The Pennsylvania State University
The Graduate School

HIERARCHICAL COARSE-GRAINING VIA A GENERALIZED YVON-BORN-GREEN FRAMEWORK: MANY-BODY CORRELATIONS, MAPPINGS, AND STRUCTURAL ACCURACY

A Dissertation in
Chemistry
by
Joseph F. Rudzinski

© 2015 Joseph F. Rudzinski

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

May 2015

The dissertation of Joseph F. Rudzinski was reviewed and approved* by the following:

Will G. Noid
Associate Professor of Chemistry
Dissertation Advisor, Chair of Committee

Barbara Garrison
Shapiro Professor of Chemistry
Head of the Chemistry Department

David Boehr
Assistant Professor of Chemistry

Janna Maranas
Associate Professor of Chemical Engineering
Associate Professor of Materials Science and Engineering

*Signatures are on file in the Graduate School.

Abstract

Atomically-detailed molecular dynamics simulations have emerged as one of the most powerful theoretic tools for studying complex, condensed-phase systems. Despite their ability to provide incredible molecular insight, these simulations are insufficient for investigating complex biological processes, e.g., protein folding or molecular aggregation, on relevant length and time scales. The increasing scope and sophistication of atomically-detailed models has motivated the development of "hierarchical" approaches, which parameterize a low resolution, coarse-grained (CG) model based on simulations of an atomically-detailed model. The utility of hierarchical CG models depends on their ability to accurately incorporate the correct physics of the underlying model. One approach for ensuring this "consistency" between the models is to parameterize the CG model to reproduce the structural ensemble generated by the high resolution model. The many-body potential of mean force is the proper CG energy function for reproducing all structural distributions of the atomically-detailed model, at the CG level of resolution. However, this CG potential is a configuration-dependent free energy function that is generally too complicated to represent or simulate. The multiscale coarse-graining (MS-CG) method employs a generalized Yvon-Born-Green (g-YBG) relation to directly determine a variationally optimal approximation to the many-body potential of mean force. The MS-CG/g-YBG method provides a convenient and transparent framework for investigating the equilibrium structure of the system, at the CG level of resolution. In this work, we investigate the fundamental limitations and approximations of the MS-CG/g-YBG method. Throughout the work, we propose several theoretic constructs to directly relate the the MS-CG/g-YBG method to other popular structure-based CG approaches. We investigate the physical interpretation of the MS-CG/g-YBG correlation matrix, the quantity responsible for disentangling the various contributions to the average force on a CG site. We then employ an iterative extension of the MS-CG/g-YBG method that improves the accuracy of a particular set of low order correlation functions relative to the original MS-CG/g-YBG model. We demonstrate that this method provides a powerful framework for identifying the precise source of error in an MS-CG/g-YBG model. We then propose a method for identifying an optimal CG representation, prior to the development of the CG model. We employ these techniques together to demonstrate that in the cases where the MS-CG/g-YBG method fails to determine an accurate model, a fundamental problem likely exists with the chosen CG representation or interaction set. Additionally, we explicitly demonstrate that while the iterative model successfully improves the accuracy of the low order structure, it does so by distorting the higher order structural correlations relative to the underlying model. Finally, we apply these methods to investigate the utility of the MS-CG/g-YBG method for developing models for systems with complex intramolecular structure. Overall, our results demonstrate the power of the g-YBG framework for developing accurate CG models and for investigating the driving forces of equilibrium structures for complex condensed-phase systems. This work also explicitly motivates future development of bottom-up CG methods and highlights some outstanding problems in the field.

Table of Contents

List of Figures

List of Tables

Acknowledgments

There are many people who have helped me to get this point. For me, this journey really began as an undergraduate at UCSB, so that is where I will begin. First, I would like to thank Professors Steven Buratto and Paul Atzberger for allowing me to work as an undergraduate researcher in their groups. I would likely not be in graduate school if it were not for these experiences and, at the time, not everyone was willing to give me a chance. I would also like to thank my mathematics academic advisor, Professor Maribel Cachadina, for helping me get into the College of Creative Studies. This allowed me to finish my math degree as well as take graduate chemistry courses, both of which had a profound affect on my graduate experience. I would like to thank Julie Standish at the MRL for helping me to get paid for doing my research. Without this funding, I would not have been able to dedicate even half the amount of time to research. I would like to thank Dan Gargas for taking the time to show me the research ropes, even in the midst of writing his thesis. I would also like to thank Katie Meihaus for helping me to kindle my passion for science in the early days. A special thanks to my mentor Jimmy O'dea. I have always admired Jimmy, but only in retrospect have I fully realized Jimmy's superior talent for teaching, immense patience, and sincere kindness. Throughout my graduate work, I have often fallen back on the skills that he helped me develop.

My graduate work in the Noid group has been both challenging and fulfilling. It has been a steadily small group, which has amplified the impact of each group member on my experience. I would like to thank the many undergraduates who have passed through the group, for helping me to hone my teaching and mentoring skills. I would like to thank Sushant Kumar for being an inspiration, both for his work ethic and his passion. I would like to thank Chris Ellis for being the Noid group spokesman, for always being available to chat (research or other), and for his collaborative effort on the Toluene paper. I would like to thank Tommy Foley for all of the time spent brainstorming research ideas, for always asking insightful questions, and for reviewing parts of this thesis. I would like to thank Nick Dunn for being the best possible office mate, for always being willing to lend a hand on a problem, for listening to me rant about research, for stepping up to the group computer duties, and for all of his time spent reviewing parts of this thesis.

I would also like to thank our collaborators in the chemical engineering department. I would like to thank Professor Scott Milner and Professor Janna Maranas for demonstrating how to carry out an efficient and successful collaboration. I would also like to thank Keran

Lu for all of his hard work on our papers and for making the collaboration so easy. I would also like to thank Janna for taking the time to be on my committee, especially on such short notice. I would like to thank my other committee members, Professor David Boehr and Professor Barbara Garrison, for taking their time for meetings and for reviewing this thesis. I would like to thank Professor Radu Roiban, who also dedicated his time to these things, but had to step down from my committee because of a sabbatical.

I would like to thank everyone who had a hand in organizing the Penn State Academic Computing Fellowship and associated meetings. I have been honored to serve as a computing fellow. The meetings have been crucial for maintaining perspective on my work and the funding from this award allowed me a great deal of freedom along my research path. Also, thank you to the other computing fellows along the way, for providing me with inspiration through their own passionate research interests. On a slightly different, but related note, I would like to thank the Penn State High Performance Computing group and, in particular, the Research Computing and Cyberinfrastructure unit. Much of this work was made possible by the hard work of this team in maintaining the phenomenal high performance computing services at Penn State.

I would especially like to thank my thesis advisor, Professor Will Noid. Will and I had an instant connection, over math, I think. When I joined his group, I knew absolutely nothing about classical simulations. Will has been my primary teacher on many important subjects including statistical mechanics, molecular simulations, coarse-graining, etc, and has helped me to find my niche in the research world. Will has been a true advisor, always making himself available for questions or to discuss research and always pushing us to improve and expand our scientific skill set. I would specifically like to thank Will for his seemingly infinite patience, especially when I had lost all of mine.

Now, I would like to add some personal acknowledgements. First, I would like to thank my family as a whole for always supporting me. I would like to thank my parents for instilling me with the importance of education from a young age and for encouraging me to follow my dreams. I would like to thank my siblings for all of their support, especially for overcoming the long distance in recent years and for dealing with my horrible communication skills. I would especially like to thank my brother, Richard, for always taking care of me (I know this is not fooling anyone, you are obviously still the smart one!). I would also like to thank Michael Vincent for being a second brother to me and to providing me with unconditional support.

Lastly, a very deep and sincere thank you to my beautiful wife, Alli. Thank you for taking a chance on me and moving across the country from Santa Barbara to State College. Thank you for dealing with the long days and nights early on and also the long days and nights later on. Thank you for understanding, but not encouraging, my reluctance to take vacation. Thank you for letting me be me. Thank you for being an unrelenting positive force. This work would not have been possible without you.

For Big Hoobs. 
auf zu unserem nächsten Abenteuer.

Chapter 1

Introduction

1.1 Molecular Simulations

From an outsider's perspective, molecular simulations may appear to be a relatively new technique, due to the somewhat recent explosion in computer technology. On the contrary, Karplus, Levitt, and Warshel's 2013 Nobel Prize in Chemistry for "the development of multiscale modeling of complex chemical systems" is indicative of the maturity of the field of molecular simulations.[1] The seminal contributions[2,3] to this work were performed more than forty years ago and, still, the underlying methodology[4,5] was introduced another two decades earlier! Despite this long history, there has never been a more exciting time for computational and theoretical chemists.

Steady improvement in computational hardware,[6] the development of powerful and reliable simulation packages,[7,8] as well as a large number of methodological advances[9–13] have elevated molecular simulations towards a more even footing with experiments.[14] It is now quite common, especially in the biochemical community, for simulations to contribute significantly to the understanding of complex chemical processes by providing concrete molecular insight that complements experimental data.[15] In particular, atomically-detailed or all-atom (AA) molecular dynamics (MD) simulations have emerged as one of the most powerful tools for modeling complex, condensed-phase systems. MD is a classical simulation method that numerically integrates Newton's equations of motion to propagate the atoms in a system according to some given, usually empirical, potential energy function or "force field."[16] The widespread popularity of MD simulations stems from both the dedicated development of accurate force fields[17–19] as well as the ease of implementation with state of the art simulation packages.[7,8]

Lower resolution, coarse-grained (CG) models are also routinely employed and have, in

many cases, contributed to the development of our basic understanding of complex molecular processes. For example, relatively simple CG models[3,20–22] have provided immense insight into the driving forces of protein folding. In recent years, there has been a resurgence of interest in developing methodology for determining more accurate CG models.[23] In particular, motivated by the increased accuracy of AA models, many of these methods aim to parameterize the CG model using information from simulations of the higher resolution model. The present work will primarily focus on these "hierarchical" or "bottom-up" CG methods.

1.2 Why Coarse-grain?

Perhaps the most obvious motivation for coarse-graining is efficiency. AA MD simulations can provide great insight into modestly sized ($\sim 10^5$ atoms) systems for hundreds of nanoseconds.[24] However, these simulations quickly become intractable for investigating processes evolving on longer time scales, e.g., protein folding or molecular aggregation. Coarse-grained models provide increased efficiency with respect to the underlying AA model for several reasons.[23,25] First, reducing the number of particles directly reduces the number of force calculations performed at each MD step. Second, eliminating the fast motions in the system allows for a larger time step in the simulation protocol. Finally, with fewer particles and, often, smoother potentials, the potential energy surface governing the motion of the CG sites is less rugged, allowing the sites to diffuse more quickly than their underlying atoms.

A less commonly cited, but potentially more important, reason for coarse-graining is to identify the driving forces of a process or the key interactions stabilizing a particular equilibrium structure. Identifying the essential motions or important interactions from an AA MD trajectory can be a highly non-trivial task, due to the large number of degrees of freedom. There is an active field dedicated to developing techniques that directly address this problem.[26] CG models simplify this task substantially, by reducing the degrees of freedom that are simulated. Ideally, a CG model will employ the minimum degrees of freedom necessary to reproduce the (pertinent) physics of the underlying model. Reducing the description of a system to the essential components is a fundamental aspect of science, that appears in various forms across disciplines.

1.3 Scope

Because of the generality of the approach, coarse-graining may be interpreted in many different ways. In the present work, we will be concerned solely with particle-based models and

tend to focus on relatively high resolution CG models that group a modest number of atoms ($\precsim$ 20) into a single CG site. Even within this context, it is difficult to classify different CG approaches, because a given method can fall into multiple categories. Additionally, it is becoming increasingly common to combine multiple methods in an attempt to utilize the strengths of each approach.

Two useful, broad categories are top-down and bottom-up approaches.[23] Top-down methods parameterize the CG model to reproduce experimentally measured properties of the system. For example, several recent models have been parameterized to reproduce the density and surface tension[27] or solvation free energies[28] for a particular set of systems. Although this parameterization is usually performed in an ad hoc manner, the resulting models often exhibit reasonable transferability (i.e., the ability to accurately apply a single model to multiple thermodynamic state points). Consequently, these models have gained wide-spread popularity for investigating a large variety of systems.[29] However, models parameterized in this way tend to poorly describe the structural features of the underlying model[30–33] and may lack the chemical specificity to accurately model the properties of a particular system, which makes them much less useful for studying molecules with complex intramolecular structure (e.g., proteins).

On the other hand, bottom-up methods parameterize the CG model based upon an underlying or higher resolution, e.g., AA, model. In particular, many of these methods specifically aim to reproduce the structure of the underlying model. Figure 1.1 schematically illustrates this "inverse" problem: Given an AA model and a transformation (i.e., mapping) from the AA to the CG representation, does there exist a potential energy function for the CG model that will reproduce the structural features of the AA model when viewed at the CG level of resolution? Several methods[34–38] have been proposed to approximately solve this problem. These parameterization methods are built from a statistical mechanical framework, often resulting in more systematic procedures than those employed with top-down methods. However, the resulting models typically do not accurately model the thermodynamics of the underlying system and tend to have limited transferability.[34,39–41] Nevertheless, these methods are ideal for modeling complex biological molecules that form specific three-dimensional structures, since they focus on reproducing structural features of the underlying model.

The accuracy of a bottom-up model is limited by the accuracy of the corresponding underlying model. Despite known deficiencies in standard AA models,[42] these models contain a great deal of physical information, which was incorporated during their parameterization using a large amount of experimental data.[17–19] In contrast to top-down models, which are underdetermined by the relatively small amount of available macroscopic information,

bottom-up models may be capable of more naturally incorporating the features of the system of interest. Consequently, bottom-up models have the potential to be predictive, but only if they can accurately retain the essential physics of the underlying model. This is the fundamental challenge of bottom-up coarse-graining and will be the primary focus of this work.

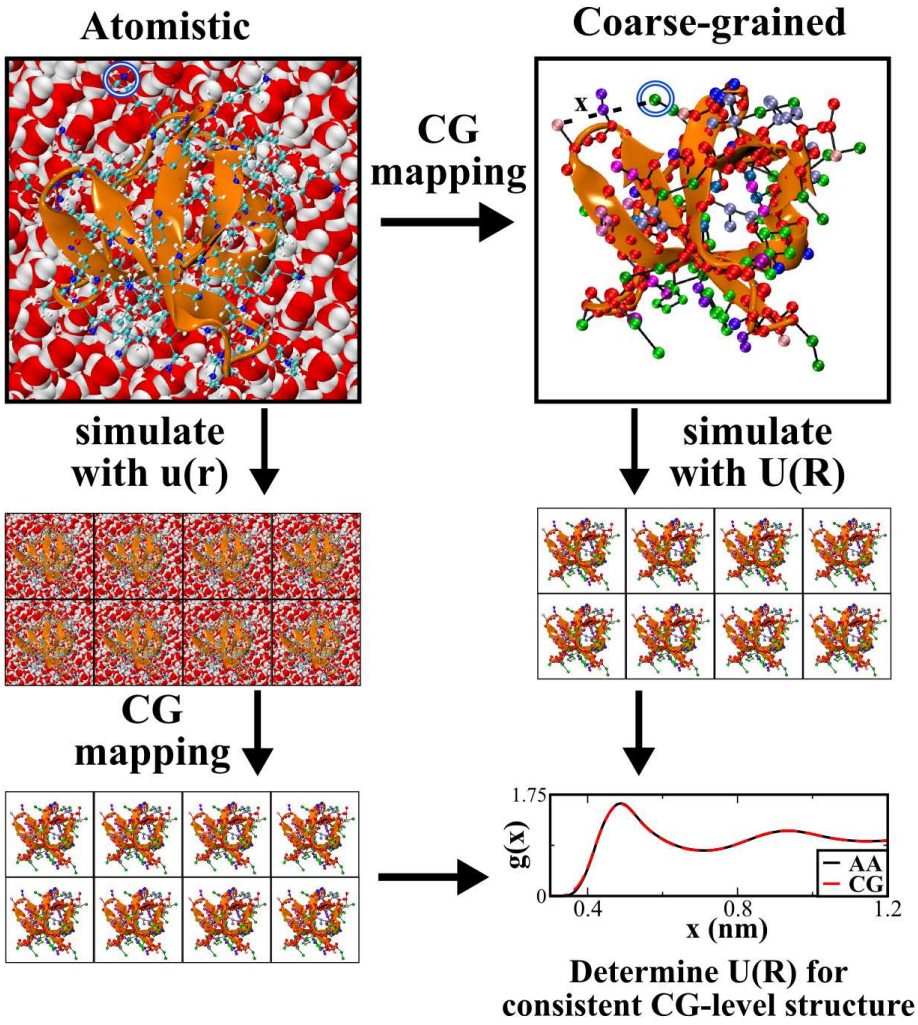

Figure 1.1. Schematic of the inverse problem of structure-based bottom-up CG methods. A transformation, or CG mapping, is defined that transforms some AA system (top left) with configuration $\mathbf{r}$ to a corresponding CG system with configuration $\mathbf{R}$. The AA system is simulated according to some given potential energy function $u(\mathbf{r})$, yielding an ensemble of AA structures (middle left). Then, the same CG mapping is applied to the AA ensemble, resulting in a CG-level ensemble of structures (bottom left) that was generated by the higher resolution model. From this mapped ensemble, structural features of the AA model, at the CG level of resolution, are determined. The CG model could also be simulated with some potential, $U(\mathbf{R})$, yielding a CG ensemble of structures (middle right). The inverse problem is to determine $U(\mathbf{R})$ such that the structural features of the CG ensemble accurately reproduce those of the mapped AA ensemble (bottom right).

1.4 Structure-based Bottom-up Coarse-graining

1.4.1 The Inverse Problem

Statistical mechanics provides a convenient language for precisely describing the inverse problem of bottom-up CG modeling. For the present discussion we will assume all systems to be in the canonical ensemble, i.e., constant composition, volume (V), and temperature (T). First, let us assume that there exists a high resolution model that accurately describes the system of interest. In the following, this will usually correspond to a standard AA model, but the arguments presented here are completely general. In this case, let the AA system contain n atoms and let its configuration be given by a $3n$-dimensional ($3n$-D) vector $\mathbf{r}$. Given an arbitrary AA potential energy function, $u(\mathbf{r})$, the probability, $p_r(\mathbf{r})$, of observing a configuration, $\mathbf{r}$, in the AA model is given by the Boltzmann distribution: $p_r(\mathbf{r}) = Z^{-1}\exp[-u(\mathbf{r})/k_BT]$; where $Z = \int d\mathbf{r}\exp[-u(\mathbf{r})/k_BT]$ is the AA canonical configuration integral and k_B is Boltzmann's constant.[43]

Now, consider a CG system with N sites ($N \leq n$) and a configuration $\mathbf{R}$ determined by a linear transformation, $\mathbf{M}$, of the AA coordinates: $\mathbf{R} = \mathbf{M}(\mathbf{r})$. Accordingly, the probability, $P_R(\mathbf{R})$, of observing a configuration, $\mathbf{R}$, in the CG model with some potential energy function, $U(\mathbf{R})$, is: $P_R(\mathbf{R}) = Z_U^{-1}\exp[-U(\mathbf{R})/k_BT]$; where $Z_U = \int d\mathbf{R}\exp[-U(\mathbf{R})/k_BT]$ is the canonical configuration integral for the CG model.

To view the AA ensemble at the CG level of resolution, we define a "mapped" probability distribution, $p_R(\mathbf{R})$, as the probability of observing an AA configuration that maps to the particular CG configuration, $\mathbf{R}$. This distribution is given by: $p_R(\mathbf{R}) = \langle\delta(\mathbf{M}(\mathbf{r}) - \mathbf{R})\rangle$, where the angular brackets denote an average according to the atomistic probability distribution. The delta-function, $\delta(\mathbf{M}(\mathbf{r}) - \mathbf{R})$, effectively eliminates contributions to the average from atomic configurations that do not map to $\mathbf{R}$. One way to ensure consistency between the AA and CG configurational ensembles (at the CG level of resolution) is to require that the mapped AA probability distribution is equal to the probability distribution of the CG model, i.e., $P_R(\mathbf{R}) = p_R(\mathbf{R})$.[41] Solving for $U(\mathbf{R})$ in this expression determines a particular CG potential energy function, $U^0(\mathbf{R}) = -k_BT\ln p_R(\mathbf{R}) + \text{const}$, that will reproduce all structure distributions of the AA model, at the level of CG resolution. $U^0(\mathbf{R})$ is known as the many-body potential of mean force[44] (PMF) and is the fundamental quantity in structure-based bottom-up CG methods. Because the PMF explicitly depends on the many-body distribution function, $p_R(\mathbf{R})$, it is generally too complicated to represent or simulate for any non-trivial system. Consequently, structure-based bottom-up CG methods aim to

approximate this function with some simpler potential energy function.

1.4.2 A *Very* Simple Example

In this section, we will illustrate the fundamental difficulties of the inverse problem and introduce important terminology and concepts that will be used later. Consider a monatomic liquid whose only interaction, $u^{(2)}(r)$, is a radially-symmetric function of the distance, r, between a pair of particles. For concreteness, let the interaction be a Lennard-Jones function (red curve in the top panel of Figure 1.2). The negative derivate of $u^{(2)}(r)$ determines the direct force, $f(r)$ (red curve in the bottom panel of Figure 1.2), on a central particle from a second particle that is a distance r away.

A simulation of this system in the NVT ensemble (T = 298 K) determines the radial distribution function (rdf), $g(r)$, shown in Figure 1.2 (black curve). $g(r)$ describes the average number of particles in a spherical shell of radius r from a central particle, relative to the corresponding number of particles in an ideal gas.[16] The first maximum in the rdf (highlighted by the orange vertical line) corresponds to the most likely contact distance between two particles. In between the two panels in Figure 1.2, a snapshot from the simulation illustrates this "first solvation shell" (orange, transparent sphere) about a central particle (red, opaque circle). Notice that the first solvation shell does *not* occur at the distance that corresponds to the minimum in the pair potential, $u^{(2)}(r)$. As the distance between two particles increases from the first solvation shell, it becomes less likely to find a particle because of the high probability of finding a particle at a slightly closer distance. Consequently, the rdf decreases to a minimum value before increasing to a second maximum (highlighted by the yellow vertical line). This second maximum corresponds to another solvation shell, where particles are again likely to be found (yellow, transparent sphere in the snapshot). The rdf can be related to experimentally measurable quantities,[45] and is one of the most common measures of structural order in condensed-phase systems.

Many CG approaches simplify the inverse problem by aiming to reproduce low order structural distributions (i.e., distributions that only depend upon the relative positions of a small number of sites, e.g., rdfs), instead of the many-body distribution function, $p_R(\mathbf{R})$. To simplify the problem even further, let's consider the present example without any reduction in the degrees of freedom (i.e., no coarse-graining). In other words, given $g(r)$, can we determine the potential energy function (i.e., $u^{(2)}(r)$ in this case) that will generate this structure? This problem represents a simple instance of the inverse problem, simplified with respect to the general case because: 1) the underlying potential, $u^{(2)}(r)$, is simple enough for

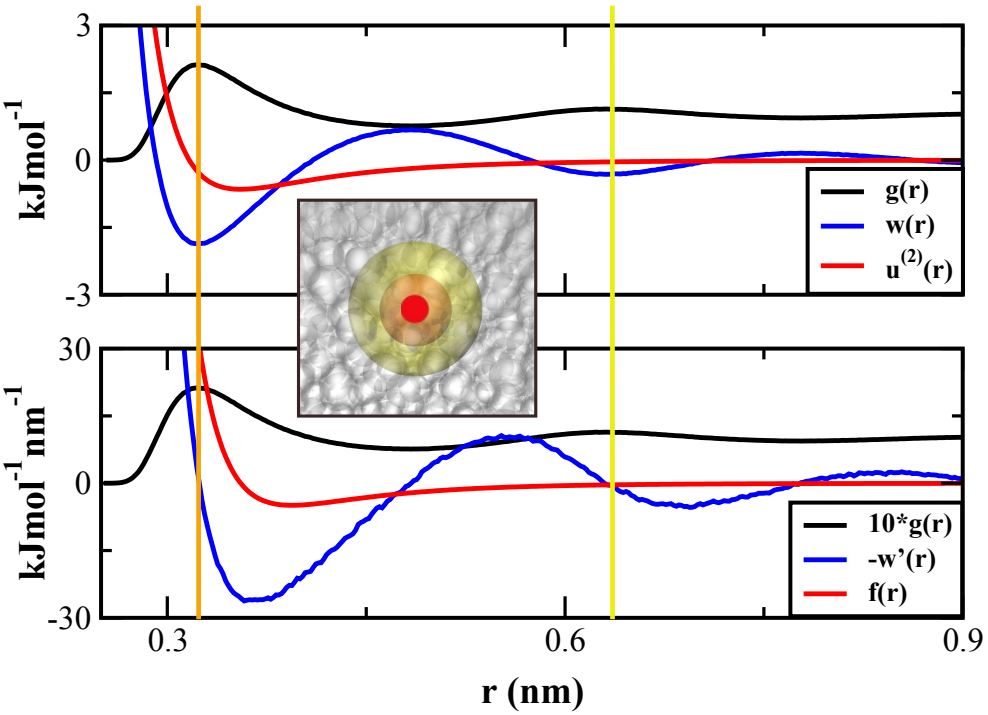

Figure 1.2. Analysis of a monatomic Lennard-Jones (LJ) model. The red curves present the LJ potential, $u^{(2)}(r) = 4\epsilon\left[\frac{\sigma^{12}}{r^{12}} - \frac{\sigma^6}{r^6}\right]$, (top panel) and the corresponding force, $f(r) = -u^{(2)'}(r)$, (bottom panel). The black curves present the rdf generated by simulations of 216 particles according to $u^{(2)}(r)$. (Note that the rdf has arbitrary units and the rdf in the bottom panel is scaled for visual convenience). The blue curves present the pair potential of mean force, $w(r) = -k_BT \ln g(r)$, (top panel) and the pair mean force, $-w'(r)$, (bottom panel). Between the two panels, a snapshot from the simulation is presented, with the first and second solvation shells illustrated as orange and yellow spheres, respectively. The radial distances of the first and second solvation shells are highlighted with the orange and yellow vertical lines, respectively.

us to represent and simulate exactly, 2) there is no coarse-graining, so we know that there exists an exact answer to the problem in this case, and 3) we are only concerned with the pair structure (i.e., $g(r)$) of the system.

Starting from $g(r)$, we can directly determine the pair potential of mean force, $w(r) = -k_BT \ln g(r)$ (blue curve in the top panel). $w(r)$ is called the pair potential of mean force because its negative derivative determines the mean (average) force on a particle when a second particle is a distance r away (blue curve in the bottom panel). Notice that by construction the maximum of $g(r)$ corresponds to the minimum of $w(r)$ or, equivalently, the distance where $-w'(r)$ passes through zero. Additionally, the fluctuating structure of the rdf is inherited by both $w(r)$ and $-w'(r)$. Moreover, the bottom panel of Figure 1.2 demonstrates that the average force, $-w'(r)$, felt by a central particle when a second particle is a particular distance, r, away is quite distinct from the direct force, $f(r)$. In other words, there are both

direct and indirect contributions to the pair mean force. The indirect contribution represents an average force generated on the particle from all possible configurations of the remaining particles when two particles are fixed at a distance r apart. Because the indirect contribution depends on this ensemble average, it is non-trivial to determine even for this extremely simple example!

In general, the inverse problem is further complicated because $g(r)$ will correspond to structural features that were generated by the underlying, high resolution model. Moreover, the CG potential energy function may contain several different interactions. Each of these interactions has a corresponding pair mean force, which will have indirect contributions from each of the other interactions. To reproduce the distribution along each CG degree of freedom, these contributions must be simultaneously disentangled to recover the proper set of pair potentials that will reproduce each pair mean force.

1.4.3 Iterative Methods

One way to implicitly disentangle the contributions to the mean force is to iteratively adjust the CG potentials until the pair structure is accurately reproduced. This is the basis of some of the most popular structure-based CG techniques.[34,35] These methods typically follow the same basic algorithm: 1) guess an initial CG potential energy function, 2) simulate the CG model according to this potential, 3) compare the resulting structure with the structure generated by the higher resolution model (mapped to the CG representation), 4) adjust the CG potential based on this comparison, 5) repeat steps 2-4 until convergence. Different iterative methods adopt various protocol for adjusting the potential in step 4. In principle, there is a unique potential which generates a given rdf (assuming that such a potential exists).[46] However, the existence of the potential is not guaranteed, especially when coarse-graining is involved, since the underlying interactions may generate a sufficiently complex structure that cannot be reproduced by some simple CG potential. Moreover, it is well known that the pair structure of a liquid is largely determined by the short-range portion of the potential.[35,47,48] In practice, this can significantly complicate the convergence and robustness of these procedures. Nevertheless, these methods have been used to develop a wide range of CG models that accurately reproduce the structure of various underlying systems.[38,49–58] The strength of these methods is that, if converged, they guarantee that the pair structure (i.e., rdfs) is quantitatively reproduced. However, for molecules with complex intramolecular structure (e.g., 3-D structures of proteins), it is unknown under what conditions reproducing the pair structure will also correspond to accurately modeling

the global structure of the system.

1.4.4 The MS-CG and G-YBG Methods

The multiscale coarse-graining (MS-CG) method, developed by Izvekov and Voth,[36,59] employs a different strategy to solve the inverse problem. This method aims to directly approximate the many-body mean force (MF), i.e., the force field determined by gradients of the PMF. To determine this approximation, a set of basis vectors is defined which directly corresponds to the chosen form of the CG potential energy function. The more complex the CG potential, the greater the range or "space" of force fields which can be represented by the basis vectors. For an arbitrarily complex potential, the space of force fields will be infinite and any force field, in particular the MF, can be represented. In general though, the CG potential will be chosen to be of a relatively simple form such that it can be simulated with standard molecular simulation software. Consequently, the corresponding basis vectors will be inadequate to represent the MF. In this case, the MS-CG method determines a CG force field that is simple enough to be represented by the basis vectors and which provides an optimal approximation to the MF.

In practice, the MS-CG force field is calculated by performing a mathematical projection of the MF into the space of force fields defined by the basis vectors. Amusingly, this projection is simply a generalization of the Pythagorean theorem defined in a vector space of force fields. Importantly, the MS-CG method determines a variational minimization problem which implies that the approximation to the MF (according to the metric employed in the MS-CG method) is guaranteed to improve as the number or complexity of the basis functions increases. Additionally, this variational problem corresponds to solving a linear least squares problem with respect to the parameters of the CG potential.[60] This is a distinguishing property of the MS-CG method because it implies that the method is direct (i.e., the optimal CG force field is determined without ever simulating the CG model) and can be implemented with standard and robust numerical methods.

An equivalent way to solve this linear least squares problem is to solve a corresponding linear system of normal equations (i.e., a matrix equation).[61] Within this framework, a correlation matrix emerges which describes the cross-correlations between CG degrees of freedom. Additionally, the target vector of these normal equations corresponds to a force correlation function, which can be re-expressed in terms of structures alone.[62] Thus, the normal equations can be solved with only structural information, without any reference to forces. Moreover, this new set of equations is a generalization of the famous Yvon-

Born-Green (YBG) equation of liquid state theory.[63,64] This generalized YBG (g-YBG) method determines the same CG force field as the MS-CG method, both in theory and in practice. For a radially-symmetric, distance-dependent potential, the g-YBG method relates the target force correlation function to the pair mean force. Consequently, an alternative interpretation of the MS-CG/g-YBG method is that it employs the cross-correlation matrix to (approximately) disentangle the contributions to the mean force along each CG degree of freedom.

The MS-CG/g-YBG method determines a variationally optimal approximation to the PMF and, consequently, recovers the PMF in the limit of a complete basis set. In the more general case of an incomplete basis set, MS-CG/g-YBG models have demonstrated remarkable accuracy for a wide range of systems.[65–67] Even though these models are not guaranteed to reproduce any particular set of structural correlation functions, in many cases they reproduce the rdfs and intramolecular 1-D distributions either quantitatively or nearly quantitatively. Considering the example above, which demonstrated that the pair structure of a liquid is non-trivially related to the underlying potential energy function, it is rather remarkable that this direct calculation can attain comparable accuracy to the iterative methods, which specifically aim to reproduce particular structural features and require a much greater computational effort. Methodological advances[68–73] continue to propel these methods, increasing their efficiency, accessibility and overall utility.

However, there are a handful of examples in the literature which reveal potential limitations of these methods. Rühle et al.[74] demonstrated that an MS-CG model of hexane, which represented each molecule with three CG sites, failed to even qualitatively reproduce the intramolecular angle distribution between the sites. Interestingly, it was later shown[75] that 2- and 4-site MS-CG models of the same underlying system reproduced the structure quite accurately. In a different study, Voth and coworkers[76,77] developed low resolution (1-3 CG sites per residue) MS-CG models of both an alpha-helix and a beta-hairpin. These models were reasonably accurate, although the hairpin structures achieved considerably lower accuracy than the helices. In a follow-up study, Thorpe et al.[78] demonstrated that the MS-CG method failed to derive a single force field that accurately modeled multiple secondary structures.

These studies, among others, motivate investigation into the fundamental approximations and limitations of the MS-CG/g-YBG method. In particular, it is generally unclear how the accuracy of an MS-CG/g-YBG model varies with, e.g., the chosen CG mapping and interaction set. The MS-CG/g-YBG correlation matrix is undoubtedly a key component of the MS-CG/g-YBG framework; yet, there has been relatively little work to understand its

precise physical significance and role in determining an optimal CG force field. Additionally, although there have been some comparisons[74,79] of the different structure-based CG methods, a deeper understanding of the connections between these methods may provide insight into the strengths and limitations of each. Specifically, precisely identifying the difficulties in applying these methods to complex biological molecules may assist in directing future investigations and developments. Finally, the utility of the MS-CG/g-YBG method also depends on the continued development of numerical methods for robust and accurate calculations.

1.5 Outline

In this work, we will investigate the MS-CG/g-YBG method. Our goal is to elucidate the fundamental approximations and limitations of the method, make rigorous connections to other CG methods, and to develop systematic and automated protocols for optimizing the accuracy of the CG model. The remaining manuscript is organized as follows.

In Chapter 2, we illustrate the typical development of a CG model using the MS-CG/g-YBG method. First, we review the essential features of the theory, focusing on the physical significance of each component. We then investigate the sensitivity of the resulting CG model to a small change in the mapping for a 3-site model of liquid toluene. Finally, we assess the higher order structural accuracy and the transferability of the models.

In Chapter 3, we investigate the relationship between two CG methods that employ distinct variational principles, namely the MS-CG and Relative Entropy methods. First, we present a unified theoretic framework for the two methods. We then show that these two seemingly disparate methods are rigorously related to the same information function. Finally, we demonstrate the consequences of this connection with a few simple numerical examples.

In Chapter 4, we investigate the MS-CG/g-YBG correlation matrix which is responsible for decomposing the mean force along each CG degree of freedom into contributions from each interaction. We demonstrate that for 1- to 3-site models of liquid heptane, the features in the correlation matrix can be interpreted in terms of basic liquid packing properties. We then explicitly demonstrate how the correlation matrix decomposes the mean force along each interaction, implicating this framework as a powerful tool for identifying key interactions within the CG model. Finally, we perform an eigenspectrum analysis to assess the sensitivity of the structure to changes in the CG potential.

In Chapter 5, we investigate the relationship between the CG mapping and the accuracy of the resulting model, both in the context of the MS-CG/g-YBG method and also more

popular iterative methods. We present a framework for interpreting iterative methods based on the MS-CG/g-YBG approach and propose a heuristic modification to provide more robust treatment of intramolecular interactions. We then propose a simple analysis to discriminate, a priori, between good and bad CG representations. Finally, we demonstrate the utility of this method for building accurate models of liquid hexane and 3-hexylthiophene and, simultaneously, apply an iterative g-YBG approach to pinpoint the source of errors in the case of bad CG representations.

In Chapter 6, we investigate the utility of the MS-CG/g-YBG method for determining minimal CG models of peptides. We demonstrate that the resulting models are quantitatively accurate when the underlying model corresponds to a well-defined structure. We use the analysis methods developed in the preceding chapters to analyze the problems that arise when the underlying model samples a more complex, heterogeneous ensemble of structures. Finally, we identify the dominant sources of errors in this case and explicitly verify them.

In Chapter 7, we provide a summary of the preceding chapters. We present these results in a broader context, reviewing other related studies whenever appropriate. We also provide an overview of practical developments that we have made along the way, but which were largely passed over in the main chapters. Finally, throughout the summary, we highlight outstanding issues which should be the focus of future investigations.

Chapter 2

Generalized-Yvon-Born-Green Model of Toluene

C. R. Ellis, J. F. Rudzinski, W. G. Noid *Macromol. Theory Simul.* **2011**, 20, 478-495[1]

Abstract

Coarse-grained (CG) models provide a highly efficient computational means for investigating complex processes that evolve on large length-scales or long time-scales. The predictive capability of these models relies upon their ability to reproduce the relevant structural properties of accurate, though prohibitively expensive, atomistic models. The many-body potential of mean force (PMF) is the appropriate potential for a CG model that quantitatively reproduces the structure of an underlying atomistic model. Because this PMF cannot be readily calculated or simulated, several methods attempt to systematically approximate this PMF with relatively simple molecular mechanics potentials. Recently, we have proposed a generalized-Yvon-Born-Green (g-YBG) approach to determine approximate potentials for accurate CG models directly from structural information. In the present work, we demonstrate the mechanism by which the g-YBG approach employs simple structural information to characterize and approximate the many-body PMF. We then employ this approach to parameterize a three site CG model for liquid toluene. We demonstrate that this model accurately reproduces the structural properties of an all-atom model. Moreover, using this model system, we demonstrate the variational nature of the method and investigate the sensitivity of the model to the CG mapping. Finally, we briefly investigate the transferability of the CG model to different temperatures.

[1]CRE and JFR are equally contributing coauthors. CRE performed simulations, determined and assessed accuracy of MSCG/g-YBG models, and contributed to the methods section. JFR analyzed force correlation functions, performed orientation calculations, and contributed to the theory section.

2.1 Introduction

Atomically detailed molecular dynamics (MD) simulations have contributed profound insight into the structure, thermodynamics, and dynamics of many condensed phase systems.[16,45] By propagating the motion of each atom within a system, these simulations provide exquisite resolution for studying molecular processes on microsecond timescales and nanometer length scales,[80] of course, subject to the accuracy of an empirical force field for a given system.[81] Despite tremendous advances in computational hardware and software, though, atomically detailed MD simulations remain prohibitively expensive for effectively investigating processes that occur on significantly longer length and time scales.

The computational expense of atomistic simulation methods has motivated the development of "coarse-grained" (CG) models that represent molecular systems in somewhat reduced detail by grouping atoms into bigger effective interaction sites.[82–86] The resulting CG models are frequently expected to be three orders of magnitude more efficient than atomically detailed models.[87,88] By focusing on the essential details of a particular system or process, CG models also significantly simplify subsequent analysis.[89] Consequently, CG models provide a powerful framework for hypothesis-driven investigations of specific interactions for a given phenomena.[90–92] Moreover, CG models have been particularly useful for investigating slow processes involving complex industrial or biological polymers, e.g., the diffusion of gases and additives through polymeric melts[93] or the association of viral capsid proteins.[94]

However, despite their computational efficiency, CG models may be misleading if they have not been carefully parameterized to reproduce the "correct physics" governing a specific process.[92,95,96] Consequently, a wide range of approaches have been developed for ensuring consistency between a CG model and either experimental data or simulations of accurate high resolution models. These approaches often focus on either thermodynamic or structural properties. Klein and coworkers[27,33,97–101] and Marrink and coworkers[28,32,102–106] have pioneered the parametrization of CG models that reproduce thermodynamic properties such as partitioning, liquid density, surface tension, or solvation free energies. In addition to accurately modeling thermodynamic properties, these models have demonstrated considerable transferability for modeling a variety of molecular systems in a range of thermodynamic conditions.[30,107,108] Although a recent study has extended the MARTINI approach to accurately model the radius of gyration of polystyrene,[31] CG models that have been parametrized to reproduce thermodynamic properties may prove less successful for modeling the conformations and interactions of proteins and other complex polymers.[32,33]

Alternatively, structure-motivated strategies parameterize CG models to quantitatively

reproduce the structural properties of a particular system. In this case, the appropriate potential for the CG model is a many-body potential of mean force (PMF),[44] i.e., a configuration-dependent free energy,[109] that reflects both the configuration distribution of the high resolution model and also the mapping from the high resolution model to the CG model. A model that employs the PMF as a conservative potential will quantitatively reproduce (at the resolution of the CG mapping) all of the structural properties of the atomistic model.[41,110] However, the many-body PMF cannot be effectively calculated, represented, or simulated.[111] Consequently, structure-motivated methods often approximate the PMF with simpler molecular mechanics-style interaction potentials.

The terms in the approximate potential are frequently parameterized to reproduce a corresponding set of simpler structural correlation functions. If the interactions in the CG model are statistically independent or only weakly coupled, then direct Boltzmann inversion determines each term in the potential immediately from the corresponding distribution function.[112–114] However, if the interactions in the CG model are more strongly coupled, then each distribution function reflects not only the corresponding interaction potential, but also the correlated forces from the surroundings. Peter, van der Vegt, and coworkers have employed a systematic procedure to directly subtract this environment mediated contribution from the potential obtained from direct Boltzmann inversion.[115–117] In addition, iterative Boltzmann inversion[35,118,119] and several related approaches, such as the inverse Monte Carlo method,[34,50,53] relative entropy framework,[37,120,121] and molecular renormalization group approach[38,122,123] iteratively refine the CG potential over multiple simulations until it accurately reproduces target distribution functions. These iterative approaches have been related to a variety of elegant variational principles[37,53,124] and have also been extended for considering density-dependent potentials.[125,126] In the case of pair additive potentials, it is possible to prove that the potentials exist and are unique,[46,127–129] although the existence and convergence properties of these iterative procedures have not been clearly determined for more complex systems.[130]

Izvekov and Voth have proposed an alternative force-based multiscale coarse-graining (MS-CG) method for approximating the many-body PMF.[36,59,131] Rather than attempting to reproduce a target set of distribution functions, the MS-CG approach employs force correlation functions determined from atomistic simulations to project the many-body mean force field (i.e., the force field determined from the PMF) onto a "basis set" determined by the form of the approximate CG potential.[41,60,62,72,132,133] These force correlation functions correspond to inner products of the mean force with the vectors in the basis set. The MS-CG force field is constructed to reproduce each inner product when evaluated using

configurations sampled from the original atomistic simulation. The resulting equations for the force field can be interpreted as relations that balance the approximate MS-CG force field and the exact many-body mean force along each basis vector.[63] The MS-CG metric tensor decomposes these projections into contributions from each interaction in the approximate CG potential.[60,62–64] By disentangling the direct and environment-mediated forces, the MS-CG method directly (i.e., noniteratively) determines a variationally optimal approximation to the many-body PMF, according to a force-based metric defined for the space of CG force fields.

Mullinax and Noid have demonstrated that the MS-CG force correlation functions can be directly determined from relatively simple structural correlation functions.[63,64,134] In fact, the normal MS-CG equations for the approximate potential are equivalent to a natural generalization of the Yvon-Born-Green (YBG) integral equation theory.[135] The conventional YBG equation can be interpreted as a statement of mechanical equilibrium that decomposes the average force on a pair of particles into a direct force between the pair and a correlated force arising from the surrounding environment. While the YBG theory has been previously extended for molecules with pair additive interactions,[136–141] the generalized-YBG (g-YBG) theory provides the first extension of this force balance relation for molecular mechanics force fields.[63,64] This g-YBG theory recovers the MS-CG approximation to the many-body PMF directly from structures, without recourse to force information, and employs the MS-CG metric tensor to address the correlations between different interactions that are neglected by straightforward Boltzmann inversion. Consequently, the g-YBG theory provides a natural link for investigating the relation between iterative structure-motivated methods and the direct force-motivated MS-CG method. More generally, this approach provides a natural perspective for quantifying the significance of many-body correlations in CG models.

The YBG equation, along with the Ornstein-Zernicke equation, is considered to be one of the fundamental relations for modeling and understanding simple liquids.[135] However, in contrast to the Ornstein-Zernicke theory, which has been extensively developed and applied for more complex molecules and polymers,[142–145] the YBG theory has seen relatively little development beyond the context of simple monatomic fluids.[136–141] Consequently, this framework also paves the way for analogous future development and application of the YBG integral equation theory for proteins, polymers, and other molecules with significant internal flexibility. Furthermore, in combination with a recently proposed extended ensemble framework,[146] the g-YBG theory provides a rigorous approach for elucidating and improving the approximations that are inherent to knowledge-based methods[147–149] for determining CG potentials from experimentally-determined structures.[150]

It should be noted that, because the PMF is defined for a particular system in a particular thermodynamic state point, CG potentials that have been optimized to reproduce the structural properties of a particular system in a particular thermodynamic state may not provide a satisfactory description for any other system or any other thermodynamic state. Many recent studies have investigated the transferability of structure-motivated CG models and some progress has been achieved.[40,70,71,125,126,146,151–157] Nevertheless, the transferability of such models is unclear a priori. Consequently, it is important to systematically and carefully characterize the transferability of structure-motivated CG models, including those developed by the g-YBG approach.

The present manuscript continues the development of the g-YBG theory and also develops a structurally accurate CG model for toluene, which is an important industrial solvent and closely related to the sidechain of polystyrene and various amino acids. The remainder of the manuscript is organized as follows: Section II reviews the g-YBG theory with special emphasis on the relationship between force and structural correlation functions. Section III provides relevant details of the following numerical calculations. Section IV clarifies the relationship between correlation functions characterizing forces and structures and presents a new CG model for toluene. The structural properties of this model are carefully compared with the original atomistic model. Additional calculations investigate the sensitivity of these results to the CG mapping, the force field basis set, and the temperature. Finally, Section V presents concluding comments.

2.2 Theory

2.2.1 Consistent Coarse-grained Models

The present work considers high and low resolution models of a molecular system. The high resolution model will be referred to as the "atomistic" model and its constituents will be referred to as "atoms." The low resolution model will be referred to as the "coarse-grained" (CG) model and its constituents will be referred to as "sites." The present formalism applies quite generally for relating high and low resolution particle-based models without rigid restraints. For clarity, lower and upper case symbols will be used for describing the atomistic and CG models, respectively. The presentation follows the development of Voth, Andersen, and coworkers[41,60] and subsequently extended by Mullinax and Noid.[63,146]

The configuration of the atomistic model is defined by the Cartesian coordinates, $\mathbf{r}$, for n atoms that interact according to a potential, $u(\mathbf{r})$, which may be of arbitrary complexity.

The atomistic configuration distribution, $p_r(\mathbf{r})$, is given by

$$p_r(\mathbf{r}) \propto \exp[-u(\mathbf{r})/k_BT]. \tag{2.1}$$

The configuration of the CG model is similarly defined by the Cartesian coordinates, $\mathbf{R}$, for N sites that interact according to a potential, $U(\mathbf{R})$. The CG configuration distribution, $P_R(\mathbf{R})$, is given by:

$$P_R(\mathbf{R}) \propto \exp[-U(\mathbf{R})/k_BT]. \tag{2.2}$$

A set of N mapping functions, $\mathbf{M}(\mathbf{r}) = \{\mathbf{M}_1(\mathbf{r}), \mathbf{M}_2(\mathbf{r}), \ldots, \mathbf{M}_N(\mathbf{r})\}$, determine the Cartesian coordinates of each site as a linear combination of coordinates for the atoms that are "involved" in the site:

$$\mathbf{M}_I(\mathbf{r}) = \sum_i c_{Ii}\mathbf{r}_i, \tag{2.3}$$

for $I = 1, \ldots, N$ and $\{c_{Ii}\}$ is a set of constants that satisfy the condition $\sum_i c_{Ii} = 1$. We will assume that each atom is involved in at most one site, although this assumption can be readily relaxed.[41]

The mapping, $\mathbf{M}$, and the atomistic probability distribution, p_r, determine the probability distribution, p_R, for sampling an atomistic configuration, $\mathbf{r}$, that maps to a fixed CG configuration, $\mathbf{R}$:

$$p_R(\mathbf{R}) = \langle \delta(\mathbf{M}(\mathbf{r}) - \mathbf{R}) \rangle , \tag{2.4}$$

where the angular brackets denote a canonical average according to $p_r(\mathbf{r})$ and $\delta(\mathbf{M}(\mathbf{r}) - \mathbf{R}) = \prod_I \delta(\mathbf{M}_I(\mathbf{r}) - \mathbf{R}_I)$. Equation (2.4) provides a natural, although certainly not unique, criteria for consistency[41] in configuration space between the models: a CG model is "consistent" (in configuration space) with a particular atomistic model if the canonical configurational distribution for the CG model, P_R, is equal to the probability distribution implied by the mapping and atomistic distribution, i.e.,

$$P_R(\mathbf{R}) = p_R(\mathbf{R}). \tag{2.5}$$

Equations (2.2), (2.4), and (2.5) imply that the appropriate potential for a consistent CG model is uniquely determined, to within an additive constant:

$$U^0(\mathbf{R}) = -k_BT \ln p_R(\mathbf{R}) + \text{const.} \tag{2.6}$$

The force field obtained from this potential, $U^0(\mathbf{R})$, may be expressed as a conditioned

canonical ensemble average of the atomistic forces evaluated for the atomistic configurations $\mathbf{r}$ that map to the given CG configuration $\mathbf{R}$:

$$\mathbf{F}_I^0(\mathbf{R}) = \langle \mathbf{f}_I(\mathbf{r}) \rangle_{\mathbf{R}} \,, \tag{2.7}$$

where $\mathbf{f}_I(\mathbf{r})$ is the net "atomistic force" on site I in configuration $\mathbf{r}$, and the subscripted angular brackets denote the conditioned canonical ensemble average

$$\langle a(\mathbf{r}) \rangle_{\mathbf{R}} = \langle a(\mathbf{r}) \delta(\mathbf{M}(\mathbf{r}) - \mathbf{R}) \rangle \,/\, \langle \delta(\mathbf{M}(\mathbf{r}) - \mathbf{R}) \rangle \,. \tag{2.8}$$

Because $\mathbf{F}_I^0$ is a mean force (MF), the potential, U^0, is referred to as the potential of mean force (PMF).[41,44] The PMF is not a conventional potential energy function, but should more properly be considered a configuration-dependent free energy that contains not only energetic, but also entropic effects arising from the distribution of configurations in Equation (2.4). The PMF is a many-body potential since it is determined from a many-body distribution function and, in general, it cannot be readily decomposed into simpler independent factors.[109,111] Consequently, CG methods typically approximate the PMF with relatively simple molecular mechanics type potentials.

2.2.2 Approximate Coarse-grained Force Fields

For the following analysis, it is convenient to consider an abstract vector space of CG force fields.[41,63,134] Each element in this space specifies a vector force on each site as a function of the CG configuration, $\mathbf{F} = \{\mathbf{F}_1(\mathbf{R}), \mathbf{F}_2(\mathbf{R}), \ldots, \mathbf{F}_N(\mathbf{R})\}$. An inner product $\odot$ can be defined between any two elements, $\mathbf{F}^{(1)}$ and $\mathbf{F}^{(2)}$, in this vector space:

$$\mathbf{F}^{(1)} \odot \mathbf{F}^{(2)} = \frac{1}{3N} \left\langle \sum_I \mathbf{F}_I^{(1)}(\mathbf{M}(\mathbf{r})) \cdot \mathbf{F}_I^{(2)}(\mathbf{M}(\mathbf{r})) \right\rangle , \tag{2.9}$$

and a corresponding norm, according to $||\mathbf{F}|| = (\mathbf{F} \odot \mathbf{F})^{1/2}$.

We consider approximate potentials of a molecular mechanics form:

$$U(\mathbf{R}) = \sum_\zeta \sum_\lambda U_\zeta(\psi_\zeta(\{\mathbf{R}\}_\lambda)), \tag{2.10}$$

where ζ indicates a particular interaction (e.g., a dihedral angle interaction) and U_ζ is the corresponding potential (e.g., a dihedral angle potential) that is a function of a single scalar variable, ψ_ζ, (e.g., a dihedral angle) that may be expressed as a function of the Cartesian

coordinates, $\{\mathbf{R}\}_\lambda$, for a set of sites, λ (e.g., the 4 successively bonded sites that form a dihedral angle).[60] Each term in Equation (2.10) is expanded as a linear combination of basis functions, $u_{\zeta d}$, with constant coefficients, $\phi_{\zeta d}$:

$$U_\zeta(x) = \sum_d \phi_{\zeta d} u_{\zeta d}(x). \tag{2.11}$$

The force on site I may be expressed:

$$\mathbf{F}_I(\mathbf{R}) = \sum_\zeta \sum_d \phi_{\zeta d}\, \mathcal{G}_{I;\zeta d}(\mathbf{R}), \tag{2.12}$$

where

$$\mathcal{G}_{I;\zeta d}(\mathbf{R}) = \sum_\lambda \frac{\partial \psi_{\zeta\lambda}(\mathbf{R})}{\partial \mathbf{R}_I} f_{\zeta d}(\psi_{\zeta\lambda}(\mathbf{R})), \tag{2.13}$$

$\psi_{\zeta\lambda}(\mathbf{R}) = \psi_\zeta(\{\mathbf{R}\}_\lambda)$, and $f_{\zeta d}(x) = -du_{\zeta d}(x)/dx$. The CG force field defined by Equation (2.12) then identifies a particular vector that can be re-expressed more simply:

$$\mathbf{F} = \sum_D \phi_D\, \boldsymbol{\mathcal{G}}_D, \tag{2.14}$$

where D is a "super index" that identifies a particular combination ζd.[60] Equation (2.14) explicitly expresses $\mathbf{F}$ as a linear combination of a set of vectors $\{\mathcal{G}_D\}$, each of which has elements given by Equation (2.12). The set of vectors included in Equation (2.14) defines an incomplete basis set that spans a subspace of the force field vector space.[41] The constants ϕ_D are both parameters for the CG potential and also coefficients that identify a particular vector in this subspace. The many-body MF defined in Equation (2.7) is an element in the vector space of CG force fields, but it is not necessarily in the subspace spanned by a given finite basis. However, because the MF is a conditional average of the atomistic force field, it follows that:

$$\mathbf{F}^0 \odot \mathbf{F} = \frac{1}{3N} \left\langle \sum_I \mathbf{f}_I(\mathbf{r}) \cdot \mathbf{F}_I(\mathbf{M}(\mathbf{r})) \right\rangle, \tag{2.15}$$

for any CG force field $\mathbf{F}$.[41]

2.2.3 Multiscale Coarse-graining Method

The MS-CG method[36,59] directly determines an approximate potential by minimizing the force-matching[158,159] functional, $\chi^2[\mathbf{F}]$:

$$\chi^2[\mathbf{F}] = \frac{1}{3N}\left\langle \sum_I |\mathbf{f}_I(\mathbf{r}) - \mathbf{F}_I(\mathbf{M}(\mathbf{r}))|^2 \right\rangle. \tag{2.16}$$

As a consequence of Equation (2.15), $\chi^2[\mathbf{F}]$ can be re-expressed:

$$\chi^2[\mathbf{F}] = \chi^2[\mathbf{F}^0] + ||\mathbf{F}^0 - \mathbf{F}||^2, \tag{2.17}$$

where $||\mathbf{F}^0 - \mathbf{F}||^2 \geq 0$, with equality holding if and only if $\mathbf{F} = \mathbf{F}^0$. Since $\chi^2[\mathbf{F}^0]$ is fixed by the atomistic model and CG mapping, $||\mathbf{F}^0 - \mathbf{F}||^2$ is minimized when $\chi^2[\mathbf{F}]$ is minimized. Consequently, if a complete basis set is employed, the MS-CG variational principle quantitatively determines the MF and the resulting CG model would quantitatively reproduce the many-body structural distribution defined by the atomistic model and the mapping. More generally, given an incomplete basis set that spans a subspace of force fields, the MS-CG procedure determines the force field in the subspace that is "closest" to the mean force, according to the norm defined above.[60] In this sense, the MS-CG method determines the force field within the subspace that provides the "optimal" approximation to the MF. Given the basis set expansion for the approximate CG force field, Equation (2.14), χ^2 becomes a quadratic function of the coefficients, $\phi \equiv \{\phi_D\}$, and the MS-CG force field can be determined as the solution to a simple linear least squares problem:[60,72]

$$\chi^2(\phi) = \chi^2(0) - 2\sum_D b_D\phi_D + \sum_D\sum_{D'} \phi_D G_{DD'}\phi_{D'}, \tag{2.18}$$

where

$$b_D = \frac{1}{3N}\left\langle \sum_I \mathbf{f}_I(\mathbf{r}) \cdot \boldsymbol{\mathscr{G}}_{I;D}(\mathbf{M}(\mathbf{r})) \right\rangle \tag{2.19}$$

$$G_{DD'} = \frac{1}{3N}\left\langle \sum_I \boldsymbol{\mathscr{G}}_{I;D}(\mathbf{M}(\mathbf{r})) \cdot \boldsymbol{\mathscr{G}}_{I;D'}(\mathbf{M}(\mathbf{r})) \right\rangle. \tag{2.20}$$

The coefficients minimizing χ^2 may be determined from the normal system of linear equations:[62]

$$\sum_{D'} G_{DD'}\phi_{D'} = b_D, \tag{2.21}$$

for each D. Notice that b_D reflects atomistic force information, while $G_{DD'}$ quantifies correlations between different interactions in mapped configurations. According to Equation (2.9), $G_{DD'}$ may be interpreted as a metric tensor corresponding to the inner product of basis vectors, i.e., $G_{DD'} = \mathcal{G}_D \odot \boldsymbol{\mathscr{G}}_{D'}$. Similarly, from Equations (2.9) and (2.15), it follows that b_D is equal to the inner product of the exact many-body MF and the basis vector $\mathcal{G}_D$, i.e., $b_D = \mathcal{G}_D \odot \mathbf{F}^0$. Consequently, it follows that the normal equations for the MS-CG approximate force field, $\mathbf{F}$, can be expressed:

$$\boldsymbol{\mathscr{G}}_D \odot \mathbf{F} = \boldsymbol{\mathscr{G}}_D \odot \mathbf{F}^0. \tag{2.22}$$

Equation (2.22) emphasizes that the normal MS-CG equations determine the projection of the MF[160–162] onto the subspace spanned by the incomplete basis set $\{\boldsymbol{\mathscr{G}}_D\}$ by requiring that the MS-CG force field and the MF have the same inner product with each basis vector.[60,63] Equations (2.19), (2.21), and (2.22) clarify the significance of atomistic force information in the MS-CG method. The force correlation functions, b_D, provide a clever and compact means for characterizing the many-body MF. However, Equations (2.6) and (2.7) demonstrate that the MF can be obtained directly from a configuration distribution function. This suggests that the MS-CG optimal approximation to the many-body PMF can be determined from appropriate structural information. The following subsection applies this line of reasoning.

2.2.4 Generalized Yvon-Born-Green Theory

To proceed further, it is convenient to reformulate the MS-CG framework in terms of a "continuous basis."[63,64] Starting from Equation (2.10), the force on site I can be re-expressed

$$\mathbf{F}_I(\mathbf{R}) = \sum_\zeta \int dx\ \phi_\zeta(x) \boldsymbol{\mathscr{G}}_{I;\zeta}(\mathbf{R}; x), \tag{2.23}$$

where $\phi_\zeta(x) = -dU_\zeta(x)/dx$ is a force function, and

$$\boldsymbol{\mathscr{G}}_{I;\zeta}(\mathbf{R}; x) = \sum_\lambda \frac{\partial \psi_{\zeta\lambda}(\mathbf{R})}{\partial \mathbf{R}_I} \delta(\psi_{\zeta\lambda}(\mathbf{R}) - x). \tag{2.24}$$

The MS-CG functional becomes

$$\begin{aligned}\chi^2[\{\phi_\zeta(x)\}] &= \chi^2[\{0\}] - 2\sum_\zeta \int dx\, b_\zeta(x)\phi_\zeta(x) \\ &+ \sum_\zeta \sum_{\zeta'} \int dx \int dx' \phi_\zeta(x) G_{\zeta\zeta'}(x,x')\phi_{\zeta'}(x'), \end{aligned} \tag{2.25}$$

where

$$b_\zeta(x) = \frac{1}{3N}\left\langle \sum_I \mathbf{f}_I(\mathbf{r}) \cdot \boldsymbol{\mathscr{G}}_{I;\zeta}(\mathbf{M}(\mathbf{r});x) \right\rangle \tag{2.26}$$

$$G_{\zeta\zeta'}(x,x') = \frac{1}{3N}\left\langle \sum_I \boldsymbol{\mathscr{G}}_{I;\zeta}(\mathbf{M}(\mathbf{r});x) \cdot \boldsymbol{\mathscr{G}}_{I;\zeta'}(\mathbf{M}(\mathbf{r});x') \right\rangle, \tag{2.27}$$

which, as before, can be interpreted as inner products of the basis vector $\boldsymbol{\mathscr{G}}_\zeta(x)$ with the MF and with $\boldsymbol{\mathscr{G}}_{\zeta'}(x')$, respectively. The normal equations are then expressed

$$\sum_{\zeta'} \int dx' G_{\zeta\zeta'}(x,x')\phi_{\zeta'}(x') = b_\zeta(x). \tag{2.28}$$

By employing Equations (2.6) and (2.7), performing integration by parts, and keeping track of appropriate Jacobian factors,[163–165] $b_\zeta(x) = \boldsymbol{\mathscr{G}}_\zeta(x) \odot \mathbf{F}^0$ can be expressed in terms of structural correlation functions:

$$b_\zeta(x) = k_B T \left(\frac{d}{dx}\bar{g}_\zeta(x) - L_\zeta(x) \right), \tag{2.29}$$

where

$$\bar{g}_\zeta(x) = \frac{1}{3N}\left\langle \sum_\lambda |\nabla\psi_{\zeta\lambda}(\mathbf{M}(\mathbf{r}))|^2 \delta(\psi_{\zeta\lambda}(\mathbf{M}(\mathbf{r})) - x) \right\rangle \tag{2.30}$$

$$L_\zeta(x) = \frac{1}{3N}\left\langle \sum_\lambda \nabla^2\psi_{\zeta\lambda}(\mathbf{M}(\mathbf{r}))\, \delta(\psi_{\zeta\lambda}(\mathbf{M}(\mathbf{r})) - x) \right\rangle, \tag{2.31}$$

$|\nabla\psi_{\zeta\lambda}(\mathbf{R})|^2 = \sum_I (\partial\psi_{\zeta\lambda}(\mathbf{R})/\partial\mathbf{R}_I)^2$, and $\nabla^2\psi_{\zeta\lambda}(\mathbf{R}) = \sum_I \partial^2\psi_{\zeta\lambda}(\mathbf{R})/\partial\mathbf{R}_I^2$. The generalized-YBG (g-YBG) equation[63] then determines the MS-CG force field directly from structural correlation functions:

$$\sum_{\zeta'} \int dx' G_{\zeta\zeta'}(x,x')\phi_{\zeta'}(x') = k_B T \left(\frac{d}{dx}\bar{g}_\zeta(x) - L_\zeta(x) \right). \tag{2.32}$$

By decomposing the metric factor, $G_{\zeta\zeta'}(x,x') = \delta_{\zeta\zeta'}\bar{g}_\zeta(x)\delta(x-x') + \bar{G}_{\zeta\zeta'}(x,x')$, Equation (2.32) separates the direct and indirect contributions to $b_\zeta(x)$:

$$\begin{aligned} \phi_\zeta(x) &+ \sum_{\zeta'} \int dx' \bar{g}_\zeta^{-1}(x)\bar{G}_{\zeta\zeta'}(x,x')\phi_{\zeta'}(x') \\ &= k_B T \left(\frac{d}{dx} \ln \bar{g}_\zeta(x) - \bar{g}_\zeta^{-1}(x) L_\zeta(x) \right). \end{aligned} \tag{2.33}$$

This result applies quite generally for any molecular mechanics force field and, in particular, for polymer models with angle and torsion potentials.

When applied to simple liquids, Equation (2.33) reduces to the YBG integral equation theory.[135] For a liquid of CG sites interacting via simple pair potentials, $U(\mathbf{R}) = \sum_\lambda U^{(2)}(\psi_\lambda(\mathbf{R}))$; λ identifies a particular pair of sites $\{I,J\}$; and $\psi_\lambda(\mathbf{R}) = \mid \mathbf{R}_I - \mathbf{R}_J \mid$ is the distance between the pair. Then $\mid \nabla\psi_\lambda(\mathbf{R}) \mid^2 = 2$, $\nabla^2\psi_\lambda(\mathbf{R}) = 4/\psi_\lambda(\mathbf{R})$, and the components of $b(x)$ simplify to:

$$\bar{g}(x) = cx^2 g(x) \tag{2.34}$$

$$L(x) = 2cxg(x) \tag{2.35}$$

in terms of $c = \frac{4\pi}{3}\rho$, the density $\rho = N/V$, and the conventional radial distribution function (RDF), $g(x)$. Equation (2.29) becomes $b(x) = k_B T c x^2 dg(x)/dx$, or, in terms of the pair potential of mean force,[135,166] $w(x) = -k_B T \ln g(x)$,

$$-w'(x) = \frac{b(x)}{cx^2 g(x)}. \tag{2.36}$$

In this case, Equation (2.33) may be re-expressed:

$$\phi(x) + \int dx' \frac{1}{cx^2 g(x)} \bar{G}(x,x')\ \phi(x') = -w'(x), \tag{2.37}$$

which quite simply asserts that the mean force, $-w'(x)$, on a given particle when there is a second particle a distance x away has two contributions:[135,166] 1) the direct force, $\phi(x)$, from the second particle; and 2) the correlated net force from the environment, which is decomposed into contributions from shells of particles at a distance x' away from the first particle. It should be noted that $\bar{G}(x,x')$ incorporates the vectorial nature of this correlated force, which, by symmetry, is aligned along the vector between the first two particles.

2.3 Methods

2.3.1 Simulation Details

All atomistic and CG simulations were performed with the Gromacs 4.0.7 simulation suite.[7,167] The Gromacs stochastic dynamics algorithm and Parrinello-Rahman barostat[168] were employed to sample the constant NVT and constant NPT ensembles, respectively. The Nose-Hoover thermostat[169,170] and the Berendsen weak-coupling thermostat and barostat[171] were employed in equilibration stages. The interactions in the atomistic model were determined from the OPLS-AA force field[18] and all bonds were flexible. Electrostatic interactions were calculated with the particle mesh Ewald method[172] and periodic boundary conditions[45] were employed in all simulations. Short-ranged van der Waals interactions and also the real space contribution to the electrostatic interactions were truncated at 1.4 nm. A 1 fs integration time step was used in both atomistic and CG MD simulations.

An atomistic model of 216 toluene molecules was simulated in the canonical ensemble at 273, 298, and 373 K. These canonical simulations were performed in a volume corresponding to the equilibrium density of the OPLS-AA toluene model at the given temperature and atmospheric pressure. The starting configuration for each canonical simulation was obtained by initially heating a lattice of toluene molecules at constant volume to 1000 K, slowly cooling the system to the target temperature, equilibrating the system volume at atmospheric pressure, and finally sampling a configuration corresponding to the average volume. The equilibrium densities of the OPLS-AA model at 1 bar pressure and temperatures of 273, 298, and 378 K were 889.3, 864.4, and 778.6 kg/m^3, respectively. These densities agree quite accurately with the experimental densities of 885.6, 862.3, and 790.1 kg/m,^{3} respectively.[173] The atomistic model was then simulated for 22 ns at each temperature. Configurations and forces were sampled every 1 ps during the last 20 ns of these simulations. These configurations and forces were mapped according to the CG representation of toluene and used to evaluate the relevant force and structural correlation functions. The atomistic simulations were extended an additional 20 ns to better sample orientational correlation functions.

The calculated CG potentials were employed in simulations sampling the canonical ensemble at 273, 298, and 398 K. Each CG system was equilibrated for 2 ns in the volume determined from corresponding atomistic simulations, after which configurations were sampled in the constant NVT ensemble every 1 ps. The resulting inter and intra-molecular correlation functions were compared with those obtained by applying the CG mapping to configurations sampled from atomistic simulations.

2.3.2 CG Mapping

As shown in Figure 2.1, the present work considers two different three site representations for toluene. Both mappings employed three sites in order to capture the planar geometry of toluene, which is expected to be essential for characterizing and accurately reproducing intermolecular packing and alignment. These sites were associated with atomic groups based upon considerations of molecular symmetry, shape, and size. Both mappings employ two equivalent CB sites that represent single C atoms, C_3 and C_5. The two mappings differ in the treatment of the third distinct site, CF. The first mapping, which is shown in Figure 2.1a and which will be referred to as the methyl mapping, defines the mapped CF coordinates as the Cartesian coordinates of the C_7 atom. The second mapping, which is shown in Figure 2.1b and which will be referred to as the COM mapping, defines the mapped CF coordinates as the center of mass for the group of atoms including C_1, C_7, and the associated hydrogen atoms. Although these two CG representations are both relatively high resolution, they are consistent with several previous CG studies that employed either three[28,31] or four sites[100] to model planar aromatic groups.

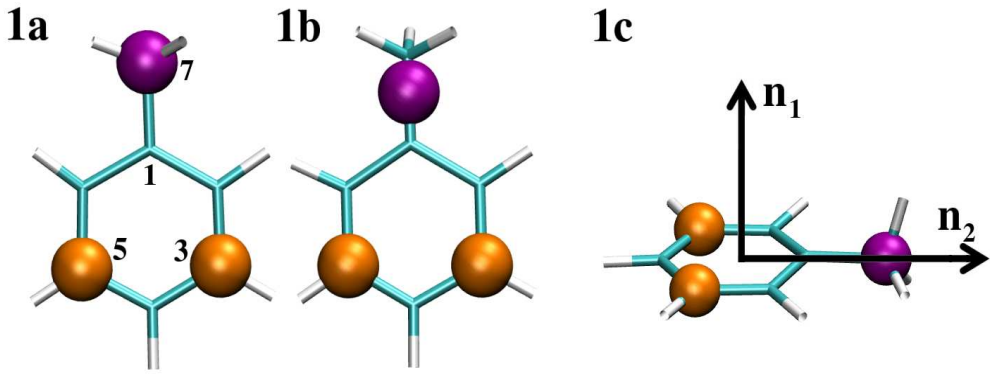

Figure 2.1. Representation of CG toluene molecules. Figures 2.1a and 2.1b present the methyl and COM mapping, respectively. Figure 2.1c presents the molecular directors employed to characterize the packing of molecules in the atomistic and CG models. These images were made with VMD.[174]

Figure 2.1c also defines the two molecular directors that will be employed in characterizing the alignment of toluene molecules in the atomistic and CG models. The first molecular director, $\hat{\mathbf{n}}_1$, is normal to the plane of the toluene molecule. The second molecular director, $\hat{\mathbf{n}}_2$, corresponds to a bisector of the CB-CF-CB bond angle and is defined as the vector pointing from the midpoint of the CB-CB bond to the CF site. These two directors approximately correspond to the first and third principle axes defined by the molecular inertia tensor.

2.3.3 Force Field Basis Set

These two different mappings determine two distinct models and also distinct many-body potentials of mean force. The force field for each CG model was calculated by projecting the corresponding many-body MF onto the force field basis set. Both models employed the same basis set, which was defined by the form of the approximate potential. In each case, the approximate potential included intramolecular bond stretch potentials between each pair of bonded sites and also short-ranged nonbonded potentials between each pair of sites in distinct molecules. These nonbonded pair potentials were truncated at 1.4 nm. The CG potential did not include explicit electrostatic interactions or angle-dependent interactions, although they may be readily treated in the present framework.[60,64,134] The majority of the calculations reported below employed linear spline functions, i.e., piecewise linear functions, to represent each term in the potential. Grid spacings of 0.001 nm and 0.01 nm were employed for bond stretch and nonbonded pair potentials, respectively.

In order to investigate the variational properties of the method, the CG force field was also calculated using a more restricted 'analytic' basis set. The calculations presented below (row 2 of Figure 2.2 and column 2 of Figure 2.5) for the restricted analytic basis set employed harmonic functions, i.e., $U_\zeta(x) = \phi_{\zeta 1}x + \phi_{\zeta 2}x^2$, to model each bond stretch and Lennard-Jones-type 12-6 functions, i.e., $U_\zeta(x) = \phi_{\zeta 1}/x^6 + \phi_{\zeta 2}/x^{12}$, to model each nonbonded pair interaction, where $\phi_{\zeta 1}$ and $\phi_{\zeta 2}$ may be either positive or negative.

