The Pennsylvania State University
The Graduate School
Eberly College of Sciences

THERMODYNAMIC ACCURACY AND TRANSFERABILITY OF COARSE GRAINED MODELS AND APPLICATIONS TO PETROCHEMICAL SYSTEMS

A Dissertation in
Chemistry
by
Nicholas J. H. Dunn

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

December 2018

The dissertation of Nicholas J. H. Dunn was reviewed and approved* by the following:

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

Vincent H. Crespi
Professor of Physics
Professor of Materials Science and Engineering
Distinguished Professor
Professor of Chemistry

Kristen Fichthorn
Merrell Fenske Professor of Chemical Engineering
Professor of Physics

Michael Hickner
Professor of Materials Science and Engineering, Chemical Engineering
Corning Faculty Fellow
Professor of Chemistry

Philip Bevilacqua
Department Head of Chemistry
Distinguished Professor of Chemistry and Biochemistry and Molecular Biology

*Signatures are on file in the Graduate School.

Abstract

Asphaltenes are a problematic fraction of crude oil known for their propensity for aggregation during oil extraction and processing. This aggregation is an expensive problem for the petroleum industry, as it is difficult to predict, prevent, or reverse. Asphaltenes have been difficult to study via experimental methods due to the complexity of the asphaltene fraction as well as its propensity for aggregating at very low concentrations. As a result, computational techniques such as molecular dynamics (MD) simulations are an appealing approach to studying asphaltenes. However, simulating relatively large-scale, slowly evolving processes such as mesoscale aggregation at an all-atom (AA) level of resolution is infeasible using even the latest high-performance computing hardware. Coarse-grained (CG) modeling has emerged as a method of reducing the computational complexity of AA MD simulations by coarsening out degrees of freedom and grouping atoms together into CG sites. This reduction in resolution simplifies simulations using the resulting models, extending the length- and time-scales accessible by MD simulation.

In order to be useful for studying real systems, CG models must incorporate the correct physics to accurately describe the system they represent. There are two main methods for incorporating these physics into CG models via parameterization: top-down and bottom-up modeling. Top-down CG models use simple functional forms for their interaction potentials and are parameterized to reproduce experimentally observable properties of the target system. The resulting models accurately reproduce the targeted properties and are transferable to other state points, but may not accurately represent the fine structural details of the system. Bottom-up CG models are parameterized using information from simulations of an underlying AA model and may use more complex functional forms. The correct potential for a bottom-up CG model is the potential of mean force (PMF). The PMF contains all of the information necessary for the CG model to reproduce all properties of the AA model at the CG level of representation. However, the PMF is too complex to determine or use in simulation, so bottom-up models almost always use a potential that is an approximation to the configuration-dependent

portion of the PMF. As a result of neglecting the state-point dependence of the PMF, these approximations generally do not provide accurate descriptions of the thermodynamic properties of the underlying AA model and are not transferable away from the state point of their parameterization.

This work implements methods for improving the thermodynamic accuracy and transferability of bottom-up CG models and studies petrochemical systems as examples for demonstrating these methods. Chapters 2, 3, and 4 of this work examine the volume-dependence of the PMF using bottom-up CG models of the the petrochemical solvents heptane and toluene as example systems. We implement the volume-dependent pressure correction devised by Das and Andersen for use with bottom-up CG models, and demonstrate that this method obtains qualitative but not quantitative agreement with the PV equation of state of the underlying AA model. We extend this pressure-matching method with a self-consistent iterative procedure that generates CG models that quantitatively reproduce the PV equation of state of the underlying AA model. We further demonstrate this method for use in parameterizing a transferable CG model that is accurate across a range of system compositions. Chapter 5 presents the open-source release of the BOCS software package used to parameterize the bottom-up models in Chapters 2-4. Finally, Chapter 6 presents a top-down toy model for asphaltenes to study nanoaggregate formation over a range of solvent conditions and molecular structures.

Contents

List of Figures

List of Tables

Acknowledgments

It has been a long journey to get here, and I would like to thank those who have helped to guide me along the way. First and foremost, I must thank my parents for their early support of my interest in science and learning. Without their encouragement and support (and patience when I took apart household items to see how they worked) I would not have been nearly as well equipped for scientific research. My path has taken me far from home, but I feel their love and support behind me even now.

I was very lucky to have many excellent professors during my undergraduate studies at Union College. I'm deeply indebted to Prof. Mary Carroll, who provided an excellent research environment in the Aerogel Lab along with guidance on how to navigate the process of academic research. Mary's clear and insightful instruction during my time in the Aerogel Lab helped me to develop many of the skills I would need to be successful in graduate school. Mary continues to be a role model for me in her approach to research, and in her skill for clear and direct communication. I would also like to extend my sincere thanks to my academic advisor and materials chemistry professor Prof. Mike Hagerman, who helped me through the transition from a pre-med track to a research chemistry focus. I feel very lucky to have been able to take several classes in materials chemistry with Mike, as his infectious enthusiasm for the content inspired my ongoing interest in self-assembly processes. And I would like to recognize Prof. Valerie Barr for officially introducing me to programming during Computer Science 104 in my senior year, where I first caught the programming bug. It was a remark she made to me during her office hours about combining chemistry and computing in my future work that inspired me to consider joining computational chemistry groups in graduate school.

During my time at Penn State I have been lucky to work alongside some excellent and brilliant researchers in the Noid group. I'm extremely grateful to Joe Rudzinski for being a wonderful office-mate, mentor, role-model, and colleague. I greatly value our many conversations about the intuition of coarse-graining, and about debugging various models and programs. I am also grateful to Tommy Foley for

providing a physicist's perspective on many of my projects and results, and for frank discussions about the challenges of graduate school. Thanks also to Kate Lebold for putting my pressure-matching framework and tutorials to the test, and for patiently helping me to identify and iron out wrinkles in both. Many thanks to Michael DeLyser for his assistance in developing and debugging the BOCS software, and for taking over the lead developer role on this project after I departed.

I would like to express my deepest appreciation to my advisor, Will Noid. I am grateful that he took a risk on me when I arrived at Penn State as a naive experimentalist with a taste for programming. Will laid out an efficient route to competency in molecular modeling and statistical mechanics through a combination of coursework, working examples in group meetings, and drawing thematic connections in the literature. Will provided much appreciated guidance, support, and patience as I worked my way through this material. Will has also been an exemplary role model for me as I developed my research and communication habits throughout graduate school. In particular, I greatly admire his teaching ability, especially his talent for clearly and concisely formulating complex concepts. Will also has a real knack for providing focused and highly effective constructive criticism that I have missed since leaving Pennsylvania. And I am grateful for the opportunity to have learned from Will's unbending rigor in every aspect of research. Everything I have taken from Will's example as a researcher has helped to make me a more effective scientist, and I will be forever grateful for his influence and guidance.

Finally, I cannot begin to express my thanks to my wife, Em, for being my rock in these stormy seas. Twice now I have uprooted us for a cross-country move to find the next stage in my career, and twice now you have enthusiastically leaped with me into the unknown. Thank you for truly being my better half. This work would not have been possible without your unwavering support.

Chapter 1
Introduction

1.1 Molecular Simulation

Molecular dynamics (MD) simulation is a powerful technique for modeling chemical systems at a high level of detail. Standard all-atom (AA) MD simulations represent each atom in a system of interest as a point mass, and numerically integrate Newton's equations of motion to propagate the system through time.[1] The interactions between atoms in classical MD models are defined by force fields, which specify the functional form and parameters of the bonded and nonbonded interactions between atoms. Molecular-mechanics-type force fields are most commonly employed, and include terms for bond stretching, angle stretching, dihedral torsion, and nonbonded pair interactions. AA force fields also include a description of electrostatics, whether by appropriately assigning charges to the point masses or by providing a more sophisticated description that includes polarizability. However, AA MD simulations are generally limited by their computational complexity to studying processes that occur on nanometer length scales and sub-microsecond timescales.[2,3] Simulations at this scale can be useful for studying a wide variety of phenomena, but fall short of what is necessary for studying phenomena that occur on longer length- or timescales such as protein conformational changes or the mesoscale aggregation of molecules in a dilute solution.[4]

1.2 Coarse-Grained Modeling

Coarse-grained (CG) modeling is one method for extending the accessible length- and timescales of MD simulations. CG models reduce the computational effort required for simulation by grouping atoms together into CG sites and defining interactions between these sites, thus reducing the number of degrees of freedom in the resulting model.[5–8] However, a CG model must incorporate the correct physics in order to accurately model the underlying system. In practice, determining the physics for a CG model can be a highly nontrivial task.

Unlike the case of AA models, there are no fully general purpose CG force fields available for use with popular simulation packages. While there exist several popular CG force fields that can be applied across a range of systems,[9–12] these models are generally limited in scope to biomolecular simulations and do not provide a mechanism to freely adjust the level of coarsening. Instead, researchers typically define and parameterize bespoke CG models for their systems of interest. This is because CG models are generally system-specific, either because of the choice of resolution and CG site definitions, or because of the lack of transferability of the CG model away from the conditions where it was parameterized.

Two general approaches to parameterizing a CG model are "top-down" and "bottom-up" modeling. In top-down CG modeling, the interactions between sites take on simple functional forms with only a few parameters to tune. The parameter values are selected so that the CG model reproduces one or more experimentally observable properties of the target system, such as the surface tension, density, or aggregation behavior.[9,13–17] Top-down models accurately reproduce the thermodynamic properties targeted in their parameterization, but may not accurately reproduce structural properties.[18,19] Top-down models are also relatively transferable to other state points without losing accuracy.

In contrast, bottom-up models use microscopic information from simulations of a more detailed (often AA) model to parameterize the CG interactions. The resulting models typically more accurately describe the structural properties of the target system, but generally fail to capture its thermodynamic properties. In principle, the many-body potential of mean force (PMF) is the correct potential for bottom-up CG models, as it exactly incorporates all structural and thermodynamic properties of the underlying AA model at the resolution of the CG model.[20–25] The

PMF (W) is determined by the reference AA model and the CG mapping:

$$W(\mathbf{R}) = -k_B T \ln z(\mathbf{R}), \tag{1.1}$$

$$z(\mathbf{R}) = \int d\mathbf{r} \exp[-u(\mathbf{r}/k_B T)] \delta(\mathbf{M}(\mathbf{r}) - \mathbf{R})) \tag{1.2}$$

where $\mathbf{M}$ maps AA configurations $\mathbf{r}$ to their CG representation, $\mathbf{r}$ is the CG configuration, $u(\mathbf{r})$ is the potential for the AA model, and $\delta(\mathbf{M}(\mathbf{r}) - \mathbf{R})$ is zero unless $\mathbf{M}(\mathbf{r}) = \mathbf{R}$. Consequently, W weights each CG configuration $\mathbf{R}$ by the corresponding Boltzmann weight for those atomistic configurations $\mathbf{r}$ that map to $\mathbf{R}$. However, the PMF is intractably complex to represent or parameterize, and thus cannot be determined or used for simulation. As a result, bottom-up modeling almost always uses approximations to the configuration-dependent portion of the PMF. Due to the limitations of available simulation engines, these approximations most frequently employ standard molecular mechanics interaction sets. This can be thought of as projecting the PMF onto an incomplete basis set that is often insufficient to describe structural cross-correlations between CG interactions, or non-structural properties of the underlying AA system. Further, typical bottom-up approaches do not take the state-point dependence of the PMF into account. This leads to CG models that accurately reproduce the simple structure of the underlying AA model,[26–32] but which fail to accurately describe its thermodynamic properties or more complex structural correlations.[33,34] Further, this approximation is the cause of the poor transferability observed for bottom-up CG models, as an accurate approximation to the PMF at one state point may not be relevant at other state points.[33,35–43]

The majority of this dissertation focuses on method and software development centered around improving the thermodynamic accuracy of bottom-up CG models derived using the multiscale coarse-graining method developed by Izvekov and Voth.[27] In particular, we focus on a general solution for correcting the description of the pressure in CG models by implementing an extension of the pressure-matching method developed by Das and Andersen.[44] This method introduces an additional volume-dependent degree of freedom to the CG model that approximates the volume-dependence of the PMF. In this work we focus our attention on petrochemical solvents, but the methodology is applicable to any system composition. We demonstrate that this approach can be used to parameterize transferable bottom-

up CG models that reproduce both the structure and the pressure-volume behavior of the underlying AA model.

1.3 Simulation Studies of Asphaltenes

Asphaltenes are a class of molecules found in crude oil known for their high propensity for aggregation. In particular, they are an expensive problem for the petrochemical industry due to their tendency to crash out of solution and coat surfaces, fouling heat-exchange equipment and clogging pipes.[45–49] It is estimated that asphaltene aggregation costs the petrochemical industry billions of dollars a year.[50–52]

Asphaltenes are defined as the fraction of crude oil that is soluble in toluene but insoluble in n-heptane. This definition as a solubility class means that there are a wide variety of molecular structures represented in any given asphaltene sample.[45,53–57] This heterogeneity combined with their high propensity for aggregation has lead to difficulty in experimentally studying the structural properties of asphaltene molecules. As a result, the molecular properties of asphaltenes have been a topic of extensive debate,[46,47,58–60] although the Yen-Mullins model of asphaltenes has recently emerged as a popular description of asphaltene behavior.[57,61] The Yen-Mullins model for asphaltene aggregation proposes that asphaltenes are relatively small molecules with an average molecular weight of approximately 750 g/mol, comprised of a central aromatic core surrounded by alkyl tails.[57,61] Figure 1.1 shows a space-filling representation of a prototypical Yen-Mullins-type asphaltene molecule with the hydrogen atoms hidden. Further, the Yen-Mullins model proposes that these molecules aggregate via a hierarchical mechanism where 7-10 individual molecules form a nanoaggregate, which then cluster together and form networks of larger particles that eventually crash out of solution.[52,61]

MD simulations of asphaltenes are appealing for the purpose of studying the microscopic behavior of proposed model asphaltene compounds. The explicit tracking of atom positions throughout the simulation removes the ambiguity associated with studying asphaltenes through experimental methods. However, experimentally relevant asphaltene concentrations for studying the onset of aggregation range from 100 mg/L for the onset of nanoaggregate formation[62–67] to 2-5 g/L for the onset of cluster formation.[68–70] At these low concentrations, simulating enough asphaltene

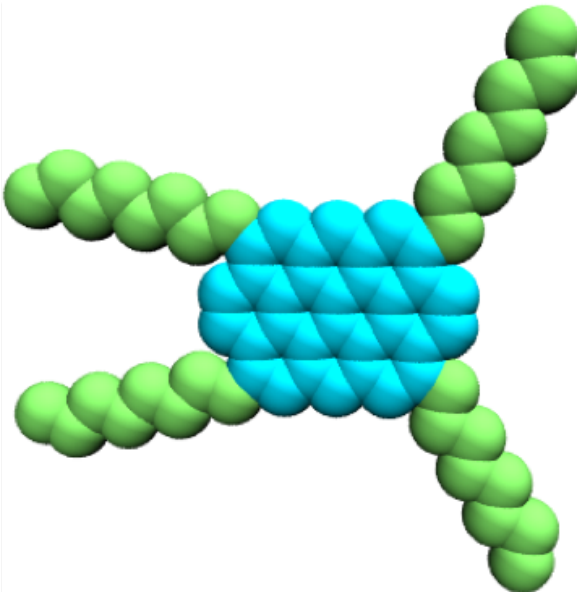

Figure 1.1. A prototypical Yen-Mullins-type asphaltene molecule with hydrogen atoms hidden. The aromatic core is colored in teal, while the alkyl tails are colored green.

molecules to form multiple nanoaggregates or clusters becomes prohibitively expensive, as the number of solvent molecules in such systems grows quickly with the number of asphaltenes.

CG models are therefore appealing in the study of asphaltenes, as they reduce the computational effort involved in MD simulations. We considered both bottom-up and top-down approaches for parameterizing CG models of asphaltenes and related molecules. The models of heptane and toluene presented in Chapters 2 and 3 were motivated by the need for an efficient, accurate model of asphaltene solvents. These bottom-up models were parameterized using a force- and pressure-matching methodology that approximated the configuration- and volume-dependent components of the PMF, and as a result accurately reproduced the structure and density distributions of the underlying models. The top-down toy model for asphaltene aggregation presented in Chapter 6 uses an implicit solvent model that greatly enhances the efficiency of the resulting simulations. We performed a parameter sweep for the nonbonded interactions in these models and examined the aggregation behavior for each set of parameters. Each set of parameters corresponds to a particular solvent quality for the asphaltene core and tail components. The particular molecular structures selected were designed to represent typical Yen-Mullins asphaltene structures.

