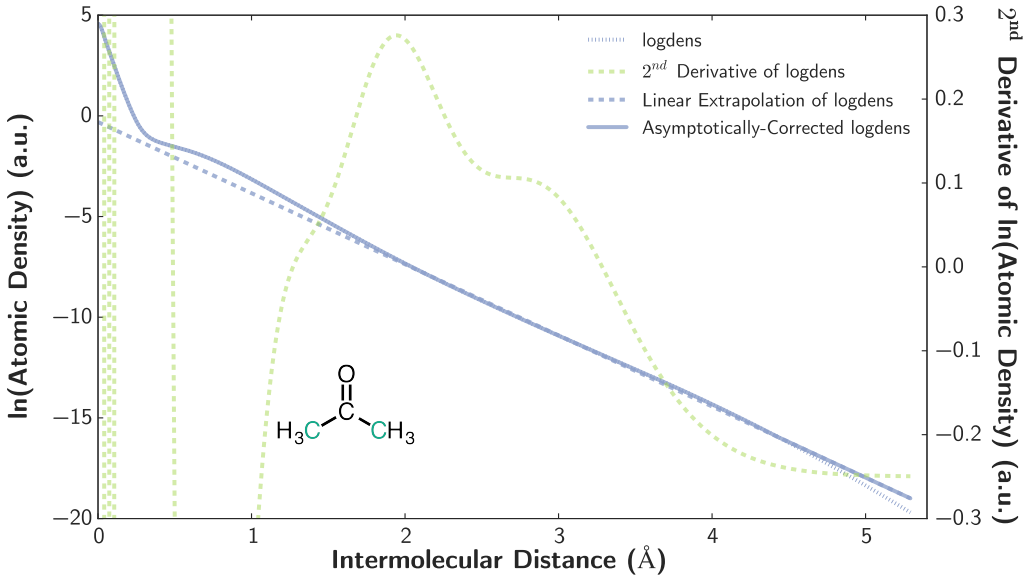


Figure 5.2: Linear extrapolation algorithm for the methyl carbon in acetone. Depicted are (in legend order) Steps 1, 2, 4, and 5 in the extrapolation algorithm. Note that some portions of the 2^{nd} derivative extend off the graph; also note that most of logdens is located underneath the asymptotically-corrected curve.

6 FORCE FIELD DEVELOPMENT FOR TWO-BODY SYSTEMS: PRINCIPLES AND PRACTICES

6.1 Overview

> The challenge for current and future work on force-field development is to improve the form of the potential energy function, to improve the methodology used in determining the potential energy parameters, and to use these advances to generate improved potential energy functions. […] That these force fields differ substantially in form and in manner of derivation serves to emphasize that force field development is still as much a matter of art as of science. Someday, consensus on the form and manner of parameterization of molecular force fields may exist, but for now much remains to be learned.
>
> — TA Halgren, 1995, adapted from Ref. 302

More than twenty years later, Halgren's perspective on biomolecular force field development remains suprisingly prescient, and the same challenges felt by early force field developers continue into the present day. The current scope of force field development is vastly complex, and Halgren's dream of 'consensus' in functional form and parameterization methodologies has yet to be realized, especially in comparing the fundamentally different approaches used in parameterization of either empirical or ab initio force fields. Despite these differences, real progress has been made to improve and standardize force field development within certain *categories* (empirical, ab initio, etc.) of development methodologies.[54,75,224,303,304] With ab initio force field development, for example, promising commonalities have emerged in how scientists tend to formulate and parameterize new molecular models.[82] Explicit polarization, originally a cost-inefficient and understudied model, is now becoming commonplace in intermolecular force fields,[63] and accurate, distributed multipolar descriptions of electrostatics seem poised to become broadly employed over the next decade.[64,67,70,129,226,228,301] Systematic methods for obtaining

distributed dispersion models have been developed within the past decade, and are constantly being improved for general use in molecular simulation.[4,83,89,90,209] Even with models for short-range interactions, where there is less consensus in terms of which functional form(s) and parameterization schemes should be used, many have begun the process of including physically-meaningful terms to describe charge penetration,[60–62,99,130,244,305–308] exchange-repulsion,[95,146] charge transfer,[97,154,244,309,310] and anisotropic effects.[54,54,89,90,99,127,151,153,208,209,221,228,233,234,241–247,263]

Building on the topics discussed in Chapter 5, our goal in this Chapter, broadly speaking, is to discuss the current state of 'consensus' regarding functional forms, parameterization methods, and best practices for the limited scope of SAPT-based force field development. Here the focus will be on both the 'scientific' and 'artistic' elements of this force field development methodology, particularly as it pertains to MASTIFF (Chapter 3) and related models (Chapter 2). As discussed below, and to use Halgren's terminology, for some asepcts of the MASTIFF development methodology there is good 'consensus' as to the required functional forms and the manner in which these forms should be parameterized. In these cases of general consensus, a somewhat black-box workflow is possible for developing new MASTIFF models, and our first goal in this Chapter is to detail the ways in which a recently developed piece of software, the POInter, can automate such tasks in the course of routine force field development.

For other aspects of force field development, such as with models for multipolar electrostatics, there can remain pratical limitations which, depending on the specific application and software package used, may requite alternative, and potentially less accurate, strategies for force field development. For these areas, a second goal in this Chapter is to outline, both conceptually and using the POInter software, current best practices for force field development in the event of limitations due to computational cost or software requirements. available functional forms.

Lastly, there remain select elements of force field development (namely the induction models discussed in Section 6.4.4) for which there has not yet been established a 'consensus' or set of best practices for force field development, either in a theoretical or practical sense. Such aspects of force field development will

need to be the subject of future work (see **??**), both in our group and across the scientific community. In the meantime, our third and final goal in this Chapter is to discuss the ambiguities involved in these 'non-consensus' aspects of force field development, and we offer some practical modeling recommendations and software tools to assist in both present and future 'next-generation' force field development.

6.2 Parameterization Overview

6.2.1 Theory

For SAPT-based force fields with functional forms similar to those in Chapters 2 and 3 – MASTIFF being a prime example* – we have discussed in Chapter 5 a number of practical approaches for beginning intermolecular force field development. In particular, we have already described strategies for optimally obtaining benchmark electronic structure theory data and for calculating some of the monomer-property-based parameters that will (vida infra) be utilized in the final force field. Nevertheless, we have not yet focused on the actual process of force field fitting, nor on strategies for assessing the accuracy and transferability of the resulting functional forms and parameters. It is to these two crucial topics that we now turn.

In regards to the force field fitting process itself, we begin by asking the obvious question: "What parameters actually need to be fit in order to obtain a final force field?" To this end, we have highlighted in Fig. 6.1 all of the parameters required to completely specify the MASTIFF force field, and have grouped these parameters according to how these parameters are calculated/optimized in practice. Specifi-

* As mentioned above, our focus in this Chapter is primarily on the MASTIFF force field. Nevertheless, most of the principles and ideas presented below should pertain generally to other force fields (Born-Mayer-IP FF, Slater-ISA FF, etc.) that are fit on a term-by-term basis to reproduce a benchmark EDA, and this Chapter should also provide a helpful set of 'best practices' for fitting and analyzing these types of force fields.

$$
\begin{aligned}
A_{ij} &= A_i A_j \\
B_{ij} &= \sqrt{B_i B_j} \\
C_{ij,2n} &= \sqrt{C_{i,2n} C_{j,2n}} \\
P(B_{ij}, r_{ij}) &= \frac{1}{3}(B_{ij} r_{ij})^2 + B_{ij} r_{ij} + 1 \\
f_{2n}(x) &= 1 - e^{-x} \sum_{k=0}^{2n} \frac{(x)^k}{k!} \\
x &= B_{ij} r_{ij} - \frac{2{B_{ij}}^2 r_{ij} + 3B_{ij}}{{B_{ij}}^2 r_{ij}^2 + 3B_{ij} r_{ij} + 3} r_{ij}
\end{aligned}
\tag{6.1}
$$

$$
\begin{aligned}
V_{ij}^{exch} &= A_{ij}^{exch} P(B_{ij}, r_{ij}) \exp(-B_{ij} r_{ij}) \\
V_{ij}^{elst} &= -A_{ij}^{elst} P(B_{ij}, r_{ij}) \exp(-B_{ij} r_{ij}) + \sum_{tu} Q_t^i T_{tu} Q_u^j \\
V_{ij}^{ind} &= -A_{ij}^{ind} P(B_{ij}, r_{ij}) \exp(-B_{ij} r_{ij}) + V_{pol}^{(2)} \\
V_{ij}^{\delta^{HF}} &= -A_{ij}^{\delta^{HF}} P(B_{ij}, r_{ij}) \exp(-B_{ij} r_{ij}) + V_{pol}^{(3-\infty)} \\
V_{ij}^{disp} &= -A_{ij}^{disp} \sum_{n=3}^{6} f_{2n}(x) \frac{C_{ij,2n}}{r_{ij}^{2n}}
\end{aligned}
$$

$$
V_{FF} = \sum_{ij} V_{ij}^{exch} + V_{ij}^{elst} + V_{ij}^{ind} + V_{ij}^{\delta^{HF}} + V_{ij}^{disp} \tag{6.2}
$$

Figure 6.1: An overview of required force field parameters for the MASTIFF and/or Slater-ISA FF force fields. All relevant equations are displayed in black, and the first instance of each parameter is shown in color according to the following scheme:

- Unconstrained parameters, which must be directly fit by POInter
- Soft-constrained parameters which, depending on user-specified settings, are treated as either soft- or hard-constraints
- Hard-Constrained parameters read in by POInter which are always treated as hard constraints

See main text for details.

cally, all force field parameters in MASTIFF can be thought of in terms of one of the following three categories:

- Unconstrained parameters: These parameters have not been specified on the basis of any monomer properties calculation, and so must be directly fit to the two-body energy itself.
- Soft-Constrained parameters: These parameters *can* be fit entirely on the basis of monomer properties, however it is sometimes advantageous to further refine these parameters with respect to the benchmark two-body energies. In this case, soft constraints[99] are often applied to the fitting process to ensure that the parameters do not deviate strongly from their original values as calculated from monomer properties.
- Hard-Constrained parameters: These parameters are calculated entirely from monomer properties, and are not futher involved in the force field fitting process except as hard constraints.

To use MASTIFF (see Chapter 3) as an example, an overview of the required parameters, and the manner in which these parameters are fit, is as follows:

- A_i^{exch}, A_i^{elst}, A_i^{ind}, $A_i^{\delta^{HF}}$: The force field energy depends linearly on a number of short-range prefactors, and in practice it is fairly straightforward to directly fit each of these prefactors to the corresponding benchmark SAPT component energy. Note that, for anisotropic atomtypes, each A coefficient may in fact involve several parameters, all of which must be directly fit:

$$A_i^{exch}(\theta_i, \phi_i) = A_{i,iso}^{exch}\left(1 + \sum_{l>0,k} a_{lk}^{exch} \mathcal{C}_{lk}(\theta_i, \phi_i)\right) \tag{6.3}$$

  (Though not entirely standard notation,[78] for clarity in this Chapter we use $\mathcal{C}$ to denote the set of renormalized spherical harmonics so as to make a clear distinction between $\mathcal{C}$, the spherical harmonics, and C, the dispersion coefficients from Section 5.5.4).

- B_i: The force field energy depends non-linearly on the short-range exponents B_{ij}, making this parameter relatively difficult to optimize without constraints. Fortunately, the B_{ij} parameters can instead be calculated on the basis of monomer properties (see Section 5.5.3), and for obtaining force fields with RMSE of ~1 kJ/mol it is often sufficient to use the ISA-obtained B_{ij} parameters without further fitting. For obtaining more accurate force fields, however, and in order to account for small uncertainties in our method of obtaining ISA-derived B_{ij} parameters (see Section 5.B), we have had good success in allowing the B_{ij} parameter to vary slightly from its ISA-derived value. In practice, this entails optimizing the B_{ij} parameters with respect to the benchmark SAPT exchange energy and subject to a harmonic penalty function.[99]

- A_i^{disp}: As with the other A pre-factors, a pre-factor can be fit to the benchmark dispersion energy so as to enhance the force field accuracy with respect to a given benchmark electronic structure theory. (Vida infra, this benchmark energy can either by DFT-SAPT or CCSD(T)). Unlike with other pre-factors, however, and because we generally have good accuracy in obtaining dispersion coefficients C (see Section 5.5.4), nominally $A_i^{disp} \approx 1$ for most systems. Still, parameters must sometimes be fit to the dispersion energy due to one or both of the following reasons:

  1. For anisotropic atoms, we must model the orientational dependence of the dispersion energy, and this model requires parameters in addition to the isotropic dispersion coefficients calculated in Section 5.5.4).
  2. Uncertainties in the iDMA-pol and/or ISA-pol dispersion coefficients can sometimes lead to inaccuracies in the isotropic dispersion coefficients, and these inaccuracies can sometimes be corrected by rescaling the isotropic dispersion coefficients themselves

  In practice, when calculating A_{ij}^{disp} we often treat the *anisotropic* dispersion coefficients a_{lk}^{disp} as free parameters, and sometimes additionally optimize an *isotropic* scale factor subject to soft constraints.* In total, this leads to the

following set of parameters and equations for the dispersion energy pre-factor:

$$A_i^{disp}(\theta_i, \phi_i) = A_{i,iso}^{disp}\left(1 + \sum_{l>0,k} a_{lk}^{disp} C_{lk}(\theta_i, \phi_i)\right) \tag{6.4}$$

where the colors serve to indicate that both free and constrained parameters are contained within the pre-factor.

- Q_t^i: Multipole moments Q_t^i can be directly calculated from monomer properties using the techniques discussed in Section 5.5.2. These ISA-based multipoles are generally quite accurate, however (vida infra) when using cheaper point charge models some care must be taken to ensure that the effective model does not lead to a deterioration in force field accuracy.

- V_{pol}: As with multipole moments, polarization parameters (Section 5.5.5) are treated as hard constraints during the force field fitting process. Currently, there is not a consensus on what functional forms and/or damping parameters should be used to model the short-range polarization energy, and this topic and its associated practical issues will be the subject of Section 6.4.4.

- $C_{i,2n}$: Dispersion coefficients are calculated via the approaches discussed in Section 5.5.4, and are generally treated as hard constraints in the force field fitting process. In some cases (vida supra), these dispersion coefficients are scaled to reproduce the SAPT energies, however we discuss in Section 6.4.5 some practical concerns involved with such scaling.

