THE UNIVERSITY OF CHICAGO

THE SENSITIVITY AND REPRESENTABILITY OF COARSE-GRAINED MODELS

A DISSERTATION SUBMITTED TO

THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES

IN CANDIDACY FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

DEPARTMENT OF CHEMISTRY

BY

JACOB WILLIAM WAGNER

CHICAGO, ILLINOIS

JUNE 2017

For my wife Sabrina and my mother Joanne,

Table of Contents

List of Figures

List of Tables

Acknowledgements

First and foremost, I would like to thank Professor Gregory A. Voth for serving as my research advisor. He provided an environment in which I able to study interesting problems relating to coarse-graining. Also, he often helped by providing direction and motivation for my projects. Also, I would also like to thank Professor Benoit Roux and Professor Timothy Berkelbach for servings as committee members for the defense of this thesis.

Additionally, I would like to thank those who have collaborated with me on the work presented in this thesis. In particular, I would like to thank those who were co-authors on my papers: Thomas Dannenhoffer-Lafage, Morris Cohen, Aleksander Durumeric, Jaehyeok Jin, Dr. James F. Dama, and Professor Gregory A. Voth.

Of course, I should thank all the people who were members of the Voth group during my tenure in the group. In particular, I would like to thank those with whom I had productive conversations: Zack Jarin, Thomas Dannenhoffer-Lafage, Morris Cohen, Aleksander Durumeric, Dr. John Grime, Dr. Glen Hocky, Dr. Rui Sun, and Dr. John Savage. Additionally, I would like to thank my colleagues who I shared an office with during my first year as a graduate student: Jeremy Tempkin, Andrew Valantine, Nolan Shepard, Paul Stanstead, Dr. Erica Strum.

Above all, I would like to thank those who supported and encouraged me so that I could make it to this point. As scientific mentors, I would like to thank Cindy Zebris of Rocky River High School, Bob Martuch of Sherwin Williams, and Craig Burkhart of the Goodyear Tire & Rubber Company. As colleagues along the way, I would like to thank Andrew Brehm, Todd MacMillan, and Vahagn Yeranossian, As personal supporters, I would like to thank my mother Joanne, my father Joe, and my wife Sabrina.

Abstract

Coarse-grained (CG) modeling is a promising way to study materials with chemical detail with a computer at minimal computational cost. Accurate and reliable CG simulation could speed-up the research and development process for the design for pharmaceuticals, tires, batteries, etc. Also, they offer the opportunity to test hypotheses with molecular resolution. However, there are several issues that must first be addressed in order to make CG modeling truly practical. Key among these is transferability and representability. In this thesis, work is presented that addresses aspects of transferability and representability. First, an approach is presented to calculate the sensitivity of coarse-grained models to changes in the model from which they are derived. This sensitivity can be used to compute first order corrections to CG interactions that extend the range over which CG models can accurately be transferred. Second, an approach is discussed that would allow one to construct CG observables that reproduce the observables of the model from which the CG model is derived. Then, a method to implement this approach is presented. Third, a new class of interactions is introduced that allows CG models to more faithfully reproduce features of the model from which it is derived. Different terms in this class are implemented and applied to liquids at interfaces as well as a protein system. Taken together, this work provides a way to further improve the transferability and representability of CG models.

Chapter 1

Introduction

The design and development of new materials is essential to the progression of technology that we become accustomed to in the modern world. For example, new drugs are designed to treat rare, chronic, and serious diseases, which helps improve life expectancy and quality of life around the world.[1-5] Likewise, advances in the materials used in tires help to improve their safety and longevity.[6-8]

For a long time, it was possible to discover these materials based solely on human experience and intuition. Now, the "low hanging fruit" for this approach has largely been exhausted. This has led to more principled experimental approaches such as combinatorial materials exploration.[9] However, the search for new materials in this way is significantly more cost and labor intensive.

As a result, efforts have been made to discover and design materials computationally through efforts such as the Materials Genome Initiative, among others.[10-13] Such approaches promise to reduce the time and cost needed for the research and development of new materials. This is possible because such an approach screens candidate materials for the desired properties without needing to source, synthesize, and physically study materials.

To this end, computer simulation is an essential element of computational materials design. While simulations of atoms via molecular dynamics (MD) simulation is cheaper than experimental methods, MD is currently limited in the system sizes and times that can be studied by computational resources. In order to circumvent these limits, simplified models called coarse-grained (CG) models can be developed, which seek to reproduce the essential features of the more fine-grained (FG), atomistic model with reduced computational cost.[14-18]

In order for CG models to fulfill their potential as tools for materials design there are a number of issues that need to be addressed. Three such fundamental issues addressed in this thesis are model fidelity, transferability, and correspondence. Fidelity refers to the ability of a CG model to reproduce the essential features (i.e., structures and distributions) of the corresponding FG model. Transferability is the ability of a CG model to maintain fidelity at conditions different from those that were used to parameterize it. A related problem to transferability is sensitivity, which refers to how a model would change as conditions change. Correspondence refers to the ability of CG observables to correspond (and reproduce) the observables of the corresponding FG model. In the literature, this problem is often referred to as representability.

The rest of this thesis is organized as follows: Chapter 2 provides background on MD simulation and CG methods. Chapter 3 discusses a low-noise, computationally efficient way to measure the sensitivity of CG models to changes in the FG model they represent. Chapters 4 and 5 discuss the correspondence between FG and CG observables as well as how to establish such correspondences in numerically simulated models. Chapters 7 and 8 discuss the introduction of extra terms into CG interactions to improve the structural fidelity of a given CG model. Finally, Chapter 9 provides conclusions and discusses future directions.

Chapter 2

Molecular Dynamics and Coarse-graining Methods

2.1 Introduction

The research presented in this thesis builds upon existing knowledge of statistical mechanics,[19] molecular dynamics (MD),[20, 21] and coarse-graining (CG) methods.[14-16, 18] While each of these topics are large subjects, this chapter will present the basic aspects of each that are directly drawn upon later in this thesis.

2.2 Statistical Mechanics

A classical system of n particles can be fully characterized given all the positions $\mathbf{r}^n$ and momenta $\mathbf{p}^n$. The Hamiltonian for such a system is the sum of kinetic K and potential V contributions:

$$H\left(\mathbf{r}^n,\mathbf{p}^n\right)=K\left(\mathbf{p}^n\right)+V\left(\mathbf{r}^n\right). \tag{2.1}$$

The kinetic contribution is simply

$$K\left(\mathbf{p}^n\right)=\sum_{i=1}^{n}\frac{p_i^2}{2m_i}\ , \tag{2.2}$$

where m_i is the mass of particle i. In principle, the potential contribution can be expressed as a many-body expansion:

$$V\left(\mathbf{r}^n\right)=\sum_i v_1\left(r_i\right)+\sum_i\sum_{j>i} v_2\left(r_i,r_j\right)+\sum_i\sum_{j>i}\sum_{k>j} v_3\left(r_i,r_j,r_k\right)+\ldots\ . \tag{2.3}$$

However, this is usually truncated at pairs, and the pair potential is often treated a solely a function of the distance between the particles: $v_2\left(r_i,r_j\right)\approx v_2\left(\left|\mathbf{r}_{ij}\right|\right)=v_2\left(r_{ij}\right)$.

A full partition function can be written for the system in the canonical (i.e., constant NVT) as

$$Q_{NVT}=\frac{1}{N!}\frac{1}{h^3}\int d\mathbf{r}^n d\mathbf{p}^n exp\left(-\beta H\left(\mathbf{r}^n,\mathbf{p}^n\right)\right), \tag{2.4}$$

where $\beta=\left(k_BT\right)^{-1}$ and k_B is Boltzmann's constant. However, the momenta can be integrated out to give ideal gas, Maxwell-Boltzmann statistics. Thus, it is still accurate but more convenient to use a simplified partition function:

$$Z(\mathbf{r}^n)=Z_{NVT}=\int dr^n exp\left(-\beta V\left(\mathbf{r}^n\right)\right). \tag{2.5}$$

With this partition function one can express the probability of a given configuration as

$$p\left(\mathbf{r}^n\right)=\frac{e^{-\beta V\left(\mathbf{r}^n\right)}}{Z_{NVT}}. \tag{2.6}$$

Likewise, ensemble-averaged quantities can be expressed as the following expectation for an arbitrary property X:

$$\left\langle X\right\rangle=\frac{\int d\mathbf{r}^n X\left(\mathbf{r}^n\right)exp\left(-\beta V\left(\mathbf{r}^n\right)\right)}{\int d\mathbf{r}^n exp\left(-\beta V\left(\mathbf{r}^n\right)\right)}. \tag{2.7}$$

Some properties used in this thesis include free energy, energy, entropy, pressure, and radial distribution functions (RDFs). For constant NVT, the Helmholtz free energy is $A=-\beta^{-1}\ln Z_{NVT}$., which is related to the energy E and entropy S through $A=E-TS$. The energy can also be expressed as derivative of the partition function: $\left\langle E\right\rangle=-\left(d\ln Z_{NVT}/d\beta\right)_{N,V}$. The entropy can

generally be expressed as $S=-k_B\int d\mathbf{r}^n p\left(\mathbf{r}^n\right)\ln p\left(\mathbf{r}^n\right)$. The pressure P can be expressed as a thermodynamic derivative:

$$P=\frac{N}{V\beta}+\beta^{-1}\left(\frac{d\ln Z_{NVT}}{dV}\right)_{N,T}=\frac{N}{V\beta}-\beta^{-1}\left(\frac{d\ln A_{NVT}}{dV}\right)_{N,T}. \tag{2.8}$$

For a system with only pair potentials, the pressure can be written via the virial expression:

$$P=\frac{N}{V\beta}-\frac{1}{3V}\sum_i\sum_{j>i} f\left(r_{ij}\right)\cdot\mathbf{r}_{ij}. \tag{2.9}$$

Finally, the RDF is expressed as

$$g\left(r\right)=V\frac{N-1}{N}\left\langle\delta\left(r-r_{ij}\right)\right\rangle. \tag{2.10}$$

The RDF is related to the potential of mean force (PMF) $w\left(r\right)=-\beta^{-1}\ln g\left(r\right)$ through the reversible work theorem.

It is important to note that these equations assume a classical, atomistic system. Specifically, these sites have no internal degrees of freedom. Thus, these expressions are not generally the valid for any other resolution. A discussion of observable expressions for other resolutions starts in Chapter 4.

2.3 Molecular Dynamics

In order to simulate complex systems, they must be simulated numerically. This amounts to discretizing newton's second law $\mathbf{f}=\mathbf{m}\cdot\mathbf{a}$, where f is the force, m is the mass, a is the acceleration, and all terms are vectors containing elements for all particles. It can be turned into a coupled system of differential equations:

$$\frac{d\mathbf{r}}{dt} = \mathbf{v} \tag{2.11}$$

$$\frac{d\mathbf{v}}{dt} = \frac{\mathbf{f}}{\mathbf{m}} \ , \tag{2.12}$$

where t is time, v is the velocity. A common way to do the integration is using the velocity Verlet algorithm:

$$\mathbf{r}\left(t+dt\right) = \mathbf{r}\left(t\right) + \mathbf{v}\left(t\right)dt + \frac{1}{2}\mathbf{a}\left(t\right)dt^2 \tag{2.13}$$

$$\mathbf{v}\left(t+dt\right) = \mathbf{v}\left(t\right) + \frac{1}{2}\left(\mathbf{a}\left(t\right) + \mathbf{a}\left(t+dt\right)\right)dt \tag{2.14}$$

where dt is the timestep. In order to maintain a constant NVT ensemble thermostats are used. Frequent choices include Berendsen thermostat and the Nose-Hoover thermostat.

The pairwise, nonbonded interactions for particles are usually the sum of a short-range van der Waals and a longer-range electrostatic interaction. The short-range interaction is usually parameterized using the Lennard-Jones potential:

$$U_{LJ}\left(r_{ij}\right) = 4\varepsilon_{ij}\left(\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right) , \tag{2.15}$$

where ε is the depth of the well, and σ is the distance at which the potential crosses 0. For computational efficiency a cutoff r_c is often used. Potential shifted versions are constructed by subtracting the value of the potential at the cutoff from the potential, which ensures continuity in

the potential at the cutoff. To ensure continuity in both the potential and the force at the cutoff, a shifted-force version can be used:

$$U_{LJ}^{SF}(r_{ij}) = U_{LJ}\left(r_{ij}\right) - U_{LJ}\left(r_c\right) + \left(r_{ij} - r_c\right)\left.\frac{U_{LJ}\left(r\right)}{dr}\right|_{r=r_c} . \tag{2.16}$$

The electrostatic interactions can be evaluated using Coulomb's law:

$$U_Q\left(r_{ij}\right) = \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} , \tag{2.17}$$

where q_i is the charge of particle i, and ε_0 is the permittivity of free space. In practice, this expression is used in combination with another method such as Ewald summation or particle-particle particle-mesh to calculate the long-range contributions. To minimize surface effects of finite simulation boxes, periodic boundary conditions (PBC) are used in practice.

Additionally, molecular systems often have bonded, angular, and dihedral interactions. Bonds connect adjacent sites, and bonded interactions are a function of the pair distance between those sites. Angles are between two sets of bonds that share a common site, and the angular interactions are a function of the arccosine of the dot product of the unit vectors of each constituent bond. The functional form for bonded and angular potentials is frequently harmonic. Dihedrals are defined as a series of 3 bonds where each bond has one site in common with exactly one other bond. Dihedral interactions are a function of the arctangent of dot product of the cross product between the central bond and one of the other bonds in the dihedral.

2.4 Coarse-graining Methods

In this thesis, coarse-graining methods is taken to refer specifically to bottom-up method, meaning those that parameterize CG models based on higher resolution, FG data. As such, they methods directly relate FG configurations to CG configurations through a mapping operator. In this thesis only linear mappings or discussed. Linear mappings are of the form $R_I = M_I\left(\mathbf{r}^n\right)\sum_i c_{iI} r_i$, where c_{iI} is the mapping coefficient for FG site i to CG site I. The only requirement is that $\sum_i c_{iI}$ for all I. Consequently, this choice of mapping determines the mass of CG sites in order to maintain consistency between FG and CG momentum space. Specifically, $M_I = \left(\sum_i c_{Ii}/m_i\right)^{-1}$.[22] Common types of mappings include center of mass (COM), center of charge (COC), and carbon-alpha (i.e., a single atomic site for a group of atoms such as an amino acid). The expressions for mapping coefficients, CG site positions, CG site forces, and CG site masses are presented in Table II.1

With the structural relationship established, bottom-up CG methods seek to determine CG interactions that reproduce the many-body PMF $W\left(\mathbf{R}^N\right) = -\beta^{-1}\ln Z(\mathbf{R}^N)$. Each of the methods discussed below do so in different ways.

Table 2-1. Mapping coefficients, CG positions, CG forces, and CG masses for center of mass (COM), center of charge (COC), and carbon-alpha (C-alpha) mappings.

Mapping	Mapping Coefficients	CG Force	CG Mass
COM	$c_{Ii} = \frac{m_i}{\sum_{i \in I} m_i}$	$F_I = \sum_{i \in I} f_i$	$M_I = \sum_{i \in I} m_i$
COC	$c_{Ii} = \frac{q_i}{\sum_{i \in I} q_i}$	$F_I = \sum_{i \in I} f_i$	$M_I = \left(\sum_{i \in I} \frac{q_i^2}{m_i \left(\sum_{i \in I} q_i \right)^2} \right)^{-1}$
C-alpha	$c_{Ii} = \delta_{ik}$	$F_I = f_k$	$M_I = m_k$

2.4.1 Distribution Matching

The first, and perhaps most obvious choice for parameterize a CG model is to reproduce the FG's model structural distributions by targeting the mapped RDF directly. The initial guess for the potential is the pair PMF:

$$U_0(R) = -\beta^{-1} \ln g_{AA}(R), \tag{2.18}$$

where g_{AA} is a mapped RDF based on an atomistic (i.e., FG) simulation. If one were to stop here, it would be called direct Boltzmann Inversion.[23]

The pair PMF neglects often-significant correlations in full PMF. So, this initial guess is often refined in an iterative fashion:

$$U_{i+1}(R) = U_i(R) + \alpha_i \ln \frac{g_i(R)}{g_{AA}(R)} \tag{2.19}$$

where g_i is calculated from a CG simulation using the former CG interaction, and α is a user-defined learning rate. The hope is that correlations neglected in prior iterations can be partially recovered in subsequent iterations. Not surprisingly, this method is called iterative Boltzmann Inversion (IBI).[15, 24]

Alternatively, one could minimize the difference in the CG and FG RDFs in another way. A residual quantifying this difference is

$$\chi^2 = \sum_k \frac{\left(g_i\left(R_k\right) - g_{AA}\left(R_k\right)\right)^2}{\sigma^2\left(R_k\right)}, \tag{2.20}$$

where the RDF is evaluated at discrete intervals and σ determines the relative importance of errors in the RDF at that point. By minimizing this residual using Netwon's method, one obtains the Inverse Monte Carlo method.

Yet another way to measure the difference between the CG and FG RDFs (or other distributions) is to use the relative entropy (RE):[25-27]

$$S_{rel} = \int d\mathbf{r}^n p_{FG}\left(\mathbf{r}^n\right) \ln \frac{p_{FG}\left(\mathbf{r}^n\right)}{p_{CG}\left(M\left(\mathbf{r}^n\right)\right)} . \tag{2.21}$$

Minimizing the RE using gradient descent is equivalent to IBI, and minimizing the RE using Newton's method is equivalent to IMC.[25, 28]

2.4.2 Force Matching

Alternatively, one could try reproducing the derivative of this distribution, would is the forces. One residual that does this is

$$\chi^2 = \int d\mathbf{r}^n \left\| f_{CG}\left(M\left(\mathbf{r}^n\right)\right) - M_I^{\dagger}\left(f_{AA}\left(\mathbf{r}^n\right)\right) \right\|^2 . \tag{2.22}$$

Since the functional forms used for the potentials in all of these methods are usually splines with linear coefficients, these coefficients become decoupled when looking at their derivative. Thus, this residual can be minimized directly in one step. This approach is commonly referred to as multi-scale coarse-graining (MS-CG).[22, 29-40]

Chapter 3

Predicting the Sensitivity of Multiscale Coarse-grained Models to their Underlying Fine-grained Model Parameters

This chapter is reprinted with permission from J. Chem. Theory Comput. 2015, 11(8), 3547-3560.[41]

3.1 Introduction

Coarse-grained (CG) models seek to capture the essential details of fine-grained (FG) models at reduced computational cost by eliminating degrees of freedom (DOFs).[14-16, 18] To further increase computational savings with CG models, it is desirable to know when one can reuse the same CG model to describe FG models similar to the original FG model that was used to parameterize the CG model. Unfortunately, CG models tend to have limited transferability because eliminating DOFs leads to state point dependent effective interactions.[42, 43] The dual problem to designing transferability is the determination of model sensitivity (i.e., how a CG model would change if it were parameterized under a different set of FG interactions or at a different state point). CG sensitivities can be used simply to determine the transferability of CG models, but a more promising and enterprising approach is to correct CG models with the changes predicted from the calculated sensitivities. However, for this to be practical one must have a method of calculating sensitivities that is more computationally cost effective than running new FG simulations to directly calculate new CG interactions at each new state point. In this chapter, we address this need for a computationally efficient method to calculate model sensitivities by proposing novel formulae for computationally efficient, low noise estimates of sensitivities from single FG simulations.

The model sensitivity of molecular systems has been extensively investigated at a single, all-atom (AA) level of resolution. Wong and Rabitz[44, 45] calculated changes in free energy as a function of changes in the input Lennard-Jones (LJ) parameters in the simplest, linear response sense of sensitivity using a finite difference (FD). Using ad hoc modifications to FDs, Rocklin et al.[46] calculated the sensitivities of binding free energies to changes in interaction parameters. More computationally efficient methods employing single state and multistate statistical reweighting[47-49] can also be used to get estimates of the sensitivity using FDs and have been used to find potential sensitivity to interaction cutoffs.[50] Fleishman and Brooks took a different approach,[51] using derivatives of the partition function and thermodynamic integration to calculate sensitivity of entropy and enthalpy via perturbation theory. Later, Wong, Thacher, and Rabitz used careful statistical mechanical derivations to determine first and second order sensitivity coefficients.[52] These sensitivity coefficients have been used in materials design to optimize binding free energies by changing cation interactions[53] and to determine which input parameters are most influential in determining observables.[54] Recently, an expanded set of sensitivity equations were applied to improve a classical water model's agreement with *ab initio* and experimental measures.[55] However, sensitivities calculated at a single level of resolution do not address the problem of CG model transferability.

Sensitivity between models of two different resolution levels has been the subject of limited study. Krishna et al.[37] used the multiscale coarse-graining (MS-CG)[22, 35, 36, 40] methodology and statistical reweighting over temperatures to get different CG potentials, which is a complementary approach to the one taken in this chapter. Lu et al.[39] used FDs with MS-CG to decompose free energies into entropic and enthalpic components using sensitivity to temperature. As an alternative, some researchers have tried to increase the transferability of CG potentials by

including extra independent variables. A few of these approaches included three-body interactions,[32, 38] temperature dependent terms,[56, 57] density dependent terms,[58-60] or concentration dependent terms.[61] However, while much work has been done to improve the transferability of CG potentials, only reweighting and FD approaches have been used to probe sensitivity in multiscale calculations.

Several systematic approaches for developing CG models could potentially be used as a starting point for developing methods to calculate sensitivities, including relative entropy (RE),[26] inverse Monte Carlo (IMC),[62] iterative Boltzmann inversion (IBI),[63] force matching (FM),[22, 35, 36, 40, 64, 65] and the generalized Yvon-Born-Green (g-YBG) equations.[66, 67] RE minimization is a general approach in which one aims to minimize the loss of Shannon information from the FG model to the CG model potential.[26] If RE is minimized using Newton's method,[62] one obtains IMC,[24] which inverts radial distribution functions (RDFs) iteratively to provide interaction potentials – though sampling noise must be taken into account.[68] Similarly, IBI inverts the RDF iteratively and is an approximate RE minimization, just as IMC is, using a fixed-point optimization that is simpler than Newton's method.[25] Likewise, FM and g-YBG, implemented as MS-CG, converge to the same result as RE in the limit of a complete basis set since FM minimizes the average of the gradient squared of the relative entropy.[28] Since RE converges to the same results as IBI, IMC, FM, and g-YBG in the appropriate limits,[25] it is important to consider which method is most appropriate given the limitations of the problem at hand. In the case that sampling is incomplete at short interaction distances and there are potential issues with basis sets to simultaneously describe the CG potential, force field, and the sensitivity of the CG force field, the local nature of MS-CG becomes appealing. Local nature here refers to the fact that, in MS-CG, the fit in each portion of the force field is linked linearly to fits in other portions

through a g-YBG equation[66, 67] rather than through a complex, nonlinear, and nonlocal dependence. This feature of MS-CG removes the nonlinearity found in the other global, distribution-matching minimization methods, thus leading to a more direct, computationally straightforward method of calculating sensitivity that is better suited to rapid prototyping.

The present work develops reweighting-free, single simulation formulae that calculate the sensitivity of CG potentials and force fields to changes in the underlying FG interaction parameters and state points at the level of linear response. The calculated sensitivities are used to develop corrections to CG models that increase model accuracy when the CG potentials and force fields are transferred alchemically across interaction parameters or thermodynamically across state points. The accuracy of these predicted sensitivities are evaluated by comparison with reweighted FDs, and the accuracy of the corrected, transferred potentials are compared against potentials without any sensitivity correction.

The remainder of this chapter is structured as follows: Section 2 describes the derivation and significance of the formulae developed in this work as well as the numerical and simulation methods used. Section 3 shows the application of these formulae to single site methanol and solvent-free sodium chloride systems and the resulting accuracy of the predicted sensitivities and potentials. Section 4 provides a general discussion of those results including suggestions for future work. Section 5 provides conclusions.

3.2 Theory and Methods

3.2.1 Sensitivity Theory

The fundamental measure of sensitivity to small changes is the derivative

$$\frac{dU(\mathbf{R}^N;\lambda)}{d\lambda} = \lim_{\delta\lambda \to 0} \frac{U(\mathbf{R}^N;\lambda+\delta\lambda) - U(\mathbf{R}^N;\lambda-\delta\lambda)}{2\delta\lambda} \quad . \tag{3.1}$$

Here, the sensitivity of the CG potential $U(\mathbf{R}^N;\lambda)$ to a FG parameter λ is calculated by finding the difference between CG potentials obtained using modified FG parameters $\lambda \pm \delta\lambda$. In the above equation, $\mathbf{R}^N$ are the CG configurational variables. This method requires the calculation of at least two CG potentials to calculate the sensitivity. However, the range of $\delta\lambda$ in which this limit is approached, known as the linear regime, is not known a priori. This means that, in fact, more than two CG potentials must be calculated to verify the meaningfulness of a single calculated sensitivity. A second problem with this FD approach is that the random noise and fluctuations in estimates of the CG potentials are magnified dramatically when $\delta\lambda$ is small, exactly where the limit is approached. For the FD approach to be feasible, therefore, one must find a $\delta\lambda$ that is both in the linear regime and sufficiently large to make pulling the sensitivity signal out of sampling noise tractable, but there is no guarantee that such a $\delta\lambda$ exists.

An alternative to the basic multi-trajectory FD (MTFD) is to use statistical reweighting to obtain the CG potentials at different $\delta\lambda$ values from a single FG simulation. Statistical reweighting reuses configurations generated using a given parameterization by applying a reweighting factor, the ratio between Boltzmann factors across parameterizations, to the results of reanalyzing the configurations using a different parameterization. For generating CG models using MS-CG or g-YBG, this amounts to weighting the FM residual from the reevaluated forces by the exponential of the difference in the CG potential calculation (or inverse temperature, if temperature is varied). The use of a single trajectory in reweighting should minimize the noise seen at small $\delta\lambda$, which makes this approach appear relatively promising compared to FD, but the range of the linear regime is still not known a priori. Another problematic condition required

for reweighting to be practical is that the original ensemble and the ensemble estimated by reweighting must have significant overlap so that the reweighted trajectory can give reliable averages. Furthermore, the averages are susceptible to bias when the original sample may not provide configurations that overlap with the reweighted ensemble *evenly*. An example of this is the application of reweighting to calculate averages at higher temperatures than the temperature of an initial simulation. Since the original trajectory explores only a subvolume of the phase space explored at higher temperatures, the reweighting procedure typically biases the resulting potentials to over-represent behavior characteristic of lower-energy conformations. For reweighting across interaction parameters, it is not always clear when this is a problem or how significant the bias may be — even after the calculation is complete.

Ideally, one would like a method of calculating sensitivities in the linear regime that does not depend on knowledge of the size of the linear regime, requires minimal computation, and is less susceptible to bias than a reweighted finite difference (RFD). We can do so by analytically evaluating the limit in the FD above, then using the resulting formulae to make our calculations. Starting from the FD formula above, one arrives at the equation

$$\frac{dU_{CG}(\mathbf{R}^N;\lambda)}{d\lambda} = \left\langle \frac{du(\mathbf{r}^n;\lambda)}{d\lambda} \right\rangle_{\mathbf{R}^N,\lambda}, \tag{3.2}$$

where $u\left(\mathbf{r}^n;\lambda\right)$ is the FG potential in terms of the FG coordinates $\mathbf{r}^n$. However, this equation is remarkably data-inefficient: it only uses one scalar value of information per sampled frame, which is the derivative of the potential with respect to λ for that frame. Fitting a many-body function with many free parameters requires a great deal of input training data, and using only one datum per frame of input data would require a huge number of frames to properly parameterize the many-body sensitivity. Therefore, we apply a trick with a long history. By

analogy to FM,[22, 35, 36, 40, 64, 65, 69] which uses more data per frame than potential matching by matching derivatives of the potential with respect to particle positions instead of matching per-frame potentials directly, one can also derive formulas for sensitivity matching that match the sensitivity of the derivatives of the potential with respect to all particle positions instead of matching the per-frame sensitivity directly. The gain in information per frame is proportional to the number of particles, which can be quite large. The remainder of this section describes the derivation of two such formulae, each of which results from a different approach to the problem of representing the many-body sensitivity in a reduced space of trial functions. *We propose that the first formula be used for practical calculation of sensitivity and the second formula be used as a theoretical diagnostic tool.*

Self-Consistent Basis (SCB) Single Point Formula

In the first derivation, we find a formula by considering the sensitivity of approximations to the many-body potential. After all, any practical CG potential will be an approximation, and we are therefore interested in the sensitivity of approximations when we talk about the sensitivity of CG models. A natural choice here is to look at the sensitivity of an approximate CG potential in the same set of trial functions used to construct the CG potential; because the basis functions for the CG potential and sensitivity are the same in this case, we call this a self-consistent basis (SCB) single-point formula. To construct this formula, one needs to start from the FM residual expression reweighted from λ to $\lambda+\delta\lambda$ with the framewise weight function

$$w_t\left(\mathbf{r}^n;\lambda,\delta\lambda\right)=\frac{\exp\left(-\beta u\left(\mathbf{r}^n,\lambda+\delta\lambda\right)+\beta u\left(\mathbf{r}^n;\lambda\right)\right)}{\frac{1}{N_t}\sum_{t=1}^{N_t}\exp\left(-\beta u\left(\mathbf{r}^n,\lambda+\delta\lambda\right)+\beta u\left(\mathbf{r}^n;\lambda\right)\right)}\ , \tag{3.3}$$

where N_t is the total number of simulation frames. Optimization of the residual with respect to basis functions in order to obtain the reweighted FM normal equations yields[40, 70]

$$\frac{1}{N_t}\sum_{t=1}^{N_t} w_t\left(\mathbf{r}^n;\lambda,\delta\lambda\right)\mathbf{F}^T\mathbf{F}\phi = \frac{1}{N_t}\sum_{t=1}^{N_t} w_t\left(\mathbf{r}^n;\lambda,\delta\lambda\right)\mathbf{F}^T\mathbf{f}\,, \tag{3.4}$$

where $\mathbf{F}$ is a matrix of configurational information about the basis function values for each particle, $\mathbf{f}$ is the 3N vector of the target forces, and ϕ is a vector of the unknown linear basis function coefficients. Then, taking a derivative of both sides with respect to $\delta\lambda$ and taking the limit $\delta\lambda \to 0$, a set of normal equations for the sensitivity emerges. Taking the limit of a long trajectory and a complete basis set and then rearranging those normal equations, the expression for the approximate sensitivity matching in terms of thermodynamic averages is

$$\frac{d\nabla U(\mathbf{R}^N;\lambda)}{d\lambda} = \left\langle M^\dagger\left(\frac{d\nabla u(\mathbf{r}^n;\lambda)}{d\lambda}\right) - \frac{\beta}{N_{CG}}\left(\frac{du(\mathbf{r}^n;\lambda)}{d\lambda} - \left\langle\frac{du(\mathbf{r}^n;\lambda)}{d\lambda}\right\rangle_\lambda\right)\left(M^\dagger\left(\nabla u(\mathbf{r}^n;\lambda)\right) - \nabla U_{CG}\left(\mathbf{R}^N;\lambda\right)\right)\right\rangle_{\mathbf{R}^N,\lambda}, \tag{3.5}$$

where $\left\langle du(\mathbf{r}^n;\lambda)/d\lambda\right\rangle_\lambda$ is the Boltzmann weighted expectation value of $du(\mathbf{r}^n;\lambda)/d\lambda$ over the entire FG ensemble, N_{CG} is the number of CG sites in order to make the sensitivity intensive, and $M^\dagger$ is the mapping operator that transforms the FG forces into CG forces (See Appendix B for derivation and details). Every term on the right can be estimated directly from simulation, so these estimates can be calculated in a single step without iteration.

Self-Consistent Iterative (SCI) Single Point Formula

An interesting alternative to the derivation in the previous section is to consider what the expression would be like if one treated the sensitivity of the many-body potential directly regardless of basis set limitations. A sensitivity estimator based on approximating the sensitivity of a full basis set potential using a finite basis set rather than on the sensitivity of approximate potentials using finite basis sets would provide a diagnostic for assessing the importance of renormalized many-body effects in CG sensitivities. As before, we look for a formula in terms of derivatives of forces instead of derivatives of potentials. Starting from the FD of the forces with the averages in the $\lambda \pm \delta\lambda$ ensembles written explicitly, then substituting the mapped FG forces for the CG forces, the transformation of these ensembles to a common λ ensemble leads to

$$\frac{d\nabla U(\mathbf{R}^N;\lambda)}{d\lambda} = \lim_{\delta\lambda \to 0} \frac{1}{2\delta\lambda} \left\langle M^\dagger\left(\nabla u(\mathbf{r}^n;\lambda+\delta\lambda)\right)\frac{e^{-\beta u(\mathbf{r}^n;\lambda)+\beta u(\mathbf{r}^n;\lambda+\delta\lambda)}}{\left\langle e^{-\beta u(\mathbf{r}^n;\lambda)+\beta u(\mathbf{r}^n;\lambda+\delta\lambda)}\right\rangle} - M^\dagger\left(\nabla u(\mathbf{r}^n;\lambda-\delta\lambda)\right)\frac{e^{-\beta u(\mathbf{r}^n;\lambda)+\beta u(\mathbf{r}^n;\lambda-\delta\lambda)}}{\left\langle e^{-\beta u(\mathbf{r}^n;\lambda)+\beta u(\mathbf{r}^n;\lambda-\delta\lambda)}\right\rangle}\right\rangle_{\mathbf{R}^N,\lambda}. \quad (3.6)$$

After expanding the exponentials in terms of $\delta\lambda$ and discarding all terms higher than linear in $\delta\lambda$ (see Appendix A for details), one obtains

$$\frac{d\nabla U(\mathbf{R}^N;\lambda)}{d\lambda} = \left\langle M^\dagger\left(\frac{d\nabla u(\mathbf{r}^n;\lambda)}{d\lambda}\right) - \frac{\beta}{N_{CG}}\left(\frac{du(\mathbf{r}^n;\lambda)}{d\lambda} - \left\langle\frac{du(\mathbf{r}^n;\lambda)}{d\lambda}\right\rangle_{\mathbf{R}^N,\lambda}\right) M^\dagger\left(\nabla u(\mathbf{r}^n;\lambda)\right)\right\rangle_{\mathbf{R}^N,\lambda}, \quad (3.7)$$

where N_{CG} is the number of CG sites in order to make the sensitivity intensive. This equation is a self-consistent iterative (SCI) single point formula because while the left hand side seems optimistically like it could be computed in a variational approximation by performing FM on the

right hand side, this is actually not correct. The term $\left\langle du(\mathbf{r}^n;\lambda)/d\lambda \right\rangle_{\mathbf{R}^N,\lambda}$ is exactly the many-body function that the formula is meant to calculate, and therefore the equation must be solved iteratively: after each variational calculation step to find the left-hand side, $\left\langle du(\mathbf{r}^n;\lambda)/d\lambda \right\rangle_{\mathbf{R}^N,\lambda}$ must be reevaluated framewise using the integrated form of the new left hand side to generate new target derivatives, and a new variational calculation must be run. The process repeats until self-consistency. Note that FM calculates a potential up to an additive constant. Normally, this constant has no physical effect, but in this case the constant is important in the nonlinear term containing $\left(du/d\lambda - \left\langle du/d\lambda \right\rangle\right)$. We therefore apply a configurationally independent constant correction to the difference $\left(du/d\lambda - \left\langle du/d\lambda \right\rangle\right)$ so that its average over all frames is zero. This amounts to a single step of direct scalar matching used to seed the iterative FM calculations; the scalar does not affect the distributions of configurations in sensitivity-corrected models.

Both the SCB and SCI single point formulae have the same first term on the right hand side, which can be considered the naïve sensitivity since it neglects any non-pairwise effects on the CG potential and force field. Interestingly, this is what one would obtain if one reanalyzed a trajectory using a different parameter set as in the RFDs but neglected to apply the reweighting factor. This is in effect what was reported by Rocklin, Mobley, and Dill.[46] One can see that if the second set of terms on the right hand side of both single point formulae were to be zero, this naïve sensitivity would, in fact, be the correct sensitivity. Thus, differences between the naïve sensitivity and the single point formulae reflect the importance of the correlation correction to the naïve sensitivity. Differences between the SCI and SCB equations reflect the importance of

basis set effects in determining which correlations should be used to correct the naïve sensitivity for approximate models.

Even though these formulae both capture correlated many-body effects of the sensitivity, they do so in different ways. In the SCI single point formula, the correction to the naïve sensitivity is like a transport term since it is the product of the mapped forces and the deviations in $du/d\lambda$; it measures the amount the FG distributions corresponding to each CG distribution are pushed around "underneath" the CG configuration. In the SCB single point formula, this correction is the product of the deviation in $du/d\lambda$ and the deviation in the forces, making it a covariance term that closely echoes the covariance corrections to the naïve sensitivity found in the literature for single resolution sensitivity.[52] A practical difference between the two single point formulae is that the SCB averages $\left\langle du(\mathbf{r}^n;\lambda)/d\lambda \right\rangle$ over the entire FG ensemble requiring no iteration, while the SCI averages $\left\langle du(\mathbf{r}^n;\lambda)/d\lambda \right\rangle$ conditional on the CG ensemble that needs to be reevaluated based on the most recent estimate from the previous iteration. These differences in averaging are consistent with the differing applications of basis sets made in the derivation of each model. Both approaches become equivalent in the limit of a complete basis set, but for a finite basis set the SCB formula describes a practically useful sensitivity and the SCI formula is better used as a diagnostic for understanding the physics of renormalized many-body effects.

3.2.2 Simulation and Fitting Details and Conditions

Molecular dynamics (MD) simulations were performed on AA methanol and 1M sodium chloride systems in LAMMPS.[71-73] All systems were run with a 1 fs timestep and used nonbonded Lennard-Jones (LJ) interactions with a radial cutoff of 1.0 nm as well as particle-

particle particle-mesh (PPPM) electrostatic interactions. Both systems were equilibrated by simulating them for 5 ns at constant NPT at 1 atm and 300K, setting their volume to the average of the last 2 ns of NPT simulation, and then simulating them for at least 1 ns at constant NVT at 300K. Subsequent sampling for modified parameters were started from this equilibrated configuration, but allowed to evolve for an additional 1 ns before sampling. The OPLS[74] methanol system of 1,000 molecules was sampled for 2 ns with configurations recorded every 250 fs, consistent with other studies.[40] For the sodium chloride system, 20 sodium and 20 chloride ions were simulated using Joung and Cheatham's[75] parameterization solvated in 1,110 SPC/E[76] water molecules for 20 ns with configurations recorded every 200 fs, consistent with other studies.[77] The methanol system was coarse-grained to one site per molecule using a center of mass mapping, as in previous work.[40] The sodium chloride system was coarse-grained by eliminating all water molecules to create a solvent free model.[77] All CG forces, potentials, and sensitivities were calculated using the MS-CG FM code with a nonbonded cutoff of 1.0 nm, and sixth order spline basis functions with a resolution of 0.07 nm. CG simulations were started from the mapped version of the final configuration for the sampling run. A total of 1,000 CG timesteps were allowed for equilibration and randomization. Configurations were sampled every 1,000 CG timesteps for both systems with sampling runs of 2×10^{6} CG timesteps for CG methanol and 2×10^{7} CG timesteps for CG solvent free sodium chloride.

Independent samples and reweighted potentials were calculated for changes to all LJ epsilon, LJ sigma, and partial charge parameters. In units of kcal/mol for LJ epsilon, Angstroms for LJ sigma, and e, the fundamental charge, for charge, CG potentials were calculated with positive or negative changes in one parameter of 0.001, 0.002, 0.005, 0.010, and 0.020. Changes to the charge of one atom type were offset by changes to the charge on an adjacent atom type in order

to keep each molecule charge neutral. For methanol, this meant moving charges on the carbon and the neighboring methyl hydrogens or the oxygen and the neighboring hydroxyl hydrogen for methanol. For sodium chloride, this meant moving charges on the ions or within the water molecule. For the graphs comparing the single point formulae to the MTFDs and RFDs, the confidence ranges for MTFD and RFD curves were determined by integrating the 95% confidence interval calculated for all pairs of FDs within 0.005 parameter units. Confidence ranges for the single point formulae and the RFDs in these curves were likewise calculated by integrating the 95% confidence interval from 5 replica simulations. The confidence range for all of the independent trajectories corresponds to the integrated 95% confidence interval of 6 replicas using the original parameterization.

Comparing the effectiveness of these sensitivity formulae for transferring potentials requires computing predicted potentials (i.e., original potentials plus the sensitivity with respect to a parameter times the change in the parameter) to the CG potentials obtained from both 1) independent trajectories using an actually modified parameter set and 2) Boltzmann reweighting the original trajectory to the modified parameter set. The difference in these modified CG potentials via sensitivity, via reweighting, and via independent trajectories from the CG potential with the original parameters is quantified by integrating the absolute difference multiplied by the RDF and divided by the range of integration. This gives a single number summary (in energy units) of how different the variously transferred potentials are from the original potential for a given change in parameters. In this section, the confidence ranges for each point were calculated by propagating the uncertainty from the potentials through each of the operations in Eq. (3.9). The uncertainty of each point of the potentials was calculated as the root mean square (RMS) fluctuations of six independent trajectories. This uncertainty in the potential was used to

calculate the uncertainty in the difference of the potentials at each point, combining the uncertainties via an RMS calculation (also referred to as error propagation in quadrature). Then, this uncertainty in the difference was scaled by the magnitude of the RDF at that point and the normalization before combining via an RMS calculation to give the uncertainty used to calculate the confidence ranges shown for each point.

3.3 Results

3.3.1 Numerical Finite Differences

Before the performance of the single point formulae developed in the work is evaluated, it is worth evaluating the noise and performance of the existing numerical FD calculations. Figure 3-1 compares the MTFD and the RFD with confidence ranges for the sensitivity of the single site methanol CG potential to changes in the charge on the hydroxyl's hydrogen. As expected, both estimates agree within the confidence ranges for sufficiently small changes to the charge, but the RFD has significantly smaller confidence ranges than the MTFD, as expected because small differences in the MTFD denominator magnify sampling noise. In fact, the RFD confidence ranges are more than 100 times smaller than for MTFD. For the purposes of initially verifying the precision of our single point formula, only RFD with confidence ranges will be shown since it is expected that any predicted sensitivity that agrees with the RFD within the confidence ranges will also agree with the MTFD. However, this is not always the case, especially when the RFD calculations are strongly biased, so to demonstrate the accuracy of the single point formulae, comparisons will be made to MTFD or independent trajectories (IT) later in this chapter.

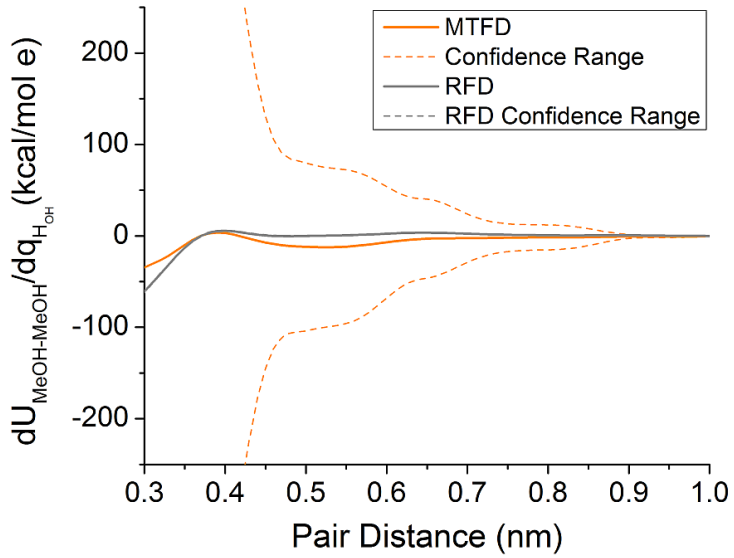

Figure 3–1. Comparison of multi-trajectory FD (MTFD) and reweighted FD (RFD) for the sensitivity of the MeOH CG potential to changes in the charge on the hydroxyl hydrogen (H_{OH}). Confidence ranges show the relative noise of each estimated sensitivity, as defined in the main text. The RFD confidence range is so small relative to the MTFD confidence range that it is not distinguishable from the RFD curve on this scale.