1.4 Outline

This dissertation explores method development for bottom-up CG modeling, and targets molecules relevant to simulation of asphaltenes and other petrochemical systems. Chapter 2 examines the impact of coarse-graining on the description of the system pressure under the CG model, and implements a correction to the CG pressure that allows bottom-up CG models to reproduce the density and compressibility of the underlying AA models. As is typical for bottom-up CG models, without correction the example CG models of heptane and toluene overestimate the pressure of the underlying model with errors on the order of 10^3 bar. This work demonstrates that the pressure correction suggested by Das and Andersen[44] qualitatively but not quantitatively corrects the pressure of the CG model. To address this, we implement a self-consistent iterative solution for obtaining a pressure correction that allows CG models to quantitatively reproduce the pressure-volume equation of state of their reference AA model. Chapter 3 extends this correction to mixtures of heptane and toluene, using an extended ensemble approach that incorporates statistics from a range of systems with varying ratios of heptane:toluene to parameterize a single CG force field. The resulting CG force field and pressure correction optimally reproduces both the structure and thermodynamics of the heptane:toluene mixtures over the range compositions included in the ensemble. Chapter 4 describes a 'van der Waals' perspective on CG modeling that suggests bottom-up models can simultaneously reproduce both structural and thermodynamic properties of an AA model, given independent variational principles for the thermodynamic quantities of interest. Chapter 5 presents the BOCS software used for deriving the CG models of heptane and toluene in Chapter 2-4 to the research community as an open source project. Chapter 6 uses toy model asphaltenes to study nanoaggregate formation over a range of solvent conditions and molecular structures. Finally, Chapter 7 provides concluding thoughts and considers the outlook of CG modeling of petrochemical systems.

Chapter 2
Bottom-up coarse-grained models that accurately describe the structure, pressure, and compressibility of molecular liquids

N. J. H. Dunn, W. G. Noid, *J. Chem Phys* **2016**, 143 (24), 243148

2.1 Abstract

The present work investigates the capability of bottom-up coarse-graining methods for accurately modeling both structural and thermodynamic properties of all-atom (AA) models for molecular liquids. In particular, we consider 1, 2, and 3-site coarse-grained (CG) models for heptane, as well as 1 and 3-site CG models for toluene. For each model, we employ the multiscale coarse-graining (MS-CG) method to determine interaction potentials that optimally approximate the configuration dependence of the many-body potential of mean force (PMF). We employ a previously developed "pressure-matching" variational principle to determine a volume-dependent contribution to the potential, $U_V(V)$, that approximates the volume-dependence of the PMF. We demonstrate that the resulting CG models describe AA density fluctuations with qualitative, but not quantitative, accuracy. Accordingly, we develop a self-consistent approach for further optimizing U_V, such that the CG models accurately reproduce the equilibrium density, compressibility, and average pressure of the AA models, although the CG models still significantly underestimate the atomic pressure fluctuations. Additionally, by comparing this array of models that accurately describe the structure and thermodynamic pressure

of heptane and toluene at a range of different resolutions, we investigate the impact of bottom-up coarse-graining upon thermodynamic properties. In particular, we demonstrate that U_V accounts for the reduced cohesion in the CG models. Finally, we observe that bottom-up coarse-graining introduces subtle correlations between the resolution, the cohesive energy density, and the "simplicity" of the model.

2.2 Introduction

Atomically-detailed molecular dynamics (MD) simulations provide tremendous insight into processes that occur on nanometer length scales and sub-microsecond timescales.[71] Nevertheless, despite great strides in computational methods and resources, traditional all-atom (AA) simulations remain prohibitively inefficient for simulating phenomena that occur on larger length and time scales.[72] These limitations have motivated tremendous interest in coarse-grained (CG) models that provide much greater efficiency by representing systems in reduced detail.[5–8] Unfortunately, it remains challenging to develop CG models that provide a realistic description of both structural and thermodynamic properties.[73–75]

Studies with CG models often adopt either "top-down" or "bottom-up" strategies.[75] Top-down approaches generally model interactions with simple functional forms that are parameterized to reproduce thermodynamic properties or other experimental observables.[9,14–17,76] The resulting models typically demonstrate relatively good transferability for modeling a wide range of thermodynamic conditions. However, they may provide a relatively poor description of structural properties.[18,19]

In contrast, bottom-up approaches employ microscopic information from an underlying AA model to parameterize the interactions in the CG model.[75] In principle, the many-body potential of mean force (PMF) is the appropriate potential for bottom-up models, since it exactly incorporates all structural and thermodynamic properties of the AA model that can be observed at the resolution of the CG model.[20–22,24,25,77] In practice, though, the many-body PMF is too complex to determine, represent, or simulate. Consequently, bottom-up approaches often approximate the configuration-dependence of the PMF (at a single state point) with potentials that accurately reproduce structural features of the AA model, such as radial distribution functions.[26–32] Unfortunately, the resulting models frequently suffer from two major deficiencies.[74]

First of all, bottom-up models often demonstrate limited and unpredictable transferability. Because the many-body PMF is a function of the thermodynamic state point, potentials that accurately approximate the PMF at one state point may provide a poor approximation at other state points. Indeed, previous studies have demonstrated that optimized structure-based potentials can vary significantly with changes in composition, density, and temperature.[33,35–43]

Furthermore, bottom-up CG models often provide a poor description of thermodynamic properties.[33,34] The van der Waals picture of liquids provides a simple explanation for this deficiency, since the local structure of fluids is primarily determined by repulsive short-ranged interactions, but relatively insensitive to the attractive long-ranged interactions that are essential for describing thermodynamic properties.[78] Consequently, bottom-up methods that focus on local structural features likely provide limited accuracy for determining these long-ranged interactions.[79] Indeed, the resulting structure-based potentials tend to dramatically overestimate the internal pressure.[28,80] More fundamentally, by integrating over atomic degrees of freedom, bottom-up coarse-graining generates many-body interactions between CG sites and transfers thermodynamic information from the configuration space into the many-body PMF.[81,82] Accordingly, it may be necessary to explicitly account for these effects when computing the pressure and other thermodynamic properties with effective CG potentials.[23,83] In particular, recent studies with integral equation theories have indicated the ramifications of coarse-graining for describing thermodynamic properties with low resolution CG polymer models.[84–86]

In practice, though, bottom-up approaches frequently modify structure-based pair potentials in order to more accurately describe the thermodynamic pressure with the standard virial expression. For instance, the iterative Boltzmann method often modifies the structure-based pair potentials with a "linear ramp" correction that increases the intermolecular cohesion and, thus, reduces the internal pressure without significantly altering local structural features.[28,34,87] Similarly, the multiscale coarse-graining (MS-CG) method often employs a virial constraint in order to ensure that the calculated potentials accurately describe the pressure of the AA model.[27,88,89] Additionally, several groups have developed pair potentials that vary with the density.[37,89,90] However, linear ramp corrections can provide a poor description of the system compressibility,[34,87] while explicitly density-dependent

potentials can lead to thermodynamic inconsistencies.[37,91–93]

Quite recently, Das and Andersen (DA) proposed a considerably different approach for modeling density fluctuations with bottom-up CG models.[44] Rather than directly modifying the interactions between sites, DA introduced into the approximate CG potential a new term, U_V, that depends upon the volume but is independent of the configuration. Moreover, DA elegantly extended the MS-CG approach to the isothermal-isobaric ensemble by determining U_V via a variational "pressure-matching" calculation. This approach is appealing because the resulting CG potential provides a variationally optimal approximation for both the configuration-dependence and also the volume-dependence of the PMF.

DA demonstrated this pressure-matching method by developing a CG model for a monatomic liquid of Lennard-Jones spheres in which 75% of the particles had been eliminated. In this case, DA quite accurately reproduced the density distribution sampled by the original Lennard-Jones fluid. However, they did not demonstrate the method for more complex molecular systems.

In this work, we applied the DA pressure-matching method to develop CG models for liquid heptane and toluene at several different resolutions. For each system and for each resolution considered, the DA approach qualitatively, but not quantitatively, reproduces the density fluctuations of the corresponding AA model. Accordingly, we developed a self-consistent approach that quantitatively reproduces the thermodynamic density and compressibility of the AA models, while also preserving an accurate description of the AA structure, for each liquid at each resolution. Given the resulting set of models, we examine the impact of resolution upon the thermodynamic behavior of CG models. We demonstrate that the magnitude of the pressure correction, $F_V = -\mathrm{d}U_V/\mathrm{d}V$, systematically increases with coarsening and that this pressure correction correlates with the reduced cohesion of the CG model. Additionally, we demonstrate that, even when accurately reproducing atomic density fluctuations, CG models significantly underestimate the pressure fluctuations of AA models, which has important ramifications for calculating material properties with structure-based, bottom-up models. Finally, we quantitatively assess the extent to which systematic bottom-up coarse-graining "simplifies" the original atomic models.

2.3 Theory

2.3.1 High resolution model

We first consider the isothermal-isobaric ensemble for a high resolution model with n atoms at constant temperature T and external pressure P_0.[94] In the following, we assume that the atomic and CG models do not include rigid constraints, although we anticipate that this assumption could be easily relaxed.[25,95] The Hamiltonian for the atomic model may be expressed:

$$h(\mathbf{r}, \mathbf{p}; V) = \kappa(\mathbf{p}) + u(\mathbf{r}; V). \tag{2.1}$$

The atomic kinetic energy, $\kappa(\mathbf{p})$, is a function of the Cartesian momenta for the n atoms, $\mathbf{p} = (\mathbf{p}_1, \ldots, \mathbf{p}_n)$,

$$\kappa(\mathbf{p}) = \sum_{i=1}^{n} \mathbf{p}_i^2 / 2m_i \tag{2.2}$$

where m_i is the mass of atom i. The atomic potential energy is a function of the Cartesian coordinates for all n atoms, $\mathbf{r} = (\mathbf{r}_1, \ldots, \mathbf{r}_n)$, and may usually be decomposed

$$u(\mathbf{r}; V) = u_2(\mathbf{r}) + u_\theta(\mathbf{r}) + u_V(V). \tag{2.3}$$

In this expression, u_2 is the set of potentials that depend upon the distance between pairs of particles and includes both intramolecular and intermolecular potentials. The second term, u_θ, is the set of bonded potentials that depend upon bending or dihedral angles, but do not depend upon inter-particle distances. These two contributions in Eq. (2.3) depend upon the system volume, V, only implicitly via periodic boundary conditions. In contrast, u_V is independent of $\mathbf{r}$, but explicitly depends upon V. This term arises in, e.g., the correction that accounts for truncating non-bonded pair potentials at a cutoff distance.[1,96]

The extended phase space distribution for $\mathbf{r}$, $\mathbf{p}$, and V factors into two statistically independent contributions. The atomic momenta are statistically independent Gaussian random variables,

$$p_p(\mathbf{p}) \quad \propto \quad \exp\left[-\beta\kappa(\mathbf{p})\right] \tag{2.4}$$

where $\beta = 1/k_B T$. The atomic configuration, $\mathbf{r}$, and the volume, V, are coupled random variables:

$$p_{rV}(\mathbf{r}, V|P_0) \quad \propto \quad \exp\left[-\beta\left(u(\mathbf{r};V) + P_0 V\right)\right] \tag{2.5}$$

for all $\mathbf{r}$ in the V-dependent atomic configuration space, $D_{AA}(V)$.

The instantaneous internal pressure of the atomic model, P_{AA}, may be expressed[94]

$$P_{AA}(\mathbf{r}, \mathbf{p}, V) = \frac{2}{3V}\kappa(\mathbf{p}) + \frac{1}{3V}w_2(\mathbf{r}) + f_V(V). \tag{2.6}$$

where

$$w_2(\mathbf{r}) \quad = \quad \sum_{i>j} r_{ij} f_{2;ij}(r_{ij}) \tag{2.7}$$

$$f_V(V) \quad = \quad -\mathrm{d}u_V(V)/\mathrm{d}V. \tag{2.8}$$

In Eq. (2.7), r_{ij} is the distance between atoms i and j, while $f_{2;ij}$ is the magnitude of the corresponding force due to $u_2(\mathbf{r})$ in Eq. (2.3). Note that u_θ does not directly contribute to the scalar virial since it is invariant with uniform scaling of the coordinates.[97] The average of P_{AA} equals the thermodynamic pressure P_0:

$$\langle P_{AA}(\mathbf{r}, \mathbf{p}, V)\rangle = P_0 \tag{2.9}$$

where the angular brackets indicate an average according to $p_{rV}(\mathbf{r}, V|P_0)p_p(\mathbf{p})$:

$$\langle a(\mathbf{r}, \mathbf{p}, V)\rangle = \int_0^\infty \mathrm{d}V \int_{V^n} \mathrm{d}\mathbf{r} \int \mathrm{d}\mathbf{p}\; p_{rV}(\mathbf{r}, V|P_0)\; p_p(\mathbf{p})\; a(\mathbf{r}, \mathbf{p}, V). \tag{2.10}$$

In Eq. (2.10) and in the following, V^n indicates an integration over the volume-dependent atomic configuration space $D_{AA}(V)$.

2.3.2 Mapped ensemble and consistency criteria

In order to relate the atomistic and CG models, we introduce a mapping, $\mathbf{M}$, that determines a CG configuration, $\mathbf{R} = (\mathbf{R}_1, ..., \mathbf{R}_N)$, for N sites as a linear function of the atomic configuration, $\mathbf{r}$, by determining the coordinates of each CG site I,

i.e.,

$$\mathbf{R}_I = \mathbf{M}_I(\mathbf{r}) = \sum_{i=1}^{n} c_{Ii}\mathbf{r}_i. \tag{2.11}$$

Equation (2.11) assumes appropriate application of periodic boundary conditions and also that the mapping coefficients are normalized: $\sum_i c_{Ii} = 1$ for each $I = 1, \ldots, N$. For simplicity, we assume that this mapping partitions the atoms into disjoint sets and associates each CG site with the mass center for one of these atomic groups. However, this last assumption can be readily relaxed by appropriate definition of the CG masses.[25] The configuration mapping operator, $\mathbf{M}$, also implies a corresponding momentum mapping operator, $\mathbf{M_P}$, that determines the CG momenta $\mathbf{P} = (\mathbf{P}_1, ..., \mathbf{P}_N)$ as a function of the atomic momenta $\mathbf{p}$.[25] We define the "mapped ensemble" by mapping each atomic microstate, $(\mathbf{r}, \mathbf{p}, V)$, to its CG representation, $(\mathbf{R}, \mathbf{P}, V)$, i.e., $\mathbf{r} \to \mathbf{R} = \mathbf{M}(\mathbf{r})$, $\mathbf{p} \to \mathbf{P} = \mathbf{M_P}(\mathbf{p})$, and $V \to V$.