---

* Currently, two constraints schemes are possible for $A_{i,iso}^{disp}$. First, we can treat this parameter as a hard constraint, which sets $A_{i,iso}^{disp} = 1$. Second, we can apply boundary conditions to treat $A_{i,iso}^{disp}$ as a free parameter within the range $0.7 \leqslant A_{i,iso}^{disp} \leqslant 1.3$. In future versions of the POInter code, we may also include the option of fitting $A_{i,iso}^{disp}$ subject to a harmonic penalty function.

6.3 The POInter Code

Having identified the required parameters that completely specify MASTIFF and other similar force fields, we now turn to a discussion of the actual fitting process itself. We begin in this section with an overview of the software used to optimize each unconstrained/soft-constrained parameter, and next (in Section 6.4) discuss principles and practices related to fitting each component of the benchmark SAPT energy.

As the name suggests, the **P**arameter **O**ptimizer for **Inter**-molecular Force Fields (POInter) is a Python package developed to aid in the fitting of (two-body + N-body polarization) intermolecular force fields. POInter is open-source and is available for download from `https://git.chem.wisc.edu/schmidt/force_fields`. Documentation and examples for using POInter are available through the wiki at `https://git.chem.wisc.edu/schmidt/force_fields/wikis/home`, but for convenience we include here a brief overview of the program input, output, usage, and main capabilities:

6.3.1 Input

Owing to the large number of parameters that serve as hard constraints in fitting the final MASTIFF force field (see Fig. 6.1), a number of input parameters files are required in POInter. Fortunately, provided the user has already executed the scripts and steps from Chapter 5, all required input scripts should have all been created automatically and copied over to the force field fitting subdirectory (`ff_fitting`) from which the POInter code is intended to be run. Thus in practice, POInter is designed to be run in combination with the Workflow so as to minimize the amount of required manual input.

In total, the following input files are required by the POInter program, where the tag <monomer> indicates that a separate file is required for each unique monomer being fit. Files highlighted in teal or red indicates that the input file sometimes or always require manual modification before running POInter, whereas files in

black are created automatically from the various scripts used in the Workflow, and usually don't require further alteration.

- **<monomer1>_<monomer2>.sapt**: Summarizes the output SAPT energies for each dimer configuration from Section 5.2, and specifies the atomtype for each atom in each monomer
- **<monomer>.disp**: Contains dispersion parameters for each monomer
- **<monomer>.drude**: Contains drude oscillator charges for each monomer
- **<monomer>.exp**: Contains short-range exponents for each monomer
- **<monomer>_<multipole_suffix>.mom**: Contains multipole moments for each monomer
- **__init__.py**: Empty file required to keep Python's module structure happy
- **<monomer>.constraints**: Constraints file, used to include hard-constraints for any A_{ij} parameters for any previously-fit atomtypes. See Section 6.C for details and Listing 6.4 for an example input file.
- **<monomer>.axes**: Axes file, used to specify the local axes and included spherical harmonics for any anisotropic atomtypes. See Section 6.4 for details and Listing 6.3 for an example input file.
- **defaults.py**: List of default settings for the POInter program; these defaults rarely need to be changed for routine force field development. See Listing 6.2 for an example input file.
- **settings.py**: List of modular settings for the POInter program; many of these settings can get changed in the course of routine force field development. See Section 6.4 for details and Listing 6.1 for an example input file.

6.3.2 Usage and Output

Once the required input files have been created/modified, running the POInter program is straightforward:

```
./run_pointer.py
```

After a few minutes of runtime, POInter will generate the following important output files (file prefixes and suffixes may differ slightly depending on the choice of input variables `file_prefix` and `file_suffix` from `settings.py`):

- **coeffs.out**: Output file containing fit parameters and error metrics
- **exchange.dat**: SAPT and force field exchange energies given in a two-column format with ordering identical to the input .sapt file
- **electrostatics.dat**: SAPT and force field electrostatic energies
- **multipoles.dat**: SAPT electrostatic and force field multipolar energies
- **induction.dat**: SAPT and force field second-order induction energies
- **dhf.dat**: SAPT and force field δHF energies
- **edrudes.dat**: polarization energies, $V_{\text{pol}}^{(2)}$ and $V_{\text{pol}}^{(3-\infty)}$, given in two-column format
- **dispersion.dat**: SAPT and force field dispersion energies
- **total_energy.dat**: SAPT and force field total energies

We now discuss specific details related to the fitting of each energy component.

6.4 Force Field Fitting: Principles and Practice

In conjunction with the POInter software, we can finally turn to the main topics of interest in this Chapter: how can we accurately and systematically develop intermolecular force fields? As discussed earlier, with MASTIFF some aspects of

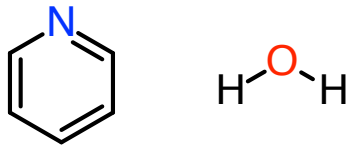


Figure 6.2: Pyridine and water – molecular examples of challenges in force field fitting

the force field development process remain an 'art', and are thus usually guided by chemical intuition, whereas increasingly more aspects of force field fitting now can be carried out in a systematic and reasonably black-box manner. Here we offer an in-depth analysis of the force field development process for MASTIFF and related force fields, paying specific attention to addressing both the 'artistic' and 'scientific' choices that must be made when developing models for new systems.

6.4.1 General

In general, a large number of development choices must be made prior to force field fitting: which benchmark energies to use, how to sample the dimer PES, which parameters to treat as hard constraints, etc. While some of these choices have been discussed in Chapter 5, the following considerations also bear mention:

Atom-typing In developing force fields for new systems (specifically with regards to the MASTIFF approach, though many of the principles below apply generally to other ab initio force fields), an initial choice must be made as to how to categorize each atom into 'atomtypes', where by definition all atoms within the same atomtype share the same force field functional form and parameters. In some cases, such as with water, atomtyping is a fairly obvious decision, and it is easy to see how two unique types should be used to describe the system. With other molecules such as pyridine (Fig. 6.2), however, this atomtyping process is difficult to treat systematically and/or universally, and an iterative guess-and-check process may be required to ascertain the number of atomtypes that are required to obtain a desired level of force field accuracy (see Section 6.4.2 for details).

Note that substantially increasing the number of free atomtypes (i.e., those atomtypes whose parameters have not been pre-fit to a different system) can sometimes lead to numerical instabilities in fitting process or to overfitting,[71] and care must be taken with large/complex systems to ensure good accuracy and transferability.

Anisotropy With the MASTIFF approach, each atomtype can be treated as being either isotropic or anisotropic, and for anisotropic atoms an arbitrary number of spherical harmonic terms can be, in principle, included in the functional form. These spherical harmonic expansions are always calculated with respect to a user-defined local atomic coordinate system, and this coordinate system should be chosen so as to maximize the symmetry (or approximate symmetry) of the system (vida infra).

For anisotropic atomtypes, in addition to specifying a local coordinate system it is necessary to specify the ranks and orders of included spherical harmonic functions $\mathcal{C}_{lk}$ that get included in Eq. (6.3). In practice, inclusion of spherical harmonics beyond rank $l = 2$ does not typically lead to worthwhile accuracy gains, and we suggest truncation at this order for most systems. Additionally, naïve inclusion of all spherical harmonics up to rank 2 can lead to numerical instabilities, and so only symmetry-allowed spherical harmonics (based on the local coordinate system) should be included. With the POInter code, these specifications for anisotropy are listed in the .axes file, with notation as in Listing 6.3.

For most atomtypes (see Chapter 3 for details), an isotropic description of the system is sufficient, and so anisotropy should typically only be necessary for atomtypes corresponding or spatially proximate to heteroatoms and/or multiple bonding environments. Still, it is always worthwhile to explicitly test the effects of treating different atoms anisotropically via comparison to the SAPT exchange energy (see Section 6.4.2 for details).

Benchmark Energies and Correction Factors In Chapter 4, we have discussed situations in which a SAPT-based energy decomposition may be insufficient for force field development, and have suggested strategies for improvement in these

cases. For most systems, however, a SAPT-based decomposition will be of good accuracy, and any deviations between SAPT and gold-standard CCSD(T) can be accounted for using a δCCSD(T) correction term as in Chapter 3. Preliminary results on CO_2, $CHCl_3$, H_2O, and NH_3 suggest that this δCCSD(T) term should be included as part of the dispersion energy (see Section 6.4.5), however more systems should be tested to see if this practice is appropriate for general force field development.

6.4.2 Exchange

From Chapter 2, and as in Eq. (6.1), our exponentially-decaying model for the exchange-repulsion energy requires two sets of parameters per atomtype, A_i^{exch} and B_i:

$$V_{ij}^{exch} = A_{ij}^{exch} P(B_{ij}, r_{ij}) \exp(-B_{ij} r_{ij}) \tag{6.5}$$

Because the exchange-repulsion energy has no long-range contributions (unlike with electrostatics, induction, and dispersion), when analyzing the final force field it is often easiest to use the exchange energy (by itself) to compare between models with differing numbers of atomtypes or treatments of anisotropy. Additionally, when fitting the B_i parameters against a harmonic penalty function, POInter takes advantage of the relative simplicity of the exchange-repulsion model and fits this B_i parameter based solely on the exchange component.

Exponent Fitting To use POInter to relax the B_i parameters from their initial ISA-based values, the following flag in `settings.py` can be set to True:

```
...
# Exchange Settings: fit_bii selects whether or not to treat the ISA
# short-range exponents are soft- (fit_bii=True) or hard-constraints
# (fit_bii=False)
fit_bii                    =    True
...
```

In general, deviations from the input B_i parameters should be no larger than 5–10%. Larger deviations may indicate problems with the calculated BS-ISA exponents or with the fitting process itself.

6.4.3 Electrostatics

Unlike with the exchange energy, the model for electrostatics must account for both the effects of multipolar interactions (at long-range) and charge penetration (at short-range):

$$V_{ij}^{elst} = -A_{ij}^{elst} P(B_{ij}, r_{ij}) \exp(-B_{ij} r_{ij}) + \sum_{tu} Q_t^i T_{tu} Q_u^j \tag{6.6}$$

A_i^{elst} parameters are fit in a similar manner to the exchange energy, though it is not recommended to attempt to re-fit the B_i parameters to the electrostatic energy. As for the multipole energy, in the simplest case (i.e. without off-sites) these parameters can simply be read in using the `settings.py` file,

```
# Electrostatic Settings: choose which multipole files the program should use
multipoles_suffix          =    '_ISA-GRID_L2.mom'
```

where the `multipoles_suffix` should point to some file `<monomer><multipoles_suffix>` in the `input/` subdirectory. In general, an L2 model is a good accuracy benchmark, and it's often useful to compare the energies obtained from point-charge models to those achievable with the L2 model.

Off-site models As described in Chapter 5, practical software limitations and issues of computational expense may sometimes require us to forego use of higher-order multipole terms. When modeling the electrostatic energy via a point charge model, best accuracy is often achieved by including off-site charges. Such off-site point charge models can be treated using POInter, though the following modifications to the standard input scripts are required:

1. Add the off-site positions to the .sapt file. Scripts for doing this are discussed in Section 5.6.

2. Modify the various monomer parameter files (located in the `input/` subdirectory to reflect the newly-added off-site positions:

   a) Add dispersion coefficients (usually all zero) to each `<monomer>.disp` file and for each atomtype

   b) Add extra blocks to each `<monomer>.exp` corresponding to the off-site positions. The .exp file(s) list exponents for each atom in the same order as the .sapt file, and the .exp and .sapt file orderings must match. Additionally, the off-site B_i parameters must be set to a non-zero value to avoid numerical errors.

   c) Add extra drude parameters to each off-site in the `<monomer>.drude` file. Atom ordering is as in the .sapt file, and (assuming you do not want the off-sites to be polarizable) the drude charge parameters should be set to zero.

   d) Assuming you do not wish short-range parameters to be fit to your off-site atoms, add the names of all off-site atomtypes to `defaults.py`:

```
lone_pair_flags                 = ['Du', 'lp']
```

Energies between the off-site point charge model and the (more standard) L2/L0 models should always be compared as a sanity check: if the errors in the offsite model are larger than the errors from the L0 model, this is usually a sign that something has gone wrong. On the other hand, if the electrostatic energies are of a similar accuracy to those obtained with the L2 model, this indicates that the point charge model is a fairly optimal description of the system's multipolar electrostatics, and can be used without further modification in molecular simulation.

6.4.4 Induction

We turn now to a discussion of the SAPT induction energy, arguably the most complicated energy component to understand and correctly model. Though there is, as of yet, no one 'best practice' for modeling the SAPT induction energy, a good induction model will need to account for all of the following physical phenomenon:

- Long-range polarization
- Polarization damping at short-range
- Charge penetration effects arising from polarization
- Charge transfer

Aside from long-range polarization, for which asymptotically-exact formulas exist and can be modeled in simulation,[77,248,249,260] there is currently little literature consensus as how best to separate out and model the various physically-meaningful induction effects. Complicating matters further, even though the SAPT induction energy is purely a 2-body effect, the polarization model we develop based on the induction energy also implicitly defines the model for *many*-body polarization, which accounts for a sizeable fraction of the total many-body energy discussed in **??**. Consequently, it is easily possible to obtain an induction model which shows good accuracy for dimer computations, but which leads to substantial inaccuracies in modeling larger clusters or bulk liquids.

Improved induction models will certainly need to be the subject of future work, and are discussed further in **??**. In the meantime, below we present a summary of common induction models that can be used to fit SAPT-based force fields.

Polarization For reasonably isotropic systems, long-range polarization effects can be described either by a Drude oscillator model or via induced dipoles.[77,311] In practice, these models tend to be numerically similar, and we use them both during parameterization and simulation.* POInter uses a Drude oscillator model for the purposes of computing the polarization energy, and the Drude charges and associated spring constants are read in as input in the `<monomer>.drude` file(s). These charges can be obtained using the methods described in Section 5.5.5, however note that these charges are sensitive to the choice of polarization damping model described below, and may need to be refit if either the functional form or parameters for the damping model are changed.

For more anisotropic systems (such as water), higher-order polarizabilities have been shown to be important, and need to be included for best accuracy.[99,248,312,313] At present, however, higher-order polarizabilities have not been implemented in most common software packages, and we are in the process of investigating how to use off-site polarizabilities for modeling highly anisotropic systems.