3.3.2 Single Site Methanol

Sensitivity Comparisons

The sensitivities calculated using the SCB single point formula are compared to the SCI single point formula as well as the RFD sensitivity estimates with confidence ranges in Figure 3-2. For LJ epsilon (Fig. 3-2a and Fig. 3-2b) and sigma (Fig. 3-2c), the SCB and SCI estimates superimpose, indicating that non-pair-representable many-body effects play little role in the pair-representable part of these sensitivities. The SCB and SCI estimates for these graphs are generally within the shown confidence range and only slightly overestimate the magnitude of the sensitivity at short interaction pair distances. For sensitivity to charge (Fig. 3-2d), the SCI

estimate is significantly different from either the SCB estimate or the actual RFD sensitivity. This difference between SCB and SCI estimates indicates that significant multibody correlations are important for charge interactions and correlations. These observations agree with and clarify prior work that indicated that CG potentials are significantly less transferable, in the naïve sense, for charge interactions than epsilon and sigma interactions because of the significant multibody correlations present.[78]

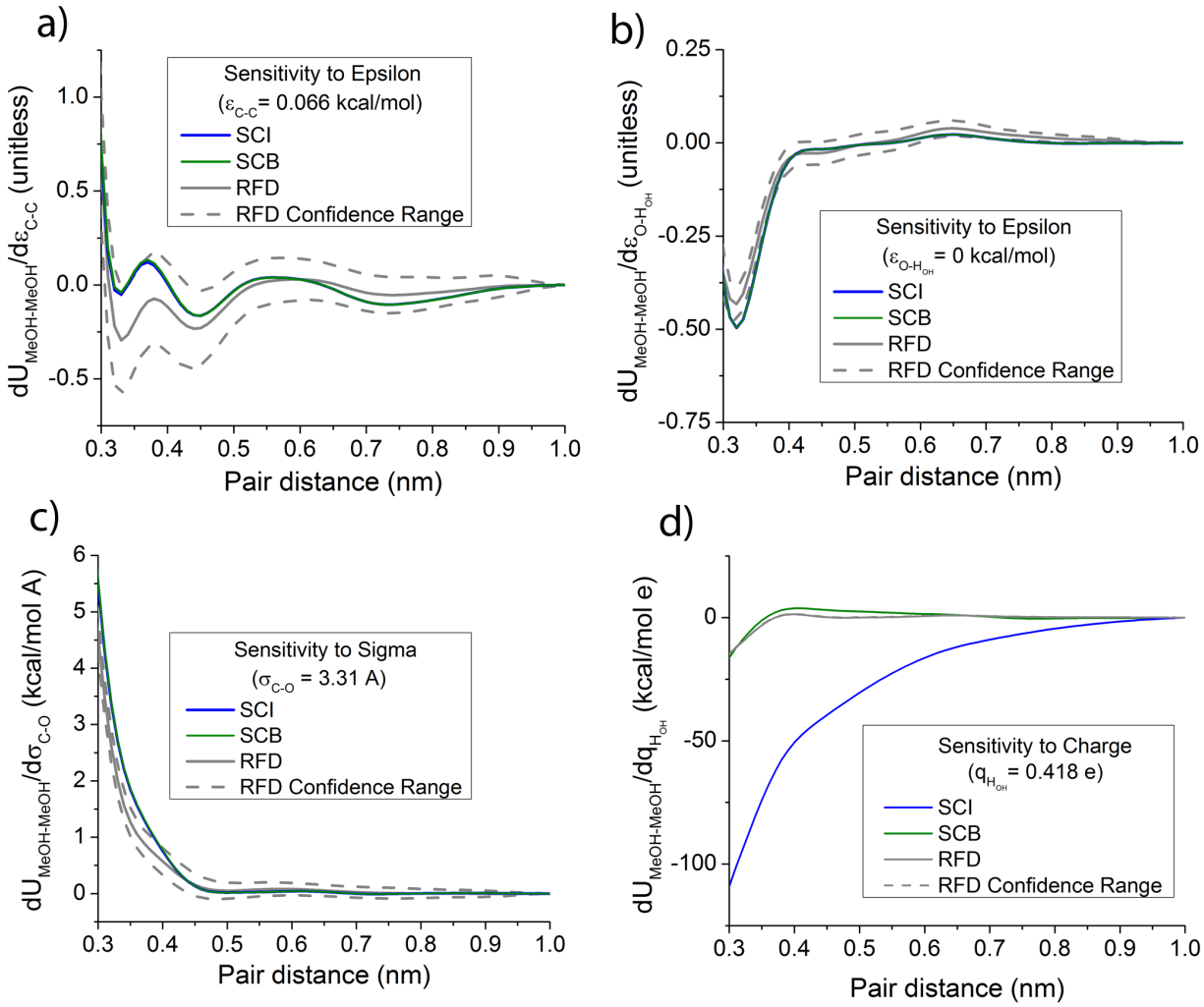

Figure 3–2. Comparison of methanol sensitivity estimates for different interaction parameters between RFD, self-consistent iterative (SCI) single point, and self-consistent

basis (SCB) single point calculations. Sensitivities are taken with respect to a) Carbon-Carbon (C-C) LJ epsilon, b) Oxygen-Hydroxyl-Hydrogen (O-H_{OH}) LJ epsilon, c) Carbon-Oxygen (C-O) LJ sigma, and d) Hydroxyl Hydrogen (H_{OH}) charge interaction parameters. RFD confidence ranges are calculated as defined in the main text. The RFD confidence range for d) is so small that it is not visible on this scale.

Predicted Potentials

As mentioned in the theory section of this chapter, potentials can be predicted using sensitivities from either of the single point formulae using

$$U\left(\mathbf{R}^N;\lambda+\delta\lambda\right)=U\left(\mathbf{R}^N;\lambda\right)+\delta\lambda\left(\frac{dU\left(\mathbf{R}^N;\lambda\right)}{d\lambda}\right), \tag{3.8}$$

which is a simple correction to the original potential that is linear in $d\lambda$. The magnitude of change in CG interaction potential from the reference (REF) parameterization as plotted in Figure 3-3 was calculated as

$$\left|\Delta U\right|=\frac{1}{R_H-R_L}\int_{R_L}^{R_H}dR\left|U_{PRED}\left(R;\lambda+\delta\lambda\right)-U_{REF}\left(R;\lambda\right)\right|g_{REF}\left(R\right), \tag{3.9}$$

where $g_{REF}(R)$ is the radial distribution function of the reference parameterization, $U_{PRED}(R;\lambda+\delta\lambda)$ is the interaction potential at a non-reference parameterization either from FMing independent FG NVT trajectories with the modified parameterization or using Eq. (3.8) with the sensitivity calculated using RFD, SCI, or SCB formulae. Figure 3-3 shows the difference in potentials for different $\delta\lambda$'s from the original ($\delta\lambda=0$) potential as described in

Section 3.2. For the epsilon graph (Fig. 3-3a), both the SCB and SCI curves agree with the reweighted curve for small differences – in the linear regime. It is remarkable that both the sensitivities have the same average slope as the curve for the CG potential determined for the new parameters with independent trajectories for changes of only 0.01 kcal/mol. For the sigma graph (Fig. 3-3b), the reweighted curves are nonlinear, but the SCB sensitivity appears to have the same average slope as the SCI sensitivity and the reweighted curve. The independent trajectories curve has a similar initial slope to the single point sensitivities, but is below the single point sensitivities for larger changes. This is somewhat expected since as perturbations increase, systems will typically make compensating changes that result in a concave response. For the charge graph (Fig. 3-3c), the reweighted curve shows nonlinearity, but the SCB curve is nonetheless reasonably consistent with the reweighted curve. As expected from the sensitivity comparisons, the SCI curve drastically overestimates the change in potential. While neither of the single point sensitivities matches the independent trajectories for charge, neither does the reweighted curve beyond 0.005 e, indicating significant sampling changes in response to charge modification that may be the result of changes in complex many-body and long range effective interactions.

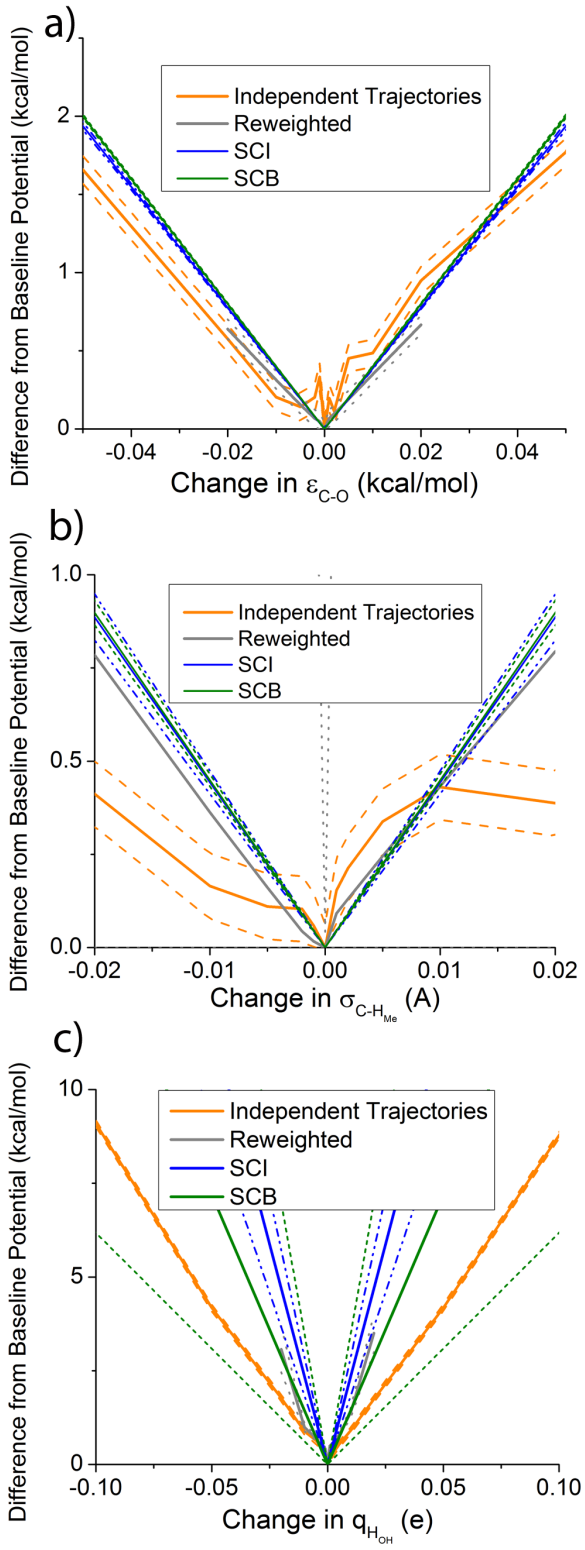

Figure 3–3. Magnitude of change in methanol CG interaction potential from OPLS parameterization, calculated (see Eq. (3.9)) as a weighted average absolute difference in

predicted potential from a reference potential weighted by the reference RDF, for predictions via independent trajectories, reweighting, and the two single point sensitivities SCI and SCB. Predictions are compared for changes in a) Carbon-Oxygen (C-O) LJ epsilon, b) Carbon-Methyl-Hydrogen ($C\text{-}H_{Me}$) LJ sigma, and c) Hydroxyl Hydrogen (H_{OH}) charge interactions.

CG Simulations

Another way to assess the accuracy of these predicted potentials is to compare the RDFs generated from CG simulations using both the actual CG potential and the potentials predicted from both sensitivity formulae. Figure 3-4 shows the RDF for a selected set of $\delta\lambda$. It is clear that any slight errors shown in Figure 3-3 for predictions across epsilon (Fig. 3-3a) and sigma (Fig. 3-3b) value do not lead to noticeable errors in the RDFs. For predictions across charge (Fig. 3-3c) values, the RDF from the SCB predicted potential has only minor deviations in the height of the first peak and valley from the actual RDF. The agreement of the RDF from the SCB predicted potential bodes well for the application of sensitivity for generating predicted potentials. However, there is a limit to how far one can use these predicted potentials, which corresponds to the breakdown of the first order approximation of the sensitivity outside the linear regime. Figs. 3-3d-f show that for sufficiently large changes in the FG interaction parameters, the magnitude of the RDF peaks and valleys differ significantly from those of the RDF obtained using the actual CG potential with the modified interaction parameters. Nonetheless, the RDFs for larger interaction parameter changes continue to show good agreement with the location of the peaks and valleys for the first two solvation shells.

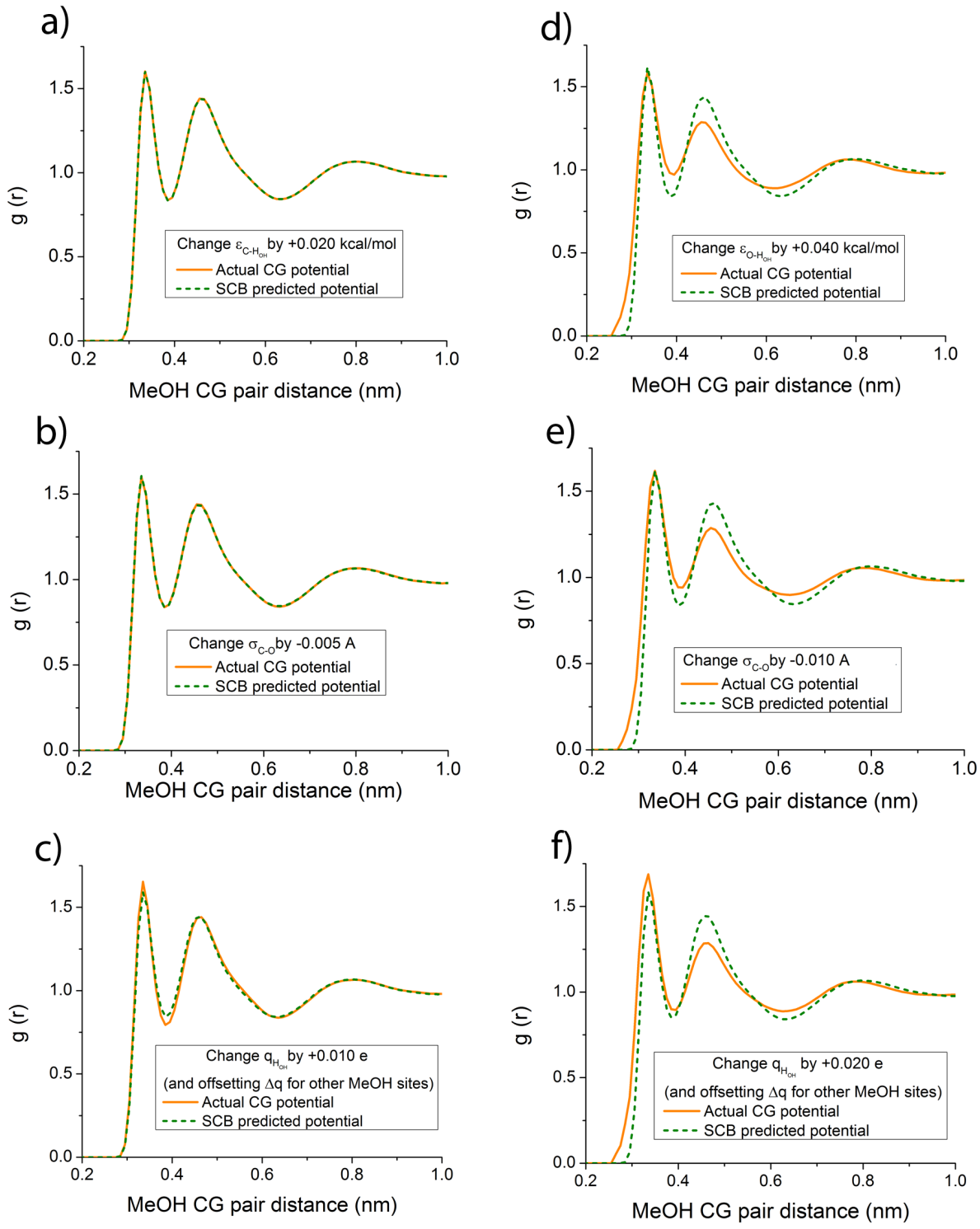

Figure 3–4. Radial distribution functions (RDFs) from CG methanol simulations. a) Changing Carbon-Hydroxyl-Hydrogen (C-H_{OH}) LJ epsilon interaction parameter by 0.020 kcal/mol. b) Changing Carbon-Oxygen (C-O) LJ sigma interaction parameter by -0.005 A. c) Changing the

Hydroxyl Hydrogen's charge by +0.010 e and applying neutralizing charges on the other methanol FG sites. d) Changing Carbon-Hydroxyl-Hydrogen ($C\text{-}H_{OH}$) LJ epsilon interaction parameter by 0.040 kcal/mol. e) Changing Carbon-Oxygen (C-O) LJ sigma interaction parameter by -0.010A. f) Changing the Hydroxyl Hydrogen's charge by +0.020 e and applying neutralizing charges on the other methanol FG sites.

3.3.3 Solvent Free Sodium Chloride

Sensitivity Comparisons

The sensitivities calculated using the SCB single point formula are compared to SCI single point formula and the RFD sensitivity estimates with confidence ranges in Figure 3-5. For epsilon (Fig. 3-5a and Fig. 3-5b) and sigma (Fig. 3-5c), the SCB and SCI sensitivities are entirely within the shown confidence ranges. It is interesting to note that the magnitude of the SCI sensitivity is less than the magnitude of the SCB sensitivity when they deviate in Fig. 3-5a-c at short interaction distances, indicating that the SCB covariance correction is greater than the SCI transport correction because of differences in the amount of sensitivity captured due to the different basis set considerations between the two formulae. For sensitivity to charge (Fig. 3-5d), the SCI estimate is significantly below the actual RFD sensitivity and the SCB sensitivity, which is above the RFD confidence range for large interaction distances and below it for intermediate interaction distances. However, both the SCB and SCI sensitivities show much better qualitative agreement for the sensitivity to charge in the sodium chloride system than in the methanol system, suggesting that the effects of electrostatics are more pair-representable in this system.

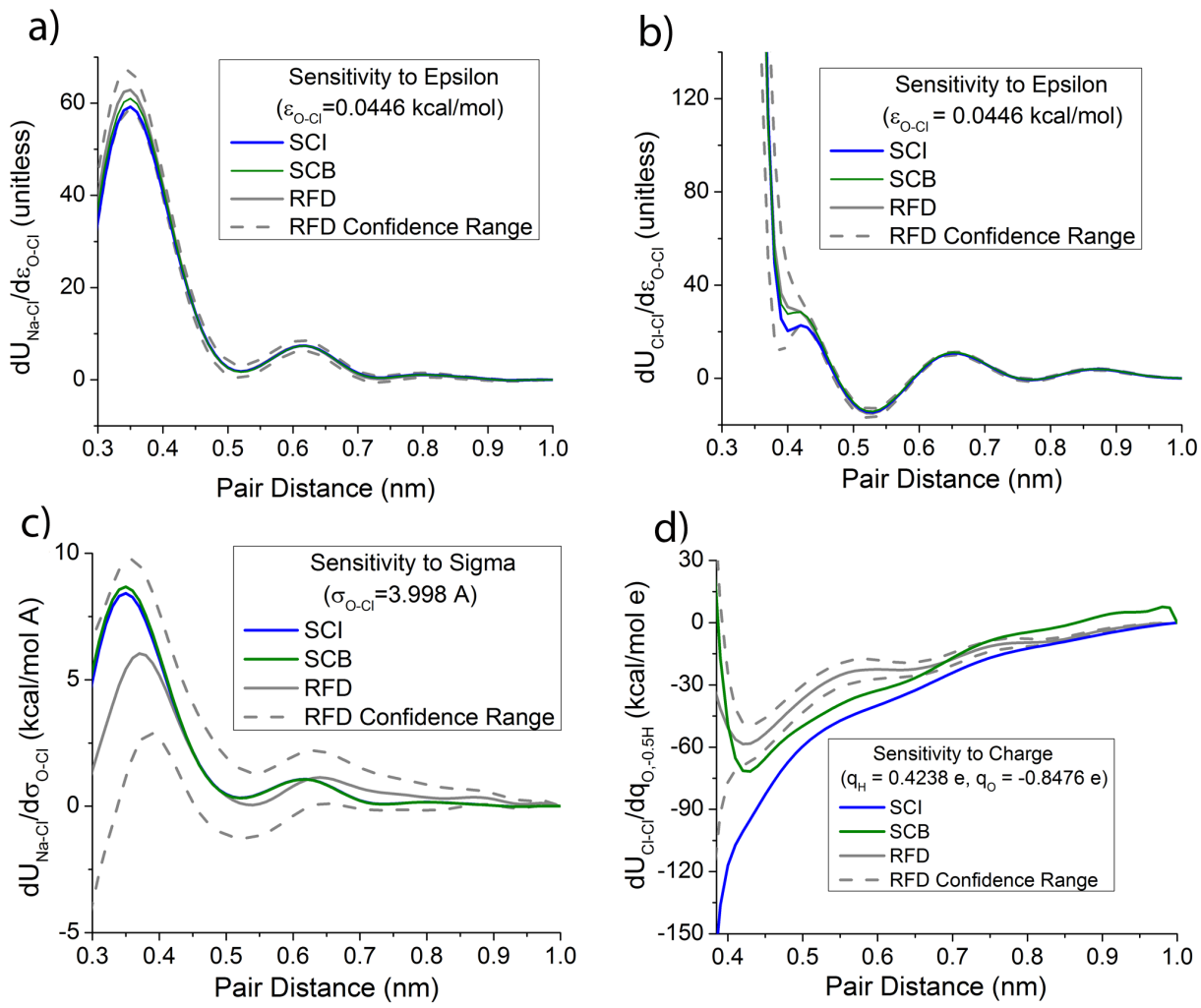

Figure 3–5. Comparison of solvent free sodium chloride Na-Na, Na-Cl, and Cl-Cl interaction potential sensitivities estimated for different interaction parameters between RFD, SCI single point, and SCB single point calculations. Sensitivities of the a) Na-Cl CG potential to the FG Oxygen-Chloride (O-Cl) LJ epsilon, b) Cl-Cl CG potential to the FG Oxygen-Chloride (O-Cl) LJ epsilon, c) Na-Cl CG potential to the Oxygen-Chloride (O-Cl) LJ sigma, and d) Cl-Cl CG potential to the water Oxygen and Hydrogen charge interactions.

Predicted Potentials

Figure 3-6 shows the difference in potentials (from Eq. (3.9)) for different $\delta\lambda$'s from the original ($\delta\lambda = 0$) potential as described in Section 3.2. For the epsilon graph (Fig. 3-6a), the SCB curve agrees with the reweighted curve for small changes until the nonlinearities appear in the reweighted curve, where bias is more of a problem. The SCB curve is also in the same range as the independent trajectories for large changes. The SCI curve deviates from both reference curves for sizable changes, but appears to have the same initial slope as the independent trajectories, which is likely within the linear regime. For the sigma graph (Fig. 3-6b), the SCB curve shows even better agreement with the reweighted curve and the average slope of the independent trajectories curves than in Figure 3-6a. The large difference between the SCI and SCB curves indicates the importance of CG basis set effects in capturing the correlations important for larger changes in parameters. For the charge graph (Fig. 3-6c), the reweighted curve looks quite linear and shows agreement with the independent trajectories only for small changes. The difference between the SCB and reweighted curves from the independent trajectories may be due to underestimation of many-body charge screening effects that the reweighted and SCB methods do not incorporate due to configurational sampling bias. The SCB curve agrees with the reweighted curve, but both overestimate changes in the potential. The SCI curve here seems to have the same average slope as the independent trajectories for positive and negative changes, indicating either less bias or a cancellation of errors.

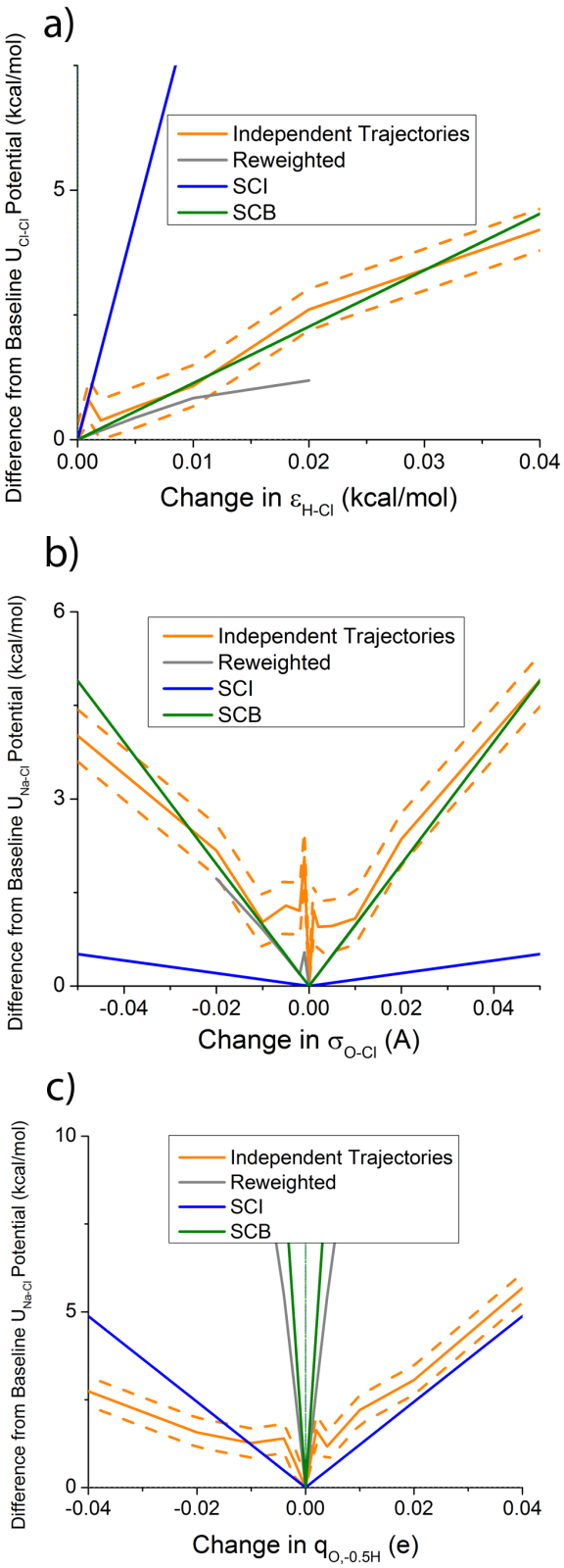

Figure 3–6. Magnitude of change in sodium chloride CG interaction potential from SPC/E water and Joung and Cheatham NaCl parameterization, calculated as a weighted average absolute

difference in predicted potential from a reference potential weighted by the reference RDF, for predictions via independent trajectories, reweighting, and the two single point sensitivities SCI and SCB. Predictions of a) $U_{Cl\text{-}Cl}$ to changes in Hydrogen-Chloride (H-Cl) LJ epsilon, b) $U_{Na\text{-}Cl}$ to changes in Oxygen-Chloride (O-Cl) LJ sigma, and c) $U_{Na\text{-}Cl}$ to changes in Water Hydrogen and Oxygen charges are compared.

CG Simulations

Figure 3-7 shows the RDFs for a selected set of $\delta\lambda$. For the sensitivity to epsilon example the heights of the first peaks predicted from the single-point sensitivities for the Na-Na and Cl-Cl RDFs (not shown, see SI) are slightly overstructured, but the opposite is true for the Na-Cl RDF (Fig. 3-7a). The opposing errors in the RDFs and potentials illustrate the additional problems of fitting the three nonbonded interactions simultaneously. For the sensitivity to sigma example (Fig. 3-7b), the RDF from the SCB predicted potential seems to be in agreement with the actual CG RDF with only minor understructuring of the contact-ion pair. When it comes to the sensitivity to charge (Fig. 3-7c), it is clear that the errors in the sensitivity of the potential carried through to the RDFs as they are all uniformly overstructured. It is not all that surprising that the charge sensitivities from the single point formulae overstructure the ions since this amounts to an underestimate of a many-body screening effect from waters that were coarse-grained out of the simulation. This continues to get worse for larger changes to the charges on the water (Fig. 3-7f). The corrections to the water screening and structure are likely manifested in many-body interactions that are beyond the range of these first order, pair-representable sensitivities. Fortunately, the agreement between the RDFs to epsilon (Fig. 3-7d) and sigma (Fig. 3-7e) parameters is quite good over a larger range of parameter changes. As with the methanol system, the heights of the RDF peaks and valley differ – only slightly so for epsilon and sigma given the

magnitude of the parameter change – while the location of the peaks and valley agree quite well. Thus, using the sensitivities from less highly correlated interactions such as the LJ nonbonded interactions appears to lead to reasonable predicted CG potentials and CG RDFs.

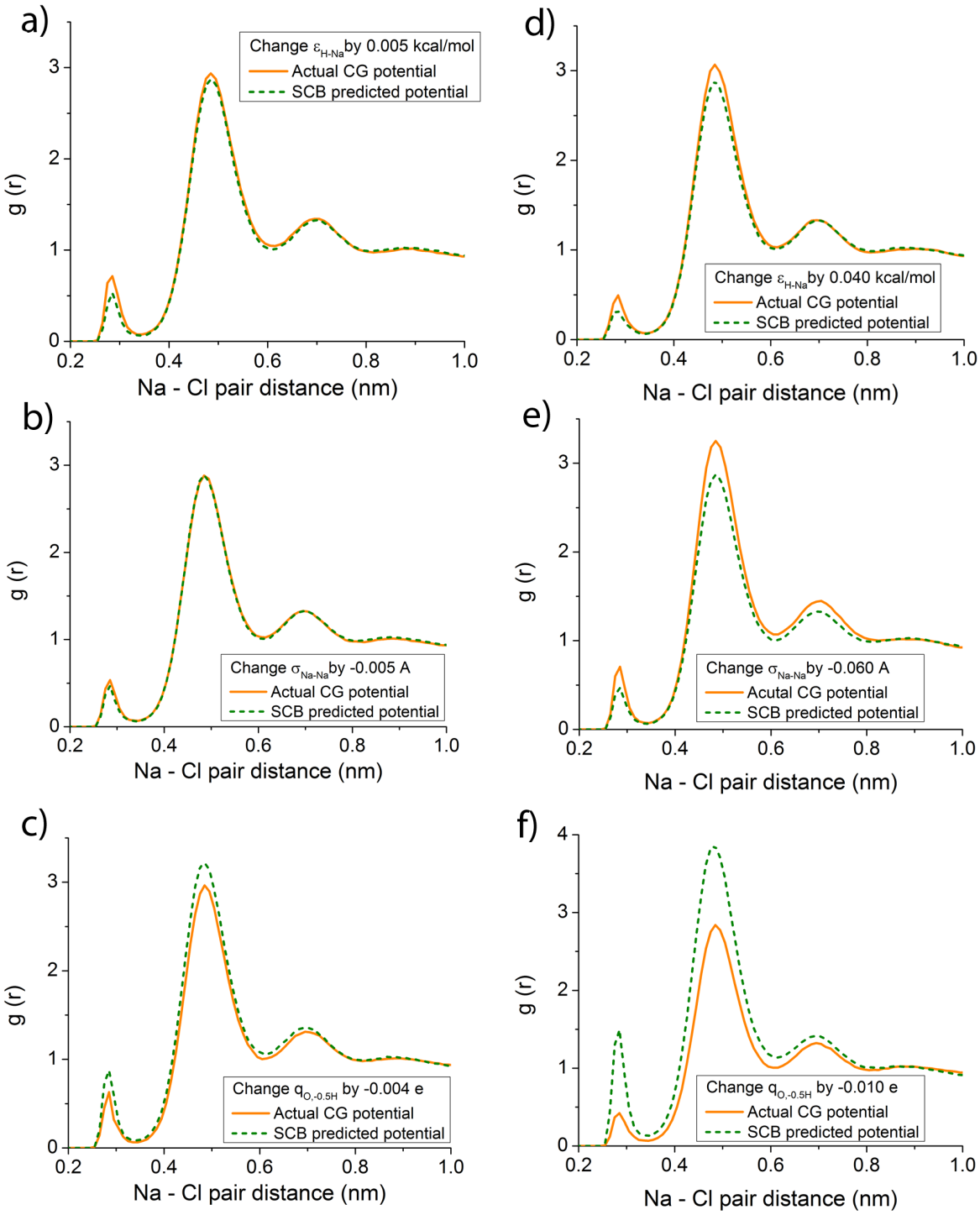

Figure 3–7. Radial distribution functions (RDFs) from CG sodium chloride simulations for the Na-Cl pair distance. a) Changing Hydrogen-Sodium (H-Na) LJ epsilon interaction parameter by 0.005 kcal/mol. b) Changing Sodium-Sodium (Na-Na) LJ sigma interaction parameter by -0.005

A. c) Changing the Water Oxygen charge by -0.004 e and the Water Hydrogen by +0.002 e. d) Changing Hydrogen-Sodium (H-Na) LJ epsilon interaction parameter by 0.040 kcal/mol. e) Changing Sodium-Sodium (Na-Na) LJ sigma interaction parameter by -0.060 A. f) Changing the Water Oxygen charge by -0.010 e and the Water Hydrogen by +0.005 e.

3.4 Discussion

In general, the sensitivities calculated with respect to LJ epsilon parameters show excellent agreement with the RFD sensitivities and the CG RDFs for both the CG methanol and solvent free sodium chloride systems. The agreement in the difference from the baseline potential in Figure 3-3a even well outside the linear regime is particularly noteworthy. The slightly more correlated sensitivities with respect to sigma parameters showed good agreement for the CG methanol system, better agreement than in the sodium chloride system that is more highly coarse-grained. However, the charge sensitivities are only qualitatively correct, reflecting the highly complex many-body correlations due to these long range interactions. In the sodium chloride system, this was reflected in the overstructuring of the CG ion pair RDFs, but the fact that sensitivities to the charge on the atoms in the implicit water molecules are reasonable is promising. Unfortunately, agreement with RFD implies that while these formulae significantly improve on the signal to noise ratio of reweighted finite differences, in many cases they do not address the problem of bias. This work reveals that bias in these estimators, not noise, is the next truly difficult problem to overcome, and it will not be overcome simply with more data as we attempted with this per-particle derivative-matching approach—bias is better dealt with by acquiring better data[79] or more sophisticated estimators.[80-82]

One can expect that the SCB sensitivities will be the most accurate and predictive when the relevant interaction is short range and does not involve many-body correlations as is the case for any short range pair interaction modeling van der Waals interactions such as the Lennard-Jones potential. In contrast, interactions that are long range and involve many-body correlations such as charge interactions are expected to be less accurate and less predictive. While the sensitivity of intramolecular interactions such as bonds, angles, and dihedrals were not investigated, physical considerations suggest that these sensitivities will not be as well-behaved as those for Lennard-Jones parameters because intra-CG-site interactions affect the CG interactions only indirectly through multi-atom correlations, but on the other hand they should be better than the sensitivities for charge interactions because intramolecular interactions are short ranged. Likewise, sensitivity of state parameters such as temperature and volume as well as their conjugate observables entropy and pressure were not investigated here, but we expect the sensitivities to these parameters to be inaccurate because the relevant interactions are long range and involve many-body correlations that may not be pair-representable.

One way to check these heuristics is to compare the SCB estimate to the SCI estimate for each parameter of interest in a given system. If the SCB and SCI estimates agree as in Fig. 3-2a-c and Fig. 3-5 a-c, then the many-body correlations that lead to the correction of the naive sensitivity which the two formulae estimate differently are well-represented in the chosen basis set. This means that the SCB sensitivity is more likely to be accurate and predictive. This feature implies that these SCI and SCB formulae can be used diagnostically to evaluate hypotheses about transferability even when they do not provide accurate linear sensitivity measurements. However, when they do, they also provide previously infeasible checks on the precise size of the

linear regime: in that case, this regime will be precisely the range over which the RFD and SCB estimates agree.

The times when the SCB and SCI single point formulae differ in their sensitivity estimates are significant because the two capture correlated many-body effects in different ways. In particular, any significant correction to the naïve sensitivity from either single point formula indicates the presence of significant multi-body effects and correlations in the sensitivity to that parameter. Differences between them, moreover, specifically indicate that the non-representable many-body effects are folded into the corresponding representable force field for a given basis, which are sensitive to the parameter under study in ways that cannot be represented within that basis set. The difference in the estimates for the sensitivity to epsilon in the solvent free NaCl system, but not the CG methanol system, reflects that the NaCl system is more highly coarse-grained as expected. Also, the difference between the SCB and SCI sensitivity estimates for charge interactions quantifies and confirms the dependence of the effective pair potential on the significant and complex many-body correlations among long range interactions that had been hypothesized previously in the literature.[78]

Noise and modeling error can still remain problematic even when these formulae perform without significant bias. Approaches previously used to improve force matching's ability to deal with these problems could also potentially be brought to bear to improve these sensitivity calculations. For instance, regularization of these sensitivity estimators similar to Lu et al.'s approach could improve the performance of these estimates with noisy data.[70, 83] Alternatively, it may be that structural differences between the CG and FG models due to the use of a finite basis set lead to prediction errors that could be ameliorated via recalculation of the input sensitivity

derivatives, $M^{\dagger}\left(d\nabla u(\mathbf{r}^n;\lambda)/d\lambda\right)$and $du(\mathbf{r}^n;\lambda)/d\lambda$, using statistics from CG sampling as well as FG sampling, as in iterative FM and iterative g-YBG approaches.[84-86]

More systematic study of the differences between calculations by these two formulae may reveal more of the character of important many-body effects in various systems. While we focused on well-known interaction parameters here, it is possible to use these formulae to investigate the addition of arbitrary biases to FG models to examine the effects of arbitrary correlations on the representable correlations in a system. This could reveal interesting experimental control parameters for fine-grained systems, e.g., in discovering manipulable fields conjugate to CG tetrahedrality correlations in water.[87-89] Furthermore, while we studied only pair-representable force fields here and considered only fixed CG basis sets, we can also use it as a criterion for basis set quality, indicating that it could also be useful for basis set design. One can use these formulae with arbitrarily complex CG basis sets and the comparisons between SCI and SCB formulae in various basis sets to choose the ones that will result in the most transferable models across a given parameter space.

Finally, the results indicate that the calculation of sensitivities to nonbonded interaction parameters is good for generating predicted potentials, especially for sensitivities to LJ epsilon and sigma interaction parameters. A direct extension of this work would be to calculate thermodynamic derivatives by repeating the derivations in the theory section with $\lambda = \beta$, the thermodynamic temperature. Since the formulae can only be used to calculate sensitivities to continuous parameters, sensitivities to volume $\lambda = V$ could be calculated in the NPT and μPT ensembles while sensitivities to concentration could be calculated via sensitivities to chemical potential $\lambda = \mu$ in the grand-canonical μPT and μVT ensembles. Another extension of the work would be to calculate the sensitivity of other CG properties or observables such as the RDF g(r)

by applying a chain rule, where the sensitivity of the CG property or observable to the CG potential would be multiplied by the sensitivity of the CG potential to FG interaction parameters as presented in this work.

3.5 Conclusion

In this chapter, new reweighting-free formulae for the calculation of the sensitivity of CG potentials and force fields to changes in the underlying FG models' interaction parameters and state point were presented that require only a single trajectory for calculation. In the results, the SCB estimates predicted sensitivities to LJ epsilon and sigma parameters that were quantitatively correct to within the confidence ranges from RFDs, representative of the practical state of the art. The single point formula does not require *a priori* knowledge of the linear sensitivity regime for a given parameter and can be useful in generating predicted CG potentials for other interaction parameters, as demonstrated by the agreement of the CG RDFs from independent trajectories with predicted potentials using this sensitivity. Of the predictive sensitivity measures examined here, the single point methods provide the lowest noise estimates of all, providing the same sensitivities as RFDs at reduced computational cost, with comparable bias, and without ambiguity concerning the size of the linear regime.

Finally, beyond their purely computational significance, these results also serve to shed light onto relatively unexplored subtleties of CG representability. Consideration of both the SCB and SCI formulae offers a new window onto the fundamental theoretical problems of representing transfer of CG models between state points, providing a vivid example of how the change from FG model to FG model in the CG-representable part of the correlations may not always be the same as the CG-representable part of the change in correlations from FG model to FG model—even at the level of linear response. While this is sometimes mentioned in discussions of the

foundations of coarse-graining, that subtlety is rarely investigated as a practical effect with fundamental physical importance in its own right. However, establishing one's foundations is also an eminently practical thing to do. We hope this work will spur deeper investigations into the theoretical interplays between representability and transferability, two of the most important challenge areas in state of the art of coarse-graining and CG modeling, in addition to providing new computational tools.

3.6 Appendix A: Derivation of the SCB single-point formula

The self-consistent basis single-point sensitivity formula describes the derivative with respect to system parameters of variationally force matched finite-basis approximations to the true many-body FES. The usual force-matching normal equations for a PMF approximated as a linear combination of a set of basis functions ψ_d with coefficients ϕ_d and λ are

$$\frac{1}{N_t N}\sum_{t=1}^{N_t}\sum_{I=1}^{N}\frac{d\psi_d}{d\mathbf{R}_I}\cdot\sum_{d'}\frac{d\psi_{d'}}{d\mathbf{R}_I}\phi_{d',\lambda}=\frac{1}{N_t N}\sum_{t=1}^{N_t}\sum_{I=1}^{N}\frac{d\psi_d}{d\mathbf{R}_I}\cdot\mathbf{F}_I(\mathbf{r}_t^n;\lambda), \tag{3.10}$$

where $\mathbf{F}_I$ are CGed forces from sampled atomistic configurations. This is an equation valid for sampling from a system with fixed parameter λ. In order to take the derivative of this with respect to λ, one must find a way to express the sampling density as a differentiable function of λ in this expression. One option is to assume all sampling is run at a reference λ_0 and then perform a weighted least-squares optimization using the reweighting factors (see Eq. (3.5) in the main text) rather than a uniformly-weighted least-squares optimization. Using the usual equations for weighted least squares, one gets the normal equations

$$\frac{1}{N_t N}\sum_{t=1}^{N_t} w(\mathbf{r}_t^n;\lambda,\lambda_0)\sum_{I=1}^{N}\frac{d\psi_d}{d\mathbf{R}_I}\cdot\sum_{d'}\frac{d\psi_{d'}}{d\mathbf{R}_I}\phi_{d',\lambda}=\frac{1}{N_t N}\sum_{t=1}^{N_t} w(\mathbf{r}_t^n;\lambda,\lambda_0)\sum_{I=1}^{N}\frac{d\psi_d}{d\mathbf{R}_I}\cdot\mathbf{F}_I(\mathbf{r}_t^n;\lambda), \tag{3.11}$$

which are in principle valid for any λ with samples taken with respect to any λ_0, though of course only practical when λ is close to λ_0. First define

$$G_{d,d'}(\lambda,\lambda_0)=\frac{1}{N_t N}\sum_{t=1}^{N_t} w\left(\mathbf{r}_t^n;\lambda,\lambda_0\right)\sum_{I=1}^{N}\frac{d\psi_d}{d\mathbf{R}_I}\frac{d\psi_{d'}}{d\mathbf{R}_I} \text{ and} \tag{3.12}$$

$$b_d(\lambda,\lambda_0)=\frac{1}{N_t N}\sum_{t=1}^{N_t} w(\mathbf{r}_t^n;\lambda,\lambda_0)\sum_{I=1}^{N}\frac{d\psi_d}{d\mathbf{R}_I}\cdot\mathbf{F}_I(\mathbf{r}_t^n;\lambda)\ . \tag{3.13}$$

Now, taking the derivative of both sides, one gets

$$\sum_{d'} G_{d,d'}(\lambda,\lambda_0)\frac{d\phi_{d',\lambda}}{d\lambda}=\frac{d}{d\lambda}\left(b_{d,d'}(\lambda,\lambda_0)-\sum_{d'} G_{d,d'}(\lambda,\lambda_0)\phi_{d',\lambda^*}\right), \tag{3.14}$$

where λ^* is equal to λ, but does not change with λ in that expression. This is a force-matching-like equation for the change in the expansion coefficients, which give the change in the PMF when multiplied by the basis functions. The equation matches the change from the true target forces, b_d, from the predicted forces with fixed PMF, $G_{d,d'}$ $\phi_{d'}$, and adjusted sampling.