2.3.4 Force Field Calculation

For each force field basis set and each mapping considered, the parameters for the MS-CG force field were calculated by solving the associated set of normal equations given by Equation (2.21). In each case, the matrix elements $G_{DD'}$ were calculated after mapping configurations sampled from all-atom MD simulations. The vector b_D was calculated according to the MS-CG method by evaluating the force correlation function in Equation (2.19). The vector b_D was also calculated from structural correlation functions according to the g-YBG method using either Equation (2.29) or a discretized version of this equation. The normal equations were solved via LU decomposition after applying left preconditioning.[175] In the following calculations, the normal equations were sufficiently well conditioned that LU decomposition was quite stable. Singular value decomposition may prove a useful alternative for systems that lead to less well conditioned normal equations. The stability of these equations reflects several considerations including the configurations sampled by the atomistic model, the CG representation of those configurations, and the force field basis set. These

numerical issues have been briefly addressed in previous studies of the MS-CG method[60,72] and remain beyond the scope of the present study. The calculated nonbonded force functions were smoothed by a running average over three consecutive grid points and then integrated to determine corresponding potentials.

2.4 Results and Discussion

Both the MS-CG and g-YBG approaches determine a CG force field by projecting the many-body MF onto an incomplete basis set.[41,63] The MS-CG method employs force correlation functions sampled from atomistic simulations to evaluate the inner product of the MF with each basis vector. In contrast, the g-YBG method determines these same inner products from structural correlation functions. The present work employs the g-YBG approach to develop a three site CG model for toluene. Figures 2.2 and 2.3 demonstrate the g-YBG procedure for determining the force field inner products. Figures 2.4-2.6 assess the structural accuracy of the resulting model at 298 K and also investigate the sensitivity of the results to the mapping and the basis set. Figures 2.7 and 2.8 investigate the transferability of the model for simulations in the canonical ensemble at 273 K and 373 K. With the exception of the second row of Figure 2.2 and the second column of Figure 2.5, which consider the reduced analytic basis set, all calculations employed linear spline basis functions with a grid spacing of 0.001 nm for bond stretch potentials and 0.01 nm for nonbonded pair potentials.

2.4.1 Accuracy of the G-YBG Approach for Calculating the MS-CG Potential

Figure 2.2 presents calculations of b_ζ (top panels) and ϕ_ζ (bottom panels) for the intramolecular CF-CB bond stretch interaction (left panels) and for the intermolecular CF-CB interaction (right panels) when considering the methyl mapping. Calculations employing forces are presented as solid black curves, while calculations employing structures are presented as dashed cyan curves. To reduce statistical noise in the figure, the results for the nonbonded interaction have been smoothed by a running average over three consecutive grid points. Nevertheless, despite the considerable noise in b_ζ for the nonbonded interaction (Figure 2.2b1), it is clear that the g-YBG approach determines these force correlation functions with essentially quantitative accuracy. Figures 2.2a2 and 2.2b2 demonstrate that, as a direct consequence of accurately recovering b_ζ, the g-YBG approach also quantitatively recovers the MS-CG force functions, ϕ_ζ. The two calculations for the CF-CB bond interaction both ob-

tain a linear restoring force and quantitatively agree over the range of bond lengths that are reasonably sampled (see Figure 2.3a1 below). The two calculations also agree with near quantitative accuracy for the CF-CB nonbonded force function. Statistical discrepancies in the calculations of b_ζ result in small shifts between the calculated force functions around $r \approx 0.75$ nm.

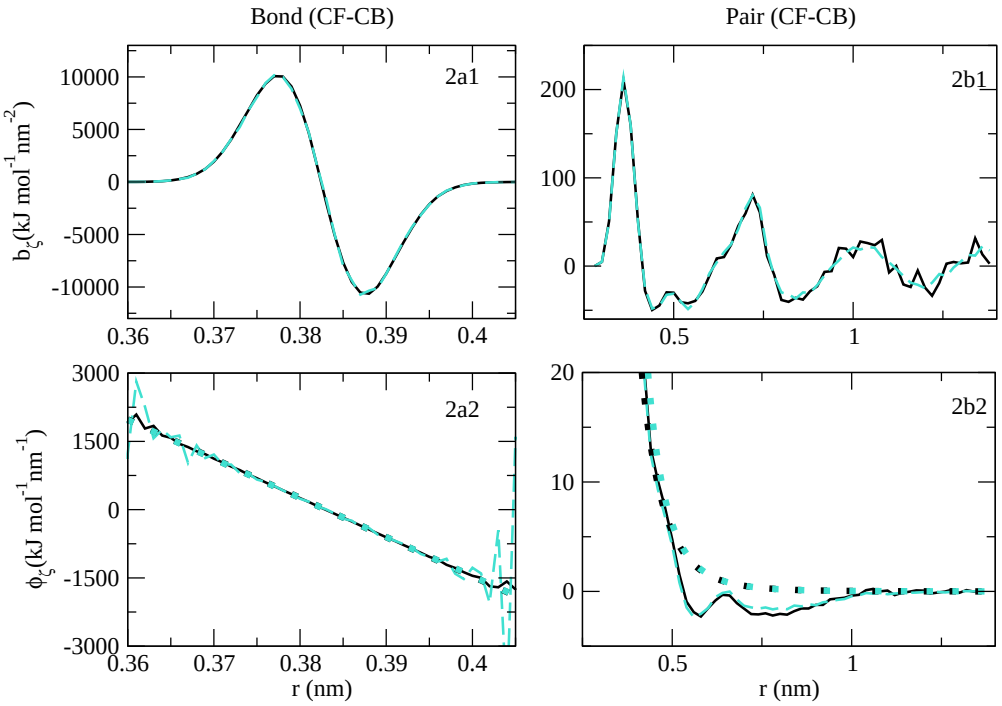

Figure 2.2. Comparison of the force correlation function $b_\zeta(x)$ (top panels) and corresponding calculated force function $\phi_\zeta(x)$ (bottom panels) for the CF-CB bond stretch interaction (left panel) and the CF-CB nonbonded interaction (right panel) when considering the methyl mapping (Figure 2.1a). Solid black curves correspond to calculations employing forces according to the MS-CG method. Dashed cyan curves correspond to calculations employing structures according to the g-YBG method. Figures 2.2a2 and 2.2b2 also present the bond stretch and nonbonded pair force functions calculated using harmonic and Lennard-Jones basis functions, respectively, when using forces (black blocks) and structures (cyan blocks).

Figures 2.2a2 and 2.2b2 also present the force functions that are calculated according to the MS-CG (black blocks) and g-YBG (cyan blocks) approaches when using the restricted analytic basis set, i.e., harmonic and Lennard-Jones basis functions for the bond and nonbonded potentials, respectively. Because the larger basis set determined a linear force function for the bond stretch and because this interaction is only weakly coupled to the nonbonded interactions, the calculations with the two basis sets agree quantitatively for the bond force

function. In contrast, by reducing the basis set for the nonbonded pair interaction from a flexible linear spline to a Lennard-Jones function, the resulting nonbonded force function changes significantly. The weakly attractive double well that is present in the linear spline calculation is replaced by a purely repulsive potential, as might be expected on the basis of Weeks-Chandler-Andersen theory.[47] As was the case for the linear spline basis set, the g-YBG calculations accurately recover the MS-CG force functions for the analytic basis set over the range of distances that are accurately sampled in the atomistic simulations.

In summary, Figure 2.2 clearly validates the g-YBG approach as an accurate approach for determining the MS-CG force field directly from structural information.[64,134]

2.4.2 Relating Force and Structural Correlation Functions

Figure 2.3 explicitly illustrates the relationship between force and structural correlation functions for the CF-CB bond stretch (left panels) and the CF-CB nonbonded interaction (right panels) when considering the methyl mapping. The solid black curves in Figures 2.3a1 and 2.3b1 present the bond stretch distribution and radial distribution function (RDF), respectively. The dashed cyan curves in Figures 2.3a1 and 2.3b1 present the potentials of mean force, $w_\zeta(x)$, that are obtained by performing direct Boltzmann inversion after normalizing the corresponding distribution by the appropriate Jacobian.[112] In each case, $w_\zeta(x)$ demonstrates attractive wells (repulsive barriers) at maxima (minima) in the corresponding distribution and diverges as the associated probability distribution vanishes.

Figures 2.3a2 and 2.3b2 compare, for each interaction, two calculations of the mean force function, $-w'_\zeta(x)$. The solid black curve presents the mean force calculated by numerically differentiating the corresponding potential of mean force from Figures 2.3a1 and 2.3b1. The dashed cyan curve presents the mean force calculated by normalizing the force correlation function, $b_\zeta(x)$, with the corresponding probability distribution function according to Equation (2.36). The two calculations agree quantitatively and validate the g-YBG identity for the MS-CG force correlation functions. As expected, the mean force functions possess the following properties: 1) they vanish at each local maxima (and local minima) in the corresponding distribution function; 2) they correspond to net forces driving each interaction towards these maxima; 3) they clearly reflect the structure present in the force correlation functions from Figures 2.2a1 and 2.2b1.

Figures 2.3a2 and 2.3b2 also compare these mean force functions with the calculated g-YBG force functions that provide the optimal approximation to the many-body MF (dashed-dotted black curves). As discussed above and demonstrated by Equation (2.22), the calcu-

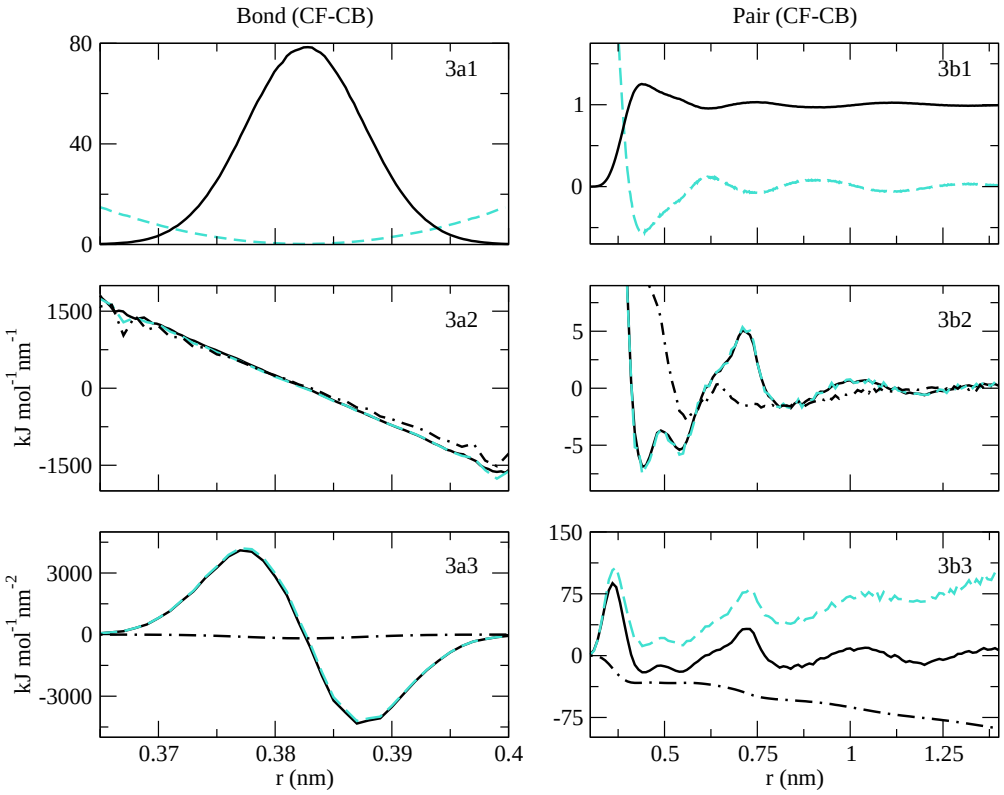

Figure 2.3. Analysis of $b_\zeta(x)$ for the CF-CB bond stretch interaction (left panels) and the CF-CB pair nonbonded interaction (right panels) when considering the methyl mapping (Figure 2.1a). The top row presents the bond stretch distribution and nonbonded radial distribution functions (solid black) and the corresponding potentials of mean force (dashed cyan). The middle row presents the corresponding mean forces calculated by numerical differentiation (solid black) and also by Equation (2.36) in terms of the force correlation function, $b_\zeta(x)$ (dashed cyan). The middle row also presents the calculated force, ϕ_ζ, for each interaction (dashed-dotted black). The bottom row compares the contributions of $k_BTd\bar{g}_\zeta(x)/dx$ (dashed cyan) and $-k_BTL_\zeta(x)$ (dashed-dotted black) to the force correlation functions, $b_\zeta(x)$ (solid black).

lated force functions, ϕ_ζ, balance the mean force along each basis vector. In the case of the CF-CB bond interaction, the spring constants for the mean and direct forces differ by approximately 12%, which demonstrates that CG bonded potentials do couple to other interactions. In the case of the intermolecular CF-CB interaction, the direct force function deviates even more dramatically from the mean force. In particular, the hard wall and first minima of the mean force, which reflect the initial rise and first maximum of the RDF, occur at significantly shorter distances than the hard wall and the first minima of the calculated direct force. Consequently, while the mean force function indicates a net attraction between the pair when they are separated by 0.4 nm $\leq r \leq$ 0.5 nm, the calculated direct force function is repulsive over this range. Similarly, while the mean force function indicates a net repulsion between the pair when they are separated by $r \approx 0.72$ nm, at this distance

the calculated direct force function is attractive. These differences between the mean force and the corresponding direct force reflect the correlated forces from the environment. In particular, preliminary calculations indicate that these differences between the direct and mean force on CF-CB pairs arise primarily from coupling to the CF-CB bond stretch and, to a lesser extent, from coupling to the CB-CB bond stretch and to nonbonded interactions. The origin and physical significance of these correlations require further study and may be considered in future investigations.

Finally, Figures 2.3a3 and 2.3b3 decompose each force correlation function $b_\zeta(x)$ (from Figures 2.2a1 and 2.2b1) according to Equation (2.29) into contributions from $k_BTd\bar{g}_\zeta(x)/dx$ (dashed cyan) and from $-k_BTL_\zeta(x)$ (dashed-dotted black). Because the bond stretch distribution varies rapidly over a very narrow distance range, the derivative term dominates b_ζ for the bond stretch interaction. In contrast, although the derivative term determines the fine structure in the nonbonded force correlation function, both terms make significant contributions.

In summary, Figure 2.3 demonstrates that the force correlation function, b_ζ, quantifies the mean net force driving each interaction to a local equilibrium. This mean force reflects both direct and environment-mediated contributions, but can be quantitatively determined directly from structural correlation functions.

2.4.3 Structural Accuracy of the G-YBG Model

Having demonstrated that the g-YBG approach quantitatively determines the MS-CG approximation to the many-body PMF, Figures 2.4-2.6 assess the structural accuracy of two distinct three site g-YBG models for toluene. In Figures 2.4 and 2.5, the left column presents results for the CG model employing the methyl mapping (shown in Figure 2.1a), while the right column presents results for the CG model employing the COM mapping (shown in Figure 2.1b). When considering Figures 2.4-2.6, it should be noted that the differences in the atomistic distribution functions for the two models reflect differences in the mapping applied to the same set of atomically-detailed configurations, while the differences in the CG distribution functions reflect differences in the configurations sampled with two distinct CG potentials that approximate different PMFs.

Figure 2.4 compares the intramolecular bond stretch (top panels) and angle distributions (bottom panels) for the atomistic (solid black curves) and CG models (dashed cyan curves). The COM mapping moves the CF site nearer to the aromatic ring and results in an equilibrium geometry that is significantly closer to an equilateral triangle. The COM mapping

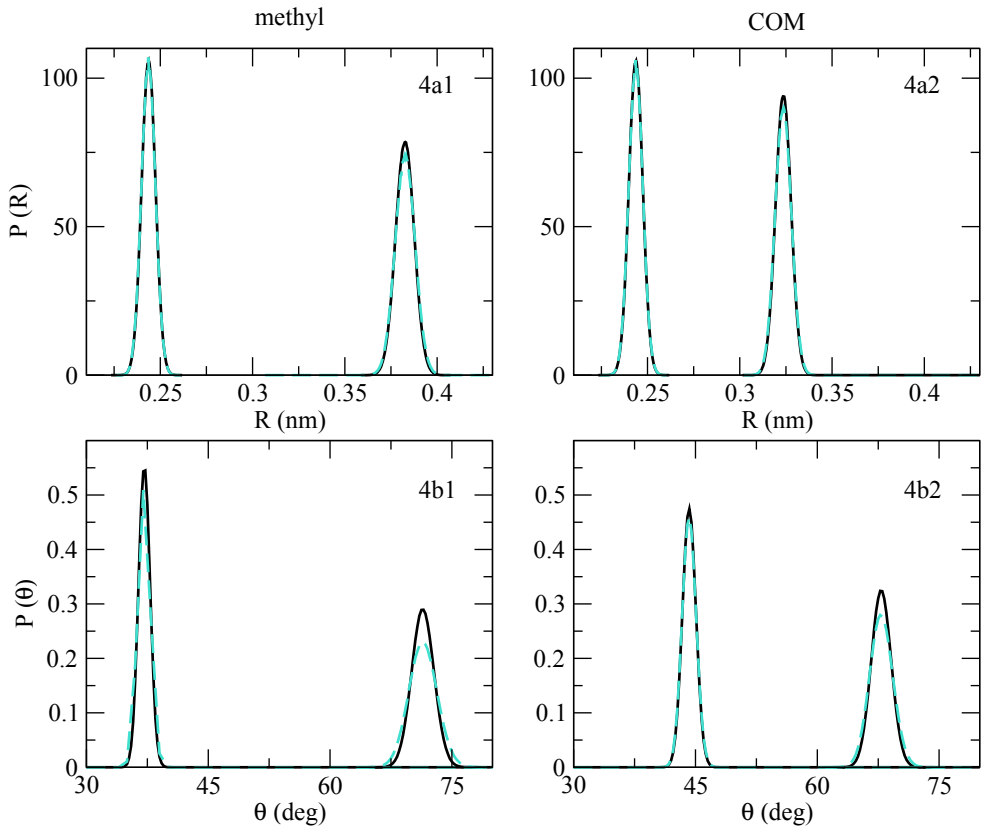

Figure 2.4. Comparison of the intramolecular distribution functions obtained from all-atom (solid black) and CG (dashed cyan) simulations. Row a presents the CB-CB and CF-CB bond stretch distributions at smaller and larger R, respectively. Row b presents the CB-CF-CB and CB-CB-CF angle distributions at smaller and larger θ, respectively. Columns 1 and 2 present results for the methyl and the COM mapping, respectively.

shortens the CF-CB bond by approximately 0.06 nm and also generates a slightly narrower distribution for this bond stretch. Figures 2.4a1 and 2.4a2 demonstrate that both CG models reproduce the corresponding bond stretch distributions with quantitative accuracy. Figures 2.4b1 and 2.4b2 demonstrate that both CG models also reproduce the CB-CF-CB angle distribution with near quantitative accuracy, though the distribution of CB-CB-CF angles is less sharply peaked in the CG models than in the atomistic model. Although both models reproduce the intramolecular toluene structure quite accurately, Figure 2.4 demonstrates that the COM model gives a slightly more accurate description than the methyl model.

Figure 2.5 compares the intermolecular site-site RDFs of the OPLS-AA (solid black curves) and CG models (dashed cyan curves) for liquid toluene. Columns 1 and 3 present results for the methyl and COM models, respectively, and employ the linear spline basis set. The atomistic CB-CB RDF (top row) demonstrates relatively little structure, lacks a sharp

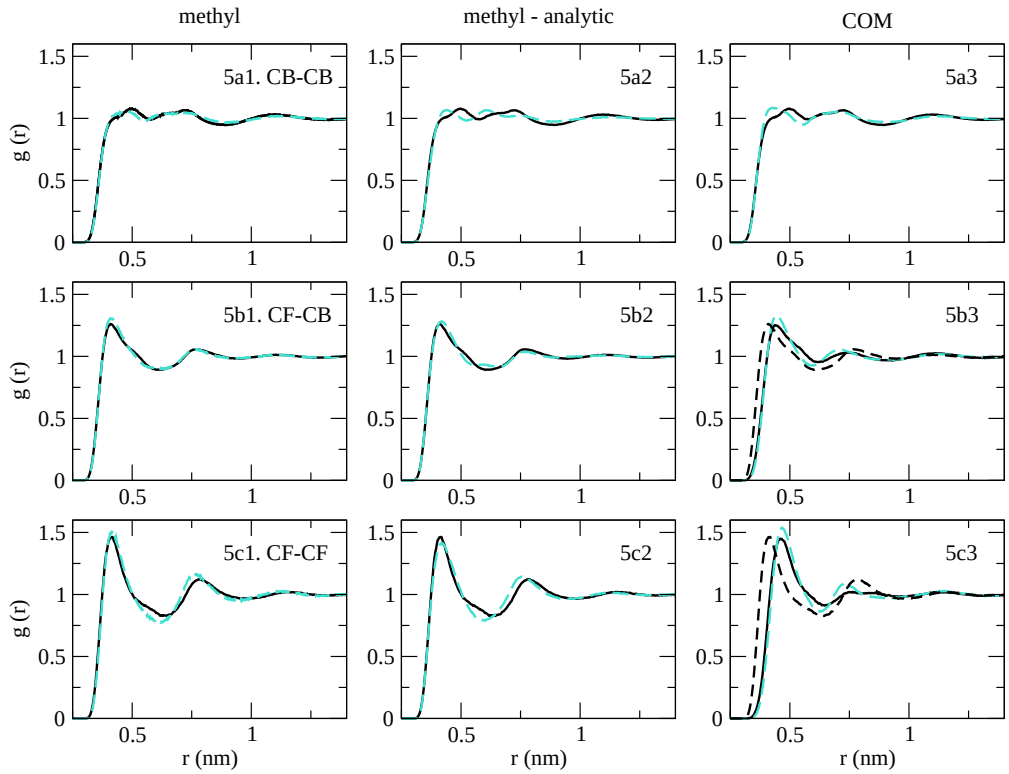

Figure 2.5. Comparison of site-site RDFs from all-atom (solid black) and CG (dashed cyan) simulations at 298 K for the CB-CB, CF-CB, and CF-CF pairs in rows a, b, and c, respectively. Columns 1 and 3 present the results for the methyl and the COM mapping, respectively, when using the linear spline basis set. Column 2 presents the results for the methyl model when using the reduced analytic basis set. Figures 2.5b3 and 2.5c3 also present atomistic RDFs for the methyl mapping as the dashed black curves.

first solvation shell peak, and instead presents a doublet-like split peak corresponding to the two CB sites in each molecule. The atomistic CF-CB RDF (middle row) demonstrates considerably more structure with significant first and second peaks. Figure 2.5b3 compares the atomistic CF-CB RDF for the methyl and COM mappings as the dashed and solid black curves, respectively. Because the COM mapping moves the CF site nearer to the aromatic ring and farther from the molecular edge, the corresponding CF-CB RDF moves the first peak to larger separation, shifts the second peak to smaller separation, and demonstrates slightly less structure relative to the CF-CB RDF for the methyl mapping. The CF-CF RDF (bottom row) continues the trend of increasing structure and also the trend of increasing differences between the two atomistic RDFs, as shown in Figure 2.5c3. Figure 2.5 demonstrates that both CG models reproduce the atomistic site-site RDFs quite accurately. With the exception of slight overstructuring in the second solvation shell of CF-CF site pairs, the methyl model nearly quantitatively reproduces the corresponding atomistic RDFs. The COM model

reproduces the corresponding atomistic RDFs with slightly less accuracy, but nevertheless clearly distinguishes between the atomistic RDFs obtained for the two mappings. The quantitative agreement between the atomistic and CG RDFs in Figure 2.5 reflects the accuracy with which the (very weakly attractive) calculated potentials reproduce the attractive minima in the corresponding pair potentials of mean force. (See Figure 2.3b2.)

Finally, column 2 of Figure 2.5 compares the atomistic and CG site-site RDFs for the methyl mapping when employing the reduced analytic basis set defined by harmonic bond potentials and Lennard-Jones-type 12-6 nonbonded potentials. Comparison of the first and second columns reveals that, for the given CG mapping, the CG force field employing the reduced analytic basis set less accurately reproduces the site-site RDFs determined by the atomistic models. However, because the 12-6 potential quite accurately captures the hard wall repulsion of the CG sites shown in Figure 2.2b2, the corresponding 12-6 basis set also provides a satisfactory representation of the atomistic RDFs. These results further reinforce the conclusion that the hard wall of the CG potential is most significant for reproducing atomistic site-site RDFs, as would be expected from Weeks-Chandler-Andersen theory.[47] Moreover, they also demonstrate the variational nature of the g-YBG calculations, i.e., for a given CG mapping, expanding the force field basis set systematically improves the MS-CG approximation to the many-body MF.[41,63]

Figure 2.6 compares the packing of toluene molecules in the atomistic model (solid curves) and in the two CG models (dashed curves), corresponding to the methyl (black curves) and COM (cyan curves) mapping, as a function of the distance between molecular centers of geometry. In contrast to Figure 2.5, the two mappings of the atomistic trajectory present almost identical results, but the results for the two CG models differ significantly. Figure 2.6a presents the intermolecular RDF between the centers of geometry for toluene molecules. Although Figure 2.6 demonstrates that the site-site RDFs are slightly more structured in the CG model, Figure 2.6a demonstrates that molecules are better packed into solvation shells in the atomistic model. Nevertheless, the COM and, to a lesser extent, also the methyl CG model reproduces the intermolecular RDF with reasonable accuracy.

Figure 2.6b presents the average of the second Legendre polynomial describing the alignment of toluene molecules along the first molecular director, $\hat{\mathbf{n}}_1$, (shown in Figure 2.1c) as a function of intermolecular separation. Figure 2.6b demonstrates that, in the atomistic model, the planes of toluene molecules are almost perfectly parallel at short distances. This alignment rapidly decays to zero by the first solvation shell and then demonstrates a second maxima near the minimum in the molecular RDF at 0.8 nm. The COM model nearly quantitatively reproduces this alignment. Perhaps as a consequence of the greater distance

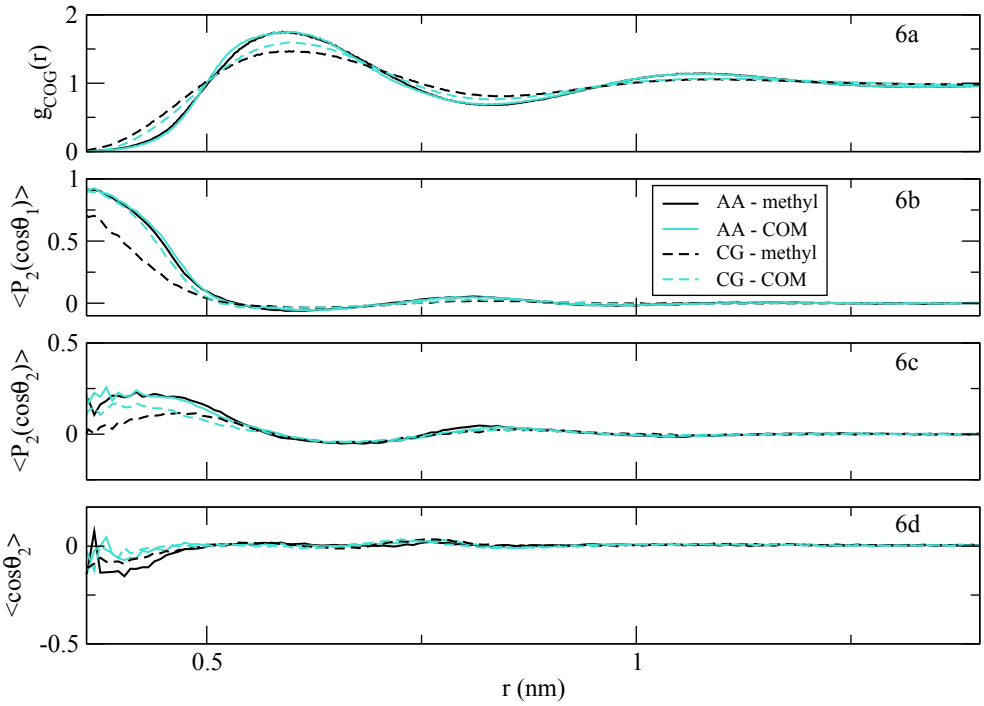

Figure 2.6. Characterization of molecular packing observed in all-atom (solid curves) and CG (dashed curves) simulations of toluene when using the methyl (black curves) and COM (cyan curves) mapping. Figure 2.6a presents the RDFs for the molecular centers of geometry. Figures 2.6b and 2.6c present the average of the second Legendre polynomial, $\langle P_2(\cos\theta)\rangle = 3/2\left\langle\cos^2\theta\right\rangle - 1/2$, for the angles, θ_1 and θ_2, formed by the molecular directors $\hat{\mathbf{n}}_1$ and $\hat{\mathbf{n}}_2$, respectively, as a function of intermolecular separation, r. Figure 2.6d presents the average, $\langle\cos\theta_2\rangle$, describing the alignment of the second molecular directors as a function of intermolecular separation. See Figure 2.1c for the definition of the molecular directors.

between the CF and CB sites, though, the methyl model only qualitatively reproduces the stacking of toluene molecules at short distances. Figure 2.6c presents the average of the second Legendre polynomial describing the alignment of toluene molecules along the second molecular director, $\hat{\mathbf{n}}_2$, shown in Figure 2.1c, as a function of intermolecular separation. In the atomistic model, toluene molecules demonstrate a slight tendency to be either aligned or anti-aligned at very short separations. This alignment decays on a slightly longer length scale than the alignment along $\hat{\mathbf{n}}_1$. The COM mapping, and to a lesser extent the methyl mapping, qualitatively reproduce this alignment. Figure 2.6d presents the average of the first Legendre polynomial describing alignment along $\hat{\mathbf{n}}_2$ as a function of intermolecular separation. Figure 2.6d demonstrates that, the atomistic model has a very slight preference for toluene molecules to be anti-aligned rather than aligned along $\hat{\mathbf{n}}_2$ at very short separations. Both CG models reproduce this preference with reasonable accuracy.

In summary, Figures 2.4-2.6 demonstrate that both CG models quite accurately repro-

duce the inter- and intra-molecular structure of the OPLS-AA toluene model in the canonical ensemble at 298 K and at a density corresponding to atmospheric pressure. Somewhat intriguingly, although Figure 2.5 demonstrates that the methyl model reproduced the atomistic site-site RDFs with greater accuracy than the COM model, Figure 2.6 demonstrates that the COM model reproduces the alignment and packing of toluene molecules more accurately.

The differences between the two CG models reflect the subtle relationship between the CG mapping, the MF, the force field basis set, and the structural accuracy of the CG model. A fixed mapping (and a given atomistic distribution) completely determines the MF. For a given MF, the accuracy of the CG approximate force field depends upon the basis set. Figure 2.5 demonstrates that, for a given MF, this approximation (and, presumably, also the accuracy of the resulting CG structure) can be systematically improved by expanding the basis set. At the same time, Figure 2.6 demonstrates that, given a fixed basis set, the accuracy with which the CG force field approximates the MF depends subtly upon the mapping. In this case, the alignment and packing of molecules in the atomistic model appear very similar when viewed through the two different mappings. However, given the same force field basis set, the aspects of the MF that determine the packing and alignment of toluene molecules are more accurately approximated by the COM force field than by the methyl force field. This relationship requires further investigation that may be facilitated by considering the relative entropy formalism.[37]

We note that DeVane et al. have recently published a four site CG model of toluene.[100] This model placed CG sites between atoms C_1 and C_7, between atoms C_2 and C_3, between atoms C_5 and C_6, and on atom C_4. Devane et al. parameterized attractive Lennard-Jones 9-6 potentials to accurately reproduce experimental density, surface tension, and interfacial tension measurements. In contrast to the present model, which slightly underestimates the molecular packing of the OPLS-AA model, the model of Devane et al. appears to slightly overestimate the atomistic packing reported by molecular RDFs.

2.4.4 Temperature Transferability of the G-YBG Model

Figures 2.7 and 2.8 assess the transferability of the g-YBG toluene model that was parameterized at room temperature (i.e., at 298 K) to temperatures slightly above the freezing temperature (i.e., at 273 K) and slightly below the boiling temperature (i.e., at 373 K). In each case, the simulations were performed in the canonical ensemble in a volume corresponding to the equilibrium density of the OPLS-AA model at atmospheric pressure. Figures 2.7 and 2.8 only present results for the methyl model, which more accurately reproduced site-site

RDFs, but less accurately reproduced molecular alignment and packing. The trends for the COM model are consistent with those shown below.

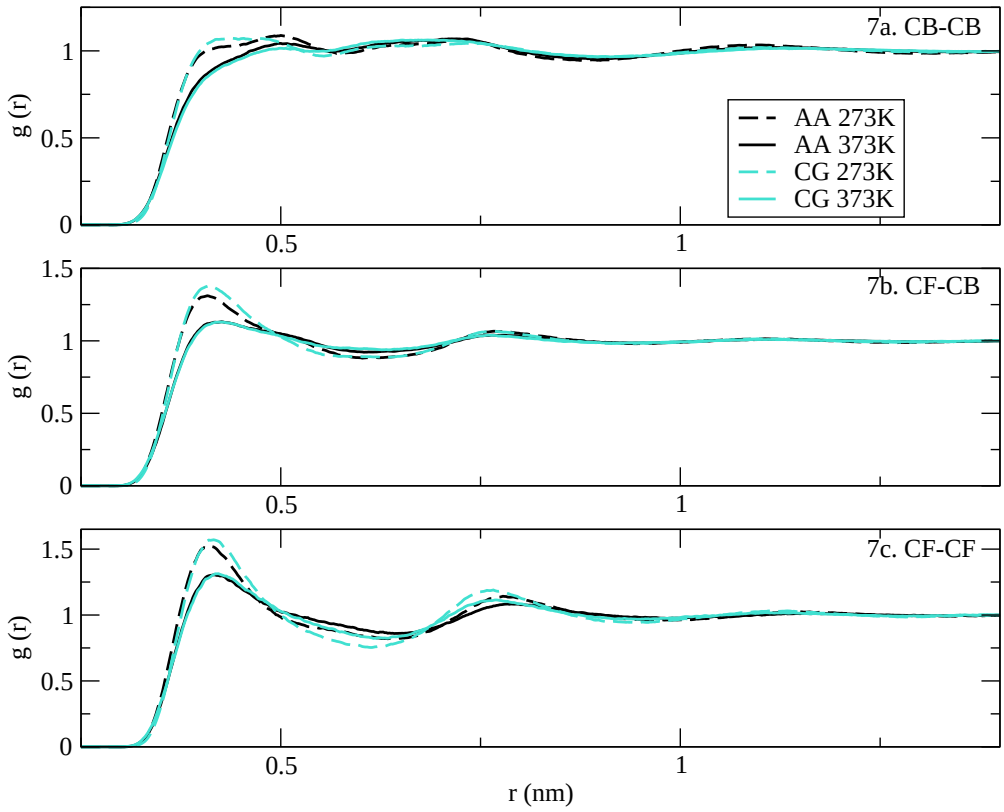

Figure 2.7. Characterization of the transferability of the CG toluene models. Figure 2.7 compares site-site RDFs from all-atom (black curves) and CG (cyan curves) simulations at 273 K (dashed curves) and 373 K (solid curves) when employing the force field calculated at 298 K for the methyl model (left columns in Figs 5-7) with the linear spline basis set. Figures 2.7a, 2.7b, and 2.7c present RDFs for the CB-CB, CF-CB, and CF-CF site pairs, respectively. Results are presented for the methyl model.

Figure 2.7 compares the site-site RDFs calculated from simulations of the methyl model (cyan curves) with those obtained from atomistic simulations (black curves) at 273 (dashed curves) and at 373 K (solid curves). As expected, simulations at decreasing temperature demonstrate more tightly packed structure, i.e., RDF peaks increase in magnitude and shift to shorter distances. Simulations of the CG model, while employing the potentials calculated for 298 K, provide an increasingly accurate description of atomistic site-site RDFs with increasing temperature. At 373 K, the CG simulations reproduce the corresponding atomistic RDFs with essentially quantitative accuracy. Even at 273 K, though, the CG model reproduces the atomistic RDFs quite accurately.

Figure 2.8 compares the alignment and packing of toluene molecules in the atomistic and CG simulations at 273 and 373 K for the methyl mapping. As shown in Figure 2.8a, the molecular RDFs follow the trends demonstrated by the site-site RDFs in Figure 2.7. Although the CG molecular RDF appears to demonstrate less temperature dependence than the atomistic RDF, the agreement between the atomistic and CG RDFs improves with increasing temperature. Figures 2.8b-2.8d characterize the alignment of toluene molecules along the molecular directors $\hat{\mathbf{n}}_1$ and $\hat{\mathbf{n}}_2$. The molecular alignment appears to be relatively temperature independent in both the atomistic and CG models.

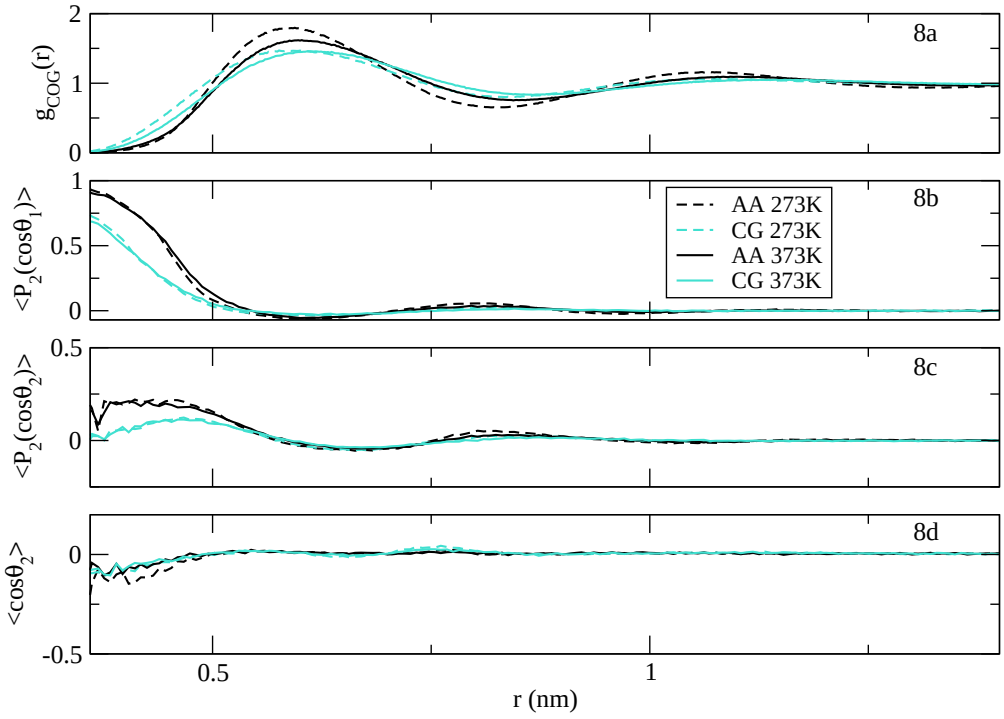

Figure 2.8. Further characterization of the transferability of the CG toluene models. Figure 2.8 compares the molecular packing observed in all-atom (black curves) and CG (cyan curves) simulations at 273 K (dashed curves) and 373 K (solid curves) when employing the force field calculated at 298 K for the methyl model (left columns in Figs 5-7) with the linear spline basis set. Figures 2.8a-2.8d present quantities corresponding to Figures 2.6a-2.6d. Results are presented for the methyl model.

In summary, Figures 2.7 and 2.8 demonstrate that the current g-YBG models appear to provide a reasonably accurate model of molecular structure in the constant NVT ensemble over the range of temperatures for which toluene remains liquid. It should be emphasized that this relatively high degree of transferability was not ensured by the parameterization

procedure for the CG model. In general, despite considerable progress in this direction, current structure-motivated CG models are not expected to necessarily provide temperature transferability. The transferability of the present models may be a consequence of the rigidity of toluene and the relatively high resolution mappings employed.

2.4.5 Efficiency

In closing, we briefly comment upon the efficiency of the present model. Although dramatic gains in computational efficiency are a primary motivation for CG modeling, the present study made no attempts to optimize the efficiency of either the atomistic or CG models. The calculated CG force field, which was optimized to reproduce structure rather than thermodynamic properties or efficiency, includes relatively complicated tabulated potentials that are characterized by a "hard" core excluded volume repulsion. The form of these potentials may preclude larger time steps. Moreover, the present work also employed a relatively fine mapping that only reduced the number of particles in each molecules from 7 C and 8 H atoms to 3 CG sites. In contrast, many CG models often employ "softer" Lennard-Jones potentials that may be more efficient to evaluate and also allow a larger integration time step. Furthermore, such models also often benefit from applying more aggressive CG mappings, although it should be noted that recent thermodynamic-motivated approaches have employed 3[28,31] or 4 sites[100] to accurately describe the packing of small aromatic groups.

In order to quantify the efficiency of the present model, it is necessary to determine the physical relationship between integration timesteps in the atomistic and CG models. While it is difficult to directly relate timescales in atomistic and CG simulations, one possible heuristic is to scale time between the two models so as to match the diffusion constant.[176] Based upon this perspective, the efficiency of the present CG model may be estimated by noting that molecules diffuse 28.3 nm/cpu-day and 0.422 nm/cpu-day in the CG and atomistic simulations, respectively. Consequently, according to this metric, the CG model provides a 67 fold gain in efficiency.

2.5 Summary and Conclusions

The present work continues the development of the g-YBG framework and parameterizes a three site CG model for simulating liquid toluene in the canonical ensemble at 298 K. Figure 2.2 demonstrates that the g-YBG calculations accurately recover the MS-CG force

field directly from structural correlation functions, i.e., the g-YBG approach minimizes a "force-matching" variational principle without requiring explicit force information. Both the MS-CG and g-YBG approach determine an "optimal approximation" to the many-body PMF by projecting the corresponding many-body MF onto a set of basis vectors that correspond to particular interactions in the CG potential. While the MS-CG method determines these projections from force correlation functions sampled in atomistic simulations, Figure 2.3 demonstrates that these projections are related to mean forces for the corresponding interactions. These mean forces of a single variable can be determined from corresponding simple structural correlation functions, e.g., RDFs. The MS-CG metric tensor allows one to decompose the mean force along a single degree of freedom into a direct force and a correlated force from the environment. Figure 2.3b explicitly demonstrates that the surrounding environment makes important contributions to the mean force for nonbonded interactions and, to a lesser extent, also for bond stretch interactions. Preliminary calculations indicate that the mean forces for nonbonded interactions reflect very important contributions from correlated bonded interactions and also contributions from coupled nonbonded interactions.

Figures 2.4-2.6 investigate the structural accuracy of the resulting three site toluene model and also the sensitivity of the model to the CG representation and the force field basis set. These calculations demonstrate that the CG model provides a faithful description of the structure present in the atomistic model, including the intramolecular structure, the intermolecular site-site RDFs, and also the molecular packing and alignment. Calculations were performed with two different CG mappings that are distinguished by the definition of the CF site. The methyl mapping defined this site by the C atom of the methyl substituent; the COM mapping defined the same site from the center of mass for the methyl group and the neighboring aromatic atom. Both representations led to reasonably accurate descriptions of the atomistic structure. The methyl model provides a better description of the intermolecular site-site RDFs, while the COM model better reproduces the molecular RDFs and the alignment of molecular axes.

The majority of calculations were performed with a flexible basis set that represented each term in the potential with linear spline functions. Calculations with reduced basis sets (quadratic functions for bonded interactions and Lennard-Jones-type functions for nonbonded interactions) provided a less accurate representation of the atomistic structure, thus demonstrating the variational nature of the method. These calculations also demonstrated that the site-site RDFs are more sensitive to the short-ranged excluded volume contributions to the potential than to the weak longer-ranged attraction or to other fine structure in the potentials, as would be expected according to Weeks-Chandler-Andersen theory.

Figures 2.7 and 2.8 demonstrate that the current model possesses reasonable temperature transferability for NVT simulations in the liquid phase. In particular, temperature dependent changes in the site-site RDFs are accurately reproduced by the CG model over a range of 100 K. The present CG model is estimated to provide a 67 fold gain in efficiency relative to the atomistic model, although we note that no efforts were attempted in the present work to optimize the efficiency of either the atomistic or CG model.

In closing, we note that the present work demonstrates that the g-YBG framework not only determines structurally accurate CG models, but also provides insight into the relationship between atomic structure and CG force fields. Moreover, the present work suggests several directions for future investigation. In particular, future work may analyze the MS-CG metric tensor to further characterize the coupling between different interactions and how these couplings are determined by molecular structure. Additional work is clearly needed to better understand the relationship between the CG mapping, force field basis set, and structural accuracy. Finally, this work suggests that the g-YBG formalism may prove useful for characterizing and modeling complex molecular systems and polymers, in analogy to the extensive development and contributions of the Ornstein-Zernicke integral equation.

Chapter 3

Coarse-graining Entropy, Forces, and Structures

J. F. Rudzinski, W. G. Noid *J. Chem. Phys.* **2011**, 135, 214101

Abstract

Coarse-grained (CG) models enable highly efficient simulations of complex processes that cannot be effectively studied with more detailed models. CG models are often parameterized using either force- or structure-motivated approaches. The present work investigates parallels between these seemingly divergent approaches by examining the relative entropy and multiscale coarse-graining (MS-CG) methods. We demonstrate that both approaches can be expressed in terms of an information function that discriminates between the ensembles generated by atomistic and CG models. While it is well known that the relative entropy approach minimizes the average of this information function, the present work demonstrates that the MS-CG method minimizes the average of its gradient squared. We generalize previous results by establishing conditions for the uniqueness of structure-based potentials and identify similarities with corresponding conditions for the uniqueness of MS-CG potentials. We analyze the mapping entropy and extend the MS-CG and generalized-Yvon-Born-Green formalisms for more complex potentials. Finally, we present numerical calculations that highlight similarities and differences between structure- and force-based approaches. We demonstrate that both methods obtain identical results, not only for a complete basis set, but also for an incomplete harmonic basis set in Cartesian coordinates. However, the two methods differ when the incomplete basis set includes higher order polynomials of Cartesian coordinates or is expressed as functions of curvilinear coordinates.

3.1 Introduction

Despite tremendous recent advances in computational hardware,[6,177–179] software,[7,8,180] and methodology,[11,12,181] many processes of fundamental interest cannot be effectively simulated with conventional atomically-detailed models. These considerations continue to motivate tremendous interest in highly efficient coarse-grained (CG) models that describe systems in reduced detail by grouping atoms into fewer effective interaction sites.[82–86,110,182] The potentials governing the site-site interactions are typically parameterized to reproduce either thermodynamic or structural properties of the system. For instance, pioneering studies by the Klein[27,97,98] and Marrink[28,32,102] groups have employed thermodynamic data to parameterize potentials that have subsequently provided considerable transferability for modeling a wide range of systems. However, despite considerable progress in this direction,[30,31] the resulting potentials may not necessarily provide quantitative accuracy for modeling the internal structure and flexibility of complex molecules.[32,33]

Alternatively, CG potentials may be parameterized to reproduce structural properties of an atomistic model for the same system. In principle, a many-body potential of mean force is the appropriate potential for a CG model that quantitatively reproduces all structural features of the atomistic model (at the resolution of the CG mapping).[44,109,110] However, in practice, the many-body potential of mean force must be approximated by simpler potentials.[111]

These approximate potentials are often parameterized to reproduce target low-order structural correlation functions (e.g., radial distribution functions) that are determined by the atomistic model and CG mapping. In some cases, these potentials can be determined by directly inverting the corresponding correlation functions.[113] However, if the interactions are coupled, CG potentials determined from direct Boltzmann inversion may not reproduce the atomistic correlation functions.[35,118,183] Motivated by renormalization group considerations,[184] the seminal early works by Lyubartsev and Laaksonen[34,53] developed an iterative Inverse Monte Carlo (IMC) method for parameterizing approximate CG potentials. The IMC method represents the approximate potential as a sum of terms, each of which is the product of a potential parameter and a conjugate function of CG coordinates, i.e., a conjugate order parameter. The IMC approach employs Newton's method to determine the set of potential parameters reproducing the target averages for the conjugate set of order parameters, i.e., the corresponding atomistic correlation functions.[50] Papoian and coworkers have expanded upon these considerations to develop a molecular renormalization group[38,122] approach that has accurately modeled complex molecules such as DNA.[123]

Shell has recently proposed an elegant relative entropy formalism for variationally determining CG potentials.[37] Shell defined the relative entropy as an average of an information function, $\phi(\mathbf{r})$, corresponding to the Kullback-Liebler divergence[185] for discriminating between the ensembles of *atomistic* configurations sampled by atomistic and CG models. Minimizing the relative entropy determines the potential parameters that reproduce target atomistic averages for the conjugate order parameters.[37,121] Moreover, the relative entropy may prove useful for estimating errors in CG models and for optimizing the CG mapping.[120]

Murtola et al. noted[124] that the relative entropy functional is closely related to the density functional employed by Chayes and coworkers[127,128] in earlier mathematical studies of the existence and uniqueness of potentials that reproduce known distribution functions. Furthermore, Murtola et al. also demonstrated that applying Newton's method to minimize the relative entropy leads to the IMC method of Lyubartsev and Laaksonen.[124] Consequently, the relative entropy formalism provides a convenient variational framework for considering several structure-based CG approaches.

Independently, Izvekov and Voth[36,59] pioneered an alternative force-based approach.[41,60,71,72,131] This multiscale coarse-graining (MS-CG) method employs forces sampled from atomistic simulations[158,159] to variationally project the many-body mean force field (i.e., the force field corresponding to the many-body potential of mean force) onto a force field "basis set" that is defined by the form of the approximate CG potential.[41,60] When the variational calculation is performed with a complete basis set, the MS-CG method determines the many-body mean force. Simulations with this force field will quantitatively reproduce all structural correlations of the atomistic model (at the level of the CG mapping).[41,62] When the variational calculation is performed with an incomplete basis set, the MS-CG method determines the force field (within the subspace spanned by the incomplete basis) that is minimum "distance" from the many-body mean force field.[41,60] The MS-CG potential can be determined directly from a system of normal linear equations that are expressed in terms of force and structural correlation functions sampled from atomistic simulations. When re-expressed in terms of structural correlation functions, the resulting equations define a generalized-Yvon-Born-Green integral equation theory for complex molecules.[63,64,134,150]

The relative entropy and MS-CG approaches both determine the same potential (to within an additive constant) when the corresponding variational calculations are performed with a complete basis set.[41,110,121] However, for calculations with an incomplete basis set, the two approaches appear quite divergent. In this case, the relative entropy approach employs multiple simulations to determine the CG potential that reproduces atomistic averages for the conjugate order parameters. In contrast, the MS-CG method does not require iterative

simulations and employs atomistic force information to directly project the many-body mean force field onto the approximate basis set, but the resulting model is not guaranteed to reproduce any particular atomistic correlation functions. (It should be emphasized, though, that for many complex systems the MS-CG model does quantitatively or semi-quantitatively reproduce the corresponding atomistic distribution functions.[76,146,151,186,187]) Ruhlë et al. have performed an insightful study that explicitly compared the potentials obtained for structure- and force-motivated models of water, methanol, and hexane.[74] Nevertheless, the general relationship between structure- and force-based methods and the resulting potentials remains relatively obscure.

The present work explores the relationship between structure- and force-based CG approaches by examining the relative entropy and MS-CG approaches. This work identifies several striking parallels between the two approaches, including the relationship between their basis sets, the variational functionals, and the uniqueness of the resulting potentials. A major conclusion of the present work is that both the relative entropy and the MS-CG functional can be expressed in terms of an information function, $\Phi(\mathbf{R})$, that discriminates between the ensembles of CG configurations sampled by the atomistic and CG models. The relative entropy corresponds to the average of Φ, while the MS-CG functional corresponds to the average of $|\nabla\Phi|^2$. In addition, we generalize the well known result of Henderson[46] by identifying conditions for the uniqueness of a set of potentials that reproduce a given set of structural correlation functions. We show that these conditions are closely related to the conditions for the uniqueness of MS-CG potentials. In the course of this analysis, we investigate the mapping entropy and generalize the MS-CG and generalized-Yvon-Born-Green (g-YBG) theory for more complex potentials. Finally, we also present numerical calculations that clarify these relationships for incomplete basis sets and particularly simple systems. We demonstrate that relative entropy and MS-CG calculations not only agree for calculations with a complete basis set, but also for calculations with a highly incomplete harmonic basis set expressed in Cartesian coordinates. In contrast, the two methods differ for either higher order Cartesian basis sets or for incomplete basis sets that are functions of non-Cartesian coordinates.

3.2 Preliminaries

The present section defines key quantities for the following analysis. We consider atomistic and CG models for a given system in the canonical ensemble and assume that neither model includes rigid constraints. Lower and upper case symbols correspond to atomistic and CG

quantities, respectively. The notation largely follows previous work.[41,63,188]

3.2.1 Atomistic Model

The configuration of the atomistic model is defined by the Cartesian coordinates, $\mathbf{r}$, for n atoms in a volume, V, that interact according to a potential, $u(\mathbf{r})$, which may be of arbitrary complexity. The atomistic configuration distribution, $p_r(\mathbf{r})$, is given by

$$p_r(\mathbf{r}) \propto \exp[-u(\mathbf{r})/k_BT]. \tag{3.1}$$

The configurational entropy for this model is given by:

$$s_{\mathbf{r}} = -k_B \int d\mathbf{r}\; p_r(\mathbf{r}) \ln \left[V^n p_r(\mathbf{r}) \right]. \tag{3.2}$$

3.2.2 Coarse-grained Model

The configuration of the CG model is similarly defined by the Cartesian coordinates, $\mathbf{R}$, for N sites in a volume, V, that interact according to a potential, $U(\mathbf{R})$. The CG configuration distribution, $P_R(\mathbf{R}|U)$, depends upon the CG potential according to:

$$P_R(\mathbf{R}|U) = V^{-N} \exp\left[-\left(U(\mathbf{R}) - F[U] \right) / k_BT \right] \tag{3.3}$$

where

$$\exp\left[-F[U]/k_BT\right] = V^{-N} \int d\mathbf{R} \exp[-U(\mathbf{R})/k_BT], \tag{3.4}$$

defines the configurational free energy, F, which is a functional of U.

The present work considers CG potentials of the following form:

$$U(\mathbf{R}) = \sum_{\zeta} \sum_{\lambda} U_\zeta(\psi_\zeta(\{\mathbf{R}\}_\lambda)), \tag{3.5}$$

where ζ indicates a particular interaction (e.g., a dihedral angle interaction) and U_ζ is the corresponding potential (e.g., a dihedral angle potential) that is a function of a single scalar variable, ψ_ζ, (e.g., a dihedral angle) that may be expressed as a function of the Cartesian coordinates $\{\mathbf{R}\}_\lambda$ for a set of CG sites, λ (e.g., the 4 successively bonded sites that form a dihedral angle).

To emphasize the similarity between CG methods and classical density functional theories, we represent the approximate potential as a sum of terms, each of which is a product

of a potential, U_ζ, and a conjugate order parameter or density, $\hat{\rho}_\zeta$:

$$U(\mathbf{R}) = \sum_\zeta \int dx\; U_\zeta(x)\hat{\rho}_\zeta(\mathbf{R};x), \tag{3.6}$$

where

$$\hat{\rho}_\zeta(\mathbf{R};x) = \sum_\lambda \delta(\psi_{\zeta\lambda}(\mathbf{R}) - x), \tag{3.7}$$

and $\psi_{\zeta\lambda}(\mathbf{R}) \equiv \psi_\zeta(\{\mathbf{R}\}_\lambda)$. Equation (3.6) assumes that the integrals are evaluated over the appropriate domain for each potential and that these potentials obey suitable smoothness properties and boundary conditions. The density fields defined in Equation (3.7) form an incomplete basis that spans a linear vector space of potentials of the form given by Equation (3.6). The functions, $U_\zeta(x)$, serve as constant coefficients that identify a potential, U, in this space.

The corresponding CG force field may be expressed:

$$\mathbf{F}_I(\mathbf{R}) = \sum_\zeta \int dx\; U_\zeta(x)\boldsymbol{\mathscr{G}}_{I;\zeta}(\mathbf{R};x) \tag{3.8}$$

where $\boldsymbol{\mathscr{G}}_{I;\zeta}(\mathbf{R};x) = -\partial\hat{\rho}_\zeta(\mathbf{R};x)/\partial\mathbf{R}_I$. Equation (3.8) may be re-expressed:

$$\mathbf{F} = \sum_\zeta \int dx\; U_\zeta(x)\boldsymbol{\mathscr{G}}_\zeta(x) \tag{3.9}$$

where both $\mathbf{F}$ and $\boldsymbol{\mathscr{G}}_\zeta(x)$ are vectors in an abstract vector space of force fields[41,60] that define a Cartesian vector force on each site as a function of the CG configuration $\mathbf{R}$. The vectors $\boldsymbol{\mathscr{G}}_\zeta(x)$ that are included in Equation (3.9) form an incomplete basis set that spans a subspace of CG force fields. The functions, $U_\zeta(x)$, serve as constant coefficients that identify a force field, $\mathbf{F}$, in this space.

The functional form of U in Equation (3.5) is immediately relevant for molecular mechanics-type force fields. Appendix A demonstrates that this treatment can be readily generalized for the case that U_ζ depends upon multiple variables, $x \to \mathbf{x} = \{x_1, x_2, \ldots\}$. Equations (3.8) and (3.9) differ from recent analyses[64,134,188] of the MS-CG method by emphasizing the potential functions, $U_\zeta(x)$, rather than the corresponding force functions, $\phi_\zeta(x) = -dU_\zeta(x)/dx$, as the coefficients of force field basis vectors. In the case that each potential function, $\{U_\zeta\}$, depends upon a single scalar variable and that each potential satisfies appropriate boundary conditions, the two representations are equivalent after integrating Equation (3.9) by parts.

However, in the case that the potential functions depend upon multiple variables, the force functions corresponding to different variables, e.g., $-\partial U_\zeta(\mathbf{x})/\partial x_i$, can no longer be treated independently in variational calculations. The present formulation is also convenient because it emphasizes the similarity between Equations (3.6) and (3.9). In particular, the densities, $\hat{\rho}_\zeta$, and corresponding force field vectors, $\mathscr{G}_\zeta$, act as basis vectors for structure- and force-motivated approaches, respectively. It should be noted that 1) the present analysis readily generalizes for the case that the potential, U, is defined by a discrete sum over terms by replacing each integral with a corresponding sum over parameters, $U_{\zeta d}$, and also that 2) the continuous relationships considered in this manuscript are directly derived from the limit of such finite discrete sums.

3.2.3 Mapping Relationships

We define a linear mapping $\mathbf{M} : \mathbf{r} \to \mathbf{R} = \mathbf{M}(\mathbf{r})$ that is assumed to satisfy conventional properties.[41] This mapping determines a probability distribution, p_R, for the atomistic model to sample a configuration $\mathbf{r}$ that maps to a given CG configuration $\mathbf{R}$:

$$p_R(\mathbf{R}) = \langle \delta(\mathbf{M}(\mathbf{r}) - \mathbf{R}) \rangle \,. \tag{3.10}$$

In Equation (3.10) and throughout this manuscript, angular brackets denote a canonical average over the atomistic configuration space according to the probability distribution $p_r(\mathbf{r})$ defined in Equation (3.1).

The probability distribution p_R defines a "different" entropy for the atomistic model. We define $s_\mathbf{R}$ as the configurational entropy of the atomistic model when evaluated as an average over the CG configuration space:

$$s_\mathbf{R} = -k_B \int d\mathbf{R}\; p_R(\mathbf{R}) \ln \left[V^N p_R(\mathbf{R}) \right] . \tag{3.11}$$

This entropy function reflects the mapping, $\mathbf{M}$, but is otherwise independent of the CG model.

Previous works[41] have considered a CG model to be consistent (in configuration space) with a particular atomistic model for the same system when the two models sample the same distribution of configurations in the CG configuration space, i.e., when $P_R(\mathbf{R}) = p_R(\mathbf{R})$ for all $\mathbf{R}$. Consequently, the probability distribution p_R defines the appropriate potential for a

consistent CG model, U^0, to within an additive constant:

$$U^0(\mathbf{R}) = -k_B T \ln \left[V^N p_R(\mathbf{R}) \right] + \text{const.} \tag{3.12}$$

The potential U^0 is named a many-body potential of mean force (PMF) because the corresponding force field corresponds to a conditioned canonical ensemble average of the atomistic force field evaluated over the atomistic configurations that map to a given CG configuration:

$$\mathbf{F}_I^0(\mathbf{R}) = \langle \mathbf{f}_I(\mathbf{r}) \rangle_{\mathbf{R}} \tag{3.13}$$

where $\mathbf{f}_I(\mathbf{r})$ is the atomistic force[41] on site I in configuration $\mathbf{r}$ and, for any quantity $a(\mathbf{r})$,

$$\langle a(\mathbf{r}) \rangle_{\mathbf{R}} = \langle a(\mathbf{r}) \delta(\mathbf{M}(\mathbf{r}) - \mathbf{R}) \rangle \, / \, \langle \delta(\mathbf{M}(\mathbf{r}) - \mathbf{R}) \rangle \,, \tag{3.14}$$

is a canonically weighted average over the set of atomistic configurations that map to $\mathbf{R}$. A CG model that employs the many body mean force (MF) field, $\mathbf{F}^0$, defined by Equation (3.13), as a conservative force field will be consistent with the atomistic model in the sense defined above.[41]

The distribution, p_R, determines several key quantities for parameterizing and analyzing CG models. For each order parameter included in Equation (3.6), we define a "target" atomistic structural correlation function:

$$p_\zeta(x) = \int d\mathbf{R} \; p_R(\mathbf{R}) \hat{\rho}_\zeta(\mathbf{R}; x). \tag{3.15}$$

We also define a corresponding correlation function that is weighted according to the equilibrium distribution for the CG model with a potential U:

$$P_\zeta(x|U) = \int d\mathbf{R} \; P_R(\mathbf{R}|U) \hat{\rho}_\zeta(\mathbf{R}; x). \tag{3.16}$$

As is well known,[34,123] P_ζ is a functional derivative of the CG free energy:
$P_\zeta(x|U) = \delta F[U] / \delta U_\zeta(x)$.

The distribution p_R also determines an inner product $\odot$ between any two elements, $\mathbf{F}^{(1)}$ and $\mathbf{F}^{(2)}$, in the vector space of CG force fields:

$$\mathbf{F}^{(1)} \odot \mathbf{F}^{(2)} = \int d\mathbf{R} \; p_R(\mathbf{R}) \; \frac{1}{3N} \sum_I \mathbf{F}_I^{(1)}(\mathbf{R}) \cdot \mathbf{F}_I^{(2)}(\mathbf{R}) \tag{3.17}$$

and a corresponding norm, according to $||\mathbf{F}|| = (\mathbf{F} \odot \mathbf{F})^{1/2}$.

3.3 Variational Methods

Having introduced appropriate notation, the present subsection analyzes and summarizes key features of the relative entropy and MS-CG approaches as prototypes for structure- and force-based CG modeling.

3.3.1 Relative Entropy

Shell introduced[37] the relative entropy in terms of the log likelihood for the CG model to reproduce the distribution of *atomistic* configurations sampled by the atomistic model:

$$S_{rel}[U] = k_B \int d\mathbf{r}\; p_r(\mathbf{r}) \ln \left[\frac{p_r(\mathbf{r})}{P_r(\mathbf{r}|U)} \right], \tag{3.18}$$

where $P_r(\mathbf{r}|U)$ is the probability of sampling the atomistic configuration $\mathbf{r}$ from a CG model with potential U. Kullback and Liebler[185] interpreted the quantity

$$\phi(\mathbf{r}|U) = \ln \left[\frac{p_r(\mathbf{r})}{P_r(\mathbf{r}|U)} \right] \tag{3.19}$$

as the information content in the configuration $\mathbf{r}$ for discriminating between the two distributions, $p_r(\mathbf{r})$ and $P_r(\mathbf{r}|U)$.

Because multiple atomistic configurations map to the same CG configuration, P_r is perhaps somewhat ambiguous. We consider the following definition:

$$P_r(\mathbf{r}|U) = \frac{g(\mathbf{r})}{\Omega(\mathbf{M}(\mathbf{r}))} P_R(\mathbf{M}(\mathbf{r})|U), \tag{3.20}$$

where

$$\Omega(\mathbf{R}) = \int d\mathbf{r}\; g(\mathbf{r})\; \delta(\mathbf{M}(\mathbf{r}) - \mathbf{R}), \tag{3.21}$$

and $g(\mathbf{r})$ is a weighting function. This definition for $P_r(\mathbf{r}|U)$ is proportional to the probability for the CG model to sample the configuration $\mathbf{M}(\mathbf{r})$ and also corresponds to defining $g(\mathbf{r})\delta(\mathbf{M}(\mathbf{r}) - \mathbf{R})/\Omega(\mathbf{M}(\mathbf{r}))$ as the normalized conditional probability of sampling an atomistic configuration $\mathbf{r}$ given a fixed CG configuration $\mathbf{R}$. Given Equation (3.20), the relative

entropy is re-expressed:

$$S_{rel}[U] = k_B \int d\mathbf{r}\; p_r(\mathbf{r}) \ln\left[\frac{V^n}{V^N}\frac{p_r(\mathbf{r})}{P_R(\mathbf{M}(\mathbf{r})|U)}\right] + S_{map} \tag{3.22}$$

where

$$S_{map} = k_B \left\langle \ln\left[\frac{V^N}{V^n}\frac{1}{g(\mathbf{r})}\Omega(\mathbf{M}(\mathbf{r}))\right]\right\rangle. \tag{3.23}$$

Shell's seminal work quite naturally defined $g(\mathbf{r}) = 1$ so that P_r gives equal weight to all atomistic configurations that map to the same CG configuration.[37] $\Omega(\mathbf{R})$ then corresponds to the total unweighted volume element of atomistic configuration space that maps to $\mathbf{R}$. Given $g = 1$, Equations (3.22) and (3.23) are equivalent to the earlier results of Shell aside from dimensional constants.[37,121]

The remainder of this work considers an alternative definition: $g(\mathbf{r}) = p_r(\mathbf{r})$. This implies that $\Omega(\mathbf{R}) = p_R(\mathbf{R})$ and

$$P_r(\mathbf{r}|U) = \frac{p_r(\mathbf{r})}{p_R(\mathbf{M}(\mathbf{r}))}P_R(\mathbf{M}(\mathbf{r})|U). \tag{3.24}$$

Equation (3.24) does not give equal weight to all $\mathbf{r}$ that map to the same $\mathbf{R}$, but weights them according to their atomistic Boltzmann weight and has the property that $P_r = p_r$ for a consistent CG model, which is not the case if $g = 1$. Given $g(\mathbf{r}) = p_r(\mathbf{r})$, the entropy of mapping, S_{map}, becomes the difference between the entropy of the atomistic model when viewed from the atomistic and the CG configuration space:

$$S_{map} = s_{\mathbf{r}} - s_{\mathbf{R}}. \tag{3.25}$$

This mapping entropy may also be expressed:

$$S_{map} = k_B \left\langle \ln\left[\frac{V^N}{V^n}\Omega_1(\mathbf{M}(\mathbf{r}))\right]\right\rangle - k_B \left\langle \ln\left[\frac{p_r(\mathbf{r})}{\bar{p}_r(\mathbf{M}(\mathbf{r}))}\right]\right\rangle \tag{3.26}$$

where $\Omega_1(\mathbf{R}) = \int d\mathbf{R}\; \delta(\mathbf{M}(\mathbf{r}) - \mathbf{R})$ is the unweighted volume element of the atomistic configuration space that maps to $\mathbf{R}$, and

$$\bar{p}_r(\mathbf{R}) = \int d\mathbf{r}\; p_r(\mathbf{r})\delta(\mathbf{M}(\mathbf{r}) - \mathbf{R})/\Omega_1(\mathbf{R}), \tag{3.27}$$

is an unweighted average of $p_r(\mathbf{r})$ evaluated over this volume element.

Equation (3.26) decomposes S_{map} into a geometric component and a component describing the smoothing of the probability distribution. The second term in Equation (3.26)

quantifies the fluctuations in the Boltzmann weight, p_r, that are smoothed out by integrating over all atomistic configurations that map to the same CG configuration. By the Gibbs inequality, this term is nonnegative and only vanishes if $p_r(\mathbf{r}) = \bar{p}_r(\mathbf{M}(\mathbf{r}))$ everywhere, i.e., if and only if all the configurations $\mathbf{r}$ that map to a given CG configuration $\mathbf{R}$ have equal Boltzmann weight in the atomistic model. This places an upper bound upon the entropy difference between the two descriptions of the atomistic model:

$$s_{\mathbf{r}} - s_{\mathbf{R}} \leq k_B \left\langle \ln \left[\frac{V^N}{V^n} \Omega_1(\mathbf{M}(\mathbf{r})) \right] \right\rangle \tag{3.28}$$

The volume element $\Omega_1(\mathbf{R})$ can be directly evaluated in terms of the mapping coefficients,[189] although the expression is somewhat complex and periodic boundary conditions may introduce additional complications. Nevertheless, it seems intuitively reasonable that Ω_1 should be independent of configuration for a system with periodic boundary conditions. It also seems clear from geometric arguments that the average in Equation (3.28) is nonnegative. In particular, the average in Equation (3.28) vanishes for any mapping that associates each site with a single atom so that $s_{\mathbf{r}} - s_{\mathbf{R}} \leq 0$, e.g., for an atomically detailed implicit solvent model.

It should be emphasized that the entropy difference, $s_{\mathbf{r}} - s_{\mathbf{R}}$, is a direct consequence of the CG mapping reducing the dimensionality of the configuration space and the associated averaging of p_r. This effect is completely independent of the CG potential. In particular, even if a CG model perfectly reproduces the configuration distribution implied by the atomistic model, the resulting model may not reproduce the configurational entropy of the atomistic model and the entropy difference between the two models will satisfy the inequality, Equation (3.28). Consequently, S_{map} may provide a useful quantitative metric for optimizing the CG mapping.

Equation (3.24) allows for a convenient expression for the relative entropy as an average over the CG configuration space:

$$S_{rel}[U] = k_B \int d\mathbf{R}\, p_R(\mathbf{R}) \ln \left[\frac{p_R(\mathbf{R})}{P_R(\mathbf{R}|U)} \right]. \tag{3.29}$$

The relative entropy has now been expressed in analogy to Equation (3.22) and without reference to a mapping entropy. In a very recent paper that considered the relative entropy for modeling CG dynamics, Español and Zúñiga derived Equation (3.29) for the relative entropy as the log-likelihood for the atomistic and CG models to sample the same distribution of *CG* configurations.[190] The present analysis clarifies the relationship between these two

conventions.

The Gibbs inequality implies that $S_{rel} \geq 0$ and that S_{rel} attains its global minimum when $p_R(\mathbf{R}) = P_R(\mathbf{R}|U)$, i.e., when the potential for the CG model is the many-body PMF. More generally, when S_{rel} is varied with respect to an incomplete basis set:

$$\frac{\delta}{\delta U_\zeta(x)} S_{rel}[U] = \Delta P_\zeta(x|U)/T, \tag{3.30}$$

where

$$\Delta P_\zeta(x|U) = p_\zeta(x) - P_\zeta(x|U) \tag{3.31}$$

is the difference between the target distribution and the distribution generated by simulations with the potential U. Thus, when considering potentials of the form given by Equation (3.5), the relative entropy is minimized when the CG potential reproduces the atomistic distribution functions for the corresponding density operators, i.e., when $P_\zeta(x|U) = p_\zeta(x)$.

As noted earlier by Murtola et al.,[124] the IMC method determines the CG potentials that minimize the relative entropy by employing Newton's method to find the parameters for which $\Delta P_\zeta(x|U)$ vanishes. Newton's method leads to a system of coupled equations for iteratively updating the potentials $\{U_\zeta(x)\} \rightarrow \{U_\zeta(x) + \delta U_\zeta(x)\}$:

$$\Delta P_\zeta(x|U) = \sum_{\zeta'} \int dx' \; M_{\zeta\zeta'}(x, x'|U) \delta U_{\zeta'}(x'). \tag{3.32}$$

where

$$\begin{aligned} M_{\zeta\zeta'}(x, x'|U) &\equiv \frac{\delta P_\zeta(x|U)}{\delta U_{\zeta'}(x')} \\ &= -\frac{1}{k_B T} \int d\mathbf{R} \; P_R(\mathbf{R}|U) \delta\hat{\rho}_\zeta(\mathbf{R}; x|U) \delta\hat{\rho}_{\zeta'}(\mathbf{R}; x'|U) \end{aligned} \tag{3.33}$$

is a susceptibility matrix describing correlated fluctuations of the order parameters in the CG model with $\delta\hat{\rho}_\zeta(\mathbf{R}; x|U) = \hat{\rho}_\zeta(\mathbf{R}; x) - P_\zeta(x|U)$. We emphasize that, because g is independent of U, Shell's earlier work[37,121] leads to the same system of equations for the CG potentials, i.e., Equations (3.30)-(3.33).