Polarization Damping At short range, the induced dipole polarizabilities must be damped in order to avoid the so-called 'polarization catastrophe', an effect in which nearby polarization sites mutually polarize each other to infinite values. While there is widespread consensus as to the importance and necessity of including polarization damping, functional forms and parameters for the polarization damping vary widely.[63,262,308,311,314] Thole-type damping functions* are some of the more commonly used, and some effort has been put forth to compare between several similar damping functions and parameterization schemes.[261,314,315]

Historically,[83] several members of our group have used an exponentially-decaying Thole function with an associated damping parameter of 2.0.[166] More recently, and due to software limitations in OpenMM, we have taken to using the 'Thole-tinker' model with a universal Thole damping parameter $a = 0.33$, which is reasonably similar to the damping parameter used by the AMOEBA force field.[263] Various Thole-type models can be specified in POInter via the `settings.py` file:

```
# Induction Settings: Choose the type and parameters for the polarization
# damping functions. Options for thole_damping_type are 'thole_tinker' and
# 'thole_exponential', and good defaults for thole_param are the 0.33 and 2.0 with
# respect to the two different damping types
# respectively
thole_damping_type          =   'thole_tinker'
thole_param                 =    0.33
```

Note that the choice of Thole damping parameter can be very important, as this modifies the relative balance between energies ascribed to polarization vs. charge transfer, in turn modifying the magnitude of the many-body polarization. In order

*In specific, Drude oscillator models have been used historically in our group for their simplicity and ease of implementation. More recently, we have begun running our simulations with the induced dipole model in order to maintain compatibility with OpenMM.

to achieve a model that achieves the correct balance between polarization and other inductive effects, future work may need to involve some of the following advances:

1. Atomtype-specific Thole damping parameters
2. New functional forms for polarization damping
3. Explicit separation between the SAPT charge-transfer and polarization energies. Several schemes have already been proposed to achieve this decomposition,[154,289,316,317] and the various available schemes should be tested for their utility in force field development.

Charge transfer and inductive charge penetration In addition to polarization damping, charge transfer and inductive charge penetration can become important at shorter intermolecular separations. Physically-motivated functional forms for these effects are generally lacking, though Misquitta[154] has suggested a double exponential decay, and work in our own group has empirically found a single exponential decay (with ISA-derived exponents B_i) to be reasonably satisfactory.

Conclusions and Recommendations Our current approach to modeling inductive effects include a sum over two contributions:

$$V_{ij}^{ind} = -A_{ij}^{ind} P(B_{ij}, r_{ij}) \exp(-B_{ij} r_{ij}) + V_{pol}^{(2)} \tag{6.7}$$

$$V_{ij}^{\delta^{HF}} = -A_{ij}^{\delta^{HF}} P(B_{ij}, r_{ij}) \exp(-B_{ij} r_{ij}) + V_{pol}^{(3-\infty)} \tag{6.8}$$

The SAPT benchmark separates the induction energy into 2nd- and higher-order (i.e. δHF) induction, and we fit both induction-like terms separately.* All parameters for the polarization model V_{pol} are currently read in as hard constraints, and the

*Be advised, most papers in the literature will simply state that a 'Thole damping function' was used, but will not make explicit which of several different Thole-type damping functions was meant.

A_i^{ind} and $A_i^{\delta^{HF}}$ prefactors (which effectively accounts for both charge transfer and charge penetration) are directly fit by POInter.

6.4.5 Dispersion

Dispersion is the last energy component that we must model in order to completely describe the two-body force field. Asymptotically, the dispersion energy follows a well-defined expansion in powers of $1/r^{2n}$, and at shorter distances the energy expression is typically damped by a Tang-Toennies function,[156,157]

$$V_{ij}^{disp} = -A_i^{disp} A_j^{disp} \sum_{n=3}^{6} f_{2n}(x) \frac{C_{ij,2n}}{r_{ij}^{2n}} \tag{6.9}$$

$$A_i^{disp}(\theta_i, \phi_i) = A_{i,iso}^{disp} \left(1 + \sum_{l>0,k} a_{lk}^{disp} \mathcal{C}_{lk}(\theta_i, \phi_i) \right) \tag{6.10}$$

where $f(x)$ is the Tang-Toennies damping function from Eq. (6.1), and the various colors highlight (as in Fig. 6.1) the different ways in which the dispersion parameters are calculated/fit. Dispersion coefficients $C_{ij,2n}$ must always be read into POInter as input, and methods for obtaining these coefficients are as described in Section 5.5.4.

In obtaining a final model for dispersion, it is important to ensure that any model is quantitatively correct in the asymptotic regime, as the least-squares optimization procedure POInter will not explicitly ensure this physically-correct behavior. Unless directly fitting atomtype-specific scale factors to the dispersion energy (vida infra), a good strategy is to (using the methods in Section 5.5.4) manually fit a universal scale factor to the dispersion coefficients in order to achieve correct asymptotic behavior. Once these scaled dispersion coefficients are read into POInter, additional anisotropic parameters can then be fit (or set to zero) by appropriate modification of the `settings.py` file:

```
# Dispersion Settings: Choose which parameters to fit to the dispersion
energies. Fit options
```

* Note that this expansion is in orders of perturbation theory, not in orders of the many-body expansion.

```
# include 'none' (to fit no parameters), 'anisotropic' (to just fit
# anisotropic dispersion parameters, but to leave isotropic dispersion
# coefficients unscaled), and 'all' (to fit both anisotropic and isotropic
# dispersion coefficients)
fit_dispersion              =    'anisotropic'
```

In select cases, such as when a δCCSD(T) correction is added to the dispersion energy, it can be worthwhile to scale the isotropic dispersion coefficients in an atomtype-specific manner. (This strategy was used in Chapter 3 to obtain MASTIFF-CC dispersion parameters for CO_2, NH_3, H_2O, and $CHCl_3$.) This behavior is also allowed in POInter using the above flags, however care must be taken to ensure that this optimization does not degrade accuracy in the asymptotic regime. In general, and in the absence of subsequent significant improvements to the overall force field fit quality, it is usually advised *not* to fit isotropic scale factors to the dispersion energy.

6.4.6 Many-Body Effects

As discussed above, some of the force field parameters that define the two-body force field also contribute to the many-body energy. Polarization in particular is inherently many-body, and (when possible) trimer energies should be computed to ensure that the polarization model defined in Section 6.4.4 leads to good three-body energies.

For less polar systems, three-body dispersion and exchange can also be important to include, and McDaniel and Schmidt[4] have shown how these parameters can be directly calculated and modeled in the many-body portion of the force field. This explicit three-body force field follows the well-known Axilrod–Teller–Muto triple dipole functional form, and Ref. 4 describes how to obtain the necessary parameters.

6.5 Force Field Validation: Assessing Fit Quality

Having obtained a final force field, the resulting model should always be assessed and validated before use in molecular simulation. We divide this topic into two sections – sanity checks and validation – to distinguish between tests that can be performed via visual inspection of the fits compared to those that require additional computations.

6.5.1 Sanity Checks

Visualization Once the POInter code has been successfully run, the resulting force fields should always be visualized, possibly with the aid of the included visualization scripts,

```
plot_compare_sapt_components.py -p <file_prefix> -s <file_suffix>'.dat' --display
plot_sapt_component_errors.py <file_prefix> <file_suffix>'.dat' --display
```

where additional visualization options are available by inserting the `-h` flag into the function call for either script. Such visualizations are shown, using the pyridine dimer as an example, in Figs. 6.3 and 6.4, and additional visuals for water and other molecules are shown in Figs. 6.5 to 6.7 and Section 3.C, rspectively.

Error Analysis of the Minimum Energy Region As discussed in Section 5.2, the mimimum energy region plays a highly important role in simulations, and thus it is crucial that our force fields correctly predict these energies. Ideally, the energy predictions should be within ± 1 kJ/mol of the benchmark energy, though larger errors are easily possible when developing isotropic force fields. Even more important than this precision, we must ensure that *on average* the force field energies are accurate in the minimum energy region. Any systematic errors in the force field will have a pronounced effect on simulation quality, particularly for studying bulk properties, as many properties depend only on the average of system energy.

Figs. 6.4 and 6.7 show errors in the minimum energy region for both pyridine and water, and serve as illustrative examples. These examples have been chosen

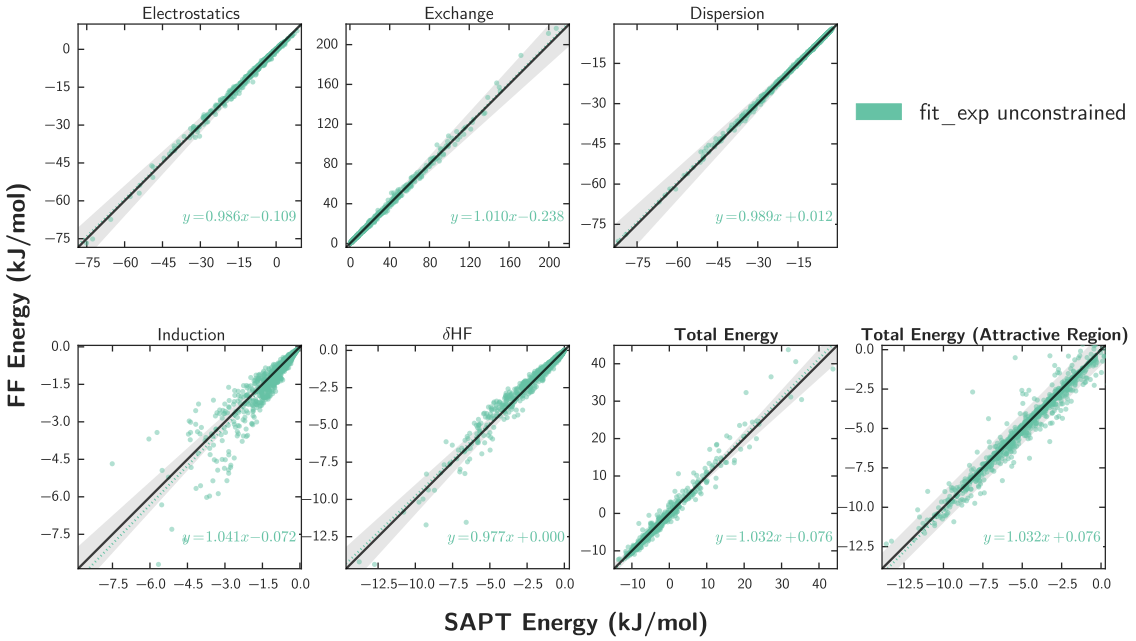


Figure 6.3: The pyridine dimer

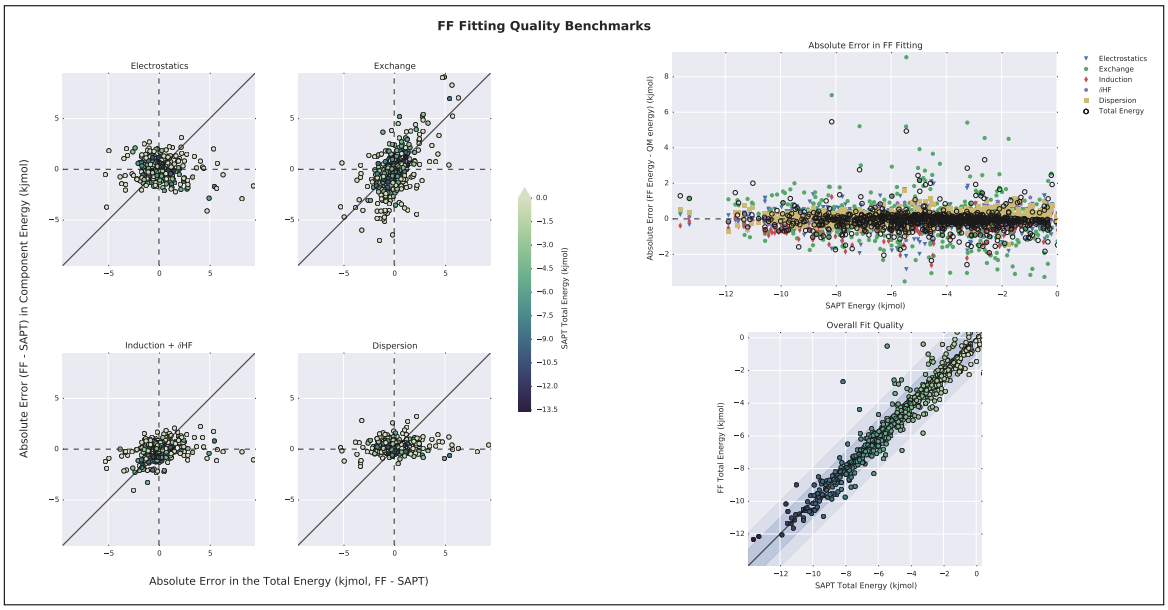


Figure 6.4: The pyridine dimer errors

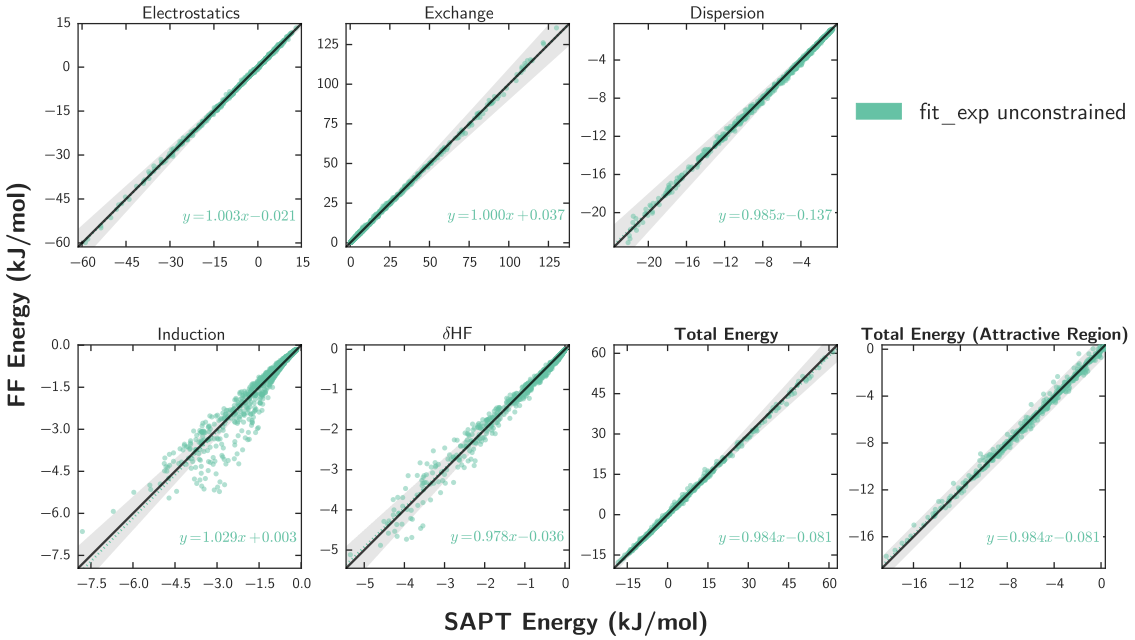


Figure 6.5: The water dimer

because both force fields are relatively good quality, but also show errors that could degrade simulation quality and be the focus of future improvement.