Evaluating the derivatives is easiest after re-expanding the new notation

$$\frac{1}{N_t N}\sum_{t=1}^{N_t} w\left(\mathbf{r}_t^n;\lambda,\lambda_0\right)\sum_{I=1}^{N}\frac{d\psi_d}{d\mathbf{R}_I}\cdot\sum_{d'}\frac{d\psi_{d'}}{d\mathbf{R}_I}\frac{d\phi_{d',\lambda}}{d\lambda}=$$

$$=\frac{d}{d\lambda}\left(\frac{1}{N_t N}\sum_{I=1}^{N_t} w\left(\mathbf{r}_t^n;\lambda,\lambda_0\right)\sum_{I=1}^{N}\frac{d\psi_d}{d\mathbf{R}_I}\cdot\left(\mathbf{F}_I\left(\mathbf{r}_t^n;\lambda\right)-\sum_{d'}\frac{d\psi_{d'}}{d\mathbf{R}_I}\phi_{d',\lambda^*}\right)\right)$$

$$+\frac{1}{N_t N}\sum_{t=1}^{N_t}\frac{dw\left(\mathbf{r}_t^n;\lambda,\lambda_0\right)}{d\lambda}\sum_{I=1}^{N}\frac{d\psi_d}{d\mathbf{R}_I}\cdot\left(\mathbf{F}_I\left(\mathbf{r}_t^n;\lambda\right)-\sum_{d'}\frac{d\psi_{d'}}{d\mathbf{R}_I}\phi_{d',\lambda^*}\right)$$

$$= \frac{1}{N_t N}\sum_{I=1}^{N_t} w\left(\mathbf{r}_t^n;\lambda,\lambda_0\right)\beta\left(-\frac{du(\mathbf{r}_t^n;\lambda)}{d\lambda}+\sum_{t'=1}^{N_t} w\left(\mathbf{r}_{t'}^n;\lambda,\lambda_0\right)\frac{du(\mathbf{r}_{t'}^n;\lambda)}{d\lambda}\right)$$
$$\cdot\sum_{I=1}^{N}\frac{d\psi_d}{d\mathbf{R}_I}\cdot\left(\mathbf{F}_I(\mathbf{r}_t^n;\lambda)-\sum_{d'}\frac{d\psi_{d'}}{d\mathbf{R}_I}\phi_{d',\lambda^*}\right)$$

$$+\frac{1}{N_t N}\sum_{I=1}^{N_t} w\left(\mathbf{r}_t^n;\lambda,\lambda_0\right)\sum_{I=1}^{N}\frac{d\psi_d}{d\mathbf{R}_I}\frac{d\mathbf{F}_I(\mathbf{r}_t^n;\lambda)}{d\lambda}$$

$$= \frac{1}{N_t N}\sum_{I=1}^{N_t} w\left(\mathbf{r}_t^n;\lambda,\lambda_0\right)\sum_{I=1}^{N}\frac{d\psi_d}{d\mathbf{R}_I}\cdot\left(\begin{array}{l}\dfrac{d\mathbf{F}_I(\mathbf{r}_t^n;\lambda)}{d\lambda}-\\ \beta\left(\dfrac{du(\mathbf{r}_t^n;\lambda)}{d\lambda}-\sum_{t'=1}^{N} w\left(\mathbf{r}_{t'}^n;\lambda,\lambda_0\right)\dfrac{du(\mathbf{r}_{t'}^n;\lambda)}{d\lambda}\right)\left(\mathbf{F}_I(\mathbf{r}_t^n;\lambda)-\sum_{d'}\dfrac{d\psi_{d'}}{d\mathbf{R}_I}\phi_{d',\lambda^*}\right)\end{array}\right)$$

, (3.15)

which is just a weighted force matching for the newly-apparent framewise sensitivities of forces in the parentheses, which are straightforward to calculate after force-matching first to find $\phi_{d,\lambda}$. Using a finite sum with some large number of samples provides a practical calculation scheme. Replacing the sums with ergodic averages, however, in the complete basis set limit these normal equations correspond to equation 5, the covariance-like SCB formula described in the main text. In a finite basis set and in the long time limit, it corresponds to the λ-derivative of the g-YBG equations.

3.7 Appendix B: Derivation of the SCI single-point formula

The self-consistent iterative single-point sensitivity formula is based on using a finite basis set to represent the per-particle-position derivatives of the full many-body sensitivity. To derive this, start with the definition of the many-body CG sensitivity

$$\frac{dU(\mathbf{R}^N;\lambda)}{d\lambda}=\left\langle\frac{du(\mathbf{r}^n;\lambda)}{d\lambda}\right\rangle_{\mathbf{R}^N,\lambda} \tag{3.16}$$

take the derivative with respect to all CG particle positions

$$\frac{d\nabla_R U\left(\mathbf{R}^N;\lambda\right)}{d\lambda}=\nabla_R\left(\frac{\int d\mathbf{r}^n\delta\left(\mathbf{R}^N-M\left(\mathbf{r}^n\right)\right)\frac{du\left(\mathbf{r}^n;\lambda\right)}{d\lambda}e^{-\beta u\left(\mathbf{r}^n;\lambda\right)}}{\int d\mathbf{r}^n\delta\left(\mathbf{R}^N-M\left(\mathbf{r}^n\right)\right)e^{-\beta u\left(\mathbf{r}^n;\lambda\right)}}\right), \tag{3.17}$$

apply the product rule to see

$$\begin{aligned}\frac{d\nabla_R U\left(\mathbf{R}^N;\lambda\right)}{d\lambda}=&\left(\frac{\int d\mathbf{r}^n\nabla_R\delta\left(\mathbf{R}^N-M\left(\mathbf{r}^n\right)\right)\frac{du\left(\mathbf{r}^n;\lambda\right)}{d\lambda}e^{-\beta u\left(\mathbf{r}^n;\lambda\right)}}{\int d\mathbf{r}^n\delta\left(\mathbf{R}^N-M\left(\mathbf{r}^n\right)\right)e^{-\beta u\left(\mathbf{r}^n;\lambda\right)}}\right),\\&-\left\langle\frac{du\left(\mathbf{r}^n;\lambda\right)}{d\lambda}\right\rangle_{\mathbf{R}^N,\lambda}\left(\frac{\int d\mathbf{r}^n\nabla_R\delta\left(\mathbf{R}^N-M\left(\mathbf{r}^n\right)\right)e^{-\beta u\left(\mathbf{r}^n;\lambda\right)}}{\int d\mathbf{r}^n\delta\left(\mathbf{R}^N-M\left(\mathbf{r}^n\right)\right)e^{-\beta u\left(\mathbf{r}^n;\lambda\right)}}\right)\end{aligned} \tag{3.18}$$

and simplify using the integration by parts formulas used by Dama et al.[90] to get

$$\begin{aligned}\frac{d\nabla_R U\left(\mathbf{R}^N;\lambda\right)}{d\lambda}=&\left\langle M^\dagger\left(\frac{d\nabla_r u\left(\mathbf{r}^n;\lambda\right)}{d\lambda}\right)\right\rangle_{\mathbf{R}^N,\lambda}\\&-\left\langle\frac{du\left(\mathbf{r}^n;\lambda\right)}{d\lambda}M^\dagger\left(\nabla_r\beta u\left(\mathbf{r}^n\right)\right)\right\rangle_{\mathbf{R}^N,\lambda}+\left\langle\frac{du\left(\mathbf{r}^n;\lambda\right)}{d\lambda}\right\rangle_{\mathbf{R}^N,\lambda}\left\langle M^\dagger\left(\nabla_r\beta u\left(\mathbf{r}^n\right)\right)\right\rangle_{\mathbf{R}^N,\lambda}\end{aligned}. \tag{3.19}$$

Finally, rearrangement and grouping leads to Eq. (3.7), the transport-like SCI equation in the main text. This corresponds to the λ-derivative of the g-YBG equations with a complete basis set.

Chapter 4

On the Representability Problem and the Physical Meaning of Coarse-grained Models

This chapter is reproduced from J. W. Wagner, J. F. Dama, A. E. P. Durumeric, and G. A. Voth, "On the Representability Problem and the Physical Meaning of Coarse-grained Models", Journal of Chemical Physics, 145, 044108 (2016),[91] with the permission of AIP Publishing.

4.1 Introduction

Models with a range of resolutions can be used to describe the same physical system, with each model's resolution providing the context to interpret its representation. For example, coarse-grained (CG) models use fewer particles than their fine-grained (FG) counterparts to represent the same system, while FG classical atomistic models explicitly represent each atom (nucleus) with a single point particle.[14-18] One "bottom-up" example of a CG model that relates the FG to CG representations is the Multiscale Coarse-graining (MS-CG) method,[22, 29-40] while another is the relative entropy approach.[25, 27, 92, 93] In all cases, these models may claim to achieve physical significance from comparison to experiment and this comparison is between experimental observables and corresponding model observables. Therefore, the relationship between each model's observables and experimental observables must first be firmly established in order for these models to be as meaningful as possible.

Often, however, the relationships between models and experiment can be unclear. Common statistical mechanical results and intuitive structural relationships establish connections between atomistic models and experiment,[20, 21, 74, 94, 95] but the model's connection to experiment ultimately depends on the resolution of the model, as some authors have attempted to make clear.[42, 43, 96] This concept of resolution-dependent interpretation has been applied to studies of how thermodynamic observable representations may change based on resolution and

thermodynamic ensemble, for example, by D'Adamo et al.,[56] Das and Andersen,[31] and Dunn and Noid.[97] However, though this literature focuses on thermodynamic observables,[31, 42, 43, 56, 96-100] the issue of CG observable representation is more fundamental: it concerns every aspect of CG model interpretation involving comparison with experiments or FG physics. Thus, *the recognition that CG observables are not simply analogs of their FG counterparts is fundamental to understanding and ultimately addressing the issue of representability in coarse-graining.* This recognition is commonly overlooked or ignored in the CG modeling and simulation literature. A more nuanced understanding is therefore needed to interpret CG models so that they have more meaningful connections to experiment.

Often, models are parameterized using observable constraints so that they reproduce a given experimental observable using a chosen observable expression.[15, 35, 62] Any one experimental observable can always be reproduced this way. However, models need to reproduce several experimental observables simultaneously, and a model cannot reproduce several observables simultaneously if their corresponding constraints conflict. For example, the Henderson uniqueness theorem guarantees a unique radial distribution given a pair potential.[101] To reproduce any additional experimental observable such as the pressure with this fixed pair potential, the model observable must be able to reproduce the experimental values using the previously determined pair potential;[42] otherwise, the observables are incompatible. While adding three-body interactions[32, 38, 87] or density dependence[58-60, 102, 103] can improve observable compatibility tradeoffs, it is not computationally tractable to add additional interactions indefinitely. Thus, one needs to choose the set of observable expressions carefully if the model is to successfully reproduce all of the corresponding experimental observables simultaneously.

Systematic coarse-graining (e.g., the multiscale coarse-graining methodology[22, 29-40] as one such case) provides an illuminating case to investigate CG observable compatibility. In systematic coarse-graining, a CG model of arbitrary resolution is defined in terms of a given FG model using a mapping and an effective potential. The mapping M defines CG configuration variables $\mathbf{R}^N$ in terms of the FG configuration variables $\mathbf{r}^n$ as the product of the individual CG particle mapping operators M_I for each CG particle $\mathbf{R}_I$, such that $\delta\left(M\left(\mathbf{r}^n\right)-\mathbf{R}^N\right)=\prod_{I=1}^{N}\delta\left(M_I\left(\mathbf{r}^n\right)-\mathbf{R}_I\right)$.[22] Then the effective CG potential $U_{CG}\left(\mathbf{R}^N\right)$, i.e., the CG configurational free energy for a specific CG configuration $\mathbf{R}^N$ can be written in terms of the FG potential $U_{FG}\left(\mathbf{r}^n\right)$ as[39]

$$e^{-\beta U_{CG}\left(\mathbf{R}^N\right)} \propto \int d\mathbf{r}^n \delta\left(M\left(\mathbf{r}^n\right)-\mathbf{R}^N\right) e^{-\beta U_{FG}\left(\mathbf{r}^n\right)}, \tag{4.1}$$

where $\beta=\left(k_B T\right)^{-1}$, k_B is Boltzmann's constant, and T is the temperature. The CG observable needs to ensure that the FG ensemble average of an observable property $A_{FG}\left(\mathbf{r}^n\right)$ in terms of the FG potential $U_{FG}\left(\mathbf{r}^n\right)$ is equal to the CG Boltzmann ensemble average of $A_{CG}\left(\mathbf{R}^N\right)$ in terms of the CG potential $U_{CG}\left(\mathbf{R}^N\right)$. One obvious choice that satisfies this requirement is

$$A_{CG}\left(\mathbf{R}^N\right)=\left\langle A_{FG}\left(\mathbf{r}^n\right)\right\rangle_{\mathbf{R}^N}=\frac{\int d\mathbf{r}^n A_{FG}\left(\mathbf{r}^n\right)\delta\left(M\left(\mathbf{r}^n\right)-\mathbf{R}^N\right)e^{-\beta U_{FG}\left(\mathbf{r}^n\right)}}{e^{-\beta U_{CG}\left(\mathbf{R}^N\right)}} \tag{4.2}$$

(see Appendix A for details). Other choices for the CG observable can also ensure that the FG and CG ensemble averages are equal; for instance, if the CG observable were defined simply as the constant ensemble average value of the FG observable, but no others are so immediately

obvious albeit trivial. In particular, this choice and no other guarantees the equality of the ensemble average between the CG and FG observables under any possible bias potential applied to the CG degrees of freedom. In addition to ensuring the equivalence of ensemble averages, Eq. (4.2) also implies that the CG observable for a given CG configuration is equal to the ensemble average of all FG configurations that map to the given CG configuration. Thus, the relationship here between FG and CG observables creates a set of compatible CG observables that can be used to simultaneously reproduce experimental observables under a range of conditions imposed at the CG level. As a result, all experimental observable relationships are present between CG observables that satisfy Eq. (4.2) since Eq. (4.2) establishes a way to identify an indefinite number of compatible CG observables. It should be noted that Eq. (4.2) is greatly simplified if the target observable *for both the FG and CG levels* depends only on the CG coordinates $\mathbf{R}^N$. One such example is if one is interested in the structure of a large biomolecular system, then the carbon-alpha atoms, for instance, might be a good enough choice to understand that structure of both the FG and CG models. However, when one wishes to calculate thermodynamic and many other structural properties, it is very rarely the case that A_{FG} and A_{CG} depend on the same set of variables.

The relationships in Eq. (4.2) also have direct implications for bottom-up CG models. In bottom-up coarse-graining, the FG model is usually parameterized to correspond with experiment (Figure 4-1a) or from "first principles" quantum calculations (which are presumed to also agree with experiment, even if this is rarely the case due to inaccuracy in the "first principles" quantum method). Then, the CG model is constructed using the mapping operator, which establishes a strict correspondence between FG and CG model configurations. This fully specifies the CG model given configurational expressions (e.g., radial distribution functions,

RDFs). So, additional observable expressions incompatible with these configurational observables cannot properly correspond with experiment (Figure 4-1b). However, Eq. (4.2) provides a way to identify observables compatible with these configurational observables using this strict correspondence between FG and CG models. The problem is that it is very difficult to know an explicit form of $A_{CG}\left(\mathbf{R}^{N}\right)$ in Eq. (4.2) beyond its formal expression appearing in that equation.

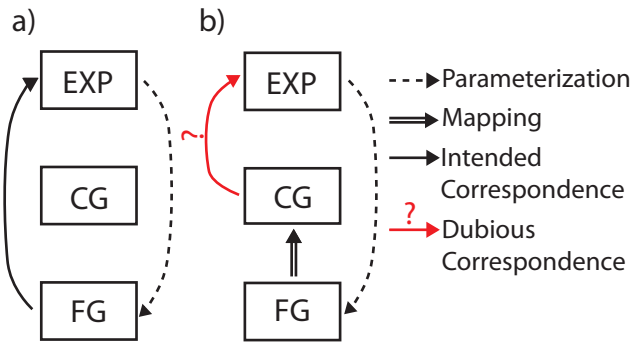

Figure 4–1. The relationships between experiment (EXP), fine-grained (FG), and coarse-grained (CG) models in bottom-up CG models: a) relationship between experiment and FG models, and b) the intended relationship between CG models and experiment. The dashed lines show parameterization. The solid lines show intended correspondences between the models, while the red line with a question mark indicates a dubious correspondence. The double line indicates the strict correspondence from FG configurations to CG configurations through the mapping operator.

On the other hand, top-down CG models do not have a strict correspondence with a specific FG model because they are parameterized using experimental data or "bulk" FG simulations directly. This loose correspondence between FG and top-down CG models means one might decide to arbitrarily choose CG expressions for physical observables with a similarly loose connection to a FG model, but even then one still needs a compatible set of CG observables in

order for the CG model to fully correspond with experiment (see Figure 4-2a), Unfortunately, the loose correspondence between FG and top-down CG models also means that the mapping needed to evaluate Eq. (4.2) is not explicitly defined (Figure 4-2b). As a result, models such as MARTINI[104-107] and mW[87] parameterized from and interpreted using incompatible observables, do not obviously correspond to an underlying atomistic model for any real physical system. That is, they are purely models at the CG level *defined* to represent certain aspects of reality. However, the representation of observable calculations must then also be viewed a part of that CG model but has no real connection to an expression for the observable in the underlying FG system.

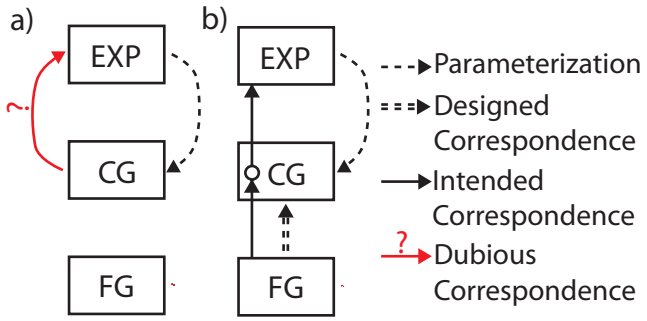

Figure 4–2. The relationships between experiment (EXP), fine-grained (FG), and coarse-grained (CG) models in top-down CG models: a) the relationship between a top-down CG model and experiment and b) the expected relationship between a FG model, a top-down CG model, and experiment. Dashed lines show parameterization. Solid lines show intended correspondences between the models, while the red line with a question mark indicates a dubious correspondence. The double dashed line indicates an intuitive, designed correspondence from the FG model to the CG model.

The present work discusses CG observable representation as an issue of model interpretation based on the correspondence between FG models, CG models, and experiment. The strict

correspondence in systematic, bottom-up coarse-graining between FG and CG models creates a compatible set of CG observables given by Eq. (4.2). In some cases, these may resemble FG observables; however, these expressions often *still differ* by the necessary introduction of a state-dependent CG potential term in the CG observable.[31, 42, 43, 56, 96-98, 100, 108, 109] On the other hand, there may be very little resemblance in general. For instance, Eq. (4.2) can define useful CG observables for what may appear to be purely FG quantities based on configuration variables integrated out of the CG model. Using results from analytical systems, we discuss the validity of previously used observable expressions, introduce new ones, and discuss implications for top-down CG models.

The remainder of this chapter is structured as follows: Section 2 discusses the correspondence in systematic coarse-graining between FG models, CG models, and experiment in depth and shows how the CG versions of FG observables often change by (and necessitate) the inclusion of extra terms. Section 3 uses analytical ideal polymer chains to demonstrate the various issues with using incorrect but seemingly intuitive CG observables in top-down and bottom-up CG models. Section 4 provides conclusions.

4.2 Theory

CG observables that satisfy Eq. (4.2) reproduce both experimental and FG model observables. These CG observables are the average of all FG observable values from FG configurations consistent with a given CG configuration. Thus the FG observable distribution projected onto the CG model is the CG observable distribution, and consequently a CG observable may have a very different functional form than its corresponding FG observable. FG observables are not the same as GG observables, just as FG force fields differ from CG force

fields.[22, 40, 110] In order to maximize the benefits of Eq. (4.2), one should consider that a CG observable's functional form may not be the same as the form of the atomistic observable expressions.

The similarity of FG and CG observable expressions is determined by the complexity of the projection of the FG observable onto the CG model. A naïve CG observable expression treats the CG configuration variables as if they were FG configuration variables, as mentioned earlier (although this is rarely the case). However, the CG observable as a function of CG configuration variables must represent contributions from the eliminated configuration variables as well as the naïve contributions from these CG configuration variables. The failure to represent the observable contributions from eliminated configuration variables prevents naïve CG observables from reproducing the corresponding FG observable distribution. Therefore, using naïve CG observable expressions can prevent researchers from accessing the full predictive power of CG models.

There are three different possible relationships between FG observables and CG observables satisfying Eq. (4.2). At one extreme, the FG and CG observable expressions may be the same. For example, the center of mass (COM) radial distribution function (RDF) from a COM CG model is directly identifiable: The COM is one of the CG configuration variables. In special cases like this, naïve CG observables faithfully represent the entire projection of the FG observable. At the other extreme, some CG observables expressions have no clear similarity to the corresponding FG expression at all. For example, the magnitude of the molecular dipole of a CG model with a single site per molecule can appear undefined: All of the FG configuration variables traditionally associated with the FG observable have been eliminated. However, a bottom-up expression for the magnitude of the molecular dipole satisfying Eq. (4.2) will capture

the average behavior of these eliminated configuration variables. By construction it reproduces the projection of the magnitude of the molecular dipole even though this observable appears to be a purely FG quantity. In between these two extremes, there is a broad category of CG observables whose expressions resemble the FG observable expression in part, but not entirely.

For example, all FG configuration variables contribute to the FG pressure. Therefore, the contributions of the eliminated configuration variables must also enter into the CG observable somehow. For pressure, neglecting contributions from eliminated configuration variables treats the CG model as if it were a system of indivisible particles (i.e., a fully atomistically resolved system). While a naïve expression for CG pressure could be used in the construction of a CG model, this choice is not compatible with the structural observables implied in the construction of bottom-up CG models. Determining if an observable representation is compatible with the system definition requires one to be aware of the model's resolution in order to understand what it actually represents. However, CG observables satisfying Eq. (4.2) correspond to the appropriate resolution and reproduce experimental observables simultaneously with the structural observables related to the system definition. In contrast, naïve CG observables defined based solely on analogy with expressions for FG observables neglect essential physics of the CG correspondence by describing a system of possibly higher effective resolution than the CG representation itself. Such observables are incompatible with the system definition and interfere with meaningful model interpretation, as its connections to an actual physical system are unclear. For example, if one uses a top-down CG model in a simulation using a CG MD software package, then simply plugs it in to the well known virial expression for the pressure, but that virial expression is used in terms of only the CG variables as if they were FG variables, then when the simulation outputs "1 atm" as its system pressure this value has no clear relationship to

the actual pressure of the real FG system that one is attempting to emulate with the top-down CG model. Indeed, that "pressure" is a part of the CG model as well, but this is purely a part of the overall model because the connection to the FG system such as in Eq. (4.2) has not been followed. (The actual FG pressure could be 1000 atm, 1 atm, or 0.1 atm.) The connections between FG and CG models afforded by the statistical mechanics inherent in Eq. (4.2) has been lost.

However, in some instances write CG observable expressions as the sum of their FG counterparts plus additional terms that capture contributions from eliminated variables. In the case of thermodynamic observables, this extra part is usually directly related to the *transferability* of a CG model, and therefore to the thermodynamic state-dependence of the CG interactions – a dependence which is often ignored but is clearly a fact based on any reasonable statistical mechanical formulation of the effective potential for a CG system. Though this approach of defining state-dependent CG potentials may not be as fundamental as Eq. (4.2), it has strong precedent in the literature on CG thermodynamics,[31, 42, 43, 56, 96-98, 100, 108, 109] and when used correctly it is can often be shown to be equivalent to Eq. (4.2). An especially revealing case is that of our previously mentioned example, i.e., pressure.[43, 98] The pressure, P, of a CG configuration if taken straight from the FG virial expression would be

$$P_{naïve}\left(\mathbf{R}^N\right)=\frac{\rho_{CG}}{\beta}+\frac{\mathbf{R}^N}{3V}\cdot\frac{dU_{CG}(\mathbf{R}^N)}{d\mathbf{R}^N}\ , \tag{4.3}$$

where $\rho_{CG}=N/V$ is the density and V is the system volume. In order to actually derive this expression from thermodynamics rather than simply using it in an *ad hoc* fashion, one must make the assumption that the effective CG interactions are *volume-independent*, so that

$$P_{CG,naïve}\left(\mathbf{R}^N\right)=-\left.\frac{dF_{CG}\left(\mathbf{R}^N\right)}{dV}\right|_{\mathbf{X}^N}=\frac{\rho_{CG}}{\beta}+\frac{\mathbf{R}^N}{3V}\cdot\frac{-dU_{CG}(\mathbf{R}^N)}{d\mathbf{R}^N}\ , \tag{4.4}$$

where F_{CG} is a naïve coarse-grained configurational free energy defined as $F_{CG}\left(\mathbf{R}^N\right)=U_{CG}\left(\mathbf{R}^N\right)-Nk_BT\log V$, $\mathbf{X}^N\equiv V^{-1/3}\mathbf{R}^N$ is the volume-scaled configuration, and $\rho_{CG}=N/V$ is the apparent CG density. This is the naïve CG observable expression defined via direct analogy to the FG observable expression. On the other hand, one could derive an expression valid even if the effective CG interaction were volume-dependent. Then, the expression for pressure from thermodynamics has an extra term from the volume derivative, which acts on both the configuration variables and the volume-dependence:

$$P_{CG}\left(\mathbf{R}^N\right)=-\left.\frac{dF_{CG}\left(\mathbf{R}^N;V\right)}{dV}\right|_{\mathbf{X}^N}=\frac{\rho_{CG}}{\beta}+\frac{\mathbf{R}^N}{3V}\cdot\left.\frac{-dU_{CG}(\mathbf{R}^N;V)}{d\mathbf{R}^N}\right|_{V}-\left.\frac{dU_{CG}(\mathbf{R}^N;V)}{dV}\right|_{\mathbf{R}^N}. \tag{4.5}$$

This expression for CG pressure with a volume-dependent effective CG interaction corresponds exactly to expressions in the literature expressed in terms of density derivatives rather than volume derivatives.[43] However, it is important to note that it satisfies Eq. (4.2), and not by accident:

$$\begin{aligned}P_{CG}\left(\mathbf{R}^N\right)=\left\langle P_{FG}\left(\mathbf{r}^n\right)\right\rangle_{\mathbf{R}^N}&=\left\langle\frac{\rho_{FG}}{\beta}+\frac{\mathbf{r}^n}{3V}\cdot\frac{-dU_{FG}\left(\mathbf{r}^n\right)}{d\mathbf{r}^n}\right\rangle_{\mathbf{R}^N}\\&=\frac{\rho_{CG}}{\beta}+\frac{\mathbf{R}^N}{3V}\cdot\frac{-dU_{CG}\left(\mathbf{R}^n;V\right)}{d\mathbf{R}^N}-\left.\frac{dU_{CG}\left(\mathbf{R}^N;V\right)}{dV}\right|_{\mathbf{R}^N}\end{aligned} \tag{4.6}$$

(see Appendix B for details). This expression makes it clear that fitting virials is not the same as fitting pressures for each CG configuration.

Now, consider the configurational internal energy, E, starting from the viewpoint of Eq. (4.2) . The configurational internal energy, $E(\mathbf{R}^N)$, defined as an average over FG configurations of the system but for a fixed CG configuration is given by

$$E_{CG}\left(\mathbf{R}^N\right)=-\frac{d\ln Z_{\mathbf{R}^N}}{d\beta}=\left\langle U_{FG}\left(\mathbf{r}^n\right)\right\rangle_{\mathbf{R}^N} , \tag{4.7}$$

where $Z_{\mathbf{R}^N}$ is the partition function of all FG configurations consistent with a given $\mathbf{R}^N$. Using Eq. (4.2), one can show that the true expression for the configurational internal energy in terms of only CG variables is

$$E_{CG}\left(\mathbf{R}^N\right)=\frac{d\beta U_{CG}\left(\mathbf{R}^N;\beta\right)}{d\beta}=U_{CG}\left(\mathbf{R}^N;\beta\right)+\beta\frac{dU_{CG}\left(\mathbf{R}^N;\beta\right)}{d\beta} \tag{4.8}$$

(see Appendix C for details). This is the result one would obtain if one recognizes that the effective CG interaction is temperature-dependent. Conveniently, it is the sum of a temperature-dependent term and the naïve CG configurational internal energy (up to the usual irrelevant constant)

$$E_{CG,naïve}\left(\mathbf{R}^N\right)=-\frac{1}{\beta}\ln p\left(\mathbf{R}^N\right)=U_{CG}\left(\mathbf{R}^N\right) , \tag{4.9}$$

where $p\left(\mathbf{R}^N\right)$ is the probability of CG configuration $\mathbf{R}^N$. It is important to note that this expression applied to the FG configuration variables instead of the CG configuration variables gives the correct FG configurational internal energy. However, it does not behave the same when applied to CG models, where it gives the CG configurational free energy.

Likewise, the configurational entropy S_{CG} of a set of CG configurational states ω_{CG} is, according to Eq. (4.2), the Gibbs entropy for all the corresponding states of the FG model that

map into those CG configurational states, $\omega_{FG} = \left\{ \mathbf{r}^n : M(\mathbf{r}^n) \in \omega_{CG} \right\}$, averaged over the eliminated configuration variables:

$$S_{CG}\left(\omega_{CG}\right) = -\int_{\mathbf{r}^n \in \omega_{FG}} d\mathbf{r}^n \delta\left(\mathbf{R}^N - M\left(\mathbf{r}^n\right)\right) p\left(\mathbf{r}^n\right) \ln p\left(\mathbf{r}^n\right) . \tag{4.10}$$

One can show that a CG expression for the configurational entropy integrand in terms of only CG variables that satisfies Eq. (4.2) for any choice of state set is

$$S_{CG,\omega_{CG}}\left(\mathbf{R}^N\right) = -k_B \ln p_{\omega_{CG}}\left(\mathbf{R}^N\right) + \frac{dU_{CG}\left(\mathbf{R}^N\right)}{dT}, \tag{4.11}$$

where $p_{\omega_{CG}}\left(\mathbf{R}^N\right)$ is the probability of configuration $\mathbf{R}^N$ among the set of configurational states ω_{CG} (see Appendix D for details). Again, this is expressed as the sum of the naïve CG expression

$$S_{CG,\omega_{CG},naïve}\left(\mathbf{R}^N\right) = -k_B \ln p_{\omega_{CG}}\left(\mathbf{R}^N\right) \tag{4.12}$$

and a temperature-dependent term. As Appendix D shows clearly, the temperature dependence in Eq. (4.11) is the complement of the temperature-dependence seen for configurational internal energy. Such ideas go back well past Stillinger's authoritative work on the subject[96] and receive focused attention in the relative entropy literature.[26, 28, 111]

Our focus here is not on the novelty of these expressions, but rather on the firmly establishing the principle that observable expressions in CG models often differ from the corresponding expressions in atomistic models. Thermodynamic observables are only one especially well-studied case, and a case where confusion about the principle is most common and best recognized.

In some cases, Eq. (4.2) will correspond to naïve CG observable expressions; however, this is not generally the case. For pressure, all configuration variables contribute to the FG observable expression, as discussed above, and so observables satisfying Eq. (4.2) must represent the contributions from all eliminated configuration variables through the additional term in Eq.(4.6). For configurational internal energy, the naïve CG configurational internal energy in Eq. (4.9) does not indicate what portion is truly internal energy; entropic effects from eliminated configuration variables are also included in the effective interaction. Consequently, Eq. (4.9) returns both entropic and energetic effects for CG models. In order to separate these effects, contributions from the eliminated configuration variables must be included using an additional term as in Eq. (4.8). For configurational entropy, the issues are similar to those for configurational internal energy. Since entropy is a measure of the distribution of states on phase space, properties of the distribution beyond the average are needed to accurately describe this observable. Consequently, all eliminated FG configuration variables make contributions to CG entropy that are captured using an additional term such as in Eq. (4.11). This is to be expected as configuration-dependent thermodynamics are projected onto the many-body potential of mean force (PMF) as a result of coarse-graining.[22, 25, 26, 28, 35, 36, 39, 40]

This has implications for the interpretation of top-down CG models as well. In top-down coarse-graining, analogy between CG particles and groups of particles in an FG model establishes sets of approximate structural observables, as depicted in Figure 4-2b. Any additional CG observables such as energies and pressures must be compatible with these structural observables if they are to reproduce both sets of experimental observables simultaneously (to say nothing of crucial structure-pressure and structure-energy cross correlation observables). In order to be compatible with the system definition, we showed that bottom-up CG observables

satisfying Eq. (4.2) will include contributions from eliminated FG configurational variables. Including these contributions leads to the introduction of *extra terms* at the CG level that did not appear in the FG observable representation. Thus, top-down CG observable expressions built by analogy to FG observables, ones that do not contain such extra terms – which is very often the case – will generally misrepresent the underlying FG or atomistic physics for the system at hand.

Top-down CG models are fundamentally inconsistent in their relationship to the real systems they are designed to model if they do not also take into account these facts about observable representability, i.e., thermodynamic state-dependence of CG potentials and the other possible complications discussed above.

4.3 Results and Discussion

In this section, simple models are used to demonstrate different aspects of CG observable representation. The analytical ideal polymer highlights how coarse-graining even a single intramolecular bond can lead to problems with naïve CG observable expressions that can then be corrected using expressions that satisfy Eq. (4.2). The discrepancies between naïve CG observables and the corresponding FG observables defy intuition and make it clear that the interpretation of naïve CG observables can be highly suspect. These problems do not go away when the number of segments is increased or even when the rigid nature of the freely jointed chain (FJC) model is relaxed. Additionally, CG observables satisfying Eq. (4.2) include contributions from segments that are eliminated when constructing an end-to-end representation of the polymer. The representationally consistent CG observables presented in Section 4.2 are compared with FG observables and naïve CG observables as appropriate for these models.

The ideal polymer chain has been a cornerstone of polymer physics because it is a simple model that displays the entropic character typical of polymers without the added complexity of non-bonded interactions between monomers[112, 113]. The FJC is an ideal polymer model where the bond vectors between segments $\vec{r}_i$ are fixed to be length b, but rotation is unimpeded. In this model, all configurations that simultaneously satisfy the bond length constraints are allowed and have zero interaction energy, while all other configurations are forbidden. As a result, the configurational internal energy is always zero and that any configurational free energy differences must be controlled by configurational entropy. A common reduced representation for this model is the magnitude of the end-to-end distance $R=\left|\sum_i \vec{r}_i\right|$, where all configuration variables other than the position of the polymer endpoints are integrated out.

For an N segment chain on a d-dimensional lattice with steps of $\pm b$ allowed in each dimension for each segment, the distribution of the end-to-end distance in the large N limit is given by

$$P_{FJC}^{d}\left(R;N\right)=\left(2\pi Nb^2\right)^{-d/2}\exp\left(\frac{-R^2}{2Nb^2}\right). \tag{4.13}$$

The properties of this model are shown in the left column of Tables 4-1 through 4-3. The naïve effective "CG" potential $U_{eff}\left(R\right)=-\beta^{-1}\ln P\left(R\right)$ for this model (actually the naïve configurational internal energy) is not independent of β, differing from the state-independent FG model. When the seemingly intuitive, naïve expression in Eq. (4.9) is used on the CG model (Table 4-1), the resulting internal energy is non-zero, which disagrees with intuition. Adding contributions from eliminated configuration variables to that internal energy by satisfying Eq. (4.2) gives Eq. (4.8), which agrees with intuition. Likewise, entropy measured using the naïve

expression in Eq. (4.12) gives zero, which violates intuition for this entropically driven model. However, the true CG entropy expression in Eq. (4.11) satisfies Eq. (4.2) and properly corresponds to the entropy of the FG model. The partitioning of configurational internal energy and entropy by the naïve observables incorrectly allocates all of the configurational free energy to internal energy instead of entropy. However, the free energies corresponding to the two sets of observables must be equal since they both describe the same CG model. As noted in Section 4.2, the fact that the entropic and energetic contributions to the configurational free energy cannot be correctly distinguished by the naïve CG observables is common. This is why an awareness of the model resolution is vital to determining the CG observable expressions that will correspond with experimental observables.

Table 4-1. Properties of freely jointed chain (FJC) models for the configurational internal energy (E) in the FG model, as well as the CG model, using both expressions for a naïve CG observable defined by direct analogy of the AA observable and the representationally consistent observable that satisfies Eq. (4.2). The CG models are end-to-end representations of the polymer chain.

Observable/ Model	*Normal FJC*	*2-segment Off-lattice*	*Blurred FJC*
$\left\langle E\left(\mathbf{r}^n\right)\right\rangle_R$	0	0	$\frac{-1}{2\beta}+\frac{k}{2Nb^2}\left\langle\left(R-l\right)^2\right\rangle$
$E_{naïve}\left(R\right)$	$\frac{-d}{2\beta}\ln\left(\frac{1}{2\pi Nb^2}\right)+\frac{R^2}{2Nb^2\beta}$	$-\frac{1}{\beta}\ln\left(\frac{3R}{4b^2}\sqrt{1-\frac{R^2}{4b^2}}\right)$	$\frac{-1}{2\beta}\ln\left(\frac{\beta k}{2\pi Nb^2}\right)$ $-\frac{1}{\beta}\ln\left(\sum_{l=-N}^{N}\exp\left(-\frac{l^2}{2Nb^2}-\frac{\beta k}{2Nb^2}\left(R-l\right)^2\right)\right)$
$E_{true}\left(R\right)$	0	0	$\frac{-1}{2\beta}+\frac{k}{2Nb^2}\left\langle\left(R-l\right)^2\right\rangle$

Table 4-2. Properties of freely jointed chain (FJC) models for the entropy (S) measured in the FG model, as well as the CG model, using both expressions for a naïve CG observable defined by direct analogy of the AA observable and the representationally consistent observable that satisfies Eq. (4.2). The CG models are end-to-end representations of the polymer chain.

Observable/ Model	*Normal FJC*	*2 segment Off-lattice*	*Blurred FJC*
$\left\langle S\left(\mathbf{r}^n\right)\right\rangle_R$	$\frac{k_B d}{2}\ln\left(\frac{1}{2\pi Nb^2}\right)-\frac{k_B R^2}{2Nb^2}$	$k_B\ln\left(\frac{3R}{4b^2}\sqrt{1-\frac{R^2}{4b^2}}\right)$	$\frac{k_B}{2}\ln\left(\frac{\beta k}{2\pi Nb^2}\right)-\frac{k_B}{2}+\frac{kk_B\beta}{2Nb^2}\left\langle\left(R-l\right)^2\right\rangle$ $-\frac{1}{\beta}\ln\left(\sum_{l=-N}^{N}\exp\left(-\frac{l^2}{2Nb^2}-\frac{\beta k}{2Nb^2}\left(R-l\right)^2\right)\right)$
$S_{naïve}\left(R\right)$	0	0	0
$S_{true}\left(R\right)$	$\frac{k_B d}{2}\ln\left(\frac{1}{2\pi Nb^2}\right)-\frac{k_B R^2}{2Nb^2}$	$k_B\ln\left(\frac{3R}{4b^2}\sqrt{1-\frac{R^2}{4b^2}}\right)$	$\frac{k_B}{2}\ln\left(\frac{\beta k}{2\pi Nb^2}\right)-\frac{k_B}{2}+\frac{kk_B\beta}{2Nb^2}\left\langle\left(R-l\right)^2\right\rangle$ $-\frac{1}{\beta}\ln\left(\sum_{l=-N}^{N}\exp\left(-\frac{l^2}{2Nb^2}-\frac{\beta k}{2Nb^2}\left(R-l\right)^2\right)\right)$

The failure of naïve CG observables to reproduce FG observables and the state-dependence acquired in the effective CG interaction is a fundamental feature of coarse-graining. To show that this is not just unique to the lattice model examined above, a 2-segment off-lattice (OL) FJC model will be investigated in 3-D. The distribution of the end-to-end distance is

$$P_{FJC}^{OL}\left(R;2\right)=\frac{3R}{4b^2}\sqrt{1-\frac{R^2}{4b^2}}\ . \tag{4.14}$$

The naïve effective CG potential has again acquired β-dependence that the FG model did not have. The properties of this model are shown in the middle column of Tables 4-1 through 4-3. This model also has zero configurational internal energy. The naïve configurational internal

energy expression does not correctly reproduce this FG value. However, the expression satisfying Eq. (4.2) correctly reproduces the FG observable. The configurational entropy shows the same behavior as above. These patterns hold even for more complicated models with non-uniform FG energies and continuous end-to-end distance, which will be demonstrated below using an elastic FJC. In this model, there is a harmonic potential $u\left(r_i\right)=k\left(r_i-b\right)^2/2$ on the length of each segment i centered at distance b with spring constant k instead of the delta function in the FJC with fixed-length bonds. The distribution of the end-to-end distance of a 1-dimensional blurred FJC is given by the proportionality

$$P_{Blur}^{1}\left(R;N\right)\propto\sqrt{\frac{\beta k}{2\pi Nb^2}}\sum_{l=-N}^{N}\exp\left(-\frac{l^2}{2Nb^2}-\frac{\beta k}{2Nb^2}\left(R-l\right)^2\right), \tag{4.15}$$

where l is a summing index needed to consider the contribution to the probability from each Gaussian centered at an end-to-end distance of l. The second term in the exponential reflects the Gaussian chain aspect of this model, which allows all Gaussians centered at the FJC distance l to have non-zero contributions to the probability of any given end-to-end distance. This model reduces back to the FJC model described in Eq. (13) in the limit that k goes to infinity. The naïve effective CG interaction for this model still has β-dependence that the FG model did not have, which comes from the FJC-like left term in the exponential. Thus, the thermodynamics are still not properly reproduced when the naïve observable expression is used for internal energy and entropy. However, the CG observables satisfying Eq. (4.2) have no problem reproducing the FG observable projected onto this CG representation. So, one should expect this behavior from any model that has some FJC-like character. The harmonic potential centered at distance b is

directly analogous to the harmonic potentials used for bonded interactions, which suggests that this problem arises whenever bonded configuration variables are coarse-grained.

Until this point, the results have focused on the thermodynamic observables presented in Eqs. (4.3) – (4.12). The pattern of creating CG observables that correspond to the FG model's observables is not just one of finding corrections to thermodynamic observables. Instead, it is illustrative of a more general aspect of this approach that can be used for any observable. Table 4-3 shows the magnitude of the average bond orientation $O = \left| N^{-1} \sum_i \hat{r}_i \right|$ for the different polymer models, where $\hat{r}_i$ is the unit vector for bond segment i. Given only the end-to-end distribution of the CG model, it would seem that the bond orientation is undefined in the CG model since configuration variables used in the FG observable were eliminated to reach the CG level of resolution. Using a naïve orientation measure on the one "CG segment" gives a value of unity since the "CG segment" is always aligned with itself. For the normal FJC and the 2-segment off-lattice FJC, there is only one value for the magnitude of the average orientation for a given end-to-end distance, which is represented correctly using an expression satisfying Eq. (4.2) . In the case of the blurred FJC, there is a distribution of magnitudes for the average orientation for a given end-to-end distance. The projection onto the CG level of resolution, however, averages this distribution for each end-to-end distance. In both cases, statements can be made about the magnitude of the average orientation. One could still improve the performance of the naïve CG observable by changing the number of sites to be the implied number of FG particles instead of using the explicitly represented CG sites. This "resolution-aware" version of the naïve CG observable expression reproduces the correct value of the observable by including information about the number of segments and the segment length. However, this observable is a special case like the COM RDF example discussed above. This works here because the

observable itself involves an aggregation step that directly follows the elimination of configuration variables in mapping from the FG model to the CG model. Although the corrected naïve observable expression satisfied Eq. (4.2) here, there are cases where a multiplicative correction to the naïve observable expression is insufficient to satisfy Eq. (4.2).