The CG momenta, $\mathbf{P}_I$, are statistically independent Gaussian random variables in the mapped ensemble:

$$p_P(\mathbf{P}) = \int \mathrm{d}\mathbf{p}\, p_p(\mathbf{p})\, \delta(\mathbf{P} - \mathbf{M_P}(\mathbf{p})) \tag{2.12}$$

$$\propto \exp\left[-\sum_{I=1}^{N} \mathbf{P}_I^2/2\sigma_I^2\right] \tag{2.13}$$

where the variances, σ_I^2, are the sum of the variances for the corresponding atomic momenta.[25,98] The mapped ensemble distribution for CG configurations and volumes is given by

$$p_{RV}(\mathbf{R}, V|P_0) = \int_{V^n} \mathrm{d}\mathbf{r}\, p_{rV}(\mathbf{r}, V|P_0)\, \delta(\mathbf{R} - \mathbf{M}(\mathbf{r})). \tag{2.14}$$

Our objective is to develop a CG model that accurately reproduces these probability distributions for the mapped ensemble. The mapped momenta distribution is easily reproduced by determining each site mass to generate the correct variance. The mapped distribution of configurations and volumes is more difficult to reproduce. The correct CG potential, W, for exactly matching this distribution is defined to within a constant (that is independent of both V and $\mathbf{R}$) by

$$\exp\left[-\beta W(\mathbf{R}, V)\right] = V_0^{N-n} \int_{V^n} \mathrm{d}\mathbf{r} \exp\left[-\beta u(\mathbf{r}; V)\right] \delta(\mathbf{R} - \mathbf{M}(\mathbf{r})), \tag{2.15}$$

where V_0 is an arbitrary reference volume that has been introduced for dimensional consistency, while ensuring that $W(\mathbf{R}, V)$ properly accounts for the "ideal gas" contribution to the pressure. Thus, W is a volume-dependent many-body potential of mean force (PMF):

$$-\partial W(\mathbf{R}, V)/\partial \mathbf{R}_I = \langle \mathbf{f}_I(\mathbf{r}) \rangle_{\mathbf{R};V} \tag{2.16}$$

$$-\partial W(\mathbf{R}, V)/\partial V = (n - N)k_B T/V + \langle w_2(\mathbf{r}) \rangle_{\mathbf{R};V} + f_V(V). \tag{2.17}$$

In these expressions, $\mathbf{f}_I(\mathbf{r})$ is the net force on CG site I in the configuration $\mathbf{r}$, while the subscripted angular brackets denote a conditioned average for the atomic model:

$$\langle a(\mathbf{r}, V) \rangle_{\mathbf{R};V} = \int_{V^n} \mathrm{d}\mathbf{r}\, p_{rV}(\mathbf{r}, V|P_0)\, \delta(\mathbf{R} - \mathbf{M}(\mathbf{r}))\, a(\mathbf{r}, V)\, /p_{RV}(\mathbf{R}, V|P_0)\, . \tag{2.18}$$

It is worth noting that Eq. (2.17) may be re-expressed:

$$-\partial W(\mathbf{R}, V)/\partial V = \langle P_{AA}(\mathbf{r}, \mathbf{p}, V) \rangle_{\mathbf{R};V} - N k_B T/V. \tag{2.19}$$

Thus, the many-body PMF, W, incorporates the non-ideality that is necessary for the CG model to reproduce the AA equation of state for the pressure. In particular, in the special case that the AA model is an ideal gas with $u(\mathbf{r}; V) = 0$, the PMF $W(\mathbf{R}, V) = -(n - N)k_B T \ln(V/V_0)$ accounts for the "apparent non-ideality" that emerges in the CG model from integrating over the atomic degrees of freedom.[44] Additionally, we note that W also depends upon T, although this dependence is not explicitly considered in the present work.[82]

2.3.3 Approximate coarse-grained model

We next consider the isothermal-isobaric ensemble for an approximate CG model with N sites at the same temperature T and external pressure P_0. The Hamiltonian for the model is

$$H(\mathbf{R}, \mathbf{P}, V) = \mathscr{K}(\mathbf{P}) + U(\mathbf{R}, V). \tag{2.20}$$

The kinetic energy is

$$\mathscr{K}(\mathbf{P}) = \sum_{I=1}^{N} \mathbf{P}_I^2/2M_I. \tag{2.21}$$

The mass, M_I, of site I is defined as the total mass of the corresponding atoms so that the CG model will reproduce the corresponding mapped momentum distribution, p_P.[25]

In general, one expects that the many-body PMF, $W(\mathbf{R}, V)$, couples the CG coordinates and volume in a complex way. However, in practice, CG models approximate the PMF with a potential, U, that is analogous to Eq. (2.3):

$$U(\mathbf{R}, V) = U_2(\mathbf{R}) + U_\theta(\mathbf{R}) + U_V(V), \tag{2.22}$$

where U_2 is a sum of bonded and non-bonded potentials that depend upon pair distances, while U_θ is a sum of bonded potentials that depend upon bond angles or dihedral angles. The first two contributions in Eq. (2.22) depend upon the system volume, V, only implicitly via periodic boundary conditions. In contrast, U_V is independent of $\mathbf{R}$, but explicitly depends upon V. Because CG interactions are often assumed to be short-ranged, previous studies have generally neglected this contribution. However, U_V is the key contribution to the CG effective potential for the DA pressure-matching algorithm.[44]

Given Eqs. (2.20)-(2.22), the instantaneous internal pressure of the CG model, P_{CG} may be expressed

$$P_{CG}(\mathbf{R}, \mathbf{P}, V) = \frac{2}{3V}\mathscr{K}(\mathbf{P}) + \frac{1}{3V}\mathscr{W}_2(\mathbf{R}) + F_V(V) \tag{2.23}$$

where

$$\mathscr{W}_2(\mathbf{R}) = \sum_{I>J} R_{IJ} F_{2;IJ}(R_{IJ}) \tag{2.24}$$

$$F_V(V) = -\mathrm{d}U_V(V)/\mathrm{d}V \tag{2.25}$$

and R_{IJ} is the distance between sites I and J, while $F_{2;IJ}$ is the corresponding force magnitude due to $U_2(\mathbf{R})$ in Eq. (2.22). For future use, we also define

$$P^0_{CG}(\mathbf{R}, \mathbf{P}) = \frac{2}{3V}\mathscr{K}(\mathbf{P}) + \frac{1}{3V}\mathscr{W}_2(\mathbf{R}), \tag{2.26}$$

which is the equation for the instantaneous pressure in conventional structure-based potentials that neglect U_V, i.e. if $U = U_2 + U_\theta$. The instantaneous pressure of the

CG model may then be expressed:

$$P_{CG}(\mathbf{R},\mathbf{P},V) = P^0_{CG}(\mathbf{R},\mathbf{P}) + F_V(V). \tag{2.27}$$

2.3.4 Pressure matching variational principle

DA demonstrated an elegant extension of the MS-CG variational approach[25,27,88,99] to determine potentials that optimally approximate both the configuration-dependence and also the volume-dependence of the many-body PMF, $W(\mathbf{R},V)$, for the isothermal-isobaric ensemble.[44] In practice, this involves the successive minimization of two functionals. DA proposed to first determine the configuration-dependent potentials, U_2 and U_θ, by minimizing

$$\chi^2_1[U_2,U_\theta] = \left\langle V^{2/3} \sum_{I=1}^{N} \left| \mathbf{f}_I(\mathbf{r};V) - \mathbf{F}_I(\mathbf{M}(\mathbf{r});V) \right|^2 \right\rangle, \tag{2.28}$$

where the angular brackets denote an average according to Eq. (2.10). The individual terms in U_2 and U_θ are represented by simple basis functions,[100] such as spline functions, and the corresponding parameters are obtained by minimizing χ^2_1, e.g., by solving the corresponding normal system of linear equations.[101] The functional χ^2_1 slightly differs from the standard MS-CG force-matching functional[25] by the factor of $V^{2/3}$ that reweights each configuration. This factor arises because DA developed the pressure-matching variational principle in scaled coordinates.[102]

After determining the intra- and intermolecular contributions to the approximate CG potential, DA proposed to parameterize U_V by minimizing a second functional:

$$\chi^2_2[U_V|U_2,U_\theta] = \left\langle \left| \delta P_0(\mathbf{r},\mathbf{p},V) - F_V(V) \right|^2 \right\rangle \tag{2.29}$$

where

$$\delta P_0(\mathbf{r},\mathbf{p},V) = P_{AA}(\mathbf{r},\mathbf{p},V) - P^0_{CG}(\mathbf{M}(\mathbf{r}),\mathbf{M}_\mathbf{P}(\mathbf{p})). \tag{2.30}$$

By minimizing χ^2_2 with respect to U_V at constant U_2 and U_θ, the pressure-matching algorithm determines the necessary additional contribution to the pressure such that the approximate CG potential reproduces the atomic pressure equation of state when averaged over the configurations sampled by the atomic model.

We note two slight differences between Eq. (2.29) and the functional proposed

by DA: 1) Equation (2.29) explicitly includes the fluctuating kinetic contribution to the instantaneous pressure, while DA employed the average of the kinetic contributions and determined U_V by minimizing the difference in the instantaneous virials. Of course, these two procedures are equivalent for determining U_V because the momenta are statistically independent of the configuration and the volume. 2) The functional in Eq. (2.29) accounts for contributions to the atomic pressure from $u_V(V)$, which were not considered by DA.

In order to numerically minimize χ_2^2, DA proposed representing U_V as a polynomial in $V/\bar{v}$ where $\bar{v}$ is the average volume of the atomic model:

$$U_V(V) = \sum_d \psi_d w_d(V), \tag{2.31}$$

where

$$w_d(V) = \begin{cases} N(V/\bar{v}) & \text{for } d = 1 \\ N\left(V/\bar{v} - 1\right)^d & \text{for } d \geq 2 \end{cases} \tag{2.32}$$

Given this functional form for U_V, χ_2^2 becomes a quadratic form in ψ_d that we minimize by solving the resulting normal system of linear equations.

2.3.5 Iterative correction

In order to obtain quantitative agreement between the density fluctuations of the atomic and CG models, we developed a simple iterative approach for further optimizing U_V. After determining U_2 and U_θ by minimizing χ_1^2, we determined our initial estimate, $U_V^{(0)}$, for U_V by minimizing χ_2^2, as proposed by DA. We then performed simulations that sampled the constant NPT ensemble for a CG model with the potential $U^{(0)} = U_2 + U_\theta + U_V^{(0)}$. From these simulations, we estimated the pressure equation of state for the CG model, $\bar{P}_{CG}^{(0)}(V)$. We then estimated the error in the CG equation of state, $\delta\bar{P}(V)$, by comparison with the AA equation of state, $\bar{P}_{AA}(V)$:

$$\delta\bar{P}(V) = \bar{P}_{AA}(V) - \bar{P}_{CG}^{(0)}(V) \tag{2.33}$$

Since F_V directly impacts the pressure in Eq. (2.23), we employed this error to determine a new estimate for U_V: $U_V^{(1)} = U_V^{(0)} + \delta U_V$ where $\delta\bar{P}(V) = \delta F_V(V) = -\mathrm{d}\delta U_V(V)/\mathrm{d}V$. Simulations of the CG model with the modified potential, $U^{(1)} = U_2 + U_\theta + U_V^{(1)}$, determine a new equation of state $\bar{P}_{CG}^{(1)}(V)$, which is compared again

with the atomic equation of state to determine a new correction. This procedure is iterated until the CG model satisfactorily reproduces the atomic equation of state.

2.4 Methods

2.4.1 Atomistic simulations

We performed all atomistic simulations with the Gromacs 4.5.3 simulation package,[103] while employing double precision floating-point arithmetic. We modeled all atomic interactions with the OPLS-AA force field,[104] while treating electrostatic interactions with the particle mesh Ewald method[105] and a Fourier grid spacing of 0.08 nm^{-1}. We truncated the short-ranged van der Waals (vdW) interactions and the real-space contribution to electrostatic interactions at 1.2 nm. We employed volume-dependent, long-ranged corrections to the energy and pressure to account for the use of truncated vdW interactions in the AA simulations. We did not employ any rigid constraints and integrated the equations of motion with the leapfrog integrator, while employing a 1 fs time step. We performed atomistic simulations for 2 systems: one containing 1200 heptane molecules and one containing 1600 toluene molecules. We simulated such large systems in order to avoid finite-size effects in the CG models because the calculated potentials for the 1-site heptane and toluene models do not vanish until nearly 3.0 nm.

We adopted the following procedure to equilibrate each system. We heated the system to 1000 K over 10 ns, simulated at 1000 K for 2 ns, and then cooled the system to 303 K over 10 ns. We next allowed each system to relax to its equilibrium density during a 2 ns simulation in the constant NPT ensemble (303 K, 1.0 bar), while employing a Berendsen thermostat and barostat with coupling constants of 0.1 and 1.0 ps, respectively.[106] The resulting equilibrium densities were 672.7 kg/m^3 for heptane and 860.7 kg/m^3 for toluene, which compare favorably with the experimentally measured densities of 679.5 kg/m^3 and 862.3 kg/m^3, respectively.[107]

After equilibration, we performed a production simulation for each system in the constant NPT ensemble. We employed the Nosé-Hoover thermostat to maintain T=303 K with a coupling constant of 0.5 ps[108,109] and employed the Parrinello-Rahman barostat to maintain P_0=1.0 bar with a coupling constant of 5 ps.[110] The

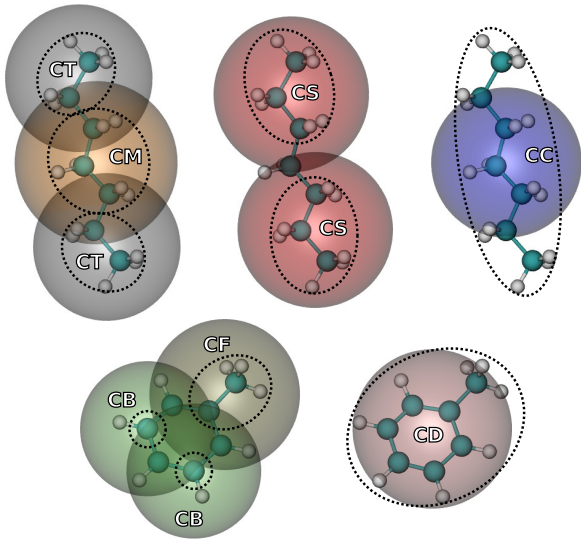

Figure 2.1. CG representations superimposed upon the corresponding atomistic representations. The CG sites, which are indicated by transparent spheres, are associated with the mass center for their constituent atoms, which are enclosed by the dashed circles. The size of each sphere indicates the excluded volume diameter of the site, which is estimated by the distance at which the corresponding site-site distribution vanishes in the AA model.

heptane and toluene systems were simulated for 100 ns and 80 ns, respectively. We discarded the first 10 ns of each simulation and employed the remainder of each simulation in order to characterize the AA models and also to parameterize the CG models.

2.4.2 CG models

We considered 1, 2, and 3-site CG models for heptane, as well as 1 and 3-site CG models for toluene. Figure 2.1 illustrates these CG representations. The dotted circles indicate the atoms that are associated with each site. The CG mapping defines each site position by the mass center for the atoms within the corresponding dotted circle. The colored spheres indicate the excluded volume of each site, which is defined by the distance at which the corresponding site-site radial distribution functions vanish.

For each CG representation, we parameterized a MS-CG interaction potential according to the theory described above. We employed bond potentials between each pair of consecutive sites within the same molecule. We employed angle potentials in

heptane to govern the intramolecular (CT-CM-CT) angles, but not in toluene. We represented these intramolecular forces with linear spline functions, while employing a grid spacing of 0.001 nm for bond potentials and 0.5 degrees for angle potentials. Additionally, we employed short-ranged pair potentials to model the nonbonded interactions between each pair of sites in distinct molecules. We represented these pair forces with cubic splines, while employing a grid spacing of 0.02 nm. We employed a cutoff of 1.4 nm for the pair potentials in the 3-site models and a cutoff of 3.0 nm for the pair potentials in the 1 and 2-site models. We determined the parameters for these potentials by minimizing the modified MS-CG functional, χ_1^2, in Eq. (2.28). In particular, we introduced the additional factor of $V^{2/3}$ into our in-house force-matching software.[31,111] We approximated the relevant ensemble averages with the configurations sampled from the AA simulations.

Despite the large systems and long simulations, we observed traces of statistical noise near the cutoff of the calculated pair forces. Since the pressure is very sensitive to this region of the pair force, we attempted to mitigate this noise by smoothing this region of the pair forces. First, we smoothed the entire calculated force function with a centered running average over 0.06 nm windows. We then determined the final potential by splicing together the original and smoothed potential at their intersection, as indicated by supplemental figure S1.[112] Figures S2-S6 present the resulting potentials.[112]

In addition to these bottom-up models, we also considered the top-down 3-site heptane model that was developed by Shinoda, DeVane, and Klein (SDK).[113] The SDK model treats the intramolecular bonds and angles with harmonic potentials that were parameterized to reproduce the average intramolecular structure of heptane. The SDK model treats non-bonded interactions with analytic 9-6 Lennard-Jones-type pair potentials of the form

$$U_{nb}(r) \;=\; \frac{27}{4}\epsilon\left[\left(\frac{\sigma}{r}\right)^9 \;-\; \left(\frac{\sigma}{r}\right)^6\right], \tag{2.34}$$

which were parameterized to reproduce the experimental bulk density and liquid-vapor interfacial tension of heptane.

2.4.3 Pressure-matching

For each bottom-up CG model, we also determined a volume-dependent contribution, U_V, to the CG potential. In each case, we represented U_V with the analytic basis functions defined in Eqs. (2.31) and (2.32). We determined the coefficients, ψ_d, by directly solving the normal system of linear equations that result from minimizing the DA pressure-matching functional, χ_2^2, in Eq. (2.29). We approximated the necessary averages with the ensembles sampled by the AA simulations. In particular, we determined P_{CG}^0 by evaluating the calculated MS-CG force field for the mapped ensemble. As previously observed by DA, our numerical calculations required only two basis functions for U_V.[44] We observed that more detailed spline representations for U_V did not significantly improve the accuracy of the CG model and appeared more sensitive to statistical noise in the simulated ensemble.

We developed an iterative approach that further optimizes U_V in order to quantitatively reproduce the AA volume distribution. In this approach, we first minimized χ_2^2 to determine an initial estimate $U_V^{(0)}$ for this volume-dependent potential. We then simulated the resulting CG model in the NPT ensemble for 20 ns in order to estimate the resulting pressure equation of state. As described in Eq. (2.33), we determined a correction, δU_V, to $U_V^{(0)}$ by comparing the AA and CG equations of state, while leaving the molecular mechanics contributions, U_2 and U_θ, to the CG potential unchanged. We repeated this procedure until the CG model satisfactorily reproduced the AA equation of state. We found that U_V converged within fewer than six iterations for each system. Figures S7 and S8 describe the calculated potentials and their convergence.[112]

2.4.4 CG simulations

All CG systems contained either 1200 heptane molecules or 1600 toluene molecules. We simulated each CG model for 20 ns. Since our focus was on investigating the thermodynamic properties of CG models, we employed a 1.0 fs time step in all of the CG simulations reported below. With this time step and without any attempt to optimize their efficiency, cursory investigations indicate that the bottom-up CG models are approximately 30 times more efficient than the OPLS-AA model. Additionally, we observe that the 1-site CG models appear stable with a 20 fs time step. However, due to the fairly stiff CG bond potentials, the 2 and 3-site CG

models become unstable when the simulation timestep is significantly larger than 1 fs.