In the case of pyridine, the force field fit shows fairly little systematic error, and could probably be used to simulate bulk properties without issue. Random errors on the order of 0.5–1.0 kJ/mol are typical for this force field, which may or may not be an issue depending on the desired accuracy level and types of simulations one intends to run. (A few outliers show large errors compared to the SAPT benchmark, however these points almost certainly reside along the repulsive wall, judging by their exchange energies, and are thus not cause for great concern.) If one desired to improve the precision of the pyridine force field, it is necessary to assess errors in the force field on a component-by-component basis. Fig. 6.4 shows how errors in the exchange component dominate the overall error, and should be the first target for improvement. Since the exchange energy only depends on a short-range exponential decay, this error could possibly be mitigated by increasing the number

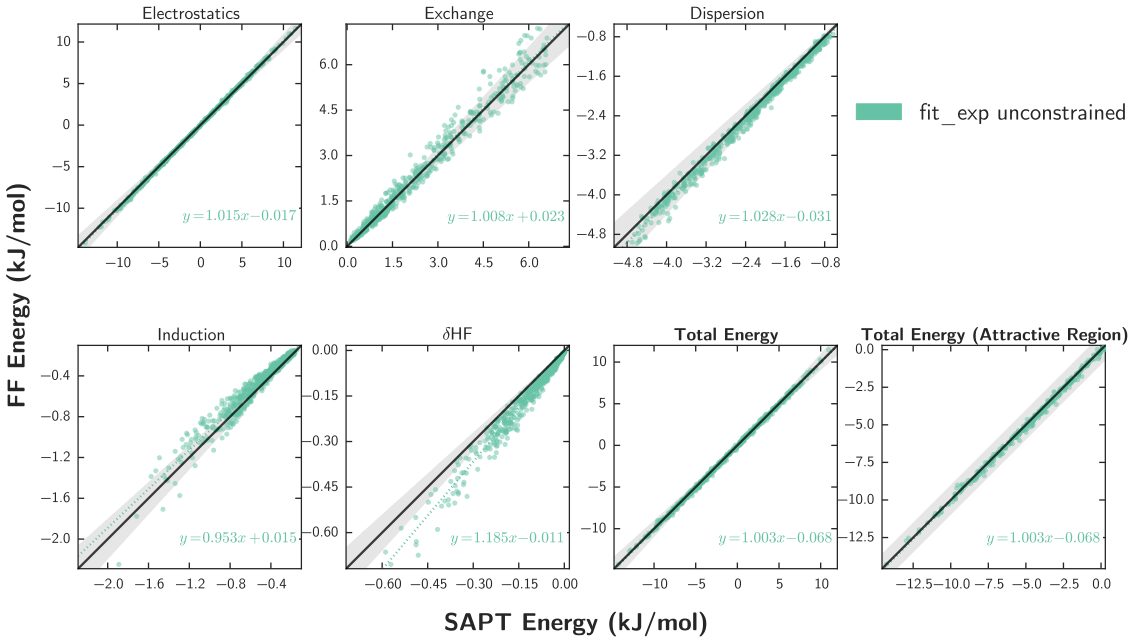


Figure 6.6: The water dimer

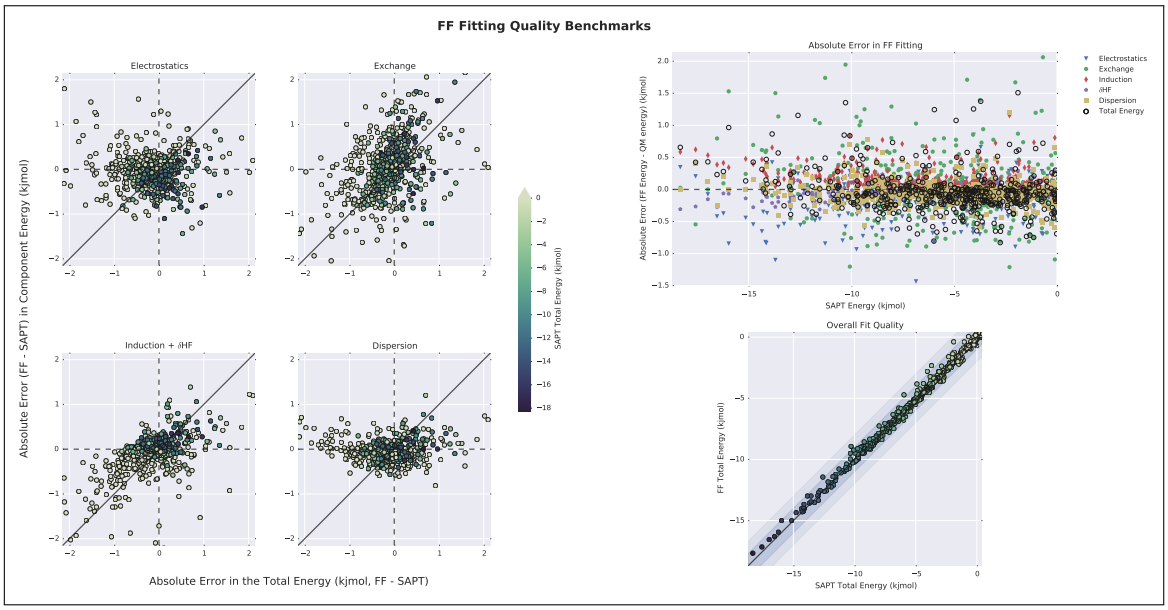


Figure 6.7: The water dimer errors

of atomtypes (as shown, this force field only has three atomtypes) or by including anisotropy on additional atomtypes, as this potential only includes anisotropy on the nitrogen atom, and neglects possibly-important anisotropies in the carbon atoms.

As for water, we see relatively little random error in the force field (especially compared to the more isotropic models discussed in Chapter 3), although again much of this random error can be attributed to the exchange energy. In contrast to pyridine, however, we see some evidence for systematic error in the water potential near the minimum energy configurations. (Indeed, after testing our water force field against the larger CC-pol database,[266] we continued to find evidence for overly-repulsive predictions in the minimum energy region.) This systematic error, small as it may seem, is exacerbated in modeling larger clusters (vida infra), and would need to be fixed before we could expect good success with this force field in general simulations involving water. Looking at the errors on a component-by-component basis, we can see that systematic errors in the potential are heavily correlated with errors in the induction energy, making this induction energy the most important target for improved modeling. As will be discussed in **??**, improving our polarization model (both in terms of the polarization damping and the long-range polarizabilities themselves) will hopefully serve to reduce these systematic errors and improve the overall quality of force fields for strongly polarizable molecules.

Error Analysis of the Asymptotic Region In addition to the minimum energy region, it is important to ensure that the asymptotic region of the potential is modeled correctly for each energy component. This asymptotic analysis is shown for the water dimer in Fig. 6.6, where we have shown force field fits for each component, but have excluded configurations with exchange energies above a certain threshold. (In general, the magnitude of the SAPT exchange energy can be used as a proxy for the 'short-rangeness' of a given configuration.)

Particularly for force fields that directly optimize the dispersion energy, it is important to ensure that each energy component displays asymptotically-correct behavior. This analysis can also be used to determine optimal scaling values for

the dispersion coefficients (see Section 5.5.4 for details).

6.5.2 Validation

Trimer and Other Cluster Interaction Energies Currently, it is hard to guarantee that our polarization models will correctly describe both the two- and three-body polarization energies. In order to validate the many-body portion of the potential, trimer interaction energies should be computed for a subset of energetically-important configurations (preferably those taken from simulation) and compared to an accurate electronic structure benchmark. These validation studies are especially critical for highly-polar systems (such as water), as in these cases the many-body polarization energy can account for a non-negligable fraction of the total system energy. For such polar systems, it can also be important to study the interactions of larger clusters in order to probe four-body and higher interactions.[312,318,319]

Simulations As the ultimate goal for many potentials is to be able to perform molecular simulations, it is useful to validate new force field parameters in relation to 'simple' experimental properties. A reasonable starting point for such comparisons is with the temperature-dependent 2nd virial coefficient, as this quantity is a direct measure of the underlying two-body potential. Methods for calculating the 2nd virial coefficient are discussed in Chapters 2 and 3 and in Refs. 83, 166.

In addition to the virial coefficient, a variety of other simulations form useful comparisons to experiment (especially for studying bulk properties), and examples of these can be found in Chapters 2 and 3 and in Refs. 4, 83, 166.

6.6 Summary and Outlook

Throughout the past two Chapters, we have outlined a force field fitting development methodology which enables an increasingly systematic parameterization of intermolecular force fields. By fitting these force field on a component-by-component basis, minimizing parameterization via frequent recourse to monomer property

calculations, and ensuring the physicality of the various functional forms that get used in the final force field, we now can reliably generate accurate and transferable force field parameters for a broad class of materials and molecular systems. Though invariably "much remains to be learned" when it comes to intermolecular force field development, it is hoped that the principles and practices outlined in this Chapter will sufficiently guide new force field developers in extending the applications of the MASTIFF methodology (or similar EDA-based force fields) to tackle new and interesting problems in molecular simulation.

6.A POInter Input Files

In total, POInter requires modification of three input files, as follows:

Listing 6.1: settings.py

```
#############################################################################
######################### General Settings ##################################
#############################################################################
# Monomer Names (should match ordering in .sapt file)
mon1                        =    'chloromethane'
mon2                        =    'chloromethane'

# Constrained Atomtype settings: Make a list of all atomtypes whose parameters
# should *not* be fit, and include parameters for these atomtypes in the
# relevant <monomer>.constraints file
constrained_atomtypes       =    []

# Names for output files
file_prefix                 =    'fit_exp_'
file_suffix                 =    '_unconstrained'

#############################################################################
#############################################################################



#############################################################################
#################### Component-Specific Settings ############################
#############################################################################
# Electrostatic Settings: choose which multipole files the program should use
multipoles_suffix           =    '_ISA-GRID_L2.mom'

# Exchange Settings: fit_bii selects whether or not to treat the ISA
# short-range exponents are soft- (fit_bii=True) or hard-constraints
# (fit_bii=False)
fit_bii                     =    True

# Induction Settings: Choose the type and parameters for the polarization
# damping functions. Options for thole_damping_type are 'thole_tinker' and
# 'thole_linear', and good defaults for thole_param are the 0.33 and 2.0 with
# respect to the two different damping types
# respectively
thole_damping_type          =    'thole_tinker'
thole_param                 =    0.33

# Dispersion Settings: Choose which parameters to fit to the dispersion energies. Fit
    options
# include 'none' (to fit no parameters), 'anisotropic' (to just fit
# anisotropic dispersion parameters, but to leave isotropic dispersion
# coefficients unscaled), and 'all' (to fit both anisotropic and isotropic
```

```
# dispersion coefficients)
fit_dispersion              =    'anisotropic'


# Residual error Settings: If set to true, fits a final A parameter to errors in the
    total
# energy in an effort to reduce systematic errors in the total energy
fit_residuals               =    False

###############################################################################
###############################################################################



###############################################################################
###################### Functional Form Settings ###############################
###############################################################################
# Radial functional forms f(r); see Stone's book for more details.
# Options are 'slater', 'stone', 'born-mayer', 'born-mayer-sisa', or 'lennard-jones'
functional_form             =    'slater'

# Combination rule settings: Select combination rules for each A prefactors, B
# exponents, and C dispersion coefficients. Options are as follows:
#   aij: 'saptff', 'waldman-hagler5', 'geometric'
#   bij: 'saptff', 'waldman-hagler5', 'geometric_mean', 'arithmetic_mean'
#   cij: 'geometric'
aij_combination_rule        =    'geometric'
bij_combination_rule        =    'geometric_mean'
cij_combination_rule        =    'geometric'

###############################################################################
###############################################################################
```

Listing 6.2: defaults.py

```
#############################################################################
########################  POInter Defaults ##################################
#############################################################################
# The following defaults should be used for most routine force field
# developement, however advanced users may wish to change some of the
# following settings:
exponent_source                     =  'ISA'
lone_pair_flags                     =  ['Du' , 'lp']
scale_weighting_temperature         =  5.0
separate_induction_exponents        =  False
springcon                           =  0.1
weighted_rmse_cutoff                =  0.0
electrostatic_damping_type          =  'None'
include_slater_charge_penetration   =  False
induction_damping_type      =    'Thole'

# Unless you know what you're doing, the following settings should only be
# changed by developers:
__version__ = '1.1.0'


#############################################################################
#############################################################################
```

Listing 6.3: **pyridine.axes**. For each anisotropic atomtype, the approximate symmetry and all terms included in the spherical harmonic expansion are listed to the right of the atomtype. Additionally, the local axis reference frame for each anisotropic atomtype is defined in the Axes subsection using the z-then-x convention employed by AMOEBA and other potentials (see Chapter 3 for details). The first column of the axes subsection denotes the index of the anisotropic atom (atom ordering as in the .sapt file), and the second column denotes whether the z or x axis is being defined. For certain local symmetries, the choice of x-axis is unimportant, and so not every anisotropic atomtype has a defined x-axis. The remaining columns define the direction vector for the axis in terms of atomic indices. The first index (often the anisotropic atom itself) lists the start of the vector, and the endpoint of the vector is defined as the midpoint of all subsequently listed atoms.

```
Pyridine
N    c2v       y10 y20 y22c

Axes
ATOM#   AXIS (z or x)   Atomic Indices defining vector (either 2 or more integers)
5 z 5 6 10
5 x 5 6
```

Listing 6.4: pyridine.constraints

```
3
EXCHANGE
H       0.424450
C       2.244624
N       1.541775   -1.000000    0.612747   -0.839073
ELECTROSTATICS
H       0.182383
C       1.421735
N       1.007315   -1.000000    0.715037   -0.778451
INDUCTION
H       0.032989
C       0.313818
N       0.000000    0.546327   -0.232167    0.184632
DHF
H       0.137793
C       0.568162
N       0.383885   -1.000000    0.336640   -0.827635
DISPERSION
H       1.000000
C       1.000000
N       1.000000   -0.163375    0.092436   -0.169124
RESIDUALS
H       0.000000
C       0.000000
N       0.000000    0.000000    0.000000    0.000000

EXPONENTS
H       2.119818
C       1.939994
N       2.110311
```