3.3.2 Multiscale Coarse-graining

The Multiscale Coarse-graining (MS-CG) method[36,59] considers a force-matching[158,159] functional:

$$\chi^2[U] = \frac{1}{3N}\left\langle \sum_I |\mathbf{f}_I(\mathbf{r}) - \mathbf{F}_I(\mathbf{M}(\mathbf{r})|U)|^2 \right\rangle \tag{3.34}$$

$$= \chi^2[U^0] + ||\mathbf{F}^0 - \mathbf{F}[U]||^2, \tag{3.35}$$

where $\mathbf{F}$ is the CG force field determined from the approximate potential, U, and $\mathbf{F}^0$ is the force field determined from the many-body PMF, U^0. The second relation follows from Equation (3.13).[41] Because the second term is necessarily nonnegative, the many-body PMF determines the unique global minimum[62] of χ^2.

Given the incomplete basis set in Equation (3.9), the MS-CG variational principle determines the force field spanned by the basis that is "closest" to the many-body mean force field.[41] Varying χ^2 with respect to the corresponding potential parameters leads to:

$$\frac{\delta}{\delta U_\zeta(x)}\chi^2[U] = 2\boldsymbol{\mathscr{G}}_\zeta(x) \odot \left(\mathbf{F}[U] - \mathbf{F}^0\right). \tag{3.36}$$

Therefore, minimizing χ^2 determines the approximate force field by geometrically projecting the MF onto the given basis set.[63,64,161,162] After evaluating Equation (3.36) at this minimum and expanding $\mathbf{F}$ according to Equation (3.9), the MS-CG potential is directly determined from the normal system[60,62] of linear equations:

$$\sum_{\zeta'} \int dx'\, G_{\zeta\zeta'}(x,x')U_{\zeta'}(x') = b_\zeta(x), \tag{3.37}$$

where $G_{\zeta\zeta'}(x,x') = \boldsymbol{\mathscr{G}}_\zeta(x) \odot \boldsymbol{\mathscr{G}}_{\zeta'}(x')$ is a metric tensor and $b_\zeta(x) = \boldsymbol{\mathscr{G}}_\zeta(x) \odot \mathbf{F}^0$ is the projection of the many-body mean force onto the corresponding basis vector.[63,64] The metric, $G_{\zeta\zeta'}(x,x')$, is a structural correlation function quantifying correlations between the different interactions and plays an analogous role to $M_{\zeta\zeta'}(x,x'|U)$ in Equation (3.33). However, $G_{\zeta\zeta'}$ is independent of U and instead determined directly from the atomistic model and CG mapping. The MS-CG method evaluates $b_\zeta(x)$ by employing forces sampled from atomistic simulations:

$$b_\zeta(x) = \frac{1}{3N}\left\langle \sum_I \mathbf{f}_I(\mathbf{r}) \cdot \boldsymbol{\mathscr{G}}_{I;\zeta}(\mathbf{M}(\mathbf{r});x) \right\rangle. \tag{3.38}$$

However, $b_\zeta(x)$ may also be evaluated in terms of structural correlation functions according

to a generalized Yvon-Born-Green equation, so that the MS-CG potential may be calculated without explicit knowledge of atomistic forces.[63,64,134,150,188]

As noted earlier, the MS-CG method has been previously discussed in terms of force functions: $\phi_\zeta(x) = -dU_\zeta(x)/dx$. If each U_ζ is a function of a single variable and if these functions satisfy appropriate boundary conditions so that each term is uniquely determined from $\phi_\zeta(x)$, then Equation (3.37) is equivalent to previous expressions. In this case, Equation (3.37) may be derived from prior expressions by differentiation with respect to x and an integration by parts with respect to x'. In particular, $b_\zeta(x) = d\widetilde{b}_\zeta(x)/dx$ and $G_{\zeta\zeta'}(x,x') = \partial^2\widetilde{G}_{\zeta\zeta'}(x,x')/\partial x\partial x'$, where $\widetilde{b}_\zeta$ and $\widetilde{G}_{\zeta\zeta'}$ are correlation functions employed in previous analyses.[41,134] However, in contrast to earlier treatments, the present expressions also readily generalize for the case that U_ζ depends upon multiple variables.

3.4 Similarities in the Variational Principles

3.4.1 Functionals

In analogy to Equation (3.19), we define

$$\Phi(\mathbf{R}|U) = \ln\left[\frac{p_R(\mathbf{R})}{P_R(\mathbf{R}|U)}\right], \tag{3.39}$$

as the information content for discriminating between the distribution of CG configurations generated by the atomistic and CG models.[185] We note that, when considering variations with an incomplete basis set,

$$\frac{\delta\Phi(\mathbf{R}|U)}{\delta U_\zeta(x)} = \delta\hat{\rho}_\zeta(\mathbf{R};x|U)/k_BT \tag{3.40}$$

where $\delta\hat{\rho}_\zeta(\mathbf{R};x|U) = \hat{\rho}_\zeta(\mathbf{R};x) - P_\zeta(x|U)$ as in Equation (3.33), while variations with respect to Cartesian coordinates may be expressed:

$$\frac{\partial\Phi(\mathbf{R}|U)}{\partial\mathbf{R}_I} = \left(\mathbf{F}_I^0(\mathbf{R}) - \mathbf{F}_I(\mathbf{R}|U)\right)/k_BT, \tag{3.41}$$

where $\mathbf{F}_I^0$ is the mean force.

According to Equation (3.29), the relative entropy may be expressed as an average of

$\Phi(\mathbf{R}|U)$ evaluated according to p_R:

$$S_{rel}[U] = k_B \int d\mathbf{R}\; p_R(\mathbf{R})\Phi(\mathbf{R}|U). \tag{3.42}$$

Moreover, as a direct consequence of Equations (3.35) and (3.41), the MS-CG functional may also be expressed in terms of $\Phi(\mathbf{R}|U)$:

$$\chi^2[U] \;=\; \frac{(k_BT)^2}{3N}\int d\mathbf{R}\; p_R(\mathbf{R})\left|\nabla\Phi(\mathbf{R}|U)\right|^2 + \chi^2[U^0] \tag{3.43}$$

where $|\nabla A(\mathbf{R})|^2 = \sum_I \left(\partial A(\mathbf{R})/\partial \mathbf{R}_I\right)^2$. The second term in Equation (3.43) depends only upon the atomistic model and the CG mapping, i.e., it is independent of U, and does not influence the variational calculation of the MS-CG potential.

The parallel between Equations (3.42) and (3.43) is a central result of the present work. The relative entropy formalism determines the CG potential that minimizes the average of Φ; the MS-CG formalism determines the CG potential that minimizes the average of $|\nabla\Phi|^2$. Although Φ may be either positive or negative, the Gibbs inequality implies that $S_{rel} \geq 0$ and vanishes if and only if U is the PMF, for which $\Phi(\mathbf{R}|U^0) = 0$. Similarly, χ^2 is also nonnegative and is minimized by the PMF for which, $|\nabla\Phi(\mathbf{R}|U^0)|^2 = 0$. $\chi^2[U^0]$ and S_{map} serve somewhat analogous roles in the MS-CG and relative entropy approaches, respectively. While $\chi^2[U^0]$ quantifies fluctuations in the atomistic force field for a given $\mathbf{R}$, S_{map} quantifies the fluctuations in the atomistic probability distribution for a given $\mathbf{R}$.

Minimizing S_{rel} with respect to $U_\zeta(x)$ then leads to the condition

$$\int d\mathbf{R}\; p_R(\mathbf{R})\; \delta\hat{\rho}_\zeta(\mathbf{R}; x|U) = 0, \tag{3.44}$$

which is equivalent to the condition of Equation (3.30), i.e., the relative entropy is minimized with respect to a set of potentials when the conjugate distributions generated by the CG model match those determined by the atomistic model. This is a self-consistent condition because it requires knowledge of $P_\zeta(x|U)$, which depends upon the potential-dependent normalization, $\exp\left[-F[U]/k_BT\right]$, and can only be determined from simulations with the CG potential U.

Minimizing χ^2 with respect to $U_\zeta(x)$ leads to the following condition:

$$\int d\mathbf{R}\; p_R(\mathbf{R})\left(\nabla\Phi(\mathbf{R}|U)\cdot\nabla\right)\delta\hat{\rho}_\zeta(\mathbf{R}; x|U) = 0. \tag{3.45}$$

As a result of Equation (3.41) and the relationship between $\hat{\rho}_\zeta(\mathbf{R};x)$ and $\mathscr{G}_\zeta(x)$, Equation (3.45) is equivalent to Equation (3.37). Equation (3.45) can be solved directly. Because $P_\zeta(x|U)$ is a configuration-independent average, the gradient operator eliminates the potential dependence from $\delta\hat{\rho}_\zeta(\mathbf{R};x|U)$. Clearly, the distinction between the two variational principles hinges upon the operator $\nabla\Phi(\mathbf{R}|U)\cdot\nabla$, which trivially vanishes when U is the many-body PMF.

3.4.2 Uniqueness

Having investigated parallels in the variational principles underlying structure-based and force-based CG approaches, we now investigate their respective uniqueness properties. Henderson[46] followed the earlier analyses of Hohenberg, Kohn,[191] and Mermin[192] to demonstrate that, in the case of simple liquids, the pair potential reproducing a given radial distribution function is unique. Johnson et al. subsequently extended this result for liquids with site-site pair potentials.[129] The present analysis further extends this result, analyzes its generality, and also observes remarkable similarities with the conditions necessary for the uniqueness of MS-CG potentials.

3.4.2.1 Structure-based Uniqueness

We consider two CG potentials U_A and U_B that may be expressed in terms of a set of potentials $\{U_{\zeta A}(x)\}$ and $\{U_{\zeta B}(x)\}$ according to Equation (3.6). We shall assume that both sets of potentials satisfy suitable boundary conditions, e.g., that non-bonded potentials vanish at long distances and that bonded potentials vanish at their minima. These two sets of potentials determine two sets of structural correlation functions $\{P_{\zeta A}(x) = P_\zeta(x|U_A)\}$ and $\{P_{\zeta B}(x) = P_\zeta(x|U_B)\}$ that are defined by Equation (3.16).

Equation (3.6) expressed the CG potential as a linear combination of density fields, $\{\hat{\rho}_\zeta(\mathbf{R};x)\}$. We define a set of fields as **linearly independent** if

$$\sum_\zeta \int dx\; c_\zeta(x)\; \hat{\rho}_\zeta(\mathbf{R};x) = 0 \tag{3.46}$$

for all $\mathbf{R}$ implies that $c_\zeta(x) = 0$ for all ζ and x. Conversely, a set of fields is linearly dependent if any field in the set is a linear combination of the other fields. If the set of fields in Equation (3.6) is linearly dependent, then U no longer determines a unique set of coefficients, $\{U_\zeta(x)\}$. It should also be noted, that unless suitable boundary conditions are required, the potentials are only unique to within an additive constant.

We assume that (1) the two sets of potentials $\{U_{\zeta A}(x)\}$ and $\{U_{\zeta B}(x)\}$ satisfy suitable boundary conditions and (2) the corresponding density fields are linearly independent. Given these assumptions, if the corresponding sets of structural correlation functions are equal, i.e., $\{P_\zeta(x|U_A) = P_\zeta(x|U_B)\}$ for all ζ and x; then it follows that the two sets of potentials are equal, i.e., $\{U_{\zeta A}(x) = U_{\zeta B}(x)\}$. Appendix B provides a complete proof of the contrapositive of this statement. It should be noted that this analysis does not prove the existence of a set of potentials reproducing a conjugate set of structural correlation functions. However, assuming that such a set of potentials exists and that conditions 1 and 2 are satisfied, then this set is unique.

3.4.2.2 Force-based Uniqueness

In contrast to the relative entropy and related iterative variational methods, the MS-CG approach is not guaranteed to reproduce any particular structural correlation functions. The MS-CG method directly projects the many-body MF onto a set of vectors employed as an incomplete basis for the CG force field according to Equation (3.9). This geometric interpretation implies that the MS-CG potential is unique as long as the vectors included in the incomplete basis are linearly independent in the space of force fields. A set of force field vectors $\{\mathscr{G}_\zeta(x)\}$ is linearly independent if

$$\sum_\zeta \int dx\; c_\zeta(x)\; \mathscr{G}_\zeta(\mathbf{R}; x) = 0 \tag{3.47}$$

for all $\mathbf{R}$ implies that $c_\zeta(x) = 0$ for all ζ and x. The linear independence of the force field basis set may be readily determined from the determinant of the MS-CG metric tensor, $G_{\zeta\zeta'}$. Moreover, analysis of $G_{\zeta\zeta'}$ also identifies which basis vectors are linearly dependent and quantifies near degeneracies which can lead to numerical instabilities in determining the MS-CG potentials.

The force field basis vectors are obtained as gradients of the corresponding density fields according to Equation (3.8). Consequently, the linear independence of the MS-CG basis set implies the linear independence of the corresponding set of density fields. Therefore, analysis of the linear independence of the force field basis set via $G_{\zeta\zeta'}$ may prove to be a useful and numerically tractable method for assessing the uniqueness of potentials obtained from relative entropy and other iterative variational methods.

3.5 Results

The present section considers several numerical examples to further investigate the relationship between force- and structure-based CG approaches. In order to facilitate visual analysis of Φ, we consider the case where the CG potential and the corresponding PMF, U^0, are functions of a single variable. We consider first the case that the variable is a Cartesian coordinate. However, calculations performed for the case that the approximate potential depends upon a curvilinear coordinate (i.e., a distance between two particles) lead to quantitatively similar results. Therefore, despite the simplicity of these examples, we expect that the resulting insight should generalize to CG models for more complex molecular systems, in which case Φ can no longer be easily visualized. Without loss of generality, we set $k_BT = 1$ in these calculations.

Due to the simplicity of the PMF, both force- and structure-based methods will determine exactly the same potential, i.e., $U^0(x)$, if a flexible spline basis set is employed. However, in order to investigate differences in the two approaches that arise for calculations employing an incomplete basis set, the present calculations represent the approximate potential with a polynomial of order m:

$$U(x) = \sum_{d=1}^{m} U_d x^d, \tag{3.48}$$

where the potential basis functions are $\rho_d(x) = x^d$ and U_d is the corresponding parameter. Because the $d = 0$ term is not considered and because the ρ_d vanish at the origin, the parameters will be unique for both force- and structure-based CG methods. According to Equation (3.30), the relative entropy functional will be minimized with respect to U_d when the CG model reproduces the d^{th} moment of the atomistic distribution for x. In contrast, the parameters that minimize the MS-CG functional satisfy the discrete analog of the normal equations, Equation (3.37). In general, the resulting MS-CG model is not guaranteed to reproduce any moments of the atomistic distribution. Consequently, we expect that that force- and structure-based models will differ for this incomplete basis set.

3.5.1 Cartesian Coordinates

We consider a CG potential that is described by a single Cartesian coordinate, x, that may represent, e.g., the pulling coordinate of an AFM experiment along a fixed direction. We arbitrarily assume that the "atomistic distribution" of x determined by the underlying

atomistic model and CG mapping is given by:

$$p(x) = A_1 \exp[-(x - x_1)^2/2\sigma_1^2] + A_2 \exp[-(x - x_2)^2/2\sigma_2^2], \tag{3.49}$$

where $A_1, A_2, x_1, x_2, \sigma_1$ and σ_2 are fixed constants. We define the length scale by setting $x_1 = 5$ and $x_2 = 7$; we consider different atomistic distributions by varying σ_1, σ_2, and A_1/A_2. This distribution is convenient because it allows for both asymmetry and bimodality. Consequently, the corresponding 'many-body' PMF $U^0(x) = -\ln p(x)$ cannot be represented by a low order polynomial in x.

3.5.1.1 Harmonic Approximation

We first consider a harmonic approximation to U^0, i.e., $m = 2$ in Equation (3.48). In this case, the parameters for a structure-based model can be directly determined by the standard relations for the mean and variance of a Gaussian distribution: $\langle x \rangle = -U_1/2U_2$ and $\langle (x - \langle x \rangle)^2 \rangle = 1/2U_2$, where the angular brackets denote averages according to the atomistic distribution, $p(x)$. The parameters for the force-based model are determined by solving the normal MS-CG equations. Remarkably, for this case, the MS-CG potential also reproduces the mean and variance of the atomistic distribution. More generally, if the approximate potential is expressed as a quadratic function of Cartesian coordinates (in any number of dimensions), the MS-CG normal equations determine the parameters appropriate for reproducing the means and covariances determined by the atomistic model. Consequently, as long as the approximate CG potential is a quadratic function of Cartesian coordinates, the relative entropy and MS-CG methods determine identical potentials. To the best of our knowledge, this corresponds to the first reported instance where force- and structure-based approaches yielded identical results for an incomplete basis set.

Figure 3.1 presents an example for which the atomistic distribution, $p(x)$, is a bimodal distribution with overlapping Gaussian peaks of similar variance. The solid black curve in Figure 3.1a corresponds to the atomistic distribution. The dotted black curve corresponds to $P_2(x)$, the distribution for the CG model determined by both force- and structure-based approaches when $m = 2$. Table 3.1 compares the first ten moments for $p(x)$ and $P_2(x)$. The dotted black curve in Figure 3.1b presents the corresponding information function $\Phi_2(x) = \ln p(x)/P_2(x)$. Because P_2 smooths over the bimodal features of p, Φ_2 is relatively small near the center of the distribution, has stationary points near the stationary points of p, and becomes increasingly negative in the wings of the distribution. Although Φ_2 assumes negative values, its average with respect to p is the relative entropy, which is necessarily

positive and equals $S_{rel} = 7.0949 \times 10^{-2}$ for P_2. The dotted black curve in Figure 3.1c presents $|d\Phi_2(x)/dx|^2$. The average of this quantity with respect to p determines $\chi^2 = 7.2927 \times 10^{-1}$. (For simplicity, the factor $3N$ is set to 1 and the contribution $\chi^2[U^0]$ set to 0 in this calculation and in the remainder of the section, since neither impacts the calculated potential or resulting distribution.) Figure 3.1d directly compares the 'match' between the resulting harmonic force field (dotted black curve) with the anharmonic 'many-body' mean force field, $F^0(x) = -dU^0(x)/dx$ (solid black curve). The dotted black curve in Figure 3.1c, which was calculated directly from the derivative of the information function, is the square of the difference between these two force fields. Visual inspection confirms that the dotted black curve in Figure 3.1d provides the best linear approximation to the many-body PMF when the weighting is performed with respect to p.

3.5.1.2 Anharmonic Approximation

We next consider an anharmonic approximation, i.e., $m = 4$, to U^0. The potential parameters determined by structure-based approaches will reproduce the first four moments of $p(x)$. These parameters can no longer be computed analytically and were instead iteratively determined by implementing the analog of Equation (3.32) in Mathematica.[193] As before, the potential parameters for the MS-CG model were directly determined by the normal MS-CG equations.

In this case, the force- and structure-based CG approaches no longer determine identical potentials. Figure 3.1a compares the resulting CG distributions with the underlying atomistic distribution, $p(x)$. In Figure 3.1a and in the following figures, the dashed red and dashed-dotted blue curves correspond to structure- and force-based models, respectively, using approximate potentials with $m = 4$. Figures 3.1b and 3.1c compare the corresponding information functions, Φ_4, and the (squared) magnitude of its derivative, $|d\Phi_4/dx|^2$.

Expanding the basis set from $m = 2$ to $m = 4$ leads to systematic improvement for both methods, as would be expected from variational methods. Figure 3.1b demonstrates that the structure-based model minimizes S_{rel} relative to the force-based model by better reproducing the height of the first peak at $x = 5$ and also the minima located near $x \approx 6.25$. Consequently, $S_{rel} = 2.1118 \times 10^{-2}$ for the structure-based model, as compared to $S_{rel} = 2.4727 \times 10^{-2}$ for the force-based model. Figure 3.1c demonstrates that the force-based model minimizes χ^2 by better reproducing the more rapidly varying wings of the distribution. As a result, $\chi^2 = 3.8926 \times 10^{-1}$ for the force-based model, as compared to $\chi^2 = 4.3677 \times 10^{-1}$ for the structure-based model. Figure 3.1d directly compares the corresponding force fields and reinforces the conclusions from Figure 3.1c. The structure-based model better reproduces

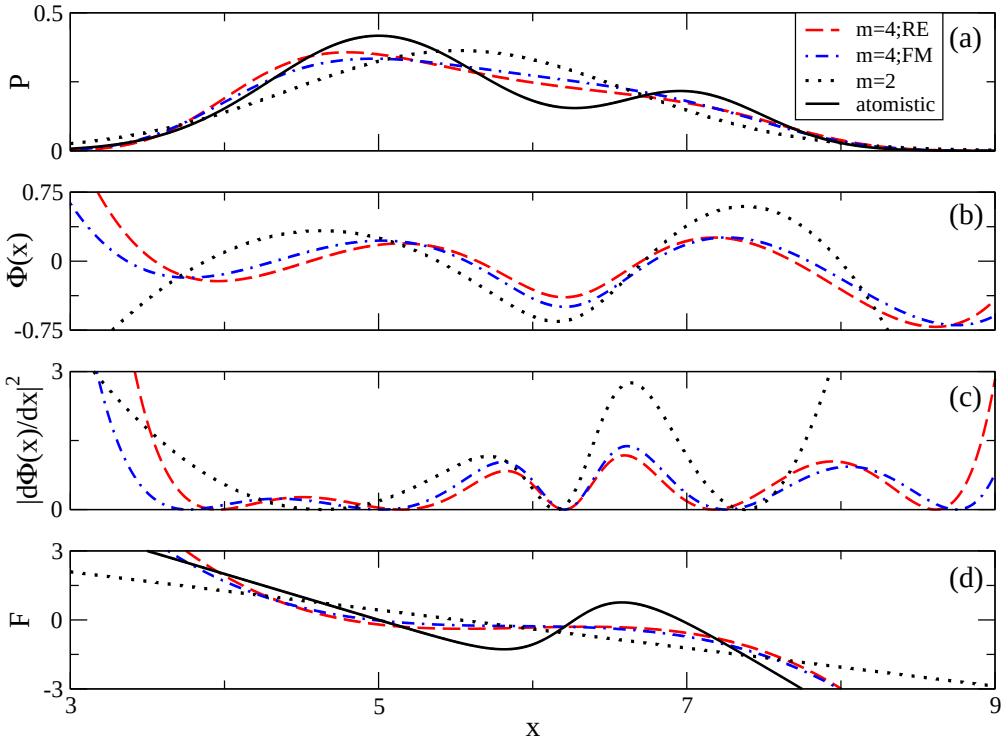

Figure 3.1. Analysis of distributions for structure- and force-based models with approximate potentials corresponding to polynomials of order $m = 2$ and $m = 4$. The solid black and dashed black curves correspond to the underlying high resolution distribution, $p(x)$, (defined by Equation (3.49)) and to results for CG models with harmonic approximate potentials, respectively. The dashed red and dashed-dotted blue curves correspond to the structure-based and force-based models, respectively, using approximate CG potentials with $m = 4$. Panel (a) compares the atomistic and CG distribution functions; panel (b) presents the corresponding information functions, $\Phi(x)$; panel (c) presents its squared gradient, $|d\Phi(x)/dx|^2$; and panel (d) presents the corresponding force functions, $-dU(x)/dx$.

the "many-body" mean force field (solid black) in the center of the distribution, but less accurately reproduces the mean force near the wings.

Table 3.1 presents the percent error in the first ten moments for the two models. The structure-based model quantitatively reproduces the first four moments of p by construction and also reproduces the next six moments to within a fraction of a percent. Somewhat surprisingly, the mean and standard deviation of the force-based model, which were quantitatively reproduced when $m = 2$, are only accurate to 1% when the basis set is expanded to $m = 4$. However, expanding the basis set significantly improves the accuracy of the force-based model to within 2% accuracy for all ten moments. The force-based model reproduces the higher moments with increasing accuracy and, in particular, reproduces the ninth and tenth moments more accurately than the structure-based model.

The results discussed above are typical for parameter sets corresponding to atomistic

Table 3.1. Comparison of distributions for the CG coordinates determined from the high resolution (AA) model and also from CG models determined by force- (FM) and structure- (RE) based approaches using approximate potentials with $m = 2$ and $m = 4$ in Equation (3.48). The first ten rows present the first ten moments for the high resolution distribution and also the percent error in the corresponding moments for each CG model. The last three rows present χ^2 (with $3N = 1$ and $\chi^2[U^0] = 0$), S_{rel}, and the entropy of each model ($S = -\langle \ln P \rangle$).

	AA	m=2	m=4	
		RE ≡ FM	RE	FM
$\langle x \rangle$	5.5224×10^{0}	0.0000	0.0000	0.5831
$\langle x^2 \rangle$	3.1704×10^{1}	0.0000	0.0000	1.0566
$\langle x^3 \rangle$	1.8886×10^{2}	0.2383	0.0000	1.3608
$\langle x^4 \rangle$	1.1641×10^{3}	0.7645	0.0000	1.4689
$\langle x^5 \rangle$	7.4001×10^{3}	1.5027	0.0095	1.3878
$\langle x^6 \rangle$	4.8334×10^{4}	2.3007	0.0476	1.1607
$\langle x^7 \rangle$	3.2318×10^{5}	2.9798	0.1392	0.8509
$\langle x^8 \rangle$	2.2046×10^{6}	3.3657	0.3130	0.5216
$\langle x^9 \rangle$	1.5297×10^{7}	3.3013	0.5884	0.2353
$\langle x^{10} \rangle$	1.0768×10^{8}	2.6560	0.9937	0.0371
χ^2	0.0000×10^{0}	7.2927×10^{-1}	4.3677×10^{-1}	3.8926×10^{-1}
S_{rel}	0.0000×10^{0}	7.0949×10^{-2}	2.1118×10^{-2}	2.4727×10^{-2}
S	1.4419×10^{0}	1.5128×10^{0}	1.4630×10^{0}	1.4724×10^{0}

distributions with overlapping peaks of similar variance. In this case, the difference between the structure and force-based approaches is subtle, but can be understood in terms of the variational principles underlying the two approaches. However, the resulting models can deviate more if the peaks in the underlying atomistic distribution function no longer overlap. Figure 3.2 presents three examples for this case. As in Figure 3.1, the solid black curve corresponds to the atomistic distribution and the dashed black curve corresponds to the distribution obtained from both force- and structure-based models when $m = 2$. The dashed red and dashed-dotted blue curves correspond to distributions obtained from structure- and force-based approaches, respectively, when $m = 4$.

Figure 3.2a considers the case that the atomistic distribution consists of two non-overlapping Gaussians of the same height and variance. In this case, the structure-based model accurately reproduces the peak heights of the atomistic distribution and balances the errors in reproducing the peak positions, the vanishing minima at $x = 6$, and the tails of the distribution. In contrast, the force-based model more accurately matches the tails of each peak in the atomistic distribution (especially the outer tails), but ignores the center of the distribution where $|d\Phi/dx|^2 \sim 0$. Consequently, the force-based model matches the outer tails and

peak positions well, but significantly underestimates the peak heights and overestimates the center of the distribution.

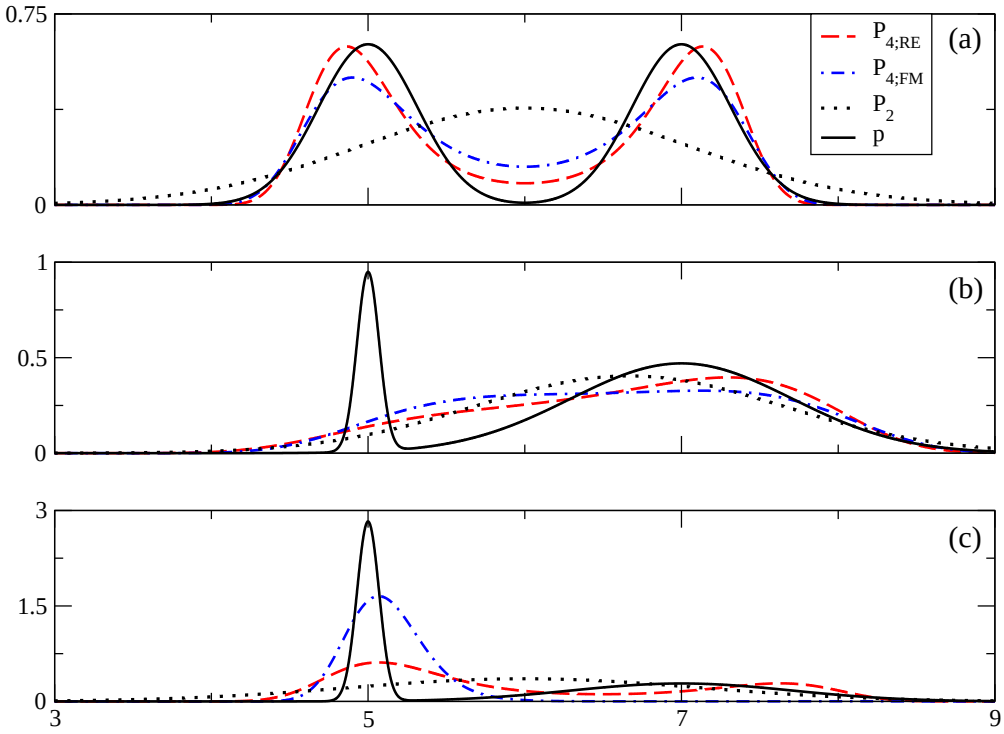

Figure 3.2. Comparison of distributions for a high resolution model (defined by Equation (3.49)) and for CG models determined by minimizing either S_{rel} (structure-based models) or χ^2 (force-based models). The solid black and dashed black curves correspond to the underlying atomistic distribution, $p(x)$, and to distributions for CG models with harmonic potentials, $P_2(x)$, respectively. The dashed red and dashed-dotted blue curves correspond to structure-based and force-based CG models, respectively, using approximate potentials corresponding to polynomials of order $m = 4$. The three panels correspond to different parameters sets for $p(x)$ - Panel (a): $\sigma_1^2 = 0.1$, $\sigma_2^2 = 0.1$, $A_1/A_2 = 1$; Panel (b): $\sigma_1^2 = 0.01$, $\sigma_2^2 = 1$, $A_1/A_2 = 2$; Panel (c): $\sigma_1^2 = 0.01$; $\sigma_2^2 = 1$; $A_1/A_2 = 10$.

Figures 3.2b and 3.2c consider cases for which the atomistic distribution consists of non-overlapping peaks with different heights and variance. In these cases, neither approach reproduces the atomistic distribution very well. In Figure 3.2b the first peak of $p(x)$ is 100 times more narrow and twice as tall as the second peak. Both approaches determine distributions that span both peaks in $p(x)$ and emphasize the second wider peak. However, because of the large derivative associated with the more narrow peak, the force-based model weights the first narrow peak slightly more . The force-based model more accurately reproduces the tails of the atomistic distribution, but the structure-based model more accurately reproduces the second wider peak.

In Figure 3.2c the first peak of $p(x)$ is 100 times more narrow and also 10 times larger than the second peak. Because of the large slope of the first peak, the force-based model shifts even more towards this peak and completely disregards the majority of the second peak. In contrast, in order to reproduce the low order moments of the atomistic distribution, the structure-based model more accurately describes the second peak and, consequently, models the larger peak less accurately.

3.5.2 Curvilinear Coordinates

In addition, we also performed numerical calculations for the case that the approximate CG potential depends upon the distance between a pair of particles. In this case, force- and structure-based CG approaches no longer give identical results for the case that the potential is a harmonic function of distance. However, the differences between the resulting harmonic potentials are numerically insignificant for the cases that we considered. Moreover, we also performed calculations corresponding to Figure 3.2 for the case that the approximate potential is an anharmonic function of interparticle distance. The results are virtually identical to those presented in Figure 3.2 for the case that the CG potential is a function of Cartesian coordinates.

In summary, our calculations suggest that structure- and force-based approaches lead to similar results for a wide range of distributions. If the approximate CG potential is a quadratic function of Cartesian coordinates, then both approaches obtain exactly identical results. If the approximate CG potential is a quadratic function of distances, then both approaches obtain nearly indistinguishable results for the cases that we considered. As the approximate potential becomes increasingly flexible, then both approaches become increasingly accurate. In many cases, the results of the two approaches are quite similar. In cases that the two methods significantly differ, the resulting models emphasize different aspects of the underlying atomistic distribution, according to the theory discussed above.

3.6 Discussion

CG models provide a highly efficient means for investigating complex processes that cannot be effectively studied with more detailed models. These models are frequently parameterized to reproduce structural properties that are determined from either experiments or from high resolution simulations. Structure-based approaches, such as the relative entropy method,[37,121] employ multiple CG simulations to reproduce target structural correlation

functions. In contrast, the MS-CG method[36,41,59,131] employs forces sampled from atomistic simulations to directly project the many-body MF onto a given basis set for the CG force field, but the resulting force field is not guaranteed to reproduce any particular target correlation functions.[41] If the CG potential is represented with a complete basis set, then both approaches determine the same potential, i.e., the many-body PMF, and will quantitatively reproduce all structural features of the atomistic model.[110] In practice, though, the CG potential is represented with a highly incomplete basis set, typically of a molecular mechanics form. In this case, the MS-CG and relative entropy methods will generally determine different approximations to the many-body PMF.[74]

Despite their obvious differences, the present work reveals remarkable similarities in the framework underlying force- and structure-based approaches to CG modeling. Both the MS-CG and relative entropy approaches determine CG potentials through variational calculations that are typically performed in a linear space spanned by a highly incomplete basis set. The force field basis vectors employed in the MS-CG approach correspond to gradients of the potential basis functions employed in the relative entropy approach. Most significantly, the functionals employed in the two variational calculations can both be expressed in terms of an information function, $\Phi(\mathbf{R})$, that discriminates between the distributions of configurations sampled by the atomistic and CG models. Shell originally defined an equivalent relative entropy as the average of an analogous information function defined on the atomistic configuration space.[37] The present work demonstrates that the MS-CG method minimizes the average $|\nabla\Phi(\mathbf{R})|^2$.

The present work generalizes the well-known result of Henderson[46] by identifying conditions for the uniqueness of CG potentials that reproduce a set of low order correlation functions. The potentials obtained from the structure- and force-based methods can only be unique if the corresponding basis vectors are linearly independent. The linear independence of basis functions for structure-based methods may be relatively difficult to test a priori. However, the linear independence of the MS-CG basis vectors may be directly tested by the metric tensor $G_{\zeta\zeta'}$, which can be readily calculated from sampled atomistic configurations.[60] Moreover, the linear independence of the MS-CG basis vectors implies the linear independence of the corresponding basis functions for structure-based potentials.

The present numerical calculations further probe the relationship between force- and structure-based methods. These calculations considered particularly simple models for which the atomistic distribution can be readily visualized and the consequences of an incomplete basis set investigated. Our calculations and analysis demonstrate that force- and structure-based approaches determine the same potential, not only in the limit of a complete basis

set, but also in the case that the potential is a quadratic function of Cartesian coordinates. However, the methods may obtain different results if the basis set is expanded to include higher order polynomials or if the approximate potential depends upon non-Cartesian coordinates, e.g., interparticle distances. Given a basis set described by m^{th}-order polynomials, the relative entropy method determines the parameters that quantitatively reproduce the first m moments of the atomistic distribution. Neither method is guaranteed to reproduce any higher order moments. Our results suggest that, depending upon the atomistic distribution, either the MS-CG or the relative entropy method may provide a more accurate description of these higher order moments. The MS-CG approach appears to more accurately reproduce the wings and rapidly varying regions of the atomistic distribution, but may be less sensitive to more slowly varying regions of the distribution. Nevertheless, our numerical investigations suggest that the two approaches lead to similar results for a wide range of atomistic distributions.

The present work also addressed several other aspects of force- and structure-based CG approaches. We reformulated and generalized the MS-CG and g-YBG approaches for more complex potentials that are functions of multiple order parameters. We demonstrated the equivalence of two definitions for the relative entropy that have been expressed as averages over the atomistic[37] or CG[190] configuration space. This analysis suggested an intriguing physical significance for the mapping entropy as the difference in the configurational entropy of the atomistic model when viewed from either the atomistic or CG configuration space. This entropy difference has a rigorous upper bound that is likely nonnegative and vanishes for the case that each site corresponds to a single atom.

We decomposed this entropy difference into two contributions. One contribution describes the geometry of the CG mapping and may be calculated analytically as a function of mapping coefficients. The second contribution results from the averaging of fluctuations in the atomistic distribution due to integrating out degrees of freedom. This second contribution is nonnegative and only vanishes when all the configurations that map to the same CG configuration have equal weight in the atomistic model. We emphasize that this entropy difference is independent of the CG potential and applies, in particular, for a perfectly consistent CG model, i.e., one for which $p_R = P_R$. Because of its significance for transferability,[146,157] representability,[39] and phase transitions, the mapping entropy may be a useful metric for optimizing CG mappings.

As mentioned above, the present work generalized the noted uniqueness result of Henderson[46] by identifying conditions for the uniqueness of structure-derived potentials. However, this result assumes the existence of CG potentials reproducing a set of distribution functions.

It is much more difficult to prove this existence and relatively little progress has been achieved in these directions.[127,128] Furthermore, although the present result is exact, it provides little immediate insight into the sensitivity of the potentials to relatively small changes in the corresponding distribution functions.[194] A number of studies have demonstrated that CG distribution functions are quantitatively similar for a wide range of potentials that differ in, e.g., the long-ranged tail of nonbonded potentials, and this insensitivity has been exploited to optimize thermodynamic properties.[35,195] We also note that the present work specifically addresses the canonical ensemble and does not directly address other ensembles or issues of transferability. Finally, we reiterate that the continuous results presented in this work were derived as the continuum limit of discrete results.

3.7 Concluding Remarks

The present work investigated the relationship between force- and structure-motivated approaches to developing CG potentials. Most significantly, our analysis demonstrated that both force- and structure-based methods can be expressed in terms of an information function that discriminates between the ensembles sampled by atomistic and CG models. We generalized the well-known result of Henderson by identifying conditions for the uniqueness of structure-based potentials. Furthermore, we demonstrated the relationship of these conditions to analogous conditions for force-based potentials. In the course of this analysis, we also generalized the MS-CG and g-YBG formalisms and also investigated the physical significance of the mapping entropy. We demonstrated that force- and structure-based methods obtain the same potential, not only in the limit of a complete basis set, but also when the approximate potential is a quadratic function of Cartesian coordinates. For more complex, but still incomplete, basis sets the two methods obtain different potentials, although the results are often quite similar.

The present work also indicates several directions for future study. Future investigations may gain greater insight into the relationship between structure-based and force-based models by carefully considering the information function $\Phi(\mathbf{R})$ and the operator, $\nabla\Phi(\mathbf{R})\cdot\nabla$. In particular, future numerical studies should be expanded for more complex model systems. In addition, we anticipate that the mapping entropy may be a useful metric for optimizing the CG mapping and for addressing the thermodynamic properties of structure-motivated models. These considerations may prove particularly important for determining CG protein models from known structural properties.

3.8 Appendix

3.8.1 More General Potentials

Recent analyses of the MS-CG and g-YBG approach have treated χ^2 as a functional of CG force functions.[134,188] While this treatment is appropriate when the terms in the CG potential depend upon a single scalar variable, it is not adequate for the case that the potential includes terms depending upon two or more independent variables. If a term in the CG potential depends upon multiple independent variables, e.g., an angle and a bond distance, then partial derivatives with respect to these variables lead to distinct force functions that can no longer be varied independently. In this case it is more convenient to perform variational calculations with the CG potential functions as the independent variables. The present appendix briefly demonstrates how the MS-CG and g-YBG formalisms may be readily generalized for potential functions that depend upon multiple variables. For convenience, we assume that the CG potential includes only one type of interaction. Distinct types of interactions can be readily treated by introducing an additional subscript (ζ) to distinguish them as in Equation (3.5). We note that Voth and coworkers have previously applied the MS-CG method to potentials depending upon the coordinates of three non-bonded particles.[133]

We consider the following potential:

$$V(\mathbf{R}) = \sum_{\lambda} U(\boldsymbol{\psi}(\{\mathbf{R}\}_\lambda)) = \int d\mathbf{x}\, U(\mathbf{x}) \hat{\rho}(\mathbf{R};\mathbf{x}) \tag{3.50}$$

where

$$\hat{\rho}(\mathbf{R};\mathbf{x}) = \sum_{\lambda} \delta(\boldsymbol{\psi}_\lambda(\mathbf{R}) - \mathbf{x}), \tag{3.51}$$

$\boldsymbol{\psi}$ denotes a set of scalar functions $\{\psi_\alpha\}$ for $\alpha = 1, 2, \ldots$, so that U can depend upon 2 or more different types of variables, $\boldsymbol{\psi}_\lambda(\mathbf{R}) = \boldsymbol{\psi}(\{\mathbf{R}\}_\lambda)$, and $\delta(\boldsymbol{\psi}_\lambda(\mathbf{R}) - \mathbf{x}) = \prod_\alpha \delta(\psi_{\lambda\alpha}(\mathbf{R}) - x_\alpha)$. The CG force vector can then be expressed as a linear combination of basis vectors as before:

$$\mathbf{F} = \int d\mathbf{x}\, U(\mathbf{x})\ \boldsymbol{\mathscr{G}}(\mathbf{x}) \tag{3.52}$$

where the basis vectors have elements:

$$\boldsymbol{\mathscr{G}}_I(\mathbf{R};\mathbf{x}) = -\nabla_I\, \hat{\rho}(\mathbf{R};\mathbf{x}) \tag{3.53}$$

$$= \sum_{\lambda} \sum_{\alpha} \left(\nabla_I \psi_{\lambda\alpha}(\mathbf{R})\right) \frac{\partial}{\partial x_\alpha} \delta(\boldsymbol{\psi}_\lambda(\mathbf{R}) - \mathbf{x}), \tag{3.54}$$

$\nabla_I = \partial/\partial \mathbf{R}_I$, and the sum over α corresponds to contributions from forces arising from each scalar variable $\psi_{\lambda\alpha}(\mathbf{R})$. The MS-CG functional χ^2 may be obtained by the substitution $x \to \mathbf{x}$ and the resulting normal equations are:

$$b(\mathbf{x}) = \int d\mathbf{x}'\, G(\mathbf{x};\mathbf{x}')\, U(\mathbf{x}'), \tag{3.55}$$

with $b(\mathbf{x}) = \boldsymbol{\mathscr{G}}(\mathbf{x}) \odot \mathbf{F}^0$ and $G(\mathbf{x},\mathbf{x}') = \boldsymbol{\mathscr{G}}(\mathbf{x}) \odot \boldsymbol{\mathscr{G}}(\mathbf{x}')$ defined as before. When expressed in terms of correlation functions, the dependence upon multiple variables results in additional partial derivatives:

$$\begin{aligned} b(\mathbf{x}) &= \frac{1}{3N}\sum_\alpha \frac{\partial}{\partial x_\alpha}\int d\mathbf{r}\; p_r(\mathbf{r}) \sum_\lambda \sum_I \Big(\mathbf{f}_I(\mathbf{r}) \cdot \nabla_I \psi_{\lambda\alpha}(\mathbf{M}(\mathbf{r}))\Big) \\ &\quad \times \delta(\boldsymbol{\psi}_\lambda(\mathbf{M}(\mathbf{r})) - \mathbf{x}) \qquad (3.56) \\ G(\mathbf{x},\mathbf{x}') &= \frac{1}{3N}\sum_\alpha \sum_{\alpha'} \frac{\partial^2}{\partial x_\alpha \partial x'_{\alpha'}}\int d\mathbf{R}\; p_R(\mathbf{R}) \sum_\lambda \sum_{\lambda'} \Big(\nabla\psi_{\lambda\alpha}(\mathbf{R}) \cdot \nabla\psi_{\lambda'\alpha'}(\mathbf{R})\Big) \\ &\quad \times \delta(\boldsymbol{\psi}_\lambda(\mathbf{R}) - \mathbf{x})\delta(\boldsymbol{\psi}_{\lambda'}(\mathbf{R}) - \mathbf{x}'), \qquad (3.57) \end{aligned}$$

where $\nabla A(\mathbf{R}) \cdot \nabla B(\mathbf{R}) = \sum_I \nabla_I A(\mathbf{R}) \cdot \nabla_I B(\mathbf{R})$.

As before the projection of the mean force onto the basis vector, $b(\mathbf{x})$, can be re-expressed in terms of structural correlation functions by straightforward integration by parts and by using $p_R(\mathbf{R})$, the probability of finding an atomistic configuration that maps to the CG configuration $\mathbf{R}$:

$$\begin{aligned} b(\mathbf{x}) &= \frac{k_B T}{3N}\int d\mathbf{R}\; p_R(\mathbf{R}) \sum_I \left(-\frac{\partial}{\partial \mathbf{R}_I} \cdot \boldsymbol{\mathscr{G}}_I(\mathbf{R};x)\right) \qquad (3.58) \\ &= k_B T\left[g(\mathbf{x}) - L(\mathbf{x})\right], \qquad (3.59) \end{aligned}$$

in terms of

$$\begin{aligned} g(\mathbf{x}) &= \frac{1}{3N}\sum_{\alpha,\alpha'} \frac{\partial^2}{\partial x_\alpha \partial x_{\alpha'}} \\ &\quad \times \left\langle \sum_\lambda \Big(\nabla\psi_{\lambda\alpha}(\mathbf{M}(\mathbf{r})) \cdot \nabla\psi_{\lambda\alpha'}(\mathbf{M}(\mathbf{r}))\Big)\, \delta(\boldsymbol{\psi}_\lambda(\mathbf{M}(\mathbf{r})) - \mathbf{x}) \right\rangle \qquad (3.60) \\ L(\mathbf{x}) &= \frac{1}{3N}\sum_\alpha \frac{\partial}{\partial x_\alpha} \left\langle \sum_\lambda \nabla^2\psi_{\lambda\alpha}(\mathbf{M}(\mathbf{r}))\; \delta(\boldsymbol{\psi}_\lambda(\mathbf{M}(\mathbf{r})) - \mathbf{x}) \right\rangle. \qquad (3.61) \end{aligned}$$

Equations (3.55) and (3.59) determine a g-YBG equation as in previous work.[63,134,188]

3.8.2 Proof of Uniqueness for Structure-based Potentials

We consider two CG potential U_A and U_B expressed in terms of a set of potentials $\{U_{\zeta A}(x)\}$ and $\{U_{\zeta B}(x)\}$ (that satisfy suitable boundary conditions and that correspond to linearly independent fields) and corresponding distribution functions $\{P_{\zeta A}(x)\}$ and $\{P_{\zeta B}(x)\}$. In subsection 3.4.2.1, we asserted that if these sets of distribution functions are equal, then the two sets of potentials are equal. In this appendix, we shall prove this result by proving its contrapositive, i.e., that if the two potentials differ in any term, $U_\zeta(x)$, then the resulting distribution functions also differ.

The Gibbs inequality[135] implies that:

$$\int d\mathbf{R}\; P_R(\mathbf{R}|U_A)\left(U_A(\mathbf{R}) - U_B(\mathbf{R})\right) \leq F[U_A] - F[U_B], \tag{3.62}$$

where $F[U]$ is the configurational free energy defined in Equation (3.4). Adding this result to the symmetric inequality that is obtained from averaging with respect $P_R(\mathbf{R}|U_B)$ leads to

$$\Psi = \sum_\zeta \int dx\; \Delta U_\zeta(x)\Delta P_\zeta(x) \leq 0, \tag{3.63}$$

where $\Delta U_\zeta(x) = U_{\zeta A}(x) - U_{\zeta B}(x)$ and $\Delta P_\zeta(x) = P_{\zeta A}(x) - P_{\zeta B}(x)$. The equality is obtained in Equations (3.62) and (3.63) if and only if $P_R(\mathbf{R}|U_A) = P_R(\mathbf{R}|U_B)$ for all $\mathbf{R}$.

The proof proceeds as follows: By hypothesis, we assume that the two potentials differ in some particular term, i.e., $\Delta U_\zeta(x) \neq 0$ for some ζ. From the assumed linear independence of the density fields, it follows that $U_A(\mathbf{R}) \neq U_B(\mathbf{R})$ for some $\mathbf{R}$. The assumed boundary conditions for U_A and U_B then imply that there exist configurations for which $P_R(\mathbf{R}|U_A) \neq P_R(\mathbf{R}|U_B)$ and the Gibbs inequality implies that $\Psi < 0$. Therefore, $\Delta P_{\zeta'}(x') \neq 0$ for some ζ' and some x'.

Thus, if the two potentials differ in some particular term, then the resulting distribution functions differ. The contrapositive must also hold true, which proves the assertion in subsection 3.4.2.1. Assuming suitable boundary conditions and the linear dependence of the density fields included in the approximate potential, the set of potentials that reproduce a given set of conjugate correlation functions is unique.

Chapter 4

The Role of Many-Body Correlations in Determining Potentials for Coarse-Grained Models of Equilibrium Structure

J. F. Rudzinski, W. G. Noid *J. Phys. Chem. B* **2012**, 116, 8621-8635

Abstract

Coarse-grained (CG) models often employ pair potentials that are parameterized to reproduce radial distribution functions (rdfs) determined for an atomistic model. This implies that the CG model must reproduce the corresponding atomistic mean forces. These mean forces include not only a direct contribution from the corresponding interaction, but also correlated contributions from the surrounding environment. The many-body correlations that influence this second contribution present significant challenges for accurately reproducing atomistic distribution functions. This work presents a detailed investigation of these many-body correlations and their significance for determining CG potentials, while using liquid heptane as a model system. We employ a transparent geometric framework for directly determining CG potentials that has been previously developed within the context of the multiscale coarse-graining and generalized-Yvon-Born-Green methods. In this framework, a metric tensor quantifies the relevant many-body correlations and precisely decomposes atomistic mean forces into contributions from specific interactions, which then determine the CG force field. Numerical investigations reveal that this metric tensor reflects both the CG representation and also subtle correlations between molecular geometry and intermolecular packing, but can be largely interpreted in terms of generic considerations. Our calculations demonstrate that contributions from correlated interactions can significantly impact the pair mean force and, thus, also the CG force field. Finally, an eigenvector analysis investigates the importance of these interactions for reproducing atomistic distribution functions.

4.1 Introduction

The complexity and computational expense of atomically detailed models has motivated considerable interest in coarse-grained (CG) models that represent systems in reduced detail.[82,88,110] However, the practical utility of these models relies upon potentials that accurately describe the interactions in the low resolution CG model.[96] While CG models that are parameterized with thermodynamic data[27,28,98,102] have provided important insight into many processes,[196,197] these models may not necessarily provide a quantitative description of structural properties.[30–33]

The correct potential for a CG model that quantitatively reproduces all structural properties of a particular atomistic model (at the resolution of an associated CG mapping) is a configuration-dependent free energy function.[109] This function is defined by the many-body probability distribution for the atomistic model to sample CG configurations (according to this CG mapping).[110] This function is referred to as a many-body potential of mean force (PMF) because its (negative) gradients define a many-body force that equals the mean atomistic force averaged over all configurations that map to the given CG configuration.[44] Simulations with this many-body mean force (MF) field will quantitatively reproduce the distribution of CG configurations sampled by the atomistic model.[41]

In general, though, the many-body PMF cannot be readily calculated, represented, or simulated.[111] Instead, structure-motivated coarse-graining methods typically approximate the PMF with much simpler molecular mechanics potentials. Each term in this approximate potential models a particular interaction with a function of a single scalar variable, e.g., the distance between a pair of sites or the angle formed by three bonded sites. The atomistic model (along with the CG mapping) defines a target probability distribution for each of these scalar variables. The approximate CG potential is then parameterized to reproduce these atomistic distributions.

Each of these distributions defines a corresponding potential of mean force (or torque[198]) along that single degree of freedom. For instance, in the case of nonbonded pair potentials, the distribution of pair distances or, equivalently, the radial distribution function (rdf), determines a pair potential of mean force. The (negative) derivative of this pair potential of mean force defines the pair mean force, which equals the average net force on a particle when a second particle is a given distance away.[166] Importantly, if a CG model reproduces the pair mean force of the atomistic model as a function of distance, then the CG model will also reproduce the corresponding atomistic rdf. Consequently, the objective of reproducing a set of atomistic rdfs is equivalent to the objective of reproducing a corresponding set

of atomistic mean forces. This highlights an important correspondence between iterative approaches that are often expressed in terms of rdfs, (e.g., Iterative Boltzmann Inversion[35] and the Inverse Monte Carlo method[34]) and noniterative approaches (e.g., the Multiscale Coarse-graining[36,59] and generalized-Yvon-Born-Green[63,64] methods) that can be expressed in terms of the corresponding mean forces.[134,188] The present work considers structure-motivated coarse-graining from the perspective of reproducing these mean forces.

To briefly introduce the significance of many-body correlations upon the pair mean force, let us consider a monatomic fluid with density, ρ, interacting via pair potentials, $U^{(2)}$. The mean force on particle 1 at position $\mathbf{r}_1$ when a second particle is located at $\mathbf{r}_2$ includes both direct and environment-mediated contributions. The direct contribution to this mean force from particle 2 is simply $-\nabla_1 U^{(2)}(\mathbf{r}_1, \mathbf{r}_2)$. However, the environment-mediated contribution reflects many-body correlations and is significantly more complicated. The presence of the two particles impacts the packing of surrounding particles, which are distributed at a position $\mathbf{r}_3$ with a probability $p_{3|2}(\mathbf{r}_3|\mathbf{r}_1, \mathbf{r}_2) = \rho g^{(3)}(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3)/g^{(2)}(\mathbf{r}_1, \mathbf{r}_2)$, where $g^{(2)}$ and $g^{(3)}$ are conventional two- and three-body correlation functions.[43,135] The Yvon-Born-Green (YBG) integral equation[43,135] provides a transparent physical picture for decomposing the mean force on particle 1 (i.e., $-\nabla_1 w^{(2)}(\mathbf{r}_1, \mathbf{r}_2) = \nabla_1 k_B T \ln g^{(2)}(\mathbf{r}_1, \mathbf{r}_2)$) into direct and environment-mediated contributions:

$$-\nabla_1 w^{(2)}(\mathbf{r}_1, \mathbf{r}_2) = -\nabla_1 U^{(2)}(\mathbf{r}_1, \mathbf{r}_2) + \int d\mathbf{r}_3 \left(-\nabla_1 U^{(2)}(\mathbf{r}_1, \mathbf{r}_3)\right) p_{3|2}(\mathbf{r}_3|\mathbf{r}_1, \mathbf{r}_2). \tag{4.1}$$

The net average force from these surrounding particles (i.e., the integral term in Equation (4.1)) drives the formation of solvation shells and generates the observed oscillations in the mean force, which correspond to the oscillations in the rdf.

Therefore, even in the simplest (and most common case) that the CG nonbonded potential includes only 2-body interactions, many-body correlations significantly impact the pair mean force and, thus, also the corresponding rdf. Some studies have addressed these effects with more complex, many-body CG potentials, although this can significantly reduce the efficiency of the model.[199,200] Regardless, any CG model that seeks to accurately reproduce atomistic rdfs must address the many-body correlations that generate the environment-mediated contribution to the mean force.

Various structure-motivated CG approaches differ in their treatment of these many-body correlations. If the density of CG sites is sufficiently low, then the environment-mediated contribution to the mean force vanishes so that the pair mean force in the CG model only reflects the corresponding direct CG force. In this case, direct Boltzmann inversion deter-

mines a CG potential that accurately reproduces the atomistic rdf.[113] In some cases, this contribution may be treated via a density expansion.[201,202] More generally, though, the contributions of many-body correlations to the CG mean forces are treated via iterative methods that perform simulations with trial CG potentials, compare the resulting CG distributions with target atomistic distributions, and then use this information to systematically refine the CG potentials.[34,35,37,38,50,53,118–124]

The Multiscale Coarse-graining (MS-CG) method[36,59] adopts a considerably different strategy. Each term in the MS-CG potential defines a corresponding force field basis vector.[41,60] As illustrated in Figure 4.1, the MS-CG force field, $\mathbf{F}$, is defined to match the projections of the many-body MF, $\mathbf{F}^0$, onto each vector, $\mathscr{G}_\zeta$, in the basis set. Because these basis vectors correspond to correlated interactions, the projections of the MS-CG force field reflect contributions from multiple basis vectors. A metric tensor, $G_{\zeta\zeta'}$, defines the "angle" between basis vectors based upon their correlation. This angle performs a central role in geometrically decomposing the projections of the many-body MF into contributions from individual MS-CG potentials.

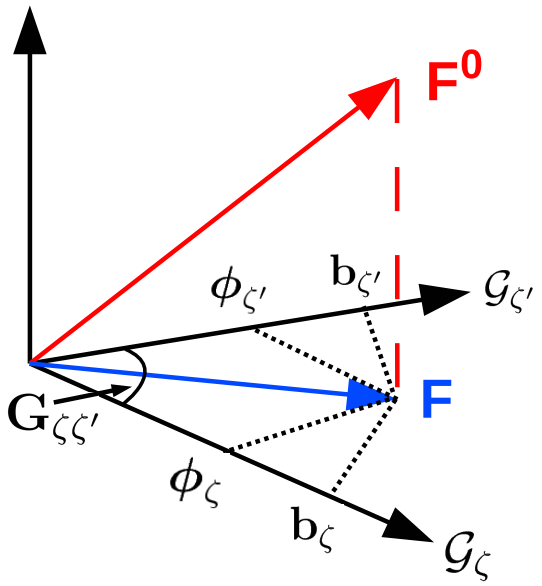

Figure 4.1. Schematic of the MS-CG procedure. The MS-CG force field, $\mathbf{F}$, is the projection of the many-body mean force, $\mathbf{F}^0$, onto the subspace spanned by the given basis set, including $\mathscr{G}_\zeta$ and $\mathscr{G}_{\zeta'}$. The coefficients, $\{\phi_\zeta\}$, of these basis vectors are determined by requiring that $\mathbf{F}$ and $\mathbf{F}^0$ have the same projections, $\{b_\zeta\}$, onto each basis vector. Since these basis vectors may correspond to correlated interactions, they form a skew coordinate system with a metric tensor $G_{\zeta\zeta'} = \mathscr{G}_\zeta \odot \mathscr{G}_{\zeta'}$.

Our group has demonstrated that, in the case of nonbonded pair interactions, the corresponding projections of the many-body MF are equivalent to pair mean forces.[134,188] Moreover, we have demonstrated that the "normal equations" for the MS-CG force field[60,62] are equivalent to a generalized-Yvon-Born-Green (g-YBG) integral equation for molecular mechanics potentials.[63,64] This g-YBG equation generalizes Equation (4.1) by directly decomposing atomistic pair mean forces into quantitative contributions from each term in the CG potential. This equation provides a computational g-YBG framework for determining the MS-CG "force-matched" forces directly from structures, i.e., without force information.

In summary, because they significantly impact pair mean forces (and thus also the resulting rdfs), all structure-motivated methods must address the many-body correlations that couple interactions in the CG model. The MS-CG metric tensor provides a direct and particularly transparent geometric framework for considering their role in determining potentials for CG models that accurately model molecular structure. Motivated by these considerations, the present work reports a detailed analysis of this metric tensor. In particular, we investigate and characterize the many-body correlations that impact the metric tensor for molecular liquids, while using a three-site model of heptane as a model system. Our analysis identifies the key robust features of this metric tensor and precisely elucidates their origin in both structural features of molecular liquids and in the CG representation. We employ the metric tensor to quantify the contributions of correlated interactions to pair mean forces. Moreover, we perform eigenvector/eigenvalue analysis to identify which correlations are most significant for determining the MS-CG force field and for reproducing atomistic structure. Our calculations and analysis significantly expand upon previous studies that have briefly considered the MS-CG metric tensor.[62,69] In addition, by focusing on the importance of mean forces in determining potentials for CG models that accurately repoduce atomistic structure, this work complements several previous studies that have compared both formal[48,79,121] and practical[74] aspects of force- and structure-based approaches to CG modeling.

The remainder of the paper is organized as follows. Section II briefly develops relevant aspects of the MS-CG and g-YBG theory. Section III summarizes the key details of our calculations, which are provided in greater detail in the Supporting Information section. Section IV presents a detailed analysis of the metric tensor. Section V discusses these results in the context of other recent efforts. Finally, Section VI summarizes the main conclusions of this work and indicates possible future directions.

4.2 Theory

We briefly summarize relevant aspects of the MS-CG[41,60,62,68,71,72,131,133,157,203] and g-YBG [63,64,134,188] theories and introduce appropriate notation. This work explicitly considers the canonical ensemble for a system with temperature T and volume V.

The configuration of an atomistic model is defined by the Cartesian coordinates, $\mathbf{r}$, for n atoms. Similarly, the configuration of a CG model for the same system is defined by the Cartesian coordinates, $\mathbf{R}$, for N sites. A mapping function, $\mathbf{M}$, determines a CG configuration as a linear function of the atomic configuration: $\mathbf{R} = \mathbf{M}(\mathbf{r})$. As previously discussed,[41] this definition is sufficiently general for typical CG mappings, e.g, center-of-mass mappings.

The appropriate potential for a CG model that quantitatively reproduces all structural properties of an atomistic model (at the resolution of the CG mapping) is a many-body potential of mean force (PMF), $U^0(\mathbf{R})$:

$$U^0(\mathbf{R}) = -k_B T \ln p_R(\mathbf{R}) + \text{const}, \tag{4.2}$$

where $p_R(\mathbf{R})$ is the probability for the atomistic model to sample a configuration that maps to the CG configuration $\mathbf{R}$. The forces, $\mathbf{F}_I^0(\mathbf{R})$, derived from the many-body PMF define the many-body mean force (MF) field and equal the mean force on site I averaged over all atomistic configurations that map to $\mathbf{R}$.

Because the PMF cannot be readily treated, we consider approximate potentials of a molecular mechanics form:

$$U(\mathbf{R}) = \sum_\zeta \sum_\lambda U_\zeta(\psi_\zeta(\{\mathbf{R}\}_\lambda)), \tag{4.3}$$

where ζ indicates a particular interaction (e.g., a dihedral angle interaction) and U_ζ is the corresponding potential (e.g., a dihedral angle potential) that is a function of a single scalar variable, ψ_ζ, (e.g., a dihedral angle) that may be expressed as a function of the Cartesian coordinates, $\{\mathbf{R}\}_\lambda$, for a set of sites, λ (e.g., the 4 successively bonded sites that form a dihedral angle).[60] The resulting force on site I may be expressed:

$$\mathbf{F}_I(\mathbf{R}) = \sum_\zeta \int dx\ \phi_\zeta(x) \boldsymbol{\mathscr{G}}_{I;\zeta}(\mathbf{R}; x), \tag{4.4}$$

where

$$\boldsymbol{\mathscr{G}}_{I;\zeta}(\mathbf{R}; x) = \sum_\lambda \mathbf{f}_{I;\zeta\lambda}(\mathbf{R}) \delta(\psi_{\zeta\lambda}(\mathbf{R}) - x), \tag{4.5}$$

$\phi_\zeta(x) = -dU_\zeta(x)/dx$ is a force function, and $\psi_{\zeta\lambda}(\mathbf{R}) = \psi_\zeta(\{\mathbf{R}\}_\lambda)$. In Equation (4.5), $\mathbf{f}_{I;\zeta\lambda}(\mathbf{R}) = \partial\psi_{\zeta\lambda}(\mathbf{R})/\,\partial\mathbf{R}_I$ determines the direction of the force on site I from a specific instance, λ, of an interaction of type ζ. We note that the present framework readily generalizes for more complex potentials.[48]

The MS-CG approach treats CG force fields as elements in an abstract vector space. Each element in this space defines a set of vector-valued functions that specify a force on each site as a function of the CG configuration. The CG force field defined by Equation (4.4) identifies a particular vector in this space that specifies the set of functions $\mathbf{F} = \{\mathbf{F}_1(\mathbf{R}), \ldots, \mathbf{F}_N(\mathbf{R})\}$. In this framework, the force field can be simply re-expressed:

$$\mathbf{F} = \sum_\zeta \int dx\ \phi_\zeta(x)\boldsymbol{\mathscr{G}}_\zeta(x). \tag{4.6}$$

Equation (4.6) defines a basis set expansion for CG force fields. The set $\{\boldsymbol{\mathscr{G}}_\zeta(x)\}$ forms a highly incomplete basis of force field vectors that are determined by the form of the approximate CG potential in Equation (4.3). The corresponding set of force functions, $\{\phi_\zeta(x)\}$, act as coefficients that identify a particular element in this vector space. An inner product, $\odot$, can be defined for any two elements, $\mathbf{F}^{(1)}$ and $\mathbf{F}^{(2)}$, in this vector space: $\mathbf{F}^{(1)} \odot \mathbf{F}^{(2)} = \frac{1}{3N}\left\langle \sum_I \mathbf{F}_I^{(1)}(\mathbf{M}(\mathbf{r})) \cdot \mathbf{F}_I^{(2)}(\mathbf{M}(\mathbf{r}))\right\rangle$, where the angular brackets denote a canonical average according to the atomistic probability distribution.

Although the many-body MF, $\mathbf{F}^0$, is an element in the vector space of CG force fields, in general, it will not be spanned by the basis set included in Equation (4.6). As illustrated in Figure 4.1, the MS-CG force field is defined by the geometric projection of the many-body MF, $\mathbf{F}^0$, onto the vector subspace of force fields spanned by this basis set. Consequently, the many-body MF and MS-CG force field have equal projections along each basis vector:

$$\boldsymbol{\mathscr{G}}_\zeta(x) \odot \mathbf{F}^0 = \boldsymbol{\mathscr{G}}_\zeta(x) \odot \mathbf{F}. \tag{4.7}$$

The inner product $b_\zeta(x) \equiv \boldsymbol{\mathscr{G}}_\zeta(x) \odot \mathbf{F}^0$ defines the projection of the many-body MF along the basis vector corresponding to the CG potential $U_\zeta(x)$. After expanding $\mathbf{F}$ according to Equation (4.6), we obtain the normal equations[41,62] for the MS-CG force functions:

$$b_\zeta(x) = \sum_{\zeta'} \int dx' G_{\zeta\zeta'}(x,x')\phi_{\zeta'}(x'). \tag{4.8}$$

In Equation (4.8), $G_{\zeta\zeta'}(x,x') \equiv \boldsymbol{\mathscr{G}}_\zeta(x) \odot \boldsymbol{\mathscr{G}}_{\zeta'}(x')$ defines a metric tensor that quantifies the "angle" formed between the different basis vectors. Figure 4.1 and Equation (4.8) demon-

strate the underlying simplicity of the MS-CG framework. While the projection $b_\zeta(x)$ quantifies an average force along the ψ_ζ degree of freedom, the metric tensor decomposes this average force into contributions from each term in the CG potential. Our previous work demonstrated that $b_\zeta(x)$ can be expressed in terms of simple structural correlation functions and that the MS-CG equations are equivalent to a generalized-Yvon-Born-Green (g-YBG) equation for complex molecular systems.[63,64] This relation provides the foundation for the g-YBG approach of determining MS-CG potentials directly from structural information.

The metric tensor can be calculated

$$\begin{aligned} G_{\zeta\zeta'}(x,x') &= \frac{1}{3N}\int d\mathbf{R}\, p_R(\mathbf{R}) \sum_{\lambda,\lambda'} \left(\sum_I \mathbf{f}_{I;\zeta\lambda}(\mathbf{R}) \cdot \mathbf{f}_{I;\zeta'\lambda'}(\mathbf{R}) \right) \\ &\quad \times \delta(\psi_{\zeta\lambda}(\mathbf{R}) - x)\delta(\psi_{\zeta'\lambda'}(\mathbf{R}) - x'). \end{aligned} \tag{4.9}$$

This average is performed over the CG configuration space, but with the weighting determined by the atomistic model according to p_R. As defined below Equation (4.5), $\mathbf{f}_{I;\zeta\lambda}(\mathbf{R})$ is the direction of the force exerted on site I by the λ instance of the ζ-type potential. The associated inner product, i.e., $\mathbf{f}_{I;\zeta\lambda}(\mathbf{R}) \cdot \mathbf{f}_{I;\zeta'\lambda'}(\mathbf{R}) = |\mathbf{f}_{I;\zeta\lambda}(\mathbf{R})||\mathbf{f}_{I;\zeta'\lambda'}(\mathbf{R})| \cos\varphi$, defines an angle φ formed by the forces exerted on the site I by the two particular interactions, $\zeta\lambda$ and $\zeta'\lambda'$. Consequently, $G_{\zeta\zeta'}$ significantly differs from a conventional 2-dimensional probability distribution and may assume positive or negative values. In particular, $G_{\zeta\zeta'}(x,x') \neq 0$ only if 1) there exist configurations, $\mathbf{R}$, for which $\psi_{\zeta\lambda}(\mathbf{R}) = x$ and $\psi_{\zeta'\lambda'}(\mathbf{R}) = x'$ for some λ and λ'; 2) the corresponding $\zeta\lambda$ and $\zeta'\lambda'$ interactions exert forces on a shared site, I; and 3) the corresponding average of $\sum_I \mathbf{f}_{I;\zeta\lambda}(\mathbf{R}) \cdot \mathbf{f}_{I;\zeta'\lambda'}(\mathbf{R})$ does not vanish. Therefore, $G_{\zeta\zeta'}$ may vanish even if the ζ and ζ' interactions are not statistically independent.

By separating out the $\lambda = \lambda'$ term in Equation (4.9), $G_{\zeta\zeta'}(x,x')$ can be decomposed into direct, $\bar{g}_\zeta(x)$, and indirect, environment-mediated, $\bar{G}_{\zeta\zeta'}(x,x')$, contributions: $G_{\zeta\zeta'}(x,x') = \bar{g}_\zeta(x)\delta_{\zeta\zeta'}\delta(x-x') + \bar{G}_{\zeta\zeta'}(x,x')$.[64] The projections of the many-body MF are then decomposed into direct and correlated contributions:

$$b_\zeta(x) = \bar{g}_\zeta(x)\phi_\zeta(x) + \sum_{\zeta'} \int dx' \bar{G}_{\zeta\zeta'}(x,x')\phi_{\zeta'}(x'). \tag{4.10}$$

In this decomposition,[64,188] $\bar{g}_\zeta(x)$ is a correlation function of a single variable that is closely related to, e.g., conventional rdfs, while $\bar{G}_{\zeta\zeta'}(x,x')$ describes the correlation between the ζ and ζ' interactions and only includes many-body correlations. In analogy to Equation (4.1), the first term in Equation (4.10) identifies the direct contribution of the $\phi_\zeta(x)$ interaction to

the projection, $b_\zeta(x)$. As illustrated in Figure 4.1, $\bar{G}_{\zeta\zeta'}(x,x')$ weights the contributions from the ζ' interaction to $b_\zeta(x)$. We emphasize that $\bar{G}_{\zeta\zeta'}$ differs from $G_{\zeta\zeta'}$ only in that the first sum in Equation (4.9) is restricted to $\lambda \neq \lambda'$. When $\zeta \neq \zeta'$ or $x \neq x'$, $G_{\zeta\zeta'}(x,x') = \bar{G}_{\zeta\zeta'}(x,x')$.

Importantly, this analysis applies for complex potentials with, e.g., angle and torsional potentials. However, in order to make this analysis more concrete and also because many of the following results consider pair nonbonded interactions, we present explicit results for a CG potential that includes only a single type of pair potential, $U^{(2)}$. In this case, $U(\mathbf{R}) = \sum_\lambda U^{(2)}(\psi_\lambda(\mathbf{R}))$, λ identifies a particular pair of sites $\{I,J\}$, $\psi_\lambda(\mathbf{R}) = |\mathbf{R}_I - \mathbf{R}_J|$ is the distance between the pair, and $\mathbf{f}_{I;\zeta\lambda}(\mathbf{R})$ is a unit vector along $\mathbf{R}_I - \mathbf{R}_J$. Then, $\bar{g}(r) = cr^2 g(r)$ in terms of the rdf, $g(r)$, and the constant $c = \frac{4\pi}{3}N/V$. Moreover, $b(r) = k_B T c r^2 dg(r)/dr = -cr^2 g(r) w'(r)$ in terms of the pair mean force, $-w'(r)$, which is defined by the pair potential of mean force,[135,166] $w(r) = -k_B T \ln g(r)$. The many-body contribution to the metric tensor then simplifies to a sum over triples:

$$\bar{G}(x,x') = \frac{1}{3N}\left\langle \sum_\lambda^{\text{triples}} \cos\varphi_\lambda\, \delta(r_{\lambda J} - x)\delta(r_{\lambda K} - x') \right\rangle, \tag{4.11}$$

where λ identifies the "central" site of a triple, $\{\lambda, J, K\}$, $r_{\lambda J}$ and $r_{\lambda K}$ are the distances from the central site to J and K, respectively, and φ_λ is the angle formed by the triple. Because the number of particles J that are a distance x away from a central particle λ scales as x^2, in this simple case, $\bar{G}(x,x')$ should scale with $(xx')^2$. The following calculations explicitly employ this scaling when considering many-body correlations that correspond to nonbonded pair interactions.