The polymer radius of gyration, R_g^2, is an example of how the correct CG observable expression differs from the naïve CG observable expression in more complicated ways. For the FG model, the radius of gyration is given by $R_g^2 = N^{-1}\sum_{i=1}^{N}\left(R_i - R_{cm}\right)^2$, where R_i is the position of bead i and $R_{cm} = N^{-1}\sum_{i=1}^{N} R_i$ is the COM. Using the naïve CG expression for the radius of gyration on the end-to-end CG representation of the 2-semgent off-lattice FJC gives

$$R_{g,CG}^{2,AA} = \frac{R^2}{4} \tag{4.16}$$

since one only knows about the effective end-to-end segment of the CG polymer. However, one can analytically evaluate the expectation of the radius of gyration conditional on the end-to-end distance. This is the projection of the radius of gyration from the FG model onto the CG model and satisfies Eq. (4.2). In this case,

$$R_{g,CG}^{2,CG} = \left\langle R_{g,AA}^{2,AA} \right\rangle_R = \frac{1}{9}\left(2b^2 + R^2\right) \tag{4.17}$$

As described by Eq. (4.17), the radius of gyration behaves differently than Eq. (4.16)would suggest. While both descriptions for R_g^2 scale quadratically with end-to-end distance, Eq. (4.16) has the wrong prefactor. Using the same sort of multiplicative scaling as above would change the denominator from 4 to 6, but an updated version of Eq. (4.16) would still not have the same

prefactor as Eq. (4.17) More notably, Eq. (4.17) shows that there is a non-zero R_g^2 at an end-to-end distance of zero. This agrees with physical intuition for any FJC, but it is not captured by Eq. (4.16). Using simple multiplicative corrections to the naïve CG observable does not fix this violation of physical intuition. Here, one can only recover the correct behavior for R_g^2 by satisfying Eq. (4.2).

Table 4-3. Properties of freely jointed chain (FJC) models for the magnitude of the average orientation measured in the FG model, as well as the CG model, using both a naïve CG observable defined by direct analogy of the FG observable and a resolution-aware (RES) observable satisfying Eq. (4.2). The CG models are end-to-end representations of the polymer chain.

Observable/ Model	*Normal FJC*	*2-segment Off-lattice*	*Blurred FJC*
$\left\langle O\left(\mathbf{r}^n\right)\right\rangle_R$	$\frac{R}{Nb}$	$\frac{R}{2b}$	$\frac{R}{Nb}$
$O_{naïve}\left(R\right)$	1	1	1
$O_{RES}\left(R\right)$	$\frac{R}{Nb}$	$\frac{R}{2b}$	$\frac{R}{Nb}$

4.4 Discussion

The example systems in Section 4.3 demonstrate problems that can arise when interpreting CG models using naïve CG observable expression. In these systems, naïve CG observables incorrectly attribute the entropic part of the configurational free energy to configurational internal energy. For the FJC, this misattribution completely changes the interpretation of the model from being governed by entropy, as intuition suggests, to being governed by internal energy. The fact that this problem does not go away for the blurred FJC suggests that it is a

general phenomenon affecting any CG model where intramolecular configuration variables are eliminated. In these systems, the expressions satisfying Eq. (4.2) were derivable because the CG interaction potential could be written analytically; however, one does not have this luxury when studying complex systems. Instead, one could numerically determine this additional term using a method such as the single-point CG sensitivity formula.[41] Alternatively, one could fit the entire expression numerically since CG observables that satisfy Eq. (4.2) are guaranteed to reproduce the projection of corresponding FG observable.

More generally, these results emphasize the importance of resolution awareness in model interpretation. For the magnitude of the average orientation, neglecting the resolution of the CG model by using the naïve observable expression for the model leads to a nonsensical answer, given the FG model that it is supposed to represent. The end-to-end representation could be constructed for a polymer of any length and segment-type, but the correct interpretation of the CG model depends on what FG model it represents. While this observable was simple enough that only minimal consideration of the FG model was needed to satisfy Eq. (4.2) and reproduce the FG value, this will not work in general. However, CG observables that satisfy Eq. (4.2) will always work regardless of the observable complexity. For example, the added complexity of the radius of gyration means that simple corrections to the naïve CG observable fail to satisfy Eq. (4.2); however, the resolution-aware CG observable that satisfies Eq. (4.2) correctly describes the scaling and asymptotes. Satisfying Eq. (4.2) also ensures that CG observables faithfully reproduce the complete projection of FG observables.

This observable projection approach avoids the observable incompatibility seen when using naïve CG observables. Of note, pressure has been a particularly problematic observable in the literature.[56, 98, 114] Depending on what observables are used to parameterize the CG model, the

compatibility of CG observable expressions with different model definitions may change. As a result, we cannot for instance resolve the dispute over which pressure expression is correct for a Debye-Huckel model effective potential since many different systems can map to each such CG representation.[108, 115-118] Instead, we assert that the change in resolution from the FG model to the CG model will determine CG observable expression compatible with the model definition. Generally, naïve CG observables not part of the model definition will be incompatible with the model definition. However, CG observables satisfying Eq. (4.2) will always be compatible with the model definition by construction. Additionally, these CG observables will adjust to be compatible with different system definitions.

The observables used to parameterize top-down CG models also need to be compatible with each other. Otherwise, the CG model will not reproduce these observables; consequently, this model will not correspond to the intended FG system. In fact, this CG model may not correspond to any physically realizable system. Without a way to determine if top-down CG observables are compatible, one may not know that their CG model is unphysical.

A natural way to avoid this complexity in top-down coarse-graining is to look for maximally transferable models, since the least state-dependent potentials imply the smallest deviation from atomistic observable forms. This is, no doubt, part of what makes transferability such an important consideration when building top-down models. Though naïvely it might seem that non-transferable models are powerful at least at the state point they describe, the principle embodied in Eq. (4.2) implies that non-transferable models cannot be interpreted according to the usual atomistic-model-based observable expressions. This no doubt accounts for some of the strong prejudices in the field against non-transferable models.

An inability to determine if top-down CG observables are compatible using Eq. (4.2) means that other approaches are needed. One could apply bottom-up CG observable expressions to top-down CG models. In this case, one needs to be aware of the different model resolutions even if the distribution of CG configurations is the same in both top-down and bottom-up CG models. Here, the bottom-up CG observable could be the basis for parameterizing a top-down CG model. Alternatively, one could try to correct naïve CG observables to be compatible with top-down CG model definitions. However, the radius of gyration example showed that there are limits to what can be done using only simple corrections to naïve CG observables.

Any comparison of top-down and bottom-up CG model properties is nonsensical unless care is taken in choosing observables.[87] Indeed, a top-down model fit using incompatible observables may reproduce a given observable using a naïve observable expression. However, that does not mean that it corresponds to the intended experimental system.[105] This problem is directly encountered when top-down CG model definitions are adjusted to reproduce additional observables such as pressure and interfacial tension without considering the compatibility of all model observables.[119]

Improvements in CG pressure can benefit constant NPT CG simulations. The integrated approach developed by Das and Andersen[31] and extended by Dunn and Noid's[97] iterative refinement procedure correctly describes volume fluctuations as the observable contributions from eliminated configuration variables are projected onto the system volume. Correct pressure fluctuations would also be captured if these contributions were projected onto the CG configuration variables instead of the system volume. However, most constant NPT CG simulations currently use a barostat based on the naïve CG virial instead of using an approach like the one developed by Das and Andersen.[97] From our earlier discussion, it is clear that the

CG virial is not compatible with any bottom-up model that is coarser than the atomistic model. Currently, the behavior of constant NPT CG models is different from the behavior of supposedly underlying FG models. Constructing barostats using an expression for CG pressure that satisfies Eq. (4.2) might help align the behavior of CG models under constant NPT to that of the FG model.

Also, mixed resolution modeling can benefit from improved CG observables.[120, 121] Since observable expressions change with model resolution, the expression for the pressure of the FG part of the system should be different from the expression for the pressure of the CG part of the system. Failure to recognize this causes unphysical density profiles at the interface between resolutions in adaptive resolution simulations that requires the introduction of thermodynamic forces to counteract this apparently pressure-induced drift.[122] This problem is even worse for adaptive resolution schemes with top-down CG models for the reasons already discussed.[123, 124] Using resolution appropriate observable expressions is an important first step towards improving these models.

4.5 Conclusion

In this chapter, we have discussed the importance of CG observables being compatible with the resolution of the CG model. Using Eq. (4.2) an indefinite number of compatible CG observables can be identified. If defined correctly, these observables may also be consistent with top-down CG model definitions. However, naïve expressions for CG observables taken directly from the FG expressions often fail to satisfy Eq. (4.2) because they neglect contributions from eliminated CG configuration variables. These neglected contributions are sometimes captured by terms that depend on a system's thermodynamic state dependence, which can be used to correct

naïve CG observable expressions; this was demonstrated analytically for analytical polymer systems. One must treat CG models differently than FG models by considering the resolution of the CG model in its interpretation. These concepts cannot be ignored in either bottom-up or top-down CG modeling if multiscale modeling is to generate physically meaningful predictions for physical systems that have their origins at a FG level.

Exploring the aspects of CG observable representation discussed here can impact research efforts in several areas. Using observable expressions that satisfy Eq. (4.2) can give new hope to work on basis set representability.[42, 43] Additionally, CG simulations in the constant NPT ensemble can more closely follow the evolution of the FG system using expressions for CG pressure that correspond to the CG model's resolution by using approaches such as the method developed by Das and Andersen or Dunn and Noid's extension. Further improvements to better describe pressure fluctuations are also possible. Likewise, mixed-resolution models will need to use a variety of properly formulated observable expressions to describe the properties of the different resolutions present in the simulation and their interfaces so that large thermodynamic artifacts are not generated. Perhaps most importantly, top-down CG models must be parameterized using compatible observable expressions that are consistent with the model's underlying FG resolution. These top-down models can benefit from bottom-up analysis to determine these observable expressions. In the end, the interpretation of CG models must ultimately depend on understanding how they relate to the actual, physical FG experimental systems they intend to describe.

4.6 Appendix A: Derivation of Equation (4.2)

In order to make the comparison of the FG and CG ensemble averages of an observable easier, one can insert a delta function involving a CG mapping operator into the FG observable ensemble average expression along with an integral over all CG configurations to get

$$\left\langle A_{FG}\right\rangle = \frac{\int d\mathbf{R}^N \int d\mathbf{r}^n \delta\left(M\left(\mathbf{r}^n\right)-\mathbf{R}^N\right) A_{FG}\left(\mathbf{r}^n\right) e^{-\beta U\left(\mathbf{r}^n\right)}}{\int d\mathbf{R}^N \int d\mathbf{r}^n \delta\left(M\left(\mathbf{r}^n\right)-\mathbf{R}^N\right) e^{-\beta U\left(\mathbf{r}^n\right)}} \tag{4.18}$$

which is still an average of the FG observable over all FG configuration space. To show the equivalence of the CG observable ensemble average and Eq. (4.18)for the definition of the CG observable given in Eq. (4.2), one substitutes Eq. (4.2) into the CG observable ensemble average, cancels the CG Boltzmann factors, and notes the equivalence of the partition functions by substituting Eq. (4.1) into the denominator of the CG observable ensemble average. This also demonstrates the equivalence of FG and CG observable ensemble averages in this case.

4.7 Appendix B: Derivation of Equation (4.6)

Based on Eq. (4.2), the expression for pressure is

$$P_{CG}\left(\mathbf{R}^N\right) = \left\langle \frac{\rho_{FG}}{\beta} + \frac{\mathbf{r}^n}{3V} \cdot \left.\frac{dU_{FG}(\mathbf{r}^n)}{d\mathbf{r}^n}\right|_V \right\rangle_{\mathbf{R}^N} . \tag{4.19}$$

One can find the difference between this expression and the CG virial expression in Eq. (3) by rewriting this as an explicit integral

$$P\left(\mathbf{R}^N\right) = \frac{\rho_{FG}}{\beta} + \frac{\int d\mathbf{r}^n \delta\left(M\left(\mathbf{r}^n\right)-\mathbf{R}^N\right) e^{-\beta U_{FG}\left(\mathbf{r}^n\right)} \dfrac{\mathbf{r}^n}{3V} \cdot \left.\dfrac{-dU_{FG}\left(\mathbf{r}^n\right)}{d\mathbf{r}^n}\right|_V}{\int d\mathbf{r}^n \delta\left(M\left(\mathbf{r}^n\right)-\mathbf{R}^N\right) e^{-\beta U_{FG}\left(\mathbf{r}^n\right)}} . \tag{4.20}$$

Introducing the scaled fine-grained configuration $\mathbf{x}^n \equiv V^{-1/3}\mathbf{r}^n$, this is

$$P\left(\mathbf{R}^N\right) = \frac{\rho_{FG}}{\beta} + \frac{\int d\mathbf{r}^n \delta\left(M\left(\mathbf{r}^n\right) - \mathbf{R}^N\right) e^{-\beta U_{FG}\left(\mathbf{r}^n\right)} \left.\frac{-dU_{FG}\left(\mathbf{r}^n\right)}{dV}\right|_{\mathbf{x}^n}}{\int d\mathbf{r}^n \delta\left(M\left(\mathbf{r}^n\right) - \mathbf{R}^N\right) e^{-\beta U_{FG}\left(\mathbf{r}^n\right)}} . \tag{4.21}$$

Rearranging the last parts of the integrand, one gets

$$P\left(\mathbf{R}^N\right) = \frac{\rho_{FG}}{\beta} + \frac{\beta^{-1}\int d\mathbf{r}^n \delta\left(M\left(\mathbf{r}^n\right) - \mathbf{R}^N\right) \left.\frac{de^{-\beta U_{FG}\left(\mathbf{r}^n\right)}}{dV}\right|_{\mathbf{x}^n}}{\int d\mathbf{r}^n \delta\left(M\left(\mathbf{r}^n\right) - \mathbf{R}^N\right) e^{-\beta U_{FG}\left(\mathbf{r}^n\right)}} , \tag{4.22}$$

where now one would like to take the derivative with respect to volume outside of the integral. To do so, notice that under a change of coordinates, the $d\mathbf{r}^n \delta\left(M\left(\mathbf{r}^n\right) - \mathbf{R}^N\right)$ portions of the integrands above and below will change identically. Under scaling the coordinates by volume, the change is simply a rescaling by a power of the volume, independent of configuration, which therefore cancels out above and below. Therefore, one has

$$P\left(\mathbf{R}^N\right) = \frac{\rho_{FG}}{\beta} + \frac{\beta^{-1}\int d\mathbf{x}^n \delta\left(M\left(\mathbf{x}^n\right) - \mathbf{X}^N\right) \left.\frac{de^{-\beta U_{FG}\left(V^{1/3}\mathbf{x}^n\right)}}{dV}\right|_{\mathbf{x}^n}}{\int d\mathbf{x}^n \delta\left(M\left(\mathbf{x}^n\right) - \mathbf{X}^N\right) e^{-\beta U_{FG}\left(V^{1/3}\mathbf{x}^n\right)}} . \tag{4.23}$$

Now the rest of the numerator integral, aside from the part in the derivative, is independent of volume. The derivative could be taken out of the integral at this point, except that it makes no sense to have a derivative constant with respect to $\mathbf{x}^n$ outside the integral over $\mathbf{x}^n$. Instead, enforcing the delta function constraint turns the volume derivative constant with respect to $\mathbf{x}^n$ inside into a volume derivative constant with respect to $\mathbf{X}^N$ outside the integral.

$$P\left(\mathbf{R}^N\right)=\frac{\rho_{FG}}{\beta}+\frac{\beta^{-1}\left.\frac{d}{dV}\right|_{\mathbf{X}^N}\int d\mathbf{x}^n\delta\left(M\left(\mathbf{x}^n\right)-\mathbf{X}^N\right)e^{-\beta U_{FG}\left(V^{1/3}\mathbf{x}^n\right)}}{\int d\mathbf{x}^n\delta\left(M\left(\mathbf{x}^n\right)-\mathbf{X}^N\right)e^{-\beta U_{FG}\left(V^{1/3}\mathbf{x}^n\right)}} \tag{4.24}$$

And, recognizing the derivative of a logarithm,

$$P\left(\mathbf{R}^N\right)=\frac{\rho_{FG}}{\beta}+\beta^{-1}\left.\frac{d}{dV}\right|_{\mathbf{X}^N}\log\int d\mathbf{x}^n\delta\left(M\left(\mathbf{x}^n\right)-\mathbf{X}^N\right)e^{-\beta U_{FG}\left(V^{1/3}\mathbf{x}^n\right)} \ . \tag{4.25}$$

And finally, converting the integral back to unscaled coordinates, one requires only a few more manipulations. First, group the volume terms:

$$\begin{aligned}P\left(\mathbf{R}^N\right)&=\frac{\rho_{FG}}{\beta}+\beta^{-1}\left.\frac{d}{dV}\right|_{\mathbf{X}^N}\ln\int V^{-n}d\mathbf{r}^nV^N\delta\left(M\left(\mathbf{r}^n\right)-\mathbf{R}^N\right)e^{-\beta U_{FG}\left(\mathbf{r}^n\right)}\\&=\frac{\rho_{FG}}{\beta}+\beta^{-1}\left.\frac{d}{dV}\right|_{\mathbf{X}^N}\left(\ln V^{-(n-N)}+\ln\int d\mathbf{r}^n\delta\left(M\left(\mathbf{r}^n\right)-\mathbf{R}^N\right)e^{-\beta U_{FG}\left(\mathbf{r}^n\right)}\right)\end{aligned} \ . \tag{4.26}$$

Then take the volume derivative of the first term and recognize the second as the (volume-dependent) CG force field,

$$P\left(\mathbf{R}^N\right)=\frac{\rho_{FG}}{\beta}-\frac{\left(n-N\right)}{\beta V}-\left.\frac{dU_{CG}\left(\mathbf{R}^N;V\right)}{dV}\right|_{\mathbf{X}^N}=\frac{\rho_{CG}}{\beta}-\left.\frac{dU_{CG}\left(\mathbf{R}^N;V\right)}{dV}\right|_{\mathbf{X}^N} \ . \tag{4.27}$$

Thus, Eq. (4.27) shows that Eq. (4.6) satisfies Eq. (4.2) for CG pressure.

4.8 Appendix C: Derivation of Equation (4.8)

The CG configurational internal energy based on Eq. (4.2) can be expressed as the naïve expression plus a correction term as

$$E_{CG}\left(\mathbf{R}^N\right)=\left\langle E_{FG}\left(\mathbf{r}^n\right)\right\rangle_{\mathbf{R}^N}=E_{CG,naïve}\left(\mathbf{R}^N;\beta\right)+\left\langle E_{FG}\left(\mathbf{r}^n\right)-E_{CG,naïve}\left(\mathbf{R}^N;\beta\right)\right\rangle_{\mathbf{R}^N} \ , \tag{4.28}$$

where the FG configurational internal energy is the FG interaction energy $E_{FG}\left(\mathbf{r}^n\right)=U_{FG}\left(\mathbf{r}^n\right)$ and the naïve CG configurational internal energy is shown in Eq. (4.9). Using the definition of the expectation

$$E_{CG}\left(\mathbf{R}^N\right)=\left\langle U_{FG}\left(\mathbf{r}^n\right)\right\rangle_{\mathbf{R}^N}=\frac{\int d\mathbf{r}^n\delta\left(M\left(\mathbf{r}^n\right)-\mathbf{R}^N\right)U_{FG}\left(\mathbf{r}^n\right)e^{-\beta U_{FG}\left(\mathbf{r}^n\right)}}{\int d\mathbf{r}^n\delta\left(M\left(\mathbf{r}^n\right)-\mathbf{R}^N\right)e^{-\beta U_{FG}\left(\mathbf{r}^n\right)}}. \tag{4.29}$$

One can write the last portion as a derivative with respect to the inverse temperature;

$$E_{CG}\left(\mathbf{R}^N\right)=\frac{\int d\mathbf{r}^n\delta\left(M\left(\mathbf{r}^n\right)-\mathbf{R}^N\right)\frac{-de^{-\beta U_{FG}\left(\mathbf{r}^n\right)}}{d\beta}}{\int d\mathbf{r}^n\delta\left(M\left(\mathbf{r}^n\right)-\mathbf{R}^N\right)e^{-\beta U_{FG}\left(\mathbf{r}^n\right)}}. \tag{4.30}$$

The derivative can be taken out of the integral easily, and then one clear has a derivative of a logarithm, giving

$$E_{CG}\left(\mathbf{R}^N\right)=-\frac{d}{d\beta}\ln\int d\mathbf{r}^n\delta\left(M\left(\mathbf{r}^n\right)-\mathbf{R}^N\right)e^{-\beta U_{FG}\left(\mathbf{r}^n\right)} \tag{4.31}$$

The logarithm is clearly related to the definition of the CG potential, giving

$$E_{CG}\left(\mathbf{R}^N\right)=\frac{d\beta U_{CG}\left(\mathbf{R}^N;\beta\right)}{d\beta}, \tag{4.32}$$

which matches Eq. (4.8).

4.9 Appendix D: Derivation of Equation (4.11)

An expression for the CG entropy can be found from the definition of the CG entropy,

$$S_{CG}\left(\omega_{CG}\right)=-k_B\int_{\mathbf{r}^n\in\omega_{FG}}d\mathbf{r}^n p_{\omega_{FG}}\left(\mathbf{r}^n\right)\ln p_{\omega_{FG}}\left(\mathbf{r}^n\right) \tag{4.33}$$

First, we split the integral into a fine-grained and coarse-grained part and then factor out the CG probabilities to find

$$S_{CG}\left(\omega_{CG}\right)=-k_B\int_{\mathbf{R}^N\in\omega_{CG}}d\mathbf{R}^N\int d\mathbf{r}^n\delta\left(\mathbf{R}^N-M\left(\mathbf{r}^n\right)\right)p_{\omega_{FG}}\left(\mathbf{r}^n\right)\ln p_{\omega_{FG}}\left(\mathbf{r}^n\right) \tag{4.34}$$

$$S_{CG}\left(\omega_{CG}\right)=-k_B\int_{\mathrm{R}^N\in\omega_{CG}}d\mathbf{R}^N p_{\omega_{CG}}\left(\mathbf{R}^N\right)\times$$

$$\int d\mathbf{r}^n\delta\left(\mathbf{R}^N-M\left(\mathbf{r}^n\right)\right)\frac{p_{\omega_{FG}}\left(\mathbf{r}^n\right)}{p_{\omega_{CG}}\left(\mathbf{R}^N\right)}\left(\ln\frac{p_{\omega_{FG}}\left(\mathbf{r}^n\right)}{p_{\omega_{CG}}\left(\mathbf{R}^N\right)}+\ln p_{\omega_{CG}}\left(\mathbf{R}^N\right)\right), \tag{4.35}$$

which simplifies to

$$S_{CG}\left(\omega_{CG}\right)=-k_B\int_{\mathrm{R}^N\in\omega_{CG}}d\mathbf{R}^N p_{\omega_{CG}}\left(\mathbf{R}^N\right)\times$$

$$\left(\ln p_{\omega_{CG}}\left(\mathbf{R}^N\right)+\int d\mathbf{r}^n\delta\left(\mathbf{R}^N-M\left(\mathbf{r}^n\right)\right)\frac{p_{\omega_{FG}}\left(\mathbf{r}^n\right)}{p_{\omega_{CG}}\left(\mathbf{R}^N\right)}\ln\frac{p_{\omega_{FG}}\left(\mathbf{r}^n\right)}{p_{\omega_{CG}}\left(\mathbf{R}^N\right)}\right) \tag{4.36}$$

because the ratio of probabilities is a normalized conditional probability. In other words, this is the entropy of the CG model's configurational distribution plus another term accounting for the entropy of the FG particles eliminated in the coarse-graining. This term can be rearranged like a usual macroscopic entropy, using the conditional probabilities for concision:

$$S_{CG,excess}\left(\mathbf{R}^N\right)=-k_B\int d\mathbf{r}^n\delta\left(\mathbf{R}^N-M\left(\mathbf{r}^n\right)\right)p_{\omega_{FG}}\left(\mathbf{r}^n\,|\,\mathbf{R}^N\right)\ln p_{\omega_{FG}}\left(\mathbf{r}^n\,|\,\mathbf{R}^N\right)=\left\langle k_B\ln p_{\omega_{FG}}\left(\mathbf{r}^n\,|\,\mathbf{R}^N\right)\right\rangle_{\mathbf{R}^N} \tag{4.37}$$

As usual for free and internal energy, $\left\langle k_B \ln p_{\omega_{FG}}\left(\mathbf{r}^n \mid \mathbf{R}^N\right)\right\rangle_{\mathbf{R}^N} = \frac{1}{T}\left(U_{CG}\left(\mathbf{R}^N;T\right) - \left\langle U_{FG}\left(\mathbf{r}^n\right)\right\rangle_{\mathbf{R}^N}\right)$, with the temperature dependence explicit for the sake of clarity. Thus

$$S_{CG,excess}\left(\mathbf{R}^N\right) = \frac{1}{T}\left(U_{CG}\left(\mathbf{R}^N;T\right) - \frac{d\beta U_{CG}\left(\mathbf{R}^N;T\right)}{d\beta} \right) = \frac{dU_{CG}\left(\mathbf{R}^N;T\right)}{dT} \tag{4.38}$$

and plugging Eq. (4.38) into Eq. (4.36) then yields Eq. (4.11).

Chapter 5

Multiscale Compatible Observable Decomposition (MS-CODE) for Coarse-grained Observable Representation

5.1 Introduction

Atomistic simulation has been successful at both describing and predicting material properties. Coarse-grained (CG) models allow one to simulate materials at reduced computational cost.[14-18] This reduced cost makes it possible to study larger systems, where emergent phenomena could be observed, than one could with a more fine-grained (FG) simulation. However, CG models are useful for the prediction of properties if one can be sure that the measured properties of the CG model actually correspond to atomistic and experimental properties.

When observables are naïvely applied to coarse-grained models, the resulting interpretation can be misleading. For example, the comparison of top-down CG models against bottom-up CG models can imply that the top-down CG model better agrees with experiment.[125] However, CG models can be parameterized to fit at most two properties (e.g., radial distribution functions (RDFs), pressure, surface tension, entropy), unless the observable expressions are compatible.[91, 101] In the case of a bottom-up CG model parameterized to reproduce a property using the atomistic observable expression,[119, 126-129] it may get that property right, but the correct expression for its other properties such as its RDFs may not have the same form as the atomistic expression. Likewise, bottom-up CG models designed to reproduce RDFs or forces will have a straightforward interpretation for its structural properties, but the correct expressions to interpret its other properties will generally not be the usual atomistic expression.[126] To this end, some

researchers have noted that observable expressions should be different based on model resolution and ensemble.[91, 96, 130]

Mutual observable compatibility is an important characteristic for CG observables.[91] It allows CG observables to be measured simultaneously. At the very least, the CG observable expression should ensure that its ensemble average in the CG model agree with the corresponding FG value. Yet, the use of the naïve CG pressure expression is quite common even though it has been shown to dramatically overestimate pressure antecdotely.[15, 131] Das and Anderson[31] and Dunn and Noid[97] have included a system-wide, state-dependent term that modifies CG pressure so that it yields the correct average pressure for a given volume. However, configuration-dependent changes in pressure are still measured through the naïve virial expression. In principle, other observables developed this way would be compatible at the level of the ensemble average.[130]

In the previous chapter,[91] we discussed a condition for obtaining configuration-level compatibility for sets of observables. It requires that each CG observable expression A_{CG} reproduce the FG observable value A_{FG} averaged over all FG configurations $\mathbf{r}^n$ that map to a given CG configuration $\mathbf{R}^N$:

$$A_{CG}\left(\mathbf{R}^N\right)=\left\langle A_{FG}\left(\mathbf{r}^n\right)\right\rangle_{\mathbf{R}^N}=\frac{\int d\mathbf{r}^n A_{FG}\left(\mathbf{r}^n\right)\delta\left(M\left(\mathbf{r}^n\right)-\mathbf{R}^N\right)e^{-\beta U\left(\mathbf{r}^n\right)}}{\int d\mathbf{r}^n\delta\left(M\left(\mathbf{r}^n\right)-\mathbf{R}^N\right)e^{-\beta U\left(\mathbf{r}^n\right)}}, \tag{5.1}$$

where U is the potential, $\beta=\left(k_B T\right)^{-1}$, k_B is Boltzmann's constant, T is temperature, and M is the mapping operator that relates FG configurations to CG configurations.

In this chapter, we investigated new ways of numerically determining CG observable expressions that reproduce the corresponding FG observable distributions. Our method, called the multiscale compatible observable decomposition (MS-CODE), is used to determine CG observable expressions for pressure and potential. Our implementation is variational, which means that adding additional terms is guaranteed not to make the resulting observable expression worse. Also, it allows for the CG observable expression to take any form, not just the naïve expression.

The rest of the chapter is structured as follows: In Section 2, we derive the minimization targets, formulate basis sets for the observables, and show how to decompose observables; In Section 3, we present the results comparing FG observable distributions with naïve and MS-CODE CG observables for the pressure of various MeOH CG models as well as the potential of 1-site CG methanol and acetonitrile systems; In Section 4, we discuss the implications of these; and In Section 5, we provide conclusions.

5.2 Theory and Methods

For the CG observable expression of a general, numerically simulated CG model, we choose to define the CG observable expression O_{CG} as the sum of linearly dependent basis functions ϕ_i and coefficients λ_i:

$$O_{CG}\left(\mathbf{R}^N\right)=\sum_i \lambda_i \phi_i\left(\mathbf{R}^N\right). \tag{5.2}$$

For the purpose of generality, these basis functions can be functions of single partlices,[132] pairs, triples, density,[133] order parameters,[134] etc. A similar approach is used in some field theory

models, where an observable expression is not directly tied to the interaction potential.[110] In the rest of this section, we will consider two different minimization targets and strategies for the parameterization of a generic expression such as Eq. (5.2).

5.2.1 Variational Minimization

First, we aim to satisfy Eq. (5.1). This requirement for CG observable expressions is directly analogous to that for the CG force field in multiscale coarse-graining (MS-CG).[22, 29-40] Continuing the analogy with MS-CG forces, the CG observable target in Eq. (5.1) can be used to parameterize an observable's basis set variationally. This parameterization requires the definition of an appropriate objective function to minimize our observable expression O_{CG}:

$$\chi^2\left[O_{CG}\right]=\frac{1}{dN}\left\langle\left\|\sum_{I=1}^{N}A_{CG,I}\left(M\left(\mathbf{r}^n\right)\right)-O_{CG,I}\left(M\left(\mathbf{r}^n\right)\right)\right\|^2\right\rangle \tag{5.3}$$

where d is the dimensionality of the observable, I is the CG particle index, and N is the number of observable values per frame. This is equivalent to minimizing the residual

$$\chi^2\left[O_{CG}\right]=\frac{1}{dN}\left\langle\left\|\sum_{I=1}^{N}M^{\dagger}\left(A_{FG,i}\left(\mathbf{r}^n\right)\right)-O_{CG,I}\left(M\left(\mathbf{r}^n\right)\right)\right\|^2\right\rangle \tag{5.4}$$

where $M^{\dagger}$ is the mapping operator of observables, which acts the same for observables as it does for forces. This is because the CG observable can at best reproduce the average of the FG observable values from configurations consistent with a given CG configuration. All other aspects of that distribution cannot be expressed as a configuration dependent term because they are averaged out (i.e., a constant at the resolution of the CG model). These residuals can be used for framewise observables (i.e., N=1). However, it is much more data-efficient to use the residual

for particlewise observables given the FG observable decomposition needed to express A_{CG} in a particlewise manner.

5.2.2 Relative Entropy Formulations

Alternatively, one could determine the CG observable expression by minimizing the relative entropy, which is one "measure" of how close a model distribution is to a target distribution.[25, 26, 135] In terms of the CG observable distribution p_{CG} relative to the FG observable distribution p_{FG}, the appropriate relative entropy is

$$S_{rel} = \int d\sigma p_{FG}(\sigma) \ln \frac{p_{FG}(\sigma)}{p_{GG}(\sigma)} , \tag{5.5}$$

where σ is an observable value. It is not productive to further partition these distributions conditional on CG configurations because a CG observable expression such as Eq. (5.2) can only produce one value while there is a distribution of FG observable values; this relative entropy would be minimized by satisfying Eq. (5.1), which is the same target as that for Eq. (5.4).

In order to minimize the relative entropy in Eq. (5.5), we must look for the place where its derivatives with respect to the linear dependent basis coefficients λ are 0. For simplicity, we will show the update rule for gradient decent, which only requires the first derivative with respect to λ; however, other minimization methods including those that require additional derivatives such as Netwon's method are possible. For gradient decent, the update rule for basis set coefficient k from iteration i to $i+1$ is simply

$$\lambda_{k,i+1} = \lambda_{k,i} - \zeta \frac{\partial S_{rel}}{\partial \lambda_k} , \tag{5.6}$$

where ζ is a step size between 0 and 1 that can be a constant or iteration dependent (such as learning rates in machine learning). Evaluating the derivative of the relative entropy with respect to basis set coefficient k,

$$\frac{\partial S_{rel}}{\partial \lambda_k} = -\left\langle \phi_k\left(\mathbf{R}^N\right) \frac{\partial}{\partial \sigma} \left(\frac{p_{FG}(\sigma)}{p_{CG}(\sigma)} \right)\Bigg|_{\sigma = O\left(\mathbf{R}^N\right)} \right\rangle \tag{5.7}$$

(see Appendix for details). Thus, the update rule is

$$\lambda_{k,i+1} = \lambda_{k,i} - \zeta \left\langle \phi_k\left(\mathbf{R}^N\right) \frac{\partial}{\partial \sigma} \left(\frac{p_{FG}(\sigma)}{p_{CG}(\sigma)} \right)\Bigg|_{\sigma = O\left(\mathbf{R}^N\right)} \right\rangle . \tag{5.8}$$

At the minimum in relative entropy, the basis coefficient does not change according to Eq. (5.8) since the relative entropy derivative is 0. We can see that this is because the probability using the current CG observable expression matches the FG value. Also, the update does not change a basis coefficient when the corresponding basis function is zero.

When the CG observable distribution is locally shifted towards higher observable values relative to the FG observable distribution at $\sigma = O\left(\mathbf{R}^N\right)$, $\frac{\partial}{\partial \sigma} \left(\frac{p_{FG}(\sigma)}{p_{CG}(\sigma)} \right)\Bigg|_{\sigma = O\left(\mathbf{R}^N\right)} > 0$. If the basis function ϕ_k corresponding to λ_k is negative there, then this contributes towards increasing λ_k; this makes sense since increasing the contribution of ϕ_k would decrease observable values, which helps reduce the difference between FG and CG observable distributions; the opposite is true if ϕ_k is positive. If the rest of the derivatives where this basis function is active are positive and ϕ_k

always non-positive, then λ_k is certainly increased. In reality, the update is much more complicated because some observable values are likely to contribute toward increasing the value of λ_k while other observable values do the opposite. Additionally, one must consider the Boltzmann weighting inherent in the CG model in order to calculate the net contributions, which ultimately determines how λ_k is changed by the update.

Based on the assumption that the mapped FG and CG configuration probabilities are equivalent, the amount of simulation required to determine the optimal CG observable expression can be minimized. Specifically, the FG simulation used to determine the CG interactions can be reused to parameterize the CG observable expression. For each CG observable iteration, the same FG observable distribution and mapped FG configurations can be used again. This is because changing the CG observable expression does not affect the distribution of configurations, unlike changes to the CG interactions.

The above derivation is primarily useful for the overall distribution of framewise values although it could be used for other levels of observable granularity. This relative entropy has a different target than the variational approach presented earlier. Here, the target is matching the overall observable distribution, not configuration specific properties. If one successfully minimizes the variational residual in Eq. (5.4), they should also get the observable distribution correct; however, the converse is not necessarily true.

If we wanted a relative entropy that actually satisfy Eq. (5.1), then the probabilities need to be conditional on a given CG configuration:

$$S^{R}_{rel} = \int d\sigma p_{FG}\left(\sigma \mid \mathbf{R}^{N}\right) \ln \frac{p_{FG}\left(\sigma \mid \mathbf{R}^{N}\right)}{p_{CG}\left(\sigma \mid \mathbf{R}^{N}\right)} \ . \tag{5.9}$$

In this expression, the CG probability is simply a delta function at the observable value for the CG configuration using the current CG observable expression. In order to minimize this relative entropy we need the derivative of the relative entropy with respect to basis set coefficient k, which is

$$\frac{\partial S_{rel}}{\partial \lambda_k} = -\left\langle \phi_k\left(M\left(\mathbf{r}^n\right)\right)\frac{\partial \ln}{\partial \sigma}\left(\frac{p_{FG}\left(\sigma \mid M\left(\mathbf{r}^n\right)\right)}{p_{CG}\left(\sigma \mid M\left(\mathbf{r}^n\right)\right)}\right)\Bigg|_{\sigma=O\left(\mathbf{r}^n\right)}\right\rangle \tag{5.10}$$

(see Appendix B for details). This gives an update rule for gradient descent of

$$\lambda_{k,i+1} = \lambda_{k,i} - \zeta\left\langle \phi_k\left(M\left(\mathbf{r}^n\right)\right)\frac{\partial \ln}{\partial \sigma}\left(\frac{p_{FG}\left(\sigma \mid M\left(\mathbf{r}^n\right)\right)}{p_{CG}\left(\sigma \mid M\left(\mathbf{r}^n\right)\right)}\right)\Bigg|_{\sigma=O\left(\mathbf{r}^n\right)}\right\rangle . \tag{5.11}$$

There are three major differences between this expression and the one in Eq. (5.8). First, the derivative is evaluated is at FG observable values instead of CG observable values. This will likely yield a more stable minimization since dramatic changes in the CG observable expression will not change the derivatives to be evaluated here as much. The second difference is that the derivative is of the logarithm of the probability ratio instead of the ratio itself. The implications of these differences are largely overshadowed by the final difference. Since the CG distribution for a given CG configuration is a delta function, the CG observable expression will be minimized when it reproduces the mode of the FG observable distribution for each CG configuration. This is similar to Eq. **Error! Reference source not found.** in that it targets the conditional average of the FG observable distribution, but Eq. **Error! Reference source not found.** uses the mean and this method uses the mode.

5.2.3 New Basis Sets

An important term in our CG observable expressions is the one-body term, which can reproduce the correct ensemble average. An expression for the one-body term $O^{(1)}$ is

$$O^{(1)}\left(\mathbf{R}^N\right)=\sum_i \lambda_i^{(1)}\alpha_i\left(\mathbf{R}^N\right)=\sum_i\sum_I \lambda_i^{(1)}\alpha_i\left(\mathbf{R}_I\right), \tag{5.12}$$

where $\lambda_i^{(1)}$ is the basis coefficient for the one-body basis function α_i. This is a configuration-independent contribution based solely on the particle type. By using a particle-wise value here, these terms can be used for systems of different sizes or compositions directly.

The most basic configuration-dependent term in any CG-observable expression is a function of the pair distance. While it would be possible to use an anti-symmetric (i.e., force-like) formulation, an anti-symmetric formulation is more generally applicable for observables. By symmetric, we mean that both particles in a pair receive observable contributions of the same sign and magnitude. This allows the configuration-dependent terms to contribute to the net framewise observable value. Also, it makes intuitive physical sense. For example, the tensor virial for pressure is the trace of the cross product of the vector between the particles and the force. Since both the vector and the force are anti-symmetric (i.e., equal and opposite), each component-wise product is symmetric because the opposite signs cancel. A formulation for this symmetric basis set is

$$O^{(2)}\left(\mathbf{R}^N\right)=\sum_i \lambda_i^{(2)}\xi_i\left(\mathbf{R}^N\right)=\sum_i\sum_{I<J} \lambda_i^{(2)}\xi_i\left(\mathbf{R}_I,\mathbf{R}_J\right), \tag{5.13}$$

where $\lambda_i^{(2)}$ is the basis coefficient for the one-body basis function ξ_i. To bring out the symmetric nature of the pair symmetric basis function, Eq. (5.13) can be rewritten as

$$O^{(2)}\left(\mathbf{R}^N\right)=\sum_i\sum_{I<J}\lambda_i^{(2)}\eta_i\left(R_{IJ}\right)\left|\hat{\mathbf{R}}_{IJ}\right|, \tag{5.14}$$

where $\xi_i\left(\mathbf{R}_I,\mathbf{R}_J\right)=\eta_i\left(R_{IJ}\right)\left|\hat{\mathbf{R}}_{IJ}\right|$, R_{IJ} is the scalar distance between CG sites I and J, and $\left|\hat{\mathbf{R}}_{IJ}\right|$ is the component-wise absolute value of the unit vector between CG sites I and J. In this form, the pair symmetric basis set is the same as the force from the usual MS-CG pair basis set, but with the absolute value around the unit vector. With this recognition, one can easily create symmetric versions of the usual interactions (e.g., angle, dihedral, three-body,[32, 38] density,[133] order parameter[134]) by applying absolute values to every unit vector. When the observable is a scalar, the unit vector becomes one-dimensional (i.e., equal to 1). For a vector observable with the same dimensionality as the system, the one body contribution is distributed equally across those dimensions.

5.2.4 Observable Decompositions

In order to keep with the stricter, more data-efficient version discussed in section 5.2.A, a particlewise decomposition is needed for atomistic observable contributions. These contributions are then transformed according to the selected CG mapping to obtain the target values that serve as inputs to the parameterization step. For atomistic observables with a pairwise formulation such the pressure and the potential energy for pairwise interactions, the decomposition is straightforward. For other situations, it can be significantly more tricky, but possible to define such a decomposition. One convenient approximation is to decompose contributions of a many-body observable by dividing a given contribution equally among all of the particles involved. In other cases where the target property is a function of another observable, it may make more sense to try decomposing the observable instead of the property. In any case, it would be possible to use the approach discussed in section 5.2.B on the probability distribution.

For the pressure of an atomistic system with pairwise non-bonded and bonded interactions, the usual expression for the configuration part of the virial is

$$P\left(\mathbf{r}^n\right)=\frac{1}{3V}\sum_{i<j}\mathbf{f}_{ij}\times\mathbf{r}_{ij}\,, \tag{5.15}$$

where $\mathbf{f}_{ij}$ F is the force between atomistic particles i and j. The decomposition is formulated such that the total pressure is the sum of per-particle contributions:

$$P\left(\mathbf{r}^n\right)=\sum_{i=1}^{n}P\left(\mathbf{r}_i\right). \tag{5.16}$$

This is achieved by simply dividing the observable contribution from a pair of particles equally between them:

$$P\left(\mathbf{r}_i\right)=\frac{1}{2}\frac{1}{3V}\sum_{i\neq j}\mathbf{f}_{ij}\times\mathbf{r}_{ij}\,. \tag{5.17}$$

Likewise, the potential energy U of an atomistic system is decomposed such that

$$U\left(\mathbf{r}^n\right)=\sum_{i=1}^{n}U\left(\mathbf{r}_i\right)\,, \tag{5.18}$$

where each particles observable contribution $U\left(\mathbf{r}_i\right)$ is determined by dividing each of the potential terms equally between each of the particles involved in the interaction:

$$U\left(\mathbf{r}_i\right)=\frac{1}{2}\sum_{i\neq j}u_{pair}\left(\mathbf{r}_{ij}\right)+\frac{1}{3}\sum_{i\neq j\neq k}u_{angle}\left(\mathbf{r}_i,\mathbf{r}_j,\mathbf{r}_k\right)+\frac{1}{4}\sum_{i\neq j\neq k\neq l}u_{dihedral}\left(\mathbf{r}_i,\mathbf{r}_j,\mathbf{r}_k,\mathbf{r}_l\right)+\ldots\,. \tag{5.19}$$

5.2.5 Properties of Good CG Observables

First and foremost, any CG observable should reproduce the FG ensemble average. In our approach, this is ensured by the configuration-independent one-body terms. For a given system volume, the approaches of Das and Anderson[31] and Dunn and Noid[97] achieve this same effect.