6.B POInter Output Files

Listing 6.5: coeffs.out

```
########################## FF Fitting Summary ############################
-------------------------------------------------
Program Version: 1.1.0
Short-range Functional Form: slater-ff
Combination Rules: aij = geometric
                   bij = geometric_mean
                   cij = geometric
Electrostatic Damping Type: None
Thole          Damping Type: thole-tinker
Thole Param: 0.33
Fitting weight: eff_mu = 0.0 Ha
                eff_kt = 0.0259 Ha
Weighted RMSE cutoff: 0.0 Ha
Anisotropic Atomtypes: N
-------------
Exponents (Optimized):
H(0)        2.119818
C(0)        1.939994
N(0)        2.110311
-------------
Monomer 1 Multipole File:
pyridine_ISA-GRID_L2.mom
Monomer 2 Multipole File:
pyridine_ISA-GRID_L2.mom
-------------------------------------------------
Exchange Parameters:
    Functional Form =
        E(exch)_ij = A*K2(rij)*(1 + a_yml*Y_ml)*exp(-bij*rij)
    where the a coefficient for each spherical harmonic term Y_ml
    is listed in the parameters below and
        K2(rij) = 1/3*(bij*rij)**2 + bij*rij + 1
Fitted Atomtypes
                A
H(0)         0.424450
                A
C(0)         2.244624
                A             a_y10           a_y20           a_y22c
N(0)         1.541775        -1.000000        0.612747       -0.839073
Constrained Atomtypes
  None
-------------
Exchange RMS Error: 6.34645e-04
Exchange Weighted RMS Error: 3.78772e-04
Exchange Weighted Mean Signed Error: 1.15033e-05
Exchange Weighted Least-Squares Error: 9.61033e-05
```

```
-------------------------------------------------
Electrostatic Parameters:
    Functional Form =
        E(elst)_ij = f_damp*qi*qj/rij - A*K2(rij)*(1 + a_yml*Y_ml)*exp(-bij*rij)
Fitted Atomtypes
                A
H(0)        0.182383
                A
C(0)        1.421735
                A           a_y10           a_y20           a_y22c
N(0)        1.007315      -1.000000        0.715037      -0.778451
Constrained Atomtypes
  None
-------------
Electrostatics RMS Error: 2.46881e-04
Electrostatics Weighted RMS Error: 1.75164e-04
Electrostatics Weighted Mean Signed Error: 1.66766e-05
Electrostatics Weighted Least-Squares Error: 3.47690e-05
-------------------------------------------------
Drude oscillator energy has been calculated using the following method: multipole-
    gradient
Induction Parameters:
    Functional Form =
        E(ind)_ij = shell_charge - A*K2(rij)*(1 + a_yml*Y_ml)*exp(-bij*rij)
Fitted Atomtypes
                A
H(0)        0.032989
                A
C(0)        0.313818
                A           a_y10           a_y20           a_y22c
N(0)        0.000000        0.546327      -0.232167        0.184632
Constrained Atomtypes
  None
-------------
Induction RMS Error: 1.75440e-04
Induction Weighted RMS Error: 1.32125e-04
Induction Weighted Mean Signed Error: 4.84152e-05
Induction Weighted Least-Squares Error: 1.57535e-05
-------------------------------------------------
DHF Parameters:
    Functional Form =
        E(dhf)_ij = - A*K2(rij)*(1 + a_yml*Y_ml)*exp(-bij*rij)
Fitted Atomtypes
                A
H(0)        0.137793
                A
C(0)        0.568162
```

```
                A           a_y10           a_y20           a_y22c
N(0)        0.383885      -1.000000        0.336640      -0.827635
 Constrained Atomtypes
   None
-------------
Dhf RMS Error: 1.36805e-04
Dhf Weighted RMS Error: 8.45054e-05
Dhf Weighted Mean Signed Error: -9.08316e-06
Dhf Weighted Least-Squares Error: 9.71484e-06
-----------------------------------------------
Dispersion Parameters:
    Functional Form =
        E(disp)_ij = sum_(n=6,8,10,12){A*fdamp_n*(Cij_n/r_ij^n)}
                C6            C8            C10           C12
H           1.498365      6.488286      33.240633     143.508377
C           5.315289     20.555612      99.173881     391.059721
N           4.242425     20.639587     150.575182     765.384698
 Fitted Atomtypes
                A
H(0)        1.000000
                A
C(0)        1.000000
                A           a_y10           a_y20           a_y22c
N(0)        1.000000      -0.163375        0.092436      -0.169124
 Constrained Atomtypes
   None
-------------
Dispersion RMS Error: 1.61515e-04
Dispersion Weighted RMS Error: 1.09295e-04
Dispersion Weighted Mean Signed Error: -4.00563e-05
Dispersion Weighted Least-Squares Error: 1.59366e-05
-----------------------------------------------
Total Energy:
-------------
Total Energy RMS Error: 4.50690e-04
Total Energy Weighted RMS Error: 2.30673e-04
Total Energy Weighted Mean Signed Error: 2.74556e-05
Total Energy Weighted Least-Squares Error: 9.22127e-05
-----------------------------------------------
###############################################################################
```

6.C Additional Fitting Options

Electrostatic damping types

slater charge penetration
exact vs approximate slater functional form
weighting temperature
constrained A params

7 CONCLUSIONS AND FUTURE DIRECTIONS

In this dissertation, we have presented a systematic methodology for improved ab initio force field development on the basis of Symmetry-Adapted Perturbation Theory (SAPT) and Iterated Stockholder Atoms (ISA) calculations. Critically, the strategies developed herein allow for accurate and physically-based modeling of short-range effects in intermolecular force fields, and enable accurate, transferable, and cost-effective treatment of the important atomic-level anisotropies commonly found in organic compounds. These improved methodologies for ab initio force field development have culminated in our MASTIFF model for intermolecular interactions, and we are already approaching the stage where MASTIFF can be used in the generation of broadly-applicable force fields for large-scale molecular simulation.

Before such large-scale simulations can become a possibility for strongly polarizable systems, fundamental limitations in the induction model for MASTIFF will be need to be adddressed. Future work on polar systems should focus on improved and more accurate treatment of the long-range polarization, the development of new models to more physically treat the polarization damping, and the decomposition of the total induction energy into charge-transfer and polarization contributions. Though much work is needed to outline specific strategies to tackle each of these issues, it is hoped that regularized SAPT methods (Ref. 154), which can perform the charge-transfer/polarization decomposition, and the ISA-pol method (Chapter 5), which can naturally partition the long-range polarization energies and quantify higher-order and/or anisotropic polarizabilities, might be of good service in this endeavor.

Assuming challenges in modeling the induction energy can be met, a second goal for future study should be the application of MASTIFF to the wide range of chemical problems where atomic-level anisotropy is particularly important. These applications can initially consist of ab initio force field development for specific molecules, and we are already in the process of developing and testing models

for industrially important compounds where isotropic force fields have known accuracy issues, such as with benzene and ethene. Emphasis should also be placed on generating transferable anisotropic ab initio force fields for general *classes* of molecules, such that the MASTIFF model can ultimately be employed as a general, all-purpose model for accurate molecular simulation. The development of these general force fields will require us to address various challenges not yet considered with the MASTIFF methodology, such as the treatment of flexible monomer geometries and the generalization of atom-specific anisotropic parameters into transferable and general atom type parameters. Nevertheless, and assuming these challenges can be overcome, it is hoped that MASTIFF and other 'next-generation' ab initio force fields will lead to increasingly accurate and robust models for molecular potential energy surfaces (PESs), such that the complex, inherently anisotropic details of intermolecular interactions may be routinely studied in large-scale molecular simulation.

Part IV

Codes

A FORCE FIELD DEVELOPMENT WORKFLOW

Purpose

Derive a first-principles, SAPT-based force field.

Relevant Literature

- VanVleet2016: 10.1021/acs.jctc.6b00209
- VanVleet2017: TBA
- McDaniel2013: 10.1021/jp3108182
- Schmidt2015: 10.1021/ar500272n
- Yu2011: 10.1021/jp204563n

Overview

To generate a SAPT-based force field, the following inputs are required: 1. Benchmark dimer energies from SAPT, computed for a variety of dimer configurations 2. Long-range multipole moments, induced dipoles, and dispersion parameters, computed from monomer properties (and BS-ISA in particular) 3. Short-range exponents computed from monomer properties (and BS-ISA in particular) 4. Short-range pre-factors fit to dimer energies

The following scripts are designed to simplify (as much as is possible) the workflow for force field generation.

Method

1. Generate the necessary input files upon which the scripts in step #2 depend. The following files must be manually created/edited, and can all be found in the templates subdirectory (with an example set of input files given for the pyridine dimer):
    1. dimer_info.dat
        - For each monomer, list the monomer's name and the charge on the
        - monomer. The appropriate file format should be clear from the pyridine example.
        - In the manner described in dimer_info.dat, list all midbonds that should be added between monomers. Midbonds are important for running accurate SAPT calcuations; see Yu2011 for details.
    2. generate_grid_settings.inp
        - This is the input file for GenerateGridPoints, which generates the dimer configurations for running SAPT calculations. The input file is commented so as to be self-explanatory; you will need to change (at the very least) the 1st, 3rd, and 4th input sections based on the identities of the two monomers
    3. MONA_MONB.inp (where MONA and MONB are replaced by the monomer names listed in dimer_info.dat)
        - This file contains a title line (line 1), and (for each monomer) the number of atoms followed by a list of coordinates in .xyz format. See pyridine_pyridine.inp for an example.

4. MONA.atomtypes, MONB.atomtypes
    - Each .atomtypes file has the format of a .xyz file, where the element names have been replaced by atomtypes. This file will be used to generate the CamCASP input files needed for ISA calculations, and is also necessary for pre-processing the input files for force field fitting.

2. To generate all files necessary to run force field calculations, run the following pre-processing scripts (from this main directory).

```
./scripts/make_geometries.sh

./scripts/get_global_coordinates.py

./scripts/submit_ip_calcs.py
```

(wait until IP calculation is finished)

```
./scripts/make_sapt_ifiles.py

./scripts/make_isa_files.py

./scripts/make_dispersion_files.py
```

3. Submit all SAPT and ISA calculations to relevant locations. At the time of this writing, SAPT calculations should preferably be run on HCTC (Condor). ISA and dispersion calculations should be run on Phoenix using Camcasp 5.8. Copy all output files back to Pople.

4. Workup the results of the SAPT and ISA calculations by running the following post-processing scripts:

```
./scripts/workup_sapt_energies.py

./scripts/workup_dispersion_files.sh
```

(Depending on the force field, dynamic polarizabilities may need to be added to templates/dispersion_base_constraints.index before running this script. See Jesse McDaniel's thesis and \cite{McDaniel2013} for a full description of the paramterization process for dispersion coefficients.)

```
./scripts/workup_drude_files.sh
```

(Depending on the force field, static polarizabilities may need to be added to templates/drude_base_constraints.index before running this script. See Jesse McDaniel's thesis and \cite{McDaniel2013} for a full description of the paramterization process for drude oscillator charges.)

```
./scripts/workup_isa_charges.py

./scripts/workup_isa_exponents.py
```

After running these scripts, you should have the SAPT energies, long-range coefficients, and short-range exponents required to run the force fitting code (which is needed to generate short-range pre-factors, see \cite{VanVleet2016}). The proper running of this code is described in the POInter documentation, see

https://git.chem.wisc.edu/schmidt/force_fields/wikis/home

Overview of Important Files

- dimer_info.dat <- monomer names and midbond positions
- dispersion_template.clt <- CamCASP input file for getting induction and
- dispersion paramters
- generate_grid_settings.inp <- geometry configuration settings
- isa_template.clt <- CamCASP input file for getting ISA exponents
- pbe0_template.com <- DF-DFT-SAPT template for the PBE0 functional
- pyridine.atomtypes <- change elements to atomtypes; only matters for
- dispersion
- pyridine_pyridine.inp <- monomer geometries

For most systems, only dimer_info.dat, the .inp files, and the .atomtypes file will need to be changed. The examples provided for these files should hopefully make the format self-explanatory.

System Requirements

Python dependencies: * numpy * scipy * chemistry (mvanvleet package; not standard, so this needs to be * downloaded and added to your `$PYTHONPATH`)

Figure A.1: An overview of the semi-automated force field development process. The full workflow and required scripts can be found at `https://github.com/mvanvleet/workflow-for-force-fields`.

A.1 Monomer Geometries

[4] McDaniel, J. G.; Schmidt, J. R. *J. Phys. Chem. B* **2014**, *118*, 8042–8053.

[5] Witt, R. K.; Kemp, J. D. *J. Am. Chem. Soc.* **1937**, *59*, 273–276.

[6] Riddick, J. A.; Bunger, W. B.; Sakano, T. K. *Techniques of Chemistry, Vol. II: Organic Solvents: Physical Properties and Methods of Purification*, 4th ed.; Wiley-Interscience: New York, 1986.

[7] Span, R.; Wagner, W. A new EOS for CO2 covering the fluid region from the triple point temperature to 1100K at pressures up to 800MPa.pdf. 1996.

[8] Guest, M. F.; Sherwood, P.; Nichols, J. A. *New Horizons Comput. Sci.* **2001**, *263*, 153–168.

[9] Chan, H. S.; Dill, K. A. *Phys. Today* **1993**, *46*, 24–32.

[10] Dymond, J. H.; Smith, E. B. *The Virial Coefficients of Pure Gases and Mixtures*, 2nd ed.; Clarendon Press: Berlin Heidelberg, 1980.

[11] Rosen, N. *Phys. Rev. Lett.* **1931**, *38*, 255–276.

[12] Tai, H. *Phys. Rev. A* **1986**, *33*, 3657–3666.

[13] Duška, M.; Hrubý, J. *EPJ Web Conf.* **2013**, *45*, 01024.

[14] Tillner-Roth, R.; Harms-Watzenberg, F.; Baehr, H. D. *Dkv Tragungsbericht* **1993**, *20*, 67.

[15] Massucci, M.; Wormald, C. *J. Chem. Thermodyn.* **1998**, *30*, 919–927.

[16] Hellmann, R. *J. Chem. Phys.* **2017**, *146*, 054302.

[17] Kalugina, Y. N.; Buryak, I. A.; Ajili, Y.; Vigasin, A. A.; Jaidane, N. E.; Hochlaf, M. *J. Chem. Phys.* **2014**, *140*, 234310.

[18] van Gunsteren, W. F.; Berendsen, H. J. C. *Angew. Chemie Int. Ed. English* **1990**, *29*, 992–1023.

[19] Hospital, A.; Goñi, J. R.; Orozco, M.; Gelpi, J. *Adv. Appl. Bioinforma. Chem.* **2015**, *8*, 37–47.

[20] Chen, L.-Q.; Chen, L.-D.; Kalinin, S. V.; Klimeck, G.; Kumar, S. K.; Neugebauer, J.; Terasaki, I. *npj Comput. Mater.* **2015**, 1–2.

[21] Karplus, M.; McCammon, J. A. *Nat. Struct. Biol.* **2002**, *9*, 646–652.

[22] Ciccotti, G.; Ferrario, M.; Schuette, C. *Molecular Dynamics Simulation*; 2014; pp 1–617.