More generally, if we consider a particular nonbonded pair potential, ζ, within a more complex CG potential and apply the preceding analysis, Equation (4.10) may be re-expressed:

$$-w'_\zeta(r) = \phi_\zeta(r) + \sum_{\zeta'} \int dx' \frac{1}{c_\zeta g_\zeta(r) r^2} \bar{G}_{\zeta\zeta'}(r,x')\phi_{\zeta'}(x'), \tag{4.12}$$

where c_ζ is a constant, $g_\zeta(r)$ is the atomistic rdf, and $w_\zeta(r)$ is the atomistic pair potential of mean force for the ζ nonbonded pair interaction. In complete analogy with Equation (4.1), this statement of the g-YBG equation decomposes the pair mean force into a direct force between the pair and correlated contributions from each term in the CG potential. The force functions, $\phi_\zeta(x)$, that satisfy this decomposition of the atomistic pair mean force, $-w'_\zeta(r)$, then determine the MS-CG force field. This relation further clarifies the central role of the metric tensor in the MS-CG method. In particular, if $\bar{G}_{\zeta\zeta'} = 0$, then the MS-CG method

reduces to direct Boltzmann inversion.

4.3 Methods

The present section briefly outlines essential details of the following calculations. The Supporting Information section provides additional details.

4.3.1 Simulation Details

All molecular dynamics (MD) simulations were performed using the Gromacs 4.5.3 simulation suite[7] according to standard procedures.[168–172] MD simulations were performed for three classes of systems: 1) atomistic models of heptane and of water, 2) one-, two-, and three-site MS-CG models of heptane, and 3) model fluids of mono-, di-, and tri-atomic molecules with similar geometries to the one-, two-, and three-site MS-CG heptane models. The present section briefly outlines the key details of these simulations, which are described in much greater detail in the Supporting Information section.

The atomistic heptane simulations considered 267 molecules and modeled all interactions with the OPLS-AA force field.[18] The heptane model was first equilibrated in the NPT ensemble at 1 bar and 298K to determine an equilibrium volume of $V = (4.08 \text{ nm})^3$, which agrees within 3% of the experimentally measured density.[204] The system was then simulated in the NVT ensemble at this volume in order to determine canonical correlation functions for the following analysis. The Supporting Information section also reports details and results for a 20 ns simulation of 216 SPC/E[205] molecules in the NVT ensemble at 298 K and $V = (1.86 \text{ nm})^3$.

The following subsection describes calculations of the MS-CG force field for each CG heptane representation. An initial configuration for each CG model was obtained by mapping a configuration from all-atom simulations. This configuration was energy minimized and simulated for 24 ns in the NVT ensemble at 298 K, with the first 4 ns serving for equilibration.

In order to identify generic features of $\bar{G}_{\zeta\zeta'}$, we simulated model fluids of mono-, di-, and tri-atomic molecules. In each case, all nonbonded interactions were modeled with identical Lennard-Jones pair potentials that generated a stable liquid phase. Each system was simulated in the NVT ensemble at 298K after equilbration in the constant NPT ensemble at 298K and 1 bar pressure to determine the equilibrium volume.

4.3.2 Mapping

The atomically detailed simulations of heptane were mapped to one-, two-, and three-site CG representations, as shown in Figure 4.2. The one-site mapping defined a single (CC) site at the molecular center of mass. The two-site mapping defined two equivalent (CS) sites at the centers of mass of the two terminal $CH_2CH_2CH_3$ groups. The three-site mapping defined two equivalent (CT, blue) sites at the centers of mass for the two terminal CH_2CH_3 groups and a single (CM, orange) site at the center of mass for the middle $CH_2CH_2CH_2$ group. Molecular graphics in Figure 4.2 and elsewhere were rendered with VMD.[174]

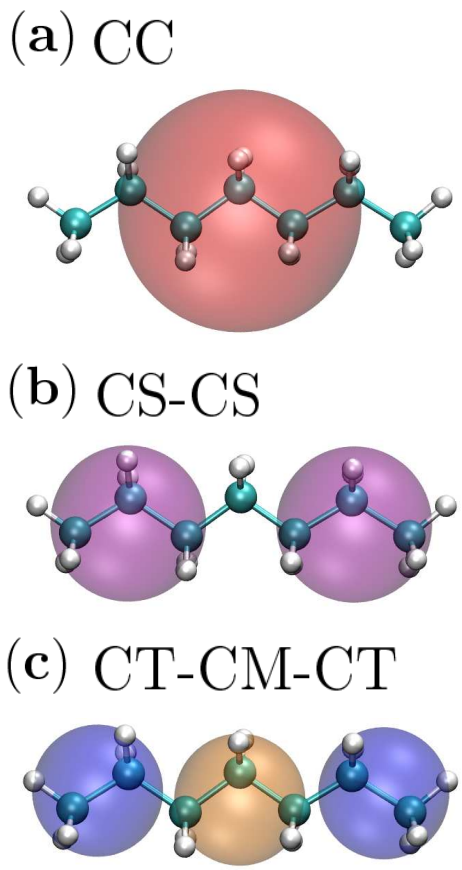

Figure 4.2. CG representations of heptane. Figure 4.2a: The one-site model defines one CC site (red) for each molecule. Figure 4.2b: The two-site model defines two CS sites (purple) for the two terminal $CH_3CH_2CH_2$ groups. Figure 4.2c: The three-site model defines two CT sites (blue) for the terminal CH_3CH_2 groups and a CM site (orange) for the central $CH_2CH_2CH_2$ group. The coordinates of each site are defined by the center of mass for the associated atomic group.

4.3.3 Force Field Calculations

For each CG representation of heptane, the MS-CG potential was calculated from a discrete version of the normal equations presented in Equation (4.8). As appropriate, the intramolecular potentials included bond and angle terms. All intermolecular interactions were modeled with central pair potentials. Each potential function was represented by a discrete set of basis functions of a single variable.[60] The CG force field is then expanded in a corresponding discrete basis set: $\mathbf{F} = \sum_D \phi_D \mathcal{G}_D$, where D identifies a particular basis function used to represent a particular potential function, $U_\zeta(x)$, $\mathcal{G}_D$ is the resulting force field basis vector, and ϕ_D is the coefficient of that vector, which also determines the corresponding force function, $\phi_\zeta(x)$. These basis vectors determine discrete representations of $G_{\zeta\zeta'}(x, x')$ and $b_\zeta(x)$, i.e., $G_{DD'}$ and b_D. $G_{DD'}$ was calculated from configurations sampled by all-atom MD simulations. b_D was calculated from atomistic forces, according to the MS-CG approach,[60,62,72] and from structural correlation functions, according to the g-YBG approach.[63,64,134,188] Both methods quantitatively agreed, as expected. These correlation functions then determine a discrete set of normal equations for the MS-CG force field parameters:

$$\sum_{D'} G_{DD'}\phi_{D'} = b_D. \tag{4.13}$$

These equations are explicitly derived in the Supporting Information and have been previously discussed.[60,62,63,72,188]

In most calculations, the potentials were represented on a grid, i.e., with piecewise constant basis functions.[62] The bond, angle, and nonbonded pair potentials were represented with grid spacings of 0.0005 nm, 0.5 deg, and 0.005 nm, respectively. In this basis set, $G_{DD'}$ and b_D correspond to representing $G_{\zeta\zeta'}(x, x')$ and $b_\zeta(x)$, respectively, on a set of discrete grid points. In order to investigate the sensitivity of simulated structural properties to the form of the CG potential, the CG force field was also calculated using Lennard-Jones-type functions to represent each nonbonded pair potential, i.e., $U_\zeta^{n-m}(x) = \phi_{\zeta 1}/x^n + \phi_{\zeta 2}/x^m$, where $\phi_{\zeta 1}$ and $\phi_{\zeta 2}$ may be either positive or negative. We report results for $n - m = 12 - 6$ and $8 - 4$, but similar results were obtained for $n - m = 11 - 6, 10 - 6, 10 - 4, 9 - 6$, and $8 - 6$.

The normal equations (Equation (4.13)) were solved via LU decomposition after applying right preconditioning.[175] The calculated force functions were smoothed with a running average over three consecutive grid points and then integrated to determine corresponding potentials. These potentials were interpolated and employed to simulate each CG model. Pair, three-body, and many-body correlation functions were calculated from these trajecto-

ries and compared to those calculated from the mapped, all-atom trajectories. The Supporting Information demonstrates that the MS-CG models quantitatively reproduced the structure of the atomistic OPLS heptane model.

4.3.4 Molecular Interpretation of $\bar{G}_{\zeta\zeta'}$

As defined above Equation (4.10), the MS-CG metric tensor $G_{\zeta\zeta'}(x, x')$ can be decomposed into a correlation function of a single variable, $\bar{g}_\zeta(x)$, and a many-body correlation function, $\bar{G}_{\zeta\zeta'}(x, x')$, that describes the coupling between the ζ and ζ' interactions. As noted above, $\bar{G}_{\zeta\zeta'} = G_{\zeta\zeta'}$ except for the case that $\zeta = \zeta'$ and $x = x'$. In order to characterize the relevant many-body correlations, we calculated $\bar{G}_{\zeta\zeta'}$ for each CG mapping shown in Figure 4.2, while using configurations sampled from atomistic simulations.

Equation (4.9) demonstrates that $\bar{G}_{\zeta\zeta'}$ differs from conventional many-body correlation functions in two ways: 1) $\bar{G}_{\zeta\zeta'}$ can adopt either positive or negative values based upon the direction of forces from the ζ and ζ' interactions; and 2) $\bar{G}_{\zeta\zeta'}$ only reflects correlations between interactions that exert forces on a shared particle. In order to better understand $\bar{G}_{\zeta\zeta'}$ and, in particular, to understand this directional effect, we compared $\bar{G}_{\zeta\zeta'}(x, x')$ to a more conventional many-body correlation function:

$$P_{\zeta\zeta'}(x, x') = \frac{1}{3N}\int d\mathbf{R}\, p_R(\mathbf{R}) \sum_{\lambda\neq\lambda'}{}^{*}\, \delta(\psi_{\zeta\lambda}(\mathbf{R}) - x)\delta(\psi_{\zeta'\lambda'}(\mathbf{R}) - x'), \tag{4.14}$$

where λ and λ' identify particular instances of the ζ and ζ' interactions, respectively, and the star indicates that the sum is restricted to the set of distinct interactions, $\zeta\lambda$ and $\zeta'\lambda'$ that exert forces on a shared particle. With this definition, both $\bar{G}_{\zeta\zeta'}$ and $P_{\zeta\zeta'}$ include contributions from the same set of interactions. However, $P_{\zeta\zeta'}$ weights each contribution with equal (positive) weight, while $\bar{G}_{\zeta\zeta'}$ weights each contribution based upon the geometry of the interaction.

As discussed above, if ζ corresponds to a nonbonded pair interaction, then both $\bar{G}_{\zeta\zeta'}(x, x')$ and $P_{\zeta\zeta'}(x, x')$ scale according to x^2. In analogy to the definition of the rdf[135,166] and in order to focus on the local many-body correlations that influence pair mean forces and CG force fields, the following section presents $\bar{G}_{\zeta\zeta'}(x, x')$ and $P_{\zeta\zeta'}(x, x')$ after rescaling both according to $(r_\zeta(x)r_{\zeta'}(x'))^{-2}$ where $r_\zeta(x) = x$ when ζ corresponds to a nonbonded interaction and 1 otherwise. These correlation functions are presented as intensity plots using gnuplot 4.4.0.[206]

4.3.5 Eigenvalue Analysis of $G_{\zeta\zeta'}$

In order to investigate the role of many-body correlations in determining the CG force field, we performed eigenvalue/eigenvector analysis of $G_{\zeta\zeta'}$. Upon determining the eigenvalues, $\{\lambda_i\}$, and associated eigenvectors, $\{\mathbf{v}_i\}$, of $G_{DD'}$, Equation (4.13) may be expressed

$$\mathbf{b} = \sum_i \lambda_i \phi_{\lambda_i} \mathbf{v}_i, \tag{4.15}$$

where $\mathbf{b}$ is a vector with elements b_D and ϕ_{λ_i} is the component along the eigenvector $\mathbf{v}_i$ of an analogous vector of force field coefficients, $\boldsymbol{\phi}$. These eigenvectors identify correlated forces that contribute to the projections of the CG force field.

The following section considers eigenvalues and eigenvectors calculated after rescaling $G_{\zeta\zeta'}(x,x')$ by $1/r_\zeta^2(x)r_{\zeta'}^2(x')$. We employ this scaling for several reasons: 1) This rescaling corresponds to a simple redefinition of the basis vectors and, equivalently, to natural left and right preconditioning of $G_{\zeta\zeta'}(x,x')$ to solve Equation (4.13); 2) This rescaling corresponds to the representation employed in visualizing the structural correlation function $\bar{G}_{\zeta\zeta'}(x,x')$; and 3) If rescaling is not applied, then the calculated eigenvalue spectrum is dominated by long-ranged nonbonded interactions, simply due to the x^2 scaling of nonbonded pairs, as discussed above.

We employed LAPACK (Linear Algebra PACKage)[207] to calculate the eigenvalues and eigenvectors of $G_{\zeta\zeta'}(x,x')/\, r_\zeta^2(x)r_{\zeta'}^2(x')$ for the three-site heptane representation. Given this set of calculated eigenvalue/eigenvector pairs, we defined a participation fraction, $n_{i;\zeta}$, to quantify the significance of the ζ interaction to eigenvector i. For each eigenvector i, we calculated $n_{i;\zeta}$ by adding the absolute values of the eigenvector elements for the basis vectors, $\{\boldsymbol{\mathscr{G}}_\zeta(x)\}$, associated with the potential $U_\zeta(x)$. We then normalized this quantity for each eigenvector i by the sum of the absolute values for all the elements in the eigenvector.

We calculated an analogous quantity to detemine the contributions to each eigenvector from short-, medium-, and long-ranged nonbonded interactions. In this calculation, interactions were considered short-ranged if they corresponded to distances smaller than the first minima in the corresponding calculated CG potential. This distance is approximately 0.5 nm for each interaction. Interactions were considered medium-ranged if they corresponded to distances between the first potential minimum and the first local maximum in the force function at approximately 0.8 nm. Interactions corresponding to greater distances were considered long-ranged.

4.4 Results

Structure-motivated CG approaches must address the effects of many-body correlations when determining potentials for reproducing atomistic structure. The MS-CG metric tensor, $G_{\zeta\zeta'}(x, x')$, provides a transparent, geometric framework for both characterizing these correlations and also for quantifying their impact upon the CG potential. The following calculations provide a detailed analysis of $G_{\zeta\zeta'}$ in the context of determining MS-CG potentials for the one-, two-, and three-site models of liquid heptane that are shown in Figure 4.2. Figures 4.3 and 4.4 precisely relate robust aspects of $G_{\zeta\zeta'}$ to specific structural features of the atomistic model and to the CG representation, respectively. Figure 4.5 demonstrates the role of $G_{\zeta\zeta'}$ in decomposing atomistic pair mean forces into contributions from the various terms in the approximate CG force field. Finally, Figures 4.6 and 4.7 employ eigenvalue/eigenvector analysis to investigate the relationship between the metric tensor, the calculated MS-CG force field, and the equilibrium structure of the resulting CG model. Although it is not the focus of the present work, the Supporting Information Section demonstrates that the one-, two-, and three-site MS-CG heptane models each accurately reproduce the structure of the atomistic model.

4.4.1 Molecular Interpretation of $\bar{G}_{\zeta\zeta'}$

The function $\bar{G}_{\zeta\zeta'}$ quantifies the contributions of many-body correlations to the MS-CG metric tensor. In order to relate $\bar{G}_{\zeta\zeta'}$ to specific structural features of molecular liquids, Figure 4.3 compares $\bar{G}_{\zeta\zeta'}$ with the more conventional many-body correlation function $P_{\zeta\zeta'}$, which is defined in Equation (4.14). Both correlation functions were computed using configurations that were sampled from atomistic simulations of the OPLS heptane model and then mapped to the three-site heptane representation shown in Figure 4.2c. As discussed above, in the cases that ζ (or ζ') correspond to potentials of a distance, r_ζ, the correlation functions $P_{\zeta\zeta'}$ and $\bar{G}_{\zeta\zeta'}$ have been rescaled by r_ζ^2 in order to highlight the relevant local structural features.

Figure 4.3a considers the case that ζ and ζ' both correspond to the CT-CT nonbonded pair potential. Figure 4.3a1 presents an intensity map of $P_{\zeta\zeta'}(r, r')$, which quantifies the probability of observing a set, $\{\lambda, J, K\}$, of three CT sites in which sites J and K (shadowed in Figure 4.3a3) are distances r and r', respectively, from the central λ site (blue in Figure 4.3a3). In this intensity plot (and in Figure 4.3b1), blue, red, and white identify regions of zero, high, and intermediate probability, respectively. Figure 4.3a1 demonstrates regions of high and low probability in $P_{\zeta\zeta'}$ that correspond to solvation shells in the CT-CT rdf.

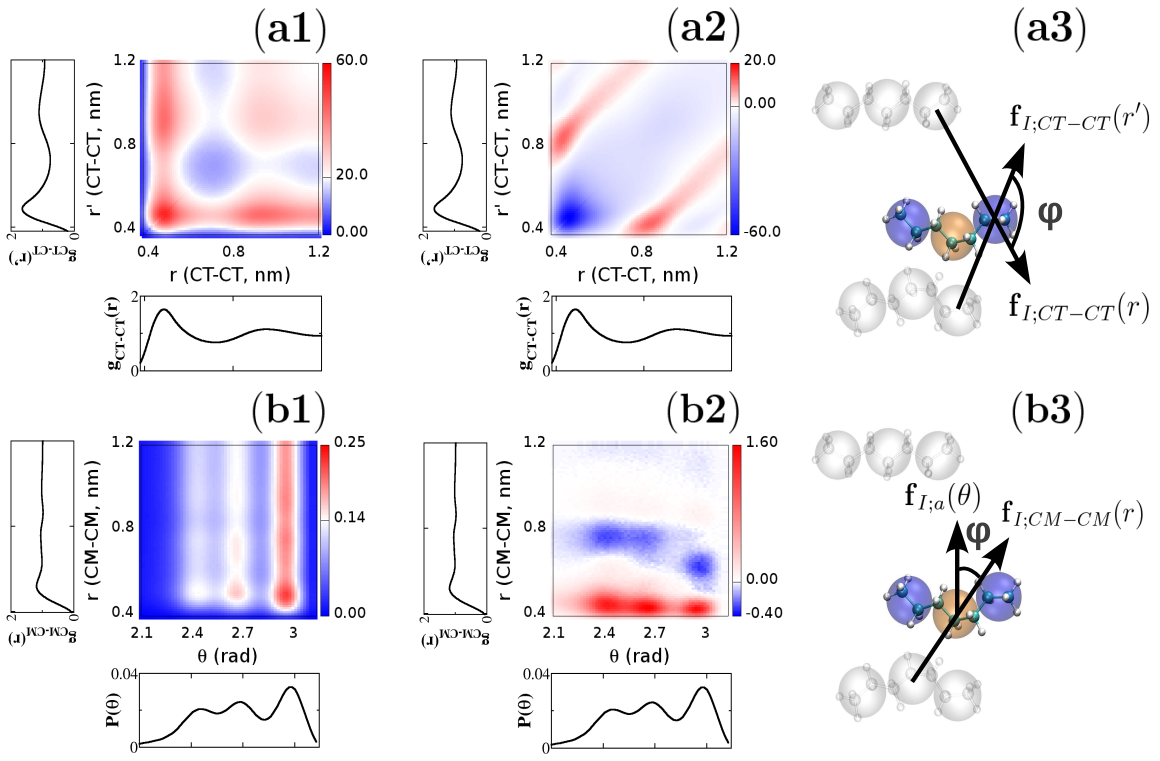

Figure 4.3. Many-body correlations in the OPLS-AA heptane model. Row a describes correlations between CT-CT and CT-CT pair nonbonded interactions. Row b describes correlations between the CM-CM pair nonbonded interaction and the intramolecular angle interaction. Columns 1 and 2 quantify these correlations with intensity plots of $P_{\zeta\zeta'}$ and $\bar{G}_{\zeta\zeta'}$, respectively. Column 3 illustrates the angle, φ, formed by the force vectors, $\mathbf{f}_{I;\zeta\lambda}$, that contribute to $\bar{G}_{\zeta\zeta'}$ in each case. For clarity of presentation, $P_{\zeta\zeta'}$ and $\bar{G}_{\zeta\zeta'}$ are rescaled by factors of 10^2 and 10^3, respectively, in Figures 4.3 and 4.4.

Figure 4.3a2 presents the corresponding block of $\bar{G}_{\zeta\zeta'}(r,r')$. In Figures 4.3 and 4.4, negative, positive, and zero values of $\bar{G}_{\zeta\zeta'}(r,r')$ are indicated by blue, red, and white regions, respectively. By definition, both $P_{\zeta\zeta'}(r,r')$ and $\bar{G}_{\zeta\zeta'}(r,r')$ include contributions from the same sets of triples. However, while each triple contributes equal (positive) weight to $P_{\zeta\zeta'}(r,r')$, the triple, $\{\lambda,J,K\}$ contributes $\cos\varphi$ to $\bar{G}_{\zeta\zeta'}(r,r')$, with φ being the angle between the nonbonded CT-CT force vectors, as illustrated in Figure 4.3a3. In particular, Equation (4.11) demonstrates that $\bar{G}_{\zeta\zeta'}(r,r')$ vanishes, if for given r and r', all angles, φ, are equally sampled. The nonvanishing regions in Figure 4.3a2 indicate preferred angles, φ, for particular distances, r and r'.

The most striking feature of Figure 4.3a2 is a large negative peak located at $r \approx r' \approx$ 0.5 nm. This feature corresponds to triples for which CT sites J and K are both in the

first solvation shell of the central site, λ. Supporting Figure S10 demonstrates that the excluded volume of sites J and K prevents their overlap, excluding configurations for which $\cos\varphi \geq 0.75$, while also promoting configurations for which $\cos\varphi = -1$. The resulting slight preference for the sites to pack on opposite sides of λ and for the force vectors to form an obtuse angle (i.e., $\cos\varphi$ ¡ 0) generates the observed negative peak. This steric effect generates a negative band for the diagonal $r \approx r'$ region, but becomes decreasingly significant with increasing r.

Figure 4.3a2 also demonstrates positive bands at $r' \approx r \pm \sigma$, where σ is the site diameter, which is estimated by the first peak in the corresponding rdf. These positive bands also reflect the solvation shell structure. Supporting Figure S10 demonstrates that, for configurations with $r' \approx r \pm \sigma$, the sites J and K are slightly more likely to be on the same side of the central site, λ. Consequently, the angle φ is slightly more likely to be acute and the resulting average will be positive. This effect is most pronounced when r' is slightly less than $r + \sigma$, since the triplet configuration is more likely to be slightly staggered than perfectly colinear.

Columns 1 and 2 of Figure 4.3b present a similar comparison of $P_{\zeta\zeta'}$ and $\bar{G}_{\zeta\zeta'}$ for the case that ψ_ζ corresponds to the intramolecular CT-CM-CT angle, θ, and $\psi_{\zeta'}$ corresponds to the CM-CM pair distance, r. Figure 4.3b1 demonstrates three distinct bands in $P_{\zeta\zeta'}(\theta, r)$ that correspond to heptane rotamer states, with the largest band corresponding to an all-trans configuration. Within these bands, regions of high and low intensities correspond to the maxima and minima in the CM-CM rdf.

Figure 4.3b2 presents the corresponding block of $\bar{G}_{\zeta\zeta'}(\theta, r)$. Figure 4.3b3 demonstrates the direction of the force vectors from ζ and ζ' interactions that act on a shared CM site. In the configurations mapped from the atomistic simulation, the force vector from the angle potential approximately bisects the intramolecular angle, θ. As in Figure 4.3a, the nonbonded CM-CM pair force is directed along the vector connecting the two sites.

Although the three bands observed in $P_{\zeta\zeta'}(\theta, r)$ remain intact, the peaks in these bands now demonstrate alternating signs in $\bar{G}_{\zeta\zeta'}(\theta, r)$. Supporting Figure S12 demonstrates that, as CM sites approach one another, the flanking CT sites must project away. As illustrated in Figure 4.3b3, the organization of CT sites away from the approaching CM site results in an acute angle (i.e., $\cos\varphi > 0$) between the associated force vectors and generates a positive peak in $\bar{G}_{\zeta\zeta'}$. This steric effect is most pronounced for molecules with a relatively small intramolecular angle, $\theta \approx 2.4$ rad. Consequently, in comparison to $P_{\zeta\zeta'}(\theta, r)$, the band of $\bar{G}_{\zeta\zeta'}(\theta, r)$ that corresponds to small intramolecular angles, $\theta \approx 2.4$ rad, grows in magnitude relative to the band for $\theta \approx 3.0$ rad. The alternating signs of the peaks in each band result from the packing of molecules into solvation shells on opposite sides of a central molecule,

as described in the context of Figure 4.3a2. Supporting Figure S13 suggest that the shift in this negative peak for $\theta \approx 3.0$ rad results from subtle packing effects.

In summary, Figure 4.3 demonstrates that $\bar{G}_{\zeta\zeta'}$ reflects relatively subtle aspects of intermolecular packing and its coupling to intramolecular configuration. However, these features result from generic properties of soft condensed-phase systems, e.g., steric interactions, solvation shell packing, and molecular geometry. In order to investigate the generality of these features and also the sensitivity of the metric tensor to the CG representation, we calculated $\bar{G}_{\zeta\zeta'}$ for model mono-, di-, and tri-atomic fluids and compared these results with calculations for the OPLS heptane simulations, while using one-, two-, and three-site mappings (Figure 4.2).

Figure 4.4 presents these calculations. In the first row, columns a, b, and c present results obtained by mapping atomistic heptane simulations to one-, two-, and three-site representations, respectively. In the second row, columns a, b, and c present calculations for simulations of model mono-, di-, and tri-atomic fluids, respectively. The third row illustrates the relevant interactions for each calculation. In each column, ζ and ζ' identify nonbonded pair potentials for corresponding site pairs in the two models. The Methods and Supporting Information Sections provide details of these calculations.

Column a of Figure 4.4 compares calculations of $\bar{G}_{\zeta\zeta'}$ for a one-site (CC) representation of atomistic heptane simulations (Figure 4.4a1) and for a model monatomic fluid (Figure 4.4a2). The size and asymmetry of heptane result in much broader and more diffuse bands than are observed for the spherically symmetric monatomic fluid. Nevertheless, both Figures 4.4a1 and 4.4a2 demonstrate the same generic features observed in Figure 4.3a2.

Column b of Figure 4.4 compares calculations of $\bar{G}_{\zeta\zeta'}$ for a two-site (CS-CS) representation of atomistic heptane simulations (Figure 4.4b1) and for a model diatomic fluid (Figure 4.4b2). Figures 4.4a1 and 4.4b1 demonstrate that $\bar{G}_{\zeta\zeta'}$ has similar structure for the one- and two-site representations of heptane. However, Figure 4.4b2 reveals a new positive (red) feature along the diagonal of $\bar{G}_{\zeta\zeta'}$ for the model diatomic fluid. As indicated in Figure 4.4b3, if the intramolecular bond length is sufficiently short, then the presence of a single site at a distance r away significantly increases the probability of a second site at a similar distance. The force vectors from these bonded sites form an acute angle (i.e., $\cos\varphi > 0$) and result in this positive feature for $r \approx r'$. The width of this feature corresponds to the intramolecular bond length and its magnitude decays with increasing r. Supporting Figure S16 demonstrates that this positive band vanishes if the bond length of the model diatomic becomes sufficiently large relative to the corresponding site diameters. This is indeed the case for the two-site heptane mapping, in which the bond length and the site diameter are both

approximately 0.5 nm

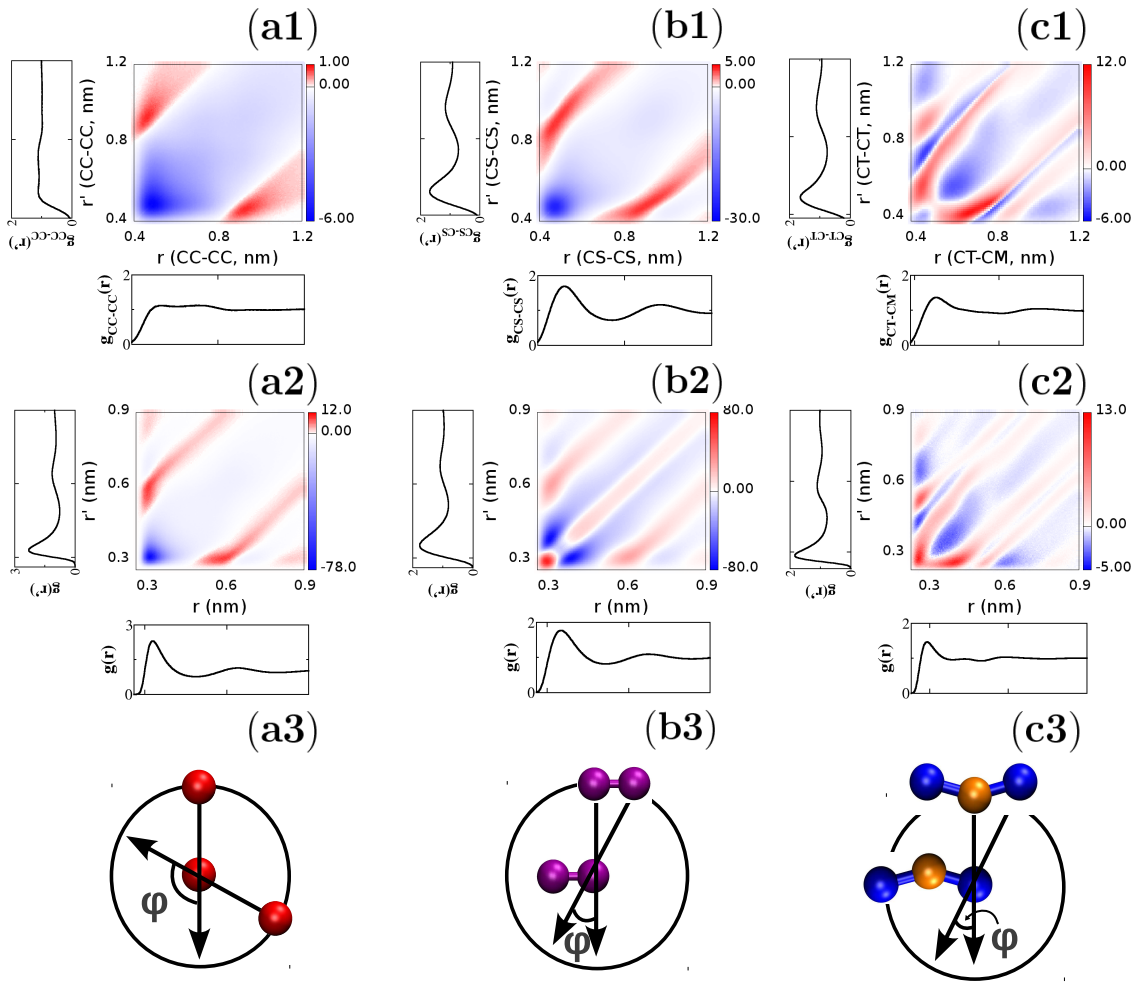

Figure 4.4. Intensity plots of $\bar{G}_{\zeta\zeta'}$ for nonbonded interactions in heptane and associated model fluids. Columns 1, 2, and 3 correspond to one-, two-, and three-site models. Rows 1 and 2 present correlation functions calculated using configurations sampled for the OPLS heptane model and for corresponding model fluids, respectively. The interactions in row 1 are identified by site types labeled in Figure 4.2. Finally, row 3 illustrates the angle, φ, that contributes to $\bar{G}_{\zeta\zeta'}$.

Finally, column c of Figure 4.4 compares calculations of $\bar{G}_{\zeta\zeta'}$ for a three-site (CT-CM-CT) representation of atomistic heptane simulations (Figure 4.4c1) and for a model triatomic fluid (Figure 4.4c2). In this case, ζ and ζ' correspond to the CT-CM and CT-CT pair nonbonded interactions. The model triatomic corresponding to Figure 4.4c2 was designed so that the ratio of the bond length, d, to the site diameter, σ, is very similar for the two models. As a result, the features of $\bar{G}_{\zeta\zeta'}(r, r')$ in Figures 4.4c1 and 4.4c2 are remarkably similar. In both cases, the bond length is sufficiently large that the positive diagonal feature observed

in Figure 4.4b2 has disappeared. Instead, the region corresponding to $r' \approx r$ reflects simple excluded volume effects similar to Figure 4.3a2. In addition, both Figures 4.4c1 and 4.4c2 demonstrate a positive band at $r' \approx r \pm \sigma$ corresponding to the second solvation shell, as in Figure 4.3a2. However, both Figures 4.4c1 and 4.4c2 also demonstrate a new positive feature at $r' \approx r \pm d$ (with $d < \sigma$) corresponding to the forces from the two CT sites that flank each central CM site and that exert forces in the same direction, as suggested by Figure 4.4c3.

4.4.2 Decomposition of Pair Mean Forces

Figure 4.1 and Equation (4.8) demonstrate that $G_{\zeta\zeta'}$ decomposes projections, $b_\zeta(x)$, of the many body MF into contributions from each force field basis vector. In the case that ζ corresponds to a bonded or nonbonded pair potential, this projection is closely related to the atomistic pair mean force, $-w'_\zeta(r)$, according to Equation (4.12). The decomposition of this atomistic mean force then defines the MS-CG force functions, $\{\phi_\zeta(r)\}$.

Columns 1, 2, and 3 of Figure 4.5 demonstrate this decomposition for the CT-CT, CM-CM, and CT-CM pair nonbonded interactions, respectively. In each panel, the solid black curve presents the pair mean force, while the solid red curve presents the contribution from the corresponding direct force. In each case, the marked difference between the mean and direct forces quantifies the significance of many-body correlations for determining additional contributions to the mean force. These differences are particularly significant at short distances near the first peak of each rdf. At these distances, the mean forces are much more attractive than the corresponding direct forces. (Note that the corresponding atomistic rdf attains local maxima or minima when the corresponding pair mean force vanishes and, in each case, achieves its first maxima near 0.5 nm.)

The dashed green and dashed blue curves in Figure 4.5a present the contributions to the nonbonded pair mean forces from intramolecular angle and bond forces, respectively. These bonded interactions contribute significantly, especially at short distances, and demonstrate the importance of the many-body correlations in Figure 4.3b for coupling intra- and intermolecular interactions.

Figure 4.5b quantifies the contributions to each nonbonded pair mean force (solid black) from the corresponding direct force (solid red) and from correlated CT-CT (dashed green), CM-CM (dashed cyan), and CT-CM (dashed magenta) nonbonded forces. In particular, Figure 4.5b1 demonstrates that, when a pair of CT sites is separated by 0.8 nm, the direct force between the pair is attractive, but the average force on each site is actually repulsive. This repulsive mean force results largely from correlated forces due to other CT sites in the

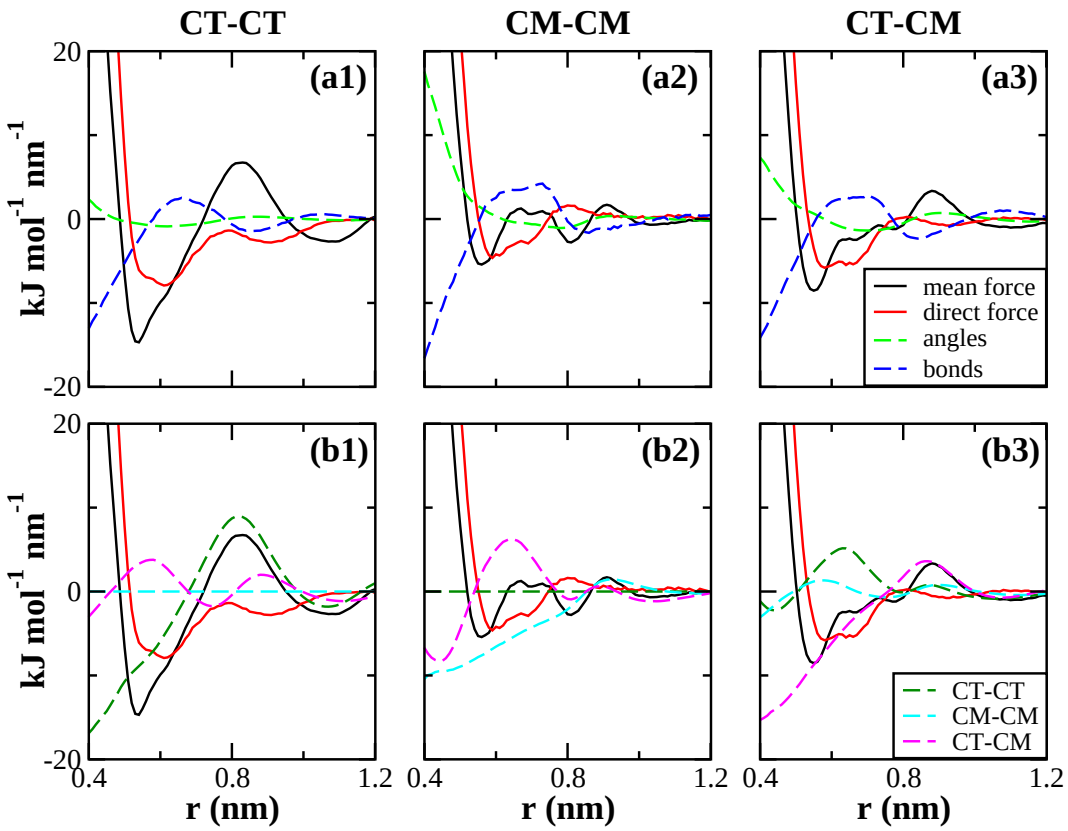

Figure 4.5. Decomposition of mean forces for the CT-CT (left), CM-CM (center), and CT-CM (right) nonbonded pair interactions. The solid black and red curves represent the calculated mean, $-w'_\zeta(r)$, and direct, $\phi_\zeta(r)$, forces, respectively. The dashed light green and blue curves present correlated contributions to the mean force from intramolecular angle and bond interactions, respectively. The dashed dark green, cyan, and magenta curves present contributions to the mean force from correlated CT-CT, CM-CM, and CT-CM pair nonbonded interactions, respectively.

first solvation shell. (See Supporting Figure S18.) Figure 4.5b1 demonstrates that correlated forces from these CT sites also significantly reduce the CT-CT mean force at short distances. Similar short-ranged effects are observed for each pair interaction.

Figure 4.5b demonstrates that correlated CT-CM and CT-CT interactions significantly impact both CT-CT and the CM-CM mean forces (Figures 4.5b2 and b3, respectively), suggesting that these interactions are particularly important for reproducing the pair structure of the system. In contrast, the CM-CM interaction contributes relatively little to the CT-CM mean force. Finally, although the CM-CM and CT-CT pair interactions are statistically correlated (Supporting Figure S19), the corresponding block of the metric tensor vanishes because these interactions never exert forces on the same particle. Consequently, the CM-CM and CT-CT interactions do not contribute to the CT-CT (Figure 4.5a2) and CM-CM (Figure 4.5b2) mean forces, respectively.

4.4.3 Eigenvalue Analysis of $G_{\zeta\zeta'}$

The MS-CG metric tensor, $G_{\zeta\zeta'}(x,x')$, defines a linear transformation from CG force functions, $\phi_\zeta(x)$, to projections, $b_\zeta(x)$, of the CG force field onto basis vectors, $\mathscr{G}_\zeta(x)$. Figure 4.6 presents eigenvalue/eigenvector analysis of the metric tensor to identify and quantify the correlated interactions that contribute to matching projections of the many-body MF. Each eigenvector identifies a set of correlated interactions that contribute to b_ζ, while the corresponding eigenvalue weights this contribution. For reasons discussed in the methods section, we calculated the 977 eigenvalue/eigenvector pairs for a discrete matrix representation of $G_{\zeta\zeta'}(x,x')/\, r_\zeta^2(x) r_{\zeta'}^2(x')$, rather than for $G_{\zeta\zeta'}(x,x')$. Figure 4.6c2 presents the resulting eigenvalue spectrum, which is bounded above by one (as is the eigenvalue spectrum for $G_{\zeta\zeta'}$) and demonstrates two inflection points near $i \approx 350$ and $i \approx 800$. Consequently, eigenvalues (and the associated eigenvectors) will be characterized as small, medium, or large based upon whether $350 > i$, $800 > i > 350$, or $i > 800$, respectively.

As described in the Methods section, we calculated a participation fraction, $n_{i;\zeta}$, to quantify the contribution from each interaction ζ to eigenvector i. The participation fraction, $n_{i;\zeta}$ vanishes if eigenvector i does not include any contribution from the ζ interaction and is normalized so that $n_{i;\zeta} = 1$ if eigenvector i only includes contributions from U_ζ. Intermediate values of $n_{i;\zeta}$ identify eigenvectors that reflect correlated contributions to b_ζ.

Figures 4.6a1, a2, and a3 present $n_{i;\zeta}$ as a function of eigenvector index i for ζ corresponding to the CT-CT, CM-CM, and CT-CM nonbonded interactions, respectively. Figures 4.6b1, b2, and b3 present an analogous participation fraction that quantifies the contribution to eigenvector i from short- ($r \lesssim 0.5$nm), medium- (0.5 nm $\lesssim r \lesssim$ 0.8 nm), and long-ranged ($r \gtrsim 0.8$ nm) interactions, respectively. Figures 4.6c1 and c2 present this participation fraction, $n_{i;\zeta}$, for ζ corresponding to intramolecular bond and angle interactions, respectively.

Figure 4.6a demonstrates that the largest eigenvectors , corresponding to the most significant contributions to b_ζ, reflect CT-CT and CT-CM nonbonded interactions. Medium eigenvectors reflect CM-CM nonbonded interactions that are coupled with CT-CT and CT-CM interactions, as well as with intramolecular angle interactions. The smallest eigenvectors correspond to contributions from intramolecular bond and angle interactions. Figure 4.6b demonstrates that the very largest eigenvectors of the normalized matrix reflect short-ranged interactions, while the next largest eigenvectors reflect correlations between short- and medium-ranged interactions. Eigenvectors corresponding to $350 < i < 800$ primarily reflect coupling between medium- and long-ranged interactions. While Figure 4.6b1 suggests that very small eigenvectors also reflect short-ranged interactions, these small eigen-

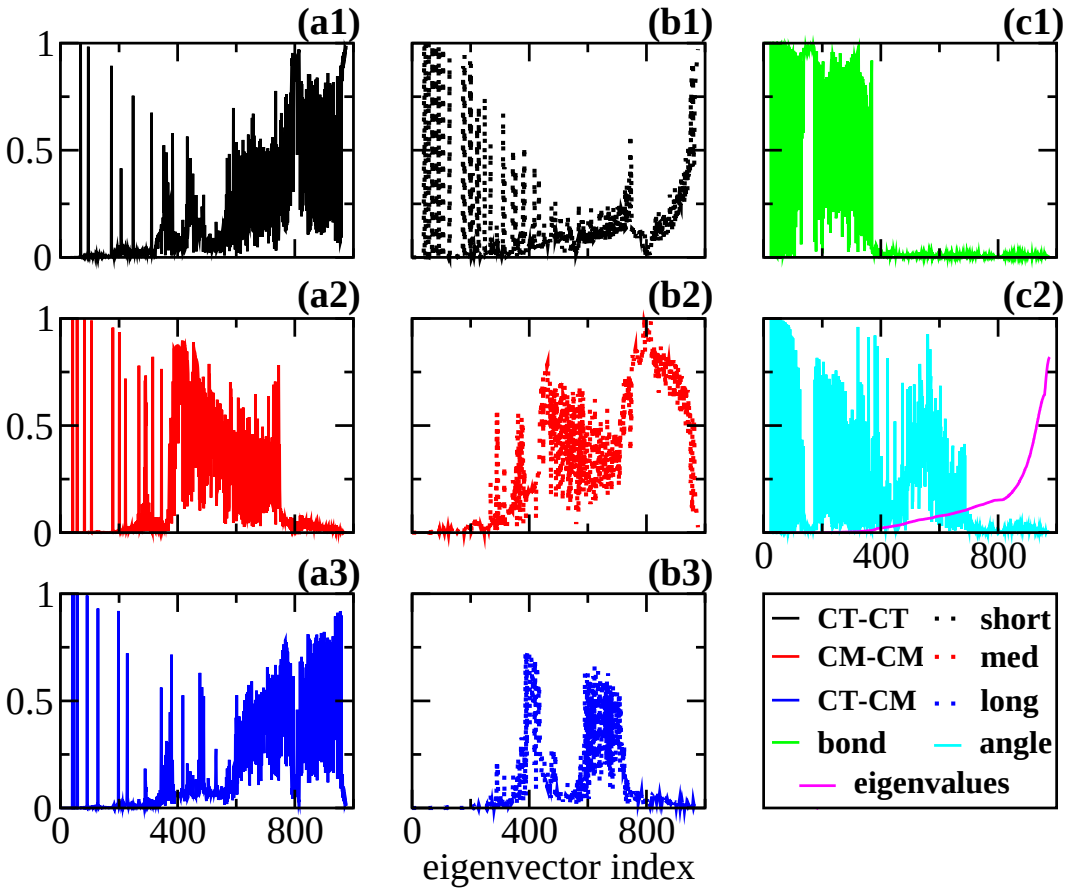

Figure 4.6. Analysis of the eigenvectors and eigenvalues for the normalized metric tensor calculated using the three-site heptane mapping. In column a, rows 1, 2, and 3 present the participation fractions for CT-CT, CM-CM, and CT-CM nonbonded pair interactions, respectively. In column b, rows 1, 2, and 3 present corresponding quantities for short-, medium-, and long-ranged nonbonded interactions, respectively. In column c, rows 1 and 2 present corresponding quantities for intramolecular bond and angle interactions, respectively. Finally, Figure 4.5c2 also presents the eigenvalue spectrum in magenta.

vectors correspond to very short-ranged interactions that are only rarely sampled.

The inverse of the metric tensor determines a linear transformation from projections, $b_\zeta(x)$, of the many-body MF field to MS-CG force functions, $\phi_\zeta(x)$. While Figure 4.5 employed the metric tensor to decompose atomistic pair mean forces into specific contributions from MS-CG force functions, the present eigenanalysis decomposes these MS-CG forces into contributions from individual eigenvectors. In particular, the largest eigenvalues correspond to those interactions that are most significant for reproducing projections, $b_\zeta(x)$, of the many-body MF and that are, consequently, most robustly determined by the MS-CG method. The smallest eigenvalues correspond to interactions that are least significant for reproducing $b_\zeta(x)$ and that are least well determined. By systematically eliminating the contributions of spe-

cific eigenvectors from the MS-CG force field, we investigate the sensitivity of the MS-CG model to various features of the CG potential and also to many-body correlations present in the atomistic model.

Figure 4.7 presents this analysis. In each case, the solid black curves correspond to the MS-CG force field calculated by numerically solving Equation 4.13. The dashed red and dashed blue curves correspond to results after eliminating the contributions to the MS-CG nonbonded forces from the smallest 350 and smallest 800 eigenvectors, respectively. The left, center, and right columns of Figure 4.7 correspond to results for the CT-CT, CM-CM, and CT-CM nonbonded interactions, respectively.

Row a of Figure 4.7 quantifies the sensitivity of the MS-CG nonbonded forces to these eigenvectors. The solid black curves present the MS-CG force functions, which correspond to the solid red curves in Figure 4.5. The dashed red curves in Figure 4.7a demonstrate that the smallest 350 eigenvectors are relatively insignificant for representing the CT-CM force function, but are more significant for the CT-CT and CM-CM forces. As expected from Figure 4.6, the smallest eigenvectors contribute minimally to repulsive short-ranged forces, but are important for representing longer-ranged forces. The dashed blue curves demonstrate that only 177 eigenvectors are necessary to accurately represent the short-ranged repulsion between CT-CT and CT-CM pairs, as suggested by Figure 4.6. In contrast, the 177 largest eigenvectors are not sufficient to represent even the short-ranged component of the CM-CM force.

Figure 4.7b presents projections of these CG force fields onto force field basis vectors for the nonbonded pair interactions. In each case, these projections were calculated by numerically evaluating the left hand side expression of Equation (4.8). The solid black curves present projections of the MS-CG force field, which quantitatively agree with the corresponding projections of the many-body MF, i.e., $\mathbf{F}^0 \odot \boldsymbol{\mathscr{G}}_\zeta(x)$ in Equation (4.7). As described above, these projections are closely related to the atomistic pair mean forces (black curves) in Figure 4.5. In particular, the first peaks in the atomistic rdfs correspond to $b_\zeta(r) = 0$ at $r \approx 0.5$ nm.

The dashed red curves in Figure 4.7b demonstrate that these projections of the many-body MF can be quite accurately reproduced after eliminating the contributions of 350 eigenvectors from the CG nonbonded force functions. (In calculating Figures 4.7b and 4.7c, the contributions of these eigenvectors were not eliminated from the intramolecular force field.) In particular, $b_\zeta(r)$ is reproduced with near quantitative accuracy for the CT-CT and CT-CM interactions. In contrast, the dashed blue curves demonstrate that the 177 largest eigenvectors are not sufficient to accurately reproduce $b_\zeta(r)$, though the resulting force field

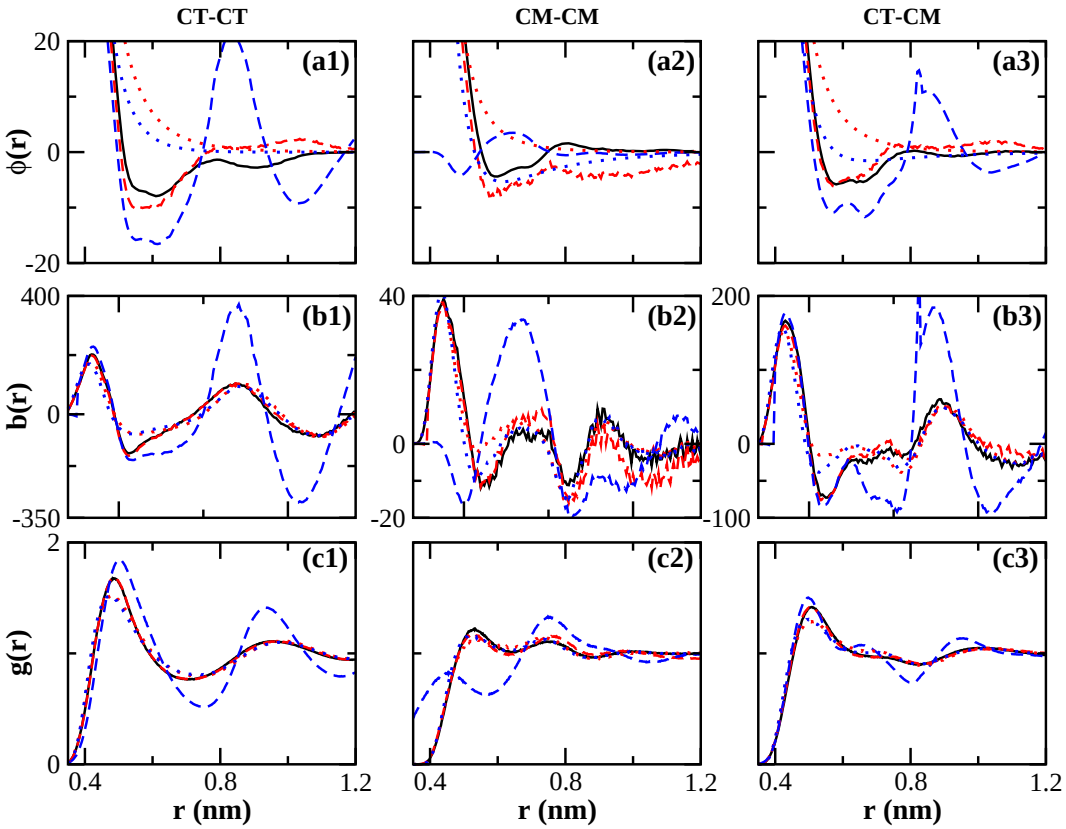

Figure 4.7. Eigenvector analysis of the MS-CG nonbonded pair forces (row 1), corresponding force projections (row 2), and resulting rdfs (row3). Rows 1 and 2 are presented in units of kJ mol^{-1} nm^{-1}. Columns 1, 2, and 3 present results for the CT-CT, CM-CM, and CT-CM nonbonded pair forces, respectively. In each case, the solid black curves present results for the MS-CG force field when nonbonded forces were represented on a grid, while the dashed red and dashed blue curves present results after eliminating the contributions of 350 and 800 eigenvectors, respectively, from these nonbonded forces. The dotted red and dotted blue curves present results for the MS-CG force field, when using 12-6 and 8-4 Lennard-Jones-type functions to represent the nonbonded forces.

reasonably estimates $b_\zeta(r)$ for the CT-CT and CT-CM interactions at short distances.

Finally, the third row of Figure 4.7 presents rdfs calculated from simulations with each CG force field. The black curves present rdfs from simulations of the MS-CG model. Supporting Figure S7 demonstrates that these rdfs agree quantitatively with corresponding atomistic rdfs. The dashed red (blue) curves present rdfs from simulations with the CG force field after eliminating the smallest 350 (800) eigenvectors from the nonbonded force functions. The dashed red curves demonstrate that these contributions from the 350 smallest eigenvectors are not necessary to reproduce the atomistic mean forces. In fact, the resulting CG simulations quantitatively reproduce the atomistic rdfs. However, as expected, the dashed blue curves demonstrate that the nonbonded interactions defined by only 177 eigenvectors are insufficient to reproduce the atomistic CM-CM rdf. Nevertheless, this very reduced CG

force field is sufficient to qualitatively reproduce the atomistic rdfs for the CT-CT and CT-CM pairs. Rows b and c suggest that, in the case of liquid heptane, the accuracy of the CG model in reproducing the projections of the many-body MF provides a predictive indicator of the accuracy with which simulations of the CG model will reproduce atomistic rdfs.

We have also performed similar eigenvector analysis of $\bar{G}_{\zeta\zeta'}(x,x')/\, r_\zeta^2(x) r_{\zeta'}^2(x')$. Supporting Figure S21 presents this correlation function after eliminating the smallest 350 and 800 eigenvectors. The correlation functions corresponding to the CT-CT and CT-CM interactions are only slightly impacted by the removing these eigenvectors. On the other hand, the correlation functions related to the CM-CM interaction are more significantly impacted. After eliminating the contributions from 800 eigenvectors, these CM-CM correlation functions are not even qualitatively described.

These results suggest that projections of the CG force field and also the rdfs generated by CG simulations are quite insensitive to many features of the CG force field. In practice, it may be possible to accurately reproduce atomistic rdfs with relatively simple potentials. To test this hypothesis, we recalculated the MS-CG force field, while using simple Lennard-Jones-type $U_\zeta^{n-m}(x)$ functions to represent nonbonded pair potentials. In this case, the three nonbonded interactions were determined by a total of six parameters, instead of the 527 parameters required in the tabulated representation. (As before, intramolecular interactions were tabulated on a grid.) The dotted red and dotted blue curves in Figure 4.7 present the results for $n-m$ = 12-6 and 8-4, respectively.

Figure 4.7a demonstrates that these Lennard-Jones-type nonbonded force functions are almost completely repulsive and differ significantly from the force functions calculated for the more flexible discrete delta basis set (solid black curves), although the sets of forces agree quite well in the repulsive hard-sphere regions. Nevertheless, despite these differences, Figure 4.7b demonstrates that the $n-m$ force functions quite accurately reproduce the projections, b_ζ, of the many body MF (solid black curve). These results demonstrate that force field projections calculated with $G_{\zeta\zeta'}$ diminish differences between CG force fields, which is consistent with the earlier observation that the corresponding eigenvalues are all less than one. Finally, the dotted curves in Figure 4.7c demonstrate that MD simulations with these $n-m$ force functions also reproduce the atomistic rdfs with reasonable accuracy. These results provide further evidence that the rdfs calculated from CG MD simulations are similarly insensitive to many details of the CG force fields.

4.5 Discussion

Structure-motivated coarse-graining approaches typically approximate the many-body PMF with molecular mechanics potentials. Each term in this potential models a particular interaction with a function of a single scalar variable. These approximate potentials are parameterized to reproduce atomistic distribution functions corresponding to each of these individual variables, e.g., rdfs. This objective is equivalent to requiring that the CG model should reproduce the atomistic mean force[43] (or torque[198]) for each interaction. However, the mean forces generated by the CG model include not only a direct contribution from the corresponding interaction, but also correlated contributions from the environment.[62] The many-body correlations that generate these correlated contributions represent a significant challenge for developing CG models that accurately reproduce atomistic structure. An improved understanding of the impact of many-body correlations for determining CG potentials and for the structural accuracy of the resulting models should lead to optimized CG mappings,[114] improved approximations to the many-body PMF, and perhaps even meaningful *a priori* estimates of errors in CG models.[120]

Various approaches treat these correlations in different ways. Direct Boltzmann inversion assumes that nonbonded pair mean forces in the CG model only reflect the corresponding direct force. This approach completely neglects the effects of correlated interactions and, depending upon the density of CG sites, may provide limited accuracy for modeling the structure of complex condensed phases. Methods such as IBI and IMC require multiple simulations to assess the significance of many-body correlations in the CG model and to systematically refine the CG potential. Although these methods typically focus upon rdfs, Soper[183] demonstrated that IBI updates the CG force functions with the difference between the atomistic and CG pair mean forces. At the same time, the force-based MS-CG method directly determines CG forces based upon atomistic pair mean forces via equations that are equivalent to a g-YBG equation.[63,64] Consequently, mean forces provide an intriguing connection between iterative structure-based and direct force-based methods for determining CG potentials.

As illustrated in Figure 4.1, the MS-CG and g-YBG methods address the effects of many-body correlations via a geometric framework that is approximate, but direct and transparent. This approach defines projections, $b_\zeta(x)$, of the many-body MF onto each vector, $\boldsymbol{\mathscr{G}}_\zeta(x)$, in an incomplete basis set that is determined by the approximate CG force field.[41,60] The metric tensor, $G_{\zeta\zeta'}(x,x')$, decomposes each projection $b_\zeta(x)$ into 1) a direct contribution from the corresponding force function, $\phi_\zeta(x)$; and 2) correlated contributions from the other CG force

functions, $\{\phi_{\zeta'}(x')\}$.

The present calculations demonstrate that $G_{\zeta\zeta'}$ reflects not only local packing properties in the atomistic model, but also the CG representation. In particular, $G_{\zeta\zeta'}$ is quite sensitive to relatively subtle effects, such as the impact of rotameric states upon the asymmetry of surrounding solvation shells. This coupling between inter- and intra-molecular interactions suggests that the common practice of determining intramolecular potentials from direct Boltzmann inversion may adversely impact the intermolecular structure of a CG model. In the case of heptane, $G_{\zeta\zeta'}$ can be interpreted by simple models with similar geometry (Figure 4.4). In more complex systems, though, strong specific interactions may bias particular configurations and generate distinct features in $G_{\zeta\zeta'}$. For instance, calculations for the SPC/E water model[205] demonstrate that hydrogen bonding generates features that cannot be attributed to generic packing properties. (Supporting Figure S26.) Consequently, we anticipate that $G_{\zeta\zeta'}$ may be particularly useful, not only for determining accurate CG potentials, but also for identifying interactions that are essential for stabilizing complex molecular structures.

In addition to providing efficient computational models, structure-motivated CG approaches that approximate the many-body PMF can also provide quantitative insight into the physical forces that underly particular phenomena. The many-body PMF is a configuration dependent free energy function that reflects both energetic and entropic contributions.[41,44,109] Consequently, several insightful studies have employed the individual terms in approximate CG potentials to quantify the thermodynamic forces underlying, e.g., hydrophobic self-assembly[208] and cellulose dissolution,[209] and to decompose these interactions into energetic and entropic components.[203,208] However, this analysis is somewhat complicated, since it is only in combination that these effective potentials rigorously approximate a free energy function.

The present work suggests a somewhat more precise analysis. The pair potential of mean force for a particular interaction describes the reversible work, i.e., free energy change, along that reaction coordinate.[43] Equation (4.12) decomposes the corresponding pair mean force into specific contributions from each term in the CG potential. In the specific case of a 3-site model for heptane, Figure 4.5 demonstrates that correlated interactions significantly impact the mean force. In particular, at short distances near the first peak of the rdf, these correlated forces dramatically reduce the mean force relative to the corresponding direct force. In addition, this analysis highlighted the effects of nearby neighbors upon the mean force for particles that are separated by greater distances. We note that this analysis depends on the interactions considered in the decomposition and, thus, must be interpreted with

caution. Nevertheless, we anticipate that this framework may prove useful in quantifying specific contributions to potentials of mean force for complex molecular systems.

In order to further investigate the role of the many-body correlations in determining the MS-CG force field, we performed an eigenvector analysis after rescaling $G_{\zeta\zeta'}$ according to Figure 4.3 and Figure 4.4. The largest eigenvalues of the rescaled matrix, which contribute most to projections of the CG force field, reflect short- and medium-ranged components of the CT-CT and CT-CM interactions. In contrast, most of the smallest eigenvalues reflect intramolecular and long-ranged interactions.

We employed these eigenvectors to represent the MS-CG force field and, in particular, to eliminate the contributions of small eigenvectors from the nonbonded force functions. These calculations further demonstrated that projections, i.e., $\mathscr{G}_\zeta(x) \odot \mathbf{F}$, onto force field basis vectors, $\mathscr{G}_\zeta(x)$, are relatively insensitive to many features of the CG force field, $\mathbf{F}$. In particular, additional calculations with Lennard-Jones-type potentials demonstrated that markedly different pair potentials generated quantitatively similar force field projections. These results are consistent with the observation that the eigenvalue spectrum of $G_{\zeta\zeta'}$ is bounded above by one. Consequently, when determining these projections, $G_{\zeta\zeta'}$ contracts the difference between CG force fields. This may be a relatively general property of $G_{\zeta\zeta'}$, although further analysis is required to assess this supposition.

We also performed MD simulations with these different CG force fields to assess their accuracy in reproducing atomistic rdfs. Simulations with the MS-CG force field (defined by a flexible discrete delta basis) quantitatively reproduced the structure of the OPLS-AA heptane model. However, additional MD simulations demonstrated that the CG rdfs are remarkably insensitive to the contributions of the 350 smallest eigenvectors to the MS-CG pair potentials. MD simulations with significantly different Lennard-Jones-type pair potentials (determined by MS-CG calculations) also reproduced the atomistic rdfs with near quantitative accuracy. These results are consistent with many previous studies indicating that simulated rdfs are sensitive to short-ranged repulsive potentials, but relatively insensitive to other features of the CG potential.[35,47]

Our results also indicate that, at least for heptane, the accuracy of a CG force field in reproducing the projections, $b_\zeta(x) \equiv \mathscr{G}_\zeta(x) \odot \mathbf{F}^0$, of the many body MF, $\mathbf{F}^0$, provides a reasonable indicator for the accuracy of the CG force field in reproducing the atomistic rdfs. This observation is consistent with the remarkable success of the MS-CG method in modeling atomistic structure.[36,59,132,151,186,210–212] However, this observation is also somewhat surprising. The CG rdfs are generated by MD simulations and reflect a complicated and nonlinear relationship between the CG potentials and the resulting rdfs. In contrast,

according to Equation (4.8), the projections of the CG force field are linearly related to the CG potential.

As noted above, the MS-CG method determines a CG force field that quantitatively matches projections of the many-body MF, i.e., $\boldsymbol{\mathscr{G}}_\zeta(x) \odot \mathbf{F} = \boldsymbol{\mathscr{G}}_\zeta(x) \odot \mathbf{F}^0$. In the case of non-bonded interactions, these projections have obvious physical significance because $\boldsymbol{\mathscr{G}}_\zeta(x) \odot \mathbf{F}^0$ corresponds to an atomistic mean force.[63,188] If simulations of the CG model reproduce these atomistic mean forces, then the CG model will also reproduce the corresponding rdfs. Neverthless, the MS-CG potentials are not guaranteed to reproduce atomistic rdfs. This is because the calculated projections of the CG force field, $\boldsymbol{\mathscr{G}}_\zeta(x) \odot \mathbf{F}$, weight the contributions of each force function, $\phi_\zeta(x)$, according to $G_{\zeta\zeta'}$. (See Equation (4.8).) Equation (4.9) demonstrates that $G_{\zeta\zeta'}$ is defined by an ensemble average according to the atomistic probability distribution or, equivalently, the many-body PMF, i.e., $G_{\zeta\zeta'} \equiv G_{\zeta\zeta'}[U^0]$. Thus these force field projections are calculated by applying the MS-CG force field to configurations sampled by the atomistic model. However, when attempting to reproduce atomistic mean forces, it is not the many-body correlations in the atomistic model that are crucial, but rather the many-body correlations generated by the CG model. Since simulations with the approximate CG force field will not generate an identical distribution of configurations to the many-body PMF, the MS-CG calculation may incorrectly estimate the contributions of correlated interactions to the CG mean forces. Consequently, MD simulations with the MS-CG force field may not perfectly reproduce the atomistic mean forces or, equivalently, rdfs.

These considerations suggest a self-consistent formulation of the MS-CG method for reproducing the atomistic mean forces. The CG potential, U^*, that reproduces a set of atomistic mean forces must satisfy Equation (4.8) for b_ζ when the contribution of each force function is weighted according to $G_{\zeta\zeta'}[U^*]$, i.e., the corresponding ensemble average for U^*. Clearly, the normal MS-CG equations approximate this self-consistent equation by $G_{\zeta\zeta'}[U^*] \approx G_{\zeta\zeta'}[U^0]$. A more accurate estimate of $G_{\zeta\zeta'}[U^*]$ should lead to a more accurate estimate of U^* and a more accurate reproduction of atomistic rdfs. This self-consistent condition underlies the iterative-YBG framework developed by Cho and Chu.[69] In their approach, the normal MS-CG equations are iteratively solved, each time using a new estimate of $G_{\zeta\zeta'}[U^*]$ that is determined by the previous approximate CG potential. In this context, it is particularly intriguing that the eigenvalues of $G_{\zeta\zeta'}[U^0]$ are often bounded by 1, which may play an important role in the accuracy and robustness of the MS-CG method, as well as the relative insensitivity of rdfs to many features of CG potentials.[35] We also note that Weeks and coworkers have developed a considerably different self-consistent framework for determining short-ranged effective potentials that is also based upon the YBG equation.[213,214]

Finally, we anticipate that the present development of the g-YBG framework may prove useful for quantifying and systematically improving the approximations underlying knowledge-based CG protein models.[215] Knowledge-based methods often determine potentials between pairs of amino acids based upon the frequency of observing the pair at a given distance in protein databank structures.[147,216,217] The contributions of the surrounding environment to the corresponding mean force are typically estimated by a "reference state," although the precise definition of this reference state remains somewhat unclear.[218,219] Mullinax and Noid[150] have demonstrated that, by directly addressing these contributions, an extended ensemble[146] version of the g-YBG theory[63,64] can quantitatively recover the underlying potential from a model protein databank generated with a popular protein potential.[220] In this context, Equation (4.12) provides a quantitative framework for identifying the contribution of this reference state to the mean force. Future work will investigate these contributions in the context of protein structures.

4.6 Summary and Conclusions

The present work reports a detailed analysis of many-body correlations in determining CG models that accurately reproduce atomistic distribution functions. In particular, these many-body correlations enter into the MS-CG and g-YBG approaches as a metric tensor that defines the angle between force field vectors, based upon correlations between the corresponding interactions in the atomistic model. This metric tensor decomposes atomistic mean forces into direct and correlated contributions from the CG force field. Our calculations for liquid heptane demonstrate that this metric tensor reflects relatively generic features of molecular packing. Moreover, our calculations demonstrate that correlated interactions make significant contributions to these mean forces, although these mean forces appear relatively insensitive to many features of the CG force field.

Our results suggest several future directions. Since the present results correspond to a relatively simple system with weak alignment properties, it will be interesting to further consider this framework for more complex systems. In particular, we anticipate that the MS-CG metric tensor may prove useful for quantifying the interactions and correlations that are important for stabilizing complex molecular structures. Additionally, the g-YBG framework may prove useful for characterizing and improving the approximations inherent to knowledge-based protein potentials. More generally, we anticipate that improved understanding of many-body correlations will contribute to developing CG potentials that more accurately model both structural and thermodynamic properties of complex condensed-phase systems.

Supporting Information Available

The Supporting Information provides a detailed description of all methods, simulations, and calculations employed in this work, as well as additional analysis of these calculations. This information is free of charge via the Internet at `http://pubs.acs.org`.

Chapter 5

Investigation of Coarse-grained Mappings via an Iterative Generalized Yvon-Born-Green Method

J. F. Rudzinski, W. G. Noid *J. Phys. Chem. B* **2014**, 118, 8295-8312

Abstract

Low resolution coarse-grained (CG) models enable highly efficient simulations of complex systems. The interactions in CG models are often iteratively refined over multiple simulations until they reproduce the one-dimensional (1-D) distribution functions, e.g., radial distribution functions (rdfs), of an all-atom (AA) model. In contrast, the multiscale coarse-graining (MS-CG) method employs a generalized Yvon-Born-Green (g-YBG) relation to determine CG potentials directly (i.e., without iteration) from the correlations observed for the AA model. However, MS-CG models do not necessarily reproduce the 1-D distribution functions of the AA model. Consequently, recent studies have incorporated the g-YBG equation into iterative methods for more accurately reproducing AA rdfs. In this work, we consider a theoretical framework for an iterative g-YBG method. We numerically demonstrate that the method robustly determines accurate models for both hexane and also for a more complex molecule, 3-hexylthiophene. By examining the MS-CG and iterative g-YBG models for several distinct CG representations of both molecules, we investigate the approximations of the MS-CG method and their sensitivity to the CG mapping. More generally, we explicitly demonstrate that CG models often reproduce 1-D distribution functions of AA models at the expense of distorting the cross-correlations between the corresponding degrees of freedom. In particular, CG models that accurately reproduce intramolecular 1-D distribution functions may still provide a poor description of the molecular conformations sampled by the AA model. We demonstrate a simple and predictive analysis for determining CG mappings that promote an accurate description of these molecular conformations.

5.1 Introduction

Atomically-detailed, or all-atom (AA), molecular simulations provide a powerful tool for studying complex condensed phase systems. However, despite significant computational advances, AA models remain prohibitively expensive for simulating the length and time scales that are relevant for many complex systems.[24] Consequently, lower resolution, coarse-grained (CG) models have attracted considerable interest in recent years.[23,25,49,67,82,88,96,110,221–225] In many cases, though, CG models will only be useful if they adequately describe the correct physics for the system of interest. In particular, many methods for parameterizing CG models have primarily focused on reproducing either thermodynamic[27,28,98,102] or structural[34,35] properties.

The many-body potential of mean force (PMF) is the appropriate potential for a CG model that reproduces all structural distribution functions that are determined by mapping the AA model to the representation of the given CG model.[41,44,109] However, this PMF is a free energy function that depends upon the coordinates of all CG sites and is too complex to calculate or simulate in practice.[111] Consequently, the interaction potentials for many CG models, i.e., the CG potentials, are often determined as approximations to the many-body PMF.[23]

In particular, several structure-based methods parameterize approximate CG potentials in order to reproduce a set of one-dimensional (1-D) distribution functions, such as the radial distribution functions (rdfs), for the CG sites that are determined by mapping the AA ensemble to the given representation.[34,35,123] Because there exists no general way for directly determining potentials that reproduce rdfs for condensed phase systems,[118,183,226] these approaches iteratively refine the CG potential until the target distributions are adequately reproduced. Iterative methods have successfully modeled complex liquids, polymers, and membranes.[38,49–58] However, they do not guarantee the reproduction of higher order correlations and, to date, have enjoyed less success for modeling systems, such as proteins, with important higher order structure.[227–229]

Recently, several variational methods have been developed for systematically approximating the many-body PMF. In particular, the relative entropy method[37,48,121] determines the CG potential by minimizing an information functional[185] that quantifies the difference between the configurational distributions for the CG sites that are generated by the AA and CG models. This calculation is also generally iterative and, under certain circumstances,[124] is equivalent to the inverse Monte Carlo method[52] for reproducing 1-D distribution functions.