Additionally, it is desirable to describe configuration-dependent fluctuations about the average value. In principle, one could get the same spread of the FG distribution by adding in noise. However, we are focused on how much of the variation in FG observable value is explained by the CG observable value. In order to differentiate between two histograms with similar spread, we look at the time series of the FG and CG observables for the same set of configurations. For simplicity, we quantify the correlation between these values using the standard deviation of the difference between the FG and CG observable values: $std\left(FG\left(t\right)-CG\left(t\right)\right)$. This standard deviation can be viewed as the uncaptured variation. By comparing that value to the standard deviation of the FG observable alone, we can see if and how much the CG observable is helping to explain the variation in the FG observable values. For MS-CODE observable expressions, the standard deviation of the difference is guaranteed to not be higher than that of the FG observable alone because MS-CODE is variationally minimized. Furthermore, the inclusion of additional basis sets is guaranteed not to increase this standard deviation for the same reason. Thus, the upper bound in terms of uncaptured variation for MS-CODE is that of the standard deviation of the FG observable In contrast, naïve observables as well as the configuration-dependent part of Das and Anderson's[31] and Dunn and Noid's[97] approach have no such guarantee. As a result, they can actually lead to a less faithful description than simply using a one-body term.

Also, there are practical advantages to MS-CODE. The observable decomposition is general in that the procedure discussed in the previous sub-section can be applied to any framewise observable. Additionally, it is based on local contributions instead of system level properties, which should improve its transferability to different size systems, systems of different composition, and even – to a lesser extent – different state points. The local decomposition also means that the method is data efficient because it uses N or 3N data points per frame where N is the number of CG sites as opposed to 1 data point in the first relative entropy approach outlined in Sec. 5.2.B or other system level approaches. The local nature of the resulting observable contributions also makes it easier to extract physical meaning from the resulting observable expressions.

5.2.6 Simulation and Fitting Details

Molecular dynamics (MD) simulation were performed on bulk, atomistic methanol (MeOH) and acetonitrile using LAMMPS.[71, 73] All systems were run with a 1 fs time step. Each system contained 1,000 molecules and interaction parameters were taken from the OPLS force field.[74, 136] The nonbonded interactions were calculated using Lennard-Jones (LJ) interactions with a radial cutoff of 1.0 nm as well as particle-particle particle-mesh (PPPM) electrostatic interactions. Both systems were equilibrated for 5 ns under constant NPT at 1 atm and 300 K. For subsequent constant NVT simulation, the box size was set to the average volume of the last 2 ns of constant NPT simulation. Then, the system was equilibrated for an additional 1 ns using LJ shifted-force[20] interactions at constant NVT at 300K before sampling every 250 fs for 2 ns.

Two additional methanol systems were created based on the equilibrated bulk system. The atomistic, liquid vapor methanol system was created expanding the box of equilibrated bulk MeOH 40 A in the z-dimension without rescaling the coordinates. This system was equilibrated

for an additional 2 ns at constant NVT at 300 K. Finally, FG frames were sampled every 250 fs for 2 ns at constant NVT at 300K. The 8,000 molecule atomistic methanol system was created by replicating the bulk 1,000 molecule system eight times. The system was allowed to re-equilibrate for 1 ns before frames were sampled every 250 fs for 2ns at constant NVT at 300K.

Per-particle decompositions of pressure and potential energy were obtained using the stress/atom and pe/atom computes in LAMMPS, respectively. Partially explicit and explicit observable contributions are those that arise from interactions between particles that are not in the same CG site; implicit observable contributions arise from interactions between particles that are entirely within the same CG bead. For 1-site center of mass (COM) CG model, the explicit contributions are those originating from nonbonded LJ and electrostatic terms. For the 2-site CG model, the partially explicit and explicit contributions also include the bond between the atoms in different CG beads for pressure.

All CG interactions and observable expressions were determined using the a modified version of the MS-CG force matching (FM) code that includes the one-body and symmetric basis functions mentioned in this chapter. All CG interactions and observables had a cutoff of 10 A. For the CG force fields, pair interactions were fit using sixth order B-splines with a binwidth of 0.6 A. For the CG pressure expression, one-body and two-body symmetric basis sets were used to fit the on-diagonal terms of the pressure tensor. For the CG potential expression, one-body and two-body symmetric basis sets were used to the scalar potential energy. The two-body symmetric interactions were fit with fourth order B-splines and a binwidth of 0.3 A. CG observable distributions were obtained using the rerun command in LAMMPS to apply the observable fields to the mapped trajectories.

5.3 Results

In this section, established CG models are used to demonstrate the performance of MS-CODE pressure and potential. Established models for methanol and acetonitrile are used so that we can focus on the observables instead of the forces. Pressure is a chosen because it has be the focus for much of the discussion about representability. Also, it demonstrates the ability to parameterize a vector observable expression since we reproduce the on-diagonal terms of the pressure tensor. From this observable, other properties such as the surface tension can be calculated. Potential energy is chosen because it relates to thermodynamic properties that can be used to resolve the enthalpy-entropy partitioning in the CG model. Also, it demonstrates the ability to parameterize a scalar observable expression. Additionally, we use this as an opportunity to explore the differences in the CG pressure expression between different mappings for methanol: 1-site center of mass (COM), 2-site COM, and 2-site center of charge (COC). Likewise, we compare the difference in the CG potential expression between 1-site COM methanol and acetonitrile models. Within each subsection, we present histograms at the level of granularity of (5.5), which serves as a validation of MS-CODE and the one-body terms in particular. The tables of standard devations that follow give more detailed information about the correlation between the FG and CG observable value, which is only a function of the configuration-dependent two-body symmetric terms here.

5.3.1 Pressure of MeOH CG Models

5.3.1.1 1-site Center of Mass

Figure 5-1 shows histograms for the pressure of a 1-site COM CG model. The distribution of pressure values for all FG contributions is shown in subpanel a. The offset of the FG distribution

is the result of choosing a box size using non-shifted force interactions, but sampling at that box size using shifted force interactions. The spread of FG distribution is extremely large suggesting that there are huge fluctuations in pressure. In comparison, both the naïve and the MS-CODE pressure distributions are relatively tight. However, the MS-CODE observable correctly reproduces the correct average pressure while the naïve CG pressure is centered 5500 atm higher.

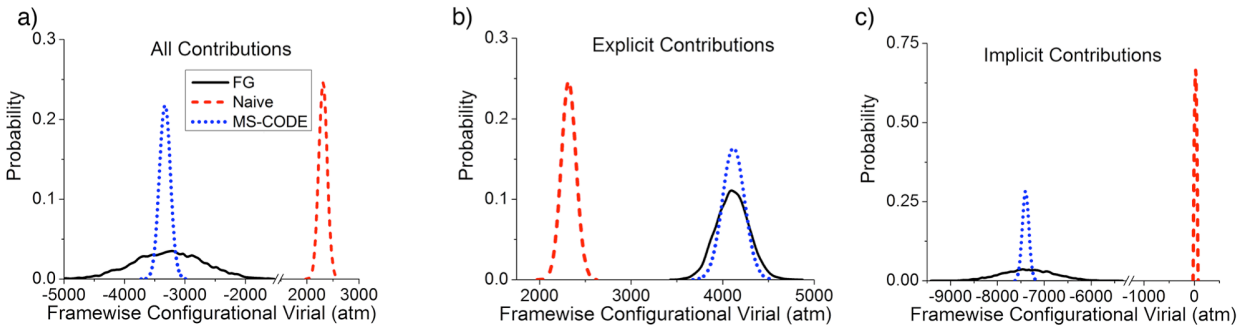

Figure 5–1. Pressure distribution histograms for 1-site center of mass (COM) MeOH models using a) all, b) explicit, and c) implicit FG contributions for FG observable, naïve CG observable, and the MS-CODE observable expressions.

In order to determine the cause of the spread in FG pressure values, the contributions to the pressure are divided into explicit contributions and implicit contributions as described in Sec. II.G. Looking at subpanel b, the spread of the FG observable distribution is only slightly wider than the MS-CODE distribution while the spread of the naïve CG observable distribution is still too narrow. Again, the MS-CODE and FG observable distributions are centered at the same value. In looking at subpanel c, it is clear that the wider spread of the FG pressure is primarily from implicit contributions. It makes sense that it would be difficult for any CG observable expression to reproduce contributions from inside the CG bead; in fact, the naïve CG expression does not reproduce any of these contributions.

To see how the pairwise CG contributions lead to this increase spread, we look at Figure 5-2. Indeed, the implicit contributions from the MS-CODE expression are nonzero while the naïve expression (not shown) is zero everywhere. It appears that there is a net negative pressure contribution at short distances, which suggests that there is some preferential alignment between molecules at this distance before excluded volume contributions dominate at shorter distances. In looking at the explicit pressure contributions from each expression in subpanel b, the naïve expression oscillates significantly more than the MS-CODE and the naïve expression has very large positive contributions at many distances. This is one explanation for why the total naïve pressure is too high. In looking at the net observable expressions in subpanel a, the shape of the MS-CODE expression is determined by the competition between attractive implicit contributions and primarily repulsive explicit contributions. Taken together with the fluctuating naïve expression, this suggests that the net pressure value is the result of many large contributions that largely cancel. In light of this, it is notable that the MS-CODE distribution is as close to the FG observable distribution as it is.

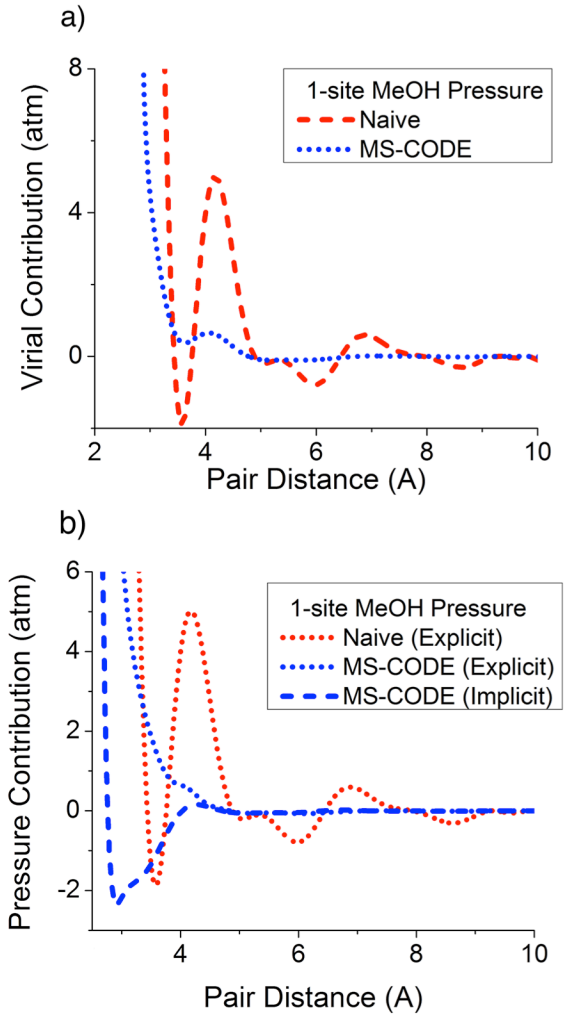

Figure 5–2. Pairwise pressure contributions for 1-site center of mass (COM) MeOH models for naïve and MS-CODE observable expressions using a) all FG contributions as well as the b) non-zero explicit and implicit FG contributions.

5.3.1.2 2-site Center of Mass

Moving on to a higher resolution CG model, we expect both observable distributions to look more like the FG observable expression than they did for the lower resolution model. Figure 5-3 shows the pressure distributions for a 2-site COM MeOH model. Indeed, both the naïve and MS-CODE pressure distributions are much broader here than in Figure 5-1. The overall MS-CODE

distribution is slightly narrower than the FG distribution. This improvement is largely because the MS-CODE expression captures the partially explicit and explicit contributions in subpanel b extremely well. However, the implicit contributions shown in subpanel c still have a very wide spread that it is hard for the CG observable distributions to capture.

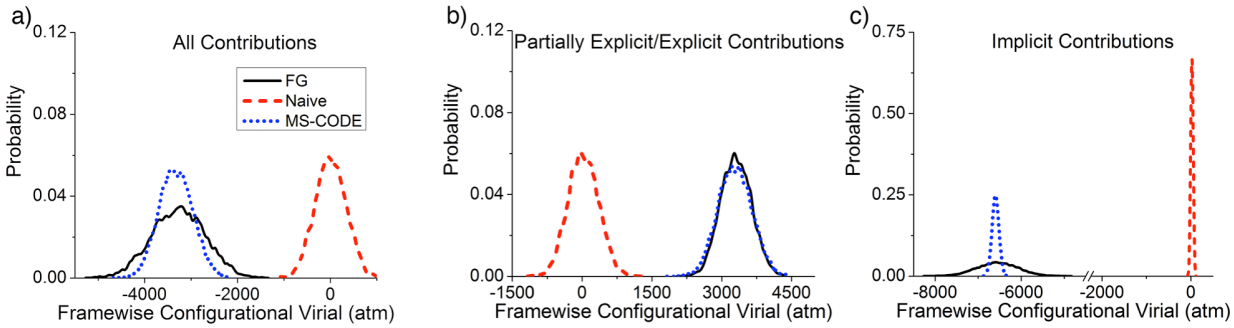

Figure 5–3. Pressure distribution histograms for 2-site center of mass (COM) MeOH models using a) all, b) partially explicit and explicit, and c) implicit FG contributions for FG observable, naïve CG observable, and the MS-CODE observable expressions.

2-site Center of Charge

An alternative mapping for 2-site MeOH is COC. The pressure distributions for this model are shown in Figure 5-4. The broadness of both the naïve and MS-CODE distributions are intermediate between those in subpanels a and b. The naïve distribution looks much more like the other distributions both in terms of spread and center for this model than for any of the others. If this increased similarity between the naïve and MS-CODE pressure values is true for other COC models, then COC models more directly represent pressure contributions to pressure, which means that the magnitude of additional state-dependent term described in our previous

work[91] would be lower here. This would also suggest that COC models might be more transferable than COM model.

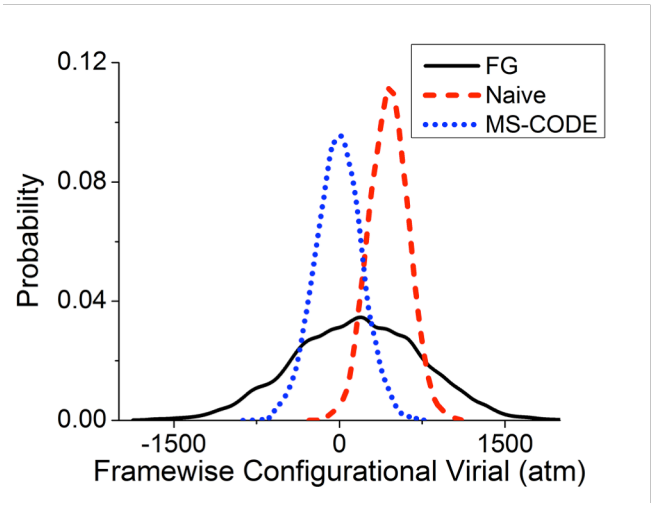

Figure 5–4. Pressure distribution histogram for 2-site center of charge (COC) MeOH model using all FG contributions for FG observable, naïve CG observable, and the MS-CODE observable expressions.

5.3.2 Potential of 1-site CG Models

5.3.2.1 Methanol

Moving on to potential, we start by looking at a 1-site COM MeOH model shown in Figure 5-5. The FG distributions are all quite narrow compared to the pressure distributions. The MS-CODE distribution is once again centered about the same value as the FG distribution in all subpanels. Surprisingly, the MS-CODE distribution is more narrow than the naïve CG observable in subpanels a and b.

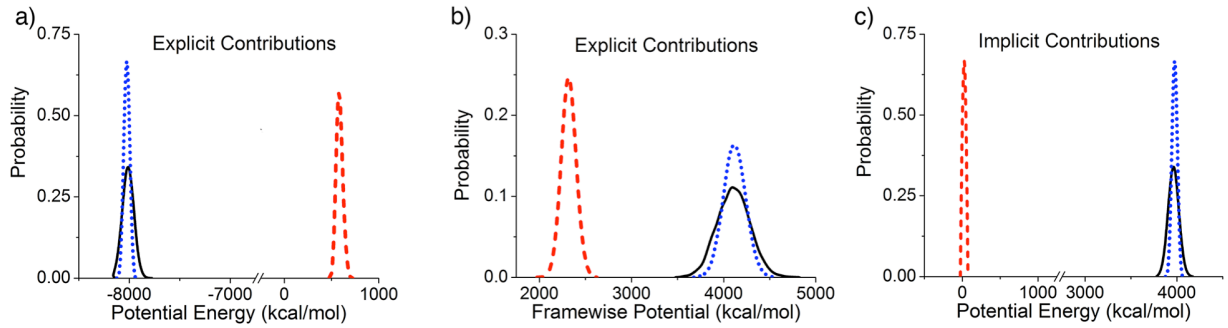

Figure 5–5. Potential distribution histograms for 1-site center of mass (COM) MeOH models using a) all, b) explicit, and c) implicit FG contributions for FG observable, naïve CG observable, and the MS-CODE observable expressions.

To find out why we look at the potential contributions from each CG expression in Figure 5-6. The MS-CODE expression is zero everywhere except for very small distances. In contrast, the naïve CG expression has features and goes to larger values starting at a larger pair distance. Here, the naïve CG potential is the potential of the MS-CG interaction. The decomposition in panel b does not add much. Again, only the explicit contributions lead to nonzero naïve CG potential contributions. The explicit and implicit MS-CODE potential contributions largely cancel leaving a net flat contribution for all but the shortest distances.

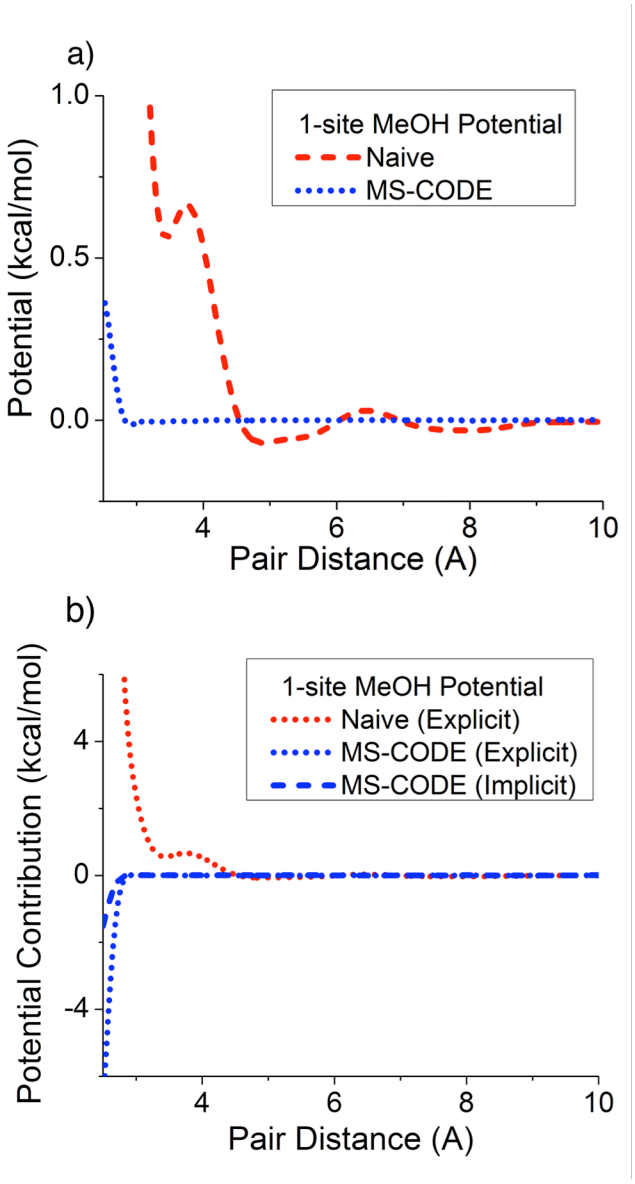

Figure 5–6. Pairwise potential contributions for 1-site center of mass (COM) MeOH models for naïve and MS-CODE observable expressions using a) all FG contributions as well as the b) non-zero explicit and implicit FG contributions.

5.3.2.2 Acetonitrile

The potential distributions for 1-site COM acetonitrile models are shown in Figure 5-7. The FG distributions are even narrower here than it was for MeOH in Figure 5-5. This makes the

narrowness of the MS-CODE distribution look better here by comparison to MeOH. Again, the MS-CODE distribution is narrower than the naïve distribution in subpanels a and b.

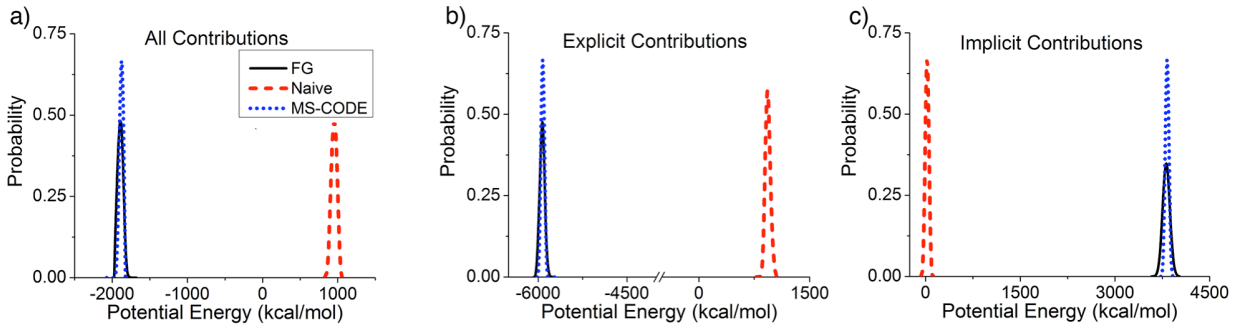

Figure 5–7. Potential distribution histograms for 1-site center of mass (COM) acetonitrile models using a) all, b) explicit, and c) implicit FG contributions for FG observable, naïve CG observable, and the MS-CODE observable expressions.

The potential contributions for the naïve and MS-CODE CG expressions are shown in Figure 5-8. They look very similar to those in Figure 5-6. One difference is that the naïve CG expression is smoother for acetonitrile than it was for MeOH. Another difference is that the net MS-CODE contribution is entirely positive for acetonitrile while it has a small range of negative contributions for MeOH.

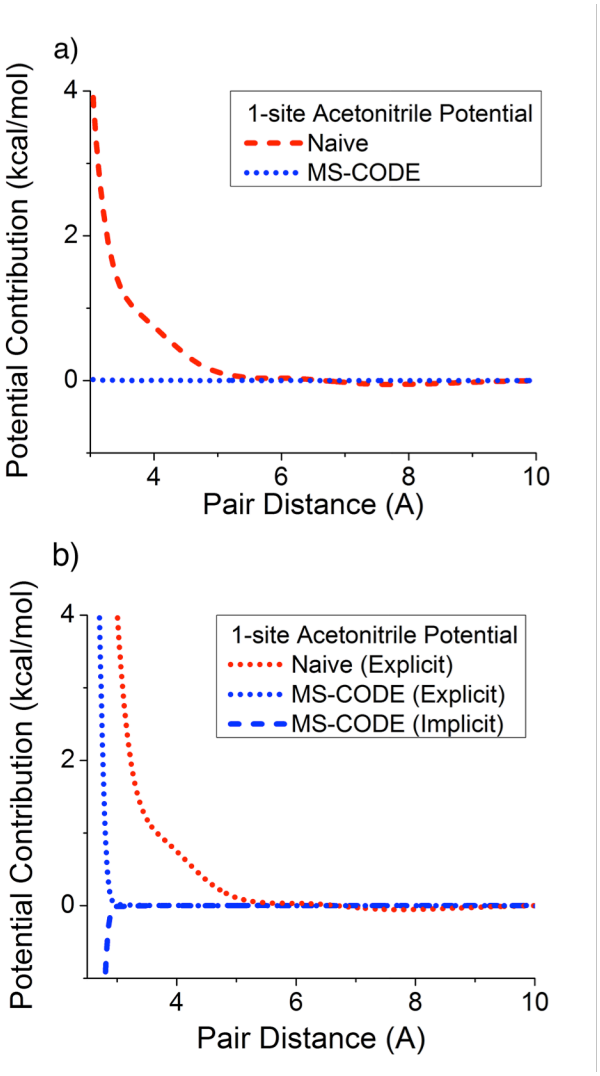

Figure 5–8. Pairwise potential contributions for 1-site center of mass (COM) acetonitrile models for naïve and MS-CODE observable expressions using a) all FG contributions as well as the b) non-zero explicit and implicit FG contributions.

5.3.3 Uncaptured Variation

In order to determine if the increased spread in some of the CG observable distributions is a result of those observables capturing more configuration-dependent variation, we look at the uncaptured variation. First we look at those values in Table 5-1 for the pressure expressions of the different methanol models examined earlier. The uncaptured variation for both the naïve and

MS-CODE observables is the lowest for the 2-site COM CG model. This result implies that pressure is better resolved in COM models than COC models. Also, the broader MS-CODE distributions relative to the naïve distributions in Figure 5-1 are indeed a reflection of the fact that the MS-CODE pressure expression explains more of the variation in the FG pressure of 1-site and 2-site COM CG models.

Table 5-1. Uncaptured variation for the pressure of MeOH using different center of mass (COM) and center of charge (COC) mappings (variation measured as standard deviation). The upper bound for MS-CODE is the variation of the FG model.

System	*Upper Bound*	*Naive*	*MS-CODE*
1-site COM	6.47(9)	6.41(5)	6.38(5)
2-site COM	6.47(9)	5.40(6)	5.38(2)
2-site COC	6.47(9)	6.19(9)	6.28(1)

In both of the potential distributions (Figures 5-5 and 5-7), the naïve expression had broader distributions than MS-CODE. In order to find out if this is because the naïve expression is actually explaining more of the variation in the FG potential, we look at the uncaptured variations in Table 5-2. It shows that the uncaptured variation for the naïve expression is actually larger than both the MS-CODE and "upper bound" values for both MeOH and acetonitrile. This means that the configuration-dependent variation in the naïve CG potential is not correlated with the FG potential value; in fact, it appears to be slightly anti-correlated. In contrast, the uncaptured variation of the MS-CODE expression is slightly below or at the variational upper bound. In this situation, the MS-CODE expression actually does a better job than the naïve expression even though the MS-CODE distribution is narrower. It is important to note that this

has nothing to do with the one-body value; thus, the average-corrected naïve expressions that one would get by applying the approach of Das and Andersen[31] and Dunn Noid[97] would not be any better.

Table 5-2. Uncaptured variation of the potential of MeOH and acetonitrile 1-site center of mass models (variation measured as standard deviation). The upper bound for MS-CODE is the variation of the FG model.

System	*Upper Bound*	*Naïve*	*MS-CODE*
MeOH	0.71(3)	**0.72(3)**	0.71(2)
Acetonitrile	0.61(2)	**0.64(6)**	0.61(2)

5.3.4 Surface Tension

To demonstrate the use of an MS-CODE CG expression for one observable to calculate another CG observable, we calculate the surface tension of a MeOH liquid slab using the MS-CODE pressure expressions from bulk. For an interface normal to the z-dimension, the surface tension γ is defined as

$$\gamma = \left\langle L_z \left(P_{zz} - \frac{1}{2}\left(P_{xx} + P_{yy} \right) \right) \right\rangle , \tag{5.20}$$

where L_z is the length of the interface along the z-dimension, P_{zz} is the normal pressure, and P_{xx} and P_{yy} are the lateral, on-diagonal components of the pressure tensor. The surface tension is completely independent from one-body pressure terms since those terms cancel. Thus, the surface tension is solely a function of the configuration-dependent parts of the CG pressure expression.

The surface tension measurements are shown in Table III. The experimental values of methanol's surface tension are between 20 mN/m and 25 mN/m,[137-140] while surface tensions between 15 mN/m and 20 mN/m have been reported from atomistic simulations.[141-143] The FG value measured from our simulation is in line with the literature simulation values. Likewise, our MS-CODE surface tension value is between the simulation and experimental values. On the other hand, the naïve CG surface tension is a very negative, unphysical value. This shows how important it is to have a compatible pressure measure at level of CG configurations.

Table 5-3. Surface tension of MeOH liquid-vapor interface.

System	Surface Tension (mN/m)
FG	17.(7)
Naïve	-28(8).
MS-CODE	21.(4)

5.3.5 Size Extensivity

To demonstrate the use of MS-CODE on a larger system than the size that it was parameterized using, we use the CG observable expressions from the 1,000-molecule system on an 8,000-molecule system. Figure 5-9 shows the pressure and potential distributions for this larger system. As expected from basic thermodynamics,[19] the distributions are narrower than they were for the smaller systems (Figures 5-1 and 5-5). Again, both the CG observable distributions are narrower than the FG distribution. Likewise, the MS-CODE observable is centered at the same value as the FG distribution. This shows that the MS-CODE observables are applicable to larger systems than the size that it parameterized using.

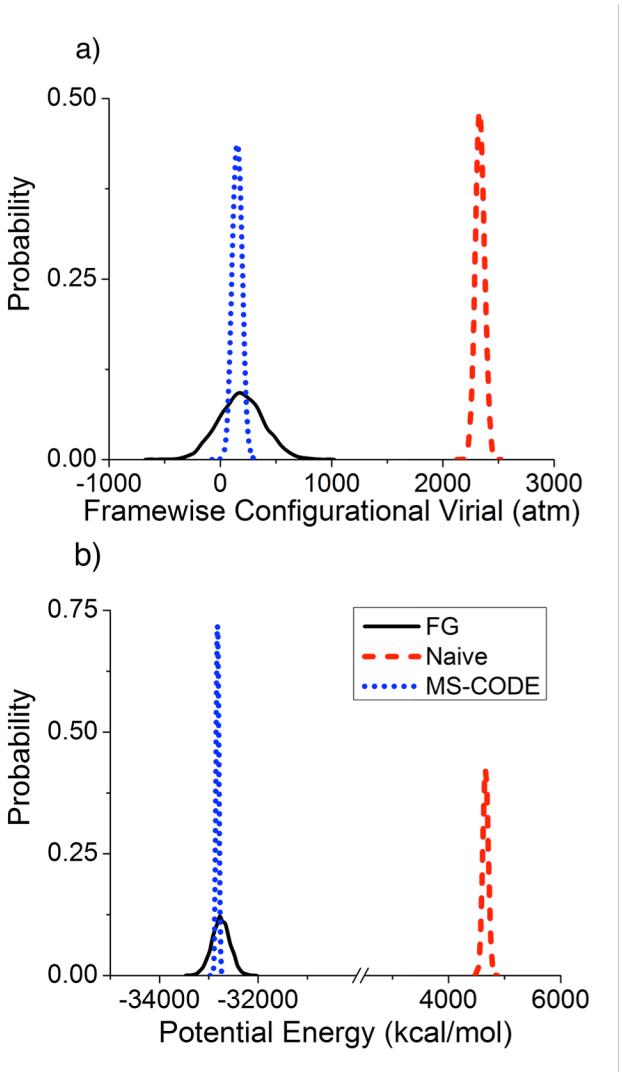

Figure 5-9. Pressure distribution histograms for a system of 8,000 MeOH molecules comparing the actual FG observable, the smaller system's naïve CG observable, and the smaller system's MS-CODE observable expression for a) pressure and b) potential.

5.4 Discussion

We have shown that MS-CODE CG observable expressions are able to correctly reproduce the average FG observable value. In addition, the MS-CODE CG pressure expressions were able to explain more the variation in the FG observable value than the naïve or average-corrected[31, 97]

observable expressions for both 1-site and 2-site COM CG MeOH models. In the case of CG potential, MS-CODE expressions performed better than the naïve expressions. Since MS-CODE expressions are variationally optimized, the will always get the average and will never do worse than that. Consequently, the observable are compatible at the level of the ensemble average and possibly at higher granularity such as the observable distribution or even the CG configuration.

A naïve conception of observable representation would be that only explicit contributions could be represented by CG observable expressions. However, we showed in Figure 2b that MS-CODE is capable of capturing implicit contributions that the naïve or average-corrected[31, 97] observable expressions are not able to capture. While some implicit contributions can be captured, there is a lower limit to how much variation CG observables can explain. This is the variability of an FG observable for a given CG configuration. One could try to measure this using a constraint like the one employed by Davtyan et al.,[144] but this is beyond the scope of this work. Nonetheless, the use of additional basis sets and iterative refinement[84-86] is likely to improve the fraction of implicit contributions that can be represented in a configuration-dependent way.

All of the naïve CG pressure expressions we examined overestimate the true value, in line with what has been anecdotally observed.[15, 131] The histograms of the implicit contributions (Figures 1b and 3b) showed that the FG observable had large negative contributions that were completely missed by the naïve observable. Since these implicit contributions arise from intramolecular interactions, we can say that these contributions are from the restoring force of the harmonic bonds used. Thus, we have evidence to suggest that the naïve pressure will always overestimate the true pressure.

It is not as clear what the net trend is for the naïve CG potential. Like for pressure, the naïve CG potential does not represent any of the implicit contributions. These implicit contributions are very positive because each intramolecular feature has positive contributions to the potential energy unless it is at its equilibrium position. However, the naïve CG potential expression also missed the general location of the explicit contributions including the average sign. In the cases we examined, the excluded volume-type explicit contributions appeared to dominate. However, it is hard to know if this is generally true.

The type of observable histograms used in this chapter can provide information about CG model selection. First, it is a natural analog to existing figures of merit for CG models such as RDFs and density distributions.[91, 133] Also, it can be viewed as a measure for the naïve transferability of a given observable. For example, the increased similarity of the naïve CG pressure distribution to the FG and MS-CODE distributions for the 2-site COC model indicates that the implicit contributions to pressure are of smaller magnitude there, and these implicit contributions can be represented as a state-dependent correction to the naïve expression.[91] Of course, less state-dependence corresponds to increased transferability. In order to construct MS-CODE expressions with more transferability, one could combine the first order estimates from CG sensitivity[41] or combine it with a method that captures state-specific variations.[31, 97]

Also, we have provided further evidence of the problems that arise when using naïve CG observables. Not only did the naïve CG potential miss the implicit contributions, it was actually anticorrelated with the FG pressure. Furthermore, the naïve CG surface tension was completely unphysical. One can only imagine what the "actual" surface tension is for CG models that claim to correctly reproduce the surface tension naïvely and were parameterized based on structural properties. Likewise, the "actual" meaning of CG site and structural properties are likely to be

unintuitive, if such a connection to physically realizable systems exists, for top-down CG models parameterized to reproduce the naïve surface tension. If the observable expressions used to parameterize the top-down model were incompatible, then the physical meaning of the model is indeterminate. Furthermore, neither MS-CODE nor the relative entropy variance could be directly applied to top-down CG models because the observable distributions – let alone the observable decompositions – are not available from experiment, which measure ensemble averaged quantities. Also, it is questionable at best if MS-CODE expressions from structure-based bottom-up CG models would be compatible with an observable-based top-down CG model or an indeterminate top-down CG model.

In this chapter, we also proposed two related relative entropy approaches. The first aimed to reproduce the FG observable at the level of the distribution. It would be interesting to see if this approach performed better than MS-CODE did for the histograms we showed. However, reproducing the observable distributions does not mean that it will do better than MS-CODE at reducing the amount of uncaptured variation (increasing the explained variation). However, if one simply wanted to reproduce the FG observable histogram, they could broaden the CG distributions by adding in fluctuations that are uncorrelated with the physical FG fluctuations. For the second relative entropy approach, the same minimum should be reached as MS-CODE in the limit of infinite sampling and basis sets. One possible advantage of the relative entropy approach is that one does not require a per-particle decomposition of a given observable, which may make it easier to apply to more observables. Nonetheless, both the relative entropy approach and MS-CODE will do better than naïve CG observable or the average-corrected[31, 97] expressions in the literature.

In the literature, much ado has been made about the inability of a bottom-up CG model to simultaneously reproduce the RDFs, pressure, and potential.[42, 43] In this chapter, we have shown how to reproduce any number of FG observables in a bottom-up CG model. In fact, the observables that we demonstrated this for were the very same pressure and potential discussed in the literature.

An important extension of the MS-CODE pressure expressions used here is integration with the state-dependent terms of Das and Anderson[31] and Dunn and Noid.[97] This would combine the configuration-dependent resolution of MS-CODE with volume-dependent terms that reproduce the compressibility and reproduce the average FG pressure at different volumes. This improved CG pressure could then be used as the barostat for constant NPT and related simulations. Now, the structural properties of CG simulations under constant NPT would be more trustworthy since the volume distribution and volume fluctuations would actually reproduce those of a corresponding FG system. This could also have big impacts membrane simulations as the zero tension conditions required by simplified Canham-Helfrich theory could reliably be obtained.[145-147]

Another CG observable that would be particularly useful to have proper representation for is entropy. In particular, scaling relationships have been prosed in the literature between excess entropy and the dynamical speed-up of that system.[148-150] If one could measure the excess entropy locally, one might be able to obtain local speed-up factors that would clarify the different timescales present in the CG system. Additionally, one might be able to construct a configuration-dependent speed-up factor to effectively relate CG time to FG and experimental time. Relatedly, exploration of how the CG potential and pressure behave could have connections to work done on R-simple systems and pseudoisomorphs.[97, 151, 152]

The approach discussed in this chapter could also be applied to other bottom-up methods. MS-CODE is directly applicable to MS-CG[22, 29-40, 153] and g-YBG[86, 154, 155] methods. Likewise, the relative entropy approaches we discussed are obviously applicable to relative entropy[25, 26] methods. Based on the relative entropy approaches, one could develop an inverse Monte Carlo (IMC)[131] or iterative Boltzmann inversion (IBI)[63] method that inverts the observable distribution iteratively to determine the appropriate CG observable expression.

MS-CODE and its related variants could also be applied to other observables. While we have focused on thermodynamic observables here, it could be applied to other CG observables such as the net dipole moment and local dielectric constant. The local decompositions used in MS-CODE allow one to obtain local and even anisotropic properties from these expressions for heterogeneous systems.

5.5 Conclusion

In this chapter, we presented ways to numerically determine CG observable expressions. Our approach, MS-CODE, is a general, data efficient, systematically improvable method to determine such expressions. We showed that is better than naïve CG observables or the average corrected variants[31, 97] for the pressure and potential of center of mass CG models. This method is able to capture implicit contributions and allows compatible sets of CG observables to be constructed.[91]

Future work could include the addition of other observables and the further investigation of those used here. As discussed, combination of MS-CODE with the work of Das and Anderson[31] and Dunn and Noid.[97] Its incorporation into barostats can improve CG NPT, mixed resolution,[120, 121] and adaptive resolution[122-124] simulations. Additionally, the investigation of excess entropy

and the associated dynamical speed-up could further our understanding of the ways in which coarse-graining effects CG dynamics. Also, there is now the ability to investigate CG thermodynamics using MS-CODE CG expressions in a way that was not possible before. Of course, the implementation of other observables can only improve the applicability of CG modeling.

5.6 Appendix A: Derivation of Equation (5.7)

For this derivation, we assume that the usual bottom-up consistency equations have been satisfied. For multiscale coarse-graining (MS-CG), this implies that

$$e^{-\beta U\left(\mathbf{R}^N\right)} \propto \int d\mathbf{r}^n \delta\left(M\left(\mathbf{r}^n\right)-\mathbf{R}^N\right) e^{-\beta U\left(\mathbf{r}^n\right)} \tag{5.21}$$

This allows us to define the observable distributions as

$$p_{FG}\left(\sigma\right)=\frac{\int d\mathbf{r}^n \delta\left(o\left(\mathbf{r}^n\right)-\sigma\right) e^{-\beta U\left(\mathbf{r}^n\right)}}{\int d\mathbf{r}^n e^{-\beta U\left(\mathbf{r}^n\right)}} \tag{5.22}$$

and

$$\begin{aligned} p_{CG}\left(\sigma\right) &= \frac{\int d\mathbf{R}^N \delta\left(O\left(\mathbf{R}^N\right)-\sigma\right) e^{-\beta U\left(\mathbf{R}^N\right)}}{\int d\mathbf{R}^N e^{-\beta U\left(\mathbf{R}^N\right)}} \\ &= \frac{\int d\mathbf{R}^N \int d\mathbf{r}^n \delta\left(O\left(\mathbf{R}^N\right)-\sigma\right)\delta\left(M\left(\mathbf{r}^n\right)-\mathbf{R}^N\right) e^{-\beta U\left(\mathbf{r}^n\right)}}{\int d\mathbf{R}^N \int d\mathbf{r}^n \delta\left(M\left(\mathbf{r}^n\right)-\mathbf{R}^N\right) e^{-\beta U\left(\mathbf{r}^n\right)}} \end{aligned} . \tag{5.23}$$

Evaluating the derivative of the relative entropy with respect to basis set coefficient k,

$$\frac{\partial S_{rel}}{\partial \lambda_k} = -\int d\sigma \frac{p_{FG}(\sigma)}{p_{CG}(\sigma)} \frac{\int d\mathbf{R}^N e^{-\beta U(\mathbf{R}^N)} \frac{\partial \delta(O(\mathbf{R}^N)-\sigma)}{\partial \lambda_k}}{\int d\mathbf{R}^N e^{-\beta U(\mathbf{R}^N)}} \quad . \tag{5.24}$$

Using a derivative trick similar to that in MS-CG I,[22] one can change what the derivative is with respect to from λ_k to σ by recognizing that

$$\frac{1}{\phi_k(\mathbf{R}^N)} \frac{\partial \delta(O(\mathbf{R}^N)-\sigma)}{\partial \lambda_k} = -\frac{\partial \delta(O(\mathbf{R}^N)-\sigma)}{\partial \sigma} \quad . \tag{5.25}$$

Using this substitution and switching the order of the integrals,

$$\frac{\partial S_{rel}}{\partial \lambda_k} = \frac{\int d\mathbf{R}^N e^{-\beta U(\mathbf{R}^N)} \phi_k(\mathbf{R}^N) \int d\sigma \frac{p_{FG}(\sigma)}{p_{CG}(\sigma)} \frac{\partial \delta(O(\mathbf{R}^N)-\sigma)}{\partial \sigma}}{\int d\mathbf{R}^N e^{-\beta U(\mathbf{R}^N)}} \quad . \tag{5.26}$$

Now, one can use integration by parts to transform the σ terms into

$$\int d\sigma \frac{p_{FG}(\sigma)}{p_{CG}(\sigma)} \frac{\partial \delta(O(\mathbf{R}^N)-\sigma)}{\partial \sigma} = \left[\frac{p_{FG}(\sigma)}{p_{CG}(\sigma)} \delta(O(\mathbf{R}^N)-\sigma) \right]_{-\infty}^{+\infty} - \int d\sigma \delta(O(\mathbf{R}^N)-\sigma) \frac{\partial \frac{p_{FG}(\sigma)}{p_{CG}(\sigma)}}{\partial \sigma}$$

$$= \int d\sigma \frac{p_{FG}(\sigma)}{p_{CG}(\sigma)} \frac{\partial \delta(O(\mathbf{R}^N)-\sigma)}{\partial \sigma} = -\frac{\partial}{\partial \sigma} \left(\frac{p_{FG}(\sigma)}{p_{CG}(\sigma)} \right) \Bigg|_{\sigma=O(\mathbf{R}^N)} \quad . \tag{5.27}$$

after applying the delta condition to the derivative of the ratio of the probabilities. Substituting this into Eq. (5.26) and expressing the result as an expectation value over CG variables we obtain Eq. (5.7).