[23] Warshel, A. *Annu. Rev. Biophys. Biomol. Struct.* **2003**, *32*, 425–443.

[24] Levitt, M.; Warshel, A. *Nature* **1975**, *253*, 694–698.

[25] Lane, T. J.; Shukla, D.; Beauchamp, K. A.; Pande, V. S. *Curr. Opin. Struct. Biol.* **2013**, *23*, 58–65.

[26] Piana, S.; Klepeis, J. L.; Shaw, D. E. *Curr. Opin. Struct. Biol.* **2014**, *24*, 98–105.

[27] Perez, A.; Morrone, J. A.; Simmerling, C.; Dill, K. A. *Curr. Opin. Struct. Biol.* **2016**, *36*, 25–31.

[28] De Vivo, M.; Masetti, M.; Bottegoni, G.; Cavalli, A. *J. Med. Chem.* **2016**, *59*, 4035–4061.

[29] Jiang, J.; Babarao, R.; Hu, Z. *Chem. Soc. Rev.* **2011**, *40*, 3599–3612.

[30] Maurin, G. *Chem. Met. Fram. Synth. Charact. Appl.* **2016**, 765.

[31] Bereau, T.; Andrienko, D.; Kremer, K. *APL Mater.* **2016**, *4*.

[32] Kalikmanov, V. I. *Nucleation theory*; 2013; p 316.

[33] Wilding, N. *Am. J. Phys.* **2001**, *69*, 1147.

[34] Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Oxford Science Publ; Clarendon Press, 1989.

[35] Karplus, M. *Angew. Chemie Int. Ed.* **2014**, *53*, 9992–10005.

[36] Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. *J. Chem. Phys.* **1953**, *21*, 1087–1092.

[37] Harrison, R. L.; Granja, C.; Leroy, C. Introduction to Monte Carlo Simulation. AIP Conf. Proc. 2010; pp 17–21.

[38] Boltzmann, L. *Vorlesungen {ü}ber Gastheorie: Th. Theorie van der Waals'; Gase mit zusammengesetzten Molek{ü}len; Gasdissociation; Schlussbemerkungen*; Vorlesungen {ü}ber Gastheorie; J. A. Barth, 1898.

[39] Schneider, G.; Fechner, U. *Nat. Rev. Drug Discov.* **2005**, *4*, 649–663.

[40] Jorgensen, W. L. *Science* **2004**, *303*, 1813–8.

[41] Ballone, P. *Entropy* **2014**, *16*, 322–349.

[42] Lopes, P. E. M.; Guvench, O.; MacKerell, A. D. *Methods*; 2015; Vol. 1215; pp 47–71.

[43] Saunders, M. G.; Voth, G. a. *Annu. Rev. Biophys.* **2013**, *42*, 73–93.

[44] De Carvalho, F. F.; Bouduban, M. E. F.; Curchod, B. F. E.; Tavernelli, I. *Entropy* **2014**, *16*, 62–85.

[45] Lei, H.; Duan, Y. *Curr. Opin. Struct. Biol.* **2007**, *17*, 187–191.

[46] Grossfield, A.; Zuckerman, D. M. *Annu Rep Comput Chem* **2009**, *1400*, 23–48.

[47] Theodorou, D. N. *Ind. Eng. Chem. Res.* **2010**, *49*, 3047–3058.

[48] E, W.; Vanden-Eijnden, E. *Annu. Rev. Phys. Chem.* **2010**, *61*, 391–420.

[49] Pande, V. S.; Beauchamp, K.; Bowman, G. R. *Methods* **2010**, *52*, 99–105.

[50] Rohrdanz, M. a.; Zheng, W.; Clementi, C. *Annu. Rev. Phys. Chem.* **2013**, *64*, 295–316.

[51] Chaplin, M. Protein Folding and Denaturation. 2017; `http://www1.lsbu.ac.uk/water/protein{_}denatured.html`.

[52] Cramer, C. J. *Essentials Comput. Chem.*; 2004; Vol. 42; pp 334–342.

[53] Freddolino, P. L.; Harrison, C. B.; Liu, Y.; Schulten, K. *Nat. Phys.* **2010**, *6*, 751–758.

[54] Stone, A. J.; Misquitta, A. J. *Int. Rev. Phys. Chem.*; 2007; Vol. 26; pp 193–222.

[55] Caleman, C.; Hong, M.; Costa, L. T.; Maaren, P. J. V.; Hub, J. S. *Cell* **2011**, 61–74.

[56] Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Proteins Struct. Funct. Bioinforma.* **2006**, *65*, 712–725.

[57] Rauscher, S.; Gapsys, V.; Gajda, M. J.; Zweckstetter, M.; de Groot, B. L.; Grubmüller, H. *J. Chem. Theory Comput.* **2015**, 151009180807001.

[58] Cisneros, G. A.; Wikfeldt, K. T.; Ojamäe, L.; Lu, J.; Xu, Y.; Torabifard, H.; Bartók, A. P.; Csányi, G.; Molinero, V.; Paesani, F. *Chem. Rev.* **2016**, *116*, 7501–7528.

[59] Jorgensen, W. L.; Tirado-Rives, J. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 6665–6670.

[60] Parker, T. M.; Sherrill, C. D. *J. Chem. Theory Comput.* **2015**, *11*, 4197–4204.

[61] Sherrill, C. D.; Sumpter, B. G.; Sinnokrot, M. O.; Marshall, M. S.; Hohenstein, E. G.; Walker, R. C.; Gould, I. R. *J. Comput. Chem.* **2009**, *30*, 2187–2193.

[62] Zgarbová, M.; Otyepka, M.; Sponer, J.; Hobza, P.; Jurecka, P. *Phys. Chem. Chem. Phys.* **2010**, *12*, 10476–10493.

[63] Cieplak, P.; Dupradeau, F.-Y.; Duan, Y.; Wang, J. *J. Phys. Condens. Matter* **2009**, *21*, 333102.

[64] Cardamone, S.; Hughes, T. J.; Popelier, P. L. a. *Phys. Chem. Chem. Phys.* **2014**, *16*, 10367.

[65] Abrahamson, A. A. *Phys. Rev.* **1963**, *130*, 693–707.

[66] Mackerell, A. D. *J. Comput. Chem.* **2004**, *25*, 1584–1604.

[67] Albaugh, A. et al. *J. Phys. Chem. B* **2016**, acs.jpcb.6b06414.

[68] Simmonett, A. C.; Pickard, F. C.; Shao, Y.; Cheatham, T. E.; Brooks, B. R. *J. Chem. Phys.* **2015**, *143*, 074115.

[69] Simmonett, A. C.; Pickard, F. C.; Ponder, J. W.; Brooks, B. R. *J. Chem. Phys.* **2016**, *145*, 164101.

[70] Demerdash, O.; Yap, E.-H.; Head-Gordon, T. *Annu. Rev. Phys. Chem.* **2014**, *65*, 149–74.

[71] Hawkins, D. M. **2004**, 1–12.

[72] Johnson, E. R.; Mackie, I. D.; DiLabio, G. a. *J. Phys. Org. Chem.* **2009**, *22*, 1127–1135.

[73] Taylor, D. E. et al. *J. Chem. Phys.* **2016**, *145*, 124105.

[74] Chalasinski, G.; Szczesniak, M. M. *Chem. Rev.* **2000**, *100*, 4227–4252.

[75] Schmidt, J. R.; Yu, K.; McDaniel, J. G. *Acc. Chem. Res.* **2015**, *48*, 548–556.

[76] Elrod, M. J.; Saykally, R. J. *Chem. Rev.* **1994**, *94*, 1975–1997.

[77] Rick, S. W.; Stuart, S. J. *Potentials and Algorithms for Incorporating Polarizability in Computer Simulations*; 2002; Vol. 18.

[78] Stone, A. J. *The Theory of Intermolecular Forces*, 2nd ed.; OUP Oxford, 2013.

[79] Szalewicz, K. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2012**, *2*, 254–272.

[80] Jeziorski, B.; Moszynski, R.; Szalewicz, K. *Chem. Rev.* **1994**, *94*, 1887–1930.

[81] McDaniel, J. G. Development and Application of Physically-Motivated First-Principles Force Fields for Complex Chemical Systems. Ph.D. thesis, UW-Madison, 2014.

[82] McDaniel, J. G.; Schmidt, J. R. *Annu. Rev. Phys. Chem.* **2016**, *67*, 467–488.

[83] McDaniel, J. G.; Schmidt, J. R. *J. Phys. Chem. A* **2013**, *117*, 2053–2066.

[84] Misquitta, A. J.; Stone, A. J. CamCASP: a program for studying intermolecular interactions and for the calculation of molecular properties in distributed form, version 5.8. University of Cambridge, 2015.

[85] Metz, M. P.; Piszczatowski, K.; Szalewicz, K. *J. Chem. Theory Comput.* **2016**, acs.jctc.6b00913.

[86] Stone, A. J. *Chem. Phys. Lett.* **1981**, *83*, 233–239.

[87] Stone, A. J. *J. Chem. Theory Comput.* **2005**, *1*, 1128–1132.

[88] Misquitta, A. J.; Stone, A. J. *J. Chem. Phys.* **2006**, *124*, 024111.

[89] Williams, G. J.; Stone, A. J. *J. Chem. Phys.* **2003**, *119*, 4620–4628.

[90] Misquitta, A. J.; Stone, A. J. *Mol. Phys.* **2008**, *106*, 1631–1643.

[91] Misquitta, A. J.; Stone, A. J.; Fazeli, F. *J. Chem. Theory Comput.* **2014**, *10*, 5405–5418.

[92] Lillestolen, T. C.; Wheatley, R. J. *Chem. Commun.* **2008**, *7345*, 5909–5911.

[93] Lillestolen, T. C.; Wheatley, R. J. *J. Chem. Phys.* **2009**, *131*, 144101.

[94] Hirshfeld, F. L. *Theor. Chim. Acta* **1977**, *44*, 129–138.

[95] Van Vleet, M. J.; Misquitta, A. J.; Stone, A. J.; Schmidt, J. R. *J. Chem. Theory Comput.* **2016**, *12*, 3851–3870.

[96] Verstraelen, T.; Vandenbrande, S.; Vanduyfhuys, L.; Heidar-Zadeh, F.; Ayers, P. W.; Waroquier, M.; Van Speybroeck, V. **2016**, *XXX*, in preperation.

[97] Vandenbrande, S.; Waroquier, M.; Speybroeck, V. V.; Verstraelen, T. **2016**,

[98] Verstraelen, T.; Vandenbrande, S.; Ayers, P. W. *J. Chem. Phys.* **2014**, *141*, 194114.

[99] Misquitta, A. J.; Stone, A. J. *J. Chem. Theory Comput.* **2016**, *12*, 4184–4208.

[100] Margenau, H.; Kestner, N. R. *Theory of Intermolecular Forces*; International series of monographs in natural philosophy; Pergamon Press: Oxford, 1969.

[101] Riley, K. E.; Pitonˇcák, M.; Jurecˇcka, P.; Hobza, P. *Chem. Rev.* **2010**, *110*, 5023–5063.

[102] Dykstra, C. E.; Lisy, J. M. *J. Mol. Struct. THEOCHEM* **2000**, *500*, 375–390.

[103] Stone, A. J.; Tough, R. *Chem. Phys. Lett.* **1984**, *110*, 123–129.

[104] Dehez, F.; Chipot, C.; Millot, C.; Ángyán, J. G. *Chem. Phys. Lett.* **2001**, *338*, 180–188.

[105] Raghavachari, K.; Trucks, G. W.; Pople, J. A.; Head-Gordon, M. *Chem. Phys. Lett.* **1989**, *157*, 479–483.

[106] Grimme, S.; Djukic, J. P. *Inorg. Chem.* **2011**, *50*, 2619–2628.

[107] Lennard-Jones, J. *Proc. Phys. Soc.* **1931**, *43*, 461–482.

[108] Born, M.; Mayer, J. E. *Zeitschrift für Phys.* **1932**, *75*, 1–18.

[109] Buckingham, R. A. *Proc. R. Soc. A Math. Phys. Eng. Sci.* **1938**, *168*, 264–283.

[110] Nezbeda, I. *Mol. Phys.* **2005**, *103*, 59–76.

[111] Galliero, G.; Boned, C. *J. Chem. Phys.* **2008**, *129*.

[112] Gordon, P. A. *J. Chem. Phys.* **2006**, *125*.

[113] Ruckenstein, E.; Liu, H. *Society* **1997**, 3927–3936.

[114] Galliero, G.; Boned, C.; Baylaucq, A.; Montel, F. *Chem. Phys.* **2007**, *333*, 219–228.

[115] Wu, G.-W.; Sadus, R. J. *Fluid Phase Equilib.* **2000**, *170*, 269–284.

[116] Errington, J. R.; Panagiotopoulos, A. Z. *J. Chem. Phys.* **1998**, *109*, 1093–1100.

[117] McGrath, M. J.; Ghogomu, J. N.; Tsona, N. T.; Siepmann, J. I.; Chen, B.; Napari, I.; Vehkamaki, H. *J. Chem. Phys.* **2010**, *133*.

[118] Bastea, S. *Phys. Rev. E. Stat. Nonlin. Soft Matter Phys.* **2003**, *68*, 031204.

[119] Errington, J. R.; Panagiotopoulos, A. Z. *J. Phys. Chem. B* **1999**, *103*, 6314–6322.

[120] Ross, M.; Ree, F. H. *J. Chem. Phys.* **1980**, *73*, 6146.

[121] Halgren, T. A. *J. Am. Chem. Soc.* **1992**, *114*, 7827–7843.

[122] Kim, Y. S.; Kim, S. K.; Lee, W. D. *Chem. Phys. Lett.* **1981**, *80*, 574–575.

[123] Misquitta, A. J.; Stone, A. J. Ab initio atom-atom potentials using CamCASP: Theory. 2015; `https://arxiv.org/abs/1512.06150v2`.

[124] Nyeland, C.; Toennies, J. P. *Chem. Phys. Lett.* **1986**, *127*, 3–8.

[125] Ihm, G.; Cole, M. W.; Toigo, F.; Klein, J. R. *Phys. Rev. A* **1990**, *42*, 5244–5252.

[126] Duke, R. E.; Starovoytov, O. N.; Piquemal, J.-P.; Cisneros, G. A. *J. Chem. Theory Comput.* **2014**, *10*, 1361–1365.

[127] Elking, D. M.; Cisneros, G. A.; Piquemal, J. P.; Darden, T. A.; Pedersen, L. G. *J. Chem. Theory Comput.* **2010**, *6*, 190–202.

[128] Chaudret, R.; Gresh, N.; Narth, C.; Lagardere, L.; Darden, T. A.; Cisneros, G. A.; Piquemal, J. P. *J. Phys. Chem. A* **2014**, *118*, 7598–7612.