We determined the initial configurations for the CG simulations by mapping a simulated AA configuration to the relevant CG representation. This AA configuration was selected from the last 10 ns of the NPT production simulation and sampled a volume within one standard deviation of the mean AA volume. We discarded the first 1.0 ns of each CG simulation as an equilibration period and analyzed the remainder of the trajectory.

We simulated all bottom-up models with a version of the LAMMPS software package[114] (17Jun13) that we modified to incorporate the pressure correction, F_V, in calculating the internal pressure. We employed the velocity-Verlet algorithm with a 1.0 fs timestep to integrate the Martyna-Tuckerman-Tobias-Klein equations of motion[96,115] at T=303 K and P_0 = 1.0 bar with coupling constants of 0.1 ps and 1.0 ps for the thermostat and barostat, respectively, while employing a Nosé-Hoover chain[116] with the default length of three. The non-bonded pair potentials were truncated at the non-bonded cutoffs adopted in the force-matching procedure. The CG models did not include explicit electrostatic interactions. Aside from the calculated volume-dependent potential, U_V, we did not employ long-ranged corrections for the energy or pressure.

We simulated the top-down 3-site SDK heptane model with Gromacs 4.5.3, while adopting many of the same settings that were employed in the AA simulations. The only differences between the simulation settings used for the AA and SDK models were that 1) the SDK model does not include electrostatic interactions; 2) the SDK model adopts a cutoff of 1.5 nm for the non-bonded pair potentials; and 3) the SDK model does not include long-ranged corrections for the energy or pressure.

2.5 Results

In the following, we investigate the capability of bottom-up CG methods to accurately model both the structure and thermodynamic properties of molecular liquids. We first focus on bottom-up 3-site CG heptane models, which can be directly compared with the top-down SDK model. We next analyze a series of bottom-up models for both heptane and toluene at several different resolutions. Finally, having developed an array of models that accurately model both the structure and density

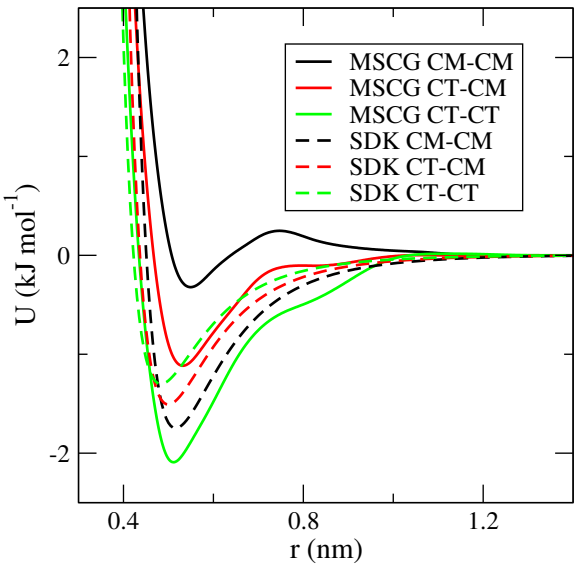

Figure 2.2. Comparison of the calculated MS-CG pair potentials (solid lines) and the published SDK pair potentials (dashed lines) for 3-site CG heptane models. The black, red, and green curves indicate the CM-CM, CT-CM, and CT-CT pair potentials, respectively.

fluctuations of these liquids at different resolutions, we investigate the impact of bottom-up coarse-graining upon thermodynamic properties.

Figure 2.2 compares the nonbonded pair potentials for two different 3-site CG models for heptane. The solid curves indicate the calculated MS-CG pair potentials, while the dashed curves indicate the published SDK pair potentials. Figure S2 demonstrates that the additional $V^{2/3}$ factor in χ_1^2 has minimal impact upon the calculated MS-CG potentials.[112] While the SDK potentials adopt a simple Lennard-Jones-type form and are quite similar, the MS-CG potentials demonstrate more complex features and greater differences. In the SDK model, the CM-CM pair potential is most attractive and the CT-CT potential is least attractive, but all three pair potentials have very similar minima. This trend is reversed and, moreover, the differences in the minima are much more pronounced in the MS-CG model. While the CT-CT MS-CG potential is more attractive than any of the SDK potentials, the CM-CM MS-CG potential is almost purely repulsive. Moreover, the site diameters demonstrate qualitatively different trends in the MS-CG and SDK models. In particular, the CT site is smaller than the CM site in the SDK model, while this trend is reversed in the MS-CG model.

Table 2.1 presents the calculated coefficients ψ_d for the pressure corrections,

Table 2.1. Pressure correction coefficients for the DA basis representation. All coefficients are given in units of J/mol.

Model	ψ_1 DA	ψ_1 DN	ψ_2 DA	ψ_2 DN
3-site hep	3.63	3.20	0.678	-4.54
2-site hep	12.8	12.7	-1.61	-6.33
1-site hep	40.0	39.9	-1.77	-1.52
3-site tol	4.94	5.30	1.14	-1.63
1-site tol	33.7	33.6	0.757	-1.47

$F_V = -\mathrm{d}U_V/\mathrm{d}V$, that are obtained from the Das-Andersen (DA) pressure matching variational principle and from the iterative procedure that we develop in this work (DN). According to Eq. (2.32), the resulting contribution to the CG pressure may be expressed:

$$F_V(V) = -(N/\bar{v})\left[\psi_1 + 2\psi_2(V-\bar{v})/\bar{v}\right] \tag{2.35}$$

where N is the number of CG sites and $\bar{v}$ is the average volume of the AA model. Consequently, the coefficients ψ_1 and ψ_2 correspond to average corrections for the pressure, $\Delta\bar{P}$, and compressibility, $\Delta\kappa_T$, respectively:

$$\Delta\bar{P} = -N\psi_1/\bar{v} \tag{2.36}$$

$$\Delta\kappa_T^{-1} = 2N\psi_2/\bar{v}. \tag{2.37}$$

Table I indicates that both the DA pressure-matching approach and the present iterative approach (DN) determine very similar reductions in the average pressure. However, for the 3-site model, the DA and DN approaches modify the compressibility in qualitatively different manners.

Figure 2.3 presents the simulated volume distributions and compressibilities for 3-site heptane models at an external pressure P_0=1 bar. In the absence of a virial constraint or pressure correction, the MS-CG model (FM) overestimates the system volume by more than 10%. (Note the discontinuity in the x-axis of Fig. 2.3.) After including the DA pressure correction, the bottom-up model (DA) reproduces the average AA volume within 1.2%. By iteratively refining this pressure correction to self-consistency, the bottom-up model (DN) quantitatively reproduces the AA volume distribution.

Figure 2.3 also presents an "experimental volume distribution" that is determined

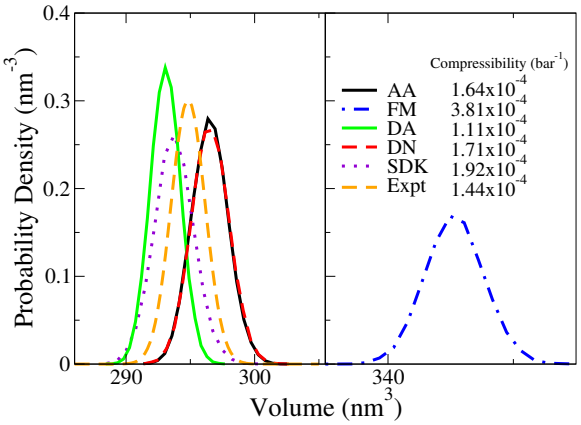

Figure 2.3. Simulated volume distributions for various heptane models. The solid black curve presents the simulated distribution for the OPLS-AA model. The dashed-dotted blue, solid green, dashed red, and dotted purple curves indicate simulated distributions for the MS-CG, DA, DN, and SDK 3-site models, respectively. The dashed orange curve indicates the normal distribution that is constructed from the experimentally known density and compressibility of heptane.

by the Gaussian distribution with the correct mean and variance,

$$p_e(V|P_0) = \frac{1}{\sqrt{2\pi\sigma_e^2}} \exp\left[-(V-\bar{V}_e)^2/2\sigma_e^2\right] \tag{2.38}$$

where $\bar{V}_e = N_{mol}/\rho_e$ and $\sigma_e^2 = k_B T\, \bar{V}_e\, \kappa_T$ are determined from the number of simulated molecules, N_{mol}, along with the experimentally known density, ρ_e, and compressibility, κ_T, for heptane at T=303 K and $P_0 = 1$ bar.[107] The AA, DA, DN, and SDK models all reproduce the experimental density and volume distribution with similar accuracy.

Figure 2.4a presents local estimates of the pressure equations of state that are obtained from simulating these heptane models. Figure 2.4a also presents a local estimate of the experimental equation of state:

$$\bar{P}_e(V) = P_0 - \kappa_T^{-1}\left(V/\bar{V}_e - 1\right) \tag{2.39}$$

The AA model slightly underestimates the density and overestimates the compressibility of heptane. As expected from Fig. 2.3, the DN model reproduces the AA equation of state with nearly quantitative accuracy. The DA model underestimates the compressibility of the AA model, while the SDK model is the most compress-

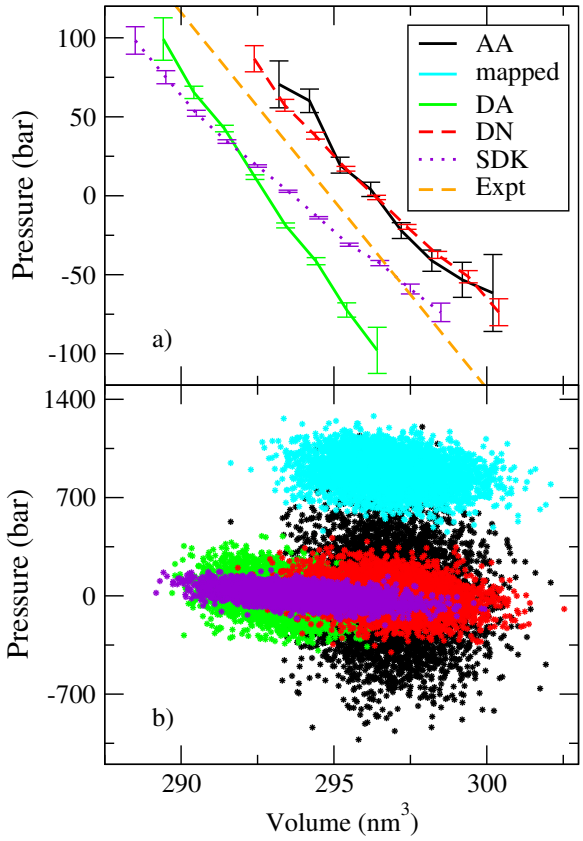

Figure 2.4. Comparison of the pressure-volume behavior for the AA heptane model and for different 3-site CG heptane models. The black, green, red, and purple curves correspond to the models of Fig. 3. Panel a) presents the equation of state for each model, which is estimated from the mean pressure at each volume in the simulated constant NPT ensemble. The error bars indicate the standard error in the simulated means. The orange curve indicates the equation of state that is determined from the experimentally known density and compressibility of heptane. Panel b) presents a scatter plot of the simulated pressure and volume. The cyan points correspond to the pressure, P^0_{CG}, that is determined by applying the MS-CG potential to the mapped ensemble.

ible of these models. Interestingly, over the simulated range of thermodynamic conditions, the bottom-up DN model matches the estimated experimental equation of state with accuracy that is comparable to (or possibly better than) the top-down SDK model, which was explicitly parameterized to reproduce the experimental density and surface tension, but not the compressibility.

While the AA, DN, and SDK models demonstrate very similar volume fluctuations, they sample significantly different pressure fluctuations. Figure 2.4b presents simulated scatter plots of the instantaneous pressure and volume for the different

Table 2.2. Standard deviations describing fluctuations in the pressure, σ, and non-ideal contribution to the pressure, σ_W, for different heptane models. All standard deviations are given in units of bar.

Model	σ	σ_W
AA	312	310
3-site mapped	118	117
3-site DA	116	115
3-site DN	113	113
3-site SDK	52	51
2-site mapped	72	72
2-site DN	69	68
1-site mapped	33	32
1-site DN	31	30

models. In Fig. 2.4b, the cyan points correspond to the pressure, P^0_{CG}, that is determined by evaluating the MS-CG force field (without the pressure correction) for the mapped ensemble. As expected from Fig. 2.3, P^0_{CG} dramatically overestimates the internal pressure of the mapped AA configurations. More interestingly, Fig. 2.4b and Table 2.2 demonstrate that all of the CG models dramatically underestimate the magnitude of the pressure fluctuations in the AA model. Because they employ the MS-CG potential to model interactions between sites, the MS-CG, DA, and DN models all sample similar pressure fluctuations. In contrast, the SDK model, which employs Lennard-Jones-type pair potentials, samples even smaller pressure fluctuations. Consequently, the reduction in pressure fluctuations depends significantly on both the model representation and the potential. We expect that the differences in pressure fluctuations primarily reflect the relative smoothness of the CG potential surfaces, since the majority of these fluctuations arise from the virial contribution to the pressure.

We next assess the equilibrium structure that is generated by the DA, DN, and SDK 3-site heptane models. Figure 2.5 compares the site-site radial distribution functions (rdfs) for the mapped AA ensemble and for the CG models. Although this is not explicitly incorporated in their parameterization, all three models reasonably reproduce the AA structure. The MS-CG-based models reproduce the AA site-site rdfs quite accurately.[117] As discussed in our prior studies,[117–119] the slight discrepancies in the MS-CG site-site rdfs result from the inability of the

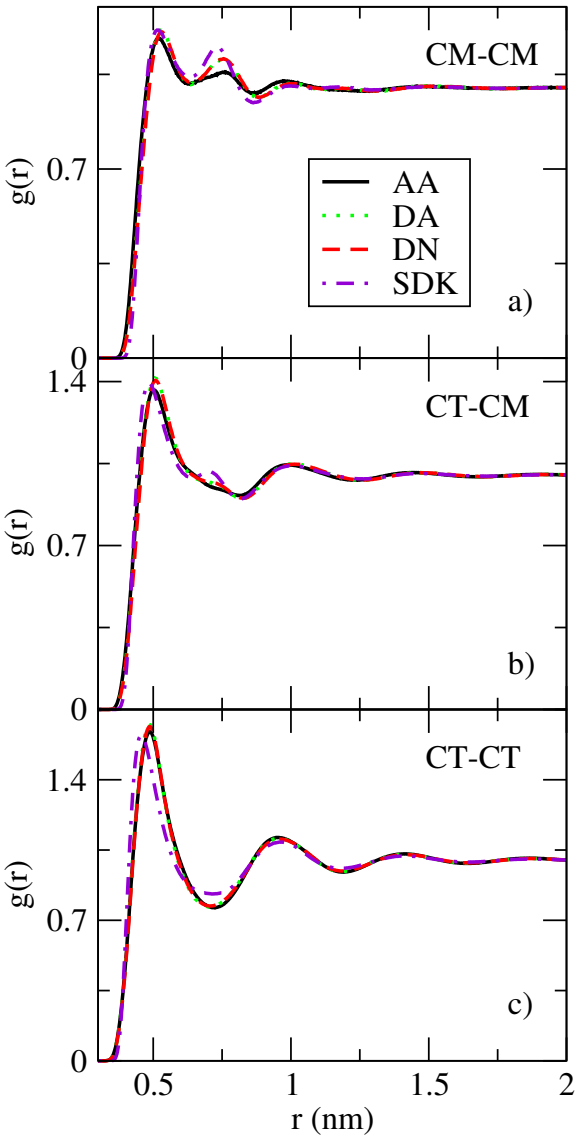

Figure 2.5. Comparison of site-site rdfs for 3-site heptane representations. The solid curves correspond to the mapped AA ensemble, while the green dotted, red dashed, and purple dashed-dotted curves correspond to the DA, DN and SDK models, respectively.

approximate CG potential to reproduce higher order correlations in the mapped ensemble. Despite slight differences in their density, the DA and DN models sample nearly indistinguishable structural distributions. In contrast, the SDK model appears to slightly underestimate the size of the CT site, which may be traced to the shift in the potential minima of Fig. 2.1. Otherwise, the SDK model reproduces the AA site-site rdfs with reasonable accuracy.