5.7 Appendix B: Derivation of Equation (5.10)

For this derivation, we assume that the usual bottom-up consistency equations have been satisfied just as it was in Appendix A. Here, the observable distributions conditional on the CG configuration are

$$p_{FG}\left(\sigma\,|\,\mathbf{R}^N\right)=\frac{\int d\mathbf{r}^n\delta\left(o\left(\mathbf{r}^n\right)-\sigma\right)\delta\left(M\left(\mathbf{r}^n\right)-\mathbf{R}^N\right)e^{-\beta U\left(\mathbf{r}^n\right)}}{\int d\mathbf{r}^n\delta\left(M\left(\mathbf{r}^n\right)-\mathbf{R}^N\right)e^{-\beta U\left(\mathbf{r}^n\right)}} \tag{5.28}$$

and

$$p_{CG}\left(\sigma\,|\,\hat{\mathbf{R}}^N\right)=\frac{\int d\mathbf{R}^N\delta\left(O\left(\mathbf{R}^N\right)-\sigma\right)\delta\left(\mathbf{R}^N-\hat{\mathbf{R}}^N\right)e^{-\beta U\left(\mathbf{R}^N\right)}}{\int d\mathbf{R}^N\delta\left(\mathbf{R}^N-\hat{\mathbf{R}}^N\right)e^{-\beta U\left(\mathbf{R}^N\right)}}=\delta\left(O\left(\hat{\mathbf{R}}^N\right)-\sigma\right). \tag{5.29}$$

Evaluating the derivative of the relative entropy with respect to basis set coefficient k,

$$\frac{\partial S_{rel}}{\partial\lambda_k}=-\int d\sigma\int d\mathbf{R}^N\,\frac{p_{FG}\left(\sigma\,|\,\mathbf{R}^N\right)}{p_{CG}\left(\sigma\,|\,\mathbf{R}^N\right)}\frac{\partial\delta\left(O\left(\mathbf{R}^N\right)-\sigma\right)}{\partial\lambda_k}\ . \tag{5.30}$$

Using a derivative trick similar to that in MS-CG I,[22] one can change what the derivative is with respect to from λ_k to σ by recognizing that

$$\frac{1}{\phi_k\left(\mathbf{R}^N\right)}\frac{\partial\delta\left(O\left(\mathbf{R}^N\right)-\sigma\right)}{\partial\lambda_k}=-\frac{\partial\delta\left(O\left(\mathbf{R}^N\right)-\sigma\right)}{\partial\sigma}\ . \tag{5.31}$$

Using this substitution and switching the order of the integrals,

$$\frac{\partial S_{rel}}{\partial\lambda_k}=\int d\mathbf{R}^N\phi_k\left(\mathbf{R}^N\right)\int d\sigma\,\frac{p_{FG}\left(\sigma\,|\,\mathbf{R}^N\right)}{p_{CG}\left(\sigma\,|\,\mathbf{R}^N\right)}\frac{\partial\delta\left(O\left(\mathbf{R}^N\right)-\sigma\right)}{\partial\sigma}\ . \tag{5.32}$$

Now, one can use integration by parts to transform the σ terms into

$$\int d\sigma \frac{p_{FG}\left(\sigma|\mathbf{R}^N\right)}{p_{CG}\left(\sigma|\mathbf{R}^N\right)} \frac{\partial\delta\left(O\left(\mathbf{R}^N\right)-\sigma\right)}{\partial\sigma}$$

$$= \left[\frac{p_{FG}\left(\sigma|\mathbf{R}^N\right)}{p_{CG}\left(\sigma|\mathbf{R}^N\right)}\delta\left(O\left(\mathbf{R}^N\right)-\sigma\right)\right]_{-\infty}^{+\infty} - \int d\sigma\delta\left(O\left(\mathbf{R}^N\right)-\sigma\right)\frac{\partial\frac{p_{FG}\left(\sigma|\mathbf{R}^N\right)}{p_{CG}\left(\sigma|\mathbf{R}^N\right)}}{\partial\sigma}$$

$$= \int d\sigma p_{FG}\left(\sigma|\mathbf{R}^N\right)\frac{\partial \ln}{\partial\sigma}\left(\frac{p_{FG}\left(\sigma|\mathbf{R}^N\right)}{p_{CG}\left(\sigma|\mathbf{R}^N\right)}\right). \tag{5.33}$$

after applying the delta condition to the derivative of the ratio of the probabilities. Substituting this into Eq. (5.32) and expressing the result as an expectation value over CG variables we obtain Eq. (5.10).

Chapter 6

Introducing Order Parameter Dependent Interactions for Multiscale Coarse-graining

6.1 Introduction

Coarse-grained (CG) models are an attractive alternative to simulating systems with atomistic models. Using CG models, one can study the same system at reduced computational cost, making it possible to explore longer time- and length-scales than with fine-grained (FG), atomistic models.[14-18] However, this speed-up possible with CG models is only relevant if the CG model can reproduce the behavior of the FG model with sufficient fidelity to study the phenomena of interest.

The CG basis set determines what CG interactions are present in the model, which in turn limits how well bottom-up CG models can reproduce mapped FG structures. Pair basis sets are sufficient to reproduce an RDF,[101] and pair basis sets can reproduce a number of surprising structures.[156] However, systems such as liquid-vapor interfaces[157] and proteins[155, 158] cannot be adequately described using pair interactions alone.

Other systems that cannot be described using only pair interactions, such as single-site CG models of water, can be improved using three-body interactions.[32, 38, 125] While these three-body interactions help, they come with dramatically increased computational cost. In systems where two-body and three-body basis sets are insufficient, one could continue the many body expansion with four-body interactions, etc.; however, the computational cost would soon be prohibitively expensive.

Instead of continuing the many-body expansion, some researchers have jumped directly to the N-body term: Density. Density interactions have been used in CG models ions with implicit solvent,[58, 59, 102] polymers with implicit solvent,[133] models of explosive materials,[60, 159, 160] and

many-body dissipative particle dynamics.[161-165] When correctly implemented, local density interactions can be faster to compute than three-body interactions since density scales a constant multiple of pair interactions while three-body require an additional nested neighbor search. Other researchers have introduced explicit dependence on state variables[43, 56, 96] such as the system volume,[31, 97, 130, 166] is isomorphic with local density. While these models have been successful, these works do not give any indication of what to do if these interactions are insufficient on their own.

Thankfully, there is a smarter way to think about adding interactions. For a long time, researchers have used order parameters (OPs) or collective variables (CVs) to describe phenomena. We mean OPs in the most general sense possible. This includes the different physical meanings ascribed to reaction coordinates,[167, 168] liquid crystal OPs,[169] the local Steinhardt OPs,[93, 170-173] characterizations of the glass transition,[174, 175] and the countless CVs used in enhanced sampling methods.[176, 177] These OPs describe the phenomena in an intuitive way that has direct physical meaning. When the behavior can clearly be grouped into different states, the ultra-coarse-graining (UCG) approach can be used.[90, 153, 178] More generally, however, one may wish to determine additional interactions that do not depend on a discrete number of states.

In this chapter, we introduce new interactions into the multiscale coarse-graining (MS-CG)[22, 29-40] methodology that depend on an arbitrary OP. These new interactions do not depend on state definitions. Additionally, the forces include the appropriate derivatives with respect to the OP needed to correspond a physical potential. The coupling between the usual pair distances and the OP can be either additive or multiplicative, depending on the physical meaning of the derivatives and the complexity necessary to capture the desired phenomena of interest.

MS-CG is a bottom-up CG method that seeks to approximate the many-body potential of mean force (PMF). It converges to the same result as other bottom-up CG methods such as iterative Boltzmann Inversion (IBI),[63] inverse Monte Carlo (IMC),[179] relative entropy (RE),[26] and generalized Yvon−Born−Green (g-YBG)[66, 67] in the limit of infinite sampling and infinite basis set.[25, 28] However, the limited nature of practical basis sets motivates us to choose a method that is data efficient such as MS-CG.[41] Additionally, MS-CG is non-iterative because it usually used with basis sets that are linearly dependent on the coefficients to be optimized.

The rest of the chapter is structured as follows: In Section 2, we define different classes of OPs and formulate basis sets for both additive and multiplicative OP couplings. In Section 3, we present the results of applying local density as an additive local OP as well as distance from a wall as an additive global OP to methanol and acetonitrile systems. In Section 4, we discuss the implications of the results presented in Section 3. In Section 5, we provide conclusions.

6.2 Theory and Methods

OPs can be divided into three major categories based on the locality of the particles needed to compute them. First, local OPs only depend on particles in the neighborhood of interest. Examples of local OPs include local density,[174, 175, 180, 181] the Steinhardt OPs,[93, 170-173, 176] and nearby bonds or particles. Second, molecular OPs depend on particles within the same molecule regardless of how far apart these particles are. Examples of molecular OPs include the radius of gyration, end-to-end distance, and fraction of secondary structure.[112, 113, 182] Third, global OPs depend on a specific reference that is fixed relative to the individual motions of CG particles. An example of a global OP is the distance from a hard-wall such as an electrode.

In order to understand how OP might be incorporated into MS-CG, it is necessary to briefly summarize how basis sets are determined. In MS-CG, the CG interaction potential $U_{CG}\left(\mathbf{R}^N\right)$ is the sum of basis functions $\phi_i\left(\mathbf{R}^N\right)$ multiplied by their coefficients λ_i:

$$U_{CG}\left(\mathbf{R}^N\right)=\sum_{i=1}^{B}\lambda_i\phi_i\left(\mathbf{R}^N\right)\ , \tag{6.1}$$

where $\mathbf{R}^N$ is set the CG positions that are related to the FG positions $\mathbf{r}^n$ through the mapping operator M and i runs through all B position-dependent basis functions. The force on CG site k, $F_{CG}\left(R_k\right)$, is then

$$F_{CG}\left(R_k\right)=-\nabla_k U_{CG}\left(\mathbf{R}^N\right)=-\sum_{i=1}^{B}\lambda_i\frac{d\phi_i\left(\mathbf{R}^N\right)}{dR_k}=-\sum_{i=1}^{B}\lambda_i\sum_{l,m}\frac{d\phi_i\left(\mathbf{R}^N\right)}{d\mathbf{R}_{lm}}\hat{\mathbf{R}}_{lm}\left(\delta_{lk}-\delta_{mk}\right)\ , \tag{6.2}$$

where δ is the Kronecker delta and $\hat{\mathbf{R}}_{lm}$ is the unit vector between CG sites l and m. The coefficients are determined by variationally minimizing the objective function χ^2, which is the normalized sum of squared differences between the mapped FG forces and the CG forces.

Including dependence on an order parameter P in the CG potential introduces an extra layer of complexity. Essentially, there are two ways to decompose this extra dependence that maintain the linear dependence of the coefficients on the basis functions.

6.2.1 Additive Basis Sets

One way to decompose this extra dependence is to assume that the OP-dependent terms are additive with the usual (i.e., pair) terms, meaning that the potentials are completely separate. The CG interaction potential with an additive OP dependence is

$$U_{CG}\left(\mathbf{R}^N,\mathbf{P}^M\right)=U_{CG,pair}\left(\mathbf{R}^N\right)+U_{CG,OP}\left(\mathbf{P}^M\right)=\sum_{i=1}^{B}\lambda_i\phi_i\left(\mathbf{R}^N\right)+\sum_{j=1}^{C}\mu_j\varphi_j\left(\mathbf{P}^M\right), \tag{6.3}$$

where $\mathbf{P}^M$ is the set of M OPs that can be calculated for the system, φ_j is a basis function that depends only on the order parameters, μ_j is the corresponding undetermined coefficient analogous to λ_i, and j runs through all C OP-dependent basis functions. The corresponding force is

$$F_{CG}\left(\mathbf{R}_k\right)=-\sum_{i=1}^{B}\lambda_i\sum_{l,m}\frac{d\phi_i\left(\mathbf{R}^N\right)}{d\mathbf{R}_{lm}}\hat{\mathbf{R}}_{lm}\left(\delta_{lk}-\delta_{mk}\right)-\sum_{i=1}^{C}\mu_i\sum_{j=1}^{M}\frac{d\varphi_i\left(\mathbf{P}^M\right)}{dP_j}\sum_{l,m}\frac{dP_j}{d\mathbf{R}_{lm}}\hat{\mathbf{R}}_{lm}\left(\delta_{lk}-\delta_{mk}\right). \tag{6.4}$$

As expected, the usual forces act separately from the OP-dependent forces.

In order for the OP-dependent part of this interaction to be physically meaningful, the derivatives with respect to CG sites in the definition of the OP must be physically meaningful. For example, density, a local OP, has derivatives that are local and make sense in terms of how one would describe a particle move to a different OP value. In contrast, a molecular OP such as the radius of gyration only gives derivatives that act relative to the center of mass of the molecule. Here, the molecule will expand or contract, but not necessarily in a physically reasonable way.

6.2.2 Multiplicative Basis Sets

Another way to decompose this extra dependence is to assume that the OP-dependent basis sets are multiplicatively coupled with the usual basis sets, meaning that the potentials are

multiplied by each other but are independent in a statistical sense. The CG interaction potential with a multiplicative OP dependence is

$$U_{CG}\left(\mathbf{R}^N,\mathbf{P}^M\right)=U_{CG,pair}\left(\mathbf{R}^N\right)*U_{CG,OP}\left(\mathbf{P}^M\right)=\sum_{i=1}^{B}\lambda_i\phi_i\left(\mathbf{R}^N\right)*\sum_{j=1}^{C}\mu_j\varphi_j\left(\mathbf{P}^M\right)=\sum_{i,j}^{B,C}\alpha_{ij}\phi_i\left(\mathbf{R}^N\right)\varphi_j\left(\mathbf{P}^M\right) \qquad (6.5)$$

where $\alpha_{ij}=\lambda_i\mu_j$. This composite coefficient α_{ij} is what would be determined through MS-CG. The corresponding force is

$$F_{CG}\left(\mathbf{R}_k\right)=-\sum_{i,j}^{B,C}\alpha_{ij}\sum_{l,m}\sum_{p=1}^{M}\left(\frac{d\phi_i\left(\mathbf{R}^N\right)}{d\mathbf{R}_{lm}}\varphi_j\left(\mathbf{P}^M\right)+\phi_i\left(\mathbf{R}^N\right)\frac{d\varphi_j\left(\mathbf{P}^M\right)}{dP_p}\frac{dP_p}{d\mathbf{R}_{lm}}\right)\hat{\mathbf{R}}_{lm}\left(\delta_{lk}-\delta_{mk}\right) \qquad (6.6)$$

Here, the force from the derivative of the usual basis sets is modified by the OP-dependent basis set and visa versa. This intimate coupling means that completely new sorts of interactions are possible here. However, this coupling also means that one needs to be careful about the how both the basis sets and their derivatives behave in order to keep the computational costs of evaluating such interactions under control. There are three primary assumptions about where position-OP basis set products could be non-zero, which correspond to different cutoffs for this coupling.

At one extreme, all possible products could be evaluated. This Cartesian product corresponds to a cutoff of half the box length. It is the most expensive to calculate and is unlikely to be physically meaningful because there is a certain non-locality to this coupling that is not motivated by the definition of the OP itself.

Another option is to only consider position-OP basis set products if there exist particles involved in each basis set that are within a finite distance. This implements a certain locality to the coupling. Nonetheless, it is likely to only be computationally worthwhile if this coupling

cutoff is shorter than the usual pair cutoff. This type of interaction could be used to capture changes in solvation structure near a molecule or a wall.

At the other extreme, one could only consider position-OP basis set products if the position and OP basis sets have a particle in common. This "specific" interaction corresponds to a coupling cutoff of 0. A special case of this interaction is coupling two pair interactions multiplicatively, which could simplify to be the sort of three-body interaction used in MS-CG IX.[32] Also, this sort of interaction could be used to couple a molecular OP with nearby solvent particles to capture very local changes in solvation.

To make sure that the multiplicative position-OP coupling is computationally worthwhile, there are some consideration that need to be taken into account. For example, one should make sure that the phenomena are not adequately described using a cheaper interaction such as an additive OP. Also, it is important to avoid situations where the multiplicative coupling could lead to both terms in Eq. (6.6) acting in the same way. For example, coupling pair and density interactions could create forces from pair derivatives that are modulated by the density basis and density derivatives that are modulated by the pair basis. Both of these contributions modify pairwise forces based on the density, but in slightly different ways. Thus, it is possible that the fitting may not be properly constrained and could lead to bad CG interactions.

6.2.3 Order Parameters Used In This Chapter

For our first exploration of OP-dependent basis functions, we will focus on the computationally cheaper additive basis set formulation. As discussed above, we believe that additive local and global OPs can improve on pair-only interactions through the introduction of physically meaningful OP-dependent interactions.

Local Density

The local density ρ at CG site I is the sum weight function contributions from neighboring CG sites:

$$\rho_{AB}\left(R_I\right)=\sum_{I\neq J\in B} m_J w_{AB}\left(R_{IJ}\right) , \tag{6.7}$$

where the A and B refer to groups of CG sites, I is in the group A, and m_J is the density weight of CG site J. Here, there is the ability to distinguish between where the density is calculated (group A) and what contributes to that density (group B). Also, the density weight allows for different types of densities such as number (m=1), mass (m= mass), and charge (m= charge) densities. In this chapter, we will only use number density, which is would give the same net interactions as mass density for our 1-site CG models.

The force for this additive-density interaction on CG site k is

$$F_{CG}\left(\mathbf{R}_k\right)=-\sum_{i=1}^{B}\lambda_i\sum_{l,m}\frac{d\phi_i\left(\mathbf{R}^N\right)}{d\mathbf{R}_{lm}}\hat{\mathbf{R}}_{lm}\left(\delta_{lk}-\delta_{mk}\right)-\sum_{i=1}^{C}\mu_i\sum_{A,B}^{\{\mathbf{A},\mathbf{B}\}}\sum_{l\in A}^{N}\frac{d\varphi_i\left(\rho^M\right)}{d\rho_{AB}\left(R_l\right)}\sum_{l\neq q\in B} m_q\frac{dw_{AB}\left(\mathbf{R}_{lq}\right)}{d\mathbf{R}_{lq}}\hat{\mathbf{R}}_{lq}\left(\delta_{lk}-\delta_{qk}\right)$$

, (6.8)

where $\left\{\mathbf{A},\mathbf{B}\right\}$ is the set of all possible combinations of A and B groups that define the different densities calculated at CG site l. It is important to note that the density-dependent part of this force is a combination of contributions from the densities calculated at CG site k and from CG site k's contribution to the densities calculated at nearby CG sites. Using definitions of the A and B groups in Eqs. (6.7) and (6.8), one can effectively control whether a CG site's density-dependent force is from only densities calculated at other CG sites, densities calculated at its location, both other CG sites and its location, or even if it has no density-dependent force at all.

With this control, one can tune both the data requirements to fit these interactions and the computational cost of those interactions. Regardless of how many different densities are calculated, this interaction can be evaluated with only two passes through the neighbor lists: first calculate the density at each CG site, and then calculate contributions to the force. This means that these interactions should scale as well as pair interactions, albeit with a bigger prefactor.

For the weight function, we use the Lucy function popularized through its use in smooth particle hydrodynamics (SPH):[183 184]

$$w_{AB}(R_{IJ}) = \left(R_{AB,C} - R_{IJ}\right)^3 \left(R_{AB,C} + 3R_{IJ}\right) / R_{AB,C}{}^4 \quad , \tag{6.9}$$

where R_C is the weight function's cutoff. This weight function is computationally cheap to evaluate, goes to zero at the cutoff, and its derivative goes to zero at the cutoff by construction. The weight function's length scale is based on the cutoff value. If the weight function cutoff is taken to be same as the pair CG cutoff as it is in this work, then there are effectively no free parameters.

For methods that calculate a truly local density such as our approach and that of Sanyal and Shell,[133] the weight function used to calculate density is of critical importance. The weight function derivatives control where significant non-zero forces can be. For local density as an additive OP, these derivatives are only the manifestation of density-dependence. There are three general groups of weight functions: step, switching, and continuously varying.

The most intuitive option is a step function because it directly corresponds to the binned density N/V. However, a step function is actually a very bad choice for an additive OP because it has a discontinuous derivative and is zero elsewhere, which means that it would not produce any finite forces.

Another option is a steep switching function, which is a continuous approximation of a step function. Sanyal and Shell[133] employed such a switching function that was computationally cheap and had continuous derivatives everywhere, including at the cutoff. One issue with this choice is that the weight function derivatives and thus the forces are only significantly non-zero is the small window where the weight function switches. In the polymer system studied by Sanyal and Shell,[133] this worked out because the polymer almost always had a bead in the zone where the switching function derivatives were significantly non-zero that could transmit the density-dependent forces to the rest of the polymer through its bonds. More generally, however, this narrow of zone of significantly non-zero forces is not guaranteed to produce physically reasonable results. CG sites closer than this zone and CG sites outside of this zone do not feel any significant forces from this density-dependent interaction, meaning that only CG sites within this zone are directly affected.

In contrast, continuously varying weight functions such as a Gaussian or the Lucy function used in this study have significant non-zero derivatives everywhere between 0 and the cutoff. They can also be set-up so that they go to zero and have a derivative that go to zero at the cutoff. In our opinion, this type of switching function is the most physically reasonable choice for additive density-dependent interactions because all CG sites within the weight function cutoff can feel significant non-zero forces. This sort of analysis also applies to other additive OP interactions: The most physically reasonable definition for additive OPs will have significantly non-zero derivatives everywhere inside their cutoff.

Now that we have presented our approach to local density and justifications for that choice, comparisons can be made with other approaches in the literature. As was noted by Sanyal and Shell,[133] the approaches of Allen and Rutledge's[58, 59, 102] density dependent implicit solvent model

and Izvekov et al.'s[60, 159, 160] density-modulated pair potentials are derived from several systems at different global densities that are then used to determine interactions at local densities. In particular, the interpolation between densities used by Izvekov et al.[60, 159, 160] does not include the density derivative that would be required for the sort of interactions we described above. Both of these approaches are in contrast to truly local densities used determine the density-dependent interactions by us and Sanyal and Shell.[133]

With that being said for additive OPs, the same is not true for multiplicative OPs. Due to the coupling between the usual (i.e., pair) basis sets and the OP-dependent basis set, it is probably better if the OP has limited derivatives since it will still be able to modify the usual basis sets without encountering the issue mentioned in Section 6.2.B. For local density, this means that relatively steep switching functions like that used in a recent UCG paper[178] are probably the most reasonable choice for a multiplicative OP-dependent basis set.

In addition to the weight function advantage, our approach has other advantages over the relative entropy approach.[133] As mentioned before, MS-CG is more data efficient, local, and designed to be non-iterative. These mean that MS-CG could be parameterized with relatively less FG data and no need for any CG simulation. Additionally, extrapolation like that discussed in MS-CG X[34] can be applied as opposed to simply setting the interaction or potential to 0 outside of the sampled region.[133]

Distance from a Wall

The distance from a hard wall is a global OP since each CG site's OP value is with respect to the same reference. If we assume that the wall is in the xy-plane, then its position can be characterized by a z-value z_0. Then, the OP value for CG site I with respect to the wall is simply $z_{I0} = \left| z_I - z_0 \right|$. The corresponding force is

$$F\left(\mathbf{R}_k\right) = -\sum_{i=1}^{B} \lambda_i \sum_{l,m} \frac{d\phi_i\left(\mathbf{R}^N\right)}{d\mathbf{R}_{lm}} \hat{\mathbf{R}}_{lm}\left(\delta_{lk} - \delta_{mk}\right) - \sum_{i=1}^{C} \mu_i \frac{d\varphi_i\left(z_{0k}\right)}{z_{0k}} \hat{z}_{0k}, \tag{6.10}$$

where the OP-dependent force only acts in the z-direction. A cutoff is implemented for this interaction since it is expected to help reproduce structures near the wall that differ from bulk. Obviously, the OP-dependent force can be implemented cheaply with a simply loop through all CG sites.

6.2.4 Simulation and Fitting Details

Molecular dynamics (MD) simulations were performed using all-atom (AA) models of methanol (MeOH) and acetonitrile in LAMMPS.[71, 73] The interactions for both molecules were taken from the OPLS AA force field.[74, 136] Short-ranged non-bonded interactions were evaluated using Lennard-Jones (LJ) interactions with a radial cutoff of 10 A. Electrostatics were evaluated using particle-particle particle-mesh (PPPM). The AA time step was 1 fs.

Before creating the AA interface systems studied in this chapter, bulk systems of pure acetonitrile and pure methanol each containing 1,000 molecules were created and equilibrated. The box size was determined by averaging the last 2 ns of a 5 ns-long simulation at constant NPT at 1 atm and 300 K. The bulk system was then further equilibrated for 2 ns at constant NVT at 300K.

The AA (i.e., FG) liquid-vapor system was created by expanding the box 40 A in the z-dimension without rescaling the coordinates. This system was equilibrated for an additional 2 ns at constant NVT at 300 K. Finally, FG frames were sampled every 250 fs for 2 ns at constant NVT at 300K.

The AA acetonitrile system confined by a hard wall was created by deleting molecules from the bulk simulation that were within 2.5 A of the edge of the box in the z-dimension. Then hard

walls were added at the maximum and minimum z-values to make the system non-periodic. These hard walls interacted with the acetonitrile atoms through Weeks-Chandler-Andersen (WCA)[185] interactions with σ = 2.00452 A and corresponding cutoff of 2.25 A. The system with the hard walls was allowed to equilibrated for 2 ns at constant NVT at 300 K before FG frames were sampled every 250 fs for the next 2 ns at constant NVT at 300K.

Each system was mapped to a 1-site, center of mass (COM) model. CG interactions were determined using the MS-CG force matching (FM) code. All CG interactions had a cutoff of 10 A. For MeOH, pair interactions were fit using sixth order B-splines with a binwidth of 0.6 A and density interactions were fit using fourth order B-splines with a binwidth of 0.5 density units using a Lucy weight function with a cutoff of 10 A. For acetonitrile, pair interactions were fit using fourth order B-splines with a binwidth of 0.2 A and density interactions were fit using fourth order B-splines with a binwidth of 0.5 density units using a Lucy weight function with a cutoff of 10 A. In the acetonitrile-wall system, the CG wall interactions were determined by subtracting out the CG pair and density interactions before using a custom version of the FM code to fit interactions only in the z-dimension with fourth order B-splines using a binwidth of 0.1 A.

CG MD simulations were also performed using LAMMPS. Starting configurations were created by mapping an FG configuration. The CG system was equilibrated for $1x10^6$ steps at constant NVT at 300 K before a $2x10^6$ step production run. Frames were sampled from the production run every 250 CG steps.

6.3 Results

6.3.1 MeOH Liquid-Vapor Interface

Previous researchers have established and studied the ability of AA MeOH models to form a liquid-vapor interface in agreement with experimental values.[142, 143, 186-189] In contrast, the only study that looked at the ability of a CG MeOH model to form a liquid-vapor interface found that pair interactions were not adequate to structurally reproduce the FG (i.e., AA) model from which it was derived.[157] Thus, we start by investigating if additive local density CG interactions can improve the agreement between FG and CG MeOH liquid-vapor interfaces.

Figure 6-1 shows the how the different 1-site CG MeOH models perform at reproducing the FG MeOH liquid-vapor interface's radial distribution function (RDF), profile across the interface, and density distribution. As expected, the CG model with only a pair interaction is insufficient to reproduce the key features of this interface. While the pair CG model does a decent job of reproducing the mapped FG RDF, the density distribution is much too broad skewing to the left and the most probable density value is too high relative to the mapped FG distribution. Using pair and density interactions that were fit simultaneously, the (pair + density) CG model does an impeccable of reproducing the FG profile. Additionally, it almost perfectly reproduces the density distribution, and the reproduction of the mapped FG RDF is no worse than the pair CG model. To make sure that this is not simply from the density interactions, we made a CG model with only density interactions. It gives virtually no structure in the RDF or profile, and it is not even close to getting the density distribution close. So, clearly the combination of pair interactions with the additive local density-dependent interaction is necessary to reproduce the mapped FG MeOH liquid-vapor interface well.

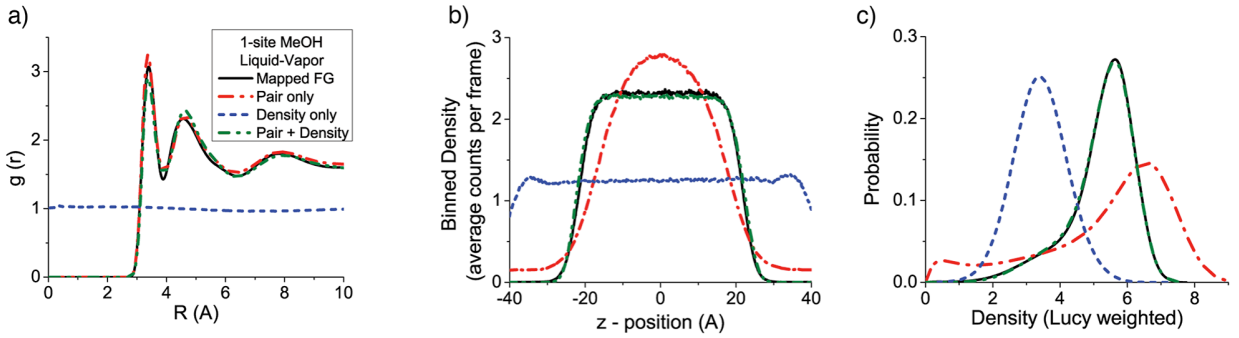

Figure 6–1 - Comparison of the a) radial distribution function (RDF), b) density profile across the liquid-vapor interface, and c) density distribution (as measured using the Lucy weight function with a cutoff of 10 A) of the mapped FG system against CG models with pair, density, and pair + density CG interactions for 1-site MeOH liquid-vapor system.

Now, we look at the different CG interactions, shown in Figure 6-2, to investigate how these models are different. The pair potential from the pair CG model most closely resembles that which would be obtained from Boltzmann inversion of the RDF. Likewise, the density potential from the density CG model generally resembles what would be obtained from Boltzmann inversion of the density distribution. The pair potential from the pair CG model is uniformly more attractive than that from bulk. This suggests that the more favorable pair interaction from the pair CG model is trying to represent the cohesion necessary to create a liquid phase. For the pair and density CG model, this is taken care of by the density interactions. Surprisingly, the pair potential from the pair and density CG model closely resembles the CG pair potential from bulk. This similarity suggests that the addition of the additive density interactions improved the naive transferability of the pair interaction.

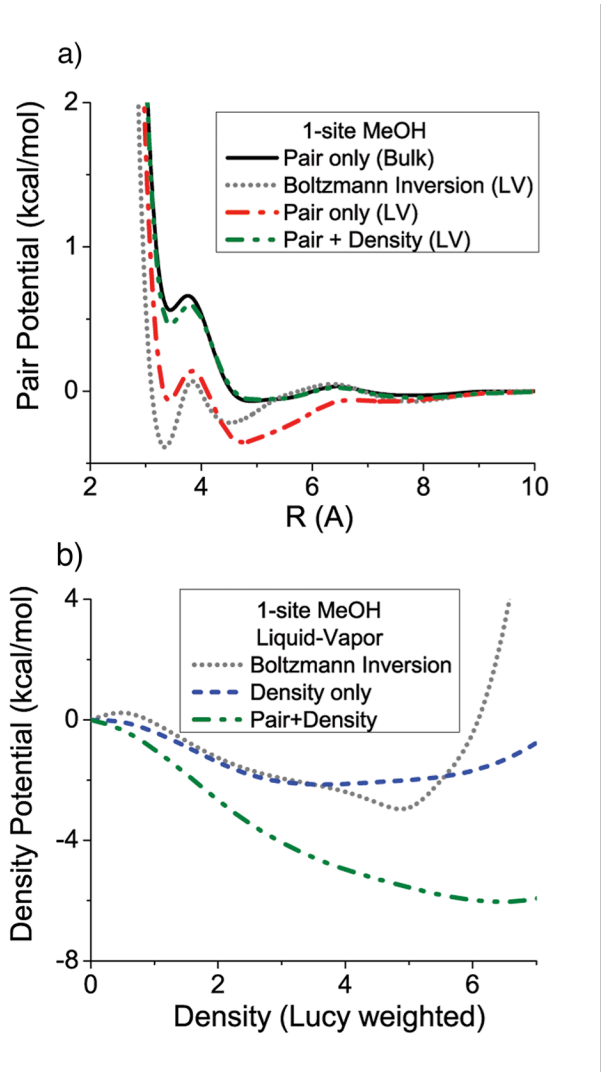

Figure 6–2 – Comparison of the a) pair CG potentials and b) density CG potentials between 1-site MeOH CG models employing MS-CG pair, density, and pair + density CG potentials for the MeOH liquid-vapor system. The MS-CG pair CG potential from bulk MeOH and the Boltzmann inverted CG pair and density potentials for the MeOH liquid-vapor system are also shown.

6.3.2 Acetonitrile Liquid-Vapor Interface

Next, we turn our attention to acetonitrile, a molecule with a stronger dipole moment than MeOH. The effects of this stronger dipole on the liquid structure should be more challenging to capture using a 1-site CG model. In the literature, it appears that there is limited FG simulation

of its bulk structure[190-192] and limited CG simulation using united atom (UA), 3-site CG models,[193-196] but no reports of a bulk 1-site CG model for acetonitrile. So, we begin by creating a 1-site COM CG model for acetonitrile. The RDF and density distribution for bulk 1-site CG acetonitrile are shown in the Appendix, and the pair interaction from this CG model will be use for comparison later.

Figure 6-3 shows the how the different 1-site CG acetonitrile models perform at reproducing the FG acetonitrile liquid-vapor interface's RDF, profile across the interface, and density distribution. First, it is worth noting that mapped acetonitrile has a less concentrated liquid phase than MeOH in terms of COM density (comparing Figure 6-1c and Figure 6-3c). Also, the first peak in the acetonitrile COM RDF is further out than it is in MeOH's COM RDF. Instead, the acetonitrile COM RDF has a pronounced shoulder before the first peak.

Looking at the CG models, neither the pair CG interaction nor the density CG interaction is sufficient to create an interface at all. Correspondingly, both models have density distributions that are significantly shifted to the left. However, the pair CG RDF at least has the same features as the mapped FG RDF. The pair and density CG model is the only one that creates a liquid-vapor interface at all. That being said, the CG liquid phase is slightly too dense which shifts the density distribution towards higher densities, widens the vapor phase, and the make RDF systematically too high. Nonetheless, the pair and density CG model has the correct features and shape in both the RDF and the density distribution. Here, the supplementation of the pair interactions with an additive density interaction makes it possible to create an interface that is qualitatively correct and would have otherwise been impossible to attempt with a CG model.

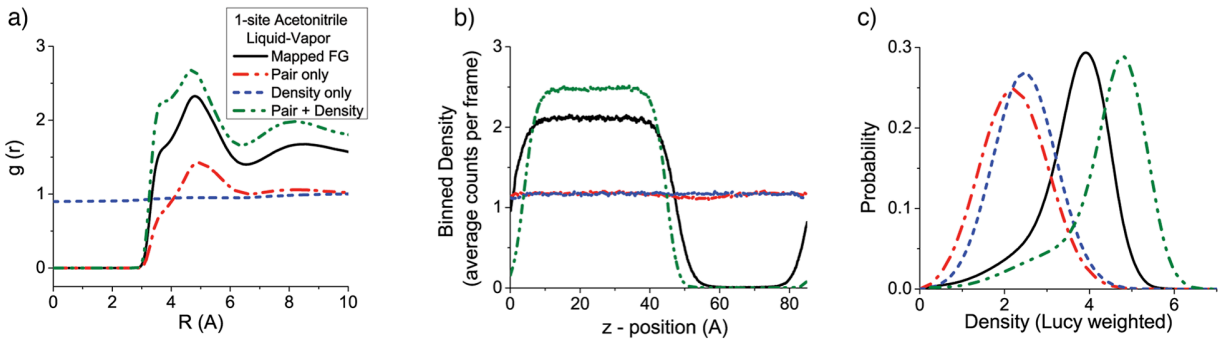

Figure 6–3 - Comparison of the a) radial distribution function (RDF), b) density profile across the liquid-vapor interface, and c) density distribution (as measured using the Lucy weight function with a cutoff of 10 A) of the mapped FG system against CG models with pair, density, and pair + density CG interactions for 1-site acetonitrile liquid-vapor system.

Figure 6-4 shows the CG interactions for the acetonitrile liquid-vapor system. Unlike the MeOH liquid-vapor interface system, the potentials from Boltzmann inversion do not particularly resemble any of the MS-CG potentials. Additionally, the pair potential from the pair and density CG model does not resemble the bulk pair potential. Instead, the pair and density CG model is significantly more attractive than the pair CG model for both the liquid-vapor and bulk systems. Likewise, the density potential from the pair and density CG model is more attractive than that from the density CG model. Here, it seems that both the pair and density interactions are needed to make the CG model attractive enough to create an interface.

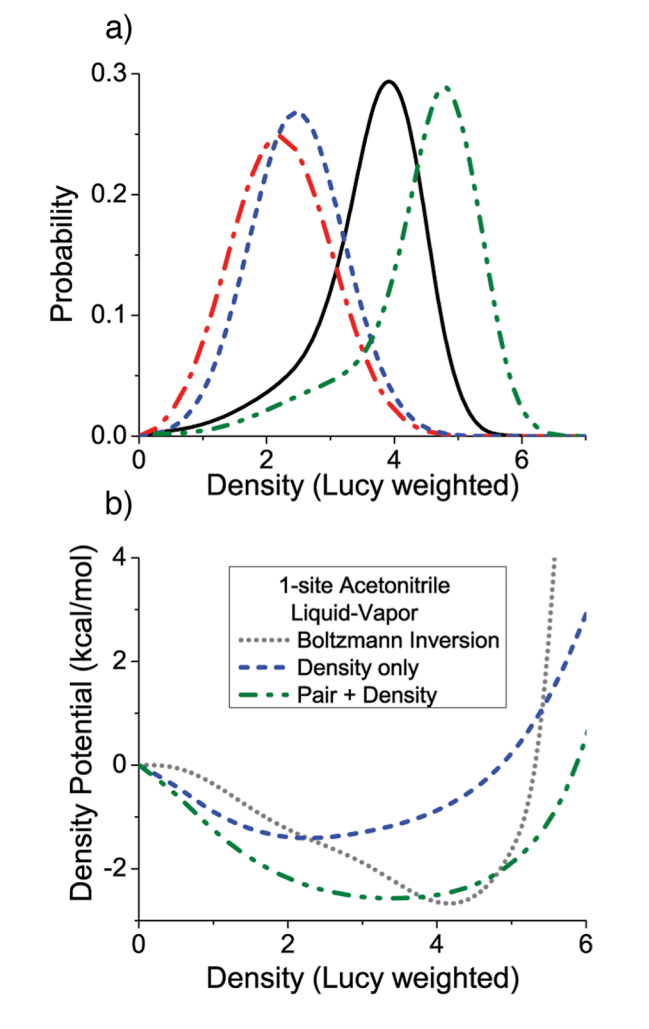

Figure 6–4 - Comparison of the a) pair CG potentials and b) density CG potentials between 1-site acetonitrile CG models employing MS-CG pair, density, and pair + density CG potentials for the acetonitrile liquid-vapor system. The MS-CG pair CG potential from bulk acetonitrile and the Boltzmann inverted CG pair and density potentials for the acetonitrile liquid-vapor system are also shown.

6.3.3 Acetonitrile Liquid-Wall Interface

Another type of interface is a liquid-solid interface. For the purposes of demonstration, we of choose to represent the solid by hard walls that confine acetonitrile liquid. The structural

properties of the different models for this system are shown in Figure 6-5. The z-profiles for the AA system is a very strongly peaked 1.5 A away from the walls with less pronounced, boarder peaks at 3 A, 5 A, and 10 A away from the walls.

The CG sites for all of the systems in Figure 6-5 have the same interaction with wall as the FG sites did. The CG model with only a density interaction once again fails to get the RDF and even has a bimodal density distribution, which corresponds to the peak closest to the walls and the density in the middle of the z-profile. The pair CG model has an RDF and a density distribution that is slightly too high. It also has the peak closest to the wall, but it has too high of a density in the middle of the z-profile. The pair and density CG model improves on the pair CG model's RDF by decreasing its intensity at intermediate distances, which is reflected by an improved density distribution. Likewise, the z pair and density CG model improves on the pair CG model with a density in the middle of the z-profile that is closer to the mapped FG value. Additionally, the pair and density CG model seems to have the features at 5 A and 10 A that were not captured by any of the CG models.

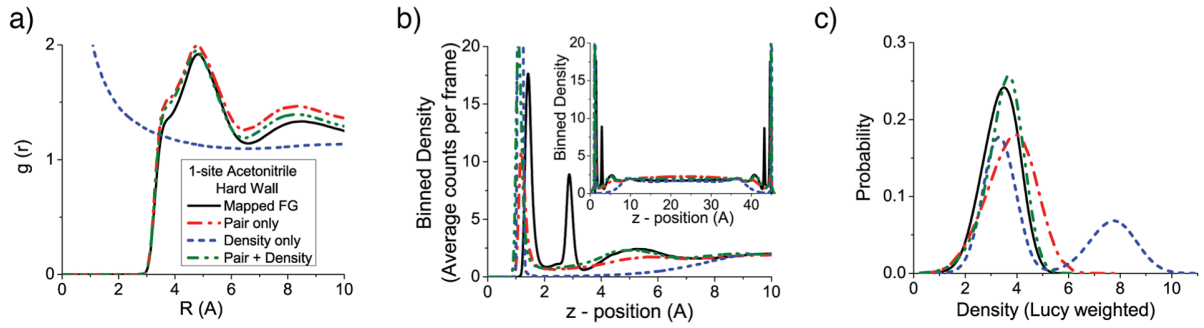

Figure 6–5 - Comparison of the a) radial distribution function (RDF), b) density profile (full-profile inset), and c) density distribution (as measured using the Lucy weight function with a cutoff of 10 A) of the mapped FG system against CG models with pair, density, and pair + density CG interactions for 1-site acetonitrile confined by two hard walls. In this figure, the walls

have the same WCA interaction with the CG sites as they did with the atoms in the FG (i.e., AA) system.

Looking at the corresponding CG potentials, Figure 6-6 shows that all of the MS-CG pair potentials for the wall system are all more attractive than for the bulk. They are also very different than the Boltzmann inversion interaction. In keeping with the behavior observed for the acetonitrile liquid-vapor system, the pair and density CG model once again improves on the pair CG model with an interaction that more attractive overall. Likewise, the density interaction from the pair and density CG model is more attractive than the density interactions from both the density CG model and Boltzmann inversion.

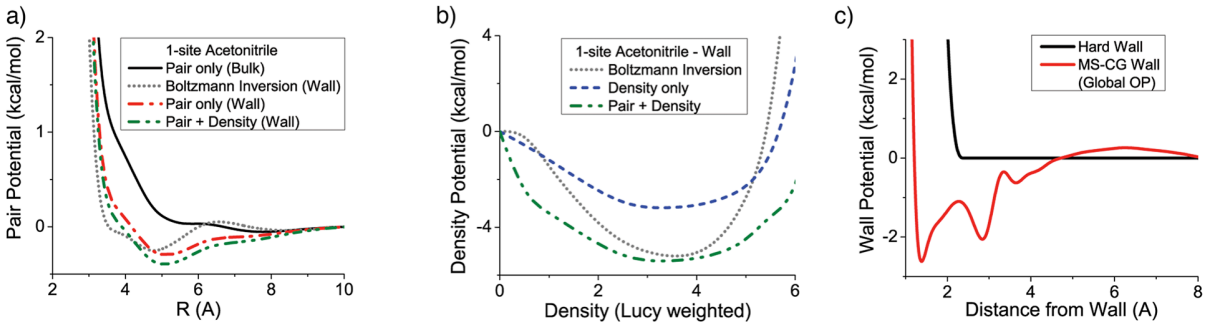

Figure 6–6 - Comparison of the a) pair CG potentials and b) density CG potentials between 1-site acetonitrile CG models employing MS-CG pair, density, and pair + density CG potentials for acetonitrile confined by two hard walls. These potentials are the same for both types of wall interactions. The MS-CG pair CG interaction from bulk acetonitrile and the Boltzmann inverted CG pair and density potentials for acetonitrile confined by hard walls are also shown. c) Comparison of the WCA hard wall and MS-CG (global OP) wall potentials.