[129] Chaudret, R.; Gresh, N.; Cisneros, G. A.; Scemama, A.; Piquemal, J.-p. *Can. J. Chem.* **2013**, *91*, 804–810.

[130] Öhrn, A.; Hermida-Ramon, J. M.; Karlström, G. *J. Chem. Theory Comput.* **2016**, *12*, 2298–2311.

[131] Gresh, N.; Cisneros, G. A.; Darden, T. A.; Piquemal, J.-P. *J. Chem. Theory Comput.* **2007**, *3*, 1960–1986.

[132] Gordon, M. S.; Freitag, M. A.; Bandyopadhyay, P.; Jensen, J. H.; Kairys, V.; Stevens, W. J. *J. Phys. Chem. A* **2001**, *105*, 293–307.

[133] Xie, W.; Gao, J. *J. Chem. Theory Comput.* **2007**, *3*, 1890–1900.

[134] Xie, W.; Orozco, M.; Truhlar, D. G.; Gao, J. *J. Chem. Theory Comput.* **2009**, *5*, 459–467.

[135] Patil, S. H.; Tang, K. T. In *Asymptotic methods in quantum mechanics*; Scäfer, F. P., Toennies, J. P., Zinth, W., Eds.; Springer, 2000.

[136] Hoffmann-Ostenhof, M.; Hoffmann-Ostenhof, T. *Phys. Rev. A* **1977**, *16*, 1782–1785.

[137] Amovilli, C.; March, N. H. *J. Phys. A. Math. Gen.* **2006**, *39*, 7349–7357.

[138] Bunge, A. V.; Esquivel, R. O. *Phys. Rev. A* **1986**, *34*, 853.

[139] Rappe, A. K.; Casewit, C.; Colwell, K.; Goddard, W. A. I.; Skiff, W. *J. Am. Chem. Soc.* **1992**, *114*, 10024–10035.

[140] Waldman, M.; Hagler, A. T. *J. Comput. Chem.* **1993**, *14*, 1077–1084.

[141] McDaniel, J. G.; Schmidt, J. R. *J. Phys. Chem. C* **2012**, *116*, 14031–14039.

[142] Podeszwa, R.; Bukowski, R.; Szalewicz, K. *J. Phys. Chem. A* **2006**, *110*, 10345–10354.

[143] Bukowski, R.; Szalewicz, K.; Groenenboom, G.; van der Avoird, A. *J. Chem. Phys.* **2006**, *125*, 044301.

[144] Jeziorska, M.; Cencek, W.; Patkowski, K.; Jeziorski, B.; Szalewicz, K. *J. Chem. Phys.* **2007**, *127*, 124303.

[145] Sum, A. K.; Sandler, S. I.; Bukowski, R.; Szalewicz, K. *J. Chem. Phys.* **2002**, *116*, 7637.

[146] Konieczny, J. K.; Sokalski, W. A. *J. Mol. Model.* **2015**, *21*, 197.

[147] Kita, S.; Noda, K.; Inouye, H. *J. Chem. Phys.* **1976**, *64*, 3446–3449.

[148] Kuechler, E. R.; Giese, T. J.; York, D. M. *J. Chem. Phys.* **2015**, *143*.

[149] Giese, T. J.; York, D. M. *J. Chem. Phys.* **2007**, *127*, 194101.

[150] Giese, T. J.; Chen, H.; Dissanayake, T.; GiambaÅŸu, G. M.; Heldenbrand, H.; Huang, M.; Kuechler, E. R.; Lee, T.-S.; Panteva, M. T.; Radak, B. K.; York, D. M. *J. Chem. Theory Comput.* **2013**, *9*, 1417–1427.

[151] Day, G. M.; Price, S. L. *J. Am. Chem. Soc.* **2003**, *125*, 16434–16443.

[152] Nobeli, I.; Price, S. L.; Wheatley, R. J. *Mol. Phys.* **1998**, *95*, 525–537.

[153] Totton, T. S.; Misquitta, A. J.; Kraft, M. *J. Chem. Theory Comput.* **2010**, *6*, 683–695.

[154] Misquitta, A. J. *J. Chem. Theory Comput.* **2013**, *9*, 5313–5326.

[155] Misquitta, A. J.; Stone, A. J. Ab initio atom-atom potentials using CamCASP: Application to Pyridine. 2015; `https://arxiv.org/abs/1512.06155v2`.

[156] Tang, K. T.; Toennies, J. P. *J. Chem. Phys.* **1984**, *80*, 3726–3741.

[157] Tang, K. T.; Toennies, J. P. *Surf. Sci.* **1992**, *279*, L203–L206.

[158] Wheatley, R. J.; Price, S. L. *Mol. Phys.* **1990**, *69*, 507–533.

[159] Mitchell, J. B. O.; Price, S. L. *J. Phys. Chem. A* **2000**, *104*, 10958–10971.

[160] Söderhjelm, P.; Karlström, G.; Ryde, U. *J. Chem. Phys.* **2006**, *124*, 244101.

[161] Tkatchenko, A.; Distasio, R. A.; Car, R.; Scheffler, M. *Phys. Rev. Lett.* **2012**, *108*, 1–5.

[162] Tkatchenko, A.; Scheffler, M. *Phys. Rev. Lett.* **2009**, *102*, 6–9.

[163] Cole, D. J.; Vilseck, J. Z.; Tirado-Rives, J.; Payne, M. C.; Jorgensen, W. L. *J. Chem. Theory Comput.* **2016**, *12*, 2312–2323.

[164] Manz, T. A.; Sholl, D. S. *J. Chem. Theory Comput.* **2010**, *6*, 2455–2468.

[165] Manz, T. A.; Sholl, D. S. *J. Chem. Theory Comput.* **2012**, *8*, 2844–2867.

[166] Yu, K.; McDaniel, J. G.; Schmidt, J. R. *J. Phys. Chem. B* **2011**, *115*, 10054–10063.

[167] Levy, M.; Perdew, J. P.; Sahni, V. *Phys. Rev. A* **1984**, *30*, 2745–2748.

[168] Kitaigorodsky, A. *Molecular crystals and Molecules*; Physical Chemistry; Elsevier Science: New York, 2012.

[169] Misquitta, A. J.; Szalewicz, K. *Chem. Phys. Lett.* **2002**, *357*, 301–306.

[170] Misquitta, A. J.; Jeziorski, B.; Szalewicz, K. *Phys. Rev. Lett.* **2003**, *91*, 033201.

[171] Misquitta, A. J.; Podeszwa, R.; Jeziorski, B.; Szalewicz, K. *J. Chem. Phys.* **2005**, *123*.

[172] Heßelmann, A.; Jansen, G.; SchuÌˆtz, M. *J. Chem. Phys.* **2005**, *122*, 014103.

[173] Podeszwa, R.; Bukowski, R.; Szalewicz, K. *J. Chem. Theory Comput.* **2006**, *2*, 400–412.

[174] Heßelmann, A.; Jansen, G. *Chem. Phys. Lett.* **2002**, *362*, 319–325.

[175] Heßelmann, A.; Jansen, G. *Chem. Phys. Lett.* **2003**, *367*, 778–784.

[176] Heßelmann, A.; Jansen, G. *Chem. Phys. Lett.* **2002**, *357*, 464–470.

[177] Jansen, G.; Hesselmann, A.; Williams, H. L.; Chabalowski, C. F. *J. Phys. Chem. A* **2001**, *105*, 11156–11158.

[178] Podeszwa, R.; Szalewicz, K. Accurate interaction energies from perturbation theory based on Kohn-Sham model. 2005; `http://arxiv.org/abs/physics/0501023`.

[179] Jeziorska, M.; Jeziorski, B.; Čížek, J. *Int. J. Quantum Chem.* **1987**, *32*, 149–164.

[180] Drude, P.; Riborg, C.; Millikan, R. A. *The Theory of Optics... Translated from the German by CR Mann and RA Millikan*; London; New York [printed], 1902.

[181] Lamoureux, G.; Roux, B. *J. Chem. Phys.* **2003**, *119*, 3025.

[182] NIST Computational Chemistry Comparison and Benchmark Database, NIST Standard Reference Database Number 101. 2015; `http://cccbdb.nist.gov/`.

[183] Shoemake, K. *Graph. Gems 3*; 1992; Chapter 6, pp 124–132.

[184] Werner, H.-J.; Knowles, P. J.; Knizia, G.; Manby, F. R.; Schütz, M. *WIREs Comput Mol Sci* **2012**, *2*, 242–253.

[185] Aidas, K. et al. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2014**, *4*, 269–284.

[186] Stone, A. J.; Dullweber, A.; Engkvist, O.; Fraschini, E.; Hodges, M. P.; Meredith, A. W.; Nutt, D. R.; Popelier, P. L. A.; Wales, D. J. ORIENT: a program for studying interactions between molecules, version 4.8. 2015; `http://www-stone.ch.cam.ac.uk/programs/orient.html`.

[187] Ferenczy, G. G.; Winn, P. J.; Reynolds, C. a. *J. Phys. Chem. A* **1997**, *101*, 5446–5455.

[188] Eastman, P. et al. *J. Chem. Theory Comput.* **2013**, *9*, 461–469.

[189] Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225–11236.

[190] Sebetci, A.; Beran, G. J. O. *J. Chem. Theory Comput.* **2010**, *6*, 155–167.

[191] Misquitta, A.; Welch, G.; Stone, A.; Price, S. **2008**, *456*, 105–109.

[192] Price, S. L.; Leslie, M.; Welch, G. W. A.; Habgood, M.; Price, L. S.; Karamertzanis, P. G.; Day, G. M. *Phys. Chem. Chem. Phys.* **2010**, *12*, 8478–8490.

[193] Totton, T. S.; Misquitta, A. J.; Kraft, M. *J. Chem. Theory Comput.* **2010**, *6*, 683–695.

[194] Hermida-Ramón, J. M.; Ríos, M. A. *Chem. Phys.* **2000**, *262*, 423–436.

[195] Nyeland, C. *Chem. Phys.* **1990**, *147*, 229–240.

[196] Vogel, E.; Jäger, B.; Hellmann, R.; Bich, E. *Mol. Phys.* **2010**, *108*, 3335–3352.

[197] McDaniel, J. G.; Yu, K.; Schmidt, J. R. *J. Phys. Chem. C* **2012**, *116*, 1892–1903.

[198] Mayo, S. L.; Olafson, B. D.; Goddard, W. A. I. *J. Phys. Chem.* **1990**, *101*, 8897–8909.

[199] Lim, T. *Zeitschrift für Naturforschung-A* **2009**, *64*, 200–204.

[200] Van Duin, a. C. T.; Dasgupta, S.; Lorant, F.; Goddard, W. a. *J. Phys. Chem. A* **2001**, *105*, 9396–9409.

[201] Eramian, H.; Tian, Y.-H.; Fox, Z.; Beneberu, H. Z.; Kertesz, M. *J. Phys. Chem. A* **2013**, *117*, 14184–14190.

[202] Badenhoop, J. K.; Weinhold, F. *J. Chem. Phys.* **1997**, *107*, 5422.

[203] Kim, H.; Doan, V. D.; Cho, W. J.; Madhav, M. V.; Kim, K. S. *Sci. Rep.* **2014**, *4*, 1–8.

[204] Wheatley, R. J.; Gopal, A. A. *Phys. Chem. Chem. Phys.* **2012**, *14*, 2087–2091.

[205] Rezáč, J.; Hobza, P. *Chem. Rev.* **2016**, *116*, 5038–5071.

[206] Jan, S.; Martin, M. L.; Chem, P.; Phys, C. *Phys. Chem. Chem. Phys.* **2016**, *18*, 20905–20925.

[207] Hassanali, A. A.; Cuny, J.; Verdolino, V.; Parrinello, M. **2014**,

[208] Stone, A. J.; Price, S. L. *J. Phys. Chem.* **1988**, *92*, 3325–3335.

[209] Price, S. L. *Rev. Comput. Chem.* **2000**, *14*, 225–289.

[210] Coppens, P.; Guru Row, T. N.; Leung, P.; Stevens, E. D.; Becker, P. J.; Yang, Y. W. *Acta Crystallogr. Sect. A* **1979**, *35*, 63–72.

[211] Bondi, A. *J. Phys. Chem.* **1964**, *68*, 441–451.

[212] Nyburg, S. C.; Faerman, C. H. *Acta Crystallogr. Sect. B Struct. Sci.* **1985**, *B41*, 274–279.

[213] Batsanov, S. S. *Inorg. Mater. Transl. from Neorg. Mater. Orig. Russ. Text* **2001**, *37*, 871–885.

[214] Auffinger, P.; Hays, F. A.; Westhof, E.; Ho, P. S. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101*, 16789–94.

[215] Lommerse, J. P. M.; Stone, A. J.; Taylor, R.; Allen, F. H. *J. Am. Chem. Soc.* **1996**, *118*, 3108–3116.

[216] Kramer, C.; Spinn, A.; Liedl, K. R. *J. Chem. Theory Comput.* **2014**, *10*, 4488–4496.

[217] Badenhoop, J. K.; Weinhold, F. *J. Chem. Phys.* **1997**, *107*, 5422.

[218] Bankiewicz, B.; Palusiak, M. *Struct. Chem.* **2012**, *24*, 1297–1306.

[219] Chessari, G.; Hunter, C. A.; Low, C. M. R.; Packer, M. J.; Vinter, J. G.; Zonta, C. *Chem. - A Eur. J.* **2002**, *8*, 2860–2867.

[220] Šponer, J.; Šponer, J. E.; Mládek, A.; Jurečka, P.; Banáš, P.; Otyepka, M. *Biopolymers* **2013**, *99*, 978–988.

[221] Bartocci, A.; Belpassi, L.; Cappelletti, D.; Falcinelli, S.; Grandinetti, F.; Tarantelli, F.; Pirani, F. *J. Chem. Phys.* **2015**, *142*, 184304.

[222] Rendine, S.; Pieraccini, S.; Forni, A.; Sironi, M. *Phys. Chem. Chem. Phys.* **2011**, *13*, 19508–19516.

[223] Politzer, P.; Murray, J. S.; Concha, M. C. *J. Mol. Model.* **2008**, *14*, 659–665.

[224] Hagler, A. T. *J. Chem. Theory Comput.* **2015**, *11*, 5555–5572.

[225] Ren, P.; Ponder, J. W. *J. Phys. Chem. B* **2003**, *107*, 5933–5947.

[226] Ponder, J. W.; Wu, C.; Pande, V. S.; Chodera, J. D.; Schnieders, M. J.; Haque, I.; Mobley, D. L.; Lambrecht, D. S.; Distasio, R. A.; Head-gordon, M.; Clark, G. N. I.; Johnson, M. E.; Head-gordon, T. *J. Phys. Chem. B* **2010**, *114*, 2549–2564.