Figure S10 compares the intramolecular structure of the heptane models.[112] The AA model samples a bimodal bond distribution and a trimodal angle distribution with a maximum peak that corresponds to nearly linear conformations. The MS-CG-based models quantitatively reproduce the AA bond distribution and also reproduce the peak positions, but not the peak heights, of the complex AA angle distribution. These discrepancies in the angle distributions of the MS-CG-based models are due to the inability of the approximate potential to reproduce the bond-angle correlations in the mapped AA ensemble.[117,118,120–122] In contrast, the harmonic bond and angle potentials of the SDK model generate simple Gaussian distributions and cannot even qualitatively match the multimodal features of the AA distributions.

Figure 2.6a demonstrates that both the bottom-up DN and top-down SDK 3-site heptane models qualitatively reproduce the center of mass (com) rdf for the mapped AA ensemble. Because site-site and com rdfs are not directly related, the accurate reproduction of site-site rdfs does not ensure accurate reproduction of com rdfs.[122] Indeed, neither CG model quantitatively reproduces the peak positions or peak heights of the AA com rdf, although the DN model appears slightly more accurate than the SDK model. We expect that these errors in the com rdfs may stem from errors in modeling the intramolecular conformations. In particular, both 3-site CG models significantly underestimate the tendency of heptane to sample nearly linear conformations. Since these linear conformations favor close contact between molecules, this error in the angle distribution may also account for the slight shift in the first peak of the com rdf.

We also employed the same bottom-up strategy to develop 1 and 2-site heptane models. For each model, we minimized the modified MS-CG force-matching functional to determine the interaction potentials, $U_2 + U_\theta$, and employed the iterative pressure-matching method (DN) to optimize U_V. Figure 2.6b presents the com rdfs for the 1 and 2-site models.[117] The 1-site model reproduces the mapped AA com rdf with similar accuracy to the 3-site model. However, although Fig. S9 demonstrates that the 2-site model accurately reproduces both the AA site-site rdf and bond distribution,[112] Fig. 2.6b indicates that this model provides a significantly less accurate description of the AA com rdf. We expect that this relatively large discrepancy occurs because the 2-site model cannot describe the shape and, thus, the packing of heptane molecules.

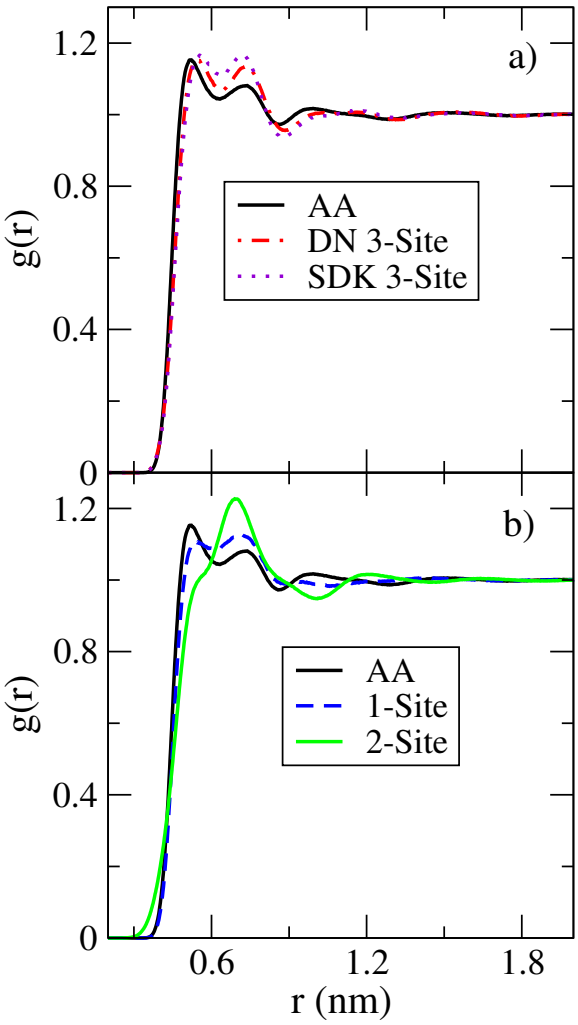

Figure 2.6. Comparison of center-of-mass rdfs for a) 3-site heptane representations, and b) 1- and 2-site heptane representations. In both panels, the solid black curves correspond to the mapped AA ensemble. In panel a), the dashed-dotted red and dotted purple curves correspond to the 3-site DN and SDK models, respectively. In panel b), the dashed blue and solid green curves correspond to the 1 and 2-site DN models, respectively.

Figure 2.7 compares the volume distributions and pressure equations of state that are estimated by constant NPT simulations of the AA model and the 1, 2, and 3-site DN heptane models. Fig. 2.7a demonstrates that all three bottom-up models quantitatively reproduce the AA volume fluctuations in the constant NPT ensemble. Additionally, Fig. 2.7b demonstrates that the three DN models also reasonably reproduce the AA pressure equation of state near the simulated pressure of $P_0 = 1$ bar.

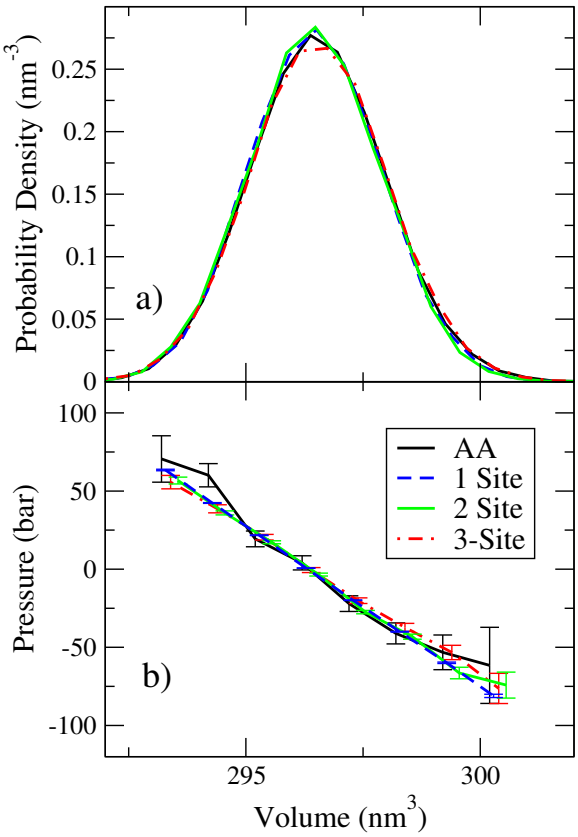

Figure 2.7. Comparison of the pressure-volume behavior for the AA heptane model and for 1, 2, and 3-site bottom-up DN heptane models. Panel a) presents the volume distribution sampled by each model in simulations at constant pressure. Panel b) presents the equation of state for each model, which is estimated from the mean pressure at each volume in the simulated constant NPT ensemble. The error bars indicate the standard error in the simulated means.

We also applied the same bottom-up procedure to parameterize 1 and 3-site CG models for toluene. Figure 2.8 demonstrates that both CG models also reasonably reproduce the com rdf for the AA model. In particular, the 1-site model reproduces the AA com rdf with nearly quantitative accuracy. Figures S11 and S12 demonstrate that the 3-site model quite accurately reproduces the intramolecular structure and site-site rdfs of the AA model.[112] However, Fig. 2.8 demonstrates that this model provides a somewhat less accurate description of the AA com rdf.[123]

Figure 2.9 compares the volume distributions and pressure equations of state that are estimated by NPT simulations of these toluene models. Figure 2.9a demonstrates that both CG toluene models reproduce the AA volume distribution with nearly quantitative accuracy. Figure 2.9b demonstrates that the CG models

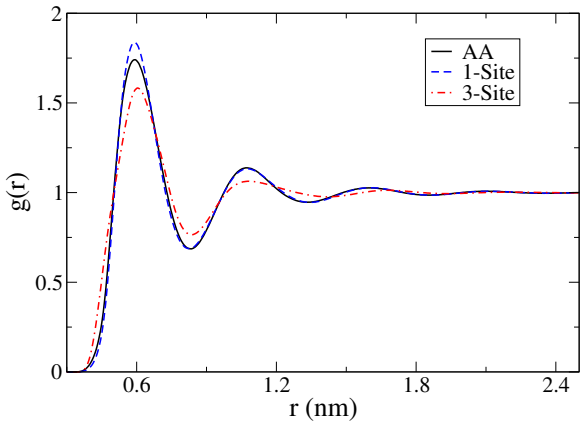

Figure 2.8. Comparison of com rdfs for AA and CG toluene models. The solid curve corresponds to the mapped AA ensemble, while the dashed and dashed-dotted curves correspond to the 1 and 3-site DN models, respectively.

also reasonably reproduce the simulated AA pressure equation of state for toluene, although the 3-site model appears slightly more compressible than the AA model.

The comparison of the different bottom-up models leads to several interesting observations: 1) Figures 2.2, S3, and S5 demonstrate that the calculated MS-CG nonbonded pair potentials become increasingly long ranged and repulsive with coarsening.[112] 2) Figure S7 suggests that the DN pressure corrections become larger and often less volume-dependent with coarsening.[112] 3) Table I and Fig. S8 demonstrate that the variational DA approach often becomes increasingly accurate with coarsening.[112]

Figure 2.10 investigates the correlation between the necessary DN pressure correction per molecule, $\left\langle \bar{F}_V \right\rangle$, and the difference in the intermolecular potential between the AA and CG models, $\langle \Delta U_{inter} \rangle$. In each CG model, $\langle \Delta U_{inter} \rangle < 0$ and $\left\langle \bar{F}_V \right\rangle < 0$. Furthermore, the required pressure correction and reduced intermolecular potential are highly correlated and systematically increase in magnitude with coarsening. Consequently, Fig. 2.10 quantifies an intuitive van der Waals perspective upon bottom-up coarse-graining, i.e., the pressure correction provides the necessary attractive force on the simulated volume in order to account for the cohesive energy density that is lost in bottom-up structure-based coarse-graining.

Finally, Figure 2.11 assesses the impact of coarsening upon the "simplicity" of bottom-up models. One mechanism for quantifying simplicity is the Pearson correlation, R, between fluctuations in the virial and the total potential. Models

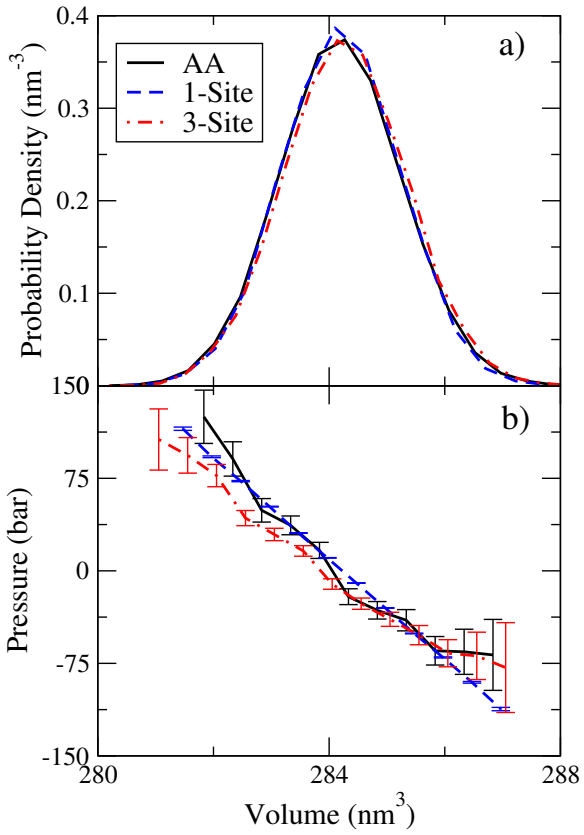

Figure 2.9. Comparison of the pressure-volume behavior for the AA toluene model and for different bottom-up DN toluene models. Panel a) presents the volume distribution sampled by each model in simulations at constant pressure. Panel b) presents the equation of state for each model, which is estimated from the mean pressure at each volume in the simulated constant NPT ensemble. The error bars indicate the standard error in the simulated means.

with a high correlation ($R \geq 0.9$) are referred to as Roskilde-simple (R-simple) and demonstrate intriguing relations between dynamic, structural, and thermodynamic properties.[124] While the origin of these correlations remains somewhat unclear, simple liquids governed by a Lennard-Jones potential are known to be R-simple at positive internal pressures due to the dominance of the repulsive term in the potential.[124,125] Indeed, the cyan and red crosses in Fig. 2.11 demonstrate the impact of pressure upon R for a system of Lennard-Jones spheres that was previously parameterized to reproduce the experimental density of heptane at 1 bar pressure.[126]

Figure 2.11 presents a scatter plot of the average intermolecular energy (per molecule) and the R-simplicity for the two AA models and six CG models that have been considered above. Figure 2.11 suggests a modest and nonlinear correlation

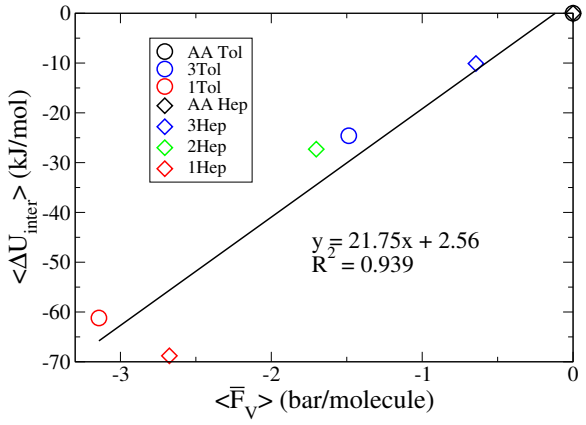

Figure 2.10. Scatter plot correlating the missing cohesive energy with the pressure correction required for each model. The x-coordinate indicates the average of $\bar{F}_V$, which is the calculated DN pressure correction per molecule. The y-coordinate indicates the average difference between the AA and CG intermolecular potentials and has been normalized per molecule. Circles and diamonds indicate toluene and heptane models, respectively. Black symbols indicate AA models, while the blue, green, and red symbols indicate CG models with 3, 2, and 1-sites per molecule.

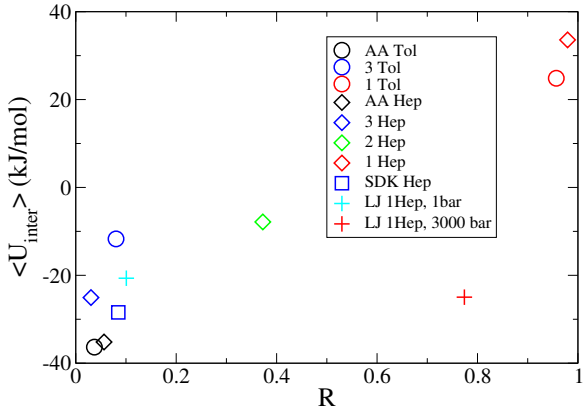

Figure 2.11. Scatter plot correlating the cohesive energy density (per molecule) with the Pearson correlation between potential and virial fluctuations in the models investigated. The x-coordinate indicates the Pearson product-moment coefficient (R) relating fluctuations in the potential energy and virial of each model. The y-coordinate indicates the average intermolecular potential (per molecule) of each model. The crosses indicate results for a 1-site heptane model with a Lennard-Jones potential that has been parameterized to reproduce the experimental density of heptane at 1 bar external pressure (cyan). The red cross indicates results for this model when simulated at 3000 bar external pressure.

between the intermolecular potential, model resolution, and R-simplicity. Among these models, the AA models demonstrate the greatest cohesive attraction and also the greatest complexity (i.e., the least simplicity). The 3-site CG models demonstrate considerably reduced cohesion, but only slightly greater simplicity. In these cases, the asymmetry and internal flexibility of the models preclude a strong correlation between the energy and virial. Interestingly, the top-down SDK 3-site heptane model, which employs Lennard-Jones-type 9-6 potentials that have been parameterized to reproduce the density and interfacial tension, demonstrates only slightly greater R-simplicity and cohesion than the corresponding bottom-up 3-site model. With further coarsening, though, the bottom-up potentials become increasingly repulsive, which leads to diminishing intermolecular cohesion, increasingly large pressure corrections, and greater R-simplicity. In the limit of the 1-site models, the CG models become simple liquids experiencing high internal pressures, P^0_{CG}, which are corrected by F_V in order to preserve the liquid density.