For the acetonitrile-wall system, it is also possible to introduce a global OP in the form of MS-CG interactions for the wall. The MS-CG interaction for the wall is compared with the WCA interaction from the FG model in Figure 6-6c. While the WCA interaction was purely repulsive, the MS-CG wall interaction has wells corresponding to the peaks in the FG z-profile seen at 1.5 A and 3 A from the wall.

The behavior of the CG models using the MS-CG wall interaction is shown in Figure 6-7. The RDFs are largely the same between Figure 6-5a and Figure 6-7a, which is expected since none of the other interactions changed. However, the MS-CG wall interaction appears to shift the density distributions for pair CG and pair and density CG models to slightly higher density. This is likely because of the attractive wells in the MS-CG wall interaction. For the pair and density CG model, the MS-CG wall interaction seems to make the z-profile worse by making the features that were already captured by that model in Figure 6-5b over-pronounced in Figure 6-7b. However, the MS-CG wall does make the z-profile for the pair CG model better in Figure 6-7b than it was in Figure 6-5b. Specifically, MS-CG wall makes it possible for pair CG model to have a pronounced feature at 3 A.

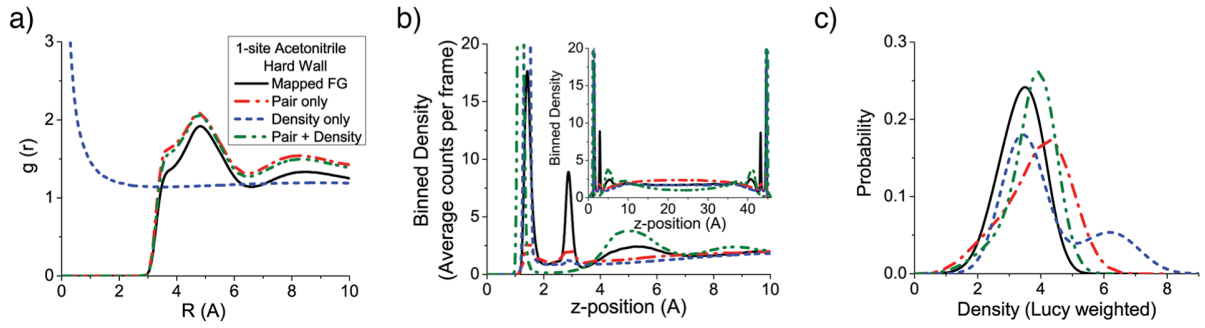

Figure 6–7 - Comparison of the a) radial distribution function (RDF), b) density profile (full profile inset), and c) density distribution (as measured using the Lucy weight function with a cutoff of 10 A) of the mapped FG system against CG models with pair, density, and pair +

density CG interactions for 1-site acetonitrile confined by two hard walls. In this figure, the wall interaction with CG sites was determined using MS-CG.

6.3.4 Timing

Now that we have shown that additive local and global OP interactions improve pair MS-CG models, it worth commenting on the computational cost of these interactions. The speed-up factors for the CG models relative to the FG model are shown for MeOH in Table 6-1. Combining an additive local OP (i.e., density) CG interaction with a pair CG interaction speeds up the evaluation of each frame by 22 times. Given that the properties in Figure 6-1 can still be obtained with a 10 fs time step, the total speed-up factor is 210. While a speed-up of around 730 is possible using only pair CG interactions, the speedup is meaningless since that model does not get the interface profile correct. The computational cost of the global OP (i.e. MS-CG wall) is essentially zero because each MS-CG wall is equivalent to adding one more CG site that has half of the usual interactions. So, significant speed-ups over FG models of more than 200 times are still possible for the systems studied here. Higher speed-ups would, of course, be possible for systems where more degrees of freedom are eliminated in the creation of the CG model.

Table 6-1. Computational Speed-up Relative to FG MeOH

CG Interactions (CG time step)	Speed-up factor
Pair (1 fs)	95
Pair (10 fs)	730
Pair + Density (1 fs)	22
Pair + Density (10 fs)	210

6.4 Discussion

The inclusion of the additive OP-dependent interactions explored in this chapter clearly improved the structural reproduction of the CG models as compared to pair-only CG models. For the MeOH liquid-vapor system, the pair and density (additive local OP) CG model was able reproduce the structural properties extremely well. For the acetonitrile systems (Figures 6-3 through 6-6), the pair and density CG models were able to capture features that the pair only CG model could not. Likewise, adding the MS-CG wall interaction to the pair CG model allowed it to reproduce an extra feature in the acetonitrile-wall system. Overall, these additive OP-dependent interactions extend the ability of CG models to reproduce structural features in interfacial systems.

A direct extension of these results would the application of local density interactions for other systems. One obvious choice would be liquid-liquid interfaces. Using the ability to selectively define groups that interact, one component could interact with density of its own kind, density of the other component, or both. Also, the local density interactions could be used in solvent-free models.

Also, the ability of the CG models to better describe interfacial system through the introduction of OP-dependent interaction makes it all the more important to have CG observables that properly correspond to experiment. While properties such as pressure, compressibility, and interfacial tension are important observables to measure in interfacial systems,[130] it is becoming more important to have a complete set of compatible observables that can be used to simultaneously measure additional CG observables.[91]

Even with just the OPs used in this work, there is possible room for improvement. The two signs of this are the overcorrection of the pair and density CG model for density distribution of

the acetonitrile liquid-vapor interface and the fact that combining the additive density and MS-CG wall interactions did not perform better than either separately. Perhaps, the iterative MS-CG[84-86] methods could be used to refine the combination of pair and OP interactions.

Nonetheless, the successful inclusion of OP-dependent interactions in MS-CG suggests that OP-dependent interactions could be included in other bottom-up methods. The similarity between MS-CG and g-YBG makes it trivial to include in g-YBG.[66, 67] In our results, we showed a Boltzmann inversion of the density distribution. Presumably the distribution of OP-value could inverted to get an OP-dependent potential that is updated with a rule that is similar to that for the pair potential to make an IBI or IMC version of this approach.[63, 179] The already existing density-dependent interaction in relative entropy is good guide of how other OP-dependent interactions could be included there.[26, 133]

However, the extension of this approach to top-down methods is not nearly as clear. To start, there is limited experimental data that can be used to fit interactions. Thus, fitting additional interactions such as OP-dependent interactions would require additional types of experimental data, which is obviously not available for most or even all materials. Moreover, experiments measure ensemble quantities, which makes it even harder to get OP-specific details that would be used to parameterize and validate such a model. What is more, it becomes even more unclear as to the appropriate observable expressions that should be used with such models.[43]

These results have implications for the transferability of CG models. For the MeOH liquid-vapor interface, the pair potential for the pair and density CG model very closely resembled the pair CG potential for bulk MeOH. However, this was not true for any of the other systems. It is possible that when an additive OP-dependent interaction makes it possible for the CG model to reproduce the structural properties of the mapped FG system, it removes the sensitivity of the

pair interaction to changes along that OP; thus, the pair interaction would be transferable across that OP in the naïve sense. This would be an interesting hypothesis to test either through additional coarse-graining or through the calculation of the CG sensitivity to the OP. [41] Even if the sensitivity is not zero, the range where the pair potential is transferable can be further extended using the first order correct from the aforementioned sensitivity calculation.

6.5 Conclusion

In this chapter, we presented a way to include OP-dependent interactions into MS-CG models. We find that the introduction of additive OP CG interactions made it possible to capture features that were not described using the usual pair CG interactions alone. This suggests that the structural reproduction of FG models can be improved through the introduction of these OPs and others. Additionally, we discussed how these OP-dependent interactions could be incorporated in other bottom-up CG methods, but it is less obvious how top-down CG methods could benefit from this approach. It seems that it may be possible to improve the naïve transferability of pair CG interactions with the addition of carefully selected and system-motivated OP-dependent interactions. With the increasing ability of CG models to reproduce structure, it is becoming more important to have good CG observables that correspond to experiment.

Future work could explore other additive OPs and even multiplicative OPs. For example, additive Steinhardt,[93, 170-173] additive liquid crystal[169], and additive spherical harmonic[93] OPs are worth testing in systems where there is clear ordering in the mapped structure of liquids and ionic melts.[197] Additionally, these additive local OPs could make it possible to study nucleation, crystallization, melting, and other such phenomena that occur at liquid-solid interfaces with CG models.[198] On the other hand, multiplicative molecular OPs such as the radius of gyration are a promising choice for improving the behavior of both solvent-free CG protein models[155] and the

solvation structure around CG protein models.[199] Additionally, it would interesting to see if multiplicative local OPs such as local density using a switching weight function could improve on our results for the acetonitrile system. Multiplicative global OPs such as the distance from an electrode could be an interested approach to capturing the variation in effective screening from double layers and other features found in batteries and fuel cells.[200-202]

6.6 Appendix: 1-Site CG Model Of Acetonitrile

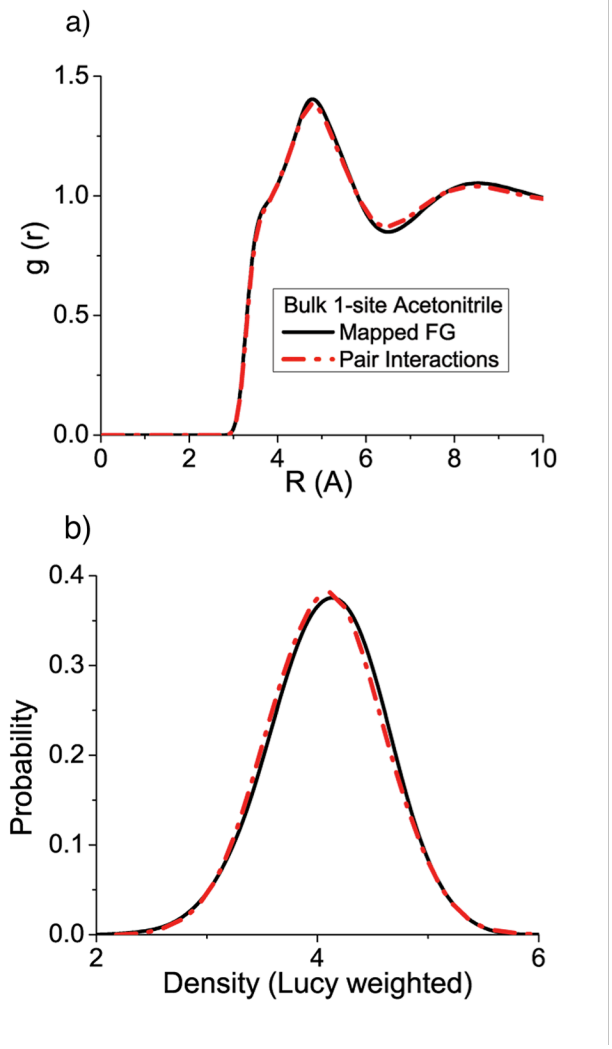

Figure 6–8. Comparison of the a) radial distribution function (RDF) b) density distribution (as measured using the Lucy weight function with a cutoff of 10 A) of the mapped FG system against a CG model with pair interactions for bulk 1-site acetonitrile.

Chapter 7

Order Parameter Dependent Interactions for Polypeptides

7.1 Introduction

Molecular simulations can provide detailed structural information and the distribution of conformations in a way that is generally inaccessible to experiments. However, atomistic simulations are generally limited to timescales on the order of microseconds that only allow the comprehensive study of small and quick-folding proteins. Given that many proteins are large and can fold on the millisecond timescale or longer, more computationally efficient approaches are needed.

Coarse-grained (CG) models aim to reproduce key features of more detailed fine-grained (FG), often atomistic, models using reduced resolution models.[14-18] In order to retain the chemical identity of the residues in polypeptides and proteins, a CG resolution of 1 CG site per residue seems appropriate. At this resolution, simple CG models have had problems reproducing the secondary structure, characteristic conformational states, and self-assembly behavior. As a result, researchers often need to increasing the computational cost of such CG models by using many interaction types,[155] ad-hoc secondary structure constraints[203] to increase the descriptiveness of such models. Even so, CG models can still be limited to a small region of phase space.[204] In fact, Rudzinski and Noid[155] found that they needed 8 separate solvent-free CG models with each employing 22 different interactions to accurately model the two-dimensional (2D) potential of mean force (PMF) of dodeca-alanine in water.

In the previous chapter, we discussed how additional interactions could be incorporated using non-traditional, yet descriptive basis sets in CG model. Specifically, that paper used local density and global distance from an interface as order parameter in interfacial liquid-liquid and liquid-

solid systems. In that case, density was a natural choice to describe the differing behavior of the phases. For polypeptides and proteins, molecular based order parameters (OPs) are natural choices. They have the potential to improve the region of phase space that can be described by a model while simultaneously reducing the number of interactions needed in that model.

In this chapter, we explore the use of molecular order parameters to improve the description of dodeca-alanine in water starting from a minimalist, solvent-free CG model. The rest of this chapter is structured as follows: in Section 2, we review the OP approach and present the OPs used in this paper. In Section 3, we present and discuss our results of using those OP for the solvent-free CG model of dodeca-alanaine; In Section 4, we provide conclusions.

7.2 Theory and Methods

7.2.1 Review of Order Parameter Interactions

In the previous chapter, we discussed how OPs can be divided into three major categories based on the locality of the particles needed to compute them: local OPs that only depend on particles in the neighborhood of interest; molecular OPs that depend on particles within the same molecule regardless of how far apart these particles are; global OPs that depend on a specific reference that is fixed relative to the individual motions of CG particles. For these OPs, they can be used to modify traditional basis sets either additively or multiplicatively. Since this chapter is an initial exploration of molecular OPs, we will focus on additive couplings because of its simplicity.

In order to understand how the molecular OPs can be incorporated into the multiscale coarse-graining (MS-CG)[22, 29-40] methodology, it is necessary to briefly summarize how basis sets are

determined. In MS-CG, the CG interaction potential $U_{CG}\left(\mathbf{R}^N\right)$ is the sum of basis functions $\phi_i\left(\mathbf{R}^N\right)$ multiplied by their coefficients λ_i:

$$U_{CG}\left(\mathbf{R}^N\right)=\sum_{i=1}^{B}\lambda_i\phi_i\left(\mathbf{R}^N\right), \tag{7.1}$$

where $\mathbf{R}^N$ is set the CG positions that are related to the FG positions $\mathbf{r}^n$ through the mapping operator M and i runs through all B position-dependent basis functions. Including additive OPs, the potential becomes

$$U_{CG}\left(\mathbf{R}^N,\mathbf{P}^M\right)=U_{CG,pair}\left(\mathbf{R}^N\right)+U_{CG,OP}\left(\mathbf{P}^M\right)=\sum_{i=1}^{B}\lambda_i\phi_i\left(\mathbf{R}^N\right)+\sum_{j=1}^{C}\mu_j\varphi_j\left(\mathbf{P}^M\right), \tag{7.2}$$

where $\mathbf{P}^M$ is the set of M OPs that can be calculated for the system, φ_j is a basis function that depends only on the order parameters, μ_j is the corresponding undetermined coefficient analogous to λ_i, and j runs through all C OP-dependent basis functions. The corresponding force is

$$F_{CG}\left(\mathbf{R}_k\right)=-\sum_{i=1}^{B}\lambda_i\sum_{l,m}\frac{d\phi_i\left(\mathbf{R}^N\right)}{d\mathbf{R}_{lm}}\hat{\mathbf{R}}_{lm}\left(\delta_{lk}-\delta_{mk}\right)-\sum_{i=1}^{C}\mu_i\sum_{j=1}^{M}\frac{d\varphi_i\left(\mathbf{P}^M\right)}{dP_j}\sum_{l,m}\frac{dP_j}{d\mathbf{R}_{lm}}\hat{\mathbf{R}}_{lm}\left(\delta_{lk}-\delta_{mk}\right), \tag{7.3}$$

where δ is the Kronecker delta and $\hat{\mathbf{R}}_{lm}$ is the unit vector between CG sites l and m. The coefficients are determined by variationally minimizing the objective function χ^2, which is the normalized sum of squared differences between the mapped FG forces and the CG forces.

7.2.2 Order Parameters Used In This Chapter

For our exploration of additive molecular OP-dependent basis functions, we will use the radius of gyration R_g^2, fraction helical content Q_{hel}, 1-4 distances R_{14}, and 1-5 distances R_{15}. While force matching implementation details are provided, interactions for these OPs are also obtained by Bolzmann inverting the corresponding probability distribution.

Radius of Gyration

The radius of gyration R_g^2 of molecule A is the sum squared distances between each site I and the molecule's center of mass R_{COM}:

$$R_{g,A}^2(\mathbf{R}^N) = \frac{1}{N}\sum_{I\in A}\left(R_I - R_{COM,A}\right)^2, \tag{7.4}$$

where $R_{COM} = \sum_{I\in A} m_I R_I / \sum_{I\in A} m_I$ and m_I is the mass of site I. For a system of M molecules, the force for this additive-R_g^2 interaction on CG site k is

$$\begin{aligned} F_{CG}\left(\mathbf{R}_k\right) = &-\sum_{i=1}^{B}\lambda_i \sum_{l,m}\frac{d\phi_i\left(\mathbf{R}^N\right)}{d\mathbf{R}_{lm}}\hat{\mathbf{R}}_{lm}\left(\delta_{lk}-\delta_{mk}\right) \\ &-\sum_{i=1}^{C}\mu_i\sum_{A=1}^{M}\frac{d\varphi_i\left(R_g^{2^M}\right)}{dR_{g,A}^2}\sum_{I\in A}^{N}\frac{2}{N}\left(R_I - R_{COM}\right)\hat{\mathbf{R}}_{I,COM}\delta_{lk} \end{aligned}. \tag{7.5}$$

This force acts to move each site in the molecule either towards or away from that molecule's center of the mass based on the basis set coefficients. As expected, this will directly act to change the molecule's radius of gyration. For solvent-free models, interactions between sites in the molecule are the only way to target such properties since there are no explicit interactions with solvent. This interaction can be evaluated with two passes through a bond or molecule list:

The first pass calculates the center of mass and radius of gyrations while the second pass calculates the forces.

Fraction Helical Content

In addition to molecule size, polypeptides and proteins have distinct secondary structures that are essential to their function. One such structural motif is the alpha helix. Since the polyalanine system studied in this paper folds into an alpha helix, targeting an OP related to this property makes sense. One such OP is the fraction helical content:

$$Q_{hel,A}(\mathbf{R}^N) = \frac{1}{N}\sum_{\substack{I,J\in A \\ I+3=J}}^{N} \exp\left(-\left(R_{IJ}-R_0\right)^2/\sigma^2\right), \tag{7.6}$$

where indices I and J are connected through exactly 3 bonds, $R_0 = 0.5nm$ is the distance for an ideal alpha helix, and $\sigma^2 = 0.02nm^2$ is a scaling factor. The resulting force for the fraction helical content as an additive OP is

$$\begin{aligned} F_{CG}\left(\mathbf{R}_k\right) = &-\sum_{i=1}^{B}\lambda_i\sum_{l,m}\frac{d\phi_i\left(\mathbf{R}^N\right)}{d\mathbf{R}_{lm}}\hat{\mathbf{R}}_{lm}\left(\delta_{lk}-\delta_{mk}\right) \\ &+\sum_{i=1}^{C}\mu_i\sum_{A=1}^{M}\frac{d\varphi_i\left(Q_{hel}^M\right)}{dQ_{hel,A}}Q_{hel,A}\sum_{\substack{I,J\in A \\ I+3=J}}^{N}2\frac{\left(R_{IJ}-R_0\right)}{\sigma^2}\hat{\mathbf{R}}_{IJ}\left(\delta_{lk}-\delta_{IJ}\right) \end{aligned}. \tag{7.7}$$

The fraction helical content is designed to run between 1 for a complete alpha helix to 0 for no alpha-helical content. Like the radius of gyration, this interaction can be evaluated with two passes through the bond list: The first pass calculates the fraction helical content while the second pass calculates the forces. The forces from this OP act on the collective state of all 1-4 pairs in the molecule as oppose to the next OP, which acts on each such pair separately.

Additionally, this OP is a function of how far the 1-4 distance is from the "ideal" 1-4 distance, as opposed to the 1-4 distance directly.

1-4 Distances

As discussed, the 1-4 distance is used to calculate secondary structural content. So, it might be reasonable to target this quantity as a more local and perhaps more physical way to achieve a similar effect as the fraction helical content OP. However, it is possible that the 1-4 distance combined with the usual dihedral interactions could potentially over-constrain the movement of the molecule.

1-5 Distances

An alternative to the 1-4 distance is the 1-5 distance. Like the 1-4 distance, the 1-5 distance is also related to the secondary structure of the model. It is hoped that this OP can be combined with others OPs such as the fraction helical content and more traditional interactions such as dihedral interactions without the potentially issues discussed above.

7.2.3 Simulation and Fitting Details

Molecular dynamics (MD) simulations were performed using all-atom (AA) models of one deca-alanine molecule solvated by 3,824 water molecules in GROMACS.[205-207] The interactions for water were taken from the TIP3P[76] model and the protein interactions were taken from the CHARMM36[208] force field. Short-ranged non-bonded interactions were evaluated using Lennard-Jones (LJ) interactions with a radial cutoff of 12 A. Electrostatics were evaluated using particle mesh Ewald (PME) summation.[209] The AA time step was 2 fs. The thermostat used was the velocity rescale algorithm.[210]

The box size for the AA system was determined by averaging the volume from 2 ns of constant NPT at 1 atm and 300 K using the Parrinello-Rahman barostat.[211] The bulk system

relaxed at that system volume for at constant NVT at 300K for 2 ns. Then, frames were sampled every 1 ps for 600 ns at constant NVT at 300K.

The AA system was mapped to a carbon-alpha CG representation. Bonded pair, angle, and dihedral interactions were determined via Boltzmann Inversion (BI). The so-called "BI" non-bonded and OP interactions were determined by inverting the appropriate one-dimensional distributions. The force-matched (FM) CG interactions were using the MS-CG force matching (FM) code. Pair interactions had a cutoff of 10 A. The FM pair non-bonded interactions were fit using sixth order B-splines with a binwidth of 0.6 A. The FM RG interactions were fit using fourth order B-splines with a binwidth of 5 A^2. The FM fraction helical interactions were fit using fourth order B-splines with a binwidth of 0.05. The 1-4 and 1-5 interactions were fit using fourth order B-splines with a binwidth of 0.5 A.

CG MD simulations were performed using LAMMPS.[71, 73] The starting configuration was created from a mapped atomistic configuration. The CG system was equilibrated for $1x10^6$ steps at constant NVT at 300 K before a $6x10^8$ step production run. Frames were sampled from the production run every 1,000 CG steps.

7.3 Results and Discussion

Atomistic Model

In order to understand what characteristics are important for the CG polyalanine systems to capture, we need to first examine the behavior of the atomistic system. Figure 7-1 shows several 2D PMFs for the atomistic system.

In sub panel a, RMSD is plotted against the radius of gyration. The largest well is in the bottom right corner, which corresponds to an extended, unfolded conformation. The top left corner corresponds to the fully alpha-helical conformation that is the reference for the RMSD

calculation. There are small wells at 0 nm, 0.2 nm, and 0.35 nm RMSD that correspond to folded and partially folded intermediates. In bottom left corner, the conformations are compact, coiled, and unfolded. With this force field, there is not a stable well in that region. Using the OPLS[74, 136] force field, Rudzinski and Noid[155] found that there was a well for the compact, coiled conformation, but not the extended conformation. Differences of this sort are expected given the different force field. Using either force field the qualitative read is the same: There is a narrow well for the folded conformation, one stable unfolded conformation, and one unstable unfolded conformation.

In sub panel b, the fraction helical content is plotted against radius of gyration. The plot is oriented such that the coiled, extended, and folded conformations are in the same place. This y-axis makes it clear that are several distinct wells that correspond to different but discrete numbers of "helical" residues. Rudzinski and Noid[155] also observed this horizontal banding. Otherwise, the features are very similar to those in sub panel a.

In sub panel c, the 1-4 distance is plotted against the radius of gyration. Comparison between this graph and sub panel b should help elucidate any cooperative effects or correlations between the individual contributions to the fraction helical content calculation and the overall total value. In sub panel c, there are two major wells. The well in the bottom left corresponds to "helical" 1-4 distances based on the $R_0 = 0.5nm$ parameter in the fraction helical content calculation and is associated with smaller radius of gyration values just like the fully helical conformation. The other well is more diffuse and associated with larger 1-4 distances and radius of gyration values. This suggests that it corresponds to more unfolded conformations. The barrier between these two predominant wells looks to be about 2 kcal/mol from the top right to the bottom left and about 3 kcal for the reverse. Given the wells are centered at different radius of gyration values, it is likely

that there is cooperative to a certain degree in that as the radius of gyration decreases, the well at low 1-4 distance gets deeper.

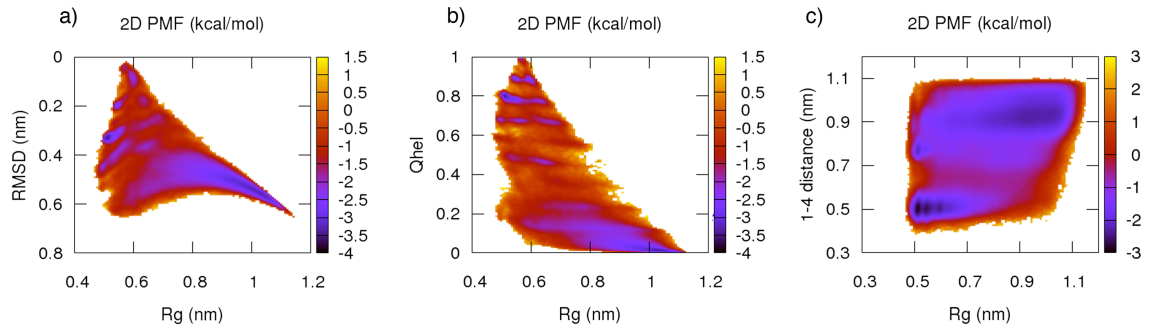

Figure 7–1. 2D PMFs for the atomistic polyalanine system. a) root mean squared deviation (RMSD) from a fully helical conformation versus radius of gyration (RG), b) fraction helical content (Qhel) versus RG, and c) 1-4 distances versus RG.

Basic CG Model

Now, we will look at a minimalistic CG model. The way that this model is different from the atomistic model allows us to identify the deficiencies of this model. This will serve as the baseline that the additive interactions will build upon.

This minimalistic CG model has only 4 interactions. The 2D PMFs for this model are shown in Figure 7-2. In sub panel a, there is a single minimum around 0.5 nm on both the x- and y-axis. While this roughly corresponds to the coiled conformation, there is no minimum in the atomistic model at this location. However, this roughly corresponds to the locations of the deepest wells along each axis. For the x-axis, this corresponds to the folded and partially folded conformations. For the y-axis, this corresponds to the extended conformation. Likewise, the minima in the other panels are all at 0.5 nm on the x-axis corresponding to the folded and partially folded

conformations; on the y-axis, the minima are at values corresponding to the atomistic minima associated with the extended conformation.

One the one hand, it is somewhat expected that Boltzmann inversion would have the correct minimum for each 1D distribution since that is how it is parameterized. Clearly, it did not get the appropriate cross-correlation between the distributions. On the other hand, it is surprising that that it got the correct locations of the minimum for radius of gyration, fraction helical content, 1-4 distance, and RMSD given that those distributions themselves were not used to parameterize the CG model.

The simplified "IS1" model presented by Rudzinksi and Noid[155] in their SI has a similar number of interactions. In top left panel of their Figure S8, the MS-CG model has only 1 minimum center at a radius of gyration of 0.5 nm and a fraction helical content of about 0.3.Their result is consistent with sub panel b of Figure 7-2.

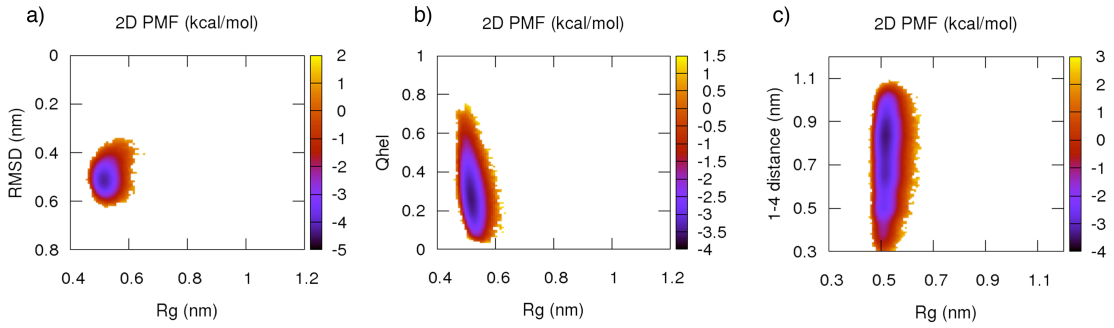

Figure 7–2. 2D PMFs for the CG model using Boltzmann Inverted pair nonbonded, bonded, angular, and dihedral interactions. a) root mean squared deviation (RMSD) from a fully helical conformation versus radius of gyration (RG), b) fraction helical content (Qhel) versus RG, and c) 1-4 distances versus RG.

CG models with a single additive OP interaction

Including the additive OP interactions to the minimalistic CG model should improve its description of the atomistic system. Figure 7-3 shows the 2D PMFs for adding one additive OP interaction with each row corresponding to a different OP. From top to bottom, the OPs are radius of gyration, fraction helical content, 1-5 distance, and 1-4 distance.

The only OP that does not improve the 2D PMFs is the 1-4 interaction. This is likely because the 1-4 and dihedral interactions could over constrain the molecule's configuration relative to either interaction alone. One reason why this happens here is that both interactions are Boltzmann inverted, which means that there the explicit double counting of shared correlations is allowed. If an MS-CG model were created using these interactions, then the double counting would presumably be avoided.

The 1-5 distance and fraction helical content OPs improve the 2D PMFs. They permit exploration of the extended conformation while still maintaining the major well at the coiled conformation. Both of these interactions avoid the problems encountered for the 1-4 interaction because they are different enough. The 1-5 interaction is a different, albeit related measure to the dihedral or 1-4 distance. Since the fraction helical content measure is a sum of transformed 1-4 distances, there is significantly flexibility in the combination of 1-4 distances or dihedrals that can produce a given fraction helical content value. Somewhat surprisingly, the 2D PMFs for the CG model with the 1-5 distance interaction and the one with the fraction helical content interaction look extremely similar. As with the 1-4 interactions, the ability of these Boltzmann inverted OPs to improve the CG description could be limited by double counting and missing cross-correlations.

The radius of gyration OP further improves the 2D PMFs. Specifically, the middle panel (sub panel b) has the same shape but a shallower well than the1-5 distance or fraction helical content OP's PMFs. This shallower well better reproduces the atomistic PMF, which has a broader and shallow well in that general area. It makes sense that the radius of gyration OP would be able to change the CG model's distribution for the radius of gyration so that it better agrees with the atomistic model's distribution. Following that line of reasoning, it is somewhat surprising that the fraction helical content OP does not noticeably improve the CG model's fraction helical content distribution.

It is worth realizing how much was gained by going from 4 interactions (Figure 7-2) to 5 interactions (Figure 7-3). This allowed the CG model to describe an additional conformational state. To get similar looking results, Rudzinski and Noid[155] need to use a MS-CG model with 21 interactions. Specifically, the middle panel of their Figure 5 shows the same general shape and energy differences as sub panels a, d, or g in Figure 7-3.

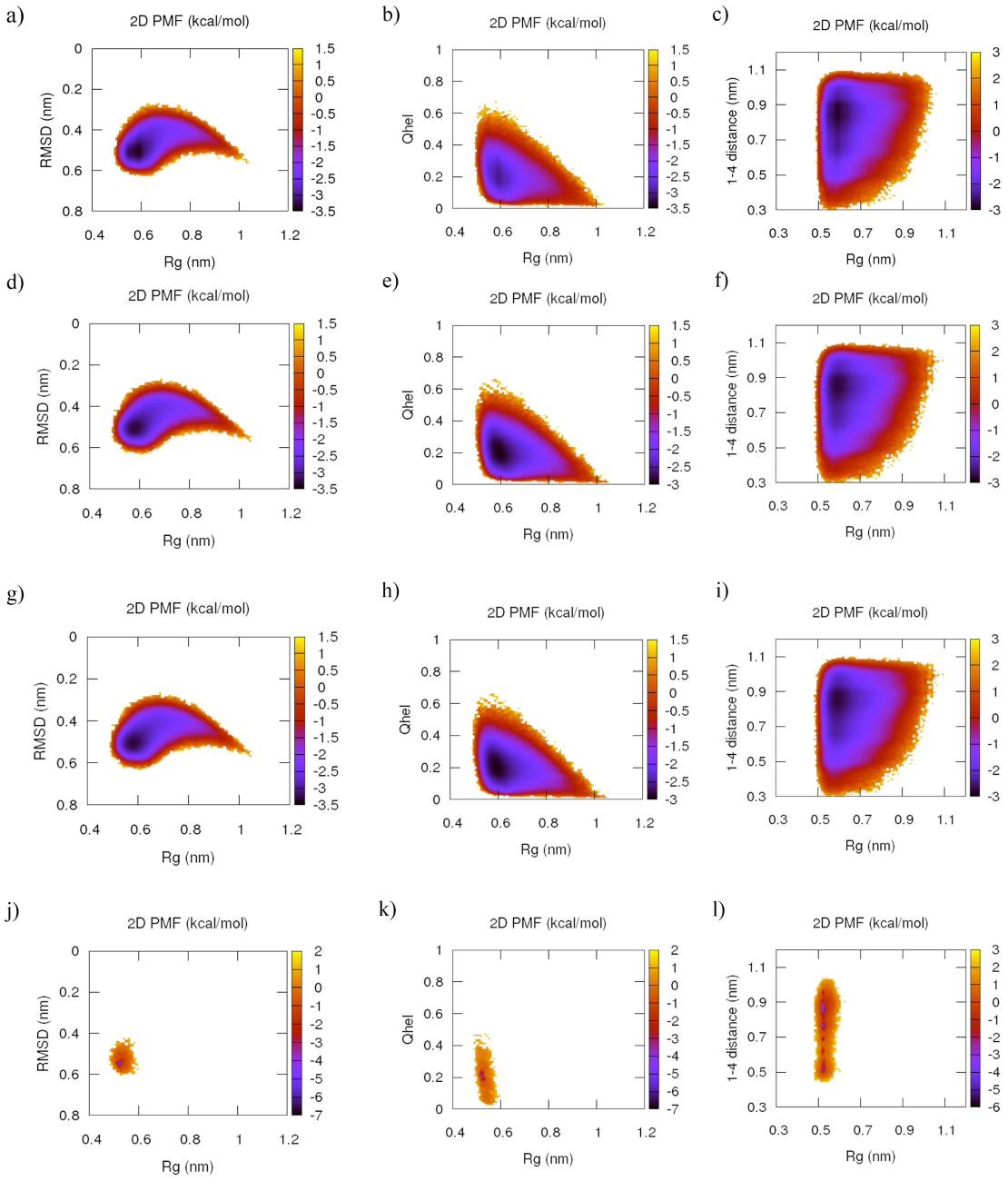

Figure 7–3. 2D PMFs for the CG polyalaine models with a single additive OP interaction. The OP is a-c) radius of gyration (RG), d-f) fraction helical content (Qhel), g-i) 1-5 distances, j-l) 1-4 distances. The 2-D PMFs in the leftmost column are root mean squared deviation (RMSD) from a fully helical conformation versus radius of gyration (RG). The 2-D PMFs in the middle column

are fraction helical content (Qhel) versus RG. The 2-D PMFs in the rightmost column are 1-4 distances versus RG.

CG models with several additive OP interactions

Given the improvement possible using a single OP, it is worth investigating if combinations of additive OPs could further improve the behavior of a CG model. Such combinations are shown in Figure 7-4 with rows corresponding to 1) radius of gyration and fraction helical content, 2) radius of gyration and 1-5 distances, 3) fraction helical content and 1-5 distances, 4) radius of gyration, fraction helical content, and 1-5 distances. All of these combinations of OPs look essentially the same. The only pronounced differences is that sub panels b and k have slightly shallower wells than sub panels e and h. This is likely the effect of the radius of gyration OP; however, it is surprising that sub panel e does not show this same effect. Nonetheless, it is likely that that double counting and missing cross-correlations discussed above prevent the CG model from better describing the atomistic behavior.

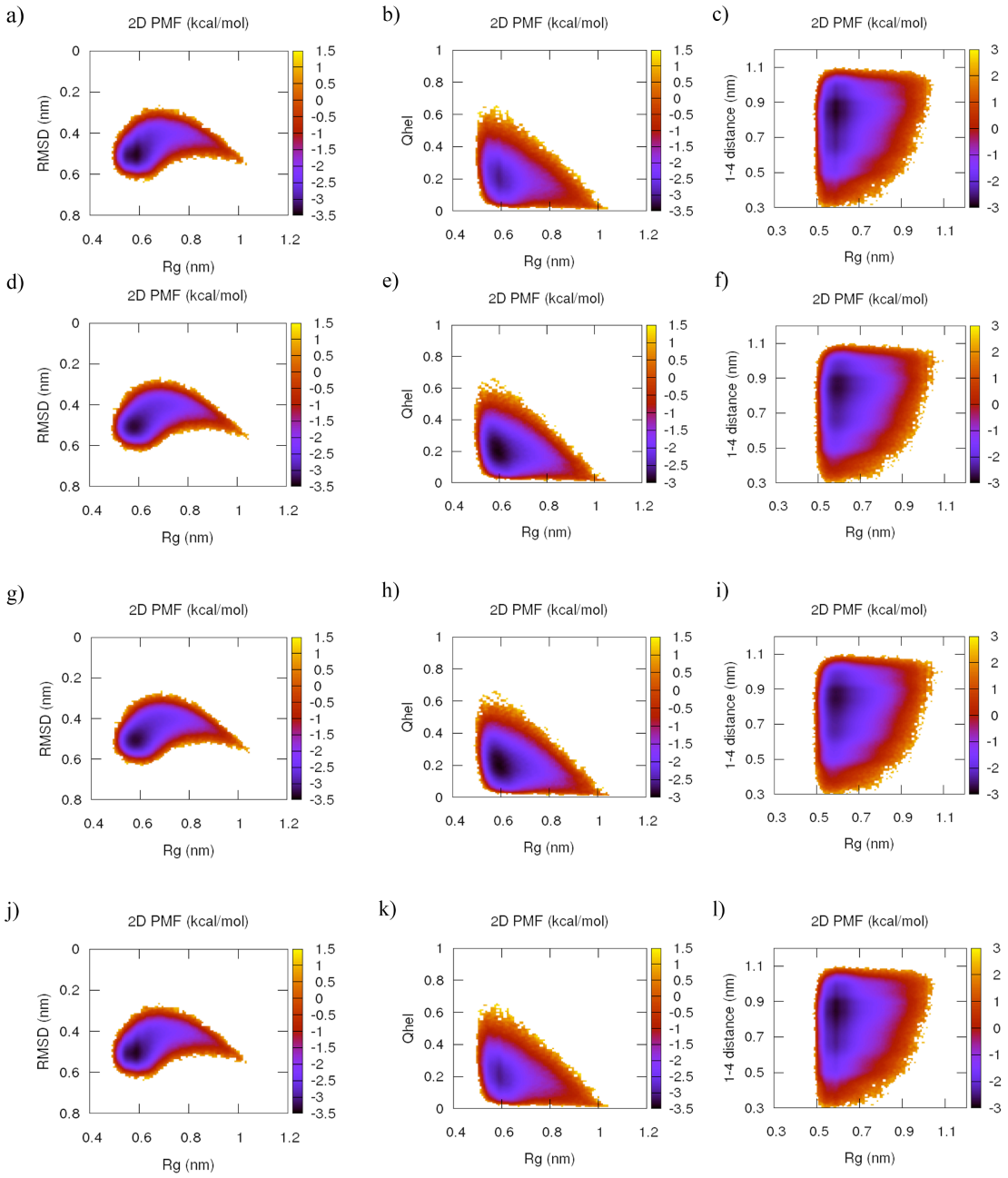

Figure 7–4. 2D PMFs for the CG polyalaine models with several additive OP interactions. The OPs are a-c) radius of gyration (RG) and fraction helical content (Qhel), d-f) RG and 1-5 distances, g-i) Qhel and 1-5 distances, j-l) RG, Qhel, and 1-5 distances. The 2-D PMFs in the leftmost column are root mean squared deviation (RMSD) from a fully helical conformation

versus radius of gyration (RG). The 2-D PMFs in the middle column are fraction helical content (Qhel) versus RG. The 2-D PMFs in the rightmost column are 1-4 distances versus RG.

7.4 Conclusion

In this chapter, the ability of molecular, additive order parameters to improve the fidelity of simple CG models was investigated. When using a single order parameter, the radius of gyration improved the CG model most while the fraction helical content also improved the CG model, but to a lesser extent. Using only 5 CG interactions, we were able to obtain a similar fidelity description as a CG model with 21 interactions reported in the literature.[155]

There is obvious room for further improvement of CG model fidelity. The improvement that could be achieved using these order parameters appears to be limited by the cross-correlation neglected by direct Boltzmann inversion. Using the same method, these interactions could be iteratively improved. Alternatively, the cross-correlations would be taken into account if MS-CG was used to fit the CG interactions. In any case, this further exploration of additive order parameter CG interactions demonstrates the broad applicability of such interactions to a variety of systems.

Chapter 8

Conclusion and Future Directions

8.1 Introduction

The research presented in this thesis is aimed at understanding how existing CG methodologies can be improved in terms of their transferability, fidelity, and correspondence in order for CG models to fulfill their potential as aids in the development of new materials. In Chapter 3, we demonstrated how our CG sensitivity measure could be used to calculate first order estimates for the transfer of CG potentials. In Chapter 4, we discussed the conditions necessary for CG observables to correspond with FG and experimental observables. In Chapter 5, we demonstrated how this correspondence could be established numerically. In Chapters 6 and 7, we demonstrated how the use of additional, order-parameter-dependent basis sets improves the fidelity of CG models. Based upon the work in this thesis, there are a number of potential future directions and implications for research related to coarse-graining.

8.2 Future Directions

Most directly, the methods in this thesis could be applied more generally. For MS-CODE, this means creating observable decompositions and applying the method to obtain more complete sets of compatible CG observables. As mentioned before, one high-value target would be excess entropy in order to better understand the associated dynamical speed-up and CG dynamics. With a more complete understanding of CG observable correspondence, aspects of CG thermodynamics could be investigated. Likewise for order parameters, this means trying other possible order parameters that make sense for novel systems as well as investigating the applicability of the specific order parameters used in Chapters 6 and 7 on other systems or even

in CG methods other than MS-CG. For sensitivity, this means developing extensions that allow one to calculate the sensitivity of other CG properties such as structural observables where the naïve CG expression is appropriate as well as using the method on different CG models to learn about how different mappings and types of interactions effect the calculated sensitivity.

Additionally, the accuracy and fidelity of CG models using the methods in this thesis could be further refined. Several iterative methods for MS-CG and related methods have been developed.[84-86] Such methods could be directly used with order parameter basis sets to further improve CG fidelity. Likewise, CG observable expressions from MS-CODE could be iteratively refined.