The multiscale coarse-graining (MS-CG) method[36,59] employs a different variational prin-

ciple that determines the CG potential to match the net forces[158] that are generated on each CG site by an AA model. The MS-CG method directly (i.e., non-iteratively) determines an optimal approximation to the PMF by projecting the corresponding force field, i.e., the many-body mean force field, into a linear space of force fields that is determined by the form of the approximate CG potential.[41,60]

We have previously interpreted the MS-CG method as a force balance relation between AA and CG models.[230] For each degree of freedom (dof) treated by the CG potential, the MS-CG method first determines the average force generated by the AA model along that dof. For central pair potentials, this average force is equivalent to the pair mean force[166] as a function of distance.[62–64,134,188] If the MS-CG model reproduces a pair mean force of the AA model, then it will also reproduce the corresponding rdf. The MS-CG procedure then employs a generalized Yvon-Born-Green (g-YBG) equation[62–64,134,188] to decompose these AA average forces into correlated contributions from each term in the CG potential. In performing this decomposition, the MS-CG method assumes that the cross-correlations of the AA model accurately approximate the cross-correlations that will be generated by the resulting MS-CG model. This approximation is a key feature of the MS-CG method since it enables the CG force field to be determined directly from AA correlations without iterative CG simulations. In general, though, the CG model will not perfectly reproduce the cross-correlations that are observed in the AA model. Consequently, MS-CG models are not guaranteed to reproduce the 1-D distribution functions of the AA model.

The MS-CG method has accurately reproduced 1-D distributions of AA models for many condensed phase systems.[65,66] However, Ruhle et al.[74] demonstrated that a 3-site MS-CG model of hexane provided a surprisingly poor description of the angle distribution for the 3 sites. In analyzing this case, Das et al.[73] concluded that the simple bond and angle terms included in the approximate CG potential were incapable of reproducing the complex bond-angle cross-correlations that are generated by mapping the AA model to the 3-site representation. Interestingly, they demonstrated that 2- and 4-site MS-CG models for hexane accurately reproduced the 1-D distributions that resulted from mapping the AA model to the corresponding representations. Prior work has also explicitly demonstrated that the accuracy and transferability of MS-CG models do not necessarily improve with increasing resolution of the CG representation.[146] These considerations motivate efforts to improve the structural accuracy of the MS-CG method and to predictively determine optimal CG mappings.

Several groups have previously investigated non-conventional force-matching methods in order to better reproduce the 1-D distribution functions of AA models. Cho and Chu

proposed iteratively solving the g-YBG equation, while replacing the cross-correlations of the AA model with the cross-correlations generated by each successive CG model.[69] They demonstrated that this iterative g-YBG method determined intermolecular pair potentials that quantitatively reproduced AA rdfs for several liquid systems. More recently, Lu et al. demonstrated an alternative approach that iteratively refines the MS-CG force field in a similar manner, but employs the cross-correlations of the AA model in each iteration.[75] However, because both studies considered CG models with relatively few intramolecular dofs, it remains unclear to what extent these approaches can be applied for more complex CG models. The present work further develops the iterative g-YBG method, clarifies the relationship between these two previous approaches, and numerically demonstrates that the method robustly determines accurate CG models for both hexane and also a considerably more complex molecule, 3-hexylthiophene (3HT).

The present work then applies this iterative g-YBG method to investigate the sensitivity of CG models to the CG mapping. The CG mapping determines the complexity of the many-body PMF and significantly influences the accuracy and transferability of CG models.[23] However, relatively little progress has been achieved for predicting optimal CG mappings. Previous studies have proposed optimizing CG mappings based upon either structural[231–235] or dynamic[236–239] features of an AA model. Nevertheless, CG mappings are most often determined by the chemical intuition of the researcher. Consequently, the present work investigates, for both hexane and 3HT, the influence of the CG mapping upon the structural accuracy of MS-CG models. We then employ the iterative g-YBG method to determine CG potentials that accurately reproduce the relevant 1-D structural distributions of the AA models. We explicitly demonstrate that an accurate reproduction of these 1-D distributions does not ensure an accurate model for either the higher order structural correlations or the molecular conformations that are sampled by the AA model. We demonstrate that the CG mapping significantly influences the relationship between 1-D and higher order correlations. On this basis, we identify distinguishing features of poor mappings and develop a predictive framework for assessing the quality of the CG mapping for reproducing molecular conformations.

The remainder of this manuscript is organized as follows. In the Theory section, we consider a framework for iterative methods that are based upon the g-YBG equation.[69,75] In the Methods section, we outline minor heuristic modifications to the iterative g-YBG method that appear to improve its stability and robustness for treating intramolecular interactions. Additionally, the Methods section describes our definition and characterization of molecular conformations. The Results section presents several CG models for both hexane and 3HT.

For each molecular system, prior to considering the CG models, we first characterize the molecular conformations that are sampled by the corresponding AA model. We then determine how these conformations are represented and distinguished by different CG mappings. We next apply both the MS-CG and the iterative g-YBG method to determine potentials for each CG representation of each system. In each case, we demonstrate that the iterative g-YBG model provides better structural accuracy than the MS-CG model and typically reproduces the 1-D distribution functions of the AA model with near quantitative accuracy. We also demonstrate that the CG mapping strongly influences the structural fidelity of the MS-CG model, as well as the reproduction of molecular conformations by the iterative g-YBG model. Finally, the Discussion section summarizes and analyzes the conclusions of this study.

5.2 Theory

This section considers a theory for iterative methods that employ the g-YBG equation.

5.2.1 Linear Space of CG Force Fields

We first consider a CG model, whose configuration is defined by the Cartesian coordinates, $\mathbf{R}$, for N sites. We assume that the model has a potential function with a molecular mechanics form:

$$U(\mathbf{R}) = \sum_{\zeta} \sum_{\lambda} U_{\zeta}(\psi_{\zeta}(\{\mathbf{R}\}_{\lambda})), \tag{5.1}$$

where ζ indicates a particular interaction (e.g., a dihedral angle interaction) and U_{ζ} is the corresponding potential (e.g., a dihedral angle potential) that is a function of a single scalar variable, ψ_{ζ}, (e.g., a dihedral angle) that may be expressed as a function of the Cartesian coordinates, $\{\mathbf{R}\}_{\lambda}$, for a set of sites, λ (e.g., the 4 successively bonded sites that form a dihedral angle).[60] Given a configuration $\mathbf{R}$, the force on each site $I = 1, \ldots, N$ may be expressed:

$$\mathbf{F}_I(\mathbf{R}) = \sum_{\zeta} \sum_{\lambda} F_{\zeta}(\psi_{\zeta}(\{\mathbf{R}\}_{\lambda}))\, \partial\psi_{\zeta}(\{\mathbf{R}\}_{\lambda})/\partial\mathbf{R}_I, \tag{5.2}$$

where $F_{\zeta}(x) = -dU_{\zeta}(x)/dx$ is a scalar function determining the magnitude of the force on the site from the ζ-type of potential, while the gradient term determines the direction of the corresponding force on each site. We represent each force function, $F_{\zeta}(x)$, by a set of basis

functions, $\{f_{\zeta d}(x)\}$, with a corresponding set of coefficients, $\{\phi_{\zeta d}\}$:

$$F_\zeta(x) = \sum_d f_{\zeta d}(x)\phi_{\zeta d}. \tag{5.3}$$

The coefficients $\{\phi_{\zeta d}\}$ will act as parameters for the force field. The force on each site I in Equation (5.2) may then be expressed:

$$\mathbf{F}_I(\mathbf{R}) = \sum_\zeta \sum_d \phi_{\zeta d}\, \boldsymbol{\mathscr{G}}_{I;\zeta d}(\mathbf{R}), \tag{5.4}$$

where $\boldsymbol{\mathscr{G}}_{I;\zeta d}(\mathbf{R}) = \sum_\lambda f_{\zeta d}(\psi_\zeta(\{\mathbf{R}\}_\lambda))\, \partial\psi_\zeta(\{\mathbf{R}\}_\lambda)/\partial\mathbf{R}_I$. As in previous work,[41,60,63] we define a force field $\mathbf{F}$ as a set of N vector valued functions $\{\mathbf{F}_1(\mathbf{R}), \ldots, \mathbf{F}_N(\mathbf{R})\}$ that specify a vector force on each site $I = 1, \ldots, N$ in each configuration, $\mathbf{R}$. This force field may be represented

$$\mathbf{F}(\boldsymbol{\phi}) = \sum_D \phi_D \mathcal{G}_D \tag{5.5}$$

where D is a "super-index" identifying a particular pair of indices ζd. We define N_D as the number of force field coefficients. Then, for each $D = 1, \ldots N_D$, $\mathcal{G}_D$ corresponds to a set of N vector valued functions $\{\boldsymbol{\mathscr{G}}_{1;D}(\mathbf{R}), \ldots, \boldsymbol{\mathscr{G}}_{N;D}(\mathbf{R})\}$ that specify the direction of the force on each site $I = 1, \ldots, N$ that is associated with the coefficient ϕ_D. The set, $\{\mathcal{G}_D\}$, of N_D such vector valued functions that are included in Equation (5.5) defines a basis for a N_D-dimensional linear space of CG force fields, which we shall refer to as "ϕ-space." A point $\boldsymbol{\phi}$ in ϕ-space determines the set, $\{\phi_D\}$, of N_D force field coefficients and thus determines a particular force field, $\mathbf{F}(\boldsymbol{\phi})$, as well as a corresponding potential, $U(\mathbf{R}; \boldsymbol{\phi})$.

This representation of CG force fields is very general. However, this paper will focus on the particularly common special case that the nonbonded contribution to the potential is represented by a sum of central pair potentials, each of which is a function of the distance r between a pair of sites. Moreover, we shall assume that each of these pair potentials is represented by relatively flexible basis functions, such as spline functions. The coefficients, $\{\phi_{\zeta d}\}$, in Equation (5.3) for each such nonbonded pair potential may then correspond to the magnitudes of the pair force at a corresponding discrete set of distances, $\{r_{\zeta d}\}$, i.e., $\phi_{\zeta d} \leftrightarrow F_\zeta(r_{\zeta d})$. This special case is particularly common for parameterizing CG models that accurately reproduce structural correlations of an AA model.[34–36,38,72,121]

5.2.2 The Generalized Yvon-Born-Green (g-YBG) Equation

For any force field $\mathbf{F}(\boldsymbol{\phi})$ in ϕ-space, the g-YBG equation[63,64] provides an exact identity that relates the force field coefficients, ϕ_D, to a corresponding set of equilibrium ensemble averages, b_D:

$$b_D(\boldsymbol{\phi}) = \sum_{D'} G_{DD'}(\boldsymbol{\phi})\phi_{D'}, \tag{5.6}$$

for each $D = 1, \ldots, N_D$, where

$$b_D(\boldsymbol{\phi}) = \int d\mathbf{R}\, P_R(\mathbf{R}|\boldsymbol{\phi}) \frac{1}{3N} \sum_I \boldsymbol{\mathscr{G}}_{I;D}(\mathbf{R}) \cdot \mathbf{F}_I(\mathbf{R};\boldsymbol{\phi}), \tag{5.7}$$

$$G_{D,D'}(\boldsymbol{\phi}) = \int d\mathbf{R}\, P_R(\mathbf{R}|\boldsymbol{\phi}) \frac{1}{3N} \sum_I \boldsymbol{\mathscr{G}}_{I;D}(\mathbf{R}) \cdot \boldsymbol{\mathscr{G}}_{I;D'}(\mathbf{R}), \tag{5.8}$$

and $P_R(\mathbf{R}|\boldsymbol{\phi})$ is the equilibrium configuration distribution for a CG model with the corresponding potential $U(\mathbf{R};\boldsymbol{\phi})$.

We have previously discussed the physical significance of these g-YBG equations, as well as their relation to various coarse-graining methods.[48,63,64,230] In the g-YBG equation, $b_D(\boldsymbol{\phi})$ is the projection of the force field $\mathbf{F}(\boldsymbol{\phi})$ onto the basis vector $\mathcal{G}_D$, while $G_{D,D'}(\boldsymbol{\phi})$ quantifies the cross-correlations between the forces due to ϕ_D and $\phi_{D'}$ that contribute to $b_D(\boldsymbol{\phi})$. In the common case that is discussed at the end of the preceding subsection, a subset of the $b_{\zeta d}$ are in 1-1 relationship with the radial distribution functions (rdfs) generated by the CG model. More precisely, if the set of nonbonded coefficients, $\{\phi_{\zeta d}\}$, correspond to non-bonded pair forces at a set of distances, $\{r_{\zeta d}\}$, then the corresponding ensemble averages, $\{b_{\zeta d}\}$, determine the pair mean forces and, thus also the rdfs, at those distances.

Given this particularly common case, the g-YBG relation (Equation (5.6)) provides a convenient means for considering the equilibrium structure generated by CG models. We shall associate a set of equilibrium ensemble averages, $\{b_D\}$ for $D = 1, \ldots, N_D$, with a vector $\boldsymbol{b}$ in an an abstract N_D dimensional space, which we shall refer to as "b-space." Each point $\boldsymbol{b}$ in b-space determines a set of rdfs generated by a CG model. The g-YBG equations then determine an operator $\hat{\mathbf{G}}$ that maps each point $\boldsymbol{\phi}$ in ϕ-space to a corresponding point $\boldsymbol{b}(\boldsymbol{\phi})$ in b-space:

$$\hat{\mathbf{G}} : \boldsymbol{\phi} \rightarrow \boldsymbol{b}(\boldsymbol{\phi}) = \mathbf{G}(\boldsymbol{\phi})\boldsymbol{\phi}, \tag{5.9}$$

where $\mathbf{G}(\boldsymbol{\phi})$ is the matrix with elements $G_{D,D'}(\boldsymbol{\phi})$ given by Equation (5.8) and the right hand operation corresponds to standard matrix-vector multiplication. Because $\mathbf{G}$ depends upon $\boldsymbol{\phi}$ through $P_R(\mathbf{R}|\boldsymbol{\phi})$, it is generally a highly nonlinear function of $\boldsymbol{\phi}$. In practice, this

mapping from ϕ-space to b-space is determined by simulating the CG model with the force field $\mathbf{F}(\boldsymbol{\phi})$ and computing the ensemble averages, $\boldsymbol{b}(\boldsymbol{\phi})$. A reverse transformation from b-space to ϕ-space can be determined by simply inverting the matrix $\mathbf{G}(\boldsymbol{\phi})$. In our experience, this inverse transformation is typically well-defined and numerically robust.[63,64,150,188] We shall assume this to be so in the present work. Figure 5.1a schematically illustrates the mapping $\hat{\mathbf{G}}$ from ϕ-space to b-space and its inverse.

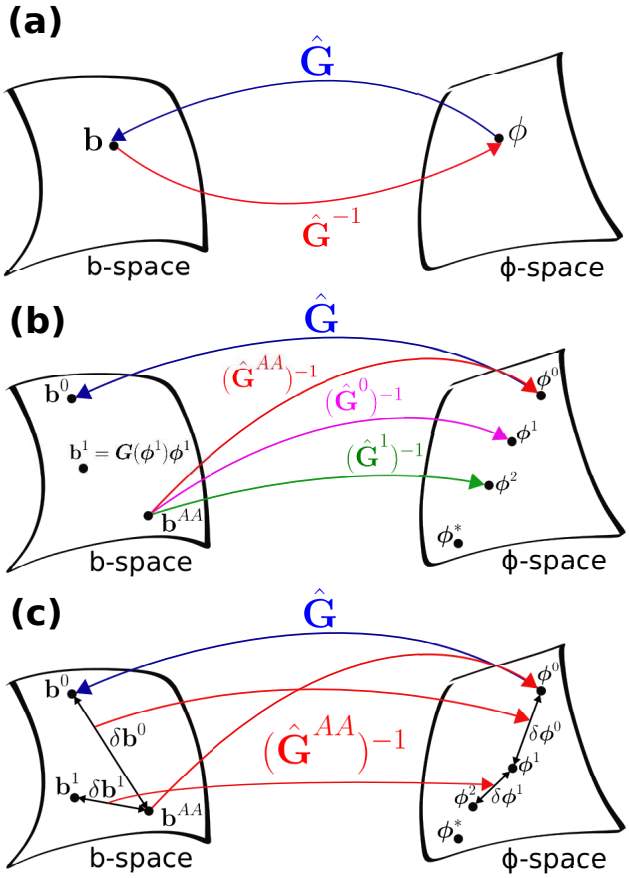

Figure 5.1. Schematics of the g-YBG relation (a), the iterative g-YBG procedure (b), and a locally linear approximation (c). The g-YBG operator $\hat{\mathbf{G}}$ maps a force field, $\boldsymbol{\phi}$, in ϕ-space to a point, $\boldsymbol{b}$, in b-space that determines corresponding rdfs, while $\hat{\mathbf{G}}^{-1}$ denotes the inverse operator.

Although $\hat{\mathbf{G}}$ generally depends upon $\boldsymbol{\phi}$ in a nonlinear manner, in certain cases, the matrix $\mathbf{G}(\boldsymbol{\phi})$ in Equation (5.6) may vary quite slowly with $\boldsymbol{\phi}$. For instance, if the equilibrium structure is largely determined via packing considerations and molecular geometry,[230] then $\mathbf{G}(\boldsymbol{\phi})$ will be quite insensitive to many features of the model potential $U(\mathbf{R};\boldsymbol{\phi})$. Over the region in ϕ-space for which $\mathbf{G}(\boldsymbol{\phi})$ is slowly varying, one expects $\hat{\mathbf{G}}$ to provide an approximately linear transformation into b-space. In particular, consider two points in ϕ-space, $\boldsymbol{\phi}^1$ and $\boldsymbol{\phi}^2$, that correspond to two points in $\boldsymbol{b}$ space, $\boldsymbol{b}^1 = \mathbf{G}(\boldsymbol{\phi}^1)\boldsymbol{\phi}^1$ and $\boldsymbol{b}^2 = \mathbf{G}(\boldsymbol{\phi}^2)\boldsymbol{\phi}^2$. If $\boldsymbol{\phi}^1$ and $\boldsymbol{\phi}^2$ are

within a domain over which $\mathbf{G}$ is slowly varying, then we may expect that $\mathbf{G}(\boldsymbol{\phi}^1) \approx \mathbf{G}(\boldsymbol{\phi}^2)$ such that $\boldsymbol{b}^2 \approx \boldsymbol{b}^1 + \mathbf{G}(\boldsymbol{\phi}^1)\delta\boldsymbol{\phi}$. This implies that

$$\delta\boldsymbol{b} \approx \mathbf{G}(\boldsymbol{\phi}^1)\delta\boldsymbol{\phi}, \tag{5.10}$$

where $\delta\boldsymbol{\phi} = \boldsymbol{\phi}^2 - \boldsymbol{\phi}^1$ and $\delta\boldsymbol{b} = \boldsymbol{b}^2 - \boldsymbol{b}^1$. We shall refer to this as a "locally linear" approximation for relating changes, $\delta\boldsymbol{\phi}$, in ϕ-space to changes, $\delta\boldsymbol{b}$, in b-space. Conversely, if the CG sites are not densely packed or if the equilibrium structure is sensitive to specific interactions, then $\mathbf{G}(\boldsymbol{\phi})$ may vary quite rapidly with $\boldsymbol{\phi}$ and this locally linear approximation may fail.

5.2.3 Multiscale Coarse-graining (MS-CG)

The preceding subsections related CG force field coefficients to the resulting equilibrium structure. This subsection considers the MS-CG method for determining optimal force field coefficients from an AA model. We consider an AA model with a configuration, $\mathbf{r}$, that is defined by the Cartesian coordinates for n atoms. We assume that a mapping function determines a configuration, $\mathbf{R}$, for the CG model as a linear function of the AA configuration, $\mathbf{r}$. The appropriate potential for a CG model that quantitatively reproduces all structural properties of the AA model (at the resolution of the CG model) is the many-body potential of mean force (PMF), $U^0(\mathbf{R})$:

$$U^0(\mathbf{R}) = -k_B T \ln p_R(\mathbf{R}) + \text{const}, \tag{5.11}$$

where $p_R(\mathbf{R})$ is the probability for the AA model to sample a configuration $\mathbf{r}$ that maps to the CG configuration $\mathbf{R}$.[23] The forces, $\mathbf{F}_I^0(\mathbf{R})$, that are derived from the PMF define the many-body mean force (MF) field, $\mathbf{F}^0$, which is the appropriate force field for a CG model that samples configurations according to $p_R(\mathbf{R})$.[41]

In general, the many-body MF, $\mathbf{F}^0$, does not correspond to any element in ϕ-space. Instead, the MS-CG method determines the point $\boldsymbol{\phi}^0$ in ϕ-space that provides an optimal approximation to $\mathbf{F}^0$. According to the MS-CG objective function,[36,41,59,60] this optimal approximation is determined by directly inverting the normal system[62] of linear equations:

$$\boldsymbol{b}^{AA} = \mathbf{G}^{AA}\boldsymbol{\phi}^0, \tag{5.12}$$

where $\boldsymbol{b}^{AA}$ and $\mathbf{G}^{AA}$ are defined in analogy to $\boldsymbol{b}(\boldsymbol{\phi})$ and $\mathbf{G}(\boldsymbol{\phi})$ in the g-YBG Equation (5.6). Despite their similarities, Equation (5.12) is significantly different from Equation (5.6), be-

cause $\boldsymbol{b}^{AA}$ and $\mathbf{G}^{AA}$ are ensemble averages weighted by the AA distribution, $p_R(\mathbf{R})$, and not by the CG distribution, $P_R(\mathbf{R}|\boldsymbol{\phi})$.

Figure 5.1b illustrates the MS-CG method in the context of ϕ-space and b-space. The matrix inverse $(\mathbf{G}^{AA})^{-1}$ determines the MS-CG force field $\boldsymbol{\phi}^0$ from $\boldsymbol{b}^{AA}$. The g-YBG equation, Equation (5.6), then determines the point $\boldsymbol{b}^0 = \boldsymbol{b}(\boldsymbol{\phi}^0)$ in b-space (and also the corresponding rdfs) that is generated by simulations of the MS-CG model with parameters $\boldsymbol{\phi}^0$:

$$\boldsymbol{b}^0 = \mathbf{G}(\boldsymbol{\phi}^0)\boldsymbol{\phi}^0, \tag{5.13}$$

where $\mathbf{G}(\boldsymbol{\phi}^0)$ is the g-YBG matrix of equilibrium cross-correlations sampled by the MS-CG model. If $\mathbf{G}(\boldsymbol{\phi}^0) = \mathbf{G}^{AA}$, then $\boldsymbol{b}^0 = \boldsymbol{b}^{AA}$ and the CG model will reproduce the AA rdfs. In some sense, this is the underlying approximation of the MS-CG method. However, this approximation is not necessarily accurate, because it would require that the CG model reproduce not only the AA rdfs, but also the higher order cross-correlations of the AA model. Consequently, the force field parameters, $\boldsymbol{\phi}^0$, are not guaranteed to reproduce the AA rdfs, as has been extensively discussed before.[230]

5.2.4 Iterative Procedures

Iterative bottom-up CG procedures seek to determine a force field $\boldsymbol{\phi}^*$ in ϕ-space that will reproduce the 1-D equilibrium distributions implied by an AA model for the relevant degrees of freedom in the CG model. In the context of the g-YBG framework, this corresponds to determining the force field coefficients $\boldsymbol{\phi}^*$ such that

$$\mathbf{G}(\boldsymbol{\phi}^*)\boldsymbol{\phi}^* = \boldsymbol{b}^{AA}. \tag{5.14}$$

Note that this equation is significantly different from the MS-CG equation, Equation (5.12). The MS-CG equation determines the force field coefficients, $\boldsymbol{\phi}^0$, that reproduce the force projections, $\boldsymbol{b}^{AA}$, of the AA model, while using the cross-correlations, $\mathbf{G}^{AA}$, of the AA model. In contrast, Equation (5.14) corresponds to a g-YBG equation that determines the force field coefficients, $\boldsymbol{\phi}^*$, that reproduce $\boldsymbol{b}^{AA}$, while using the cross-correlations $\mathbf{G}(\boldsymbol{\phi}^*)$ sampled by a CG model with the corresponding potential $U(\mathbf{R};\boldsymbol{\phi}^*)$. Equation (5.14) is the basic self-consistency criterion for the iterative g-YBG method proposed by Cho and Chu.[69]

Figure 5.1b schematically illustrates the resulting iterative g-YBG procedure. First, the MS-CG method applies $(\mathbf{G}^{AA})^{-1}$ to map (red curve) the point $\boldsymbol{b}^{AA}$ to the MS-CG force field $\boldsymbol{\phi}^0$. This force field is then mapped (via molecular simulation) to a point, $\boldsymbol{b}^0$, in b-space. In

general, $\boldsymbol{b}^0 \neq \boldsymbol{b}^{AA}$ since the cross-correlations, $\mathbf{G}^0 = \mathbf{G}(\boldsymbol{\phi}^0)$, that are generated by the MS-CG model will not exactly reproduce the AA cross-correlations, $\mathbf{G}^{AA}$. The cross-correlations, $\mathbf{G}^0$, define a new inverse map (magenta curve) that can be applied to $\boldsymbol{b}^{AA}$ to obtain a new point in ϕ-space, $\boldsymbol{\phi}^1$. Simulations with this force field then determine a new set of cross-correlations, $\mathbf{G}^1 = \mathbf{G}(\boldsymbol{\phi}^1)$, and a new point in b-space, $\boldsymbol{b}^1 = \boldsymbol{b}(\boldsymbol{\phi}^1)$, which, in turn, determine a new force field, $\boldsymbol{\phi}^2 = (\mathbf{G}^1)^{-1}\,\boldsymbol{b}^1$. The iterative g-YBG procedure repeats this process until a self-consistent solution, $\boldsymbol{\phi}^*$, is obtained. At this point, simulation with the force field $\boldsymbol{\phi}^*$ maps to $\boldsymbol{b}^{AA}$ and simultaneously generates a set of correlations $\mathbf{G}(\boldsymbol{\phi}^*)$ that define a reverse map from $\boldsymbol{b}^{AA}$ to the point $\boldsymbol{\phi}^*$. Note that this implies $\mathbf{G}(\boldsymbol{\phi}^*) \neq \mathbf{G}^{AA}$, i.e., the final model reproduces 1-D correlations of the AA model (e.g., rdfs) at the expense of modifying higher order cross-correlations. At each iteration, the g-YBG equation is applied to estimate force field coefficients that reproduce $\boldsymbol{b}^{AA}$, while using cross-correlations that are increasingly more consistent with the final CG model.

If we assume that the MS-CG parameters $\boldsymbol{\phi}^0$ are sufficiently close to the desired self-consistent parameters, $\boldsymbol{\phi}^*$, then we may employ the local linearity argument, Equation (5.10), to estimate the error, $\delta\boldsymbol{\phi} = \boldsymbol{\phi}^* - \boldsymbol{\phi}^0$, in ϕ-space based upon the corresponding error, $\delta\boldsymbol{b}^0 = \boldsymbol{b}^{AA} - \boldsymbol{b}^0$, in b-space. Moreover, in the spirit of the fundamental MS-CG assumption, Lu et al.[75] further assumed that $\mathbf{G}(\boldsymbol{\phi}^0) \approx \mathbf{G}^{AA}$ and estimated the force-field correction:

$$\delta\boldsymbol{\phi}^0 \approx (\mathbf{G}^{AA})^{-1}\delta\boldsymbol{b}^0. \tag{5.15}$$

Lu et al. proposed[75] using this estimate $\delta\boldsymbol{\phi}^0$ to determine a new set of force field parameters, $\boldsymbol{\phi}^1 = \boldsymbol{\phi}^0 + \delta\boldsymbol{\phi}^0$. Simulations with $\mathbf{F}(\boldsymbol{\phi}^1)$ determine a new point, $\boldsymbol{b}^1 = \boldsymbol{b}(\boldsymbol{\phi}^1)$, in b-space and a corresponding error $\delta\boldsymbol{b}^1 = \boldsymbol{b}^{AA} - \boldsymbol{b}^1$. They then iterated this procedure to convergence, as illustrated in Figure 5.1c. Note that, in contrast to the iterative g-YBG method of Cho and Chu,[69] this procedure employs the cross-correlation matrix, $\mathbf{G}^{AA}$, of the AA model at each refinement step, but directly considers the error $\delta\boldsymbol{b}$ in b-space.

This latter treatment is both elegant and also computationally efficient, since it avoids the necessity of accurately determining $\mathbf{G}(\boldsymbol{\phi})$ for each set of force field parameters, $\boldsymbol{\phi}$. This approximation is particularly likely to be successful for cases where the error, $\delta\boldsymbol{\phi}$, in the initial MS-CG model is sufficiently small such that the locally linear approximation holds. However, this approach may possibly prove less successful for cases in which $\mathbf{G}(\boldsymbol{\phi})$ is sensitive to specific interactions or for cases that $\delta\boldsymbol{\phi}$ is too large. For these reasons, we revisit the original proposal of Cho and Chu[69] and attempt to iteratively solve the self-consistent g-YBG equation, Equation (5.14), without assuming local linearity.

5.3 Methods

The present section briefly outlines the key details of our calculations. The Supporting Information section provides a much more detailed description.

5.3.1 Simulation Details

We performed molecular dynamics (MD) simulations with all-atom (AA) models for both hexane and 3-hexylthiophene (3HT), while using the OPLS-AA force field[18] to model all interactions. Additionally, we performed MD simulations with multiple CG models for each system. The representations and potentials used for these CG simulations are described below. All simulations were performed with the Gromacs 4.5.3 simulation suite[7] at a temperature of 298 K according to standard procedures.[45,168–172]

We performed a 100 ns production simulation of the AA hexane model that sampled the constant NVT ensemble for 267 molecules in a cubic box of volume $V = (3.89 \text{ nm})^3$ with periodic boundary conditions. This simulated density was determined as the average volume in a 10 ns simulation that sampled the constant NPT ensemble with $P = 1$ bar. This density is within 2% of the experimentally measured density.[240] The initial configuration for the NVT simulation was sampled from the NPT simulation. After the first 10 ns, we sampled configurations every 1 ps during the remainder of the production NVT simulation. Following similar protocols, we also performed a 60 ns production simulation of the AA 3HT model that sampled the constant NVT ensemble for 800 molecules in a cubic box of volume $V = (6.18 \text{ nm})^3$. This density is also within 2% of the experimentally measured density.[241]

Production simulations of CG models for hexane and 3HT were performed for 11 ns and 6 ns, respectively, in the constant NVT ensemble. In each case, we obtained a starting configuration by mapping a configuration from the AA simulation and then performing an energy minimization with the CG potential. After the first 1 ns, we sampled configurations every 0.5 ps during the remainder of each simulation.

5.3.2 CG Mapping and Interactions

We considered 3- and 4-site CG representations of hexane (Figures 5.2a1 and b1, respectively), as well as 5- and 6-site CG representations of 3HT (Figures S7 and 5.6a, respectively). As illustrated in these figures, each mapping partitions the molecule into disjoint atomic groups and associates a CG site with the mass center for each atomic group that included at least one heavy atom (i.e,. an atom other than hydrogen.)

For each CG model, the potential energy function assumed a molecular mechanics form with separate intramolecular and intermolecular potentials. The intermolecular potential included central pair potentials between each pair of sites on different molecules. A distinct potential was determined for each unique pair of site types.

The intramolecular potential included bond potentials for each pair of "bonded" sites, i.e., each pair of sites for which at least one bond connected the corresponding atomic groups. In the case of 3- and 4-site hexane models, the intramolecular potential also included angle and dihedral potentials between each set of 3 and 4 consecutive bonded sites. The intramolecular potential for the 6-site 3HT models also included 1) bond potentials between CR2 and CR3 sites to stabilize the ring, 2) angle potentials for the CT-CM-CR1 and CM-CR1-CR2 triples, and 3) dihedral potentials for the CM-CR1-CR3-S and CR1-CR2-S-CR3 quadruples. The intramolecular potentials did not include "nonbonded" pair potentials between sites in the same molecule.

5.3.3 Molecular State Analysis

To characterize the molecular conformations for each AA model, we first analyzed the distributions sampled by the corresponding AA simulation for all bond, angle, and dihedral degrees of freedom (dofs). We then identified those dofs that sampled multimodal distributions. In the case of the AA models, only dihedral angles sampled multimodal distributions. We defined discrete states for each of these dofs by choosing a strict dividing surface at each minima between peaks in the corresponding distribution. The set of molecular conformations, which we shall refer to as "molecular states," were determined by enumerating all possible combinations of the discrete states for the relevant dof. We emphasize that these AA molecular states are completely independent of the CG mapping.

In order to identify the relevant molecular states for each CG model, we first generated a "mapped ensemble" of CG configurations by applying the corresponding CG mapping to each configuration that was sampled by the AA trajectory. Given this mapped ensemble, we calculated the distribution sampled along each intramolecular dof of the CG model that is governed by a term in the CG potential. We then defined discrete states for each of these dofs in the CG model that sampled a multimodal distribution in the mapped ensemble. Finally, given these discrete states for CG dofs, we defined CG molecular states in the same way as described above for the AA model.

We calculated, for each AA molecular state, the conditional probability that it was mapped to each CG molecular state. We have plotted these conditional probabilities, $p(y|x)$,

as a function of the AA (x) and CG (y) molecular states using gnuplot.[206] We shall refer to these figures as molecular state mappings or simply state mappings.

5.3.4 Reexamination of the Iterative G-YBG (iter-gYBG) Method

As discussed above, Cho and Chu[69] developed an iterative g-YBG (iter-gYBG) method for parameterizing nonbonded pair potentials. They employed an independent "fluctuation matching" method[242] to optimize the intramolecular potentials, but only considered CG models with fairly limited intramolecular flexibility. In the present work, we initially attempted to apply their iter-gYBG method to parameterize both intra- and intermolecular potentials for molecules and CG models of considerably greater complexity. However, our preliminary calculations yielded unreasonable results for force field parameters that govern rarely sampled bond lengths and bond angles, which only weakly couple to other interactions. Accordingly, we have adopted a heuristic modification to the iter-gYBG procedure.

We have previously discussed[230] the decomposition of the g-YBG correlation matrix into direct and indirect contributions: $G_{D,D'} = \bar{g}_D \delta_{D,D'} + \bar{G}_{D,D'}$. The direct term, $\bar{g}_D$, is an ensemble average that reflects a single interaction due to the force field coefficient ϕ_D, while the indirect term, $\bar{G}_{D,D'}$, characterizes the cross-correlations between two interactions that are governed by coefficients ϕ_D and $\phi_{D'}$. Given this decomposition, Equation (5.14) can be expressed:

$$b_D^{AA} = \bar{g}_D^{CG} \phi_D + \sum_{D'} \bar{G}_{D,D'}^{CG} \phi_{D'}, \tag{5.16}$$

for each coefficient ϕ_D. As in Equation (5.14), $\bar{g}_D^{CG}$ and $\bar{G}_{D,D'}^{CG}$ are ensemble averages sampled by a CG model with force field coefficients $\boldsymbol{\phi} = \{\phi_D\}$. For each coefficient ϕ_D corresponding to a bond or angle potential, we adopted the following modification to Equation (5.16):

$$b_D^{AA} = \bar{g}_D^{AA} \phi_D + \sum_{D'} \bar{G}_{D,D'}^{CG} \phi_{D'}, \tag{5.17}$$

where $\bar{g}_D^{AA}$ is the corresponding ensemble average that is determined by the AA model and the CG mapping. Note that Equation (5.17) employs cross-correlations, $\bar{G}_{D,D'}^{CG}$, that are sampled by the CG model, which may be inconsistent with the ensemble averages $\bar{g}_D^{AA}$ that are determined from the AA model. Consequently, initial iterations may drastically over- or under-estimate the necessary changes to the force coefficients, especially for interaction coefficients that are strongly coupled. Accordingly, we solve Equation (5.17) only for bond and angle coefficients, which are more likely to be weakly coupled to other interactions. We

solve Equation (5.16) for all other interaction coefficients. To further stabilize the procedure, we employ a decelerating coefficient, λ, which scales the change in the force for a given iteration in a manner similar to the iterative Boltzmann inversion method.[35]

We adopted the following work flow for the new iter-gYBG procedure:

1. Compute $\boldsymbol{b}^{AA}$, $\bar{\mathbf{g}}^{AA}$, and $\bar{\mathbf{G}}^{AA}$ from the AA trajectory. In particular, we calculated $\boldsymbol{b}^{AA}$ from AA forces according to the MS-CG framework,[60,62,72] although it may also be calculated from structures according to the g-YBG approach.[63,64,134,188]

2. Solve Equation (5.12) to obtain the MS-CG force field, $\boldsymbol{\phi}^0$.

3. Simulate this MS-CG model to determine the relevant equilibrium ensemble averages: $\boldsymbol{\phi}^0 \rightarrow \bar{\mathbf{g}}^0, \bar{\mathbf{G}}^0$

4. Solve the "modified" g-YBG equations (Equation (5.17) for bond and angle coefficients, Equation (5.16) for all other coefficients) with the ensemble averages that are calculated in step 3 to obtain a new CG force field, $\boldsymbol{\phi}^1$.

5. Update the force field: $\tilde{\boldsymbol{\phi}}^1 = \boldsymbol{\phi}^0 + \lambda(\boldsymbol{\phi}^1 - \boldsymbol{\phi}^0)$, where $0 \leq \lambda \leq 1$. As described in the Supporting Information, λ is systematically determined at each step according to changes in the force fields coefficients.

6. Simulate the CG model with the new force field to determine the corresponding ensemble averages: $\tilde{\boldsymbol{\phi}}^1 \rightarrow \bar{\mathbf{g}}^1, \bar{\mathbf{G}}^1$.

7. Repeat steps 3-5 until $\boldsymbol{\phi}^{i+1} \approx \boldsymbol{\phi}^i$ (with $\lambda \approx 1$) and $\boldsymbol{b}^i \approx \boldsymbol{b}^{AA}$.

As described in the Supporting Information, the modified procedure is motivated by considering the iter-gYBG framework for weakly coupled interactions. Moreover, although heuristic, Equation (5.17) is generally consistent with the original iter-gYBG procedure. In particular, consider a subset of force field coefficients, $\{\phi_D\}$, that correspond to the magnitude of a central pair potential or bond potential at a set of distances, $\{r_D\}$. Then, for these coefficients, $\{D\}$, the direct contributions to the g-YBG correlation matrix, $\{\bar{g}_D\}$, can be determined immediately from the corresponding set of force projections $\{b_D\}$.[188] Consequently, since $\boldsymbol{b}_D(\boldsymbol{\phi}^*) = \boldsymbol{b}_D^{AA}$ for a model with the self-consistent coefficients, $\boldsymbol{\phi}^*$, it also follows that $\bar{g}_D(\boldsymbol{\phi}^*) = \bar{g}_D^{AA}$ for the self-consistent model. Thus, for these interactions, Equation (5.17) becomes equivalent to Equation (5.14) for the self-consistent model. We note that the relationship between $\bar{g}_D$ and b_D is not so simple for a potential that is a

function of an angle. Consequently, it is not obvious that the same reasoning applies for angle interactions.

This iterative g-YBG procedure appears numerically robust and determines accurate models for many systems. In particular, for a wide range of test cases, the procedure converges to a single accurate CG model. However, for the models presented in this manuscript, the numerical procedure did not perfectly converge to a model satisfying the self-consistent g-YBG equation, Equation (5.14). Instead, after initially converging upon an accurate model, the method began to diverge. Consequently, we selected an optimal iter-gYBG model based upon two criteria: 1) the model parameters remained stable for at least two iterations of the procedure; 2) the model accurately reproduced the AA force projections. (See Figures S1, S2, and S8.) It is possible that the method would ultimately have converged upon an even better model, but we did not exhaustively explore this possibility. The Supporting Information provides much more information regarding the convergence of the method and model selection.

5.3.5 Force Field Calculations

We applied the MS-CG and iter-gYBG methods to determine potentials for each CG representation of hexane or 3HT. For each term in the CG potential, the corresponding force function was represented by a discrete set of basis functions of a single variable.[60] Different basis functions (e.g., linear spline, cubic B-spline, etc) were employed for different interaction types. We modified the g-YBG system of linear algebraic equations in order to automate the iterative procedure, increase its robustness, and minimize user interference. As described in the Supporting Information section, these modifications included removing force field coefficients for interactions that were rarely sampled,[60] introducing constraints to ensure periodicity of dihedral potentials, and regularizing nonbonded interactions to avoid over-fitting statistical noise. We extensively tested these modifications to ensure that they minimally impacted the MS-CG calculation. We then solved the modified linear equations[207] via singular value decomposition[175] after applying right-left preconditioning[38,122] to render the linear equations dimensionless.

5.4 Results

In this study, we performed AA simulations of hexane and 3-hexylthiophene (3HT). For each of these systems, we considered several CG mappings and analyzed the corresponding mapping of molecular states. We derived MS-CG and iter-gYBG models for each mapping and then assessed the quality of these models by considering the resulting intra- and intermolecular structure. Finally, by comparing the results for various CG models, we investigated the relationship between the mapping and structural fidelity of CG models.

5.4.1 Hexane

5.4.1.1 Molecular State Analysis

Prior to calculating any CG force field, we first analyzed the molecular states sampled by the AA model for hexane. In the AA model, all bonds and angles fluctuate about well defined equilibria, while each of the three dihedral angles along the backbone samples both trans (T) and gauche (G) states. Accordingly, the molecular state of an AA hexane molecule is specified by a three letter code describing the state of each dihedral. Table S1 presents the probabilities for hexane molecules to sample these six molecular states during the AA simulation. We next considered how these AA molecular states are mapped to CG molecular states by various CG mappings. For each mapping of interest, we determined a "mapped ensemble" of CG configurations by applying the mapping to the ensemble of configurations sampled by the AA trajectory. We defined discrete states for each intramolecular dof in the CG model that sampled a multimodal distribution in this mapped ensemble. Finally, we defined CG molecular states based upon combinations of the discrete states for multimodal dofs. For simplicity, we will often refer to molecular states as simply "states."

Figure 5.2a analyzes the molecular state mapping for the 3-site mapping (CT-CM-CT) that is illustrated in Figure 5.2a1. Figure 5.2a2 presents the distributions for the two equivalent CT-CM bonds and the CT-CM-CT angle that are obtained by mapping the AA trajectory to the 3-site CG representation. We define small (s) and large (l) states for each bond, as well as small (s), medium (m), and large (l) states for the angle by dividing the mapped distributions into distinct states at each minima. The set of CG molecular states is then defined by a three letter code that describes the individual states of the bonds and the angle. Given the symmetry of the molecule, this CG mapping yields nine CG states. Figure 5.2a3 graphically demonstrates the mapping of states from the AA to the CG representation. The height of the bar graph at (x, y) is proportional to the conditional probability,

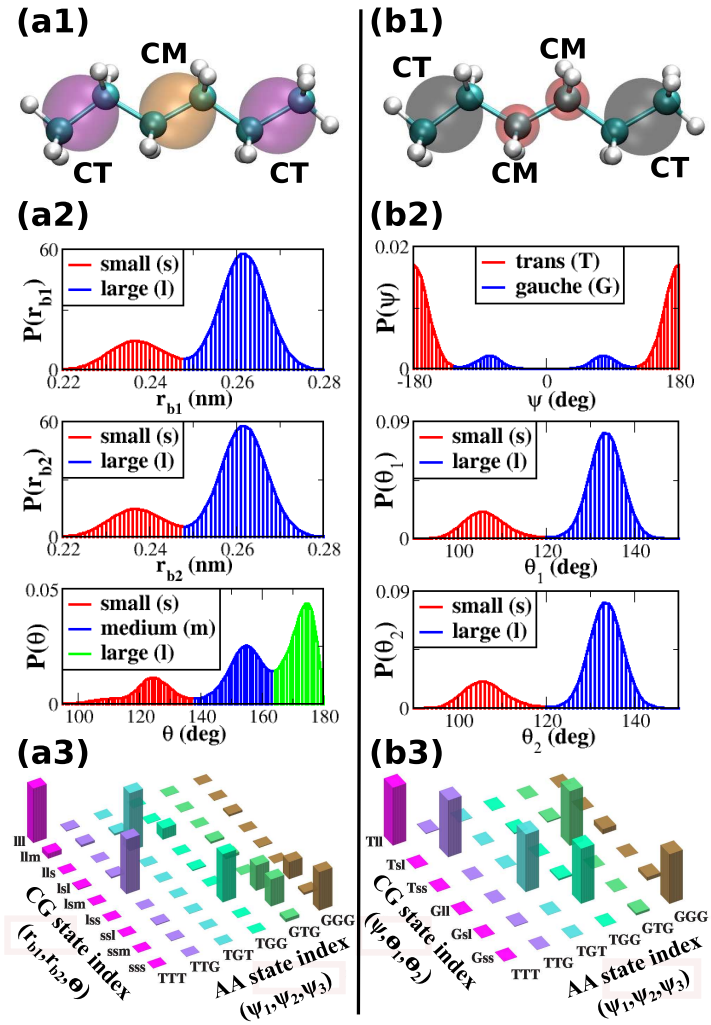

Figure 5.2. Analysis of the molecular states for 3-site (column a) and 4-site (column b) mappings of hexane. In each column, panel 1 illustrates the corresponding CG mapping, panel 2 identifies the subpopulations defined by the 1-D distributions that are sampled by the AA model along relevant CG dofs, and panel 3 analyzes the mappings of molecular states. In panel 3, the height at location (x,y) indicates the conditional probability, $p(y|x)$, that AA molecular state x is mapped to CG molecular state y.

$p(y|x)$, that AA state x is mapped to CG state y by the 3-site mapping. In this case, some AA states (e.g., TTT) are largely mapped to a single CG state, while other AA states (e.g., GGG) are "split between" (i.e., mapped to) multiple CG states. Additionally, this representation introduces several "forbidden" states (e.g., lss) into the CG model that are not sampled by the AA model. Since all CG molecular states are defined in terms of discrete states of individual dofs that are accessible to the CG model, these forbidden states reflect cross-correlations between CG dofs that emerge in the mapped ensemble as a consequence of eliminating particular details from the AA model.

Figure 5.2b presents a corresponding analysis of the mapping of states for the 4-site CG representation that is illustrated in Figure 5.2b1. The mapped bond distributions are all unimodal for this 4-site representation and thus unimportant for the molecular state analysis. Figure 5.2b2 demonstrates that the mapped angle (θ) and dihedral (ψ) distributions are bimodal. (Symmetric peaks in the dihedral distribution are indistinguishable and, thus, counted as a single state.) We divided the two angle distributions into small (s) and large (l) states. We divided the CG CT-CM-CM-CT dihedral distribution into T and G states. Figure 5.2b3 presents the molecular state mapping for this 4-site representation. The x-axis again indicates the six AA states, while the y-axis indicates the six CG states. In this case, the correspondence between AA and CG states is nearly one-to-one (1-1).

Intuitively, we expect that a "good" CG map should preserve a 1-1 correspondence between AA and CG states, as is observed for the 4-site mapping of hexane. A 1-1 mapping of molecular states provides an inherently simpler relationship between the intramolecular dofs and the molecular states of the CG model. Given such a "good" mapping, we expect that, if the CG model reproduces the mapped 1-D distributions for these dofs, then it will also reproduce the mapped distribution of states. Conversely, if the mapping introduces forbidden CG states, as in Figure 5.2a3 for the 3-site representation, then we expect that the CG model will be unlikely to reproduce the mapped distribution of molecular states. We expect that, unless the CG potential is sufficiently complex to capture the cross-correlations of the AA model and prohibit the forbidden CG states, the CG model will likely reproduce the AA 1-D distributions by sampling both allowed and forbidden states.

We have also considered the 2-site and 4-site mappings proposed by Das et al.[73] These mappings both yield only two CG states and, therefore, lose information regarding the AA states. Consequently, we have focused on the 3- and 4-site mappings in Figure 5.2, which distinguish between all six AA molecular states.

5.4.1.2 Model Assessment

Figures 5.3 and 5.4 characterize the equilibrium structure generated by several CG models of hexane. In both figures, panels a and b present results for the 3- and 4-site models, respectively. In all panels, the solid black curves present results for the AA model, which are determined by mapping the AA trajectory to the corresponding CG representation; the dashed red curves present results for the MS-CG models, which are parametrized directly (i.e., non-iteratively) from the AA simulation according to Equation (5.12); and the dashed-dotted green curves present results for the iter-gYBG models, which are determined from the iterative procedure described in the Theory and Methods sections.

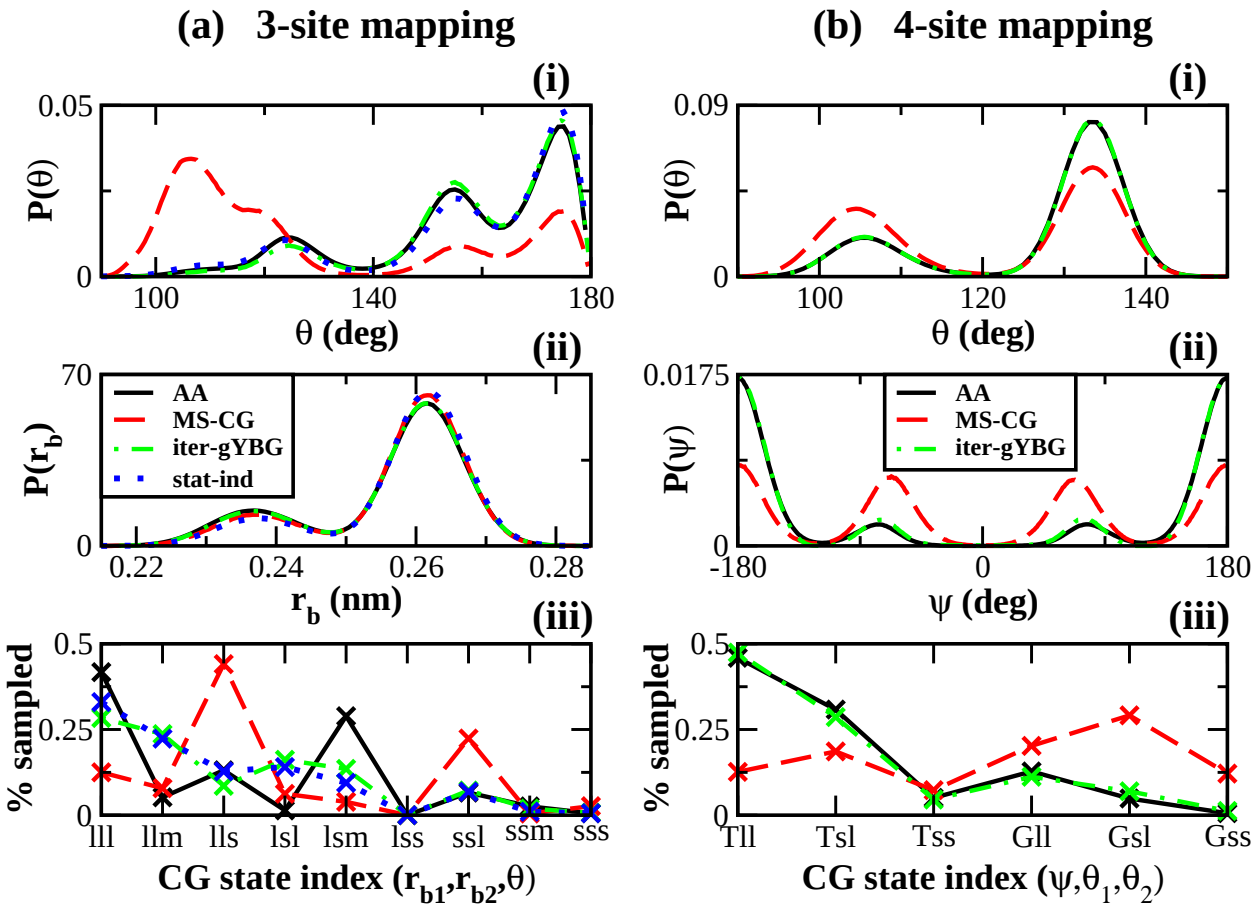

Figure 5.3. Accuracy of the intramolecular structure generated by 3-site (column a) and 4-site (column b) CG models for hexane. In each panel, solid black, dashed red, and dashed-dotted green curves correspond to the AA, MS-CG, and iter-gYBG models, respectively. The blue curves in panel a correspond to a mixed model that was calculated with simplified cross-correlations between bond and angle dofs. Panels a(i) and a(ii) present angle and bond distributions for 3-site models, while b(i) and b(ii) present angle and torsion distributions for 4-site models. Panels a(iii) and b(iii) present the sampled distributions of molecular states for 3- and 4-site models, respectively.

Panels a(i) and a(ii) of Figure 5.3 present 1-D distribution functions for the angle and bond dof, respectively, in the 3-site representation, while Figure 5.4a presents the corresponding rdfs. For this representation, the MS-CG model nearly quantitatively reproduces the bond distribution and also the three rdfs, but does not even qualitatively reproduce the angle distribution. The iter-gYBG model nearly quantitatively reproduces the angle distribution, while also improving the other distributions to nearly quantitative accuracy.

Panel a(iii) presents the sampled distributions for the CG states that are defined in Figure 5.2a2 for the 3-site representation. Although the iter-gYBG model performs somewhat better than the MS-CG model, neither model accurately reproduces the distribution of CG states that is sampled in the mapped AA ensemble. It appears that, in order to sample large

angles ($\theta \approx 175°$) with the correct weight, the iter-gYBG model significantly undersamples the lll state, which is the dominant state in the AA model, and significantly oversamples the forbidden lsl state. As expected, because the 3-site mapping introduces forbidden CG states, the accurate reproduction of the mapped 1-D distributions does not ensure that the CG model even qualitatively reproduces the mapped distribution of molecular states. Interestingly, the MS-CG model, which poorly models the angle distribution but does incorporate the relevant cross-correlations in parameterizing the CG potential, more accurately samples the forbidden molecular states (e.g., llm and lsl).

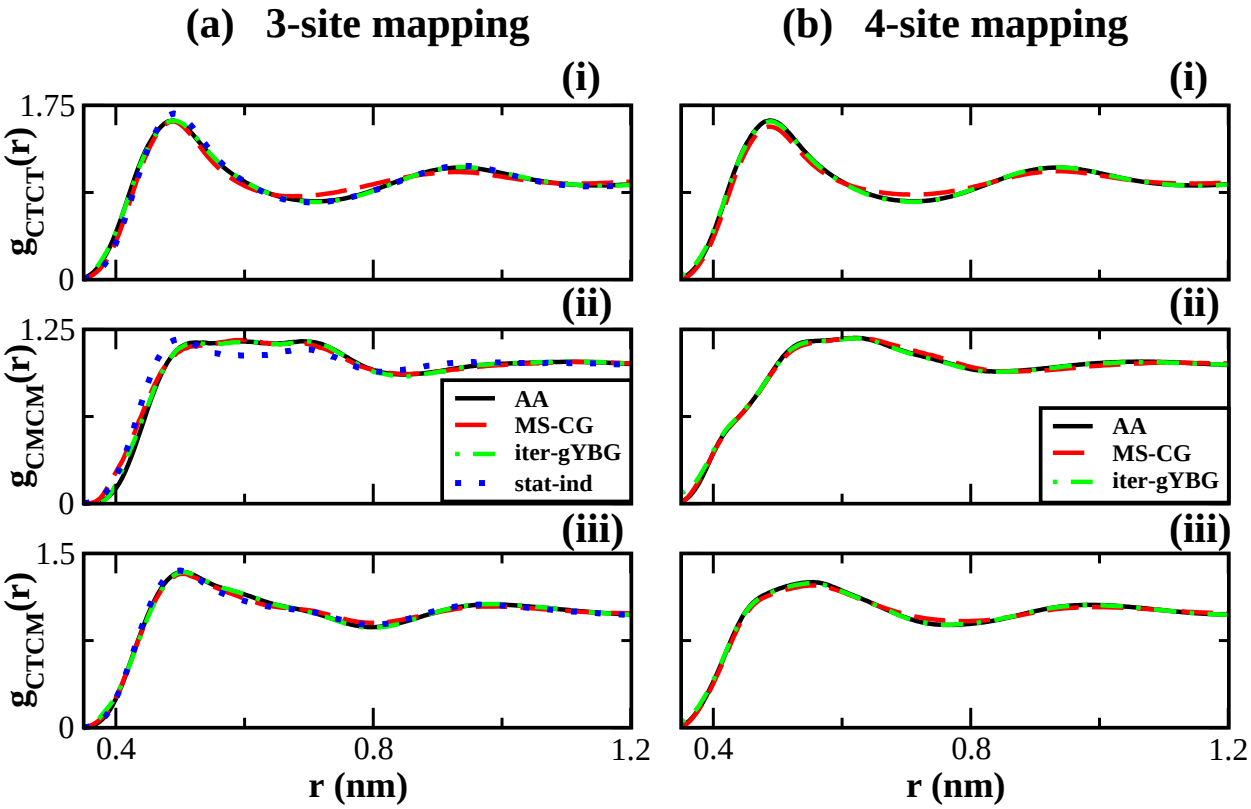

Figure 5.4. Accuracy of the intermolecular structure generated by 3-site (column a) and 4-site (column b) CG models for hexane. In each panel, solid black, dashed red, and dashed-dotted green curves correspond to the AA, MS-CG, and iter-gYBG models, respectively. The blue curves in panel a correspond to a mixed model that was calculated with simplified cross-correlations between bond and angle dofs. Panels (i), (ii), and (iii) present CTCT, CMCM, and CTCM rdfs, respectively, for the corresponding representation.

Panels b(i) and b(ii) of Figure 5.3 present 1-D distribution functions for the angle and dihedral dof, respectively, in the 4-site representation, while Figure 5.4b presents the corresponding rdfs. In comparison to the 3-site MS-CG model, the 4-site MS-CG model more accurately reproduces the mapped 1-D distributions for intramolecular dof. The 4-site MS-CG model reproduces the peak positions, but not peak magnitudes, for the AA angle and

dihedral distributions. Moreover, the MS-CG model reproduces the bond distribution and rdfs with near quantitative accuracy. The iter-gYBG model quantitatively reproduces the AA angle distribution, nearly quantitatively reproduces the AA dihedral distribution, and also improves the other distributions to quantitative accuracy.

Figure 5.3b(iii) presents the sampled distributions for the CG states that are defined in Figure 5.2b for the 4-site representation. The MS-CG model poorly reproduces the distribution of CG states. In contrast, the iter-gYBG model reproduces the distribution of CG states with nearly quantitative accuracy. Because the 4-site mapping provides a 1-1 relation between the AA and CG molecular states, the accurate reproduction of AA intramolecular distributions ensures that the iter-gYBG model also reproduces the distribution of mapped AA states.

5.4.1.3 Bond-Angle Correlation Analysis

The MS-CG equations, Equation (5.12), determine force field coefficients by considering the matrix $\bar{\mathbf{G}}^{AA}$, which describes the cross-correlations generated by mapping the AA model to the CG representation. As described above, the MS-CG method assumes that these AA cross-correlations provide a reasonable approximation to the cross-correlations generated by the resulting MS-CG model. The disagreement observed in Figure 5.3a1 between the angle distributions generated by the AA and MS-CG models, however, suggest that this approximation breaks down for the 3-site hexane model. Consequently, we analyzed the matrix block of $\bar{\mathbf{G}}^{AA}$ that describes the coupling between the angle, θ, and bond, r_b, dofs. This block can be represented[230] as a function, $\bar{G}(\theta, r_b)$, that quantifies the cross-correlations between the forces generated along the angle and bond dofs as a function of θ and r_b. Figure 5.5 presents intensity plots of $\bar{G}(\theta, r_b)$ for various models. In each panel, white regions indicate combinations (θ, r_b) that are not sampled, while blue and green regions indicate combinations with negative cross-correlations of increasing magnitude.

Figure 5.5a demonstrates the complex bond (r_b) - angle (θ) cross-correlations that are generated by mapping the AA model to the 3-site representation. In particular, in the notation of Figure 5.3, the AA model completely excludes conformations with the large angle/small bond combination, which correspond to $\theta \approx 3.05$ rad and $r_b \approx 0.237$ nm. Figures 5.5b and 5.5c present the bond-angle cross-correlations that are generated by the MS-CG and iter-gYBG models, respectively. Because the intramolecular CG potentials do not directly couple the bond and angle dofs, neither CG model reproduces the complex cross-correlations of the AA model. Interestingly, the MS-CG model, which is parameterized on the basis of the AA cross-correlations, successfully excludes these large angle/small bond

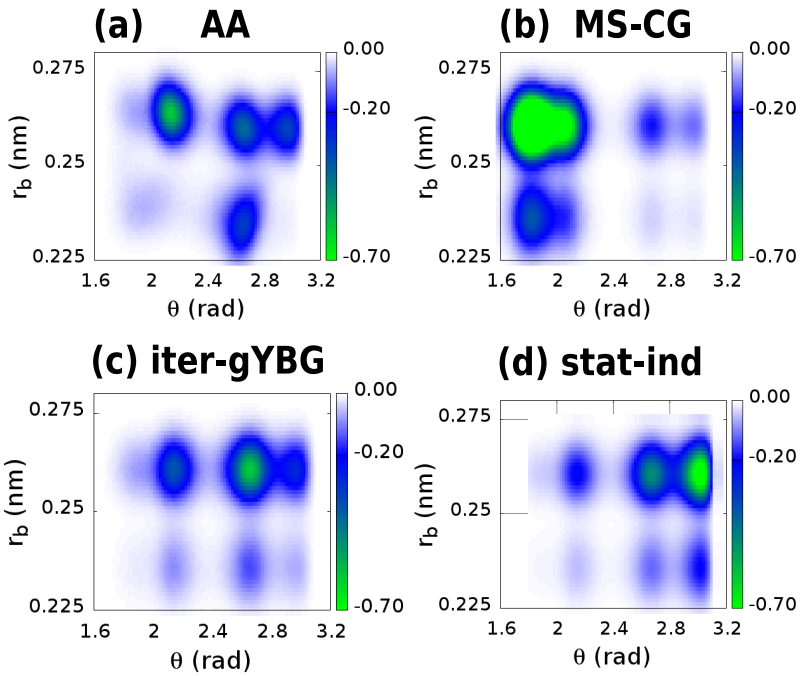

Figure 5.5. Intensity plots of the g-YBG cross-correlation matrix $\bar{G}(\theta, r_b)$, which describes cross-correlations between the angle, θ, and bond, r_b, dofs of the 3-site CG hexane model. Panels a, b, and c present cross-correlation matrices sampled by AA, MS-CG, and iter-gYBG models, respectively. Panel d presents an artificially constructed cross-correlation matrix that assumed statistical independence of θ and r_b.

conformations, but does so at the expense of over stabilizing small angle conformations. Conversely, the iter-gYBG model accurately reproduces the angle distribution of the AA model, but does so at the expense of significantly distorting the AA cross-correlations. In particular, the iter-gYBG model samples the large angle/small bond conformations that are excluded by the AA model.

Next we tested whether the use of simplified bond-angle cross-correlations (i.e., cross-correlations that are more consistent with the resulting CG model) in the MS-CG equations, Equation (5.12), would determine a force field that more accurately modeled the AA angle distribution. Accordingly, we replaced the bond-angle block of $\bar{\mathbf{G}}^{AA}$ in Equation (5.12), which corresponds to Figure 5.5a, with a matrix that was constructed by assuming that r_b and θ were statistically independent. Figure 5.5d presents an intensity map for this artificially constructed block. We then determined a new CG potential by solving the MS-CG equations, Equation (5.12), after performing this replacement. The dotted blue curves in Figures 5.3a and 5.4a present the distributions sampled by the resulting CG model. Remarkably, this CG model, which is parameterized on the basis of the simplified cross-correlations that can actually be generated by the CG model, nearly quantitatively reproduces the AA angle

distribution, although this procedure introduces some small errors into the rdfs. We note that, while Figure 5.5d indicates the cross-correlations that were employed in calculating the CG potential for this "mixed" model, simulations with this mixed model generated a set of cross-correlations that were very similar to the cross-correlations generated by the iter-gYBG model, which are presented in Figure 5.5c.

5.4.2 3HT

5.4.2.1 Molecular State Analysis

As described above for hexane, we first analyzed the molecular states sampled in AA simulations of 3-hexylthiophene (3HT). Figure 5.6a presents the atomic structure of 3HT. Because the thiophene ring is relatively rigid, the AA molecular states are determined by the five C-C-C-C dihedral angles along the backbone of the hexyl side chain. The four dihedrals that are farthest from the ring sample both trans (T) and gauche (G) states. The fifth dihedral, which connects to the ring, determines the orientation of the chain with respect to the ring and samples gauche (G) and cis (C) states. Therefore, the AA molecular states for 3HT are determined by a five letter code that describes the state of each dihedral. The resulting set of 32 AA states are naturally partitioned into 8 groups, as indicated below. Table S1 presents the probability for the AA model to sample each of these groups.

In the following, we consider two distinct 6-site representations of 3HT, which are presented in panels a1 and a2 of Figure 5.6. Both representations represent the thiophene ring with 4 sites and the hexyl side chain with 2 sites. Additionally, both representations associate an S site with the sulfur atom and associate CR2 and CR3 sites with the carbon atoms adjacent to the sulfur. The two representations differ in the placement of the 4^{th} ring site and the 2 hexyl sites. Mapping 1 (Figure 5.6a1) associates the CR1 site with the mass center for the two ring carbons that are not bonded to the sulfur and associates the two side chain sites with the mass centers for the $CH_2CH_2CH_2$ and $CH_2CH_2CH_3$ groups along the hexyl chain. Mapping 2 (Figure 5.6a2) associates the CR1 site with the ring carbon that is bonded to the hexyl chain and associates the two side chain sites with the mass centers for the CH_2CH_2 and $CH_2CH_2CH_2CH_3$ groups along the hexyl chain. The CG molecular states depend quite sensitively upon the details of the mapping, such as moving the CR1 site between two mappings. We rejected other 6-site mappings for 3HT because they poorly described the AA molecular states. The Supporting Information presents additional results for two 5-site representations of 3HT in order to correlate the structural accuracy of the CG model with the mapping of the thiophene ring.

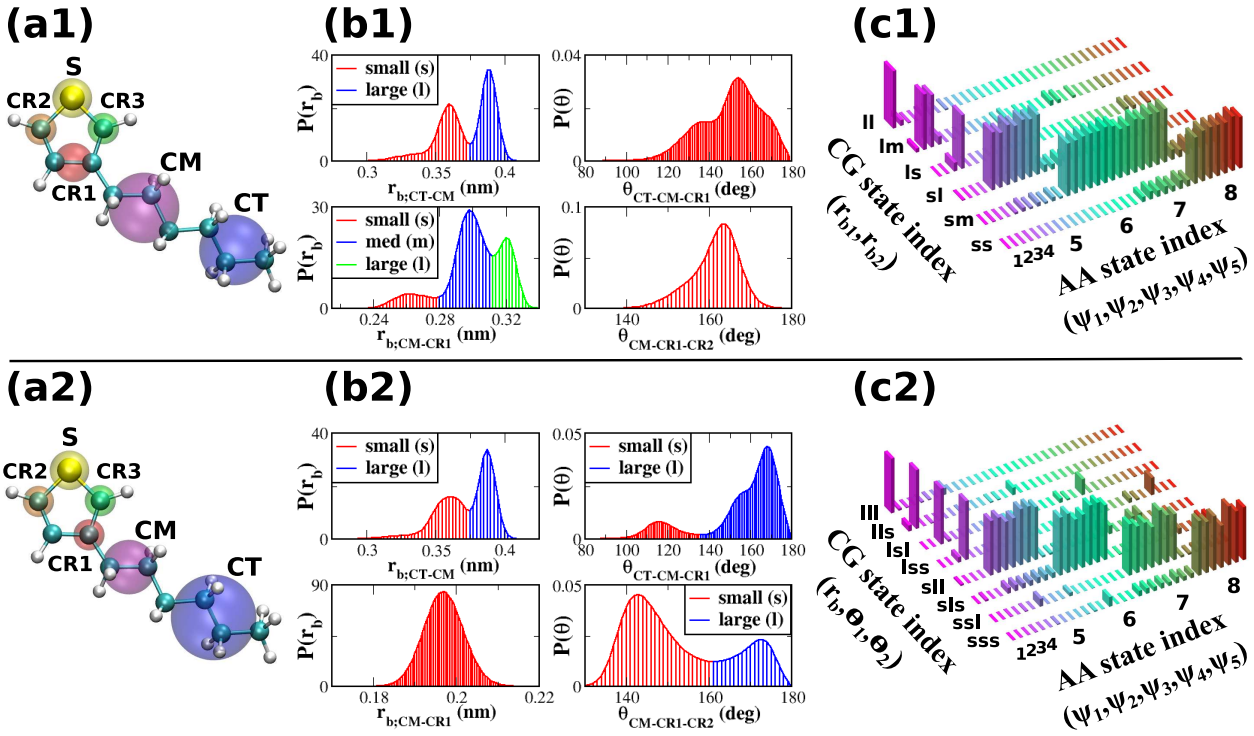

Figure 5.6. Analysis of the molecular states for two 6-site representations of 3HT. In each row, column a illustrates the corresponding CG mapping, column b identifies the subpopulations defined by the 1-D distributions that are sampled by the AA model along relevant CG dofs, and column c analyzes the mappings of molecular states. In column c, the height at location (x,y) indicates the conditional probability, $p(y|x)$, that AA molecular state x is mapped to CG molecular state y.

We next analyzed the representation of these AA molecular states by each CG mapping. As described above for hexane, for each CG mapping, we generated a mapped ensemble of CG configurations by applying the mapping to the ensemble sampled by the AA simulation. For each mapped ensemble, we then defined the CG molecular states by analyzing the resulting 1-D distributions along each intramolecular CG dof. The resulting CG states are distinguished by the conformations of two bonds (CT-CM, CM-CR1) and two angles (CT-CM-CR1, CM-CR1-CR2).

The first row of Figure 5.6 analyzes the mapped ensemble for mapping 1. The resulting CT-CM bond distribution is bimodal, the resulting CM-CR1 bond distribution is trimodal, and both resulting angle distributions are unimodal. Although the CT-CM-CR1 distribution suggests the presence of two subpopulations, the overlap in Figure 5.6b1 demonstrates that mapping 1 cannot distinguish these subpopulations. Accordingly, we treat the CT-CM-CR1 angle distribution as unimodal. Consequently, for mapping 1, we define six CG molecular states according to the states of the CT-CM and CM-CR1 bonds. Figure 5.6c1 presents the state mapping for mapping 1. Mapping 1 effectively maps each AA state to a single CG

state. Additionally, every CG state corresponds to at least one AA state, i.e., none of the CG states are forbidden in the mapped ensemble.

The second row of Figure 5.6 analyzes the mapped ensemble for mapping 2. This mapping generates bimodal distributions for the CT-CM bond and both angles, but a unimodal distribution for the CM-CR1 bond. Consequently, for mapping 2, we define eight CG molecular states according to the states of the CT-CM bond, CT-CM-CR1 angle, and CM-CR1-CR2 angle. Figure 5.6c2 presents the state mapping for mapping 2. Mapping 2 also effectively maps each AA state to a single CG state and only generates CG states that are sampled by the AA model. Accordingly, we expect that both mappings will allow for an accurate description of molecular states.

5.4.2.2 Model Assessment

Figures 5.7-5.10 and Figures S4-S6 assess the equilibrium structure generated by several 6-site CG models for 3HT. In each figure, panels a and b present results for mappings 1 and 2, respectively. In all panels, the solid black curves present results obtained by mapping the simulated AA ensemble to the given CG representation, the dashed red curves present results for the MS-CG model, and the dashed-dotted green curves present results for the iter-gYBG model.

Intramolecular Structure: Figures 5.7 and S4 characterize the intramolecular structure sampled by the 6-site 3HT models.

For mapping 1, the MS-CG model nearly quantitatively reproduces most of the AA intramolecular distributions (Figure S4), but reproduces with only qualitative accuracy the CT-CM-CR1 and CM-CR1-CR2 angle distributions (Figures 5.7a(i) and (ii), respectively), as well as the CM-CR1-CR3-S dihedral and CR2-CR3 bond distributions (Figure S4). The iter-gYBG model nearly quantitatively reproduces almost all intramolecular distributions of the mapped ensemble, with the largest errors occurring in the CT-CM-CR1 and CM-CR1-CR2 angle and the CM-CR1-CR3-S dihedral distributions. Additional analysis (not shown) indicates that the accuracy of the CT-CM-CR1 and CM-CR1-CR2 angle are anti-correlated during the iter-gYBG procedure. Consequently, the iter-gYBG model reproduces the mapped distribution for the CM-CR1-CR2 angle with slightly less accuracy than the MS-CG model. However, Figure S2 demonstrates that, on average, the iter-gYBG model reproduces the 1-D AA distributions with greater accuracy than the MS-CG model.