2.6 Discussion

The present work investigates a recent proposal by Das and Andersen (DA) for modeling the pressure with bottom-up models.[44] This approach combines two key ideas. Firstly, DA introduced an additional volume-dependent, but configuration-independent, potential, U_V, in order to approximate the volume-dependence of the PMF and to account for its contribution to the thermodynamic pressure. Indeed, as indicated in Eq. (2.3), such terms are frequently employed in AA models in order to compensate for truncating Lennard-Jones pair potentials.[1,94] However, they have been generally neglected in CG potentials. Secondly, DA extended the multi-scale coarse-graining (MS-CG) variational principle[25] to the constant NPT ensemble in order to parameterize U_V. This MS-CG/DA approach appears particularly appealing because it provides a variational framework for approximating both the configuration-dependence and also the volume-dependence of the PMF.

Our calculations demonstrate that the DA approach significantly improves the simulated pressure of (unmodified) MS-CG models. In the absence of a virial constraint,[27,88] the calculated MS-CG potentials dramatically over-estimate the thermodynamic pressure and, consequently, the resulting models significantly underestimate the equilibrium density of molecular liquids. For instance, the 3-site

MS-CG heptane model underestimates the liquid density by more than 10%, while the 1 and 2-site MS-CG heptane models quickly vaporize at 1 bar external pressure. In each case that we considered, by incorporating volume-dependent potentials, U_V, the MS-CG/DA approach reproduced the equilibrium density fluctuations of the underlying AA model with qualitative, but not quantitative, accuracy. In particular, the MS-CG/DA approach generally reproduces the average AA density within approximately 1%. The discrepancies in the simulated MS-CG/DA densities appear to result from subtle differences in the structural correlations that are present in the mapped ensemble and in the ensemble sampled by the CG model, since U_V is parameterized to reproduce the atomic pressure when evaluated for the mapped ensemble.

Accordingly, we developed an iterative, self-consistent approach that further optimizes U_V in order to reproduce the AA pressure equation of state for the configurations that are sampled in CG simulations. In combination with the calculated MS-CG interaction potentials, the iterative pressure matching approach determined bottom-up CG (DN) models that accurately reproduced the corresponding AA site-site radial distribution functions (rdfs) and bond distributions, although they provided slightly less accurate descriptions of the AA center-of-mass rdfs and angle distributions. More importantly, each DN model reproduced the AA distribution of density fluctuations with quantitative accuracy and reproduced with semi-quantitative accuracy the simulated AA pressure equation of state. In particular, the bottom-up 3-site DN heptane model described the experimental density and compressibility with comparable accuracy to a top-down 3-site model, which SDK previously parameterized to reproduce the experimental density and surface tension.[113]

The iterative pressure-matching approach can be seen as a simple adaptation of iterative Boltzmann inversion (IBI)[28,127] for reproducing the potential of mean force (or, equivalently, the mean force[117,119]) that governs volume fluctuations. Although one can envision more sophisticated algorithms that treat the correlations between the CG configuration and volume,[26,122,128,129] the present simple algorithm appears satisfactory for the present work. As demonstrated in Fig. S8, this approach rapidly converges and provides a semi-quantitative description of the AA density fluctuations within two iterations.[112] Importantly, each iteration requires a single short and efficient CG simulation in order to estimate the resulting equation of

state, but does not alter the intermolecular potentials and, consequently, minimally impacts the equilibrium structure of the model.

This approach can be readily generalized in two important ways. While we employed the MS-CG approach to optimize the intermolecular potentials directly (i.e., noniteratively) from the original AA ensemble, the extended DA approach can also be adapted for iterative structure-based CG methods.[26,28,30,119,122,127–130] For instance, one could first iteratively optimize the interaction potential, $U_2 + U_\theta$, in order to reproduce AA structural correlations in a constant volume ensemble that corresponds to the average AA density. Given this structure-based interaction potential, one could then estimate U_V by minimizing the DA pressure-matching functional over the mapped AA ensemble. If necessary, one could iteratively optimize U_V to reproduce the AA pressure equation of state. Furthermore, the present approach could also be applied to reproduce a pressure equation of state that has been determined experimentally, rather than from AA simulation. This appears a promising hybrid bottom-up/top-down approach for accurately describing both structure and thermodynamic properties.

In the course of this work, we developed a series of CG models that accurately approximate both the configuration-dependence and volume-dependence of the PMF for heptane and toluene at four different resolution. By comparing the different models, we investigated the fundamental impact of bottom-up coarse-graining upon thermodynamic properties. Table I demonstrates that, for each CG model, the iteratively derived pressure correction decreases the pressure and compressibility of the CG model relative to the (uncorrected) MS-CG model. This is consistent with previous observations that bottom-up structure-based CG potentials tend to generate higher pressures and also demonstrate greater compressibility than the underlying AA model.[33,34] Table I and Fig. S7 suggest that the iterative DN pressure corrections become larger and, usually, less volume-dependent with decreasing resolution.[112] Interestingly, Fig. S8 demonstrates that the variational DA approach often determines an increasingly accurate estimate of the optimized DN pressure correction with coarsening.[112] This presumably reflects the decreasing significance of many-body correlations in the CG models with decreasing resolution due to the corresponding decrease in site density. Furthermore, Figures 2.2, S13, and S15 demonstrate that the calculated MS-CG interaction potentials become, not only more repulsive, but also longer-ranged with coarsening.[112]

The DN pressure correction that is necessary for reproducing AA density fluctuations correlates quite strongly with the reduced cohesive energy of the CG models, which is entirely consistent with the van der Waals picture of liquids.[78] As the cohesive energy density of the model decreases, increasingly large pressure corrections are necessary to prevent the material from expanding. This is also consistent with the intuitive notion discussed in the introduction, i.e., structure-based methods accurately determine repulsive short-ranged interactions that govern local structure at the expense of attractive longer-ranged interactions that provide this cohesion. However, it is worth noting that the PMF systematically increases with coarsening, as entropy is transferred from the configurational space into the effective interactions.[82,131] Since the MS-CG potential is parameterized to approximate the PMF, it is quite possible that this entropic effect also contributes to this reduced cohesion.[41,42]

These observations are quite consistent with recent studies of thermodynamic consistency by Guenza and coworkers.[79,83–86] These studies employed integral equation theories to determine effective pair potentials for low resolution CG polymer models in which each site is "larger" than the polymer persistence length. The resulting pair potentials are characterized by a soft Gaussian core with a long-ranged repulsive tail and, at even greater distances, a very shallow attractive well.[84,85] At such low resolutions, the polymer potential determined from these site-site pair potentials appears to provide a very accurate approximation to the many-body PMF, such that the models accurately describe both structure and thermodynamic properties.[84,86] As discussed above, Guenza and coworkers have also demonstrated the transfer of entropy from configurational space into the effective interactions between CG sites.[83,86] Moreover, the effective site-site potentials become increasingly repulsive and long-ranged with increased coarsening,[84,85] which is very consistent with the MS-CG potentials that are calculated for heptane and toluene in the present work. Guenza and coworkers have also emphasized the distinct roles of the repulsive and attractive regimes of the CG potential for governing the model structure and thermodynamic properties.[85] Interestingly, the pressure correction, U_V, appears to perform a similar role for the present bottom-up models as the very weak long-ranged attractions in the CG polymer models. In both cases, these contributions provide the necessary cohesion for stabilizing condensed phases. It is quite striking to observe such remarkable similarities in

distinct coarse-graining approaches that address very disparate systems at very different resolutions.

Conversely, it is also interesting that the thermodynamic impact of structure-based coarse-graining depends upon the particular molecule considered. In particular, coarsening to a 3-site representation appears to have much greater effect upon the thermodynamic properties of toluene than heptane. However, the 1-site heptane and toluene models demonstrate quite similar energetic losses and require quite similar pressure corrections to sample the correct density. These differences presumably reflect the stronger aromatic interactions of toluene and the greater conformational flexibility of heptane, although further analysis is necessary to decisively determine the microscopic origins of their differing response to coarsening.

Intuitively, one expects that, since coarsening leads to increasingly simple representations of molecular systems, the resulting models will demonstrate increasingly simple thermodynamic behavior. One measure of simplicity, Roskilde (R)-simplicity, considers the Pearson correlation (R) between fluctuations in the potential and virial.[124] R-simple systems, such as compressed Lennard-Jones fluids, are defined by potential-virial correlations of $R \geq 0.9$. The phase diagrams for R-simple systems demonstrate "isomorphic curves," along which the structure, dynamics, and thermodynamic properties of the system are approximately invariant when expressed in reduced units.[132] If a CG model demonstrates R-simple behavior, then a simulation at a single thermodynamic state point would be sufficient to predict model properties along the associated isomorph. Thus, R-simplicity could prove a practically useful property for predicting the transferability and thermodynamic behavior of CG models.

Our results suggest that there is a modest nonlinear correlation between model resolution, cohesive energy, and R for the present bottom-up models that reproduce both structural and thermodynamic properties. Interestingly, the 3-site models, including the SDK model with Lennard-Jones-type potentials, demonstrate only slightly greater R-simplicity than the original atomic models. In contrast, the bottom-up 1-site models demonstrate extremely high correlations between the potential and virial fluctuations.

Our calculations suggest that this reflects two effects that occur with increasing coarsening: 1) the MS-CG potentials become longer ranged and more repulsive; and 2) an increasingly large pressure correction is necessary to preserve the AA

density. These observations emphasize the near equivalence of the following two types of simulations: 1) simulations that employ U_V, which systematically reduces the internal pressure of the CG model, to sample the constant NPT ensemble at an external pressure $P_0 = 1$ bar; and 2) simulations of the original MS-CG potential (without the pressure correction) when performed at a correspondingly higher external pressure that preserves the density of the original AA model. (Their inequivalence results from the second term in U_V, which preserves the compressibility of the AA model.) In the 1-site models, this large pressure correction compresses the system to such an extent that repulsive interactions dominate both the potential and virial fluctuations, which causes $R \to 1$. Thus, it appears that structure-based bottom-up approaches can simplify CG models to the extent that they demonstrate R-simple behavior at sufficiently low resolution. This emphasizes that, without appropriate care, bottom-up coarse-graining can fundamentally alter the character of the model. This clearly motivates further work investigating how AA and CG models respond to changes in thermodynamic state.[133]

Finally, although these bottom-up DN models accurately reproduce atomic density fluctuations, they dramatically underestimate the atomic pressure fluctuations, which is primarily due to reduced fluctuations in the CG virial. This discrepancy depends quite sensitively upon both the resolution and also the approximate potential of the CG model. In particular, the top-down SDK model, which employs Lennard-Jones-type potentials, demonstrates significantly smaller pressure fluctuations than the bottom-up DN models, which employ MS-CG potential functions with relatively greater complexity.

At a given volume and temperature, the variance in the pressure fluctuations may be expressed[1]

$$\left\langle \delta P^2 \right\rangle_{NVT} = \frac{k_B T}{V} \left(\frac{2N k_B T}{3V} + \langle P \rangle_{NVT} - \kappa_T^{-1} + \frac{\langle \chi \rangle_{NVT}}{V} \right). \tag{2.40}$$

Since the bottom-up DN models reasonably reproduce both the pressure and the compressibility, κ_T, at each volume, the reduced pressure fluctuations can be directly traced to the reduced number of interacting particles, N, and to the 'hypervirial'

contribution to the compression modulus[1]

$$\chi = \frac{1}{9}\sum_{i<j}^{N}\left\langle r_{ij} f_{2;ij}(r_{ij}) + r_{ij}^2 f'_{2;ij}(r_{ij}),\right\rangle \tag{2.41}$$

where r_{ij} is the distance between particles i and j, $f_{2;ij}$ is the two-body force between the pair, and $f'_{2;ij}$ indicates its derivative. Clearly, the reduced pressure fluctuations will have important practical ramifications for modeling material properties, such as the shear modulus and viscosity.[134–137] (Of course, a realistic description of viscosity also requires a more sophisticated treatment of CG dynamics.[138–140]) In order to accurately describe these material properties with bottom-up CG models, it will be necessary to introduce additional fluctuating forces into the model, e.g., with thermostats or fictitious particles.[141–143]

2.7 Conclusions

In closing, we note that this work demonstrates the promise of bottom-up modeling approaches for accurately modeling both structural and thermodynamic properties of molecular liquids in practice. Moreover, this work emphasizes the importance of considering the state-point dependence and, in particular, the volume-dependence of the PMF when describing thermodynamic properties with bottom-up models. By developing a self-consistent extension of the DA variational procedure, we determined volume-dependent potentials, U_V, for bottom-up CG models that quantitatively reproduced the equilibrium density and compressibility of AA models. This volume-dependent potential accounts for the intermolecular cohesion that is lost in structure-based potentials. This work also demonstrates that bottom-up coarse-graining will tend to underestimate pressure fluctuations, which are important for many material properties, and, in the limit of very low resolution, tend to generate models that demonstrate R-simple behavior.

Finally, this work indicates many directions for future studies. For instance, while the present approach accurately models density fluctuations of homogeneous fluids, clearly further work is necessary to extend this approach for liquid-vapor interfaces and other inhomogeneous systems. Similarly, although U_V reasonably approximates the effective density-dependent many-body contributions to the pressure at a single pressure, further work is necessary to assess its transferability to

significantly different densities. Additionally, further studies are necessary to assess the potential transferability of this approach for modeling different temperatures and compositions. Indeed, preliminary results suggest that by applying the extended ensemble approach,[39] along with simple, accurate descriptions of U_V, it may be possible to develop a single transferable potential for modeling both the equilibrium structure and density fluctuations for a wide range of liquid mixtures. Finally, this work strongly motivates further efforts to develop a simple "van der Waals framework" that coherently treats both the structural and thermodynamic aspects of bottom-up coarse-graining.

Chapter 3
Bottom-up coarse-grained models with predictive accuracy and transferability for both structural and thermodynamic properties of heptane-toluene mixtures

N. J. H. Dunn, W. G. Noid, *J Chem Phys.* **2016**, 144, 204124

3.1 Abstract

This work investigates the promise of a "bottom-up" extended ensemble framework for developing coarse-grained (CG) models that provide predictive accuracy and transferability for describing both structural and thermodynamic properties. We employ a force-matching variational principle to determine system-independent, i.e., transferable, interaction potentials that optimally model the interactions in five distinct heptane-toluene mixtures. Similarly, we employ a self-consistent pressure-matching approach to determine a system-specific pressure correction for each mixture. The resulting CG potentials accurately reproduce the site-site rdfs, the volume fluctuations, and the pressure equations of state that are determined by all-atom (AA) models for the five mixtures. Furthermore, we demonstrate that these CG potentials provide similar accuracy for additional heptane-toluene mixtures that were not included their parameterization. Surprisingly, the extended ensemble approach not only improves the transferability, but also the accuracy of the calculated potentials. Additionally, we observe that the required pressure

corrections strongly correlate with the intermolecular cohesion of the system-specific CG potentials. Moreover, this cohesion correlates with the relative "structure" within the corresponding mapped AA ensemble. Finally, the appendix demonstrates that the self-consistent pressure-matching approach corresponds to minimizing an appropriate relative entropy.

3.2 Introduction

Molecular dynamics (MD) simulations provide a powerful tool for investigating and predicting the properties of soft materials.[144] In particular, simulations with all-atom (AA) models often describe structural and thermodynamic properties with nearly quantitative accuracy.[145,146] Nevertheless, the computational expense of AA models continues to motivate tremendous interest in low-resolution coarse-grained (CG) models that more efficiently address mesoscale phenomena.[7,8,72] Moreover, coarse-graining offers researchers the opportunity to precisely tailor models for specific phenomena. However, in order to realize these computational and conceptual advantages, CG models must not only be efficient, but also predictive. Thus, the challenge of systematic coarse-graining is to determine models that accurately describe structural and thermodynamic properties, while also demonstrating predictive transferability for modeling a wide range of systems across varying thermodynamic conditions.[73–75]

Statistical thermodynamics provides a straightforward "bottom-up" framework for addressing this challenge. The many-body potential of mean force (PMF) is the central quantity in this framework.[20–22,24,25,77] The PMF is a free energy function that corresponds to the net Boltzmann weight for all AA configurations that map to a particular CG configuration.[147–149] Consequently, the PMF depends upon both the CG configuration and the thermodynamic state point. This PMF is the appropriate effective potential for a CG model that quantitatively preserves both the structure and thermodynamic properties of an underlying AA model. Unfortunately, the PMF cannot be determined for most interesting systems. Instead, bottom-up methods often determine CG potentials as systematic approximations to the PMF.[32] In particular, bottom-up approaches often approximate the configuration-dependence of the PMF with sufficient accuracy to reasonably reproduce simple structural properties, e.g., radial distribution functions, of AA models.[26,28–30]

However, bottom-up approaches have enjoyed less success for modeling thermodynamic properties. The process of coarse-graining directly transfers thermodynamic information from the AA configuration space into thermodynamic contributions to the many-body PMF.[81,82] These contributions must be addressed in order to accurately describe thermodynamic properties with CG models.[83,86,91–93] While "top-down" approaches address these effects by explicitly parameterizing CG potentials to reproduce target thermodynamic properties,[9,113,150] structure-based bottom-up approaches tend to provide a poor description of thermodynamic properties, such as the pressure.[33,34] Consequently, bottom-up approaches often introduce ad hoc corrections to the interaction potentials in order to accurately model the pressure.[27,28,80,88,89] Unfortunately, this approach can yield a poor description of the compressibility.[34,87]

Quite recently, Das and Andersen pioneered a "pressure-matching" variational principle in order to model the volume-dependence of the PMF.[44] We demonstrated that this approach provides a qualitative, but not quantitative, description of density fluctuations for molecular liquids.[151] Moreover, we developed a self-consistent pressure-matching approach that quantitatively reproduces the equilibrium density, compressibility, and pressure of molecular liquids.[151] We demonstrate in the appendix that this self-consistent pressure-matching approach corresponds to minimizing a relative entropy with respect to a volume-dependent potential. More generally, we expect that this approach may prove useful for accurately describing many thermodynamic properties with CG models.