[227] Shi, Y.; Xia, Z.; Zhang, J.; Best, R.; Wu, C.; Ponder, J. W.; Ren, P. *J. Chem. Theory Comput.* **2013**, *9*, 4046–4063.

[228] Price, S. L.; Leslie, M.; Welch, G. W. A.; Habgood, M.; Price, L. S.; Karamertzanis, P. G.; Day, G. M. *Phys. Chem. Chem. Phys.* **2010**, *12*, 8478.

[229] Day, G. M.; Motherwell, W. D. S.; Jones, W. *Cryst. Growth Des.* **2005**, *5*, 1023–1033.

[230] Price, S. L. *Int. Rev. Phys. Chem.* **2008**, *27*, 541–568.

[231] Phipps, M. J. S.; Fox, T.; Tautermann, C. S.; Skylaris, C.-K. *Chem. Soc. Rev.* **2015**, *44*, 3177–3211.

[232] Williams, D. E. *J. Comput. Chem.* **1988**, *9*, 745–763.

[233] Chaudret, R.; Gresh, N.; Narth, C.; Lagarde, L.; Darden, T. A.; Andre, G.; Piquemal, J.-p. **2014**,

[234] Cisneros, G. A.; Piquemal, J. P.; Darden, T. A. *J. Chem. Phys.* **2006**, *125*.

[235] Dixon, R. W.; Kollman, P. a. *J. Comput. Chem.* **1997**, *18*, 1632–1646.

[236] Harder, E.; Anisimov, V. M.; Vorobyov, I. V.; Lopes, P. E. M.; Noskov, S. Y.; Jr, A. D. M. **2006**, 1587–1597.

[237] Mu, X.; Wang, Q.; Wang, L.-P.; Fried, S. D.; Piquemal, J.-P.; Dalby, K. N.; Ren, P. *J. Phys. Chem. B* **2014**, *118*, 6456–6465.

[238] Wikfeldt, K. T.; Batista, E. R.; Vila, F. D.; Jónsson, H. *Phys. Chem. Chem. Phys.* **2013**, *15*, 16542–56.

[239] Piquemal, J. P.; Chelli, R.; Procacci, P.; Gresh, N. *J. Phys. Chem. A* **2007**, *111*, 8170–8176.

[240] Loboda, O.; Ingrosso, F.; Ruiz-López, M. F.; Szalewicz, K.; Millot, C. *J. Chem. Phys.* **2016**, *144*.

[241] Langhoff, P. W. *J. Chem. Phys.* **1971**, *55*, 2126.

[242] Krishtal, A.; Vannomeslaeghe, K.; Geldof, D.; Van Alsenoy, C.; Geerlings, P. *Phys. Rev. A - At. Mol. Opt. Phys.* **2011**, *83*, 3–6.

[243] Stone, A. J. *J. Am. Chem. Soc.* **2013**, *135*, 7005–7009.

[244] Duke, R. E.; Starovoytov, O. N.; Piquemal, J.-p.; Andre, G. **2014**,

[245] Gavezzotti, A. *J. Phys. Chem. B* **2003**, *107*, 2344–2353.

[246] Torheyden, M.; Jansen, G. *Mol. Phys.* **2006**, *104*, 2101–2138.

[247] Mitchell, J. B. O.; Price, S. L.; Leslie, M.; Buttar, D.; Roberts, R. J. *J. Phys. Chem. A* **2001**, *105*, 9961–9971.

[248] Misquitta, A. J.; Stone, A. J.; Price, S. L. *J. Chem. Theory Comput.* **2008**, *4*, 19–32.

[249] Holt, A.; Karlstr?m, G. *J. Comput. Chem.* **2008**, *29*, 2033–2038.

[250] Holt, A.; Boström, J.; Karlström, G.; Lindh, R. **2010**,

[251] Parker, T. M.; Burns, L. a.; Parrish, R. M.; Ryno, A. G.; Sherrill, C. D. *J. Chem. Phys.* **2014**, *140*, 094106.

[252] Werner, H.-J. et al. MOLPRO, version 2012.1, a package of ab initio programs. 2012.

[253] Oakley, M. T.; Wheatley, R. J. *J. Chem. Phys.* **2009**, *130*, 034110.

[254] Korona, T. *Theor. Chem. Acc.* **2011**, *129*, 15–30.

[255] Giauque, W. F.; Egan, C. J. *J. Chem. Phys.* **1937**, *5*, 45.

[256] Cervinka, C.; Fulem, M. *J. Chem. Theory Comput.* **2017**, *13*, 2840–2850.

[257] Heit, Y. N.; Nanda, K. D.; Beran, G. J. O. *Chem. Sci.* **2016**, *7*, 246–255.

[258] Simon, A.; Peters, K. *Acta Crystallogr. Sect. B* **1980**, *36*, 2750–2751.

[259] Yu, K.; Schmidt, J. R. *J. Chem. Phys.* **2012**, *136*, 034503.

[260] Misquitta, A. J.; Stone, A. J. *J. Chem. Theory Comput.* **2007**, 7–18.

[261] Liu, C.; Qi, R.; Wang, Q.; Piquemal, J.-P.; Ren, P. *J. Chem. Theory Comput.* **2017**, acs.jctc.7b00225.

[262] Thole, B. T. *Chem. Phys.* **1981**, *59*, 341–350.

[263] Van Vleet, M. J.; Misquitta, A. J.; Schmidt, J. R. *J. Chem. Theory Comput.* **2017**, submitted.

[264] Knizia, G.; Adler, T. B.; Werner, H.-J. *J. Chem. Phys.* **2009**, *130*, 054104.

[265] Bukowski, R.; Sadlej, J.; Jeziorski, B.; Jankowski, P.; Szalewicz, K.; Kucharski, S. A.; Williams, H. L.; Rice, B. M. *J. Chem. Phys. Addit. Inf. J. Chem. Phys. J. Homepage* **1999**, *110*.

[266] Babin, V.; Leforestier, C.; Paesani, F. *J. Chem. Theory Comput.* **2013**, *9*, 5395–5403.

[267] Desgranges, C.; Delhommelle, J. **2015**,

[268] Pérez-Sánchez, G.; González-Salgado, D.; Piñeiro, M. M.; Vega, C. *J. Chem. Phys.* **2013**, *138*, 084506.

[269] Stone, A. J. *Mol. Phys.* **1978**, *36*, 241–256.

[270] Millot, C.; Stone, A. *Mol. Phys.* **1992**, *77*, 439–462.

[271] Furukawa, H.; Cordova, K. E.; O'Keeffe, M.; Yaghi, O. M. *Science (80-.).* **2013**, *341*, 1230444–1230444.

[272] Millward, A. R.; Yaghi, O. M. *J. Am. Chem. Soc.* **2005**, *127*, 17998–17999.

[273] Dietzel, P. D. C. et al. *J. Mater. Chem.* **2009**, *19*, 7362.

[274] Dzubak, A. L.; Lin, L.-C.; Kim, J.; Swisher, J. a.; Poloni, R.; Maximoff, S. N.; Smit, B.; Gagliardi, L. *Nat. Chem.* **2012**, *4*, 810–816.

[275] Czaja, A. U.; Trukhan, N.; Müller, U. *Chem. Soc. Rev.* **2009**, *38*, 1284.

[276] Krishna, R.; van Baten, J. M. *Phys. Chem. Chem. Phys.* **2011**, *13*, 10593–10616.

[277] Getman, R. B.; Bae, Y.-s.; Wilmer, C. E.; Snurr, R. Q.; Carlo, M. *Adsorpt. J. Int. Adsorpt. Soc.* **2012**, 703–723.

[278] Yazaydin, a. O.; Snurr, R. Q.; Park, T.-H.; Koh, K.; Liu, J.; Levan, M. D.; Benin, A. I.; Jakubczak, P.; Lanuza, M.; Galloway, D. B.; Low, J. J.; Willis, R. R. *J. Am. Chem. Soc.* **2009**, *131*, 18198–9.

[279] Valenzano, L.; Civalleri, B.; Chavan, S.; Palomino, G. T.; AreaÌ'n, C. O.; Bordiga, S. *J. Phys. Chem. C* **2010**, *114*, 11185–11191.

[280] Poloni, R.; Lee, K.; Berger, R. F.; Smit, B. **2014**,

[281] Lin, L.-c.; Lee, K.; Gagliardi, L.; Smit, B. *J. Chem. Theory Comput.* **2014**, *10*, 1477–1488.

[282] Haldoupis, E.; Borycz, J.; Shi, H.; Vogiatzis, K. D.; Bai, P.; Queen, W. L.; Gagliardi, L.; Siepmann, J. I. *J. Phys. Chem. C* **2015**, *119*, 16058–16071.

[283] Mercado, R.; Vlaisavljevich, B.; Lin, L.-C.; Lee, K.; Lee, Y.; Mason, J. A.; Xiao, D. J.; Gonzalez, M. I.; Kapelewski, M. T.; Neaton, J. B.; Smit, B. *J. Phys. Chem. C* **2016**, acs.jpcc.6b03393.

[284] Becker, T. M.; Heinen, J.; Dubbeldam, D.; Lin, L.-C.; Vlugt, T. J. H. *J. Phys. Chem. C* **2017**, acs.jpcc.6b12052.

[285] McDaniel, J. G.; Li, S.; Tylianakis, E.; Snurr, R. Q.; Schmidt, J. R. *J. Phys. Chem. C* **2015**, *119*, 3143–3152.

[286] Lao, K. U.; Schaeffer, R.; Jansen, G.; Herbert, J. M. *J. Chem. Theory Comput.* **2015**, 150417132228001.

[287] Pastorczak, E.; Corminboeuf, C. *J. Chem. Phys.* **2017**, *146*, 120901.

[288] Żuchowski, P. *Chem. Phys. Lett.* **2008**, *450*, 203–209.

[289] Horn, P. R.; Head-gordon, M. *Phys. Chem. Chem. Phys.* **2016**, *18*, 23067–23079.

[290] Su, P.; Li, H. *J. Chem. Phys.* **2009**, *131*, 014102.

[291] Chen, Y.; Li, H. *J. Phys. Chem. A* **2010**, *114*, 11719–24.

[292] Su, P.; Jiang, Z.; Chen, Z.; Wu, W. *J. Phys. Chem. A* **2014**, *118*, 2531–42.

[293] Fedorov, D. G.; Kitaura, K. **2006**,

[294] Yu, K.; Kiesling, K.; Schmidt, J. R. **2012**,

[295] Verma, P.; Xu, X.; Truhlar, D. G. **2013**,

[296] Valenzano, L.; Civalleri, B.; Sillar, K.; Sauer, J. **2011**, 21777–21784.

[297] Haldoupis, E.; Borycz, J.; Shi, H.; Vogiatzis, K. D.; Bai, P.; Queen, W. L.; Gagliardi, L.; Siepmann, J. I. *J. Phys. Chem. C* **2015**, *74*, 150616135429005.

[298] Jansen, G.; Scha, R. **2012**,

[299] Frenkel, D.; Smit, B. *Acad. Press*; 2002; Vol. New York,; p 638.

[300] Guibas, L. Representing rotations with quaternions. 1992; `graphics.stanford.edu/courses/cs164-09-spring/Handouts/handout12.pdf`.

[301] Unke, O. T.; Devereux, M.; Meuwly, M. *J. Chem. Phys.* **2017**, *147*, 161712.

[302] Halgren, T. A. *Curr. Opin. Struct. Biol.* **1995**, *5*, 205–10.

[303] Wang, L.-P.; Martinez, T. J.; Pande, V. S. **2014**,

[304] Wildman, J.; Repiscák, P.; Paterson, M. J.; Galbraith, I. *J. Chem. Theory Comput.* **2016**, acs.jctc.5b01195.

[305] Wang, Q.; Rackers, J. A.; He, C.; Qi, R.; Narth, C.; Lagardere, L.; Gresh, N.; Ponder, J. W.; Piquemal, J. P.; Ren, P. *J. Chem. Theory Comput.* **2015**, *11*, 2609–2618.

[306] Freitag, M. a.; Gordon, M. S.; Jensen, J. H.; Stevens, W. J. *J. Chem. Phys.* **2000**, *112*, 7300.

[307] Wang, Q.; Rackers, J. a.; He, C.; Qi, R.; Narth, C.; Lagardere, L.; Gresh, N.; Ponder, J. W.; Piquemal, J.-P.; Ren, P. *J. Chem. Theory Comput.* **2015**, 150512142224003.

[308] SlipchenkoâŁ , L. V.; Gordon, M. S. *Mol. Phys.* **2009**, *107*, 999–1016.

[309] Gordon, M. S.; Smith, Q. A.; Xu, P.; Slipchenko, L. V. *Annu. Rev. Phys. Chem.* **2013**, *64*, 553–78.

[310] Mei, Y.; Simmonett, A. C.; Pickard, F. C.; DiStasio, R.; Brooks, B. R.; Shao, Y. *J. Phys. Chem. A* **2015**, 150506100905001.

[311] Lopes, P. E. M.; Roux, B.; MacKerell, A. D. *Theor. Chem. Acc.* **2009**, *124*, 11–28.

[312] Cisneros, G. A.; Wikfeldt, K. T.; Ojam??e, L.; Lu, J.; Xu, Y.; Torabifard, H.; Bart??k, A. P.; Cs??nyi, G.; Molinero, V.; Paesani, F. *Chem. Rev.* **2016**, *116*, 7501–7528.

[313] Welch, G. W. A.; Karamertzanis, P. G.; Misquitta, A. J.; Stone, A. J.; Price, S. L. *J. Chem. Theory Comput.* **2008**, *4*, 522–532.

[314] Wang, J.; Cieplak, P.; Li, J.; Hou, T.; Luo, R.; Duan, Y. *J. Phys. Chem. B* **2011**, *115*, 3091–3099.

[315] Wang, J.; Cieplak, P.; Cai, Q.; Hsieh, M.-J.; Wang, J.; Duan, Y.; Luo, R. *J. Phys. Chem. B* **2012**, *116*, 7999–8008.

[316] Bistoni, G.; Belpassi, L.; Tarantelli, F. *J. Chem. Theory Comput.* **2016**,

[317] Lao, K. U.; Herbert, J. M. *J. Chem. Theory Comput.* **2016**, acs.jctc.6b00155.

[318] Medders, G. R.; Götz, A. W.; Morales, M. A.; Bajaj, P.; Paesani, F. *J. Chem. Phys.* **2015**, *143*.

[319] Medders, G. R.; Babin, V.; Paesani, F. *J. Chem. Theory Comput.* **2013**, *9*, 1103–1114.