One big potential impact of MS-CODE would be the development of better barostats. While the work of Das and Anderson[31] and Dunn and Noid[97] serves as a better barostats than the naïve CG choice, its fluctuations are at the granularity of system volume. A combination of MS-CODE with such work could produce barostats that would improve CG NPT, mixed resolution,[120, 121] and adaptive resolution[122-124] simulations.

An area where my work on CG observables and model representation come together is mixed and adaptive resolution modeling.[120, 121] Since observable expressions change with model resolution, the pressure expression for different resolutions in the system should have different observable expressions for pressure. This is one cause of the apparent pressure-induced drift.[122] Additionally, the cross interactions (i.e., AA – CG interactions) are not properly calibrated in most mixed resolution models. While an understanding of this is starting to develop,[212, 213] the steps needed to correct this issue have not been taken. In particular, one could determine these cross interactions using the mixed resolution approach described by Izvekov et al.;[120] however, it is not necessarily clear that the best option for combining the atomistic, CG, and cross

interactions is a simple linear combination. Using CG sensitivity and MS-CODE, one could calculate a series of intermediate resolution interactions, which could improve the behavior of such models even more.

8.3 Remaining Challenges

Even with the contributions to CG modeling presented in this thesis, a number of big challenges still remain for computational materials design. While this work was concerned with the connection between CG and FG models, the accuracy the usually atomistic FG models needs to be improved for the structures and properties reproduced from CG models to have the same reliability as actual experiments. Additionally, much work is needed on understanding how dynamics and timescale are changed through CG. This is a surprising complex issue and cannot be solved with a simple global scaling of all apparent CG times. Also, the ability to model reactions at the CG resolution is necessary for the study of many complex materials. Finally, an understanding and implementation of how to correctly handle dynamic resolution changes and dynamic mappings will be helpful in many ways. For example, CG reactions can be handled with increased detail while still achieving the same goal of increased computational efficiency as current adaptive and mixed resolution models.

8.4 Final Thoughts

It is hoped that the work in this thesis will pave the way for future developments, which will make computational materials design more powerful. CG sensitivity highlighted a new direction for increasing transferability of CG models, where fidelity can be maintained without the need for additional atomistic simulation. The work on CG observables has highlighted the importance

of observable compatibility and model correspondence. The work on order parameters has also highlighted a new direction for increasing the fidelity of reduced resolution models. Perhaps the ability to obtain high-fidelity, high-resolution CG models with proper correspondence will make it possible to construct similarly high-fidelity but reduced resolution, more coarse CG models.

Bibliography

[1] L. Huynh, C. Neale, R. Pomes, and C. Allen, "Computational approaches to the rational design of nanoemulsions, polymeric micelles, and dendrimers for drug delivery," Nanomed. Nanotechnol. Biol. Med. **8**, 20 (2012).

[2] T. J. Marrone, J. M. Briggs, and J. A. McCammon, "Structure-based drug design: Computatational advances," Annu. Rev. Pharm. Tox. **27**, 71 (1997).

[3] J. Greer, J. Erickson, J. J. Baldwin, and M. D. Varney, "Application of the three-dimensional structures of protein target molecules in structure-based drug design," J. Med. Chem. **37**, 1035 (1994).

[4] R. Capdeville, E. Buchdunger, J. Zimmerman, and A. Matter, "Glivec (STI571, imatinib), a rationally developed, targeted anticancer drug," Nat. Rev. Drug Discov. **1**, 493 (2002).

[5] *Rational Drug Design* (Springer, New York, 1999).

[6] Y. Li, S. Tang, B. C. Abberton, M. Kröger, C. Burkhart, B. Jiang, G. J. Papakonstantopoulos, M. Poldneff, and W. K. Liu, "A predictive multiscale computational framework for viscoelastic properties of linear polymers," Polymer **53**, 5935 (2012).

[7] H. Xu, D. A. Dikin, C. Burkhart, and W. Chen, "Descriptor-based methodology for statistical characterization and 3D reconstruction of microstructural materials," Comput. Mat. Sci. **85**, 206 (2014).

[8] C. D. Wood, A. Ajdari, C. W. Burkhart, K. W. Putz, and L. C. Brinson, "Understanding competing mechanisms for glass transition changes in filled elastomers," Compos. Sci. Tech. **127**, 88 (2016).

[9] *Combinatorial Chemistry: A Practical Approach* (Oxford UP, Oxford, 2000).

[10] E. Wimmer, "Computational materials design and processing: Perspectives for atomistic approaches," Mat. Sci. Eng. D **37**, 72 (1996).

[11] S. Curtarolo, G. L. W. Hart, M. B. Nardelli, N. Mingo, S. Sanvito, and O. Levy, "The high-throughpu highway to computational materials design," Nature Materials **12**, 191 (2013).

[12] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. A. Persson, "Commentary: The materials project: A materials genome approach to accelerating materials innovation," APL Materials **1**, 011002 (2013).

[13] J. J. de Pablo, B. Jones, C. L. Kovacs, V. Ozolins, and A. P. Ramirez, "The materials genome initiative, the interplay of experiment, theory and computation," Curr. Opin. Sol. State Mat. Sci. **18**, 99 (2014).

[14] E. Brini, E. A. Algaer, P. Ganguly, C. Li, F. Rodriguez-Ropero, and N. F. A. van der Vegt, "Systematic coarse-graining methods for soft matter simulations – A review," Soft Matter **9**, 2108 (2013).

[15] F. Müller-Plathe, "Coarse-graining in polymer simulation: From the atomistic to the mesoscopic scale and back," ChemPhysChem **3**, 754 (2002).

[16] W. G. Noid, "Perspective: Coarse-grained models for biomolecular systems," J. Chem. Phys. **139**, 090901 (2013).

[17] S. Riniker, J. R. Allison, and W. F. van Gunsteren, "On developing coarse-grained models for biomolecular simulation: A review," Phys. Chem. Chem. Phys. **14**, 12423 (2012).

[18] G. A. Voth, *Coarse-graining of Condensed Phase and Biomolecular Systems* (CRC Press, Boca Raton, 2009).

[19] D. Chandler, *Introduction to Modern Statistical Thermodynamics* (Oxford UP, 1987).

[20] M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids* (Oxford UP, Oxford, 1989).

[21] D. Frenkel and B. Smit, *Understanding Molecular Simulation: From Algorithms to Applications* (Academic Press, Orlando, 2001), 2nd edn., Computational Science Series, vol 1.

[22] W. G. Noid, J.-W. Chu, G. S. Ayton, V. Krishna, S. Izvekov, G. A. Voth, A. Das, and H. C. Andersen, "The multiscale coarse-graining method. I. A rigorous bridge between atomistic and coarse-grained models," J. Chem. Phys. **128**, 244114 (2008).

[23] W. Tschöp, K. Kremer, J. Batoulis, T. Bürger, and O. Hahn, "Simulation of polymer melts. I. Coarse-graining procedure for polycarbonates," Acta Polymerica **49**, 61 (1998).

[24] D. Reith, M. Putz, and F. Müller-Plathe, "Deriving effective mesoscale potentials from atomistic simulations," J. Comput. Chem. **24**, 1624 (2003).

[25] A. Chaimovich and M. S. Shell, "Coarse-graining errors and numerical optimization using a relative entropy framework," J. Chem. Phys. **134**, 094112 (2011).

[26] M. S. Shell, "The relative entropy is fundamental to multiscale and inverse thermodynamic problems," J. Chem. Phys. **129**, 144108 (2008).

[27] M. S. Shell, "Systematic coarse-graining of potential energy landscapes and dynamics in liquids," J. Chem. Phys. **137**, 084503 (2012).

[28] J. F. Rudzinski and W. G. Noid, "Coarse-graining entropy, forces, and structures," J. Chem. Phys. **135**, 214101 (2011).

[29] Z. Cao and G. A. Voth, "The multiscale coarse-graining method. XI. Accurate interactions based on the centers of charge of coarse-grained sites," J. Chem. Phys. **143**, 243116 (2015).

[30] A. Das and H. C. Andersen, "The multiscale coarse-graining method. III. A test of pairwise additivity of the coarse-grained potential and of new basis functions for the variational calculation," J. Chem. Phys. **131**, 034102 (2009).

[31] A. Das and H. C. Andersen, "The multiscale coarse-graining method. V. Isothermal-isobaric ensemble," J. Chem. Phys. **132**, 164106 (2010).

[32] A. Das and H. C. Andersen, "The multiscale coarse-graining method. IX. A general method for construction of three body coarse-grained force fields," J. Chem. Phys. **136**, 194114 (2012).

[33] A. Das and H. C. Andersen, "The multiscale coarse-graining method. VIII. Multiresolution hierarchical basis functions and basis function selection in the construction of coarse-grained force fields," J. Chem. Phys. **136**, 194113 (2012).

[34] A. Das, L. Lu, H. C. Andersen, and G. A. Voth, "The multiscale coarse-graining method. X. Improved algorithms for constructing coarse-grained potentials for molecular systems," J. Chem. Phys. **136**, 194115 (2012).

[35] S. Izvekov and G. A. Voth, "Multiscale coarse graining of liquid-state systems," J. Chem. Phys. **123**, 134105 (2005).

[36] S. Izvekov and G. A. Voth, "A multiscale coarse-graining method for biomolecular systems," J. Phys. Chem. B **109**, 2469 (2005).

[37] V. Krishna, W. G. Noid, and G. A. Voth, "The multiscale coarse-graining method. IV. Transferring coarse-grained potentials between temperatures," J. Chem. Phys. **131**, 024103 (2009).

[38] L. Larini, L. Lu, and G. A. Voth, "The multiscale coarse-graining method. VI. Implementation of three-body coarse-grained potentials," J. Chem. Phys. **132**, 164107 (2010).

[39] L. Lu and G. A. Voth, "The multiscale coarse-graining method. VII. Free energy decomposition of coarse-grained effective potentials," J. Chem. Phys. **134**, 224107 (2011).

[40] W. G. Noid, P. Liu, Y. Wang, J. W. Chu, G. S. Ayton, S. Izvekov, H. C. Andersen, and G. A. Voth, "The multiscale coarse-graining method. II. Numerical implementation for coarse-grained molecular models," J. Chem. Phys. **128**, 244115 (2008).

[41] J. W. Wagner, J. F. Dama, and G. A. Voth, "Predicting the sensitivity of multiscale coarse-grained models to their underlying fine-grained model parameters," J. Chem. Theory Comput. **11**, 3547 (2015).

[42] M. E. Johnson, T. Head-Gordon, and A. A. Louis, "Representability problems for coarse-grained water potentials," J. Chem. Phys. **126**, 144509 (2007).

[43] A. A. Louis, "Beware of density dependent pair potentials," J. Phys.: Condens. Matter **14**, 9187 (2002).

[44] C. F. Wong, "Systematic sensitivity analyses in free energy perturbation calculations," J. Am. Chem. Soc. **113**, 3208 (1991).

[45] C. F. Wong and H. Rabitz, "Sensitivity analysis and principal component analysis in free energy calculations," J. Phys. Chem. **95**, 9628 (1991).

[46] G. J. Rocklin, D. L. Mobley, and K. A. Dill, "Calculating the sensitivity and robustness of binding free energy calculations to force field parameters," J. Chem. Theory Comput. **9**, 3072 (2013).

[47] H. Paliwal and M. R. Shirts, "A benchmark test set for alchemical free energy transformations and its use to quantify error in common free energy methods," J. Chem. Theory Comput. **7**, 4115 (2011).

[48] M. R. Shirts and J. D. Chodera, "Statistically optimal analysis of samples from multiple equilibrium states," J. Chem. Phys. **129**, 124105 (2008).

[49] M. R. Shirts and V. S. Pande, "Comparison of efficiency and bias of free energies computed by exponential averaging, the Bennett acceptance ratio, and thermodynamic integration," J. Chem. Phys. **122**, 144107 (2005).

[50] H. Paliwal and M. R. Shirts, "Using multistate reweighting to rapidly and efficiently explore molecular simulation parameters space for nonbonded interactions," J. Chem. Theory Comput. **9**, 4700 (2013).

[51] S. H. Fleischman and C. L. Brooks, "Thermodynamics of aqueous solvation: Solution properties of alcohols and alkanes," J. Chem. Phys. **87**, 3029 (1987).

[52] C. F. Wong, T. Thacher, and H. Rabitz, in *Reviews in Computational Chemistry*, edited by K. B. Lipkowitz, and D. B. Boyd (Wiley & Sons, New York, 1998), pp. 281.

[53] P. Cieplak, D. A. Pearlman, and P. A. Kollman, "Walking on the free energy hypersurface of the 18‐crown‐6 ion system using free energy derivatives," J. Chem. Phys. **101**, 627 (1994).

[54] S. B. Zhu and C. F. Wong, "Sensitivity analysis of water thermodynamics," J. Chem. Phys. **98**, 8892 (1993).

[55] L. P. Wang, T. Head-Gordon, J. W. Ponder, P. Ren, J. D. Chodera, P. K. Eastman, T. J. Martinez, and V. S. Pande, "Systematic improvement of a classical molecular model of water," J. Phys. Chem. B **117**, (2013).

[56] G. D'Adamo, A. Pelissetto, and C. Pierleoni, "Predicting the thermodynamics by using state-dependent interactions," J. Chem. Phys. **138**, 234107 (2013).

[57] H.-J. Qian, P. Carbone, X. Chen, H. A. Karimi-Varzaneh, C. C. Liew, and F. Müller-Plathe, "Temperature-transferable coarse-grained potentials for ethylbenzene, polystyrene, and their mixtures," Macromolecules **41**, 9919 (2008).

[58] E. C. Allen and G. C. Rutledge, "A novel algorithm for creating coarse-grained, density dependent implicit solvent models," J. Chem. Phys. **128**, 154115 (2008).

[59] E. C. Allen and G. C. Rutledge, "Coarse-grained, density dependent implicit solvent model reliably reproduces behavior of a model surfactant system," J. Chem. Phys. **130**, 204903 (2009).

[60] S. Izvekov, P. W. Chung, and B. M. Rice, "The multiscale coarse-graining method: Assessing its accuracy and introducing density dependent coarse-grain potentials," J. Chem. Phys. **133**, 064109 (2010).

[61] J. W. Shen, C. Li, N. F. van der Vegt, and C. Peter, "Transferability of coarse grained potentials: Implicit solvent models for hydrated ions," J. Chem. Theory Comput. **7**, 1916 (2011).

[62] T. Murtola, E. Falck, M. Karttunen, and I. Vattulainen, "Coarse-grained model for phospholipid/cholesterol bilayer employing inverse Monte Carlo with thermodynamic constraints," J. Chem. Phys. **126**, 075101 (2007).

[63] A. P. Lyubartsev and A. Laaksonen, "Calculation of effective interaction potentials from radial distribution functions: A reverse Monte Carlo approach," Phys. Rev. E **52**, 3730 (1995).

[64] L. Lu and G. A. Voth, in *Adv. Chem. Phys.*, edited by S. A. Rice, and A. R. Dinner (Wiley-Interscience, New York, 2012).

[65] W. G. Noid, G. S. Ayton, S. Izvekov, and G. A. Voth, in *Coarse-Graining of Condensed Phase and Biomolecular Systems*, edited by G. A. Voth (CRC Press, New York, 2008), pp. 21.

[66] J. W. Mullinax and W. G. Noid, "Generalized Yvon-Born-Green theory for molecular systems," Phys. Rev. Lett. **103**, 198104 (2009).

[67] J. W. Mullinax and W. G. Noid, "A generalized-Yvon−Born−Green theory for determining coarse-grained interaction potentials," J. Phys. Chem. C **114**, 5661 (2010).

[68] I. Bilionis and N. Zabaras, "A stochastic optimization approach to coarse-graining using a relative-entropy framework," J. Chem. Phys. **138**, 044313 (2013).

[69] F. Ercolessi and J. B. Adams, "Interatomic potentials from first-principles calculations: The force-matching method," Europhys. Lett. **26**, 583 (1994).

[70] L. Lu, S. Izvekov, A. Das, H. C. Andersen, and G. A. Voth, "Efficient, regularized, and scalable algorithms for multiscale coarse-graining," J. Chem. Theory Comput. **6**, 954 (2010).

[71] W. M. Brown, A. Kohlmeyer, S. J. Plimpton, and A. N. Tharrington, "Implementing molecular dynamics on hybrid high performance computers – Particle–particle particle-mesh," Comput. Phys. Commun. **183**, 449 (2012).

[72] W. M. Brown, P. Wang, S. J. Plimpton, and A. N. Tharrington, "Implementing molecular dynamics on hybrid high performance computers – short range forces," Comput. Phys. Commun. **182**, 898 (2011).

[73] S. Plimpton, "Fast parallel algorithms for short-range molecular dynamics," J. Comput. Phys. **117**, 1 (1995).

[74] W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives, "Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids," J. Am. Chem. Soc. **118**, 11225 (1996).

[75] I. S. Joung and T. E. Cheatham, 3rd, "Molecular dynamics simulations of the dynamic and energetic properties of alkali and halide ions using water-model-specific ion parameters," J. Phys. Chem. B **113**, 13279 (2009).

[76] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, "Comparison of simple potential functions for simulating liquid water," J. Chem. Phys. **79**, 926 (1983).

[77] Z. Cao, J. F. Dama, L. Lu, and G. A. Voth, "Solvent free ionic solution models from multiscale coarse-graining," J. Chem. Theory Comput. **9**, 172 (2013).

[78] I. F. Thorpe, D. P. Goldenberg, and G. A. Voth, "Exploration of transferability in multiscale coarse-grained peptide models," J. Phys. Chem. B **115**, 11911 (2011).

[79] C. D. Christ and W. F. van Gunsteren, "Enveloping distribution sampling: A method to calculate free energy differences from a single simulation," J. Chem. Phys. **126**, 184110 (2007).

[80] B. Efron, "Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods," Biometrika **68**, 589 (1981).

[81] H. L. Jones, "Jackknife estimation of functions of stratum means," Biometrika **61**, 343 (1974).

[82] A. C. Cameron, J. B. Gelbach, and D. L. Miller, "Bootstrap-based improvements for inference with clustered errors," Review of Economics and Statistics **90**, 414 (2008).

[83] L. P. Wang, J. Chen, and T. Van Voorhis, "Systematic parametrization of polarizable force fields from quantum chemistry data," J. Chem. Theory Comput. **9**, 452 (2013).

[84] H. M. Cho and J. W. Chu, "Inversion of radial distribution functions to pair forces by solving the Yvon-Born-Green equation iteratively," J. Chem. Phys. **131**, 134107 (2009).

[85] L. Lu, J. F. Dama, and G. A. Voth, "Fitting coarse-grained distribution functions through an iterative force-matching method," J. Chem. Phys. **139**, 121906 (2013).

[86] J. F. Rudzinski and W. G. Noid, "Investigation of coarse-grained mappings via an iterative generalized Yvon-Born-Green method," J. Phys. Chem. B **118**, 8295 (2014).

[87] V. Molinero and E. B. Moore, "Water modeled as an intermediate element between carbon and silicon," J. Phys. Chem. B **113**, 4008 (2009).

[88] C. A. Angell, R. D. Bressel, M. Hemmati, E. J. Sare, and J. C. Tucker, "Water and its anomalies in perspective: Tetrahedral liquids with and without liquid–liquid phase transitions," Physical Chemistry Chemical Physics **2**, 1559 (2000).

[89] M. Singh, D. Dhabal, A. H. Nguyen, V. Molinero, and C. Chakravarty, "Triplet correlations dominate the transition from simple to tetrahedral liquids," Phys. Rev. Lett. **112**, 147801 (2014).

[90] J. F. Dama, A. V. Sinitskiy, M. McCullagh, J. Weare, B. Roux, A. R. Dinner, and G. A. Voth, "The theory of ultra-coarse-graining. 1. General principles," J. Chem. Theory Comput. **9**, 2466 (2013).

[91] J. W. Wagner, J. F. Dama, A. E. Durumeric, and G. A. Voth, "On the representability problem and the physical meaning of coarse-grained models," J. Chem. Phys. **145**, 044108 (2016).

[92] S. P. Carmichael and M. S. Shell, "A new multiscale algorithm and its application to coarse-grained peptide models for self-assembly," J. Phys. Chem. B **116**, 8383 (2012).

[93] W. Lechner and C. Dellago, "Accurate determination of crystal structures based on averaged local bond order parameters," J. Chem. Phys. **129**, 114707 (2008).

[94] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman, "A second generation force field for the simulation of proteins, nucleic acids, and organic molecules," J. Am. Chem. Soc. **117**, 5179 (1995).

[95] A. D. J. MacKerell, D. Bashford, M. Bellott, R. L. Bunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. I. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wlórkiewicz-Kuczera, D. Yin, and M. Karplus, "All-atom empirical potential for molecular modeling and dynamics studies of proteins," J. Phys. Chem. B **102**, 3586 (1998).

[96] F. H. Stillinger, H. Sakai, and S. Torquato, "Statistical mechanical models with effective potentials: Definitions, applications, and thermodynamic consequences," J. Chem. Phys. **117**, 288 (2002).

[97] N. J. H. Dunn and W. G. Noid, "Bottom-up coarse-grained models that accurately describe the structure, pressure, and compressibility of molecular liquids," J. Chem. Phys. **143**, 243148 (2015).

[98] P. Ascarelli and R. J. Harrison, "Density-dependent potentials and the hard-sphere model for liquid metals," Phys. Rev. Lett. **22**, 385 (1969).

[99] C. Caccamo and G. Pellicane, "Microscopic theories of model macromolecular fluids and fullerenes: The role of thermodynamic consistency," J. Chem. Phys. **117**, 5072 (2002).

[100] C. F. Tejero and E. Lomba, "Density-dependent interactions and thermodynamic consistency in integral equation theories," Mol. Phys. **107**, 349 (2009).

[101] R. L. Henderson, "A uniqueness theorem for fluid pair correlation functions," Phys. Lett. A **49**, 197 (1974).

[102] E. C. Allen and G. C. Rutledge, "Evaluating the transferability of coarse-grained, density-dependent implicit solvent models to mixtures and chains," J. Chem. Phys. **130**, 034904 (2009).

[103] B. Smith, T. Hauschild, and J. M. Prausnitz, "Effect of a density-dependent potential on the phase behaviour of fluids," Mol. Phys. **77**, 1021 (1992).

[104] D. H. de Jong, G. Singh, W. F. D. Bennett, C. Arnarez, T. A. Wassenaar, L. V. Schäfer, X. Periole, D. P. Tieleman, and S. J. Marrink, "Improved parameters for the Martini coarse-grained protein force field," J. Chem. Theory Comput. **9**, 687 (2013).

[105] S. J. Marrink, A. H. de Vries, and A. E. Mark, "Coarse grained model for semiquantitative lipid simulations," J. Phys. Chem. B **108**, 750 (2004).

[106] S. J. Marrink, H. J. Risselada, S. Yefimov, D. P. Tieleman, and A. H. de Vries, "The Martini force field: Coarse grained model for biomolecular simulations," J. Phys. Chem. B **111**, 7812 (2007).

[107] S. J. Marrink and D. P. Tieleman, "Perspective on the Martini model," Chem. Soc. Rev. **42**, 6801 (2013).

[108] T. E. Colla, A. P. dos Santos, and L. Y, "Equation of state of charged colloidal suspensions and its dependence on the thermodynamic route," J. Chem. Phys. **136**, 194103 (2012).

[109] R. L. C. Akkermans and W. J. Briels, "A structure-based coarse-grained model for polymer melts," J. Chem. Phys. **114**, 1020 (2001).

[110] M. C. Villet and G. H. Fredrickson, "Numerical coarse-graining of fluid field theories," J. Chem. Phys. **132**, 034109 (2010).

[111] T. T. Foley, M. S. Shell, and W. G. Noid, "The impact of resolution upon entropy and information in coarse-grained models," J. Chem. Phys. **143**, 243104 (2015).

[112] M. Rubinstein and R. Colby, *Polymer Physics* (Oxford UP, Oxford, 2003).

[113] M. Doi and S. Edwards, *The Theory of Polymer Dynamics* (Oxford UP, Oxford, 1986).

[114] L. Belloni, "Colloidal interactions," J. Phys.: Condens. Matter **12**, R549 (2000).

[115] R. van Roij, M. Dijkstra, and J.-P. Hansen, "Phase diagram of charge-stabilized colloidal suspensions: Van der Waals instability without attractive forces," Phys. Rev. E **59**, 2010 (1999).

[116] K. S. Pitzer, "Electrolyte theory - Improvements since Debye and Hückel," Acc. Chem. Res. **10**, 371 (1977).

[117] E. Trizac and Y. Levin, "Renormalized jellium model for charge-stabilized colloidal suspensions," Phys. Rev. E **69**, 031403 (2004).

[118] D. Y. C. Chan, P. Linse, and S. N. Petris, "Phase separation in deionized colloidal systems: Extended Debye-Hückel theory," Langmuir **17**, 4202 (2001).

[119] M. Ndao, J. Devémy, A. Ghoufi, and P. Malfreyt, "Coarse-graining the liquid-liquid Interfaces with the Martini force field: How is the interfacial tension reproduced?," J. Chem. Theory Comput. **11**, 3818 (2015).

[120] S. Izvekov and G. A. Voth, "Mixed resolution modeling of interactions in condensed-phase systems," J. Chem. Theory Comput. **5**, 3232 (2009).

[121] M. Praprotnik, L. Delle Site, and K. Kremer, "Adaptive resolution molecular-dynamics simulation: Changing the degrees of freedom on the fly," J. Chem. Phys. **123**, 224106 (2005).

[122] H. Wang, C. Schütte, and L. Delle Site, "Adaptive resolution simulation (AdResS): A smooth thermodynamic and structural transition from atomistic to coarse grained resolution and vice versa in a grand canonical fashion," J. Chem. Theory Comput. **8**, 2878 (2012).

[123] J. Zavadlav, M. N. Melo, S. J. Marrink, and M. Praprotnik, "Adaptive resolution simulation of polarizable supramolecular coarse-grained water models," J. Chem. Phys. **142**, 244118 (2015).

[124] J. Zavadlav, M. N. Melo, A. V. Cunha, A. H. de Vries, S. J. Marrink, and M. Praprotnik, "Adaptive resolution simulation of Martini solvents," J. Chem. Theory. Comput. **10**, 2591 (2014).

[125] J. B. Lu, Y. Q. Qiu, R. Baron, and V. Molinero, "Coarse-graining of TIP4P/2005, TIP4P-Ew, SPC/E, and TIP3P to monatomic anisotropic water models using relative entropy minimization," J. Chem. Theory Comput. **10**, 4104 (2014).

[126] W. Huang and W. F. van Gunsteren, "Challenge of representing entropy at different levels of resolution in molecular simulation," J. Phys. Chem. B **119**, 753 (2015).

[127] Y. A. Perez Sirkin, M. H. Factorovich, V. Molinero, and D. A. Scherlis, "Vapor pressure of aqueous solutions of electrolytes reproduced with coarse-grained models without electrostatics," J. Chem. Theory Comput. **12**, 2942 (2016).

[128] A. Ghoufi, P. Malfreyt, and D. J. Tildesley, "Computer modelling of the surface tension of the gas-liquid and liquid-liquid interface," Chem. Soc. Rev. **45**, 1387 (2016).

[129] M. Ndao, F. Goujon, A. Ghoufi, and P. Malfreyt, "Coarse-grained modeling of the oil–water–surfactant interface through the local definition of the pressure tensor and interfacial tension," Theor. Chem. Acc. **136**, 21 (2017).

[130] N. J. H. Dunn, T. T. Foley, and W. G. Noid, "Van der Waals perspective on coarse-graining: Progress toward solving representability and transferability problems," Acc. Chem. Res. **49**, 2832 (2016).

[131] A. Lyubartsev, A. Mirzoev, L. Chen, and A. Laaksonen, "Systematic coarse-graining of molecular models by the Newton inversion method," Faraday Discuss. **144**, 43 (2010).

[132] P. B. Warren, "A manifesto for one-body terms: the simplest of all many-body interactions?," J. Phys.: Condens. Matter **15**, S3467 (2003).

[133] T. Sanyal and M. S. Shell, "Coarse-grained models using local-density potentials optimized with the relative entropy: Application to implicit solvation," J. Chem. Phys. **145**, 034109 (2016).

[134] J. W. Wagner, T. Dannenhoffer-Lafage, J. Jin, and G. A. Voth, "Introducing Order Parameter Dependent Interactions for Multiscale Coarse-Graining (MS-CG)," J. Chem. Phys., In preparation (2017).

[135] J. A. Thomas and T. M. Cover, *Elements of Information Theory* (Wiley & Sons, New York, 1991).

[136] M. L. P. Price, D. Ostrovsky, and W. L. Jorgensen, "Gas-phase and liquid-state properties of esters, nitriles, and nitro compounds with the OPLS-AA force field," J. Comput. Chem. **22**, 1340 (2001).

[137] W. J. Cheong and P. W. Carr, "The surface tension of mixtures of methanol, acetonitrile, tetrahydrofuran, isopropanol, tertiary butanol and dimethyl-sulfoxide with water at 25C," J. Liq. Chromatogr. **10**, 561 (1987).

[138] G. Vazquez, E. Alvarez, and J. M. Navaza, "Surface tension of alcohol + water from 20 to 50 C," J. Chem. Eng. Data **40**, 611 (1995).

[139] *CRC Handbook of Chemistry and Physics* (CRC, Boston, 2003), 84 edn.

[140] M. Wu, T. Cubaud, and C.-M. Ho, "Scaling law in liquid drop coalescence driven by surface tension," Phys. Fluids **16**, L51 (2004).

[141] M. Matsumoto, Y. Takaoka, and Y. Kataoka, "Liquid–vapor interface of water–methanol mixture. I. Computer simulation," J. Chem. Phys. **98**, 1464 (1993).

[142] M. Matsumoto and Y. Kataoka, "Molecular orientation near liquid–vapor interface of methanol: Simulational study," J. Chem. Phys. **90**, 2398 (1989).

[143] M. Sega, B. Fábián, G. Horvai, and P. Jedlovszky, "How Is the surface tension of various liquids distributed along the interface normal?," J. Phys. Chem. C **120**, 27468 (2016).

[144] A. Davtyan, J. F. Dama, G. A. Voth, and H. C. Andersen, "Dynamic force matching: A method for constructing dynamical coarse-grained models with realistic time dependence," J. Chem. Phys. **142**, 154104 (2015).

[145] P. B. Canham, "The minimum energy of bending as a possible explanation of the biconcave shape of the human red blood cell," J. Theor. Biol. **26**, 61 (1970).

[146] W. Helfrich, "Elastic properties of lipid bilayers: theory and possible experiments," Z. Naturforsch. C. **28**, 693 (1973).

[147] E. G. Brandt, A. R. Braun, J. N. Sachs, J. F. Nagle, and O. Edholm, "Interpretation of fluctuation spectra in lipid bilayer simulations," Biophys. J. **100**, 2104 (2011).

[148] M. Dzugutov, "A universal scaling law for atomic diffusion in condensed matter," Nature **381**, 137 (1996).

[149] J. A. Armstrong, C. Chakravarty, and P. Ballone, "Statistical mechanics of coarse graining: estimating dynamical speedups from excess entropies," J. Chem. Phys. **136**, 124503 (2012).

[150] E. Voyiatzis, F. Müller-Plathe, and M. C. Bohm, "Excess entropy scaling for the segmental and global dynamics of polyethylene melts," Phys. Chem. Chem. Phys. **16**, (2014).

[151] N. J. H. Dunn and W. G. Noid, "Bottom-up coarse-grained models with predictive accuracy and transferability for both structural and thermodynamic properties of heptane-toluene mixtures," J. Chem. Phys. **144**, 204124 (2016).

[152] A. E. Olsen, J. C. Dyre, and T. B. Schroder, "Communication: Pseudoisomorphs in liquids with intramolecular degrees of freedom," J. Chem. Phys. **145**, 241103 (2016).

[153] A. Davtyan, J. F. Dama, A. V. Sinitskiy, and G. A. Voth, "The theory of ultra-coarse-graining. 2. Numerical implementation," J. Chem. Theory Comput. **10**, 5265 (2014).

[154] W. G. Noid, J. W. Chu, G. S. Ayton, and G. A. Voth, "Multiscale coarse-graining and structural correlations: Connections to liquid-state theory," J. Phys. Chem. B **111**, (2007).

[155] J. F. Rudzinski and W. G. Noid, "Bottom-up coarse-graining of peptide ensembles and helix-coil transitions," J. Chem. Theory Comput. **11**, 1278 (2015).

[156] G. Zhang, F. H. Stillinger, and S. Torquato, "Probing the limitations of isotropic pair potentials to produce ground-state structural extremes via inverse statistical mechanics," Phys. Rev. E **88**, 042309 (2013).

[157] M. Jochum, D. Andrienko, K. Kremer, and C. Peter, "Structure-based coarse-graining in liquid slabs," J. Chem. Phys. **137**, 064102 (2012).

[158] Z. Jia and J. Chen, "Necessity of high-resolution for coarse-grained modeling of flexible proteins," J. Comput. Chem. **37**, 1725 (2016).

[159] S. Izvekov, P. W. Chung, and B. M. Rice, "Particle-based multiscale coarse graining with density-dependent potentials: application to molecular crystals (hexahydro-1,3,5-trinitro-s-triazine)," J. Chem. Phys. **135**, 044112 (2011).

[160] J. D. Moore, B. C. Barnes, S. Izvekov, M. Lisal, M. S. Sellers, D. E. Taylor, and J. K. Brennan, "A coarse-grain force field for RDX: Density dependent and energy conserving," J. Chem. Phys. **144**, 104501 (2016).

[161] P. B. Warren, "Vapor-liquid coexistence in many-body dissipative particle dynamics," Phys. Rev. E **68**, 066702 (2003).

[162] S. Merabia and I. Pagonabarraga, "Density dependent potentials: Structure and thermodynamics," J. Chem. Phys. **127**, 054903 (2007).

[163] M. Arienti, W. Pan, X. Li, and G. Karniadakis, "Many-body dissipative particle dynamics simulation of liquid/vapor and liquid/solid interactions," J. Chem. Phys. **134**, 204114 (2011).

[164] A. Ghoufi, J. Emile, and P. Malfreyt, "Recent advances in many body dissipative particles dynamics simulations of liquid-vapor interfaces," Eur. Phys. J. E **36**, 10 (2013).

[165] S. Jamali, A. Boromand, S. Khani, J. Wagner, M. Yamanoi, and J. Maia, "Generalized mapping of multi-body dissipative particle dynamics onto fluid compressibility and the Flory-Huggins theory," J. Chem. Phys. **142**, 164902 (2015).

[166] V. Agrawal, P. Peralta, Y. Li, and J. Oswald, "A pressure-transferable coarse-grained potential for modeling the shock Hugoniot of polyethylene," J. Chem. Phys. **145**, 104903 (2016).

[167] B. Peters, "Reaction coordinates and mechanistic hypothesis tests," Annu. Rev. Phys. Chem. **67**, 669 (2016).

[168] D. W. Oxtoby, H. P. Gillis, and A. Campion, *Principles of Modern Chemistry* (Brooks/Cole, Belmont, CA, 2012), 7 edn.

[169] P. G. de Gennes and J. Prost, *The Physics of Liquid Crystals* (Clarendon Press, Oxford, 1995).

[170] P. J. Steinhardt, D. R. Nelson, and M. Ronchetti, "Icosahedral bond orientational order in supercooled liquids," Phys. Rev. Lett. **47**, 1297 (1981).

[171] P. J. Steinhardt, D. R. Nelson, and M. Ronchetti, "Bond-orientational order in liquids and glasses," Phys. Rev. B **28**, 784 (1983).

[172] P. L. Chau and A. J. Hardwick, "A new order parameter for tetrahedral configurations," Mol. Phys. **93**, 511 (1998).

[173] T. Kawasaki and A. Onuki, "Construction of a disorder variable from Steinhardt order parameters in binary mixtures at high densities in three dimensions," J. Chem. Phys. **135**, 174109 (2011).

[174] M. Leocmach, J. Russo, and H. Tanaka, "Importance of many-body correlations in glass transition: An example from polydisperse hard spheres," J. Chem. Phys. **138**, 12A536 (2013).

[175] C. Xia, J. Li, Y. Cao, B. Kou, X. Xiao, K. Fezzaa, T. Xiao, and Y. Wang, "The structural origin of the hard-sphere glass transition in granular packing," Nat. Commun. **6**, 8409 (2015).

[176] T. Q. Yu, P. Y. Chen, M. Chen, A. Samanta, E. Vanden-Eijnden, and M. Tuckerman, "Order-parameter-aided temperature-accelerated sampling for the exploration of crystal polymorphism and solid-liquid phase transitions," J. Chem. Phys. **140**, 214109 (2014).

[177] O. Valsson, P. Tiwary, and M. Parrinello, "Enhancing important fluctuations: Rare events and metadynamics from a conceptual viewpoint," Annu. Rev. Phys. Chem. **67**, 159 (2016).

[178] J. F. Dama, J. Jin, and G. A. Voth, "The theory of ultra-coarse-graining. 3. Coarse-grained sites with rapid local equilibrium of internal states," J. Chem. Theory Comput., in press (2017).

[179] R. L. McGreevy and L. Pusztai, "Reverse Monte Carlo simulation: A new technique for the determination of disordered structures," Mol. Sim. **1**, 359 (1988).

[180] S. Izvekov, "Towards an understanding of many-particle effects in hydrophobic association in methane solutions," J. Chem. Phys. **134**, 034104 (2011).

[181] F. M. Schaller, M. Neudecker, M. Saadatfar, G. W. Delaney, G. E. Schroder-Turk, and M. Schroter, "Local origin of global contact numbers in frictional ellipsoid packings," Phys. Rev. Lett. **114**, 158001 (2015).

[182] A. Ardevol, F. Palazzesi, G. A. Tribello, and M. Parrinello, "General protein data bank-based collective variables for protein folding," J. Chem. Theory Comput. **12**, 29 (2016).

[183] L. B. Lucy, "A numerical approach to the testing of the fission hypothesis," Astrophys. J. **82**, 1013 (1977).

[184] R. A. Gingold and J. J. Monaghan, "Smoothed particle hydrodynamics: Theory and application to non-spherical stars," Mon. Not. R. Astron. Soc. **181**, 375 (1977).

[185] J. D. Weeks, D. Chandler, and H. C. Andersen, "Role of repulsive forces in determining the equilibrium structure of simple liquids," J. Chem. Phys. **54**, 5237 (1971).

[186] M. E. van Leeuwen and B. Smit, "Molecular simulation of the vapor-liquid coexistence curve of methanol," J. Phys. Chem. **99**, 1831 (1995).

[187] L. X. Dang and T.-M. Chang, "Many-body interactions in liquid methanol and its liquid/vapor interface: A molecular dynamics study," J. Chem. Phys. **119**, 9851 (2003).

[188] S. Patel and C. L. Brooks, 3rd, "A nonadditive methanol force field: bulk liquid and liquid-vapor interfacial properties via molecular dynamics simulations using a fluctuating charge model," J. Chem. Phys. **122**, 024508 (2005).

[189] I. F. W. Kuo, C. J. Mundy, M. J. McGrath, and J. I. Siepmann, "Structure of the methanol liquid−vapor interface: A comprehensive particle-based simulation study," J. Phys. Chem. C **112**, 15412 (2008).

[190] W. L. Jorgensen and J. M. Briggs, "Monte Carlo simulations of liquid acetonitrile with a three-site model," Mol. Phys. **63**, 547 (1988).

[191] X. Grabuleda, C. Jaime, and P. A. Kollman, "Molecular dynamics simulation studies of liquid acetonitrile: New six-site model," J. Comput. Chem. **21**, 901 (2000).

[192] S. Pothoczki and L. Pusztai, "Intermolecular orientations in liquid acetonitrile: New insights based on diffraction measurements and all-atom simulations," J. Mol. Liq. **225**, 160 (2017).

[193] T. Radnai and P. Jedlovszky, "Reverse Monte Carlo simulation of a heteronuclear molecular liquid: Structural study of acetonitrile," J. Phys. Chem. **98**, 5994 (1994).

[194] J. Richardi, P. H. Fries, R. Fischer, S. Rast, and H. Krienke, "Structure and thermodynamics of liquid acetonitrile via Monte Carlo simulation and Ornstein-Zernike theories," J. Mol. Liq. **73-74**, 465 (1997).

[195] H. J. Böhm, I. R. McDonald, and P. A. Madden, "An effective pair potential for liquid acetonitrile," Mol. Phys. **49**, 347 (2006).

[196] P. J. Gee and W. F. van Gunsteren, "Acetonitrile revisited: A molecular dynamics study of the liquid phase," Mol. Phys. **104**, 477 (2006).

[197] B. S. Jabes, D. Nayar, D. Dhabal, V. Molinero, and C. Chakravarty, "Water and other tetrahedral liquids: Order, anomalies and solvation," J. Phys.: Condens. Matter **24**, 284116 (2012).

[198] G. C. Sosso, J. Chen, S. J. Cox, M. Fitzner, P. Pedevilla, A. Zen, and A. Michaelides, "Crystal nucleation in liquids: Open questions and future challenges in molecular dynamics simulations," Chem. Rev. **116**, 7078 (2016).

[199] M. Ozboyaci, D. B. Kokh, S. Corni, and R. C. Wade, "Modeling and simulation of protein-surface interactions: Achievements and challenges," Q. Rev. Biophys. **49**, e4 (2016).

[200] E. Spohr, "Molecular simulation of the electrochemical double layer," Electrochim. Acta **44**, 1697 (1999).

[201] A. A. Franco, "Multiscale modelling and numerical simulation of rechargeable lithium ion batteries: Concepts, methods and challenges," RSC Adv. **3**, 13027 (2013).

[202] K. Kirchner, T. Kirchner, V. Ivaništšev, and M. V. Fedorov, "Electrical double layer in ionic liquids: Structural transitions from multilayer to monolayer structure at the interface," Electrochim. Acta **110**, 762 (2013).

[203] S. J. Marrink and D. P. Tieleman, "Perspective on the Martini model," Chem. Sov **42**, 6801 (2013).

[204] I. F. Thorpe, J. Zhou, and G. A. Voth, "Peptide folding using multiscale coarse-grained models," J. Phys. Chem. B **112**, 13079 (2008).

[205] M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, and E. Lindahl, "GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers," SoftwareX **1-2**, 19 (2015).

[206] S. Páll, M. J. Abraham, C. Kutzner, B. Hess, and E. Lindahl, in *Solving Software Challenges for Exascale. EASC 2014.*, edited by S. Markidis, and E. Laure (Springer, Cham, 2015), pp. 3.

[207] S. Pronk, S. Pall, R. Schulz, P. Larsson, P. Bjelkmar, R. Apostolov, M. R. Shirts, J. C. Smith, P. M. Kasson, D. van der Spoel, B. Hess, and E. Lindahl, "GROMACS 4.5: A high-throughput and highly parallel open source molecular simulation toolkit," Bioinformatics **29**, 845 (2013).

[208] R. B. Best, X. Zhu, J. Shim, P. E. Lopes, J. Mittal, M. Feig, and A. D. Mackerell, Jr., "Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone phi, psi and side-chain chi(1) and chi(2) dihedral angles," J, Chem. Theory Comput. **8**, 3257 (2012).

[209] T. Darden, L. Perera, L. Li, and L. Pedersen, "New tricks for modelers from the crystallography toolkit: The particle mesh Ewald algorithm and its use in nucleic acid simulations," Structure **7**, R55 (1999).

[210] G. Bussi, D. Donadio, and M. Parrinello, "Canonical sampling through velocity rescaling," J. Chem. Phys. **126**, 014101 (2007).

[211] M. Parrinello and A. Rahman, "Polymorphic transitions in single crystals: A new molecular dynamics method," J. Appl. Phys. **52**, 7182 (1981).

[212] A. Renevey and S. Riniker, "Improved accuracy of hybrid atomistic/coarse-grained simulations using reparametrised interactions," J. Chem. Phys. **146**, 124131 (2017).

[213] K. Kreis and R. Potestio, "The relative entropy is fundamental to adaptive resolution simulations," J. Chem. Phys. **145**, 044104 (2016).