Similarly, it remains challenging to improve or predict the transferability of bottom-up CG models. Since the PMF explicitly varies with thermodynamic conditions, approximations that are optimized for one state-point will generally provide a less accurate approximation at other state points. Indeed, many studies have documented the sensitivity of optimized bottom-up potentials to variations in, e.g., temperature and density.[33,35–38,40–42,152,153] Accordingly, an increasing number of CG models have adopted potentials that explicitly vary with thermodynamic conditions.[37,43,84,86,154–161] However, comparatively little progress has been achieved via bottom-up approaches for parameterizing interaction potentials that are transferable to chemically distinct systems. Previous bottom-up approaches have attempted to develop transferable potentials by comparing or combining potentials that were optimized for various model systems.[162–167] Conversely, other studies have

attempted to achieve greater transferability by minimizing the context-dependence of the calculated interaction potentials.[168–170]

Alternatively, we have proposed a simple "extended ensemble" approach[39] that adopts a global optimization strategy for determining system-independent, i.e., transferable, potentials. We define extended ensembles as collections of chemically distinct systems that are not at equilibrium with one another and that may be under different thermodynamic conditions. We assume an atomic equilibrium ensemble, a statistical weight, and a CG representation for each system. The resulting mapped CG ensembles define a corresponding set of PMF's. We then employ a variational principle in order to determine potentials that provide an optimal approximation to this set of PMF's. The scheme is quite general and can be applied to develop both transferable and system-specific contributions to the CG potentials.

However, previous studies have provided relatively little assessment of this approach. We demonstrated that, in combination with a generalized-Yvon-Born-Green theory,[31,111] the extended ensemble approach quantitatively determined the underlying potentials from a databank of structures for multiple model proteins.[171] In addition to this proof of principle demonstration, we also applied the extended ensemble approach to parameterize transferable interaction potentials for methanol-neopentane liquid mixtures.[39] In this case, the resulting potentials performed quite reasonably for a range of mixtures, but appeared less accurate than potentials that were optimized for specific systems. This analysis was relatively cursory, though, and not expanded to other systems. Moreover, we developed the extended ensemble approach at constant volume, which is quite unsatisfactory, since one must first determine an appropriate density for each system.

In the present work, we further investigate the promise of bottom-up methods for developing predictive CG models, i.e., transferable models that accurately describe both structural and thermodynamic properties. In particular, we integrate the self-consistent pressure-matching approach within the extended ensemble framework to calculate transferable interaction potentials and system-specific volume-dependent potentials for modeling a set of binary heptane-toluene mixtures. The resulting potentials accurately describe the structure, density, and compressibility for these liquid mixtures. We demonstrate that the system-specific volume-dependent potentials can be readily predicted for other heptane-toluene mixtures. More importantly, the resulting potentials accurately reproduce both structural and thermodynamic

properties of mixtures that were not included in the parameterization. Most surprisingly, the transferable potentials can provide greater accuracy than chemically specific potentials. Finally, we employ the resulting potentials to gain additional insight into the effect of coarsening upon both the cohesive energy density and also the "simplicity" of models.

The remainder of the manuscript is organized as follows: Section II derives the extended ensemble theory; Section III summarizes the details of the present calculations; Section IV presents results from these calculations; Section V discusses these results; and Section VI reviews the key conclusions of this work. The appendix demonstrates that the self-consistent pressure-matching method minimizes a relative entropy for the extended configuration space, while the supplementary material[172] provides a more extensive presentation of the results.

3.3 Theory

In this section, we develop the extended ensemble framework for determining transferable potentials for CG models that accurately reproduce both the structure and density fluctuations of AA models. We first introduce the notion of extended ensembles in order to develop appropriate consistency criteria. We demonstrate that a generalized potential of mean force (PMF) is the appropriate potential for ensuring consistency between AA and CG models for the extended ensemble. We employ a force-matching variational principle[25,27,88,173,174] to determine transferable interaction potentials that optimally approximate the configuration dependence of this PMF. We employ a self-consistent pressure-matching approach[44,151] to determine system-specific volume dependent potentials that accurately reproduce the volume-dependence of this PMF. The appendix briefly considers the extended ensemble approach and, in particular, the self-consistent pressure-matching approach in terms of minimizing a relative entropy.

3.3.1 Extended ensembles

We define an extended ensemble as a collection of equilibrium ensembles for multiple distinct systems. In the present work, we shall assume that each system is at the same temperature, T, and samples isotropic volume fluctuations at the same

external pressure, P_0. For simplicity, we shall also assume appropriate use of periodic boundary conditions and that there are no rigid constraints.

3.3.1.1 Atomic extended ensemble

For each system in the extended ensemble, we assume an atomically detailed model that specifies two distinct, but related components: 1) an abstract "topology" variable, γ, that both labels the system and also specifies the number, n_γ, connectivity, and identity of the n_γ atoms in the atomic model; and 2) an associated potential function, u_γ, that governs the interactions between these atoms. A configuration, $\mathbf{r}_\gamma$, for the topology γ specifies the Cartesian coordinates of the n_γ atoms in the atomic model, $\mathbf{r}_\gamma = \{\mathbf{r}_{\gamma 1}, \ldots, \mathbf{r}_{\gamma n_\gamma}\}$.

A microstate, $(\gamma, \mathbf{r}_\gamma, v)$, in the atomic extended ensemble specifies a particular topology, γ, and a configuration, $\mathbf{r}_\gamma$, that is consistent with the volume, v. We model γ as a quenched random variable that is sampled with probability p_γ. For each γ, we treat $\mathbf{r}_\gamma$ and v as dynamical random variables that sample the equilibrium distribution at the given temperature, T, and external pressure, P_0:

$$p_{rv|\gamma}(\mathbf{r}_\gamma, v) = \Delta_\gamma^{-1} \exp\left[-\beta\left(P_0 v + u_\gamma(\mathbf{r}_\gamma, v)\right)\right] \tag{3.1}$$

$$\Delta_\gamma = \int \mathrm{d}v \int_{v_\gamma} \mathrm{d}\mathbf{r}_\gamma \exp\left[-\beta\left(P_0 v + u_\gamma(\mathbf{r}_\gamma, v)\right)\right], \tag{3.2}$$

with $\Delta_\gamma = \Delta_\gamma(T, P_0)$ and $\beta = 1/k_B T$. In Eq. (3.2) and in the following, the subscript v_γ indicates integration over the volume-dependent configuration space for the system γ. Thus, the atomic extended ensemble assigns a probability

$$p_{\gamma rv}(\mathbf{r}_\gamma, v) = p_\gamma p_{rv|\gamma}(\mathbf{r}_\gamma, v) \tag{3.3}$$

for each microstate $(\gamma, \mathbf{r}_\gamma, v)$. For any function $a_\gamma(\mathbf{r}_\gamma, v)$, we define conventional ensemble averages for a single topology γ:

$$\langle a_\gamma(\mathbf{r}_\gamma, v)\rangle_\gamma = \int \mathrm{d}v \int_{v_\gamma} \mathrm{d}\mathbf{r}_\gamma \, p_{rv|\gamma}(\mathbf{r}_\gamma, v) a_\gamma(\mathbf{r}_\gamma, v). \tag{3.4}$$

We define extended ensemble averages

$$\langle a_\gamma(\mathbf{r}_\gamma, v)\rangle = \sum_\gamma p_\gamma \langle a_\gamma(\mathbf{r}_\gamma, v)\rangle_\gamma . \tag{3.5}$$

In particular, the instantaneous internal pressure of the AA model is

$$\begin{aligned} \mathbf{P}_\gamma(\mathbf{r}_\gamma, v) &= n_\gamma k_B T/v + w_\gamma(\mathbf{r}_\gamma, v) && (3.6) \\ w_\gamma(\mathbf{r}_\gamma, v) &= -\left(\partial u_\gamma(\mathbf{r}_\gamma, v)/\partial v\right)_{\rho_\gamma}, && (3.7) \end{aligned}$$

and the subscript indicates the derivative is evaluated at constant scaled coordinates, ρ_γ. In the extended ensemble at constant T and P_0

$$P_0 = \langle \mathbf{P}_\gamma(\mathbf{r}_\gamma, v)\rangle_\gamma = \langle \mathbf{P}_\gamma(\mathbf{r}_\gamma, v)\rangle . \tag{3.8}$$

3.3.1.2 Coarse-grained extended ensemble

Similarly, for each system in the extended ensemble, we consider a CG model that specifies two distinct, but related components: 1) an abstract topology variable, Γ, that both labels the system and also specifies the number, N_Γ, connectivity, and identity of the N_Γ sites in the CG model; and 2) an associated potential function, U_Γ, that governs the interactions between these sites. A configuration, $\mathbf{R}_\Gamma$, for the topology Γ specifies the Cartesian coordinates of the N_Γ sites in the CG model, $\mathbf{R}_\Gamma = \{\mathbf{R}_{\Gamma 1}, \ldots, \mathbf{R}_{\Gamma N_\Gamma}\}$.

A microstate, $(\Gamma, \mathbf{R}_\Gamma, V)$, in the CG extended ensemble specifies a particular topology, Γ, and a configuration, $\mathbf{R}_\Gamma$, that is consistent with the volume, V. As for the atomic extended ensemble, we model Γ as a quenched random variable with probability P_Γ, while $\mathbf{R}_\Gamma$ and V are dynamic random variables that sample the equilibrium distribution at constant T and P_0:

$$P_{RV|\Gamma}(\mathbf{R}_\Gamma, V) \propto \exp\left[-\beta\left(P_0 V + U_\Gamma(\mathbf{R}_\Gamma, V)\right)\right]. \tag{3.9}$$

Thus, the CG extended ensemble assigns a probability

$$P_{\Gamma RV}(\mathbf{R}_\Gamma, V) = P_\Gamma P_{RV|\Gamma}(\mathbf{R}_\Gamma, V) \tag{3.10}$$

for each microstate $(\Gamma, \mathbf{R}_\Gamma, V)$.

The instantaneous internal pressure of the CG model is

$$\begin{aligned} \mathbf{P}_\Gamma(\mathbf{R}_\Gamma, V) &= N_\Gamma k_B T/V + \mathscr{W}_\Gamma(\mathbf{R}_\Gamma, V) && (3.11) \\ \mathscr{W}_\Gamma(\mathbf{R}_\Gamma, V) &= -\left(\partial U_\Gamma(\mathbf{R}_\Gamma, V)/\partial V\right)_{R_\Gamma}, && (3.12) \end{aligned}$$

and the subscript indicates the derivative is evaluated at constant scaled coordinates, R_Γ.

3.3.2 Mapped extended ensemble

We introduce two mappings that relate atomic and CG extended ensembles. First, we define a "topology mapping," μ, that maps each atomic topology, γ, to a CG topology $\Gamma = \mu(\gamma)$. In particular, $\mu(\gamma)$ must define the number, type, and connectivity of the sites in the CG representation of γ. Secondly, we define a configuration mapping, $\mathbf{M}_\gamma$, that maps each atomic configuration $\mathbf{r}_\gamma$ for γ to a CG configuration $\mathbf{R}_{\mu(\gamma)} = \mathbf{M}_\gamma(\mathbf{r}_\gamma)$ for $\Gamma = \mu(\gamma)$. In particular, $\mathbf{M}_\gamma$ determines the Cartesian coordinates for each site I in the topology $\mu(\gamma)$ as a linear combination of atomic coordinates:

$$\mathbf{R}_{\mu(\gamma)I} = \mathbf{M}_{\gamma I}(\mathbf{r}_\gamma) = \sum_i c_{\gamma;Ii}\mathbf{r}_{\gamma i}. \tag{3.13}$$

The notation indicates that the sum is performed over all atoms i that are defined by γ. The mapping coefficients in Eq. (3.13) must be normalized, $\sum_i c_{\gamma;Ii} = 1$, for each γ and relevant site $I \in \mu(\gamma)$. For simplicity, we assume that the mapping associates the sites with disjoint sets of atoms and defines the site coordinates by the mass centers for corresponding atomic sets. Thus, each microstate $(\gamma, \mathbf{r}_\gamma, v)$ in the atomic extended ensemble is mapped to a microstate $(\Gamma, \mathbf{R}_\Gamma, V)$ for a CG extended ensemble with $\gamma \to \Gamma = \mu(\gamma)$, $\mathbf{r}_\gamma \to \mathbf{R}_{\mu(\gamma)} = \mathbf{M}_\gamma(\mathbf{r}_\gamma)$, and $v \to V = v$.

These mappings then determine probability distributions for the "mapped ensemble." In particular,

$$p_\Gamma = \sum_\gamma p_\gamma \, \delta_{\mu(\gamma),\Gamma} \tag{3.14}$$

is the probability for sampling an atomic topology γ that maps to Γ. Similarly,

$$p_{\Gamma RV}(\mathbf{R}_\Gamma, V) \;=\; \sum_\gamma \int \mathrm{d}v \int_{v_\gamma} \mathrm{d}\mathbf{r}_\gamma \, p_{\gamma rv}(\mathbf{r}_\gamma, v) \, \delta_{\mu(\gamma),\Gamma} \; \delta(v - V) \, \delta\left(\mathbf{M}_\gamma(\mathbf{r}_\gamma) - \mathbf{R}_\Gamma\right) \tag{3.15}$$

is the probability for sampling an atomic microstate, $(\gamma, \mathbf{r}_\gamma, v)$, that maps to $(\Gamma, \mathbf{R}_\Gamma, V)$.

3.3.3 Consistency criteria

Consistency between atomic and CG extended ensembles requires that the CG extended ensemble sample the same microstate distribution as the mapped atomic extended ensemble:

$$p_{\Gamma RV}(\mathbf{R}_\Gamma, V) = P_{\Gamma RV}(\mathbf{R}_\Gamma, V). \tag{3.16}$$

The correct weighting for the different CG topologies can be easily achieved by defining

$$P_\Gamma = p_\Gamma. \tag{3.17}$$

It is more challenging to reproduce the distribution of configurations $\mathbf{R}_\Gamma$ and volumes V. According to Eqs. (3.9), (3.15) and (3.16), the appropriate CG potential, W_Γ, for a consistent CG model is defined to within a constant that is independent of $\mathbf{R}_\Gamma$ and V by

$$\exp\left[-\beta W_\Gamma(\mathbf{R}_\Gamma, V)\right] = \sum_\gamma p_{\gamma|\Gamma}\, v_0^{N_\Gamma+1}\, \Delta_\gamma^{-1}\, z_\gamma(\mathbf{R}_\Gamma, V) \tag{3.18}$$

$$z_\gamma(\mathbf{R}_\Gamma, V) = \int_{V_\gamma} \mathrm{d}\mathbf{r}_\gamma \exp\left[-\beta u_\gamma(\mathbf{r}_\gamma, V)\right] \delta\left(\mathbf{M}_\gamma(\mathbf{r}_\gamma) - \mathbf{R}_\Gamma\right), \tag{3.19}$$

where $p_{\gamma|\Gamma} = p_\gamma\, \delta_{\mu(\gamma),\Gamma}\, / p_\Gamma$ and v_0 is an arbitrary constant reference volume that is introduced for dimensional consistency. For each Γ, W_Γ is a many-body potential of mean force (PMF) for both the configuration, $\mathbf{R}_\Gamma$, and for the volume, V:

$$-\left(\frac{\partial W_\Gamma(\mathbf{R}_\Gamma, V)}{\partial \mathbf{R}_{\Gamma I}}\right)_V = \left\langle \mathbf{f}_{\gamma I}(\mathbf{r}_\gamma, v)\right\rangle_{\mathbf{R}_\Gamma, V} \tag{3.20}$$

$$-\left(\frac{\partial W_\Gamma(\mathbf{R}_\Gamma, V)}{\partial V}\right)_{R_\Gamma} = \left\langle \mathbf{P}_\gamma(\mathbf{r}_\gamma, v) - N_\Gamma k_B T/V\right\rangle_{\mathbf{R}_\Gamma, V}. \tag{3.21}$$

In these expressions, $\mathbf{f}_{\gamma I}(\mathbf{r}_\gamma, V)$ denotes the net force on site I in microstate $(\gamma, \mathbf{r}_\gamma, V)$ and the subscripted angular brackets denote conditioned averages over the AA extended ensemble:

$$\left\langle a_\gamma(\mathbf{r}_\gamma, v)\right\rangle_{\mathbf{R}_\Gamma, V} = \left\langle a_\gamma(\mathbf{r}_\gamma, v)\delta_{\mu(\gamma),\Gamma}\, \delta(v - V)\delta\left(\mathbf{M}_\gamma(\mathbf{r}_\gamma) - \mathbf{R}_\Gamma\right)\right\rangle / p_{\Gamma RV}(\mathbf{R}_\Gamma, V). \tag{3.22}$$

Note that, if μ maps multiple atomic topologies γ to a single CG topology Γ, then W_Γ reflects an averaging over this degeneracy.