Figure 5.7a(iii) presents the simulated distributions for the CG molecular states that are defined for mapping 1 in Figure 5.6c1. The MS-CG model qualitatively reproduces the mapped distribution of molecular states, while the iter-gYBG model nearly quantitatively

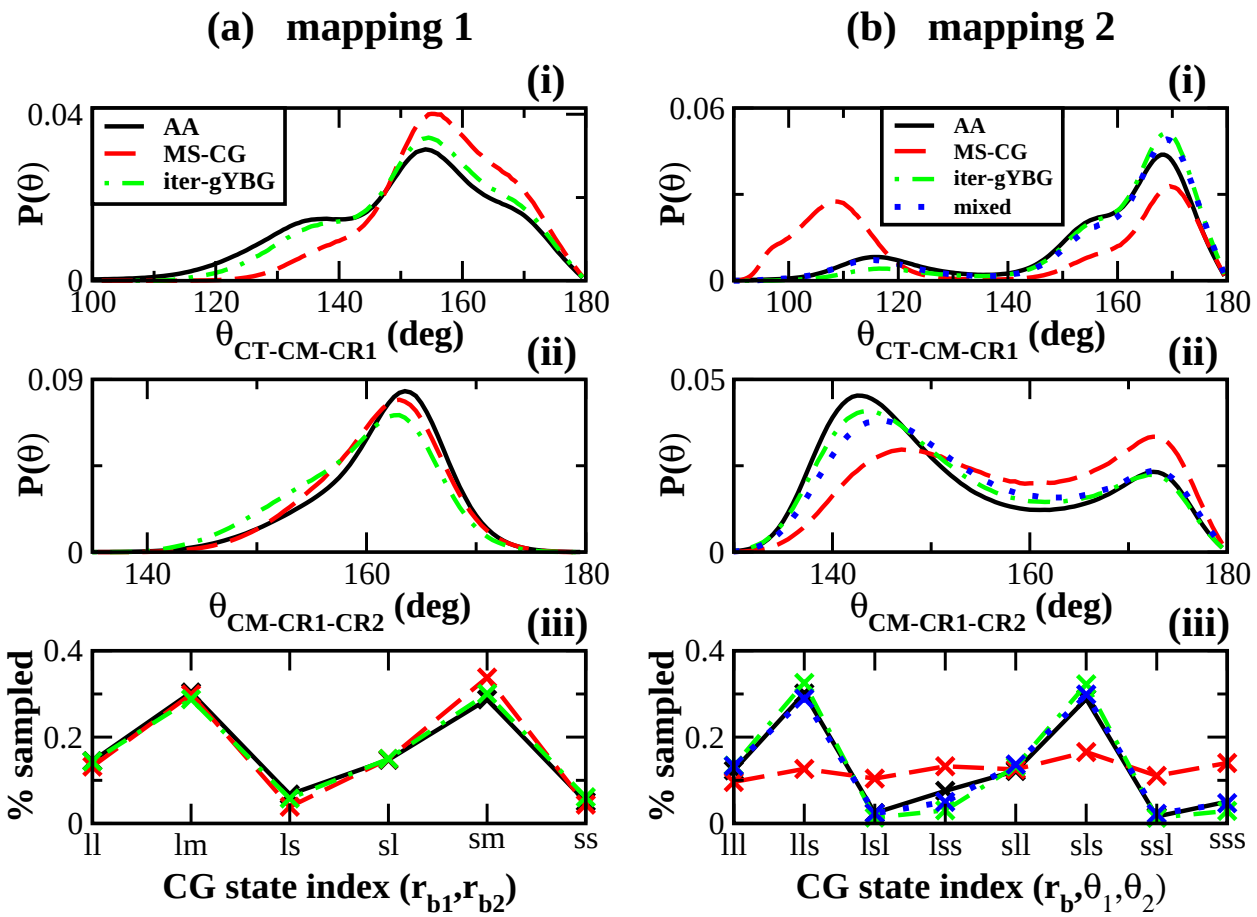

Figure 5.7. Accuracy of the intramolecular structure generated by 6-site CG models for 3HT. Columns a and b correspond to mappings 1 and 2, respectively. In each panel, solid black, dashed red, and dashed-dotted green curves correspond to the AA, MS-CG, and iter-gYBG models, respectively. The blue curves in column b correspond to a mixed model that was calculated with several simplified cross-correlations. Panels (i) and (ii) present the CT-CM-CR1 and CM-CR1-CR2 angle distributions, respectively, for each mapping. Panel (iii) presents the sampled distributions of molecular states for each mapping.

reproduces this distribution.

For mapping 2, the MS-CG model nearly quantitatively reproduces several of the AA intramolecular distributions (Figure S4), but exhibits significant errors in the CT-CM-CR1 and CM-CR1-CR2 angle distributions (Figures 5.7b(i) and (ii), respectively), as well as in the CM-CR1-CR3-S dihedral and CR2-CR3 bond distributions (Figure S4). The iter-gYBG model nearly quantitatively reproduces almost all intramolecular distributions of the mapped ensemble. The largest errors for the iter-gYBG model occur in the CT-CM-CR1 and CM-CR1-CR2 angle and the CM-CR1-CR3-S dihedral distributions.

Figure 5.7b(iii) presents the simulated distributions of CG molecular states that are defined for mapping 2 in Figure 5.6c2. Although the MS-CG model does not reproduce the

mapped distribution of states, the iter-gYBG model reproduces this distribution with nearly quantitative accuracy.

These observations are consistent with our a priori expectations based upon the mapping of molecular states. Both 6-site mappings provide a good representation of the molecular states. Consequently, the iter-gYBG models that accurately reproduce the mapped 1-D distribution functions also accurately describe the mapped distribution of molecular states.

The Supporting Information demonstrates markedly different results for two 5-site 3HT models. Figure S7 demonstrates that, in stark contrast to the two 6-site representations, the two 5-site representations provide a poor mapping of molecular states for 3HT. The two 5-site representations split many AA states between CG states and introduce several forbidden states. Figure S9 demonstrates that both 5-site iter-gYBG models reproduce the corresponding mapped 1-D distributions with reasonable accuracy. However, Figure S9 also demonstrates that these models fail to reproduce the mapped distribution of molecular states. In fact, for these "bad" mappings, the process of refining the CG potential to more accurately model the mapped 1-D distribution functions can actually lead to less accurate models for the molecular states.

Interestingly, although both 6-site iter-gYBG models quite accurately reproduce the mapped intramolecular structure of the AA model, Figure 5.7 demonstrates a striking difference in the structural accuracy of the two 6-site MS-CG models. While the MS-CG model for mapping 1 accurately modeled both the 1-D intramolecular distributions and the molecular states of the AA model, the MS-CG model for mapping 2 demonstrated significant errors in both of these structural metrics. This suggests that mapping 2 introduced complex cross-correlations between dofs in the CG model that cannot be captured by a simple molecular mechanics approximation to the many-body PMF.

Accordingly, we examined the matrix of AA cross-correlations, $\bar{\mathbf{G}}^{AA}$, that is employed in Equation (5.12) to determine the MS-CG model for mapping 2. Figure S3b demonstrates that mapping 2 introduced particularly complex cross-correlations between the CT-CM-CR1 angle and three other intramolecular dofs. We hypothesized that these complex cross-correlations caused the errors observed in Figure 5.7b for the MS-CG model for mapping 2. In order to test this hypothesis, we performed a similar analysis to that described above for hexane. We replaced these three blocks of $\bar{\mathbf{G}}^{AA}$ with the corresponding cross-correlations that were used in generating the iter-gYBG model that reproduced the AA 1-D distributions. After performing this substitution, we then solved the MS-CG equation, Equation (5.12), to determine a new CG force field. The blue curves in Figures 5.7-5.10 and Figures S4-S6 analyze the structure generated by the resulting "mixed" CG model. Figure 5.7b demonstrates

that this CG model reproduces the AA angle distributions with much better accuracy than the original MS-CG model. Interestingly, this mixed model also reproduced, with nearly quantitative accuracy, the molecular state distribution that was sampled by the AA model. These results indicate that these cross-correlations were the major source of the errors in Figure 5.7b and that, moreover, the iter-gYBG model accurately modeled the 1-D mapped distributions of the AA model at the expense of distorting the cross-correlations of the model.

Intermolecular Structure: Figures 5.8-5.10 and Figures S5-S6 characterize the intermolecular structure generated by the 6-site CG models for 3HT. In these figures, the dashed blue curves present results for the "mixed" CG model for mapping 2, which was generated by calculating the CG potential via the MS-CG equations after replacing certain complex intramolecular AA cross-correlations with the simpler cross-correlations that were employed in determining the iter-gYBG model, as described above.

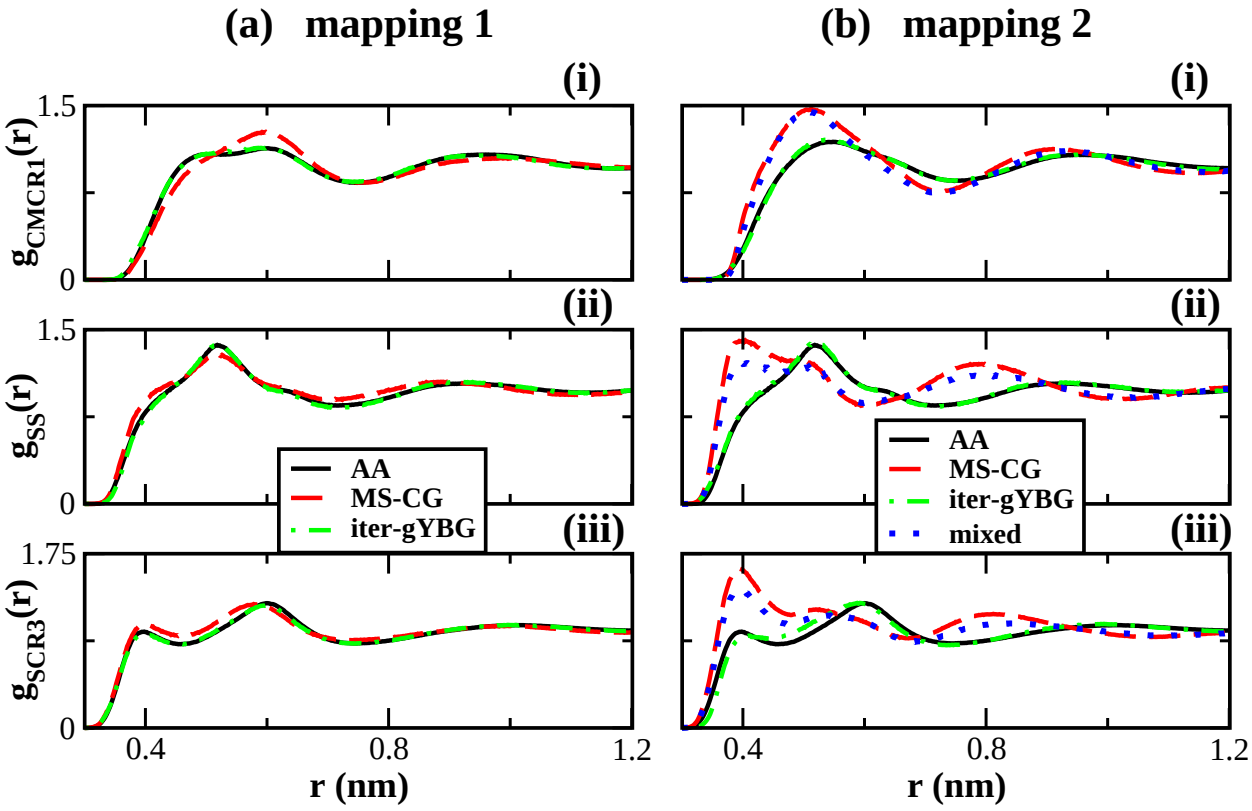

Figure 5.8. Site-site rdfs for 6-site 3HT models. Columns a and b correspond to mappings 1 and 2, respectively. In each panel, solid black, dashed red, and dashed-dotted green curves correspond to the AA, MS-CG, and iter-gYBG models, respectively. The blue curves in column b correspond to a mixed model that was calculated with several simplified cross-correlations. Panels (i)-(iii) present the CMCR1, SS and SCR3 rdfs, respectively.

Figure 5.8 presents a subset of the 21 rdfs for the CG sites that were selected to highlight the errors in the CG models. Figure S5 presents the complete set of rdfs. The MS-CG model for mapping 1 qualitatively reproduces all of the AA rdfs, while the MS-CG model for mapping 2 generates significant errors in several of the rdfs. For both mappings, the iter-gYBG model corrects these errors and nearly quantitatively reproduces all of the AA rdfs.

The mixed model for mapping 2 reproduces the AA rdfs with only slightly greater accuracy than the original MS-CG model, which suggests that the errors in the MS-CG rdfs result from complex cross-correlations involving intermolecular dofs. Since these errors do not occur for the 5-site models (Figure S10) or for the 6-site models that were calculated for the symmetric mapping 1 (Figures 5.8a and S5a), these errors likely result from the asymmetric placement of the CR1 site in mapping 2. However, further analysis is required to better understand these errors.

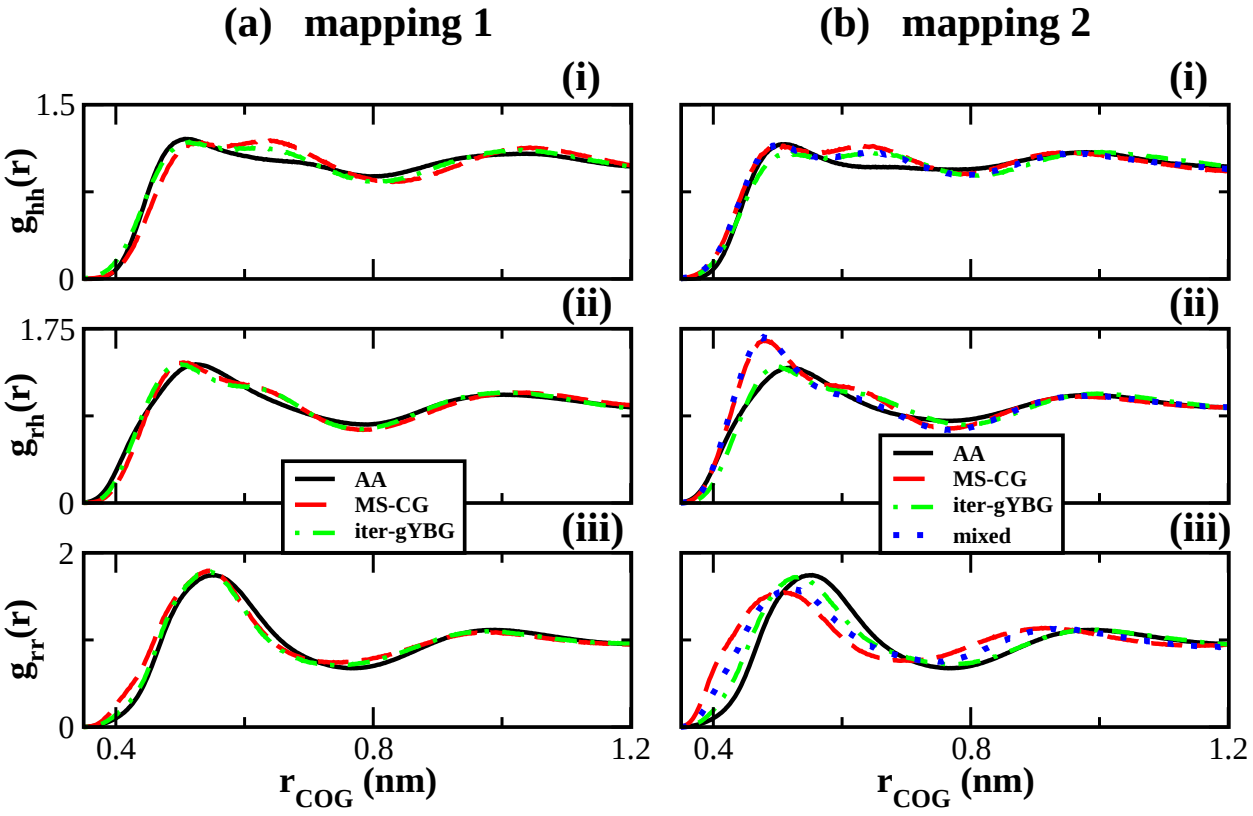

Figure 5.9. Center of geometry (cog) rdfs for 6-site 3HT models. Columns a and b correspond to mappings 1 and 2, respectively. In each panel, solid black, dashed red, and dashed-dotted green curves correspond to the AA, MS-CG, and iter-gYBG models, respectively. The blue curves in panel b correspond to a mixed model that was calculated with several simplified cross-correlations. Panels (i)-(iii) present the hexyl-hexyl, hexyl-ring and ring-ring cog rdfs, respectively.

Figure 5.9 characterizes the packing of hexyl tail groups (h), which are defined by the center of geometry (cog) for the CT and CM sites, and of ring groups (r), which are defined by the cog for the CR2 and CR3 sites. For each of the mappings, panels (i)-(iii) present the hexyl-hexyl, hexyl-ring, and ring-ring cog rdfs, respectively. For mapping 1, the small errors in the site-site rdfs of the MS-CG model result in only minor discrepancies in the cog rdfs, while the iter-gYBG model reproduces these cog rdfs with nearly quantitative accuracy. For mapping 2, the large errors in the MS-CG site-site rdfs cause significant discrepancies for the cog rdfs, especially for rdfs involving ring groups. The iter-gYBG model significantly reduces these errors and nearly quantitatively reproduces the cog rdfs. Interestingly, the blue curves in Figure 5.9 demonstrate that, by improving the description of intramolecular correlations, the mixed model also reproduces the ring-ring cog rdf with somewhat greater accuracy than the original MS-CG model.

Figure 5.10 characterizes the average relative orientation of the hexyl tails and thiophene rings as a function of distance. The orientation of hexyl tails is characterized by the molecular director, $\hat{h}$, defined as the unit vector pointing from the CM to the CT site. The orientation of ring groups is characterized by two different molecular directors: $\hat{r}_1$ is the unit vector normal to the plane of the ring and $\hat{r}_2$ is the unit vector pointing from the center of mass of the CR2 and CR3 sites to the S site. The orientational preferences are characterized by the average of the second Legendre polynomial, $\langle P_2(x)\rangle$, as a function of distance between groups. The second Legendre polynomial between directors $\hat{x}_1$ and $\hat{x}_2$ is given by $P_2(\hat{x}_1 \cdot \hat{x}_2) = P_2(\cos\theta) = \frac{3}{2}(\cos\theta)^2 - \frac{1}{2}$, where θ is the angle formed between the two directors. $P_2(\cos\theta)$ is 1 if the directors are parallel (or anti-parallel) and -0.5 if they are perpendicular.

Panels a(i) and b(i) present $\langle P_2(\hat{r}_1 \cdot \hat{r}'_1)\rangle$, which characterizes the tendency of the ring faces to align as a function of separation. This packing is largely due to the ring geometry and is well reproduced by both the MS-CG and iter-gYBG models for both 6-site mappings. In contrast, 5-site models that represent the thiophene ring with 3 sites do not model this stacking as accurately. (See Figure S12.) Panels a(ii) and b(ii) present $\left\langle P_2(\hat{h} \cdot \hat{r}'_2)\right\rangle$, which characterizes the tendency of the hexyl tails to be aligned (either parallel or anti-parallel) with the second ring director, $\hat{r}_2$. For both 6-site representations, neither the MS-CG or iter-gYBG models reproduce the hexyl-ring packing very accurately at short separation distance. However, the models derived using mapping 1 describe the hexyl-ring packing slightly more accurately. Panels a(iii) and b(iii) present $\langle P_2(\hat{h} \cdot \hat{h}')\rangle$, which characterizes the tendency of the hexyl tails to be aligned (either parallel or anti-parallel). Each of the 6-site models, qualitatively reproduces the hexyl tail alignment.

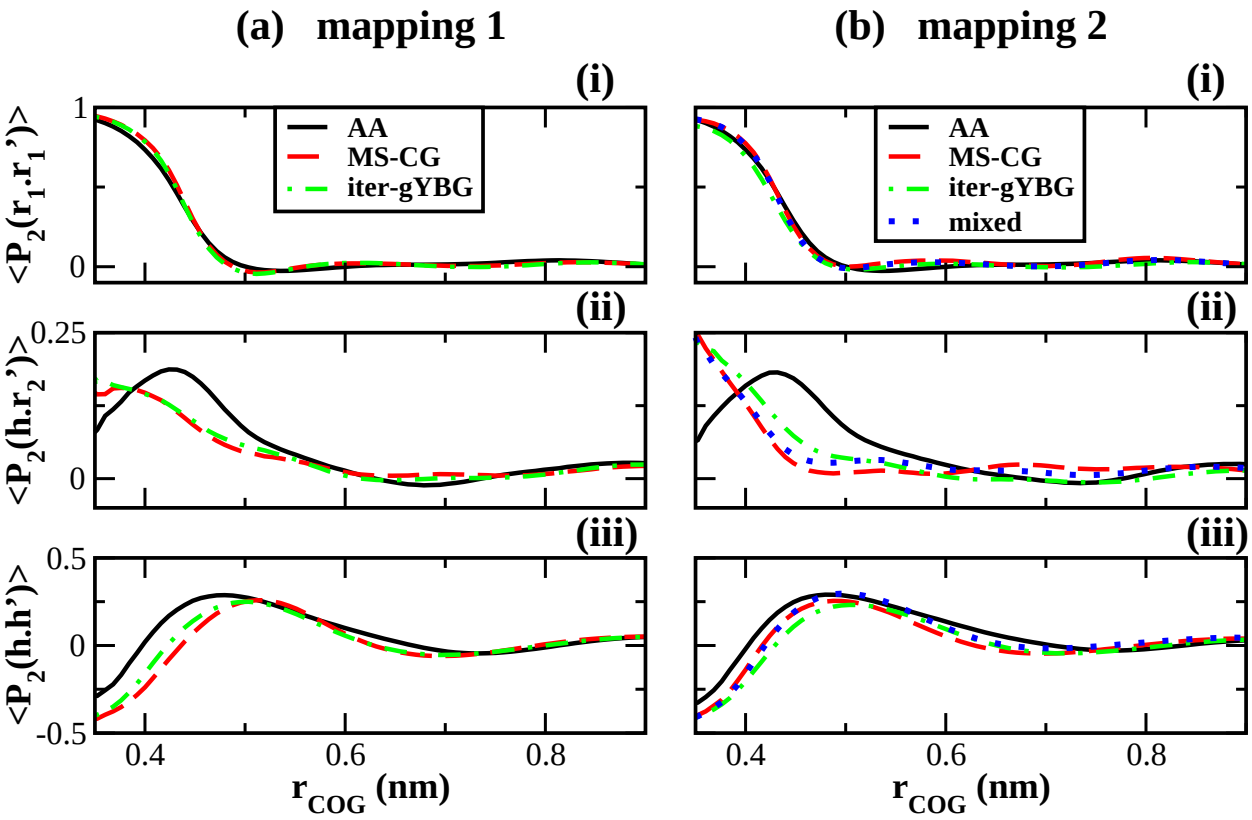

Figure 5.10. Intermolecular alignment for 6-site 3HT models. Columns a and b correspond to mappings 1 and 2, respectively. In each panel, solid black, dashed red, and dashed-dotted green curves correspond to the AA, MS-CG, and iter-gYBG models, respectively. The blue curves in panel b correspond to a mixed model that was calculated with several simplified cross-correlations. Intermolecular alignment is characterized by the average of the second Legendre polynomial, $\langle P_2(\cos\theta)\rangle$, for the angle, θ, formed by directors as a function of distance between corresponding groups. The directors, $\{\hat{h}, \hat{r}_1$, and $\hat{r}_2\}$, correspond to the hexyl tail director, the thiophene ring normal, and the thiophene ring direction.

5.5 Discussion

The interaction potentials in CG models are often parameterized to reproduce the 1-D distributions, e.g., rdfs, that are determined by mapping an AA ensemble to the CG representation.[23] However, the cross-correlations between these interactions preclude any simple means for directly determining the CG potentials to reproduce target AA distribution functions.[118,183,226] Consequently, many methods iteratively refine the CG potentials over multiple simulations.[38,49,50,225]

In contrast, the MS-CG method[36,41,59,60] directly determines CG potentials without requiring iterative simulations. The MS-CG method employs the g-YBG relation to treat these cross-correlations and determine the potentials that will reproduce the projections of

the many-body mean force (and, thus, also the AA rdfs).[62–64] However, the MS-CG method approximates the cross-correlations that will be present in the CG model with the cross-correlations that are observed in the AA model.[230] In cases that this approximation fails, the MS-CG model may not accurately reproduce the target AA distributions.

Motivated by these considerations, Cho and Chu[69] developed an iterative g-YBG procedure for systematically refining CG potentials in order to reproduce target AA rdfs. They demonstrated this method for CG models of several fairly small molecules, while employing a fluctuation matching procedure[242] to reproduce the intramolecular fluctuations observed in the AA model. This iterative g-YBG method employed the cross-correlations generated by each successive CG simulation to systematically modify the nonbonded pair potentials at each iteration. Very recently, Lu et al.[75] demonstrated a simpler version of this iterative method. Interestingly, though, this latter method employed the cross-correlations that were observed in the AA model, rather than in the CG model.

The present work developed a framework for clarifying the physical content of these iterative g-YBG methods. This framework extends the approach of Cho and Chu[69] by consistently treating both intra- and intermolecular interactions for arbitrarily complex molecular mechanics potentials, although we introduced some heuristic modifications in our numerical calculations. In addition, our analysis indicates that the method of Lu et al.[75] provides an elegant approximation to the more general method. This locally linear approximation may be particularly accurate when the original MS-CG model is sufficiently accurate or when the relevant structural features are largely determined by packing considerations. Conversely, this approximation may prove less successful in cases that strong, specific interactions, such as hydrogen bonds, determine the relevant structures.

In some ways, the iterative g-YBG method may be considered analogous to iterative Boltzmann inversion (IBI).[35,49] Neither method provides an underlying variational principle and neither method guarantees convergence. Moreover, while IBI iteratively refines potentials to match pair potentials of mean force (and, thus, also corresponding pair mean forces),[183] the iterative g-YBG method refines potentials to match projections of the many-body mean force along particular CG dofs. In the case of central pair potentials, these force projections are equivalent to the pair mean forces that are reproduced by IBI.[64,188] However, while the IBI method neglects correlations between different interactions at each refinement step, the iterative g-YBG method employs these cross-correlations when updating the potentials.

We have numerically demonstrated the iterative g-YBG method for several systems. In addition to 3- and 4-site models of hexane, we have also applied the method to parameterize

several 5- and 6-site CG models of 3-hexylthiophene, which is a considerably more complex molecule. In each case, the model obtained from the iterative g-YBG method reproduces the 1-D distribution functions of the AA model with greater accuracy than the MS-CG model. In several cases, the improvement is very significant. Moreover, in most cases, the iterative g-YBG model reproduces the AA 1-D distributions with almost quantitative accuracy.

The Supporting Information discusses the convergence, stability, and accuracy of the method. In particular, the Supporting Information demonstrates that, under certain circumstances, the method may not find a model that perfectly satisfies the self-consistent g-YBG equation or the method may appear to diverge after initially obtaining an accurate model. Further studies are clearly necessary to investigate these issues. Nevertheless, the present results suggest that the iterative g-YBG method may provide a robust method for determining CG models that reproduce 1-D AA distributions with reasonable accuracy.

Our studies also investigated the sensitivity of MS-CG models to the CG representation for both hexane and 3HT. In particular, we considered two 6-site MS-CG models of 3HT. The MS-CG model for mapping 1 reproduced the corresponding AA distribution functions quite accurately, while the MS-CG model for mapping 2 reproduced the corresponding AA distribution functions relatively poorly. We expected that these errors resulted from overly complex cross-correlations that were generated by mapping the AA ensemble to the given CG representation. We numerically verified this hypothesis in several cases by solving the MS-CG equations to determine potentials after substituting simplified cross-correlations for the more complex cross-correlations of the mapped AA ensemble. In each case, the resulting CG model more accurately reproduced the mapped 1-D distributions.

The iterative g-YBG method did determine models that reasonably reproduced the mapped 1-D distributions for these problematic cases, e.g., mapping 2 for 3HT. However, it did so at the expense of distorting the cross-correlations between CG dofs. Clearly, certain CG representations generate a mapped ensemble with complex cross-correlations between CG dofs that cannot be reproduced by simple molecular mechanics potentials. In such cases, one expects the MS-CG method will provide limited accuracy, since it assumes that the resulting model will reproduce these cross-correlations.[73] Thus, the structural accuracy of CG models and, in particular, MS-CG models depends quite sensitively upon a subtle relation between the cross-correlations present in the AA model, the CG representation, and the flexibility of the approximate CG potential. Our results strongly motivate further investigations of this relationship.

This discussion also emphasizes a well-known and very general aspect of coarse-grained modeling: Coarse-grained models that reproduce the 1-D distribution functions of an AA

model are certainly not guaranteed to reproduce the cross-correlations that are observed in the AA model.[243] In particular, even if a CG model reproduces the 1-D distribution functions of an AA model for all intramolecular dofs, the CG model may not accurately describe the molecular conformations, or molecular states, that are actually sampled in the AA model. Figure 3a explicitly demonstrates this for the 3-site model of hexane. In this case, the iterative g-YBG model reproduces all 1-D distributions of the AA model, but does not sample the correct distribution of molecular states. The Supporting Information section demonstrates similar results for two 5-site models of 3HT. These considerations may be even more problematic for biomolecules, such as proteins, with hierarchical structures that reflect significant cross-correlations.[227–229]

Importantly, our calculations clearly demonstrate that these discrepancies depend quite sensitively upon the mapping and, at least in the present cases, can be resolved by employing a "good" mapping. In particular, Figure 3b demonstrates that the 4-site iterative g-YBG model for hexane reproduces not only the AA 1-D distribution functions, but also the AA probabilities for sampling all six molecular states. Similarly, both 6-site iterative g-YBG models for 3HT reproduced the distribution of molecular states sampled by the AA model.

Finally, the most significant and generally useful outcome of this work may be the analysis of molecular states for identifying good CG maps. Our calculations indicate that good CG maps preserve a simple relationship between the molecular states that are sampled in the AA model and those that are accessible in the CG model. An ideal CG map should allow one and only one molecular state for each relevant state that is sampled in the AA model. For a good CG map, an accurate reproduction of the mapped AA distributions along individual dofs will also promote an accurate sampling of molecular states. Importantly, this analysis can be performed directly from an AA ensemble before calculating a CG potential or performing any CG simulations. Conversely, our analysis highlights three features of poor CG maps: 1) Poor CG maps split molecular states of the AA model into multiple molecular states for the CG model. 2) More importantly, poor CG maps introduce "forbidden" states that become accessible to the CG model, but that were not sampled in the AA model. These forbidden states reflect cross-correlations in the AA model that are not captured by a simple molecular mechanics CG potential. 3) Finally, poor CG maps "hide" important molecular states of the AA model from the CG model, which results in irretrievably lost information about the molecular state. Of course, the present work provided only an initial investigation of these considerations. Future work should extend this analysis for treating nonbonded interactions. Nevertheless, the present simple treatment is already clearly relevant for properly describing the conformations of fairly complex molecules such as 3HT.

Supporting Information Available

The Supporting Information provides a detailed description of all methods, simulations, and calculations employed in this work, as well as additional analysis of these calculations. This information is free of charge via the Internet at `http://pubs.acs.org`.

Chapter 6

Minimal Models for Disordered and Helical Peptide Ensembles

J. F. Rudzinski, W. G. Noid *J. Chem. Theor. Comp.* **submitted 11/2014**

Abstract

This work investigates the capability of bottom-up methods for parameterizing minimal coarse-grained (CG) models of disordered and helical peptides. We consider four distinct high resolution peptide models that sample ensembles with varying complexity. For each high resolution model, we parameterize a CG model via the multiscale coarse-graining (MS-CG) method, which employs a generalized Yvon-Born-Green (g-YBG) relation to determine potentials directly (i.e., without iteration) from the high resolution ensemble. The MS-CG method accurately describes high resolution models that fluctuate about a single conformation. However, given the minimal resolution and simple molecular mechanics potential, the MS-CG method provides a less accurate description for a high resolution peptide model that samples a disordered ensemble with multiple distinct conformations. We employ an iterative g-YBG method to develop a CG model that more accurately describes the relevant distribution functions and free energy surfaces for this disordered ensemble. Nevertheless, this more accurate model does not reproduce the cooperative helix-coil transition that is sampled by the high resolution model. By comparing the different models, we demonstrate that the errors in the MS-CG model primarily stem from the lack of cooperative interactions afforded by the minimal representation and molecular mechanics potential. This work demonstrates the potential of the MS-CG method for accurately modeling complex biomolecular structures, but also highlights the importance of more complex potentials for modeling cooperative transitions with a minimal CG representation.

6.1 Introduction

Atomically-detailed molecular dynamics (MD) simulations provide tremendous insight into protein structure and fluctuations on nanosecond time scales.[24] Nevertheless, despite great strides in computational methods and resources, atomically-detailed models remain prohibitively inefficient for investigating many complex biological processes, such as peptide aggregation, that evolve on much longer time scales.[244] Consequently, lower resolution coarse-grained (CG) models continue to enjoy surging popularity.[23,224,245,246]

In particular, minimal CG models have proven to be particularly useful for modeling protein folding and interactions. By eliminating an explicit treatment of solvent and by representing each amino acid with a single site, which is usually associated with the corresponding α-carbon, these models provide tremendous efficiency.[247] Moreover, due to the regularity of the protein backbone geometry,[248,249] this remarkably sparse minimal representation still allows for an accurate atomic reconstruction of the peptide backbone.[250,251]

Accordingly, a vast array of off-lattice minimal models have been parameterized by various means and for various purposes. For instance, the seminal studies of Thirumalai and coworkers[22,220] represented proteins with a "reduced alphabet" (i.e., amino acids are distinguished by their character as, e.g., hydrophobic or polar) and employed simple potentials to investigate universal features of protein folding. Conversely, native-based Gō models[252–254] and network models[255,256] represent proteins with an "extended alphabet" (i.e., amino acids are distinguished by their location in the protein sequence) and employ biased potentials that stabilize known structures in order to characterize the folding and fluctuations of specific proteins.[257,258] Additionally, minimal models have been employed to characterize generic aspects of cellular crowding, unfolded protein ensembles, and peptide aggregation.[259–262] These latter minimal models have often been parameterized via top-down approaches[221,263] that attempt to capture emergent, often thermodynamic, properties.

In contrast, several recent studies have employed bottom-up approaches to parameterize CG peptide models that accurately describe the ensembles sampled by all-atom (AA) models.[76–78,116,117,228,229,264–266] The many-body potential of mean force (PMF) is the appropriate potential for a CG model that reproduces all structural features of the mapped AA ensemble, i.e., the ensemble generated by mapping the AA ensemble to the CG representation. Bottom-up methods typically approximate the many-body PMF with simple molecular mechanics potentials that include separate terms for bond, angle, torsion, and pair "nonbonded" interactions. These various terms are often iteratively refined in order to reproduce target 1-D distribution functions for corresponding degrees of freedom in the

CG model.[34,35,37,38] In general, these studies have described the mapped AA ensemble with reasonable accuracy. For instance, the CG model of Bezkorovaynaya et al.[228] reproduced the relevant 1-D distribution functions of the underlying ensemble quite accurately, but did not accurately describe the cross-correlations between the angle and torsion degrees of freedom along the CG peptide backbone.

The present work expands upon previous studies by further investigating the capabilities and limitations of bottom-up coarse-graining methods for determining minimal peptide models that accurately describe AA conformational ensembles for helical and disordered peptides. In particular, we employ the multiscale coarse-graining (MS-CG) method[36,59] to determine potentials that provide a variationally optimal approximation to the many-body PMF.[41,60] In contrast to many other bottom-up structure-based approaches, which iteratively refine the model potential to reproduce particular structural features, the MS-CG method employs a generalized Yvon-Born-Green (g-YBG) equation[62–64] to directly (i.e., non-iteratively) determine the approximate CG potential from the correlations that are observed in the mapped AA ensemble.

In a certain sense, the MS-CG/g-YBG approach is quite elegant, since it provides a direct solution to the inverse problem of inferring potentials from the AA ensemble. Moreover, the MS-CG/g-YBG approach also holds considerable computational promise. While iterative methods require the solution to nonlinear optimization problems for a large number of parameters, the MS-CG/g-YBG method requires only the solution of a linear least squares problem. However, the MS-CG/g-YBG procedure rests upon the fundamental assumption that the form of the CG potential is sufficiently flexible for reproducing the relevant cross-correlations of the mapped AA ensemble.[230] Clearly this assumption depends not only upon the AA model, but also upon the complexity of the CG potential and the CG representation of the AA model.[73,267]

This assumption is quite central to bottom-up CG methods. In cases that this assumption is valid, the MS-CG model will accurately reproduce the structure of the mapped AA ensemble. However, in cases that this assumption is not valid, the MS-CG model may not accurately describe this ensemble. Moreover, in this case, iterative bottom-up methods may reproduce the target 1-D distribution functions for individual degrees of freedom, but will do so at the expense of distorting the cross-correlations between these degrees of freedom. This distortion may prove especially detrimental for describing the complex, hierarchical structures of proteins and other biomolecules.

Accordingly, the objective of the present work is to assess this basic assumption in the context of parameterizing minimal CG peptide models with implicit solvent. We demonstrate

that the MS-CG/g-YBG framework directly determines very accurate minimal models for peptides that fluctuate about a single well-defined conformation. These results complement the results of previous studies[76–78] with the MS-CG method that represented peptides with slightly higher resolution and employed explicit solvent. However, given the minimal representation and molecular mechanics potential, the MS-CG/g-YBG framework provides a less accurate description for more complex ensembles that sample multiple conformations. In order to investigate this discrepancy, we employ an iterative g-YBG (iter-gYBG) method[69,75] to parameterize CG models for these more complex ensembles. These iter-gYBG models provide a reasonably accurate description of the mapped AA ensembles and reveal the structural features of the mapped ensembles that are not consistent with the assumptions of the MS-CG/g-YBG framework. Finally, by adapting the g-YBG framework to account for these problematic features of the mapped AA ensemble, we significantly improve the accuracy of the resulting MS-CG minimal peptide models.

6.2 Theory

The MS-CG, g-YBG, and iterative g-YBG methods have been extensively discussed in previous papers.[36,41,48,59,64,69,75,188,267] We briefly summarize the details that are particularly relevant for the present work. We first consider an AA model with a configuration, $\mathbf{r}$, that is defined by the Cartesian coordinates for n atoms. We assume that a mapping function determines a configuration, $\mathbf{R}$, for the CG model as a linear function of $\mathbf{r}$. In the present work, the CG configuration $\mathbf{R}$ corresponds to the Cartesian coordinates of the α-carbons in the AA model. It is convenient to define the "mapped AA ensemble" as the ensemble of CG configurations that is generated by applying the mapping to each configuration that is sampled by the AA model. The many-body potential of mean force (PMF), $W(\mathbf{R})$, is the appropriate potential for a CG model that quantitatively reproduces all structural properties of this mapped AA ensemble.[23] The PMF may be defined, to within a configuration independent constant, by

$$W(\mathbf{R}) = -k_B T \ln p_R(\mathbf{R}) + \text{const}, \tag{6.1}$$

where $p_R(\mathbf{R})$ is the probability for the AA model to sample a configuration $\mathbf{r}$ that maps to the CG configuration $\mathbf{R}$.[23] In general, the PMF is a complex, many-body function. The MS-CG method determines the parameters, $\boldsymbol{\phi}^0$, for a molecular mechanics potential energy function, $U(\boldsymbol{\phi}^0)$, that provides a variationally optimal approximation to the PMF.[41,62] According to

the MS-CG objective function,[36,41,59,60] this optimal approximation is determined by directly inverting the normal system[62] of linear equations:

$$\boldsymbol{b}^{AA} = \mathbf{G}^{AA}\boldsymbol{\phi}^{0}. \tag{6.2}$$

In Equation 6.2, $\boldsymbol{b}^{AA}$ is a vector of ensemble averages that can be expressed either in terms of AA forces[41,62] or in terms of a corresponding set of structural correlation functions.[63,64,188]

We focus on the common case that the nonbonded contribution to the CG potential, U, is represented by a sum of central pair potentials and each of these pair potentials is represented by a set of flexible basis functions (e.g., spline functions). In this case, a subset of the elements in $\boldsymbol{b}^{AA}$ is in 1-1 relationship with the radial distribution functions (rdfs) for the CG sites in the mapped AA ensemble. $\mathbf{G}^{AA}$ is a matrix of ensemble averages that quantify the cross-correlations between pairs of CG degrees of freedom in the mapped AA ensemble. This matrix allows the MS-CG method to decompose the force correlation vector, $\boldsymbol{b}^{AA}$, into contributions from the various terms in the CG potential, $U(\boldsymbol{\phi}^{0})$.

Equation 6.2 is related to a generalized Yvon-Born-Green equation[63,64] that exactly relates a given set of potential parameters, $\boldsymbol{\phi}$, to the vector, $\boldsymbol{b}(\boldsymbol{\phi})$, of force correlation functions and to the matrix, $\mathbf{G}(\boldsymbol{\phi})$, of cross-correlations that are generated by equilibrium sampling of a CG model with potential $U(\boldsymbol{\phi})$:

$$\boldsymbol{b}(\boldsymbol{\phi}) = \mathbf{G}(\boldsymbol{\phi})\boldsymbol{\phi}. \tag{6.3}$$

According to Equations 6.2 and 6.3, the MS-CG method employs the g-YBG relation to determine potential parameters $\boldsymbol{\phi}^{0}$ that reproduce the AA force correlation vector, $\boldsymbol{b}^{AA}$, but employs the matrix, $\mathbf{G}^{AA}$, of cross-correlations that are observed in the mapped AA ensemble to approximate the cross-correlations, $\mathbf{G}(\boldsymbol{\phi}^{0})$, that will be generated by the CG potential $U(\boldsymbol{\phi}^{0})$. If $\mathbf{G}(\boldsymbol{\phi}^{0}) = \mathbf{G}^{AA}$, then $\boldsymbol{b}^{0} = \boldsymbol{b}^{AA}$ and the CG model will reproduce the corresponding AA rdfs. However, this will generally not be the case because it would require that the CG model reproduce the higher order cross-correlations of the AA model.[230] Thus, if an MS-CG model does reproduce the rdfs, this suggests that it also likely reproduces higher order correlations of the AA model.

Iterative bottom-up CG procedures seek to determine potential parameters $\boldsymbol{\phi}^{*}$ that will reproduce the 1-D equilibrium distributions of the mapped AA ensemble for the relevant degrees of freedom in the CG model. In the context of the g-YBG framework,[69,75,267] this

corresponds to determining the force field coefficients $\boldsymbol{\phi}^*$ such that

$$\boldsymbol{b}^{AA} = \mathbf{G}(\boldsymbol{\phi}^*)\boldsymbol{\phi}^*. \tag{6.4}$$

In contrast to Equation 6.2 for the MS-CG potential, Equation 6.4 corresponds to a self-consistent g-YBG equation that determines the force field coefficients, $\boldsymbol{\phi}^*$, that reproduce $\boldsymbol{b}^{AA}$, while using the cross-correlations $\mathbf{G}(\boldsymbol{\phi}^*)$ sampled by a CG model with the corresponding potential $U(\boldsymbol{\phi}^*)$.[69]

In practice, the iter-gYBG procedure first determines the MS-CG potential parameters $\boldsymbol{\phi}^0$ according to Equation 6.2. Simulations with this CG model determine the resulting matrix, $\mathbf{G}(\boldsymbol{\phi}^0)$, of cross-correlations. The iter-gYBG method then determines a new set of potential parameters by solving Equation 6.4 for $\boldsymbol{\phi}^*$, while approximating $\mathbf{G}(\boldsymbol{\phi}^*)$ with the correlations, $\mathbf{G}(\boldsymbol{\phi}^0)$, generated by the preceding CG model. This procedure is iterated until the CG potential adequately reproduces the AA force correlation vector $\boldsymbol{b}^{AA}$ and, thus, also the AA pair structure. Note that this implies $\mathbf{G}(\boldsymbol{\phi}^*) \neq \mathbf{G}^{AA}$, i.e., the final CG model reproduces rdfs of the AA model at the expense of distorting higher order cross-correlations. As discussed further below, we have heuristically modified the method to improve its robustness for systems with complex intramolecular structure.[267]

6.3 Methods

This section summarizes the key details of our calculations. The Supporting Information section provides a much more detailed description.

6.3.1 Simulation Details

All reported molecular dynamics (MD) simulations were performed in the constant NVT ensemble with the Gromacs 4.5.3 simulation suite[7] according to standard procedures.[45,168–172] All-atom (AA) peptide simulations were performed at a temperature $T = 298$ K, while employing the OPLS-AA force field[18] to model peptide interactions and the SPC/E model[205] to describe the solvent (when applicable). We employed LINCS[268] to rigidly constrain all bonds that involve H atoms. The AA peptide models were capped with an N-terminal acetyl group and a C-terminal N-methyl amide group.

6.3.1.1 High Resolution Models

We considered four distinct high resolution peptide models. Three of these high resolution models represent short alanine peptides in conventional atomic detail. The fourth high resolution model is a C-α native-based,[252–254] i.e., Gō, model. Although the Gō model represents each amino acid with a single site, it employs considerably higher resolution than the MS-CG model that we parameterize for the Gō model. While the Gō model treats each site as a distinct type and employs 17 distinct types of pair potentials, the corresponding MS-CG model treats each site equivalently and employs only 3 distinct types of pair potentials to model the same set of interactions.

6.3.1.1.1 FCP1 We performed MD simulations of a C-α Gō model[252–254] for the C-terminal residues (944-61) of the FCP1 protein.[269] This region of FCP1 is intrinsically disordered in isolation.[270] However, when interacting with the C-terminal domain of the Rap74 subunit of Transcription Factor IIF, the FCP1 residues 944-57 fold to form an α helix, while the final 4 residues remain disordered.[271] We previously[272] employed the Structure-based Models in Gromacs web-server[273] (http://smog.ucsd.edu) to construct a Gō model for the Rap74-FCP1 system from the published crystal structure of the complex (PDBID: 1J2X).[271] The resulting intramolecular potential for the C-terminal FCP1 peptide defined a high resolution model for FCP1. We employed the Gō model to sample an ensemble with an average helical content of $\langle Q_{hel} \rangle = 0.65$, where Q_{hel} is defined below. We sampled a total of 2.8 million configurations for this model and employed the last 2.5 million for subsequent analysis.

6.3.1.1.2 AA Model for Alanine 12-mer in Vacuum We performed a 320 ns AA MD simulation of a capped alanine 12-mer in vacuum. After the first 20 ns, we sampled configurations every 1 ps to obtain a total of 300,000 configurations for subsequent analysis.

6.3.1.1.3 AA Models for Solvated Alanine Oligomers We also performed AA MD simulations with explicit solvent for a capped alanine tetramer and a capped alanine 12-mer.

After equilibration, we performed five independent 200 ns simulations of the solvated alanine tetramer. We sampled configurations every 1 ps to obtain a total of 1 million configurations for subsequent analysis.

We performed a single 600 ns production simulation of the solvated alanine 12-mer. After the initial 50 ns, we sampled configurations every 1 ps, which yielded a total of 550,000 configurations for subsequent analysis. This simulation resulted in a heterogeneous ensemble

that included helical, coil, and extended structures. Over the course of the AA trajectory, we observed 6 folding events, for which Q_{hel} increased from < 0.50 to > 0.98, and 15 partial folding events, for which Q_{hel} increased from < 0.55 to > 0.82. Although the resulting AA ensemble is unlikely to be completely converged, it provides a suitable ensemble for assessing the capability of the MS-CG method to accurately model a complex disordered AA ensemble.

6.3.1.2 CG Models

We constructed at least one CG model for each high resolution peptide model. Each CG model employed an implicit treatment of solvent and incorporated solvent effects into the interactions between the peptide CG sites. The CG peptide models represented amino acids with a single site that corresponded to the residue α-carbon. The CG model for FCP1 included 15 sites for residues 944-58. The CG models for AA models of alanine peptides associated a site with each α-carbon of the AA model. These CG models did not explicitly represent the capping groups that were present in the AA models.

For each CG model, the approximate potential function assumed a molecular mechanics form with bond, angle, and dihedral potentials for each pair, triple, and quadruple, respectively, of consecutive CG sites along the peptide chain. Each model included central pair potentials between sites separated by more than 2 bonds along the chain. We employed distinct 1-4 and 1-5 potentials to model the interaction between pairs of sites separated by exactly 3 and exactly 4 bonds, respectively. We employed an additional "nonspecific" pair potential to model interactions between all pairs separated by more than 4 bonds. For each of these 3 classes of pair interactions, we determined a distinct potential for each relevant pair of site types. The CG models for FCP1 and the alanine tetramer employed a single type of CG site. In contrast, the CG models for the 12-residue alanine peptides employed 5 site types with distinct types for each of the two terminal residues on either side of the peptide.

6.3.2 Force Field Calculations

Each calculated term in the CG potential was represented by a discrete set of basis functions of a single variable.[60] Different basis functions (e.g., linear spline, cubic B-spline, etc.) were employed for different types of interactions. The coefficients for these basis functions were parameterized via either the MS-CG or iter-gYBG method. We note that the present treatment of rigidly constrained bonds is not rigorously consistent with the MS-CG theory.[41] Nevertheless, we expect that this should not substantively impact the present results.

For each high resolution ensemble, we parameterized a MS-CG model by solving Equa-

tion 6.2 for the potential parameters. For the solvated alanine 12-mer, we also parameterized MS-CG models for specific regions of configuration space by solving Equation 6.2, while determining $\boldsymbol{b}^{AA}$ and $\mathbf{G}^{AA}$ from the configurations that sampled the corresponding regions of the free energy surface (FES). Additionally, as described in the Results section, we investigated the errors in the MS-CG model for the solvated alanine 12-mer by calculating the MS-CG force field according to Equation 6.2, but using a modified correlation matrix, $\mathbf{G}^{AA\text{-}mod}$, in the place of $\mathbf{G}^{AA}$.

We employed a modified version of the iter-gYBG method[69] to parameterize CG models for the solvated alanine tetramer and 12-mer. As described in our prior work, this heuristic modification, which applies only for bond and angle interactions, employs an exact decomposition of $\mathbf{G}$ into direct and indirect contributions.[230] For bond and angle interactions, we determine the direct contribution to $\mathbf{G}$ from the mapped AA ensemble and then adapt the indirect contributions to $\mathbf{G}$ based upon the CG models that are generated during the iterative procedure. This modification increased the stability and robustness of the calculations, especially for models with complex intramolecular structure.

As described in the Supporting Information, the iter-gYBG calculations for the solvated alanine 12-mer employed reference potentials[134] for the angle and dihedral interactions in order to improve the robustness of the method. We determined the reference potentials for the CG angles via direct Boltzmann inversion of the mapped AA angle distributions. We determined the reference potentials for the CG dihedrals from the MS-CG potential for the solvated alanine 12-mer. The iter-gYBG procedure determined a correction to the reference dihedral potential, which was represented by a Fourier series expansion, but not for the reference angle potential. As in our previous work,[267] we computed the reference contribution, $\boldsymbol{b}^{\mathrm{Ref}}$, to the $\boldsymbol{b}^{AA}$ vector from the mapped AA ensemble. We then iteratively solved for the CG potential parameters after replacing $\boldsymbol{b}^{AA}$ with $\delta\boldsymbol{b} = \boldsymbol{b}^{\mathrm{AA}} - \boldsymbol{b}^{\mathrm{Ref}}$ in Equation 6.4.

Our previous study[267] demonstrated that the iter-gYBG method did not always converge. Instead, after initially converging upon an accurate model after a few iterations, the method sometimes diverged to less accurate models. In the present work we find that, over the course of many iterations, the iter-gYBG method repeatedly approaches to and diverges from an accurate force field. Accordingly, we have developed several metrics to identify the optimal iter-gYBG model. The Supporting Information section reviews these criteria and briefly assesses the convergence of the iter-gYBG method for the present peptide models.

As in our prior calculations,[267] we modified the g-YBG system of linear algebraic equations in order to automate the iterative procedure, improve its robustness, and minimize user interference as previously described.[267] As described in the Supporting Information section,

these modifications included removing force field coefficients for interactions that were rarely sampled, introducing constraints to ensure periodicity of dihedral potentials, and regularizing central pair interactions to avoid overfitting statistical noise. We extensively tested these modifications to ensure that they minimally impacted the MS-CG calculation. We then solved the modified linear equations[207] via singular value decomposition[175] after applying right-left preconditioning[38,122] to render the linear equations dimensionless.

6.3.3 Structural Analysis

In addition to analyzing structural distribution functions along individual degrees of freedom, we also examined 2-D free energy surfaces (FES's) for pairs of order parameters. We considered several order parameters that are functions of the CG configuration, $\mathbf{R}$. The fractional helical content, or helicity, is defined: $\mathrm{Q}_{\mathrm{hel}}(\mathbf{R}) = \frac{1}{N_{\mathrm{hel}}} \sum_{i-j=3} \exp[-\frac{1}{2\sigma^2}(R_{ij} - R_0)^2]$, where R_{ij} is the distance between site i and j in configuration $\mathbf{R}$, N_{hel} is the number of 1-4 pairs, $R_0 = 0.5$ nm, and $\sigma^2 = 0.02$ nm^2. The radius of gyration, $\mathrm{R_g}$, is defined:[7] $\mathrm{R_g}(\mathbf{R}) = \sqrt{\sum_{i=1}^{N} \mid \mathbf{R}_i - \mathbf{R}^{\mathrm{com}} \mid^2}$, where $\mathbf{R}_i$ is the position of site i, $\mathbf{R}^{\mathrm{com}}$ is the center of mass position, and N is the total number of CG sites. The deviation from a perfectly helical configuration, $\mathbf{R}^{\mathrm{hel}}$, is characterized by the RMSD and DRMS metrics.[7] The RMSD is defined: $\mathrm{RMSD}(\mathbf{R} \mid \mathbf{R}^{\mathrm{hel}}) = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \mid \mathbf{R}_i - \mathbf{R}_i^{\mathrm{hel}} \mid^2}$, where $\mathbf{R}_i$ and $\mathbf{R}_i^{\mathrm{hel}}$ are the coordinates of site i in configurations $\mathbf{R}$ and $\mathbf{R}^{\mathrm{hel}}$, respectively. The DRMS is defined: $\mathrm{DRMS}(\mathbf{R} \mid \mathbf{R}^{\mathrm{hel}}) = \sqrt{\frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} (R_{ij} - R_{ij}^{\mathrm{hel}})^2}$, where R_{ij} and R_{ij}^{hel} are the distances between sites i and j in structures $\mathbf{R}$ and $\mathbf{R}^{\mathrm{hel}}$, respectively. The root mean square fluctuation (RMSF) of a CG site i is defined:[7] $\mathrm{RMSF}\ (i) = \sqrt{\left\langle \left| \mathbf{R}_i - \bar{\mathbf{R}}_i \right|^2 \right\rangle}$, where $\bar{\mathbf{R}}_i$ is the average position of site i and the brackets denote an ensemble average. Before computing the RMSD or RMSF, a least squares superposition was performed for each configuration, $\mathbf{R}$, with respect to the reference structure, $\mathbf{R}^{\mathrm{hel}}$.

6.4 Results

In this study, we considered four high resolution peptide models that demonstrated varying degree of complexity and helicity. We considered two high resolution models that fluctuated about well-defined helices: 1) a Gō model[252–254] for a 15-residue segment of FCP1[269] and 2) an all-atom (AA) model for a 12-residue alanine peptide in vacuum. We also considered explicitly solvated AA models of 4- and 12-residue alanine peptides, which generated more complex conformational ensembles that demonstrate helix-coil transitions. For each of these

high resolution models, we parameterized a CG model via the MS-CG method[36,41,59,60] and, in some cases, also via the iter-gYBG method.[69,75,267] In each case, the CG model represented the peptide with sites at α-carbons, while implicitly incorporating solvent effects into the potential for the peptide CG sites. We assessed the quality of each CG model by comparison with the corresponding mapped AA ensemble. We identified specific structural features of the mapped AA ensemble that are not accurately described by the minimal peptide resolution and simple molecular mechanics potential. Finally, we demonstrated that these features are the dominant source of error in the MS-CG models for disordered peptides.

6.4.1 Structured Peptides

6.4.1.1 Flexible Helices

We first employed a C-α Gō model for the FCP1 peptide in order to generate an ensemble of conformations with simple fluctuations about a helical structure. Although this Gō model represents each residue with a single site, for our purposes it is a "high resolution" model since it employs 17 distinct pair potentials, as well as distinct bond, angle, and dihedral potentials for each instance of these interactions. These potentials explicitly bias the ensemble to sample the folded peptide conformation from the corresponding PDB structure (PDBID: 1J2X).[271] In contrast, the MS-CG model treats the interactions between the 15 sites with only 3 types of pair potentials: 1) a 1-4 potential to model interactions between sites separated by exactly 3 bonds, 2) a 1-5 potential to model interactions between sites separated by exactly 4 bonds, and 3) a "nonspecific" pair potential to model interactions between all sites separated by more than 4 bonds. Additionally, the MS-CG model employs a single bond, angle, and dihedral potential to model all instances of these interactions.

The MS-CG model reproduces the ensemble sampled by the Gō model with reasonably high accuracy. The MS-CG model quantitatively reproduces the bond, angle, dihedral, 1-4, and 1-5 distributions, and qualitatively reproduces the nonspecific pair distribution, i.e., the distribution of sites separated by more than 4 bonds. (See panel a of Figure S1.) Figure 6.1 further characterizes the ensembles that are sampled by the two models for the FCP1 peptide. Panels (a) and (b) present the free energy surface (FES) as a function of helicity, Q_{hel}, and the radius of gyration, R_g, for the Gō and MS-CG models, respectively. In comparison to the high resolution model, the MS-CG model slightly overestimates the stability of bent helices with slightly reduced R_g and Q_{hel}. Nevertheless, the MS-CG model quantitatively reproduces the average helicity, $\langle Q_{hel} \rangle = 0.65$, of the high resolution ensemble and, more generally, reproduces the corresponding FES with reasonably high accuracy.

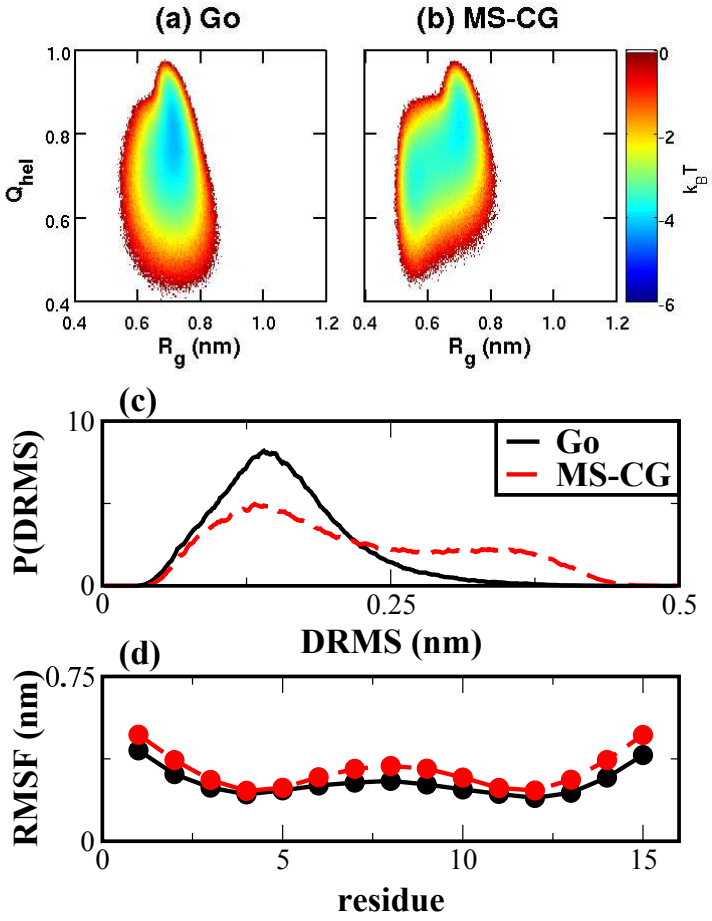

Figure 6.1. Comparison of Gō and MS-CG models for the FCP1 peptide. Panels (a) and (b) present the FES as a function of helicity, Q_{hel}, and the radius of gyration, R_g, for the Gō and MS-CG models, respectively. Panels (c) and (d) present the DRMS distribution and the RMSF of each residue, respectively, for the Gō (solid, black curves) and MS-CG models (dashed, red curves).

Panels (c) and (d) of Figure 6.1 compare the distributions that are sampled by the Gō and MS-CG models for the DRMS metric, which describes deviations from the FCP1 crystal structure, and for the RMSF metric, which quantifies the fluctuations of each residue. The MS-CG model qualitatively reproduces the DRMS distribution of the Gō model. Because the MS-CG model tends to sample bent helices, the corresponding DRMS distribution is slightly broader and demonstrates a second peak. The MS-CG model also slightly overestimates the fluctuations in the middle and the termini of the helix, but still reproduces the RMSF of the Gō model with reasonable accuracy.

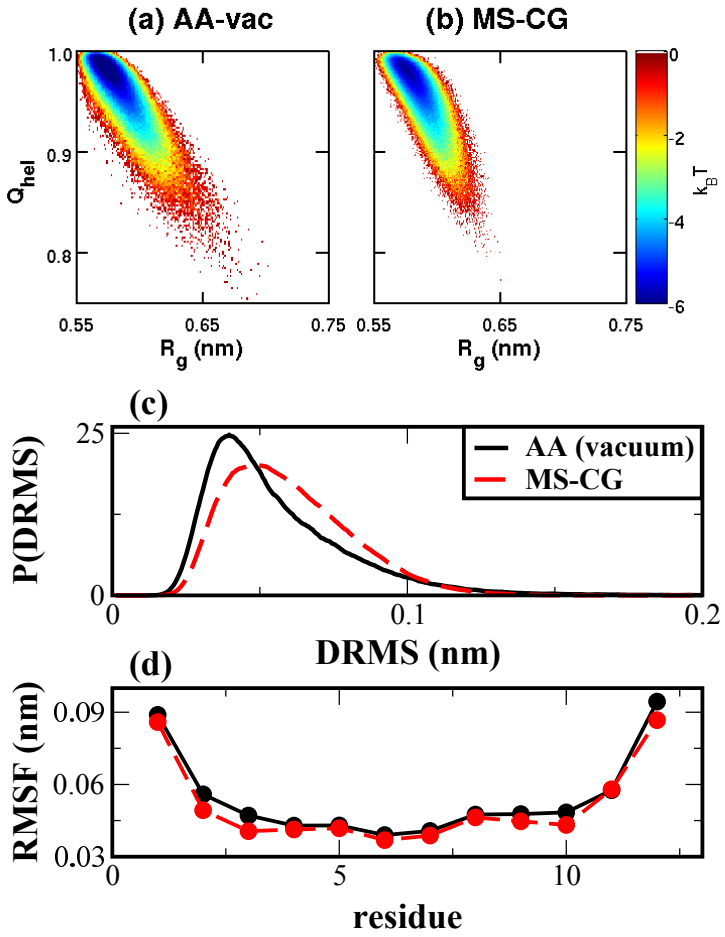

Figure 6.2. Comparison of AA and MS-CG models for the alanine 12-mer in vacuum. Panels (a) and (b) present the FES as a function of helicity, Q_{hel}, and the radius of gyration, R_g, for the AA and MS-CG models, respectively. Panels (c) and (d) present the DRMS distribution and the RMSF of each residue, respectively, for the AA (solid, black curves) and MS-CG models (dashed, red curves).

6.4.1.2 Precise Helices

In order to generate a second high resolution ensemble with well-defined helical structure, we performed simulations with the OPLS-AA model[18] for a capped 12-residue alanine peptide in vacuum. In comparison to the Gō-model, which sampled a rather "floppy" helix, this AA model sampled a very precise helix that only exhibited slight unfolding from the peptide termini. As in the preceding case, the MS-CG model included distinct potentials to model 1-4 and 1-5 interactions, as well as a nonspecific pair potential between sites separated by more than 4 bonds. Because the interior residues sampled very precise helical conformations, while the terminal residues sampled less rigid helical conformations, we found it necessary to employ distinct site types and corresponding interactions for the two terminal sites on

each end of the peptide. Thus, the MS-CG model for the alanine 12-mer in vacuum treated 5-distinct site types with 11 distinct pair potentials. When we did not distinguish between the interior and terminal sites, the resulting MS-CG model underestimated the helicity of the interior residues and overestimated the helicity of the termini.

The MS-CG model quantitatively reproduces the bond, angle, dihedral, 1-4, and 1-5 distributions of the mapped AA ensemble, and also qualitatively reproduces the corresponding nonspecific pair distribution. (See panel b of Figure S1.) In correspondence with Figure 6.1, Figure 6.2 employs the same metrics to compare the ensembles generated by the AA and MS-CG models for the alanine 12-mer in vacuum. Clearly, the MS-CG model provides a very accurate description of the mapped AA ensemble.

6.4.2 Helix-coil Transition for a Single Peptide Unit

As demonstrated by the preceding calculations, the MS-CG method determines accurate minimal models for peptides that fluctuate about a well-defined helical structure. We next considered whether bottom-up coarse-graining methods can determine minimal peptide models that accurately describe the helix-coil transition of a high resolution AA model. We constructed a high resolution ensemble for the helix-coil transition by simulating the OPLS-AA model for a capped 4-residue alanine peptide in explicit SPC/E solvent.[205] For this ensemble, we parameterized CG models with 4 sites that correspond to the α-carbons of the AA model and modeled the CG interactions with 3 equivalent bond potentials, 2 equivalent angle potentials, 1 dihedral potential, and a single 1-4 pair potential. Because the MS-CG model provided limited accuracy for this ensemble, we employed the iter-gYBG method to investigate the source of this discrepancy.

Panels a, b, and c of Figure 6.3 present the simulated distributions for the two angles, θ, the one dihedral angle, ψ, and the 1-4 distance, r_{1-4}, between the first and last CG sites, respectively. These degrees of freedom all sample multimodal distributions in the mapped AA ensemble (solid black curves). In particular, helical conformations correspond to the peaks at $\theta \approx 91$ deg, $\psi \approx 57$ deg, and $r_{1-4} \approx 0.53$ nm, while extended configurations correspond to $\theta \approx 122$ deg, $\psi \approx$ -120 deg, and $r_{1-4} \approx 0.93$ nm. The MS-CG model (dashed red curves) qualitatively reproduces these distributions, while the iter-gYBG model (dashed-dotted green curves) reproduces each distribution with nearly quantitative accuracy. The slight errors in the r_{1-4} and ψ distributions of the iter-gYBG model appear to result from a competition between accurately modeling the two interactions simultaneously. This may result either from fundamental limitations of the minimal peptide representation or, alternatively, from

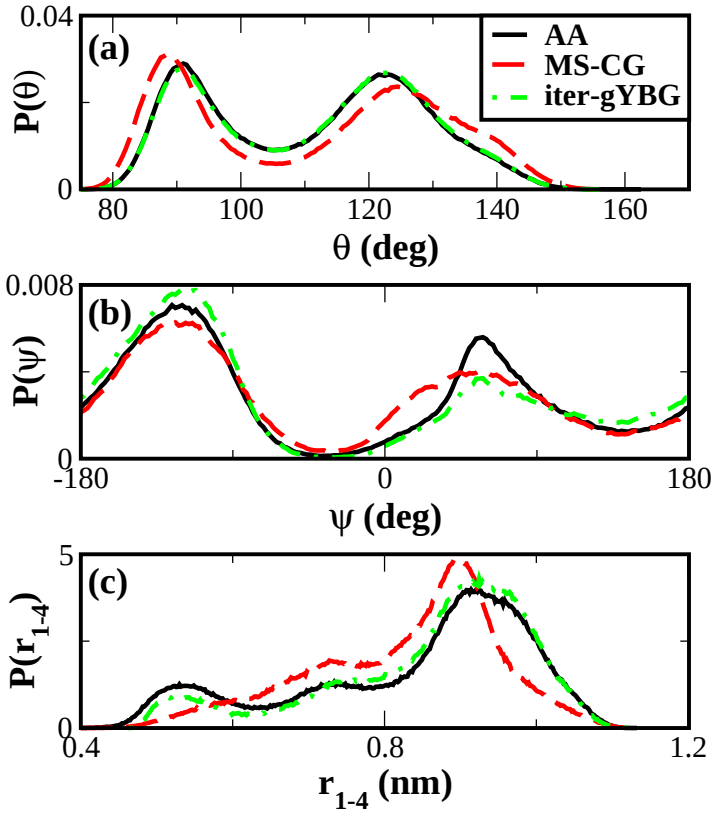

Figure 6.3. Comparison of AA, MS-CG, and iter-gYBG models for the solvated alanine tetramer. Panels (a), (b), and (c) present the 1-D distributions sampled along the CG angle, dihedral, and 1-4 degrees of freedom, respectively. The solid black, dashed red, and dashed-dotted green curves correspond to distributions sampled by the AA, MS-CG, and iter-gYBG models, respectively.

the insensitivity of the iter-gYBG procedure. (See Figure S2 of the Supporting Information section.)

Figure 6.4 presents the FES's that are sampled by the three models as a function of r_{1-4} and ψ. The AA model (panel a) samples helical (H), extended (E1 and E2), and intermediate (I) conformations. The MS-CG model (panel b) samples the I, E1, and E2 regions, albeit with incorrect propensities, but fails to significantly sample the H region. The iter-gYBG model (panel c) samples the FES with considerably greater accuracy, which is expected since it is explicitly parameterized to reproduce force correlation functions that are related to the AA distributions along r_{1-4} and ψ.

However, both the MS-CG and iter-gYBG models sample regions of the FES that were not sampled by the AA model, including extensions of the E2 region and also an intermediate/extended (IE) region. As discussed above, the MS-CG method determines potentials based upon the assumption that the chosen form of the CG potential is capable of reproducing the cross-correlations of the AA model. The IE region that is "forbidden" from the AA

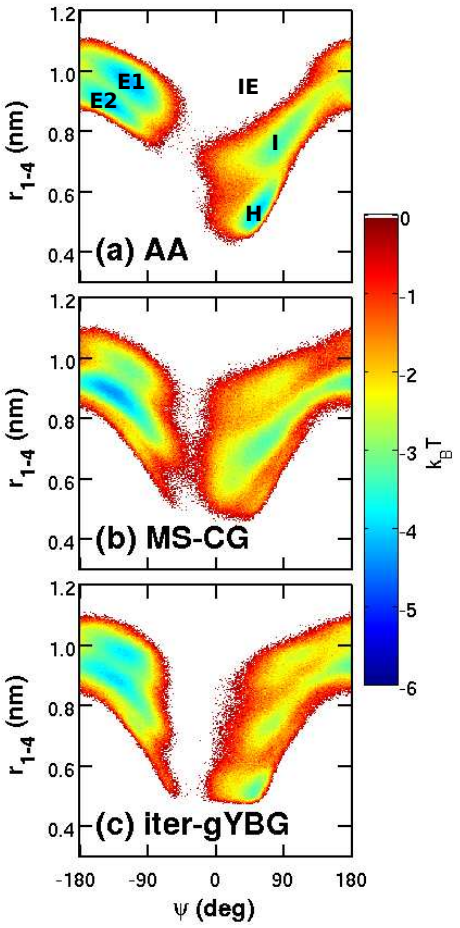

Figure 6.4. FES's as a function of 1-4 distance and dihedral angle for the solvated alanine tetramer. Panels (a), (b), and (c) present results for the AA, MS-CG, and iter-gYBG models, respectively. In panel (a), the labels identify helix (H), intermediate (I), extended (E1/E2), and intermediate-extended (IE) regions of configuration space.

model demonstrates an instance that this assumption fails. This region corresponds to values for r_{1-4} that are sampled in extended conformations and values for ψ that are sampled in helical conformations. Because the atomic geometry of the peptide backbone precludes this combination of r_{1-4} and ψ, the AA model transitions from extended to helical conformations via the I region and not through the IE region. In contrast, the simple molecular mechanics potential and minimal peptide resolution of the CG model do not adequately couple these two degrees of freedom. Thus, in order to sample both the H and E2 regions of the FES, the CG model also samples the IE region. Consequently, the MS-CG model provides a relatively poor description of the 1-D distributions of the mapped AA ensemble because $\mathbf{G}^{AA}$ provides a relatively poor approximation to the cross-correlations generated by the CG model. Moreover, iterative methods, such as the iter-gYBG method, which quite accurately reproduce the 1-D distributions of the mapped ensemble, do so by distorting the cross-correlations of

the mapped AA ensemble and, in particular, sampling the forbidden IE region of the FES. Figure S3 of the Supporting Information section explicitly compares the cross-correlations generated by the AA and iter-gYBG models and demonstrates their effect on the resulting CG potentials.

6.4.3 Disordered Peptide Ensemble

We next considered whether bottom-up CG methods can accurately describe the ensemble sampled by the OPLS-AA model for a capped 12-residue alanine peptide in explicit SPC/E solvent. This high resolution model sampled a very heterogeneous ensemble that included helical, coil, and extended structures. For this high resolution ensemble, we parameterized 12-site CG models via both the MS-CG and iter-gYBG methods. As for the 12-residue alanine peptide in vacuum, we distinguished between the interior sites and the two sites at each terminus of the peptide, while employing a total of 11 distinct pair potentials to model the interactions between sites that are separated by exactly 3, exactly 4, and more than 4 bonds.

Quite recently, Carmichael and Shell[229] also employed a bottom-up method to parameterize CG models from a simulation of an atomically-detailed model for a 15-residue alanine peptide with implicit solvent. In particular, they parameterized CG models with 1-, 2-, and 3-sites per residue by minimizing the relative entropy[37,121] with respect to the atomically-detailed ensemble. In order to make direct comparison with their study, Figure 6.5 presents simulated FES's in terms of the order parameters that they considered, i.e., the root mean square displacement (RMSD) from a perfect helical conformation and the radius of gyration, R_g.

Figure 6.5a demonstrates that our high resolution peptide model, i.e., the OPLS-AA model in explicit SPC/E solvent, samples a rather diffuse FES with several shallow minima, including a metastable minima for a nearly perfect helical structure. Figure 6.5a suggests that this explicit solvent model samples a more complex ensemble and demonstrates greater helical tendency than the implicit solvent model considered by Carmichael and Shell. Figure 6.5b demonstrates that the MS-CG minimal model samples a range of collapsed and extended structures with little tendency for helical structures. Interestingly, this MS-CG model appears to sample a similar ensemble to the minimal model parameterized by Carmichael and Shell, although it is important to recognize that the MS-CG model employed a more complex potential function and was parameterized for a mapped AA ensemble with greater complexity. Finally, Figure 6.5c demonstrates that the iter-gYBG model samples an ensem-

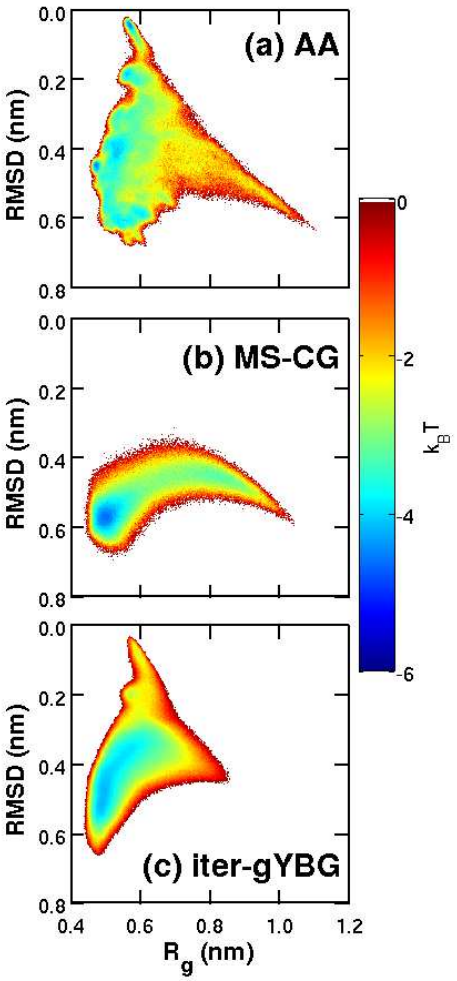

Figure 6.5. FES's as a function of RMSD and R_g for the solvated alanine 12-mer. Panels (a), (b), and (c) present results for the AA, MS-CG, and iter-gYBG models, respectively.

ble that is much more similar to the mapped AA ensemble. In particular, the iter-gYBG model samples helical conformations with much greater tendency than the MS-CG model, although with slightly less tendency than the AA model. Moreover, the iter-gYBG model clearly smooths over the fine structure of the AA FES and underestimates the stability of extended conformations.

In order to more closely compare the three models and, in particular, the helix-coil transitions that they sample, we next analyzed the ensembles as a function of Q_{hel}, and R_g. Figure 6.6 presents the corresponding FES for the explicitly solvated AA model. The AA FES demonstrates basins that correspond to helical and coil conformations, as well as a "tail" of extended conformations. Figure 6.6 presents the average structure sampled by the AA model in each of these regions of the FES, as well as representative structures that demonstrate the fluctuations sampled about these average structures. In addition, Figure 6.6

also reveals horizontal bands in the AA FES that correspond to metastable intermediates that form as the AA model transitions from coil conformations to fully helical conformations. We partitioned this FES into eight regions for subsequent analysis.

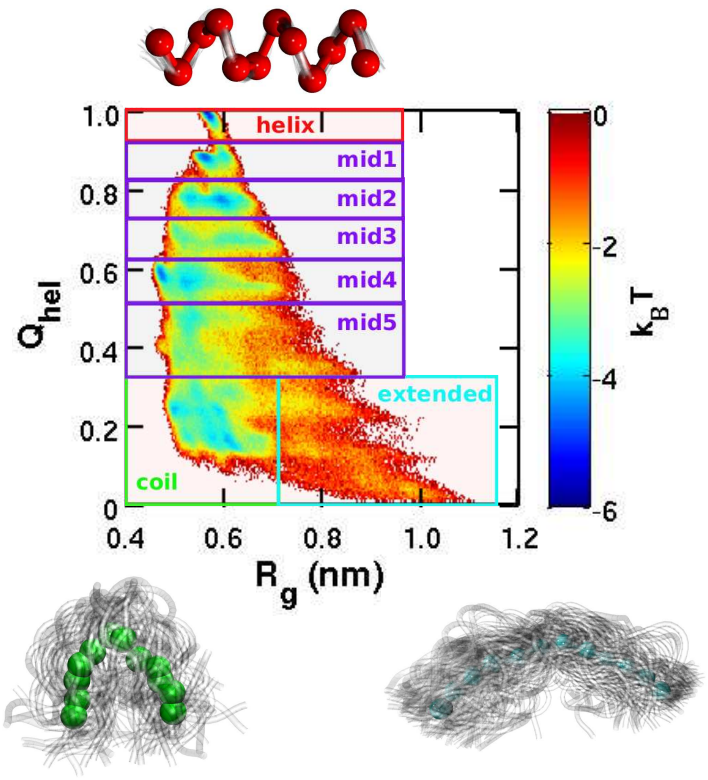

Figure 6.6. FES as a function of Q_{hel} and R_g sampled by the AA model for the solvated alanine 12-mer. The rectangles indicate 8 partitions of this FES. The opaque configurations present the average structures sampled in the helix, coil, and extended regions of the FES. The superimposed transparent images present traces of the peptide backbone for representative conformations sampled in these regions.

Panel a of Figure 6.7 presents the FES sampled by the MS-CG model for the solvated alanine 12-mer as a function of Q_{hel} and R_g. Figure 6.7a demonstrates that the MS-CG model samples coil and extended conformations, but does not sample conformations with $Q_{hel} > 0.5$. We considered several possible causes for the discrepancies between the MS-CG and AA ensembles, including: 1) the MS-CG method cannot accurately describe the helical structures that are sampled by the AA model for the solvated alanine 12-mer; 2) given the minimal peptide representation and simple molecular mechanics form, there does not exist a single potential that can sample the entire range of structures in the mapped AA ensemble; or 3) the minimal peptide representation and simple molecular mechanics potential cannot reproduce the cross-correlations of the mapped AA ensemble that characterize the transitions between different regions of the FES.

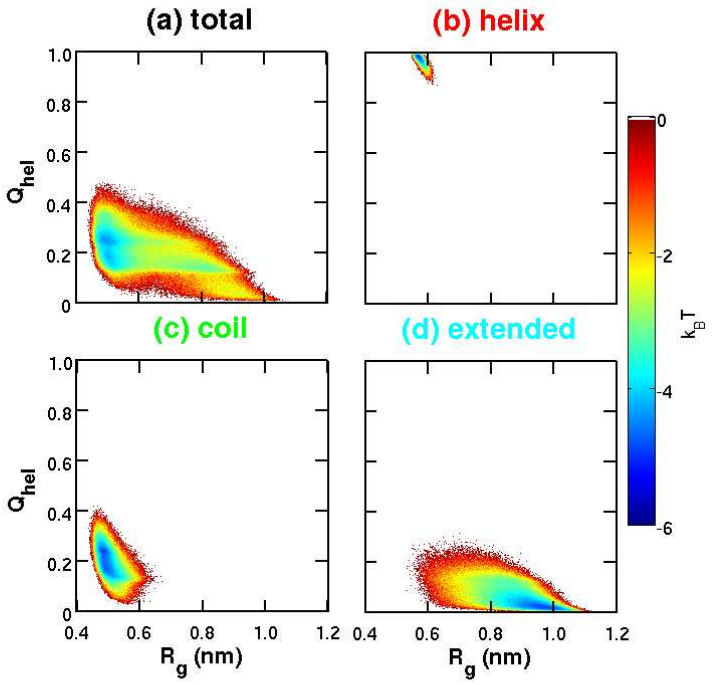

Figure 6.7. FES's as a function of Q_{hel} and R_g sampled by various CG models for the solvated alanine 12-mer. Panel (a) corresponds to the MS-CG model. Panels (b), (c), (d) correspond to models parameterized via the MS-CG method for the helix, coil, and extended regions, respectively, of the FES.

Accordingly, we employed the MS-CG method to determine CG models that were specific to the helical, coil, and extended regions of the FES. Panels b, c, and d of Figure 6.7 present the FES's sampled by these MS-CG models and demonstrate that they accurately sample the corresponding regions of configuration space. These results are consistent with the results in Figures 6.1 and 6.2 for MS-CG models of well-defined helices. Thus, minimal MS-CG models appear capable of reasonably reproducing the structure and cross-correlations within each region of configuration space that is characterized by a well-defined average structure.

Figure 6.8 presents the FES sampled by the iter-gYBG model for the solvated alanine 12-mer as a function of Q_{hel} and R_g. The iter-gYBG model reasonably reproduces the AA FES in Figure 6.6. In particular, the FES's for the iter-gYBG and AA models demonstrate similar bands in the helix-coil transition region of the FES, although the iter-gYBG FES does not demonstrate the same minima in this region. Moreover, the iter-gYBG model samples helical conformations with similar weight to the AA model. However, the iter-gYBG model overestimates the stability of the mid3, mid4, and mid5 regions of the FES, which correspond to the helix-coil transition, while underestimating the stability of the extended and coil regions of the FES. Nevertheless, Figure 6.8 suggests that a minimal model with a molecular mechanics potential is capable of sampling the stable conformations in the heterogeneous AA

ensemble. Thus, the discrepancies between the AA and MS-CG ensembles appear to stem primarily from the inability of the MS-CG method to accurately describe the transitions between the different regions of conformational space.

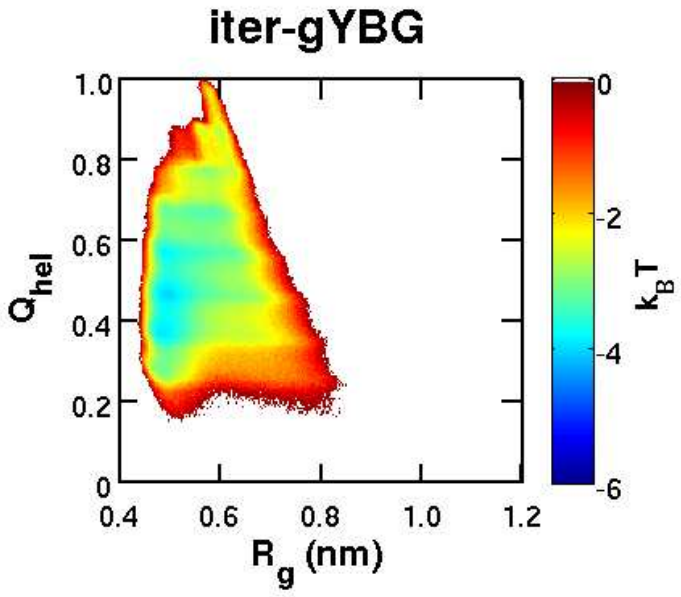

Figure 6.8. FES as a function of Q_{hel} and R_g sampled by the iter-gYBG model for the solvated alanine 12-mer.

As discussed above, the MS-CG method assumes that the CG model is capable of reproducing the cross-correlations $\mathbf{G}^{AA}$ in the mapped AA ensemble. In contrast, iterative methods, such as the iter-gYBG method, more accurately reproduce the mapped AA distributions along individual degrees of freedom by accounting for the limited ability of the CG model to reproduce the corresponding cross-correlations. Thus, by comparing the cross-correlations sampled by the AA and iter-gYBG models, we can identify specific features of the AA ensemble that cause discrepancies between the AA and MS-CG models. Our analysis of these cross-correlations indicated, perhaps unsurprisingly, that the minimal peptide representation and simple molecular mechanics potential are incapable of reproducing the cooperativity of the helix-coil transition in the AA model. In particular, we observed four interrelated manifestations of this limitation: 1) The CG model does not reproduce the complex correlations between the CG angle and the other intramolecular degrees of freedom that arise during this transition. 2) The AA and iter-gYBG models sample considerably different structural correlations in the transition regions of the FES (i.e., the mid3, mid4, and mid5 regions) in Figure 6.6. 3) The iter-gYBG model samples the helix region of the FES with reasonable weight, but significantly oversamples the transition regions and undersamples the coil regions of the FES. 4) The iter-gYBG model samples helices that are slightly less precise

than the helices of the AA model.

We hypothesized that the errors in the MS-CG model largely stem from these cross-correlations in the mapped AA ensemble that cannot be reproduced by the minimal CG model. We numerically tested this hypothesis by modifying the matrix, $\mathbf{G}^{AA}$, of cross-correlations in order to account for the limitations of the minimal peptide representation and simple molecular mechanics potential. To perform these modifications, we decomposed $\mathbf{G}^{AA}$ into distinct contributions from each region ν of the FES: $\mathbf{G}^{AA} = \sum_{\nu} w_{\nu} \mathbf{G}^{AA}_{\nu}$, where ν identifies a particular region of the FES in Figure 6.6, $\mathbf{G}^{AA}_{\nu}$ indicates the matrix of cross-correlations that is calculated from the configurations that map to region ν, and w_{ν} is the probability for the AA model to sample region ν. For each of the 4 discrepancies identified above, we made corresponding modifications to generate a new matrix of cross-correlations, $\mathbf{G}^{AA\text{-}mod}$:

1. Because the CG model appears incapable of reproducing the cross-correlations of the CG angle with the other intramolecular degrees of freedom, we treated the angle interactions independently of the remaining interactions in a manner similar to the "hybrid force-matching" procedure suggested by Rühle and Junghans.[274]

2. Because the CG model appears incapable of reproducing the cross-correlations sampled by the AA model in the mid3, mid4, and mid5 regions of the FES that correspond to the transition between coil and helical conformations, we replaced the contributions to the $\mathbf{G}$ matrix from these regions by linearly interpolating between the mid2 and coil regions, i.e., $\mathbf{G}^{AA\text{-}mod}_{\nu} = \gamma_{\nu} \mathbf{G}^{AA}_{\text{mid2}} + (1 - \gamma_{\nu}) \mathbf{G}^{AA}_{\text{coil}}$ for $\nu = \text{mid3}, \text{mid4}, \text{mid5}$, where γ_{ν} is determined by linearly interpolating the helicity Q_{hel} that is sampled at the center of region ν.

3. Because the CG model tends to overestimate the stability of the mid5 region and underestimate the stability of the coil region of the AA FES, we replaced the atomistic weights for these regions with weights more closely resembling those sampled by the iter-gYBG model.

4. Because the CG model appears incapable of reproducing the precise structural features of the AA helix, we "blurred" the corresponding contributions to the $\mathbf{G}$ matrix.[275]

The Supporting Information section (Figures S9-S12) describes these modifications in much greater detail.

We then determined a modified MS-CG model by solving the system of MS-CG normal equations for the potential parameters after replacing $\mathbf{G}^{AA}$ with the $\mathbf{G}^{AA\text{-}mod}$. In comparison

to the matrix of cross-correlations, $\mathbf{G}^{AA}$, that is determined from the mapped AA ensemble, $\mathbf{G}^{AA\text{-}mod}$ provides a much more accurate description of the cross-correlations that can be generated by the CG model. Thus, if these structural features of the mapped AA ensemble are the major sources of error in the MS-CG model, the modified MS-CG model should provide a significantly improved description of the AA FES. Figure 6.9 presents the FES's sampled by the modified MS-CG model as a function of both RMSD and R_g (top) and also Q_{hel} and R_g (bottom). Indeed, after making these modifications to $\mathbf{G}$, the MS-CG equations determine a model that describes the AA ensemble with much greater accuracy than the original MS-CG model. This strongly supports our hypothesis regarding the errors in the MS-CG model.

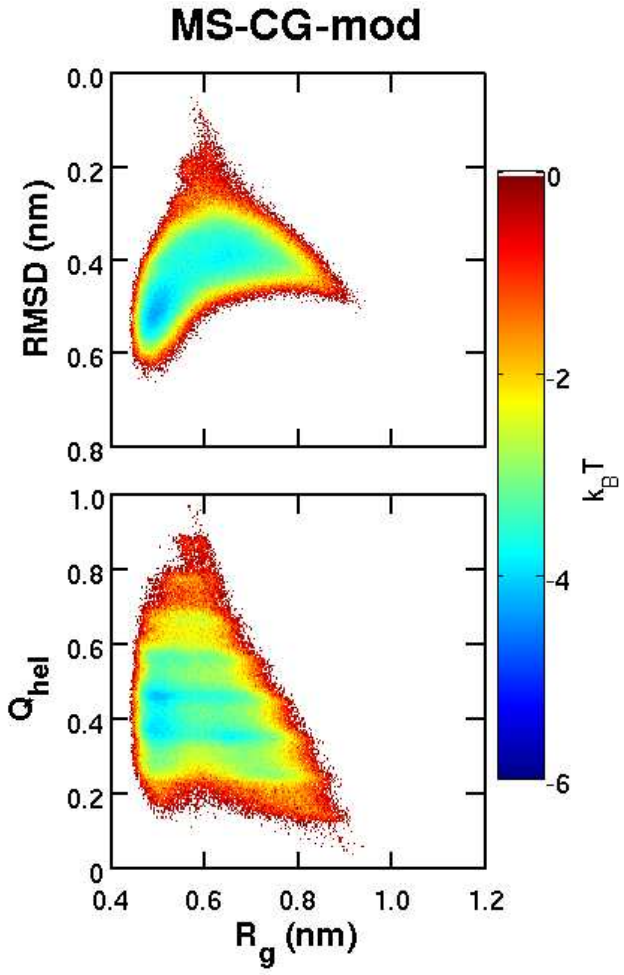

Figure 6.9. FES's sampled by the "modified" MS-CG model for the solvated alanine 12-mer. The top panel presents the FES as a function of RMSD and R_g; the bottom panel presents the FES as a function of Q_{hel} and R_g.

6.5 Discussion

The present work investigates the potential and limitations of bottom-up coarse-graining methods for accurately modeling the ensembles of structures that are sampled by high resolution peptide models. In particular, we considered several high resolution peptide models that sampled ensembles with varying degrees of complexity and helicity. For each high resolution ensemble, we employed the MS-CG method to parameterize an implicit solvent, minimal resolution CG model that represented each residue with a single site and incorporated solvent effects into the interactions between the peptide sites. In contrast to other bottom-up methods, the MS-CG method does not require simulations with multiple CG models in order to parameterize the CG potential. Rather, it employs the g-YBG relation to directly determine the CG potential from the high resolution ensemble. However, this calculation assumes that the CG model will reproduce the cross-correlations that are present in the mapped AA ensemble. Since these cross-correlations may be essential for describing hierarchical protein structures, this study assessed the validity of the MS-CG assumption in the particular case that the CG model employs a minimal peptide representation and a simple molecular mechanics potential.

The MS-CG method determined accurate minimal models for two different high resolution peptide ensembles that fluctuated about well-defined helices. The MS-CG model very accurately reproduced the ensemble of floppy helices sampled by a simple Gō model. The MS-CG model also very accurately reproduced the ensemble of precise helices sampled by an AA model for an alanine 12-mer in vacuum. In this latter case, the MS-CG model required distinct site types for the terminal residues in order to reproduce the different helicities of the terminal and interior residues.

As a third high resolution peptide model, we considered an explicitly solvated AA model for the alanine 12-mer. The solvated high resolution model sampled a much more complex ensemble that included helical, coil, and extended structures, as well as various intermediate structures. Interestingly, when we extracted subensembles with well defined average structures, the MS-CG method proved capable of determining models that accurately described helical, coil, or extended regions of conformational space. However, given the minimal representation and the simple molecular mechanics potential, the MS-CG method did not succeed in determining a single potential for sampling the entire conformational space with appropriate weight. The pioneering studies of Voth and coworkers[76–78] demonstrated similar challenges for determining transferable MS-CG models that accurately described both helical and turn conformations with a slightly higher resolution peptide model and an explicit CG

solvent model.

It is interesting to compare these results with the recent studies of Carmichael and Shell,[229] who employed a clever iterative nonlinear optimization method to parameterize implicit solvent CG peptide models that minimized the relative entropy at various levels of resolution. Figure 6.5 suggests that the present study considered a high resolution ensemble with somewhat greater complexity than the high resolution model considered by Carmichael and Shell. Interestingly, though, Figure 6.5 also suggests that the MS-CG model, which we directly determined from the high resolution ensemble, appears to sample the same regions of conformational space as the corresponding minimal model that was obtained by iteratively minimizing the relative entropy. This minimal peptide model does not adequately sample helical conformations and, instead, too frequently samples collapsed coil structures.

We employed an iterative g-YBG (iter-gYBG) method[69,75,267] to parameterize an implicit solvent, minimal CG model that more accurately reproduced the AA ensemble for the explicitly solvated alanine 12-mer. The iter-gYBG method is quite similar to other iterative bottom-up methods, since it employs multiple CG simulations in order to determine potentials that reproduce target 1-D correlation functions of the mapped AA ensemble. The resulting iter-gYBG model reasonably reproduces the 1-D distributions of the mapped AA ensemble, but does so at the expense of distorting the cross-correlations between these degrees of freedom.

Despite these distortions, the iter-gYBG model samples the various regions of the FES with much greater accuracy than the MS-CG model. The similarity of the AA and iter-gYBG ensembles is perhaps somewhat surprising, since the iter-gYBG model is parameterized to reproduce a set of 1-D correlation functions that reflect weighted averages over the various conformations in the heterogeneous AA ensemble. Since different ensembles might possibly give rise to very similar 1-D correlation functions, one might expect that iterative structure-based CG methods may be quite insensitive to the underlying distribution of conformations. In fact, the distributions of the mapped AA ensemble appear quite similar to the distributions sampled by intermediate states in the helix-coil transition and, indeed, the iter-gYBG model demonstrates too great a tendency for sampling this region of conformational space. Nevertheless, the iter-gYBG model provides a reasonably accurate description of the entire AA FES. Thus, while it may sometimes prove useful to employ different potentials for sampling different types of helices,[276] this study provides evidence that bottom-up structure-based methods can determine a single potential that reasonably samples the complex ensemble sampled by an AA peptide model.

By comparing the ensembles sampled by the AA and iter-gYBG peptide models, we iden-

tified the cross-correlations that cannot be reproduced by the CG model and, thus, invalidate the fundamental MS-CG assumption. This analysis suggested that the errors in the MS-CG model for the solvated alanine 12-mer stem primarily from 4 particular manifestations of the cooperative interactions in the AA model: 1) The minimal representation and simple molecular mechanics potential appear incapable of reproducing the cross-correlations that involve the CG angle. In fact, the MS-CG method determines a more accurate peptide model if these cross-correlations are systematically neglected, similar to the "hybrid force-matching" method of Rühle and Junghans.[274] 2) The CG model appears incapable of reproducing the precise cross-correlations that the AA model samples during the helix-coil transition, as indicated by our calculations for the solvated alanine tetramer. 3) The CG potentials that are determined via the g-YBG equation appear to systematically overestimate the stability of transition structures that arise during the helix-coil transition. 4) Minimal models may be incapable of reproducing the precise cross-correlations that result from the cooperative interactions that stabilize helices in AA models. Significantly, we were able to explicitly prove that these four considerations significantly limit the accuracy of the MS-CG model. By implementing corresponding modifications to the MS-CG cross correlation matrix, $\mathbf{G}^{AA}$, the modified MS-CG model described the mapped AA ensemble much more accurately.

The cooperative interactions that stabilize helices in the AA model generate sharp features in the $\mathbf{G}^{AA}$ cross-correlation matrix. Since the simple molecular mechanics potential and minimal CG representation do not provide sufficient coupling to reproduce these precise cross-correlations, the sharp features in $\mathbf{G}^{AA}$ determine relatively weak MS-CG forces, $\boldsymbol{\phi}^0 = \left(\mathbf{G}^{AA}\right)^{-1} \boldsymbol{b}^{AA}$, that do not adequately stabilize helical structures. This suggests an amusing "reciprocal" relation for the MS-CG method: since $\boldsymbol{b}^{AA}$ describes the structural features of the AA helix, it follows that weakening the corresponding features of $\mathbf{G}^{AA}$ leads to a stronger MS-CG forces $\boldsymbol{\phi}^0$ that more accurately stabilize helices. Thus, given the limitations of the minimal resolution and the molecular mechanics potential, the iter-gYBG model appears to stabilize helices by replacing the relatively weak, but cooperative interactions of the AA model with stronger, but less cooperative interactions.

More generally, our calculations suggest that minimal peptide models with molecular mechanics potentials may not accurately describe the helix-coil transition that is sampled by atomically-detailed models. Although the iter-gYBG model for the alanine tetramer accurately reproduces the AA distributions along each relevant degree of freedom and samples both helical and extended conformations with reasonable weight, Figure 6.4 demonstrates that this model can sample the helix-coil transition via a mechanism that is excluded by the atomic structure of the peptide backbone. Similarly, although the iter-gYBG and AA models

for the solvated alanine 12-mer sample conformations of similar size (i.e., R_g) when transitioning between helical and coil structures, the actual transition structures demonstrate considerably different cross-correlations in the two models. It is possible that alternative parameterization methods may determine minimal models that more accurately describe this transition. However, it seems more likely that the minimal representation and the simple molecular mechanics potential may be fundamentally incapable of reproducing the cooperative interactions that result from an atomically-detailed peptide backbone and explicit hydrogen bonding interactions.[228,277] These considerations strongly motivate the development of CG potentials that couple, in particular, angle and torsion degrees of freedom.[247] Moreover, they echo previous conclusions that cooperative interactions are essential for protein folding and thermodynamics.[278–280]

Finally, this work further demonstrated the utility of the iter-gYBG method for modeling complex molecular systems. As in previous work,[267] we employed a heuristic approach for robustly treating intramolecular interactions. As documented in the Supporting Information, although the iter-gYBG method determines reasonably accurate potentials after relatively few iterations, it demonstrates suboptimal convergence properties and often diverges away from, or oscillates about, these accurate potentials. This clearly motivates subsequent work to address limitations of the iter-gYBG method. Nevertheless, the iter-gYBG method provides a convenient mechanism for 1) efficiently sampling the space of CG models, 2) determining reasonably accurate models for the structure of complex molecular systems, 3) identifying structural features that limit the accuracy of the MS-CG method, and 4) revealing necessary improvements to the CG representation and interaction set.

6.6 Conclusion

This work investigated the potential and limitations of bottom-up methods for developing minimal, implicit solvent peptide models from AA simulations. The MS-CG method determines models that quite accurately reproduce the ensembles generated by high resolution models for flexible helices, rigid helices, coils, and extended structures. However, when the AA models sampled multiple distinct conformations, the MS-CG models did not correctly weight the various regions of configuration space. We demonstrated that these discrepancies in the MS-CG models for disordered peptides primarily result from the inability of the minimal peptide resolution and simple molecular mechanics potential to reproduce the complex cross-correlations that arise in the cooperative helix-coil transition of the AA model. Consequently, the MS-CG model can be improved by smoothing over the cross-correlations in

the mapped AA ensemble, e.g., by neglecting cross-correlations involving the angle degree of freedom or by blurring the cross-correlations that arise in precise helical conformations.

We also demonstrated that the iter-gYBG method reasonably reproduces the 1-D distributions of the mapped AA ensemble for the disordered peptide. Moreover, although it clearly averages over the fine details of the AA FES and overestimates the stability of intermediate transition structures, the iter-gYBG method determines a CG potential that reasonably reproduces the FES of the AA model. This is somewhat surprising since the 1-D distributions may provide relatively little information regarding either the distribution of conformations in the ensemble or the global structure of the peptide.

More generally, these calculations demonstrate that CG models with minimal resolution and simple molecular mechanics potentials are unlikely to provide an accurate description of the cooperative helix-coil transition that occurs in AA models. In particular, minimal CG models appear to stabilize helical conformations by replacing the weak, but cooperative interactions of the AA model with stronger, but less cooperative interactions. These results provide insight into the potential of bottom-up methods for accurately modeling biomolecular structure and motivate future investigations of the relationship between the CG representation, the complexity of the CG potential, and the accuracy of the CG model.

Supporting Information Available

The Supporting Information provides a detailed description of all methods, simulations, and calculations employed in this work, as well as additional analysis of these calculations. This information is free of charge via the Internet at `http://pubs.acs.org`.

Chapter 7

Conclusions and Outlook

In this work we presented a number of developments in the theory and methodology of structure-based bottom-up coarse-graining (CG). In particular, we investigated the utility of the multiscale coarse-graining (MS-CG) and generalized Yvon-Born-Green (g-YBG) methods for modeling complex condensed-phase systems, aiming to elucidate the fundamental approximations and limitations of the methods. We made connections to other CG methods and drew more general conclusions from our work whenever possible. These investigations were directly built upon the foundational work of Voth and coworkers[36,41,59,60,62,68,71,133,157,203] and of Mullinax and Noid[63,64,134,146,150] in developing the MS-CG and g-YBG methods, respectively. Along the way, we were also significantly motivated and influenced by the quickly evolving CG literature.[37,69,74,75,229,274] Despite significant advances in the field, there are a number of open problems which, in our opinion, severely hinder structure-based bottom-up CG models from achieving their full potential. With these considerations in mind, this chapter will summarize the present work, while identifying other, closely related, developments and also highlighting more general, outstanding problems.

7.1 Overview

Bottom-up CG methods parameterize the CG model directly from simulations of a higher resolution, e.g., all-atom (AA), model. For methods aiming to reproduce the structure of the underlying model, the many-body potential of mean force (PMF) is the appropriate potential for simulating the CG model such that all structural features of the mapped AA ensemble (i.e., the ensemble generated by the AA model and then mapped to the CG representation) are exactly reproduced. In general, the PMF cannot be explicitly represented or simulated, because it is a high dimensional function that depends upon the coordinates of

all CG sites in the system as well as the thermodynamic state (i.e., temperature, pressure, chemical identity, etc.) of the underlying model. Consequently, structure-based bottom-up CG methods typically attempt to approximate the PMF with a potential energy function of a relatively simple form that can be easily simulated in standard molecular dynamics (MD) software packages.

The most widely used structure-based bottom-up CG approaches make no direct reference to the PMF, but implicitly approximate it by iteratively tuning the CG potentials to reproduce a set of low order distribution functions (i.e., distributions that only depend on the relative positions of a small number of CG sites, e.g., radial distribution functions) of the mapped AA ensemble. Methods of this type differ in how they update the CG potentials at each step, e.g., by ignoring correlations[35] or by employing a linear response formalism.[34] These iterative methods are conceptually simple and have been successfully employed to build reasonably accurate models for investigating a wide range of systems.[38,49–58] Despite their success, the convergence properties of these methods are not well characterized. Additionally, since these methods solve a nonlinear optimization problem, they may be of limited use as the complexity and number of parameters of the model increases.

A different class of methods for developing structure-based bottom-up CG models employ variational principles to determine a more direct approximation to the many-body PMF. Shell[37] proposed the Relative Entropy (RE) method, which parameterizes the CG model by minimizing an information function that measures the difference between the CG and mapped AA probability distributions. The RE method is also iterative and becomes equivalent to the Inverse Monte Carlo method[34] under certain conditions. Despite initial concerns about efficiency, advancements[121,281] in implementation have led to the development of RE models for studying complex molecular systems, e.g., peptide aggregation.[229]

Izvekov and Voth[36,59] proposed a different variational method for developing CG models based on a force-matching approach.[158] This Multiscale Coarse-graining (MS-CG) method is distinct from the methods described above in that it directly (i.e., non-iteratively) determines a variationally optimal approximation to the many-body PMF.[41] The MS-CG method projects the many-body mean force field (i.e., the force field determined by gradients of the many-body PMF) into the space of force fields spanned by the basis vectors chosen to represent the CG potential energy function. This projection corresponds to a linear least squares problem to determine the set of parameters for the CG potential.[60] Equivalently, one can solve the corresponding set of normal equations for the CG parameters.[61]

The normal equations employ a correlation matrix that describes the cross-correlations between pairs of CG degrees of freedom. In Chapter 4,[230] we interpreted features of this

matrix in terms of basic packing properties of liquids. The target vector of the normal equations corresponds to a set of equilibrium force correlation functions that can also be expressed solely in terms of structural correlations (i.e., force-matching without forces!).[62] Moreover, this new set of equations corresponds to a generalization of the Yvon-Born-Green equation from liquid state theory.[63]

7.2 The Generalized Yvon-Born-Green Method

The Generalized Yvon-Born-Green (g-YBG) method[63,64] solves the MS-CG normal equations while calculating the target vector from structures only, without any reference to forces. As demonstrated in Chapter 2,[188] this method yields the same CG force field as the MS-CG method, both in theory and in practice. (Although, we note that the g-YBG method typically requires significantly more sampling of the AA model.) Additionally, the g-YBG method employs the MS-CG correlation matrix to directly decompose equilibrium structural correlation functions into contributions from each term in the CG potential. Thus, this method provides a convenient and transparent framework for identifying the key interactions or driving forces for stabilizing a particular equilibrium structure.

In Chapter 5,[267] we presented a different interpretation of the MS-CG/g-YBG method. Consider a model system with a potential energy function completely specified by a set of parameters $\boldsymbol{\phi}$. Simulation of the system according to this potential generates a set of structural (or force) correlation functions, $\boldsymbol{b}(\boldsymbol{\phi})$, and also a correlation matrix, $\mathbf{G}(\boldsymbol{\phi})$. The g-YBG framework presents an exact relation between $\boldsymbol{\phi}$ and $\boldsymbol{b}(\boldsymbol{\phi})$, namely, $\mathbf{G}(\boldsymbol{\phi})\boldsymbol{\phi} = \boldsymbol{b}(\boldsymbol{\phi})$. Note that the generation of $\boldsymbol{b}(\boldsymbol{\phi})$ from $\boldsymbol{\phi}$ via molecular simulation is a highly nonlinear procedure. In practice, given the ensemble of structures generated from simulations (and the form of the potential), one can calculate $\mathbf{G}(\boldsymbol{\phi})$ and $\boldsymbol{b}(\boldsymbol{\phi})$ and then invert the g-YBG equations to determine $\boldsymbol{\phi}$.[150,282] In this case, no CG mapping was employed, the basis set is complete (i.e., it can exactly represent the underlying potential energy function), and, consequently, the g-YBG equations exactly relate the potential to the structural correlations resulting from the simulation.

More generally, the MS-CG/g-YBG method employs a higher resolution trajectory that is mapped to the chosen CG representation to generate the correlation matrix and target correlation functions, i.e., $\mathbf{G} \equiv \mathbf{G}^{\mathrm{AA}}$, $\boldsymbol{b} \equiv \boldsymbol{b}^{\mathrm{AA}}$. In this case, inverting the g-YBG equations determines the CG potential, $\boldsymbol{\phi}$, that will generate the target set of force correlation functions, $\boldsymbol{b}^{\mathrm{AA}}$, assuming that the correlations generated by the AA model (mapped to the CG representation), $\mathbf{G}^{\mathrm{AA}}$, accurately approximate the correlations that will be generated

by the resulting CG model. This can be considered the fundamental approximation of the MS-CG/g-YBG method.

7.3 Connections to Other Methods

7.3.1 The Relative Entropy Method

Although the MS-CG and RE methods both employ variational principles to approximate the many-body PMF, these two methods appear to be quite different. The MS-CG method performs a direct calculation, while the RE method is iterative. Accordingly, the RE method assures (assuming convergence) that a particular set of structural distributions of the mapped AA ensemble is quantitatively reproduced. On the other hand, the MS-CG framework has no such guarantee.

Several recent studies have compared these two methods, from either a practical or theoretic standpoint. Rühle et al.[74] explicitly compared CG models resulting from these two methods for a particular set of systems, demonstrating that the methods can recover quite different potentials in general. In contrast, Krishna and Larini[79] presented a very general theoretical framework for considering structure-based CG methods.

In Chapter 3,[48] we presented a rigorous theoretic comparison of the MS-CG and RE methods. Our analysis demonstrated that the two methods are intimately connected. In particular, both variational functionals can be expressed in terms of the same information function. While the RE method minimizes the average of this information function over the mapped ensemble, the MS-CG method minimizes the average of the squared gradient of the function. Both methods are constructed to recover the PMF in the limit of a complete basis set. The relationship derived in this work implies that both methods will also yield the same answer in particular scenarios with an incomplete basis set. Additionally, we demonstrated that, for very simple model systems, each method is biased to reproduce particular features of the PMF. These results motivate further investigations into the relationship between these methods to further elucidate their strengths and limitations.

7.3.2 The Iterative g-YBG Method

Iterative methods that parameterize the CG potential to reproduce a set of 1-D structural distributions of the mapped AA ensemble are often employed to parameterize pairwise, distance-dependent potentials. In this case, these methods guarantee (upon convergence) the reproduction of the corresponding set of radial distribution functions (rdfs). This is

highly desirable because rdfs are one of the most common measures of structural order in condensed-phase systems. Moreover, rdfs can be related to experimentally measurable quantities.[45] In contrast to these iterative methods, the MS-CG/g-YBG method does not guarantee the reproduction of the rdfs. Although MS-CG/g-YBG models of a wide variety of systems very accurately reproduce the rdfs,[65–67] there are examples in the literature where these models fail to accurately reproduce the 1-D distributions of the mapped AA ensemble.[74]

These examples have motivated the development of non-conventional force-matching techniques, which aim to improve the accuracy of MS-CG/g-YBG models. Rühle and Junghans[274] proposed two hybrid force-matching methods, which attempt to decouple the contributions from inter- and intramolecular interactions. These methods resulted in models with improved accuracy, relative to the MS-CG model, in the cases considered. However, the source of the improvement remained somewhat elusive.

Cho and Chu[69] proposed an iterative approach based on the g-YBG framework to ensure that a set of 1-D force correlation functions (directly related to the rdfs for pairwise, distance-dependent potentials) were accurately reproduced. This iterative g-YBG (iter-gYBG) method solves the g-YBG equations repetitively, replacing the g-YBG correlation matrix with a corresponding matrix generated from simulations of the CG model at each step. Later, Lu et al.[75] proposed an elegant locally linear approximation to this approach, which saved computational expense by avoiding the calculation of the correlation matrix at each iteration.

In Chapter 5,[267] we proposed a theoretical framework for these two iterative force-matching-based methods and also an extension of the iter-gYBG method to more robustly treat systems with complex intramolecular structure. We demonstrated that, in the examples that we considered, the iter-gYBG method improves the structural accuracy of the model with respect to the MS-CG model. For many systems, convergence of the iter-gYBG method is efficient and robust. However, in general, the convergence properties of the method appear to be less than ideal. For the complex systems of interest in Chapters 5[267] and 6,[283] the iter-gYBG procedure typically diverged after finding an accurate CG force field. In cases where we could perform a large number of iterations, we found that the procedure oscillates about the optimal model. Clearly, a more detailed investigation into the numerical properties of this procedure is necessary. Nevertheless, and perhaps most importantly, we also demonstrated that the iter-gYBG framework provides a powerful tool for identifying specific structural features of the correlation matrix that result in errors in the MS-CG model.

7.4 Transferability

As noted above, the many-body PMF rigorously depends on the thermodynamic state point of the underlying model. Since developing a distinct CG model for each state point of interest could quickly become intractable, there is tremendous interest in understanding the transferability properties (i.e., the accuracy at which a single model can be applied to multiple thermodynamic state points or over a range of similar systems) of CG models.

7.4.1 The Extended Ensemble Framework

In some simple cases, it has been shown that MS-CG models are modestly transferable over, e.g., a small temperature range (Chapter 3[188]). Additionally, Krishna et al.[157] proposed a simple method for predicting the temperature dependence of an MS-CG model. Later Mullinax and Noid[146] proposed a rigorous method for deriving optimal potentials over a set of distinct ensembles. This so-called extended ensemble (EE) method defines a generalized PMF and then employs a variational principle in parallel with the MS-CG method.

7.4.1.1 Liquids

The EE method was first tested over a range of concentrations for liquid neopentane/methanol mixtures.[146] In this case, the resulting model appeared very similar to the MS-CG model derived for the "middle" concentration (i.e., a weighted average of the MS-CG potentials over concentration). In the same study, an EE united atom model for short alkane-alcohol mixtures was developed. In this case, the EE potential did not correspond to a weighted average of MS-CG potentials. We have also, very recently,[282] demonstrated that the EE method can be used to derive a lower resolution model for alkanes to be transferable over different chain lengths.

7.4.1.2 Model Protein Databank

Later, Mullinax and Noid[150] demonstrated a very different application of the EE framework. In this study, a model protein databank (PDB) of protein-like structures was generated using a simple 1-site per residue model.[220] This databank consisted of α, β, and α/β structures. By employing the EE g-YBG method with a complete basis set, the exact parameters of the protein model were recovered from the structural data in the model PDB.

These results explicitly demonstrate that the EE g-YBG framework could be used to inform the development of a knowledge-based (KB) protein model. KB methods use sta-

tistical analysis of the PDB to parameterize the CG potential energy function.[147] These models rely on the ad-hoc[284,285] assumption that the structures in the PDB obey Boltzmann statistics. Additionally, when determining the interaction energy between a pair of sites, KB procedures define a reference state to avoid "overcounting" (i.e., to account for the indirect contribution to the pair mean force). Many different reference states have been proposed based on varying arguments,[148,216,286–289] resulting in a large number of distinct KB models.

The g-YBG framework directly decomposes direct and indirect contributions to the pair mean force for each interaction in the CG model. In the case of a database of structures, this decomposition corresponds to a direct calculation of an optimal reference state using information from the database itself. Moreover, the EE framework provides a rigorous and systematic method for treating statistics from different protein structures. However, there are several outstanding issues with the direct application of the EE framework to the PDB. These issues range from theoretical (e.g., understanding how the features of the PMF change with protein secondary structure) to practical (e.g., constructing dihedral potentials from limited sampling). Some of these problems will be discussed in the Practical Considerations section below.

7.4.1.3 Ionomers

Very recently, we have demonstrated[290] another application of the EE framework for developing models to investigate the conductivity properties of ionomer systems. Ionomers are solid polymer electrolytes with anion groups covalently bound to the polymer, preventing charge polarization present in many polymer/salt systems.[291] Ionomers have received serious interest as an alternative to organic electrolytes for large-scale energy storage.[292,293] However, these systems typically have low room-temperature conductivity, prohibiting them from general use in battery applications.[294] Developing ionomer materials with higher conductivity is difficult because the molecular-level structures formed by ion aggregates are not well understood and are difficult to characterize experimentally.[295–297] Consequently, molecular simulations have the potential to provide useful insight into these systems. AA MD simulations have been applied to study ionomers,[298] but become quickly intractable due to the long relaxation times of the polymers.

Recently, Lu et al.[299] employed the g-YBG method to derive a very coarse model for a PEO-based ionomer at a particular temperature and ion concentration. The resulting model demonstrated very good accuracy, qualitatively reproducing the rdfs between ions and the bond distribution between anions. Moreover, the drastic increase in efficiency supplied by the CG model allowed a detailed investigation of the ionic aggregates formed by this system.

This analysis informed an analogy of the aggregate formation to a worm-like micelle theory,[300] quantifying the competition between enthalpic and entropic driving forces.

However, to gain insight into the conductivity properties of these systems, it is essential that the CG model can accurately describe a range of different systems and thermodynamic conditions. We have employed[290] the EE framework to extend the model of Lu et al. over a range of temperatures and ion concentrations. In particular, we employed a reference electrostatic potential with a state-point-dependent dielectric to derive a single, optimal short-range potential over several different state points. Remarkably, the resulting model qualitatively reproduced the trends in the rdfs over both temperature and ion concentration, even for states outside of the parameterization set.

7.5 Coarse-grained Mappings

In addition to being state-point-dependent, the many-body PMF also rigorously depends on the chosen CG mapping. Consequently, the accuracy of a given approximation to the PMF is intimately and non-trivially related to the mapping. Although a handful of methods[231–239] have been developed for choosing a mapping in particular situations, there has been relatively little progress in developing both systematic and general methods for choosing a mapping. In most cases, especially for commonly used high resolution representations of liquid systems, chemical intuition is the standard method for defining a CG mapping.

However, there are several cases in the literature that demonstrate the limitations of this chemical intuition. For example, in Chapter 2[188] we demonstrated that moving the placement of a single CG site by just 0.5 $\mathring{A}$ significantly impacts the accuracy of the resulting model. Additionally, Mullinax and Noid[146] demonstrated that the accuracy of a CG model does not necessarily increase with increasing resolution. These results are both subtle and unintuitive, motivating a deeper understanding of the relationship between the CG mapping and the accuracy of the resulting model.

7.5.1 Internal State Analysis

In Chapter 5,[267] we proposed a simple a priori (i.e., before calculating the CG potential) method for identifying an optimal CG mapping, among a given set of mappings, based on the internal conformations sampled by the AA model. We found that CG internal conformations that are forbidden (i.e., never sampled) by the AA model can cause significant errors in the MS-CG model. This is because the MS-CG method assumes that the correlations generated

by the AA model accurately reflect correlations that will be generated by the resulting CG model. Forbidden CG internal conformations imply complex cross-correlations between CG degrees of freedom. The simple molecular mechanics interactions that are commonly employed in CG models are unlikely to be able to reproduce.

In both examples considered in Chapter 5, we found that we could remedy this issue by choosing a mapping that did not generate forbidden CG states. Moreover, we demonstrated that the internal conformation analysis successfully identified mappings that resulted in accurate MS-CG models. We also developed corresponding iter-gYBG models, which accurately reproduced the set of 1-D CG distributions. We found that iter-gYBG models derived from "good" mappings more accurately reproduced the distribution of internal conformations. This is a direct consequence of the simplified relationship between the 1-D distributions and higher order correlations implied by the lack of forbidden CG states.

These results demonstrate that, in the case that the MS-CG model fails to accurately reproduce the 1-D CG distributions generated by the underlying model, the iter-gYBG method can be employed to improve the accuracy of the model with respect to these distributions. However, the iter-gYBG model achieves this improved accuracy at the cost of distorting the higher order, cross-correlations between CG degrees of freedom. In some sense, this implies that when there are errors in an MS-CG model, there is likely a fundamental flaw in the representation or interaction set of the model.

This study provided a simple, but instructive, examination of a set of criteria to more systematically choose a CG mapping. The proposed method is clearly limited to relatively high resolution models of modestly sized molecules with some internal flexibility. However, extension of this analysis to intermolecular states would significantly expand the utility of the method. This should be a point of future investigation.

7.5.2 Mapping Entropy

Shell[37] demonstrated that the relative entropy can be expressed as a sum of two terms; a term that depends on the CG potential and a "mapping entropy" which is independent of the potential. In Chapter 3,[48] we defined a slightly different mapping entropy and proved that it quantified the difference in entropies between the AA and CG models. Additionally, we demonstrated that this mapping entropy could be rewritten as a sum of two, physically meaningful, terms and also derived an upper bound for this quantity. The mapping entropy describes the inherent loss of information due to the CG mapping alone and, consequently, may be useful for developing a general, systematic mapping optimization procedure. This

should also be a point of future investigations.

7.6 Coarse-grained Interaction Sets

7.6.1 Molecular Liquids

Based on the general success of the MS-CG method, the molecular mechanics interactions commonly employed to model the CG potential energy function are often sufficient to accurately reproduce the structural features of the underlying model. However, there are exceptions to this success, e.g., the 1-site MS-CG model of SPC/E water,[59] which fails to accurately reproduce the characteristic tetrahedral structure of the underlying model. Das and Andersen[301] demonstrated that more complex, three-body interactions could be employed within the MS-CG framework to improve the structural accuracy of this model.

In Chapter 5[267] we demonstrated that the errors in a 3-site MS-CG model of hexane were due to the inability of the simple molecular mechanics potentials employed to reproduce the complex cross-correlations generated by the underlying model. It was also demonstrated[75,267] that changing the mapping to a lower or higher representation alleviated this problem by either simplifying the PMF (i.e., reducing the amount of information in the model) or by decreasing the complexity of the cross-correlations, respectively. One could take a different approach and include a term in the CG potential energy function for the 3-site model that explicitly coupled the bond and angle degrees of freedom. This coupling potential would surely be more capable of reproducing the complex cross-correlations of the underlying model. This approach was beyond the scope of the present study, but may be essential for systems with significant internal flexibility or for lower resolution models of larger molecules.

7.6.2 Minimal Models of Peptides

In Chapter 6,[283] we examined the utility of the MS-CG method for determining minimal models for peptides. We found that the MS-CG method produced quantitatively accurate models when the underlying ensemble corresponded to a well-defined structure (e.g., a helix). However, when the underlying model sampled a more complex, disordered ensemble, the resulting MS-CG model failed to even qualitatively reproduce the propensity of different structures. We demonstrated that correlations generated by cooperative helix formation in the AA model could not be reproduced by a minimal CG model with simple molecular mechanics interactions. By making physically motivated modifications to the correlation matrix, based on an iter-gYBG model, we verified the source of error in the MS-CG model.

These studies motivate the development and implementation of more complex basis functions into the MS-CG/g-YBG framework, which explicitly couple CG degrees of freedom and allow for a better representation of cross-correlations generated by AA models. In particular, explicitly accounting for coupling between intramolecular interactions, e.g., between degrees of freedom governed by an angle and a dihedral angle, will be essential for accurately modeling molecules with complex intramolecular structure at a minimal resolution. Additionally, although several CG protein models employ interactions that couple degrees of freedom,[109,111,302,303] the implementation of these types of complex potentials, which lend themselves to lower resolution models, into standard MD simulation packages will significantly expand the breadth of such modeling efforts.

7.7 Practical Considerations

Throughout this work, we have mainly focused on theoretical and methodological developments of the MS-CG/g-YBG method. Along the way, we also faced many practical challenges, which were largely confined to the Supporting Information documentation of Chapters 4,[230] 5,[267] and 6.[283] In this section, we will review these practical issues and corresponding advances of the MS-CG/g-YBG method.

7.7.1 Sources of Numerical Problems

7.7.1.1 Interaction Set

Similar to the CG mapping, there is no general method for choosing the set of interactions governing terms in the CG potential. In Chapter 3,[48] we proved that the MS-CG/g-YBG method determines a unique CG potential if and only if the basis vectors are linearly independent. Linear dependence of the basis vectors can be assessed by examining the MS-CG/g-YBG correlation matrix. In particular, linearly dependent basis vectors will result in some number of zero eigenvalues in the MS-CG correlation matrix. Moreover, nearly redundant interactions defined in the CG potential energy function will result in very small eigenvalues and, consequently, numerical instabilities in the MS-CG/g-YBG procedure. In practice, this issue is typically easy to avoid by 1) carefully choosing the set of CG interactions and 2) monitoring the condition number (i.e., the largest eigenvalue divided by the smallest eigenvalue) of the MS-CG correlation matrix.

We have also found that employing (semi-)redundant interactions with the iter-gYBG method can lead to a bias towards particular regions of configuration space (not published).

Therefore, it is generally important to choose an interaction set with caution. The development of more automated and systematic methods for choosing the set of CG interactions should be a point of future study.

7.7.1.2 Basis Function Representation

Once a set of interactions is defined, one must choose how to represent the corresponding force functions. Force functions have been represented by both analytic and tabulated functions within the MS-CG framework. The analytic functions typically have a smaller number of parameters, but significantly restrict the functional form compared with a tabulated representation. Additionally, we have found that the condition number increases tremendously when employing an analytic representation.

For many problems, any of these functions are sufficient to determine an accurate force field. However, the more complex basis functions (e.g., B-splines) have been demonstrated to be very useful for modeling complex systems with limited sampling of the underlying model.[72] In general, the proper representation (functional form, grid spacing, cut-off, etc.) must be chosen for each particular application. Moreover, the numerical properties and resulting accuracy of the calculation may be dependent on these choices. However, we have found that for the same approximate level of coarse-graining, the optimal representation of the force functions is very similar for a wide range of distinct applications. Das and Andersen[304] proposed an automated algorithm for optimizing the force function representation. This method could be very useful as the variety of systems treated with the MS-CG/g-YBG method expands.

7.7.1.3 Insensitivity of Structure

It is well known that the pair structure of a liquid is largely determined by the short-range portion of the pairwise interaction potentials.[35,47,48] In Chapter 4,[230] we demonstrated that in the context of the MS-CG procedure, this insensitivity can result in a wide range of interaction potentials that can yield the same pair structure within numerical precision, even though in theory there is a unique potential which generates this structure.[46,48] In principle, this insensitivity could lead to instabilities in the MS-CG/g-YBG procedure. In practice, we do not observe this problem, perhaps because the MS-CG/g-YBG correlation matrix employs higher order information which can effectively distinguish between different potentials that yield nearly the same structure. However, this further motivates the development of specialized interaction functions for CG modeling that would restrict the space of nearly

identical solutions.

This insensitivity is even more alarming in the context of the iterative methods, which are built upon the assumption that there exists a unique potential which generates a given pair structure. In practice, this could lead to unpredictable variance in the calculated potentials via these methods. For example, in Chapters 5[267] and 6[283] we found that the iter-gYBG method diverged from or oscillated about an optimal CG model. Further investigation into the general impact of this problem and the robustness of the iterative methods is warranted.

7.7.2 Tools for Numerical Assessment

One of the most appealing aspects of the MS-CG/g-YBG method is that it is built upon a well-characterized numerical framework. Throughout this work, we have employed standard numerical analysis techniques to assess the numerics of the MS-CG/g-YBG method for any given problem. We have already implicated the condition number of the MS-CG correlation matrix as an important quantity to assess the numerical stability of the method. In the case that semi-redundant basis vectors cause very small eigenvalues of the correlation matrix, the calculations may result in physically unreasonable force functions. This problem can sometimes be avoided by removing the lowest eigenvalues with Singular Value Decomposition,[175] although we do not advise this approach in general. It is more useful to adjust the interaction set to avoid redundant interactions.

It is also quite standard to assess the numerics of a system of linear equations by analyzing the eigenspace of the matrix. In Chapter 4,[230] we employed this type of analysis to gain intuition about the importance of particular basis vectors. However, we note that a particular preconditioning[122] is necessary to make the matrix dimensionless and to put the basis vectors on an even footing. Further implementation of numerical techniques (e.g., a rigorous error analysis) to assess the robustness of MS-CG/g-YBG calculations will continue to propel the utility of the method.

7.7.3 Solutions for Numerical Problems

7.7.3.1 Reference Potentials

Within the context of the MS-CG/g-YBG method, Mullinax and Noid[134] formally derived a theory for subtracting out contributions from a pre-defined, fixed term in the CG potential. They demonstrated that this "reference" potential can significantly simplify the computational complexity of the force field calculation. Additionally, they demonstrated that the

method provided a convenient framework for utilizing CG potentials that have already been validated for describing a particular type of interaction. Utilizing this theory, they developed a CG model that accurately described the solvation structure of a hydrophobic solute in methanol.

In Chapters 5[267] and 6,[283] we employed this methodology to avoid problematic interactions within the iter-gYBG framework. We have also applied it to the systematic development of dihedral angle potentials in cases where particular angles are never sampled (Chapter 6). The numerical simplification provided by reference potentials may be essential for robust MS-CG/g-YBG calculations of complex systems with many, highly coupled interactions.

7.7.3.2 Regularization

To increase the robustness of MS-CG/g-YBG and, especially, iter-gYBG calculations, it is useful to add a regularization term to the normal equations that penalizes the force functions from becoming too large. This regularization helps prevent numerical noise from accumulating throughout the iterative procedure, particularly at the ends of the distributions where sampling is more sparse. Additionally, regularization avoids the over-fitting of noise, especially for interactions represented with B-spline basis functions. In Chapters 5[267] and 6,[283] we employed a regularization scheme that is quite simple compared to others proposed,[305] but worked well to prevent these numerical problems. In many simple test cases, we found that the MS-CG results were quite robust to the parameter that provides the weight of the penalty to large forces. However, in more complicated systems, where regularization was important for dealing with numerical noise, the results were much more sensitive to this parameter. Consequently, great caution should be taken when using this, or other, regularization schemes. Utilization of sophisticated numerical techniques within the MS-CG/g-YBG implementation will be essential for the continued treatment of more complex systems.

7.7.3.3 Constraints

In many cases, the MS-CG/g-YBG method robustly and accurately determines a CG model for a given mapping, interaction set, and basis function representation. However, in certain cases, the decomposition of the mean force provided by the MS-CG/g-YBG correlation matrix may determine unphysical force functions. For example, for dihedral angle interactions that are treated with a tabulated basis function, the normal equations have no inherent requirement that the calculated dihedral force function should integrate to zero. However, in

practice, this requirement must be satisfied since the dihedral potential must be periodic. For simple systems, solving the normal equations may yield a force function that approximately integrates to zero. In general, however, this is not the case and one must add a constraint to the normal equations which requires that each set of tabulated dihedral coefficients sum to zero. In Chapters 5[267] and 6,[283] we have demonstrated that adding this constraint results in accurate and numerically robust calculations for several different applications.

Implementing reference potentials or reducing the complexity of the basis function (e.g., from tabulated to analytic) is another form of effectively constraining the normal equations. For building accurate CG models of complex systems (e.g., proteins), it will be crucial to simplify the determination of the optimal force field as much as possible. In particular, as we have already discussed, the development of specialized interaction potentials or basis function representations for accurately treating effective interactions in molecular systems should be the focus of future investigations.

7.7.4 Software

Although the MS-CG/g-YBG method provides an efficient and transparent framework for investigating CG models of complex systems, its implementation is relatively complex. For this method to be generally useful, there needs to be accessible software that performs accurate and robust calculations. VOTCA[74] is an open-source software package which was developed for determining CG models using the Iterative Boltzmann Inversion,[35] Inverse Monte Carlo,[34] and MS-CG methods. VOTCA employs a block averaging linear least squares algorithm[60] to calculate the MS-CG force field, while employing a cubic spline basis to represent the force function for each interaction. This implementation is useful for building accurate MS-CG models for a wide range of systems. Moreover, VOTCA provides a platform for seamless comparisons of the CG models resulting from each of these methods, as demonstrated by Rühle and coworkers.[74,274] These comparisons may be extremely useful in achieving the optimal accuracy of a particular CG model or for understanding the limitations of a particular representation or interaction set.

However, the present work has demonstrated that accurate MS-CG models can be developed for a wider range of systems by employing various specialized techniques and analysis. Utilization of these techniques will likely lead to both more accurate CG models for a larger variety of systems and further development of bottom-up CG methodology. However, to achieve this advancement, there must be reliable software that can be used with moderate (i.e., non-expert) knowledge of the MS-CG/g-YBG method. We are currently preparing to

release such a software package[282] and will demonstrate its utility for performing robust and automated MS-CG/g-YBG calculations for a variety of applications.

7.7.4.1 Tools for Gaining Physical Intuition

The MS-CG/g-YBG framework provides a transparent view of the interactions governing the equilibrium structure of the system. Several studies have utilized this framework to elucidate the physical driving forces of complex processes, e.g., hydrophobic association[208] and cellulose deconstruction.[209] In Chapter 2,[188] we explicitly demonstrated how the MS-CG/g-YBG method decomposes the mean force along each CG degree of freedom into contributions from each interaction governing a term in the CG potential. Subsequently, in Chapter 4,[230] we demonstrated how to analyze this decomposition to determine the dominant forces governing the packing of molecules in a complex molecular liquid.[230] We also demonstrated that visualization of sub-blocks of the correlation matrix, along with an understanding of the force vectors arising from each interaction, can provide physically relevant information about the equilibrium structure of the system. In Chapters 5[267] and 6,[283] we extended this visualization analysis in conjunction with the iter-gYBG method to identify precise correlations that cause errors in the MS-CG/g-YBG models. These tools, along with others that we have developed for investigating the physics of CG models, will be included within our software package.[282]

7.8 Final Outlook

At the outset of this dissertation work, the MS-CG/g-YBG framework had already been developed and tested on many systems. We set forth to gain a deeper understanding of this method, to elucidate its strengths and limitations, and to expand the variety and complexity of systems that could be accurately treated. We developed several theoretic constructs that made rigorous connections to other structure-based bottom-up CG methods, revealing the particular strengths of each. As illustrated in Chapter 5, these connections directly led to a shift in our interpretation of the MS-CG/g-YBG method as we began to better understand its fundamental approximation.

We utilized this insight to develop a framework for determining the precise source of error in cases that MS-CG/g-YBG models lack sufficient accuracy. In some cases, these investigations identified general principles for developing accurate CG models. Moreover, in each case that we considered, we were able to overcome the initial problem, either by improving the accuracy of the model (a posteriori) or by avoiding the problem altogether (a

priori). Finally, these advances allowed us to begin robustly and accurately treating more complex systems.

Top-down CG models have already contributed great insight into many universal phenomena.[3,20–22] More recently, bottom-up methods have been applied to develop accurate CG models for systems with increasing complexity, including hierarchical biological structures.[122,306] This dissertation clearly demonstrates both the power and the potential of bottom-up CG methods. In particular, we demonstrated that the statistical mechanical basis for bottom-up approaches allows systematic investigation and reduction of errors that arise in the CG model. For example, using the methods and insight from this work, we recently constructed exceedingly accurate and efficient models for investigating the conductivity properties of PEO-based ionomers.[290,299]

Despite largely focusing on molecular liquids, this work was mainly motivated by applications for investigating complex biological molecules, e.g., proteins. As we discussed earlier in this chapter, the g-YBG framework has enormous potential for identifying an optimal reference state within the context of knowledge-based methods. However, there remain theoretical, methodological, and practical issues that need to be resolved in order to realize this potential. We reviewed many of these issues within the last chapter, summarized our contribution to a number of them, and highlighted some outstanding problems.

More generally, the holy grail of CG biomolecular modeling seems to be the development of a fully transferable CG protein model. Ideally, this CG model will retain many features of a standard AA model: 1) it should be transferable to any amino-acid sequence, 2) it should be transferable over thermodynamic state points (e.g., temperature), and 3) it should employ relatively simple "molecular mechanics" potentials. Additionally, the CG model is expected to provide dramatically greater efficiency than the AA model by, e.g.: 1) representing the solvent molecules implicitly and 2) reducing the resolution of the protein. This appears to be a formidable challenge for any coarse-graining approach.

In particular, the many-body PMF depends on both the chemical identity of the system as well as the thermodynamic state point. Additionally, in this work we demonstrated that the chosen CG representation can have a profound and counterintuitive impact upon the accuracy of the CG model. Without a better understanding of these considerations, it is impossible to ensure that a CG model will exhibit a significant amount of transferability. However, in this work, we also proposed a simple procedure for identifying an optimal CG representation and developed a framework for precisely identifying fundamental limitations of a given CG model. These methods, along with the extended ensemble method, proposed by Mullinax and Noid[146] and reviewed in the last chapter, provide a powerful framework for

investigating the transferability properties of a CG model.

Another significant difficulty may arise because molecular mechanics potentials may be inadequate for accurately modeling 3-D protein structures while employing a low resolution protein representation. As degrees of freedom are removed from the system, the underlying atomically-detailed interactions give rise to many-body, effective interactions between CG sites. If a CG model is expected to reproduce the configurational ensemble of the underlying model (even for a single system and thermodynamic state point), it must be able to accurately approximate these effective interactions. In this work, we explicitly demonstrated that molecular mechanics potentials, along with a minimal CG representation were incapable of reproducing the cooperative formation of helical turns observed in atomically-detailed models. However, we also demonstrated that more complex potentials, which couple CG degrees of freedom, would likely remedy this problem.

The need for more complex potentials for accurately modeling CG interactions is largely recognized and, occasionally, explicitly addressed.[109,111,302,303] The implementation of more complex potentials into standard MD software packages will surely increase the number of CG models that employ them. However, a better understanding of the specific functional forms which lend themselves to accurately modeling particular types of CG effective interactions would significantly advance CG modeling efforts. Moreover, the methods proposed in this work for identifying limitations of the interaction set or basis functions are particularly poised to systematically tackle this problem. Therefore, it seems clear that bottom-up methods have much to offer for the development of transferable CG protein models.

In the last chapter we identified a number of outstanding challenges which should be addressed for the more general advancement of CG modeling. In addition to the proposed future developments discussed in the preceding paragraph, the development of systematic and automated methods for determining the optimal CG mapping and interaction set would represent a momentous advance for CG methodology. Additionally, the continued use of sophisticated numerical methods for robust and accurate determination of CG force fields as well as the implementation of these methods into specialized, but accessible, software packages will significantly expand the breadth of coarse-graining efforts. Finally, this work has explicitly demonstrated that the development of CG models, in itself, provides detailed insight into the physics of the underlying system. Therefore, the continued development of CG models will lead to further insight into the driving forces of complex molecular processes.

Bibliography

[1] (2013) *The Nobel Prize in Chemistry 2013*, Nobelprize.org. Nobel Media AB 2014. URL: http://www.nobelprize.org/nobel_prizes/chemistry/laureates/2013/

[2] Szabo, A. and M. Karplus (1972) "Mathematical-Model for Structure-Function Relations in Hemoglobin," *J. Mol. Biol.*, **72**(1), 163–197. DOI: 10.1016/0022-2836(72)90077-0

[3] Levitt, M. (1976) "A Simplified Representation of Protein Conformations for Rapid Simulation of Protein Folding," *J. Mol. Biol.*, **104**(1), 59–107. DOI: 10.1016/0022-2836(76)90004-8

[4] Metropolis, N., A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller (1953) "Equation of State Calculations by Fast Computing Machines," *J. Chem. Phys.*, **21**(6), 1087. DOI: 10.1063/1.1699114

[5] Alder, B. J. and T. E. Wainwright (1959) "Studies in Molecular Dynamics. 1. General Method." *J. Chem. Phys.*, **31**(2), 459–466. DOI: 10.1063/1.1730376

[6] Shaw, D. E., P. Maragakis, K. Lindorff-Larsen, S. Piana, R. O. Dror, M. P. Eastwood, J. A. Bank, J. M. Jumper, J. K. Salmon, Y. B. Shan, and W. Wriggers (2010) "Atomic-Level Characterization of the Structural Dynamics of Proteins," *Science*, **330**(6002), 341–346. DOI: 10.1126/science.1187409

[7] Hess, B., C. Kutzner, D. van der Spoel, and E. Lindahl (2008) "GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation," *J. Chem. Theor. Comp.*, **4**(3), 435–447. DOI: 10.1021/ct700301q

[8] Case, D. A., I. T. E. Cheatham, T. Darden, H. Gohlke, R. Luo, J. K. M. Merz, A. Onufriev, C. Simmerling, B. Wang, and R. Woods (2005) "The Amber Biomolecular Simulation Programs," *J. Comp. Chem.*, **26**(16), 1668–1688. DOI: 10.1002/jcc.20290

[9] Brooks, C. (1995) "Methodological Advances in Molecular-Dynamics Simulations of Biological-Systems," *Curr. Opin. Struc. Biol.*, **5**(2), 211–215. DOI: 10.1016/0959-440X(95)80078-6

[10] Sugita, Y. and Y. Okamoto (1999) "Replica-Exchange Molecular Dynamics Method for Protein Folding," *Chem. Phys. Lett.*, **314**(1-2), 141–151. DOI: 10.1016/S0009-2614(99)01123-9

[11] Shirts, M. R. and J. D. Chodera (2008) "Statistically Optimal Analysis of Samples from Multiple Equilibrium States," *J. Chem. Phys.*, **129**(12), 124105. DOI: 10.1063/1.2978177

[12] Noe, F., C. Schutte, E. Vanden-Eijnden, L. Reich, and T. R. Weikl (2009) "Constructing the Equilibrium Ensemble of Folding Pathways from Short Off-Equilibrium Simulations," *Proc. Natl. Acad. Sci. USA*, **106**(45), 19011–19016. DOI: 10.1073/pnas.0905466106

[13] Durrant, J. D. and J. A. McCammon (2011) "Molecular Dynamics Simulations and Drug Discovery," *BMC Biol.*, **9**, 71. DOI: 10.1186/1741-7007-9-71

[14] (2011) *Materials Genome Initiative for Global Competativeness*, National Science and Technology Council, Office of Science and Technology Policy, contact: Tom Kalil and Cyrus Wadia. URL: http://www.whitehouse.gov/mgi

[15] Karplus, M. and J. A. McCammon (2002) "Molecular Dynamics Simulations of Biomolecules," *Nat. Struct. Mol. Biol.*, **9**(10), 646–652. DOI: 10.1038/nsb1002-788a

[16] Frenkel, D. and B. Smit (2002) *Understanding Molecular Simulation: From Algorithms to Applications*, second ed., Academic Press, San Diego, CA USA.

[17] Brooks, B. R., R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus (1983) "CHARMM - A Program for Macromolecular Energy, Minimization, and Dynamics Calculations," *J. Comp. Chem.*, **4**(2), 187–217. DOI: 10.1002/jcc.540040211

[18] Jorgensen, W. L., D. S. Maxwell, and J. Tirado-Rives (1996) "Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids," *J. Am. Chem. Soc.*, **118**(45), 11225–11236. DOI: 10.1021/ja9621760

[19] Wang, J. M., R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case (2004) "Development and Testing of a General Amber Force Field," *J. Comp. Chem.*, **25**(9), 1157–1174. DOI: 10.1002/jcc.20035

[20] Zimm, B. and J. Bragg (1959) "Theory of the Phase Transition between Helix and Random Coil in Polypeptide Chains," *J. Chem. Phys.*, **31**(2), 526–535. DOI: 10.1063/1.1730390

[21] Lifson, S. and A. Roig (1961) "On the Theory of Helix-Coil Transition in Polypeptides," *J. Chem. Phys.*, **34**(6), 1963–1974. DOI: 10.1063/1.1731802

[22] Honeycutt, J. D. and D. Thirumalai (1992) "The Nature of Folded States of Globular Proteins," *Biopolymers*, **32**(6), 695–709. DOI: 10.1002/bip.360320610

[23] Noid, W. G. (2013) "Perspective: Coarse-Grained Models for Biomolecular Systems," *J. Chem. Phys.*, **139**(9), 090901. DOI: 10.1063/1.4818908

[24] Schlick, T., R. Collepardo-Guevara, L. A. Halvorsen, S. Jung, and X. Xiao (2011) "Biomolecular Modeling and Simulation: A Field Coming of Age," *Quart. Rev. Biophys.*, **44**(2), 191–228. DOI: 10.1017/S0033583510000284

[25] Ingólfsson, H. I., C. A. Lopez, J. J. Uusitalo, D. H. de Jong, S. M. Gopal, X. Periole, and S. J. Marrink (2014) "The Power of Coarse Graining in Biomolecular Simulations," *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, **4**(3), 225–248. DOI: 10.1002/wcms.1169

[26] Rohrdanz, M. A., W. Zheng, and C. Clementi (2013) "Discovering Mountain Passes via Torchlight: Methods for the Definition of Reaction Coordinates and Pathways in Complex Macromolecular Reactions," *Annu. Rev. Phys. Chem.*, **64**, 295–316. DOI: 10.1146/annurev-physchem-040412-110006

[27] Shelley, J. C., M. Y. Shelley, R. C. Reeder, S. Bandyopadhyay, and M. L. Klein (2001) "A Coarse Grain Model for Phospholipid Simulation," *J. Phys. Chem. B*, **105**(19), 4464–4470. DOI: 10.1021/jp010238p

[28] Marrink, S. J., H. J. Risselada, S. Yefimov, D. P. Tieleman, and A. H. de Vries (2007) "The MARTINI Force Field: Coarse Grained Model for Biomolecular Simulations," *J. Phys. Chem. B*, **111**(27), 7812–7824. DOI: 10.1021/jp071097f

[29] Marrink, S. J. and D. P. Tieleman (2013) "Perspective on the Martini Model," *Chem. Soc. Rev.*, **42**(16), 6801–6822. DOI: 10.1039/c3cs60093a

[30] Maerzke, K. A. and J. I. Siepmann (2011) "Transferable Potentials for Phase Equilibria-Coarse-Grain Description for Linear Alkanes," *J. Phys. Chem. B*, **115**(13), 3452–3465. DOI: 10.1021/jp1063935

[31] Rossi, G., L. Monticelli, S. R. Puisto, I. Vattulainen, and T. Ala-Nissila (2011) "Coarse-Graining Polymers with the MARTINI Force-Field: Polystyrene as a Benchmark Case," *Soft Matter*, **7**(2), 698–708. DOI: 10.1039/c0sm00481b

[32] Monticelli, L., S. K. Kandasamy, X. Periole, R. G. Larson, D. P. Tieleman, and S.-J. Marrink (2008) "The MARTINI Coarse-Grained Force Field: Extension to Proteins," *J. Chem. Theor. Comp.*, **4**(5), 819–834. DOI: 10.1021/ct700324x

[33] DeVane, R., W. Shinoda, P. B. Moore, and M. L. Klein (2009) "Transferable Coarse Grain Nonbonded Interaction Model for Amino Acids," *J. Chem. Theor. Comp.*, **5**(8), 2115–2124. DOI: 10.1021/ct800441u

[34] Lyubartsev, A. P. and A. Laaksonen (1995) "Calculation of Effective Interaction Potentials from Radial Distribution Functions: A Reverse Monte Carlo Approach," *Phys. Rev. E*, **52**(4,A), 3730–3737. DOI: 10.1103/PhysRevE.52.3730

[35] Müller-Plathe, F. (2002) "Coarse-Graining in Polymer Simulation: From the Atomistic to the Mesoscopic Scale and Back," *ChemPhysChem*, **3**(9), 754–769. DOI: 10.1002/1439-7641(20020916)3:9<754::AID-CPHC754>3.0.CO;2-U

[36] Izvekov, S. and G. A. Voth (2005) "A Multiscale Coarse-Graining Method for Biomolecular Systems," *J. Phys. Chem. B*, **109**(7), 2469–2473. DOI: 10.1021/jp044629q

[37] Shell, M. S. (2008) "The Relative Entropy is Fundamental to Multiscale and Inverse Thermodynamic Problems," *J. Chem. Phys.*, **129**(14), 144108. DOI: 10.1063/1.2992060

[38] Savelyev, A. and G. A. Papoian (2009) "Molecular Renormalization Group Coarse-Graining of Electrolyte Solutions: Applications to Aqueous NaCl and KCl," *J. Phys. Chem. B*, **113**(22), 7785–7793. DOI: 10.1021/jp9005058

[39] Louis, A. A. (2002) "Beware of Density Dependent Pair Potentials," *J. Phys.: Condens. Matter*, **14**(40), 9187–9206. DOI: 10.1088/0953-8984/14/40/311

[40] Silbermann, J., S. H. L. Klapp, M. Shoen, N. Channamsetty, H. Block, and K. E. Gubbins (2006) "Mesoscale Modeling of Complex Binary Fluid Mixtures: Towards an Atomistic Foundation of Effective Potentials," *J. Chem. Phys.*, **124**(7), 074105. DOI: 10.1063/1.2161207

[41] Noid, W. G., J.-W. Chu, G. S. Ayton, V. Krishna, S. Izvekov, G. A. Voth, A. Das, and H. C. Andersen (2008) "The Multiscale Coarse-Graining Method. I. A Rigorous Bridge between Atomistic and Coarse-Grained Models," *J. Chem. Phys.*, **128**(24), 244114. DOI: 10.1063/1.2938860

[42] Beauchamp, K. A., Y.-S. Lin, R. Das, and V. S. Pande (2012) "Are Protein Force Fields Getting Better? A Systematic Benchmark on 524 Diverse NMR Measurements," *J. Chem. Theor. Comp.*, **8**(4), 1409–1414. DOI: 10.1021/ct2007814

[43] Hill, T. L. (1997) *An Introduction to Statistical Thermodynamics*, Addison Wesley Publishing Company.

[44] Kirkwood, J. G. (1935) "Statistical Mechanics of Fluid Mixtures," *J. Chem. Phys.*, **3**(5), 300–313. DOI: 10.1063/1.1749657

[45] Allen, M. P. and D. P. Tildesley (1987) *Computer Simulation of Liquids*, Oxford Press, New York, NY USA.

[46] Henderson, R. L. (1974) "A Uniqueness Theorem for Fluid Pair Correlation Functions," *Phys. Lett. A*, **49**(3), 197–198. DOI: 10.1016/0375-9601(74)90847-0

[47] Andersen, H. C., D. Chandler, and J. D. Weeks (1976) "Roles of Repulsive and Attractive Forces in Liquids : The Equilibrium Theory of Classical Fluids," *Adv. Chem. Phys.*, **34**, 105–156. DOI: 10.1002/9780470142530.ch2

[48] Rudzinski, J. F. and W. G. Noid (2011) "Coarse-Graining Entropy, Forces, and Structures," *J. Chem. Phys.*, **135**(21), 214101. DOI: 10.1063/1.3663709

[49] Karimi-Varzaneh, H. A. and F. Müller-Plathe (2012) "Coarse-Grained Modeling for Macromolecular Chemistry," *Top. Curr. Chem.*, **307**, 295–321. DOI: 10.1007/128_2010_122

[50] Lyubartsev, A., A. Mirzoev, L. J. Chen, and A. Laaksonen (2010) "Systematic Coarse-Graining of Molecular Models by the Newton Inversion Method," *Faraday Disc.*, **144**, 43–56. DOI: 10.1039/b901511f

[51] Murtola, T., E. Falck, M. Patra, M. Karttunen, and I. Vattulainen (2004) "Coarse-Grained Model for Phospholipid/Cholesterol Bilayer," *J. Chem. Phys.*, **121**(18), 9156–9165. DOI: 10.1063/1.1803537

[52] Lyubartsev, A. P. (2005) "Multiscale Modeling of Lipids and Lipid Bilayers," *Eur. Biophys. J.*, **35**(1), 53–61. DOI: 10.1007/s00249-005-0005-y

[53] Lyubartsev, A. P. and A. Laaksonen (1997) "Osmotic and Activity Coefficients from Effective Potentials for Hydrated Ions," *Phys. Rev. E*, **55**(5), 5689–5696. DOI: 10.1103/PhysRevE.55.5689

[54] Hadley, K. R. and C. McCabe (2010) "A Structurally Relevant Coarse-Grained Model for Cholesterol," *Biophys. J.*, **99**(9), 2896–2905. DOI: 10.1016/j.bpj.2010.08.044

[55] Hadley, K. R. and C. McCabe (2010) "A Coarse-Grained Model for Amorphous and Crystalline Fatty Acids," *J. Chem. Phys.*, **132**(13), 134505. DOI: 10.1063/1.3360146

[56] Wang, Z. J. and M. Deserno (2010) "A Systematically Coarse-Grained Solvent-Free Model for Quantitative Phospholipid Bilayer Simulations," *J. Phys. Chem. B*, **114**(34), 11207–11220. DOI: 10.1021/jp102543j

[57] Mukherjee, B., L. Delle Site, K. Kremer, and C. Peter (2012) "Derivation of Coarse Grained Models for Multiscale Simulation of Liquid Crystalline Phase Transitions," *J. Phys. Chem. B*, **116**(29), 8474–8484. DOI: 10.1021/jp212300d

[58] Jochum, M., D. Andrienko, K. Kremer, and C. Peter (2012) "Structure-Based Coarse-Graining in Liquid Slabs," *J. Chem. Phys.*, **137**(6), 064102. DOI: 10.1063/1.4742067

[59] Izvekov, S. and G. A. Voth (2005) "Multiscale Coarse Graining of Liquid-State Systems," *J. Chem. Phys.*, **123**(13), 134105. DOI: 10.1063/1.2038787

[60] Noid, W. G., P. Liu, Y. T. Wang, J.-W. Chu, G. S. Ayton, S. Izvekov, H. C. Andersen, and G. A. Voth (2008) "The Multiscale Coarse-Graining Method. II. Numerical Implementation for Molecular Coarse-Grained Models," *J. Chem. Phys.*, **128**(24), 244115. DOI: 10.1063/1.2938857

[61] Charles L. Lawson, R. J. H. (1974) *Solving Least Squares Problems*, Prentice-Hall, Englewood Cliffs, NJ USA.

[62] Noid, W. G., J.-W. Chu, G. S. Ayton, and G. A. Voth (2007) "Multiscale Coarse-Graining and Structural Correlations: Connections to Liquid State Theory," *J. Phys. Chem. B*, **111**(16), 4116–4127. DOI: 10.1021/jp068549t

[63] Mullinax, J. W. and W. G. Noid (2009) "Generalized Yvon-Born-Green Theory for Molecular Systems," *Phys. Rev. Lett.*, **103**(19), 198104. DOI: 10.1103/PhysRevLett.103.198104

[64] Mullinax, J. W. and W. G. Noid (2010) "A Generalized Yvon-Born-Green Theory for Determining Coarse-grained Interaction Potentials," *J. Phys. Chem. C*, **114**(12), 5661–5674. DOI: 10.1021/jp9073976

[65] Noid, W. G., G. S. Ayton, S. Izvekov, and G. A. Voth (2008) "The Multiscale Coarse-Graining Method: A Systematic Approach to Coarse Graining," in *Coarse-graining of condensed phase and biomolecular systems* (G. A. Voth, ed.), chap. 3, CRC Press, 21–40.

[66] Lu, L. and G. A. Voth (2012) "The Multiscale Coarse-Graining Method," *Adv. Chem. Phys.*, **149**, 47–81. DOI: 10.1002/9781118180396.ch2

[67] Saunders, M. G. and G. A. Voth (2013) "Coarse-Graining Methods for Computational Biology," *Ann. Rev. Biophys.*, **42**, 73–93. DOI: 10.1146/annurev-biophys-083012-130348

[68] Das, A. and H. C. Andersen (2009) "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**(3), 034102. DOI: 10.1063/1.3173812

[69] Cho, H. M. and J. W. Chu (2009) "Inversion of Radial Distribution Functions to Pair Forces by Solving the Yvon-Born-Green Equation Iteratively," *J. Chem. Phys.*, **131**(13), 134107. DOI: 10.1063/1.3238547

[70] Wang, Y. T., W. G. Noid, P. Liu, and G. A. Voth (2009) "Effective Force Coarse-Graining," *Phys. Chem. Chem. Phys.*, **11**(12), 2002–2015. DOI: 10.1039/b819182d

[71] Das, A. and H. C. Andersen (2010) "The Multiscale Coarse-Graining Method. V. Isothermal-Isobaric Ensemble," *J. Chem. Phys.*, **132**(16), 164106. DOI: 10.1063/1.3394862

[72] Lu, L. Y., S. Izvekov, A. Das, H. C. Andersen, and G. A. Voth (2010) "Efficient, Regularized, and Scalable Algorithms for Multiscale Coarse-Graining," *J. Chem. Theor. Comp.*, **6**(3), 954–965. DOI: 10.1021/ct900643r

[73] Das, A., L. Lu, H. C. Andersen, and G. A. Voth (2012) "The Multiscale Coarse-Graining Method. X. Improved Algorithms for Constructing Coarse-Grained Potentials for Molecular Systems," *J. Chem. Phys.*, **136**(19), 194115. DOI: 10.1063/1.4705420

[74] Rühle, V., C. Junghans, A. Lukyanov, K. Kremer, and D. Andrienko (2009) "Versatile Object-Oriented Toolkit for Coarse-Graining Applications," *J. Chem. Theor. Comp.*, **5**(12), 3211–3223. DOI: 10.1021/ct900369w

[75] Lu, L., J. F. Dama, and G. A. Voth (2013) "Fitting Coarse-Grained Distribution Functions Through an Iterative Force-Matching Method," *J. Chem. Phys.*, **139**(12), 121906. DOI: 10.1063/1.4811667

[76] Zhou, J., I. F. Thorpe, S. Izvekov, and G. A. Voth (2007) "Coarse-Grained Peptide Modeling Using a Systematic Multiscale Approach," *Biophys. J.*, **92**(12), 4289–4303. DOI: 10.1529/biophysj.106.094425

[77] Thorpe, I. F., J. Zhou, and G. A. Voth (2008) "Peptide Folding Using Multiscale Coarse-Grained Models," *J. Phys. Chem. B*, **112**(41), 13079–13090. DOI: 10.1021/jp8015968

[78] Thorpe, I. F., D. P. Goldenberg, and G. A. Voth (2011) "Exploration of Transferability in Multiscale Coarse-Grained Peptide Models," *J. Phys. Chem. B*, **115**(41), 11911–26. DOI: 10.1021/jp204455g

[79] Krishna, V. and L. Larini (2011) "A Generalized Mean Field Theory of Coarse-Graining," *J. Chem. Phys.*, **135**(12), 124103. DOI: 10.1063/1.3638044

[80] Gumbart, J., L. G. Trabuco, E. Schreiner, E. Villa, and K. Schulten (2009) "Regulation of the Protein-Conducting Channel by a Bound Ribosome," *Structure*, **17**(11), 1453–1464. DOI: 10.1016/j.str.2009.09.010

[81] Best, R. B., N.-V. Buchete, and G. Hummer (2008) "Are Current Molecular Dynamics Force Fields Too Helical?" *Biophys. J.*, **395**(1), L07–09. DOI: 10.1529/biophysj.108.132696

[82] Voth, G. A. (ed.) (2009) *Coarse-Graining of Condensed Phase and Biomolecular Systems*, CRC Press, Boca Raton, FL USA.

[83] Partial special issue of J. Chem. Theory Comp. Volume 2, Issue 3 May 2006.

[84] Phys. Chem. Chem. Phys. Themed issue on coarse-grained modeling soft condensed matter. Volume 11, Issue 12 2009.

[85] Soft Matter themed issue on modeling soft matter. Volume 2, Issue 22 2009.

[86] Faraday Discussions on multiscale simulation of soft matter systems Volume 144 2010.

[87] Peter, C. and K. Kremer (2009) "Multiscale Simulation of Soft Matter Systems - From the Atomistic to the Coarse-Grained Level and Back," *Soft Matter*, **5**(22), 4357–4366. DOI: 10.1039/b912027k

[88] Peter, C. and K. Kremer (2010) "Multiscale Simulation of Soft Matter Systems," *Faraday Disc.*, **144**, 9–24. DOI: 10.1039/b919800h

[89] Levitt, M. and A. Warshel (1975) "Computer Simulation of Protein Folding," *Nature*, **253**(5494), 694–698. DOI: 10.1038/253694a0

[90] ten Wolde, P. R. and D. Chandler (2002) "Drying-Induced Hydrophobic Polymer Collapse," *Proc. Natl. Acad. Sci. USA*, **99**(10), 6539–6543. DOI: 10.1073/pnas.052153299

[91] Reynwar, B. J., G. Illya, V. A. Harmandaris, M. M. Muller, K. Kremer, and M. Deserno (2007) "Aggregation and Vesiculation of Membrane Proteins by Curvature-Mediated Interactions," *Nature*, **447**(7143), 461–464. DOI: 10.1038/nature05840

[92] Clementi, C. (2008) "Coarse-Grained Models of Protein Folding: Toy Models or Predictive Tools?" *Curr. Opin. Struc. Biol.*, **18**(1), 10–15. DOI: 10.1016/j.sbi.2007.10.005

[93] Fritz, D., C. R. Herbers, K. Kremer, and N. F. A. van der Vegt (2009) "Hierarchical Modeling of Polymer Permeation," *Soft Matter*, **5**(22), 4556–4563. DOI: 10.1039/b911713j

[94] Hagan, M. and D. Chandler (2006) "Dynamic Pathways for Viral Capsid Assembly," *Biophys. J.*, **91**(1), 42–54. DOI: 10.1529/biophysj.105.076851

[95] Hagler, A. T. and B. Honig (1978) "On the Formation of Protein Tertiary Structure on a Computer," *Proc. Natl. Acad. Sci. USA*, **75**(2), 554–558. DOI: 10.1073/pnas.75.2.554

[96] Ayton, G. S., W. G. Noid, and G. A. Voth (2007) "Multiscale Modeling of Biomolecular Systems: In Serial and in Parallel," *Curr. Opin. Struc. Biol.*, **17**(2), 192–198. DOI: 10.1016/j.sbi.2007.03.004

[97] Nielsen, S. O., C. F. Lopez, G. Srinivas, and M. L. Klein (2003) "A Coarse Grain Model for n-Alkanes Parameterized from Surface Tension Data," *J. Chem. Phys.*, **119**(14), 7043–7049. DOI: 10.1063/1.1607955

[98] Shinoda, W., R. Devane, and M. L. Klein (2007) "Multi-Property Fitting and Parameterization of a Coarse Grained Model for Aqueous Surfactants," *Mol. Sim.*, **33**(1-2), 27–36. DOI: 10.1080/08927020601054050

[99] Khurana, E., R. H. DeVane, A. Kohlmeyer, and M. L. Klein (2008) "Probing Peptide Nanotube Self-Assembly at a Liquid-Liquid Interface with Coarse-Grained Molecular Dynamics," *Nanoletters*, **8**(11), 3626–3630. DOI: 10.1021/nl801564m

[100] DeVane, R., M. L. Klein, C. C. Chiu, S. O. Nielsen, W. Shinoda, and P. B. Moore (2010) "Coarse-Grained Potential Models for Phenyl-Based Molecules: I. Parametrization Using Experimental Data," *J. Phys. Chem. B*, **114**(19), 6386–6393. DOI: 10.1021/jp9117369

[101] Chiu, C. C., R. DeVane, M. L. Klein, W. Shinoda, P. B. Moore, and S. O. Nielsen (2010) "Coarse-Grained Potential Models for Phenyl-Based Molecules: II. Application to Fullerenes," *J. Phys. Chem. B*, **114**(19), 6394–6400. DOI: 10.1021/jp9117375

[102] Marrink, S. J., A. H. de Vries, and A. E. Mark (2004) "Coarse Grained Model for Semiquantitative Lipid Simulations," *J. Phys. Chem. B*, **108**(2), 750–760. DOI: 10.1021/jp036508g

[103] Lopez, C. A., A. J. Rzepiela, A. H. de Vries, L. Dijkhuizen, P. H. Hunenberger, and S. J. Marrink (2009) "Martini Coarse-Grained Force Field: Extension to Carbohydrates," *J. Chem. Theor. Comp.*, **5**(12), 3195–3210. DOI: 10.1021/ct900313w

[104] Monticelli, L., E. Salonen, P. C. Ke, and I. Vattulainen (2009) "Effects of Carbon Nanoparticles on Lipid Membranes: A Molecular Simulation Perspective," *Soft Matter*, **5**(22), 4433–4445. DOI: 10.1039/b912310e

[105] Periole, X., M. Cavalli, S. J. Marrink, and M. A. Ceruso (2009) "Combining an Elastic Network with a Coarse-Grained Molecular Force Field: Structure, Dynamics, and Intermolecular Recognition," *J. Chem. Theor. Comp.*, **5**(9). DOI: 10.1021/ct9002114

[106] Lee, H., A. H. de Vries, S. J. Marrink, and R. W. Pastor (2009) "A Coarse-Grained Model for Polyethylene Oxide and Polyethylene Glycol: Conformation and Hydrodynamics," *J. Phys. Chem. B*, **113**(40). DOI: 10.1021/jp9058966

[107] Mognetti, B. M., L. Yelash, P. Virnau, W. Paul, K. Binder, M. Müller, and L. G. MacDowell (2008) "Efficient Prediction of Thermodynamic Properties of Quadrupolar Fluids from Simulation of a Coarse-Grained Model: The Case of Carbon Dioxide," *J. Chem. Phys.*, **128**(10), 104501. DOI: 10.1063/1.2837291

[108] Mognetti, B. M., P. Virnau, L. Yelash, W. Paul, K. Binder, M. Müller, and L. G. MacDowell (2009) "Coarse-Grained Models for Fluids and their Mixtures: Comparison of Monte Carlo Studies of their Phase Behavior with Perturbation Theory and Experiment," *J. Chem. Phys.*, **130**(4), 044101. DOI: 10.1063/1.3050353

[109] Liwo, A., S. Oldziej, M. R. Pincus, R. J. Wawak, S. Rackovsky, and H. A. Scheraga (1997) "A United-Residue Force Field for Off-Lattice Protein-Structure Simulations. 1. Functional forms and Parameters of Long-Range Side-Chain Interaction Potentials from Protein Crystal Data," *J. Comp. Chem.*, **18**(7), 849–873. DOI: 10.1002/(SICI)1096-987X(199705)18:7<849::AID-JCC1>3.0.CO;2-R

[110] Murtola, T., A. Bunker, I. Vattulainen, M. Deserno, and M. Karttunen (2009) "Multiscale Modeling of Emergent Materials: Biological and Soft Matter," *Phys. Chem. Chem. Phys.*, **11**(12), 1869–1892. DOI: 10.1039/b818051b

[111] Liwo, A., C. Czaplewski, J. Pillardy, and H. A. Scheraga (2001) "Cumulant-Based Expressions for the Multibody Terms for the Correlation between Local and Electrostatic Interactions in the United-Residue Force Field," *J. Chem. Phys.*, **115**(5), 2323–2347. DOI: 10.1063/1.1383989

[112] Tschop, W., K. Kremer, J. Batoulis, T. Burger, and O. Hahn (1998) "Simulation of Polymer Melts. I. Coarse-Graining Procedure for Polycarbonates," *Acta Poly.*, **49**(2-3), 61–74.
DOI: 10.1002/(SICI)1521-4044(199802)49:2/3<61::AID-APOL61>3.0.CO;2-V

[113] Baschnagel, J., K. Binder, P. Doruker, A. A. Gusev, O. Hahn, K. Kremer, W. L. Mattice, F. Muller-Plathe, M. Murat, W. Paul, S. Santos, U. W. Suter, and V. Tries (2000) "Bridging the Gap between Atomistic and Coarse-Grained Models of Polymers: Status and Perspectives," *Adv. Poly. Sci.*, **152**, 41–156.

[114] Harmandaris, V. A., N. P. Adhikari, N. F. A. van der Vegt, and K. Kremer (2006) "Hierarchical Modeling of Polystyrene: From Atomistic to Coarse-Grained Simulations," *Macromolecules*, **39**(19), 6708–6719. DOI: 10.1021/ma0606399

[115] Villa, A., C. Peter, and N. F. A. van der Vegt (2010) "Transferability of Nonbonded Interaction Potentials for Coarse-Grained Simulations: Benzene in Water," *J. Chem. Theor. Comp.*, **6**(8), 2434–2444. DOI: 10.1021/ct100228t

[116] Villa, A., N. F. A. van der Vegt, and C. Peter (2009) "Self-Assembling Dipeptides: Including Solvent Degrees of Freedom in a Coarse-Grained Model," *Phys. Chem. Chem. Phys.*, **11**(12), 2068–2076. DOI: 10.1039/b818146m

[117] Villa, A., C. Peter, and N. F. A. van der Vegt (2009) "Self-Assembling Dipeptides: Conformational Sampling in Solvent-Free Coarse-Grained Simulation," *Phys. Chem. Chem. Phys.*, **11**(12), 2077–2086. DOI: 10.1039/b818144f

[118] Schommers, W. (1983) "Pair Potentials in Disordered Many-Particle Systems - A Study of Liquid Gallium," *Phys. Rev. A*, **28**(6), 3599–3605.
DOI: 10.1103/PhysRevA.28.3599

[119] Reith, D., M. Putz, and F. Muller-Plathe (2003) "Deriving Effective Mesoscale Potentials from Atomistic Simulations," *J. Comp. Chem.*, **24**(13), 1624–1636.
DOI: 10.1002/jcc.10307

[120] Chaimovich, A. and M. S. Shell (2010) "Relative Entropy as a Universal Metric for Multiscale Errors," *Phys. Rev. E*, **81**(6,Part 1), 060104.
DOI: 10.1103/PhysRevE.81.060104

[121] Chaimovich, A. and M. S. Shell (2011) "Coarse-Graining Errors and Numerical Optimization Using a Relative Entropy Framework," *J. Chem. Phys.*, **134**(9), 094112. DOI: 10.1063/1.3557038

[122] Savelyev, A. and G. A. Papoian (2009) "Molecular Renormalization Group Coarse-Graining of Polymer Chains: Applications to Double-Stranded DNA," *Biophys. J.*, **96**(10), 4044–4052. DOI: 10.1016/j.bpj.2009.02.067

[123] Savelyev, A. and G. A. Papoian (2010) "Chemically Accurate Coarse Graining of Double-Stranded DNA," *Proc. Natl. Acad. Sci. USA*, **107**(47), 20340–20345. DOI: 10.1073/pnas.1001163107

[124] Murtola, T., M. Karttunen, and I. Vattulainen (2009) "Systematic Coarse Graining from Structure Using Internal States: Application to Phospholipid/Cholesterol Bilayer," *J. Chem. Phys.*, **131**(5), 055101. DOI: 10.1063/1.3167405

[125] Allen, E. C. and G. C. Rutledge (2009) "Evaluating the Transferability of Coarse-Grained, Density-Dependent Implicit Solvent Models to Mixtures and Chains," *J. Chem. Phys.*, **130**(3), 034904. DOI: 10.1063/1.3055594

[126] Allen, E. C. and G. C. Rutledge (2009) "Coarse-Grained, Density Dependent Implicit Solvent Model Reliably Reproduces Behavior of a Model Surfactant System," *J. Chem. Phys.*, **130**(20). DOI: 10.1063/1.3139025

[127] Chayes, J. T., L. Chayes, and E. H. Lieb (1984) "The Inverse Problem in Classical Statistical Mechanics," *Comm. Math. Phys.*, **93**(1), 57–121. DOI: 10.1007/BF01218639

[128] Chayes, J. T. and L. Chayes (1984) "On the Validity of the Inverse Conjecture in Classical Density Functional Theory," *J. Stat. Phys.*, **36**(3-4), 471–88. DOI: 10.1007/BF01010992

[129] Johnson, M. E., T. Head-Gordon, and A. A. Louis (2007) "Representability Problems for Coarse-Grained Water Potentials," *J. Chem. Phys.*, **126**(14), 144509. DOI: 10.1063/1.2715953

[130] Megariotis, G., A. Vyrkou, A. Leygue, and D. N. Theodorou (2011) "Systematic Coarse Graining of 4-Cyano-4-pentylbiphenyl," *Ind. Eng. Chem. Res.*, **50**(2), 546–556. DOI: 10.1021/ie901957r

[131] Izvekov, S., P. W. Chung, and B. M. Rice (2010) "The Multiscale Coarse-Graining Method: Assessing its Accuracy and Introducing Density Dependent Coarse-Grain Potentials," *J. Chem. Phys.*, **133**(6), 064109. DOI: 10.1063/1.3464776

[132] Lu, L. and G. A. Voth (2009) "Systematic Coarse-graining of a Multicomponent Lipid Bilayer," *J. Phys. Chem. B*, **113**(5), 1501–1510. DOI: 10.1021/jp809604k

[133] Larini, L., L. Y. Lu, and G. A. Voth (2010) "The Multiscale Coarse-Graining Method. VI. Implementation of Three-Body Coarse-Grained Potentials," *J. Chem. Phys.*, **132**(16), 164107. DOI: 10.1063/1.3394863

[134] Mullinax, J. W. and W. G. Noid (2010) "Reference State for the Generalized Yvon-Born-Green Theory: Application for Coarse-Grained Model of Hydrophobic Hydration," *J. Chem. Phys.*, **133**(12), 124107. DOI: 10.1063/1.3481574

[135] Hansen, J.-P. and I. R. McDonald (1990) *Theory of Simple Liquids*, 2 ed., Academic Press, San Diego, CA USA.

[136] Whittington, S. G. and L. G. Dunfield (1973) "A Born-Green-Yvon Treatment of Polymers with Excluded Volume," *J. Phys. A: Math., Nucl., Gen.*, **6**(4), 484. DOI: 10.1088/0305-4470/6/4/012

[137] Gubbins, K. E. (1980) "Structure of Nonuniform Molecular Fluids - Integrodifferential Equations for the Density-Orientation Profile," *Chem. Phys. Lett.*, **76**(2), 329–32. DOI: 10.1016/0009-2614(80)87034-5

[138] Attard, P. (1995) "Polymer Born-Green-Yvon Equation with Proper Triplet Superposition Approximation - Results for Hard-Sphere Chains," *J. Chem. Phys.*, **102**(13), 5411. DOI: 10.1063/1.469269

[139] Taylor, M. P. and J. E. G. Lipson (1993) "A Site-Site Born-Green-Yvon Equation for Hard Sphere Dimers," *J. Chem. Phys.*, **100**(1), 518. DOI: 10.1063/1.466966

[140] Taylor, M. P. and J. E. G. Lipson (1995) "A Born-Green-Yvon Equation for Flexible Chain-Molecule Fluids. I. General Formalism and Numerical Results for Short Hard-Sphere Chains," *J. Chem. Phys.*, **102**(5), 2118. DOI: 10.1063/1.468734

[141] Taylor, M. P. and J. E. G. Lipson (1995) "A Born-Green-Yvon Equation for Flexible Chain-Molecule Fluids. II. Applications to Hard-Sphere Polymers," *J. Chem. Phys.*, **102**(15), 6272. DOI: 10.1063/1.469073

[142] Andersen, H. C. and D. Chandler (1972) "Optimized Cluster Expansions for Classical Fluids. I. General Theory and Variational Formulation of the Mean Spherical Model and Hard Sphere Percus-Yevick Equations," *J. Chem. Phys.*, **57**(5), 1918–1929. DOI: 10.1063/1.1678512

[143] Andersen, H. C. and D. Chandler (1972) "Optimized Cluster Expansions for Classical Fluids. II. Theory of Molecular Liquids," *J. Chem. Phys.*, **57**(5), 1930–1937. DOI: 10.1063/1.1678513

[144] Schweizer, K. S. and J. G. Curro (1987) "Integral Equation Theory of the Structure of Polymer Melts," *Phys. Rev. Lett.*, **58**(3), 246–249. DOI: 10.1103/PhysRevLett.58.246

[145] Schweizer, K. S. and J. G. Curro (1997) "Integral Equation Theories of the Structure, Thermodynamics, and Phase Transitions of Polymer Fluids," *Adv. Chem. Phys.*, **93**, 1–142. DOI: 10.1002/9780470141571.ch1

[146] Mullinax, J. W. and W. G. Noid (2009) "Extended Ensemble Approach for Deriving Transferable Coarse-Grained Potentials," *J. Chem. Phys.*, **131**(10), 104110. DOI: 10.1063/1.3220627

[147] Tanaka, S. and H. A. Scheraga (1976) "Medium- and Long-Range Interaction Parameters between Amino Acids for Predicting Three-Dimensional Structures of Proteins," *Macromolecules*, **9**(6), 945–950. DOI: 10.1021/ma60054a013

[148] Sippl, M. J. (1995) "Knowledge-Based Potentials for Proteins," *Curr. Opin. Struc. Biol.*, **5**(2), 229–235. DOI: 10.1016/0959-440X(95)80081-6

[149] Jernigan, R. L. and I. Bahar (1996) "Structure-Derived Potential and Protein Simulations," *Curr. Opin. Struc. Biol.*, **6**(2), 195–209. DOI: 10.1016/S0959-440X(96)80075-3

[150] Mullinax, J. W. and W. G. Noid (2010) "Recovering Physical Potentials from a Model Protein Databank," *Proc. Natl. Acad. Sci. USA*, **107**(46), 19867–19872. DOI: 10.1073/pnas.1006428107

[151] Liu, P., S. Izvekov, and G. A. Voth (2007) "Multiscale Coarse-Graining of Monosaccharides," *J. Phys. Chem. B*, **111**(39), 11566–11575. DOI: 10.1021/jp0721494

[152] Ghosh, J. and R. Faller (2007) "State Point Dependence of Systematically Coarse-Grained Potentials," *Mol. Sim.*, **33**(9-10), 759–767. DOI: 10.1080/08927020701275050

[153] Vettorel, T. and H. Meyer (2006) "Coarse Graining of Short Polythylene Chains for Studying Polymer Crystallization," *J. Chem. Theor. Comp.*, **2**(3), 616–629. DOI: 10.1021/ct0503264

[154] Qian, H.-J., P. Carbone, X. Chen, H. A. Karimi-Varzaneh, C. C. Liew, and F. Müller-Plathe (2008) "Temperature-Transferable Coarse-Grained Potentials for Ethylbenzene, Polystyrene and their Mixtures," *Macromolecules*, **41**(24), 9919–29. DOI: 10.1021/ma801910r

[155] Carbone, P., H. A. K. Varzaneh, X. Chen, and F. Müller-Plathe (2008) "Transferability of Coarse-Grained Force Fields: The Polymer Case," *J. Chem. Phys.*, **128**(6), 064904. DOI: 10.1063/1.2829409

[156] Fischer, J., D. Paschek, A. Gieger, and G. Sadowski (2008) "Modeling of Aqueous Poly(oxyethylene) Solutions. 2. Mesoscale Simulations," *J. Phys. Chem. B*, **112**(43), 13561–13571. DOI: 10.1021/jp805770q

[157] Krishna, V., W. G. Noid, and G. A. Voth (2009) "The Multiscale Coarse-Graining Method. IV. Transferring Coarse-Grained Potentials between Temperatures," *J. Chem. Phys.*, **131**(2), 024103. DOI: 10.1063/1.3167797

[158] Ercolessi, F. and J. B. Adams (1994) "Interatomic Potentials from First-Principles Calculations: The Force-Matching Method," *Europhys. Lett.*, **26**(8), 583. DOI: 10.1209/0295-5075/26/8/005

[159] Izvekov, S., M. Parrinello, C. J. Burnham, and G. A. Voth (2004) "Effective Force Fields for Condensed Phase Systems from *Ab Initio* Molecular Dynamics Simulation: A New Method for Force-Matching," *J. Chem. Phys.*, **120**(23), 10896–10913. DOI: 10.1063/1.1739396

[160] Chorin, A. J., O. H. Hald, and R. Kupferman (2000) "Optimal Prediction and the Mori-Zwanzig Representation of Irreversible Processes," *Proc. Natl. Acad. Sci. USA*, **97**(7), 2968–73. DOI: 10.1073/pnas.97.7.2968

[161] Chorin, A. J. (2003) "Conditional Expectations and Renormalization," *Multiscale Model. Simul.*, **1**(1), 105–118. DOI: 10.1137/S1540345902405556

[162] Chorin, A. J. and O. H. Hald (2006) *Stochastic Tools in Mathematics and Science*, Springer, New York, NY USA.

[163] Ruiz-Montero, M. J., D. Frenkel, and J. J. Brey (1997) "Efficient Schemes to Compute Diffusive Barrier Crossing Rates," *Mol. Phys.*, **90**(6), 925–942. DOI: 10.1080/002689797171922

[164] Sprik, M. and G. Ciccotti (1998) "Free Energy from Constrained Molecular Dynamics," *J. Chem. Phys.*, **109**(18), 7737–7744. DOI: 10.1063/1.477419

[165] Ciccotti, G., R. Kapral, and E. Vanden-Eijnden (2005) "Blue Moon Sampling, Vectorial Reaction Coordinates, and Unbiased Constained Dynamics," *ChemPhysChem*, **6**(9), 1809–1814. DOI: 10.1002/cphc.200400669

[166] Hill, T. L. (1987) *Statistical Mechanics: Principles and Selected Applications*, Dover reprint.

[167] van der Spoel, D., E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. C. Berendsen (2005) "GROMACS: Fast, Flexible, and Free," *J. Comp. Chem.*, **26**(16), 1701–1718. DOI: 10.1002/jcc.20291

[168] Parrinello, M. and A. Rahman (1982) "Strain Fluctuations and Elastic Constants," *J. Chem. Phys.*, **76**(5), 2662–2666. DOI: 10.1063/1.443248

[169] Nose, S. (1984) "A Molecular Dynamics Method for Simulations in the Canonical Ensemble," *Mol. Phys.*, **52**(2), 255–268. DOI: 10.1080/00268978400101201

[170] Hoover, W. G. (1985) "Canonical Dynamics: Equilibrium Phase-Space Distributions," *Phys. Rev. A*, **31**(3), 1695–1697. DOI: 10.1103/PhysRevA.31.1695

[171] Berendsen, H. J. C., J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak (1984) "Molecular Dynamics with Coupling to an External Bath," *J. Chem. Phys.*, **81**(8), 3684–3690. DOI: 10.1063/1.448118

[172] Darden, T., D. York, and L. Pedersen (1993) "Particle Mesh Ewald: An N log(N) Method for Ewald Sums in Large Systems," *J. Chem. Phys.*, **99**(12), 8345–8348. DOI: 10.1063/1.464397

[173] Kashiwagi, H., T. Hashimoto, Y. Tanaka, H. Kubota, and T. Makita (1982) "Thermal Conductivity and Density of Toluene in the Temperature Range 273-373K at Pressures up to 250 MPa," *Int. J. Thermophysics*, **3**(3), 201–215. DOI: 10.1007/BF00503316

[174] Humphrey, W., A. Dalke, and K. Schulten (1996) "VMD: Visual Molecular Dynamics," *J. Mol. Graph.*, **14**(1), 33–38. DOI: 10.1016/0263-7855(96)00018-5

[175] Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery (1992) *Numerical Recipes in FORTRAN: The Art of Scientific Computing*, Cambridge University Press, New York, NY USA.

[176] Fritz, D., K. Koschke, V. A. Harmandaris, N. F. A. van der Vegt, and K. Kremer (2011) "Multiscale Modeling of Soft Matter: Scaling of Dynamics," *Phys. Chem. Chem. Phys.*, **13**(22). DOI: 10.1039/c1cp20247b

[177] Shirts, M. and V. S. Pande (2000) "Computing - Screen Savers of the World Unite!" *Science*, **290**(5498), 1903–1904. DOI: 10.1126/science.290.5498.1903

[178] Mervis, J. (2001) "Advanced Computing - NSF Launches TeraGrid for Academic Research," *Science*, **293**(5533), 1235–1236. DOI: 10.1126/science.293.5533.1235

[179] Okimoto, N., N. Futatsugi, H. Fuji, A. Suenaga, G. Morimoto, R. Yanai, Y. Ohno, T. Narumi, and M. Taiji (2009) "High-Performance Drug Discovery: Computational Screening by Combining Docking and Molecular Dynamics Simulations," *PLoS Comput. Biol.*, **5**(10), e1000528. DOI: 10.1371/journal.pcbi.1000528

[180] Phillips, J. C., R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kale, and K. Schulten (2005) "Scalable Molecular Dynamics with NAMD," *J. Comp. Chem.*, **26**(16), 1781–1802. DOI: 10.1002/jcc.20289

[181] Bowman, G. R., V. A. Voelz, and V. S. Pande (2011) "Atomistic Folding Simulations of the Five-Helix Bundle Protein lambda(6-85)," *J. Am. Chem. Soc.*, **133**(4), 664–667. DOI: 10.1021/ja106936n

[182] Peter, C. and K. Kremer (2010) "Multiscale Simulation of Soft Matter Systems," *Faraday Disc.*, **144**, 9–24. DOI: 10.1039/b919800h

[183] Soper, A. K. (1996) "Empirical Potential Monte Carlo Simulation of Fluid Structure," *Chem. Phys.*, **202**(2-3), 295–306. DOI: 10.1016/0301-0104(95)00357-6

[184] Swendsen, R. H. (1979) "Monte-Carlo Renormalization Group," *Phys. Rev. Lett.*, **42**(14), 859–861. DOI: 10.1103/PhysRevLett.42.859

[185] Kullback, S. and R. A. Leibler (1951) "On Information and Sufficiency," *Ann. Math. Stat.*, **22**(1), 79–86. DOI: 10.1214/aoms/1177729694

[186] Izvekov, S., A. Violi, and G. A. Voth (2005) "Systematic Coarse-Graining of Nanoparticle Interactions in Molecular Dynamics Simulation," *J. Phys. Chem. B*, **109**(36), 17019–17024. DOI: 10.1021/jp0530496

[187] Izvekov, S. and G. A. Voth (2006) "Multiscale Coarse-Graining of Mixed Phospholipid/Cholesterol Bilayers," *J. Chem. Theor. Comp.*, **2**(3), 637–648. DOI: 10.1021/ct050300c

[188] Ellis, C. R., J. F. Rudzinski, and W. G. Noid (2011) "Generalized-Yvon-Born-Green Model of Toluene," *Macromol. Theory Sim.*, **20**(7,SI), 478–495. DOI: 10.1002/mats.201100022

[189] Bradley, D. M. and R. C. Gupta (2002) "On the Distribution of the Sum of Non-Identically Distributed Uniform Random Variables," *Ann. Inst. Statist. Math.*, **54**(3), 689–700. DOI: 10.1023/A:1022483715767

[190] Espanol, P. and I. Zuniga (2011) "Obtaining Fully Dynamic Coarse-Grained Models from MD," *Phys. Chem. Chem. Phys.*, **13**(22), 10538–10545. DOI: 10.1039/c0cp02826f

[191] Hohenberg, P. and W. Kohn (1964) "Inhomogeneous Electron Gas," *Phys. Rev. B*, **136**(3B), 864–871. DOI: 10.1103/PhysRev.136.B864

[192] Mermin, N. D. (1965) "Thermal Properties of Inhomogeneous Electron Gas," *Phys. Rev. A*, **137**(5A), 1441–1443.

[193] Wolfram Research, Inc. (2010) *Mathematica Edition: Version 8.0*, Wolfram Research, Inc., Champaign, IL USA.

[194] Weeks, J. D. (2003) "External Fields, Density Functionals, and the Gibbs Inequality," *J. Stat. Phys.*, **110**(3-6), 1209–1218. DOI: 10.1023/A:1022157229397

[195] Wang, H., C. Junghans, and K. Kremer (2009) "Comparative Atomistic and Coarse-Grained Study of Water: What Do We Lose by Coarse-Graining?" *Eur. Phys. J. E*, **28**(2), 221–229. DOI: 10.1140/epje/i2008-10413-5

[196] Srinivas, G., S. Nielsen, P. Moore, and M. Klein (2006) "Molecular Dynamics Simulations of Surfactant Self-Organization at a Solid-Liquid Interface," *J. Am. Chem. Soc.*, **128**(3), 848–853. DOI: 10.1021/ja054846k

[197] Rzepiela, A. J., D. Sengupta, N. Goga, and S. J. Marrink (2010) "Membrane Poration by Antimicrobial Peptides Combining Atomistic and Coarse-Grained Descriptions," *Faraday Disc.*, **144**, 431–443. DOI: 10.1039/b901615e

[198] Lifson, S. and I. Oppenheim (1960) "Neighbor Interactions and Internal Rotations in Polymer Molecules. 4. Solvent Effect on Internal Rotations," *J. Chem. Phys.*, **33**(1), 109–115. DOI: 10.1063/1.1731064

[199] Dias, C. L., T. Ala-Nissila, M. Grant, and M. Karttunen (2009) "Three-Dimensional "Mercedes-Benz" Model for Water," *J. Chem. Phys.*, **131**(5), 054505. DOI: 10.1063/1.3183935

[200] Warren, P. (2003) "Vapor-Liquid Coexistence in Many-Body Dissipative Particle Dynamics," *Phys. Rev. E*, **68**(6,2), 066702. DOI: 10.1103/PhysRevE.68.066702

[201] Louis, A. A., P. G. Bolhuis, J. P. Hansen, and E. J. Meijer (2000) "Can Polymer Coils be Modeled as "Soft Colloids"," *Phys. Rev. Lett.*, **85**(12), 2522–2525. DOI: 10.1103/PhysRevLett.85.2522

[202] Bolhuis, P., A. Louis, J. Hansen, and E. Meijer (2001) "Accurate Effective Pair Potentials for Polymer Solutions," *J. Chem. Phys.*, **114**(9), 4296–4311. DOI: 10.1063/1.1344606

[203] Lu, L. and G. A. Voth (2011) "The Multiscale Coarse-Graining Method. VII. Free Energy Decomposition of Coarse-Grained Effective Potentials," *J. Chem. Phys.*, **134**(22), 224107. DOI: 10.1063/1.3599049

[204] Papaloannou, D., D. Zlakas, and C. Panayiotou (1991) "Volumetric Properties of Binary-Mixtures. 1. 2-Propanone + 2,2,4-Trimethylpentane and Normal-Heptane + Ethanol Mixtures," *J. Chem. Eng. Data*, **36**(1), 35–39. DOI: 10.1021/je00001a011

[205] Berendsen, H., J. Grigera, and T. Straatsma (1987) "The Missing Term in Effective Pair Potentials," *J. Phys. Chem.*, **91**(24), 6269–6271. DOI: 10.1021/j100308a038

[206] Williams, T., C. Kelley, and many others (2010), "Gnuplot 4.4: an interactive plotting program," http://gnuplot.sourceforge.net/.

[207] Anderson, E., Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. D. Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen (1999) *LAPACK Users' Guide*, SIAM, Philadelphia, PA USA.

[208] Izvekov, S. (2011) "Towards an Understanding of Many-Particle Effects in Hydrophobic Association in Methane Solutions," *J. Chem. Phys.*, **134**(3), 034104. DOI: 10.1063/1.3521480

[209] Cho, H. M., A. S. Gross, and J.-W. Chu (2011) "Dissecting Force Interactions in Cellulose Deconstruction Reveals the Required Solvent Versatility for Overcoming Biomass Recalcitrance," *J. Am. Chem. Soc.*, **133**(35), 14033–41. DOI: 10.1021/ja2046155

[210] Chu, J.-W., S. Izvekov, and G. A. Voth (2006) "The Multiscale Challenge for Biomolecular Systems: Coarse-Grained Modeling," *Mol. Sim.*, **32**(3-4), 211–218. DOI: 10.1080/08927020600612221

[211] Izvekov, S. and G. A. Voth (2009) "Solvent-Free Lipid Bilayer Model Using Multiscale Coarse Graining," *J. Phys. Chem. B*, **113**(13), 4443–4455. DOI: 10.1021/jp810440c

[212] Wang, Y. T. and G. A. Voth (2005) "Unique Spatial Heterogeneity in Ionic Liquids," *J. Am. Chem. Soc.*, **127**(35), 12192–12193. DOI: 10.1021/ja053796g

[213] Chen, Y. G., C. Kaur, and J. D. Weeks (2004) "Connecting Systems with Short and Long Ranged Interactions: Local Molecular Field Theory for Ionic Fluids," *J. Phys. Chem. B*, **108**(51), 19874–19884. DOI: 10.1021/jp0469261

[214] Rodgers, J. M., C. Kaur, Y. G. Chen, and J. D. Weeks (2006) "Attraction between Like-Charged Walls: Short-Ranged Simulations Using Local Molecular Field Theory," *Phys. Rev. Lett.*, **97**(9), 097801. DOI: 10.1103/PhysRevLett.97.097801

[215] Skolnick, J. (2006) "In Quest of an Empirical Potential for Protein Structure Prediction," *Curr. Opin. Struc. Biol.*, **16**(2), 166–171. DOI: 10.1016/j.sbi.2006.02.004

[216] Miyazawa, S. and R. L. Jernigan (1985) "Estimation of Effective Interresidue Contact Energies from Protein Crystal Structures: Quasichemical Approximation," *Macromolecules*, **18**(3), 534–552. DOI: 10.1021/ma00145a039

[217] Sippl, M. J. (1990) "Calculation of Conformational Ensembles from Potentials of Mean Force. An Approach to the Knowledge-Based Prediction of Local Structures in Globular Proteins," *J. Mol. Biol.*, **213**(4), 859–883. DOI: 10.1016/S0022-2836(05)80269-4

[218] Thomas, P. D. and K. A. Dill (1996) "Statistical Potentials Extracted from Protein Structures: How Accurate are They?" *J. Mol. Biol.*, **257**(2), 457–469. DOI: 10.1006/jmbi.1996.0175

[219] Betancourt, M. R. (2009) "Another Look at the Conditions for the Extraction of Protein Knowledge-Based Potentials," *Proteins*, **76**(1), 72–85. DOI: 10.1002/prot.22320

[220] Honeycutt, J. D. and D. Thirumalai (1990) "Metastability of the Folded States of Globular Proteins," *Proc. Natl. Acad. Sci. USA*, **87**(9), 3526–3529. DOI: 10.1073/pnas.87.9.3526

[221] Nielsen, S. O., C. F. Lopez, G. Srinivas, and M. L. Klein (2004) "Coarse Grain Models and the Computer Simulation of Soft Materials," *J. Phys.: Condens. Matter*, **16**(15), R481. DOI: 10.1088/0953-8984/16/15/R03

[222] Guenza, M. G. (2008) "Theoretical Models for Bridging Timescales in Polymer Dynamics," *J. Phys.: Condens. Matter*, **20**(3). DOI: 10.1088/0953-8984/20/03/033101

[223] Takada, S. (2012) "Coarse-Grained Molecular Simulations of Large Biomolecules," *Curr. Opin. Struc. Biol.*, **22**(2), 130–137. DOI: 10.1016/j.sbi.2012.01.010

[224] Riniker, S., J. R. Allison, and W. F. van Gunsteren (2012) "On Developing Coarse-Grained Models for Biomolecular Simulation: A Review," *Phys. Chem. Chem. Phys.*, **14**(36), 12423–30. DOI: 10.1039/c2cp40934h

[225] Brini, E., E. A. Algaer, P. Ganguly, C. Li, F. Rodriguez-Ropero, and N. F. A. van der Vegt (2013) "Systematic Coarse-Graining Methods for Soft Matter Simulations - A Review," *Soft Matter*, **9**(7), 2108–2119. DOI: 10.1039/C2SM27201F

[226] Reatto, L., D. Levesque, and J. J. Weis (1986) "Iterative Predictor-Corrector Method for Extraction of the Pair Interaction from Structural Data for Dense Classical Liquids," *Phys. Rev. A*, **33**(5), 3451–3465. DOI: 10.1103/PhysRevA.33.3451

[227] Májek, P. and R. Elber (2009) "A Coarse-Grained Potential for Fold Recognition and Molecular Dynamics Simulations of Proteins," *Prot. Struct. Func. Bioinfo.*, **76**(4), 822–836. DOI: 10.1002/prot.22388

[228] Bezkorovaynaya, O., A. Lukyanov, K. Kremer, and C. Peter (2012) "Multiscale Simulation of Small Peptides: Consistent Conformational Sampling in Atomistic and Coarse-Grained Models," *J. Comp. Chem.*, **33**(9), 937–949. DOI: 10.1002/jcc.22915

[229] Carmichael, S. P. and M. S. Shell (2012) "A New Multiscale Algorithm and its Application to Coarse-Grained Peptide Models for Self-Assembly," *J. Phys. Chem. B*, **116**(29), 8383–93. DOI: 10.1021/jp2114994

[230] Rudzinski, J. F. and W. G. Noid (2012) "The Role of Many-Body Correlations in Determining Potentials for Coarse-Grained Models of Equilibrium Structure," *J. Phys. Chem. B*, **116**(29), 8621–35. DOI: 10.1021/jp3002004

[231] Tschop, W., K. Kremer, O. Hahn, J. Batoulis, and T. Burger (1998) "Simulation of Polymer Melts. II. From Coarse-Grained Models Back to Atomistic Description," *Acta Poly.*, **49**(2-3), 75–79. DOI: 10.1002/(SICI)1521-4044(199802)49:2/3<75::AID-APOL75>3.0.CO;2-5

[232] Arkhipov, A., P. Freddolino, and K. Schulten (2006) "Stability and Dynamics of Virus Capsids Described by Coarse-Grained Modeling," *Structure*, **129**(12), 1767–1777. DOI: 10.1016/j.str.2006.10.003

[233] Harmandaris, V. A., D. Reith, N. F. A. Van der Vegt, and K. Kremer (2007) "Comparison between Coarse-Graining Models for Polymer Systems: Two Mapping Schemes for Polystyrene," *Macromol. Chem. Physic.*, **208**(19-20), 2109–2120. DOI: 10.1002/macp.200700245

[234] Zhang, Z. and G. A. Voth (2010) "Coarse-Grained Representations of Large Biomolecular Complexes from Low-Resolution Structural Data," *J. Chem. Theor. Comp.*, **6**(9), 2990–3002. DOI: 10.1021/ct100374a

[235] Sinitskiy, A. V., M. G. Saunders, and G. A. Voth (2012) "Optimal Number of Coarse-Grained Sites in Different Components of Large Biomolecular Complexes," *J. Phys. Chem. B*, **116**(29,SI), 8363–8374. DOI: 10.1021/jp2108895

[236] Gohlke, H. and M. F. Thorpe (2006) "A Natural Coarse Graining for Simulating Large Biomolecular Motion," *Biophys. J.*, **91**(6), 2115–2120. DOI: 10.1529/biophysj.106.083568

[237] Stepanova, M. (2007) "Dynamics of Essential Collective Motions in Proteins: Theory," *Phys. Rev. E*, **76**(5), 051918. DOI: 10.1103/PhysRevE.76.051918

[238] Zhang, Z. Y., L. Y. Lu, W. G. Noid, V. Krishna, J. Pfaendtner, and G. A. Voth (2008) "A Systematic Methodology for Defining Coarse-Grained Sites in Large Biomolecules," *Biophys. J.*, **95**(11), 5073–5083. DOI: 10.1529/biophysj.108.139626

[239] Guttenberg, N., J. F. Dama, M. G. Saunders, G. A. Voth, J. Weare, and A. R. Dinner (2013) "Minimizing Memory as an Objective for Coarse-Graining," *J. Chem. Phys.*, **138**(9), 094111. DOI: 10.1063/1.4793313

[240] Ramos-Estrada, M., G. Iglesias-Silva, and K. Hall (2006) "Experimental Measurements and Prediction of Liquid Densities for n-Alkane Mixtures," *J. Chem. Thermodyn.*, **38**(3), 337–347. DOI: 10.1016/j.jct.2005.05.020

[241] Yaws, C. L. (2008) *Thermophysical Properties of Chemicals and Hydrocarbons*, William Andrew, Norwich, NY USA.

[242] Chu, J.-W. and G. A. Voth (2006) "Coarse-Grained Modeling of the Actin Filament Derived from Atomistic-Scale Simulations," *Biophys. J.*, **90**, 1572–1582.

[243] Evans, R. (1990) "Comment on Reverse Monte Carlo Simulation," *Mol. Sim.*, **4**(6), 409–411. DOI: 10.1080/08927029008022403

[244] Morriss-Andrews, A. and J.-E. Shea (2014) "Simulations of Protein Aggregation: Insights from Atomistic and Coarse-grained Models," *J. Phys. Chem. Lett.*, **5**(11), 1899–1908. DOI: 10.1021/jz5006847

[245] Hyeon, C. and D. Thirumalai (2011) "Capturing the Essence of Folding and Functions of Biomolecules Using Coarse-Grained Models," *Nat. Commun.*, **2**, 487. DOI: 10.1038/ncomms1481

[246] Potoyan, D. A., A. Savelyev, and G. A. Papoian (2013) "Recent Successes in Coarse-Grained Modeling of DNA," *WIREs Comput. Mol. Sci.*, **3**(1), 69–83. DOI: 10.1002/wcms.1114

[247] Tozzini, V. (2005) "Coarse-Grained Models for Proteins," *Curr. Opin. Struc. Biol.*, **15**(2), 144–150. DOI: 10.1016/j.sbi.2005.02.005

[248] Tozzini, V., W. Rocchia, and J. McCammon (2006) "Mapping All-Atom Models Onto One-Bead Coarse-Grained Models: General Properties and Applications to a Minimal Polypeptide Model," *J. Chem. Theor. Comp.*, **2**(3), 667–673. DOI: 10.1021/ct050294k

[249] Tozzini, V. (2010) "Minimalist Models for Proteins: A Comparative Analysis," *Quart. Rev. Biophys.*, **43**(3), 333–371. DOI: 10.1017/S0033583510000132

[250] Purisima, E. and H. Scheraga (1984) "Conversion from a Virtual-Bond Chain to a Complete Polypeptide Backbone Chain," *Biopolymers*, **23**(7), 1207–1224. DOI: 10.1002/bip.360230706

[251] Rey, A. and J. Skolnick (1992) "Efficient Algorithm for the Reconstruction of a Protein Backbone from the Alpha-Carbon Coordinates," *J. Comp. Chem.*, **13**(4), 443–456. DOI: 10.1002/jcc.540130407

[252] Taketomi, H., Y. Ueda, and N. Go (1975) "Studies on Protein Folding, Unfolding and Fluctuations by Computer-Simulation. 1. Effect of Specific Amino-Acid Sequence Represented by Specific Inter-Unit Interactions," *Int. J. Pept. Protein Res.*, **7**(6), 445–459.

[253] Go, N. (1983) "Theoretical-Studies of Protein Folding," *Annu. Rev. Biophys. Bio.*, **12**, 183–210. DOI: 10.1146/annurev.bb.12.060183.001151

[254] Nymeyer, H., A. Garcia, and J. Onuchic (1998) "Folding Funnels and Frustration in Off-Lattice Minimalist Protein Landscapes," *Proc. Natl. Acad. Sci. USA*, **95**(11), 5921–5928. DOI: 10.1073/pnas.95.11.5921

[255] Tirion, M. M. (1996) "Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis," *Phys. Rev. Lett.*, **77**(9), 1905–1908. DOI: 10.1103/PhysRevLett.77.1905

[256] Bahar, I., A. R. Atilgan, and B. Erman (1997) "Direct Evaluation of Thermal Fluctuations in Proteins Using a Single-Parameter Harmonic Potential," *Fold. Des.*, **2**(3), 173–181. DOI: 10.1016/S1359-0278(97)00024-2

[257] Hills, R. D., Jr. and C. L. Brooks, III (2009) "Insights from Coarse-Grained Go Models for Protein Folding and Dynamics," *Int. J. Mol. Sci.*, **10**(3), 889–905. DOI: 10.3390/ijms10030889

[258] Whitford, P. C., K. Y. Sanbonmatsu, and J. N. Onuchic (2012) "Biomolecular Dynamics: Order-Disorder Transitions and Energy Landscapes," *Rep. Prog. Phys.*, **75**(7). DOI: 10.1088/0034-4885/75/7/076601

[259] Friedel, M., A. Baumketner, and J.-E. Shea (2006) "Effects of Surface Tethering on Protein Folding Mechanisms," *Proc. Natl. Acad. Sci. USA*, **103**(22), 8396–8401. DOI: 10.1073/pnas.0601210103

[260] Enciso, M. and A. Rey (2012) "Simple Model for the Simulation of Peptide Folding and Aggregation with Different Sequences," *J. Chem. Phys.*, **136**(21), 215103. DOI: 10.1063/1.4725883

[261] Enciso, M., C. Schutte, and L. Delle Site (2013) "A pH-Dependent Coarse-Grained Model for Peptides," *Soft Matter*, **9**(26), 6118–6127. DOI: 10.1039/C3SM27893J

[262] Ghavami, A., E. van der Giessen, and P. R. Onck (2013) "Coarse-Grained Potentials for Local Interactions in Unfolded Proteins," *J. Chem. Theor. Comp.*, **9**(1), 432–440. DOI: 10.1021/ct300684j

[263] Muller, M., K. Katsov, and M. Schick (2006) "Biological and synthetic membranes: What can be learned from a coarse-grained description?" *Phys. Rep.*, **434**(5-6), 113–176. DOI: 10.1016/j.physrep.2006.08.003

[264] Bereau, T. and M. Deserno (2009) "Generic coarse-grained model for protein folding and aggregation," *J. Chem. Phys.*, **130**, 235106.

[265] Ni, B. and A. Baumketner (2013) "Reduced Atomic Pair-Interaction Design (RAPID) Model for Simulations of Proteins," *J. Chem. Phys.*, **138**(6), 064102. DOI: 10.1063/1.4790160

[266] Hills, R. D., L. Y. Lu, and G. A. Voth (2010) "Multiscale Coarse-Graining of the Protein Energy Landscape," *PLoS Comput. Biol.*, **6**(6), e1000827. DOI: 10.1371/journal.pcbi.1000827

[267] Rudzinski, J. F. and W. G. Noid (2014) "Investigation of Coarse-Grained Mappings via an Iterative Generalized Yvon-Born-Green Method," *J. Phys. Chem. B*, **118**(28), 82958312. DOI: 10.1021/jp501694z

[268] Hess, B., H. Bekker, H. J. C. Berendsen, and J. G. E. M. Fraajie (1997) "LINCS: A linear constraint solver for molecular simulations," *J. Comp. Chem.*, **18**, 1463–72.

[269] Archambault, J., G. Pan, G. Dahmus, M. Cartier, N. Marshall, S. Zhang, M. Dahmus, and J. Greenblatt (1998) "FCP1, the RAP74-Interacting Subunit of a Human Protein Phosphatase that Dephosphorylates the Carboxyl-Terminal Domain of RNA Polymerase IIO," *J. Biol. Chem.*, **273**(42), 27593–27601. DOI: 10.1074/jbc.273.42.27593

[270] Nguyen, B., K. Abbott, K. Potempa, M. Kobort, J. Archambault, J. Greenblatt, P. Legault, and J. Omichinski (2003) "NMR Structure of a Complex Containing the TFIIF Subunit RAP74 and the RNA Polymerase II Carboxyl-Terminal Domain Phosphatase FCP1," *Proc. Natl. Acad. Sci. USA*, **100**(10), 5688–5693.

[271] Kamada, K., R. Roeder, and S. Burley (2003) "Molecular Mechanism of Recruitment of TFIIF-Associating RNA Polymerase C-Terminal Domain Phosphatase (FCP1) by Transcription Factor IIF," *Proc. Natl. Acad. Sci. USA*, **100**(5), 2296–2299. DOI: 10.1073/pnas.262798199

[272] Kumar, S., S. A. Showalter, and W. G. Noid (2013) "Native-Based Simulations of the Binding Interaction Between RAP74 and the Disordered FCP1 Peptide," *J. Phys. Chem. B*, **117**(11), 3074–3085. DOI: 10.1021/jp310293b

[273] Noel, J. K., P. C. Whitford, K. Y. Sanbonmatsu, and J. N. Onuchic (2010) "SMOG@ctbp: Simplified Deployment of Structure-Based Models in GROMACS," *Nucl. Acids Res.*, **38**(2), W657–W661. DOI: 10.1093/nar/gkq498

[274] Rühle, V. and C. Junghans (2011) "Hybrid Approaches to Coarse-Graining using the VOTCA Package: Liquid Hexane," *Macromol. Theory Sim.*, **20**(7,SI), 472–477. DOI: 10.1002/mats.201100011

[275] Shapiro, L. and G. Stockman (2001) *Computer Vision*, Prentice Hall.

[276] Lia, G., B. Spampinato, G. Maccari, and V. Tozzini (2014) "Minimalist Model for the Dynamics of Helical Polypeptides: A Statistic-Based Parametrization," *J. Chem. Theor. Comp.*, **10**(9), 38853895. DOI: 10.1021/ct5004059

[277] Larini, L. and J.-E. Shea (2012) "Coarse-Grained Modeling of Simple Molecules at Different Resolutions in the Absence of Good Sampling," *J. Phys. Chem. B*, **116**(29,SI), 8337–8349. DOI: 10.1021/jp2097263

[278] Kolinski, A., W. Galazka, and J. Skolnick (1996) "On the Origin of the Cooperativity of Protein Folding: Implications from Model Simulations," *Prot. Struct. Func. Bioinfo.*, **26**(3), 271–287. DOI: 10.1002/(SICI)1097-0134(199611)26:3<271::AID-PROT4>3.0.CO;2-H

[279] Klimov, D. K. and D. Thirumalai (1998) "Cooperativity in Protein Folding: From Lattice Models with Sidechains to Real Proteins," *Fold. Des.*, **9**(2), 127–139. DOI: 10.1016/S1359-0278(98)00018-2

[280] Kaya, H. and H. S. Chan (2000) "Polymer Principles of Protein Calorimetric Two-State Cooperativity," *Prot. Struct. Func. Bioinfo.*, **40**(4), 637–661. DOI: 10.1002/1097-0134(20000901)40:4<637::AID-PROT80>3.0.CO;2-4

[281] Bilionis, I. and N. Zabaras (2013) "A Stochastic Optimization Approach to Coarse-Graining using a Relative-Entropy Framework," *J. Chem. Phys.*, **138**(4), 044313. DOI: 10.1063/1.4789308

[282] Dunn, N. J., J. F. Rudzinski, and W. G. Noid (2015) "MS-CG/g-YBG Force Field Code Release (tentative title)," Manuscript in progress.

[283] Rudzinski, J. F. and W. G. Noid (2014) "Minimal Models for Disordered and Helical Peptide Ensembles," *J. Chem. Theor. Comp.*, Manuscript submitted 11/2014.

[284] Thomas, P. and K. Dill (1996) "An Iterative Method for Extracting Energy-Like Quantities from Protein Structures," *Proc. Natl. Acad. Sci. USA*, **93**(21), 11628–11633. DOI: 10.1073/pnas.93.21.11628

[285] Ben-Naim, A. (1997) "Statistical Potentials Extracted from Protein Structures: Are These Meaningful Potentials?" *J. Chem. Phys.*, **107**(9), 3698–3706. DOI: 10.1063/1.474725

[286] Betancourt, M. R. and D. Thirumalai (1999) "Pair Potentials for Protein Folding: Choice of Reference States and Sensitivity of Predicted Native States to Variations in the Interaction Schemes," *Protein Sci.*, **8**(2), 361–369.

[287] Lu, H. and J. Skolnick (2001) "A Distance-Dependent Atomic Knowledge-Based Potential for Improved Protein Structure Selection," *Prot. Struct. Func. Gen.*, **44**(3), 223–232. DOI: 10.1002/prot.1087

[288] Zhou, H. and Y. Zhou (2002) "Distance-Scaled, Finite Ideal-Gas Reference State Improves Structure-Derived Potentials of Mean Force for Structure Selection," *Protein Sci.*, **11**(11), 2714–2726. DOI: 10.1110/ps.0217002

[289] Shen, M. Y. and A. Sali (2006) "Statistical Potential for Assessment and Prediction of Protein Structures," *Protein Sci.*, **15**(11), 2507–2524. DOI: 10.1110/ps.062416606

[290] Rudzinski, J. F., K. Lu, S. T. Milner, J. K. Maranas, and W. G. Noid (2015) "Transferable Models of PEO-Based Ionomers Using the Extended Ensemble G-YBG Method (tentative title)," Manuscript in progress.

[291] Armand, M. (1983) "Polymer Solid Electrolytes - An Overview," *Solid State Ionics*, **9-10**(DEC), 745–754. DOI: 10.1016/0167-2738(83)90083-8

[292] Arico, A., P. Bruce, B. Scrosati, J. Tarascon, and W. Van Schalkwijk (2005) "Nanostructured Materials for Advanced Energy Conversion and Storage Devices," *Nat. Mater.*, **4**(5), 366–377. DOI: 10.1038/nmat1368

[293] Slater, M. D., D. Kim, E. Lee, and C. S. Johnson (2013) "Sodium-Ion Batteries," *Adv. Funct. Mater.*, **23**(8,SI), 947–958. DOI: 10.1002/adfm.201200691

[294] Bruce, P. and C. Vincent (1993) "Polymer Electrolytes," *J. Chem. Soc., Faraday Trans.*, **89**(17), 3187–3203. DOI: 10.1039/ft9938903187

[295] Laurer, J. and K. Winey (1998) "Direct Imaging of Ionic Aggregates in Zn-Neutralized Poly(ethylene-co-methacrylic acid) Copolymers," *Macromolecules*, **31**(25), 9106–9108. DOI: 10.1021/ma981503g

[296] Seitz, M. E., C. D. Chan, K. L. Opper, T. W. Baughman, K. B. Wagener, and K. I. Winey (2010) "Nanoscale Morphology in Precisely Sequenced Poly(ethylene-co-acrylic acid) Zinc Ionomers," *J. Am. Chem. Soc.*, **132**(23), 8165–8174. DOI: 10.1021/ja101991d

[297] Choi, U. H., M. Lee, S. Wang, W. Liu, K. I. Winey, H. W. Gibson, and R. H. Colby (2012) "Ionic Conduction and Dielectric Response of Poly(imidazolium acrylate) Ionomers," *Macromolecules*, **45**(9), 3974–3985. DOI: 10.1021/ma202784e

[298] Lin, K.-J. and J. K. Maranas (2012) "Cation Coordination and Motion in a Poly(ethylene oxide)-Based Single Ion Conductor," *Macromolecules*, **45**(15), 6230–6240. DOI: 10.1021/ma300716h

[299] Lu, K., J. F. Rudzinski, W. G. Noid, S. T. Milner, and J. K. Maranas (2014) "Scaling Behavior and Local Structure of Ion Aggregates in Single-Ion Conductors," *Soft Matter*, **10**(7), 978–989. DOI: 10.1039/c3sm52671b

[300] Cates, M. and S. Candau (1990) "Statics and Dynamics of Worm-Like Surfactant Micelles," *J. Phys.: Condens. Matter*, **2**(33), 6869–6892. DOI: 10.1088/0953-8984/2/33/001

[301] Das, A. and H. C. Andersen (2012) "The Multiscale Coarse-Graining Method. IX. A General Method for Construction of Three Body Coarse-Grained Force Fields," *J. Chem. Phys.*, **136**(19), 194114. DOI: 10.1063/1.4705417

[302] Buchete, N. V., J. E. Straub, and D. Thirumalai (2003) "Anisotropic Coarse-Grained Statistical Potentials Improve the Ability to Identify Native-Like Protein Structures," *J. Chem. Phys.*, **118**(16), 7658–7671. DOI: 10.1063/1.1561616

[303] Buchete, N. V., J. E. Straub, and D. Thirumalai (2004) "Orientational Potentials Extracted from Protein Structures Improve Native Fold Recognition," *Protein Sci.*, **13**(2), 862–874. DOI: 10.1016/j.polymer.2003.10.093

[304] Das, A. and H. C. Andersen (2012) "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**(19), 194113. DOI: 10.1063/1.4705384

[305] Liu, P., Q. Shi, H. Daume, and G. A. Voth (2008) "A Bayesian Statistics Approach to Multiscale Coarse Graining," *J. Chem. Phys.*, **129**(21), 214114. DOI: 10.1063/1.3033218

[306] Naome, A., A. Laaksonen, and D. P. Vercauteren (2014) "A Solvent-Mediated Coarse-Grained Model of DNA Derived with the Systematic Newton Inversion Method," *J. Chem. Theor. Comp.*, **10**(8), 3541–3549. DOI: 10.1021/ct500222s

Vita

Joseph F. Rudzinski

Born and raised in southern California, Joseph F. Rudzinski received Bachelors degrees in Chemistry and Mathematics from the University of California, Santa Barbara (UCSB) in 2009. At UCSB, Joseph worked for a total of three years as a research assistant in the experimental physical chemistry group of Steve Buratto and the applied mathematics group of Paul Atzberger. During this time, he worked on a diverse set of projects: optimizing the optical properties of electrochemically etched silicon Bragg mirrors, investigating electrochemical deposition of Pt on the efficiency of H_2 fuel cells, and modeling aptamer binding in micro-fluidic systems. Joseph's graduate mentors, Jimmy Odea and Dan Gargas played a critical role in his development as a researcher during this time.

Joseph began his graduate studies at The Pennsylvania State University in the theoretical chemistry group of Will Noid in the summer of 2009. In addition to his central research work, Joseph took on a number of other diverse challenges while in graduate school. These included performing the majority of his graduate coursework in the physics department, participating in several outreach projects, mentoring five undergraduate and three graduate students, and leading a graduate seminar in theoretical chemistry.

Joseph met his wife, Allison Fox, at UCSB in 2007. They have been married since August 2014 and have enjoyed traveling around the east coast during their time in State College. Joseph, Allison, and their two cats (Cauchy and Finley) will be relocating to Mainz, Germany, where Joseph has accepted a postdoctoral research appointment at the Max Planck Institute for Polymer Research beginning in January 2015.