3.3.4 Variational principles

In constructing a CG model for the extended ensemble, we define a set of approximate potentials, $U = \{U_\Gamma\}$, for the various systems, Γ, in the CG extended ensemble. Following Das and Andersen,[44] we introduce two functionals for variationally determining the potentials, U_Γ, as optimal approximations to W_Γ.

3.3.4.1 Force matching

First, we define an extended ensemble force-matching (FM) functional[173,174] for approximating the configuration dependence of W:

$$\chi_1^2[U] = \left\langle \frac{1}{3N_{\mu(\gamma)}} v^{2/3} \sum_{I \in \mu(\gamma)} \left| \mathbf{f}_{\gamma I}(\mathbf{r}_\gamma, v) - \mathbf{F}_{\mu(\gamma)I}(\mathbf{M}_\gamma(\mathbf{r}_\gamma), v) \right|^2 \right\rangle , \tag{3.23}$$

where $\mathbf{F}_{\Gamma I}(\mathbf{R}_\Gamma, V) = -\left(\partial U_\Gamma(\mathbf{R}_\Gamma, V)/\partial \mathbf{R}_{\Gamma I}\right)_V$. For each AA system, γ, the sum in χ_1^2 is performed over the sites that are defined in the corresponding CG representation, $\Gamma = \mu(\gamma)$. Also, note that Eq. (3.23) includes the $v^{2/3}$ factor introduced in Ref. 44, such that χ_1^2 will weight more heavily contributions from the larger systems in the extended ensemble. Consequently, the calculated potential may potentially be biased towards more accurately approximating the PMF for larger systems at the expense of smaller systems in the extended ensemble. However, in the present calculations, this bias should be negligible since all systems sample very similar volumes.

It follows from Eq. (3.20) that the set of generalized PMF's, $W = \{W_\Gamma\}$, minimizes χ_1^2:

$$\chi_1^2[U] = \chi_1^2[W] + \left\langle \frac{1}{3N_{\mu(\gamma)}} v^{2/3} \sum_{I \in \mu(\gamma)} \left| \Delta\mathbf{F}_{\mu(\gamma)I}(\mathbf{M}_\gamma(\mathbf{r}_\gamma), v) \right|^2 \right\rangle \tag{3.24}$$

$$\geq \chi_1^2[W], \tag{3.25}$$

where $\Delta\mathbf{F}_{\Gamma I} = \mathbf{F}_{\Gamma I} - \mathbf{F}^0_{\Gamma I}$ and $\mathbf{F}^0_{\Gamma I}$ indicates the forces derived from the PMF, W_Γ. Thus, minimizing χ_1^2 provides a variational procedure for determining potentials that optimally approximate the configuration dependence of the many-body PMF within the extended ensemble.

3.3.4.2 Pressure matching

Secondly, we define an extended ensemble pressure-matching (PM) functional for approximating the volume dependence of W_Γ:

$$\chi_2^2[U] = \left\langle \left| \mathbf{P}_\gamma(\mathbf{r}_\gamma, v) - \mathbf{P}_{\mu(\gamma)}(\mathbf{M}_\gamma(\mathbf{r}_\gamma), v) \right|^2 \right\rangle, \tag{3.26}$$

where $\mathbf{P}_\gamma$ and $\mathbf{P}_\Gamma$ are the internal pressures of AA and CG models, as defined in Eqs. (3.6) and (3.11), respectively. It follows from Eq. (3.21) that W minimizes χ_2^2

$$\begin{aligned} \chi_2^2[U] &= \chi_2^2[W] + \left\langle \left| \Delta\mathscr{W}_{\mu(\gamma)}(\mathbf{M}_\gamma(\mathbf{r}_\gamma), v) \right|^2 \right\rangle & (3.27) \\ &\geq \chi_2^2[W], & (3.28) \end{aligned}$$

where $\Delta\mathscr{W}_\Gamma = \mathscr{W}_\Gamma - \mathscr{W}_\Gamma^0$ and $\mathscr{W}_\Gamma^0$ is the volume derivative of the PMF: $\mathscr{W}_\Gamma^0(\mathbf{R}_\Gamma, V) = -\left(\partial W_\Gamma(\mathbf{R}_\Gamma, V)/\partial V\right)_{R_\Gamma}$. Thus, minimizing χ_2^2 provides a variational procedure for determining potentials that optimally approximate the volume dependence of the many-body PMF within the mapped extended ensemble.

3.3.5 Approximate potentials

In practice, we cannot determine the many-body PMF, W_Γ. Rather, we employ the FM and PM variational principles to determine a potential, U_Γ, that optimally approximates the configuration- and volume-dependence of W_Γ. For each CG topology, Γ, the approximate CG potential, U_Γ, may be expressed:

$$U_\Gamma(\mathbf{R}_\Gamma, V) = U_{\Gamma R}(\mathbf{R}_\Gamma) + U_{\Gamma V}(V). \tag{3.29}$$

The "interaction potential," $U_{\Gamma R}$, depends upon the volume, V, only implicitly via periodic boundary conditions. We shall define $U_{\Gamma R}$ by a sum of simpler "transferable," i.e., system-independent, potentials that each govern a particular type of interaction. In contrast, the second term, $U_{\Gamma V}$, depends explicitly upon both the volume and also the particular system, but is independent of the CG configuration. As proposed by Das and Andersen,[44] we first optimize $U_{\Gamma R}$ by minimizing χ_1^2. Then, given this interaction potential, we optimize $U_{\Gamma V}$ by minimizing χ_2^2 for fixed $U_{\Gamma R}$.

We shall further assume that $U_{\Gamma R}$ adopts the common, simple form

$$U_{\Gamma R}(\mathbf{R}_\Gamma) = U_{\Gamma 2}(\mathbf{R}_\Gamma) + U_{\Gamma\theta}(\mathbf{R}_\Gamma), \tag{3.30}$$

where $U_{\Gamma 2}$ is a sum of transferable pair potentials that depend only upon the distance between pairs of sites, while $U_{\Gamma\theta}$ is a sum of transferable bonded potentials that depend upon bond or dihedral angles and, thus, do not directly contribute to the isotropic pressure.[97] The internal pressure of the CG model may then be expressed:

$$\mathbf{P}_\Gamma(\mathbf{R}_\Gamma, V) = \mathbf{P}_{\Gamma R}(\mathbf{R}_\Gamma, V) + F_{\Gamma V}(V), \tag{3.31}$$

where $F_{\Gamma V} = -\mathrm{d}U_{\Gamma V}(V)/\mathrm{d}V$ is the "pressure correction" due to $U_{\Gamma V}$, while $\mathbf{P}_{\Gamma R}$ is the internal pressure due to the interaction potential, $U_{\Gamma R}$:

$$\mathbf{P}_{\Gamma R}(\mathbf{R}_\Gamma, V) = N_\Gamma k_B T/V + \frac{1}{3V}\sum_{(I,J)\in\Gamma} R_{IJ} F_{2;IJ}(R_{IJ}). \tag{3.32}$$

In this expression, R_{IJ} is the minimum image distance between the (I, J) pair of sites, $F_{2;IJ}$ indicates the magnitude of the corresponding 2-body force, and the sum is performed over all relevant pairs for Γ.

3.3.5.1 Interaction potential

For each Γ, we define $U_{\Gamma R}$ as a sum of system-independent potentials, U_ζ, each of which depends upon a scalar function, ψ_ζ, of the coordinates, $\mathbf{R}_{\Gamma\lambda}$, for a particular subset, λ, of sites in the CG topology, Γ:

$$U_{\Gamma R}(\mathbf{R}_\Gamma) = \sum_\zeta \sum_{\lambda\in\Gamma} U_\zeta(\psi_\zeta(\mathbf{R}_{\Gamma\lambda})). \tag{3.33}$$

Thus, $U_{\Gamma R}$ depends upon the CG topology, Γ, only for specifying the particular interactions that are relevant for the system. In each system, we employ the same transferable potential, U_ζ, for modeling the ζ interaction.

Each transferable interaction potential, U_ζ, is expressed as a linear combination of simple system-independent basis functions, $u_{\zeta d}(x)$, e.g., splines, with coefficients, $\phi_{\zeta d}$:

$$U_\zeta(x) = \sum_d \phi_{\zeta d} u_{\zeta d}(x) \tag{3.34}$$

The force on each site, I, is then

$$\mathbf{F}_{\Gamma I}(\mathbf{R}_\Gamma) = \sum_\zeta \sum_d \phi_{\zeta d} \mathscr{G}_{\Gamma I;\zeta d}(\mathbf{R}_\Gamma) \tag{3.35}$$

where

$$\mathscr{G}_{\Gamma I;\zeta d}(\mathbf{R}_\Gamma) = \sum_{\lambda \in \Gamma} f_{\zeta d}(\psi_\zeta(\mathbf{R}_{\Gamma\lambda}))\ \partial \psi_\zeta(\mathbf{R}_{\Gamma\lambda})/\partial \mathbf{R}_{\Gamma I} \tag{3.36}$$

and $f_{\zeta d}(x) = -\mathrm{d}u_{\zeta d}(x)/\mathrm{d}x$. For notational convenience, we introduce a "super-index" D for identifying a particular combination ζd:

$$\mathbf{F}_{\Gamma I}(\mathbf{R}_\Gamma) = \sum_D \phi_D \mathscr{G}_{\Gamma I;D}(\mathbf{R}_\Gamma). \tag{3.37}$$

Given this basis set expansion for $\mathbf{F}_\Gamma$, χ_1^2 becomes a simple quadratic form in the potential parameters, ϕ_D:

$$\chi_1^2(\phi) \;=\; \chi_1^2(0) - 2\sum_D b_D \phi_D + \sum_D \sum_{D'} G_{DD'} \phi_D \phi_{D'} \tag{3.38}$$

where

$$b_D \;=\; \left\langle \frac{1}{3N_{\mu(\gamma)}} v^{2/3} \sum_{I \in \mu(\gamma)} \mathbf{f}_{\gamma I}(\mathbf{r}_\gamma, v) \cdot \mathscr{G}_{\mu(\gamma)I;D}(\mathbf{M}_\gamma(\mathbf{r}_\gamma)) \right\rangle \tag{3.39}$$

$$G_{DD'} \;=\; \left\langle \frac{1}{3N_{\mu(\gamma)}} v^{2/3} \sum_{I \in \mu(\gamma)} \mathscr{G}_{\mu(\gamma)I;D}(\mathbf{M}_\gamma(\mathbf{r}_\gamma)) \cdot \mathscr{G}_{\mu(\gamma)I;D'}(\mathbf{M}_\gamma(\mathbf{r}_\gamma)) \right\rangle. \tag{3.40}$$

The optimal interaction potential is then determined by the coefficients that satisfy the normal system of linear equations

$$\sum_{D'} G_{DD'} \phi_{D'} = b_D. \tag{3.41}$$

Because the same potential parameters, ϕ_D, are used for each system, these normal equations average over multiple atomic systems.

3.3.5.2 Volume-dependent potential

While we assume that the interaction potentials are transferable, we shall determine system-specific volume-dependent potentials, $U_{\Gamma V}$, for each CG topology, Γ. We

define

$$U_{\Gamma V}(V) = \sum_d \psi_{\Gamma V d}\, u_{\Gamma V d}(V), \tag{3.42}$$

where

$$u_{\Gamma V d}(V) = \begin{cases} N_\Gamma(V/\bar{v}_\Gamma) & \text{for } d = 1 \\ N_\Gamma\left(V/\bar{v}_\Gamma - 1\right)^d & \text{for } d \geq 2 \end{cases} \tag{3.43}$$

and $\bar{v}_\Gamma$ is the average volume in the mapped ensemble for Γ. The pressure correction $F_{\Gamma V}(V) = -\mathrm{d}U_{\Gamma V}(V)/\mathrm{d}V$ is then

$$F_{\Gamma V}(V) = \sum_d \psi_{\Gamma V d}\, f_{\Gamma V d}(V) \tag{3.44}$$

where $f_{\Gamma V d}(V) = -\mathrm{d}u_{\Gamma V d}(V)/\mathrm{d}V$. In practice, we only include the $d = 1$ and 2 terms in this sum:

$$F_{\Gamma V}(V) = -\left(N_\Gamma/\bar{v}_\Gamma\right)\left[\psi_{\Gamma V 1} + 2\psi_{\Gamma V 2}\left(V - \bar{v}_\Gamma\right)/\bar{v}_\Gamma\right] \tag{3.45}$$

where the two parameters correspond to corrections for the pressure, $\Delta\mathbf{P}_\Gamma$, and the inverse compressibility, $\Delta\kappa_{T;\Gamma}^{-1}$, when $V = \bar{v}_\Gamma$:

$$\Delta\mathbf{P}_\Gamma = -N_\Gamma\psi_{\Gamma V 1}/\bar{v}_\Gamma \tag{3.46}$$

$$\Delta\kappa_{T;\Gamma}^{-1} = 2N_\Gamma\psi_{\Gamma V 2}/\bar{v}_\Gamma. \tag{3.47}$$

Having determined $U_{\Gamma R}$ by minimizing χ_1^2, we then determine the parameters for $U_{\Gamma V}$ by minimizing χ_2^2:

$$\chi_2^2[U_{\Gamma V}|U_{\Gamma R}] = \sum_\gamma p_\gamma\, \chi_{2\gamma}^2(\psi_{\mu(\gamma)}) \tag{3.48}$$

where $\chi_{2\gamma}^2(\psi_\Gamma)$ is the PM functional for a single topology, γ:

$$\chi_{2\gamma}^2(\psi_{\mu(\gamma)}) = \left\langle \left|\mathbf{P}_\gamma(\mathbf{r}_\gamma, v) - \mathbf{P}_{\mu(\gamma)}(\mathbf{M}_\gamma(\mathbf{r}_\gamma), v)\right|^2 \right\rangle_\gamma \tag{3.49}$$

$$= \left\langle \left|\delta\mathbf{P}_{\gamma R}(\mathbf{r}_\gamma, v) - \sum_d f_{\mu(\gamma)Vd}(v)\psi_{\mu(\gamma)Vd}\right|^2 \right\rangle_\gamma, \tag{3.50}$$

where $\delta\mathbf{P}_{\gamma R}(\mathbf{r}_\gamma, v) = \mathbf{P}_\gamma(\mathbf{r}_\gamma, v) - \mathbf{P}_{\mu(\gamma)R}(\mathbf{M}_\gamma(\mathbf{r}_\gamma), v)$, such that χ_2^2 is a simple quadratic form in $\psi_{\Gamma V d}$. Since the parameters, $\psi_{\Gamma V d}$, for $U_{\Gamma V}$ are system-specific

and we have assumed a 1-1 relationship between γ and Γ in this work, we perform a separate pressure-matching calculation for each different AA topology, γ.

CG simulations with the parameters, $\psi_{\Gamma V d}$, that minimize χ_2^2 reproduce atomic density fluctuations with qualitative, but not quantitative, accuracy. As discussed in our prior work,[151] since minimizing χ_2^2 corresponds to matching the atomic pressure over the mapped ensemble, we believe that this discrepancy results from subtle differences between the mapped and simulated CG extended ensembles. Consequently, we have adopted an iterative self-consistent procedure to further optimize the parameters, $\psi_{\Gamma V d}$, for each CG system, Γ. Given the pressure correction, $F_{\Gamma V}(V)$, for Γ that is obtained from minimizing χ_2^2, we perform a short simulation at constant external pressure in order to estimate the pressure equation of state, $\mathbf{P}_\Gamma(V)$, for the CG model. We compare $\mathbf{P}_\Gamma(V)$ with the AA pressure equation of state, $\mathbf{P}_\gamma(V)$, determined for the corresponding AA system, i.e., $\mu(\gamma) = \Gamma$. The difference between the AA and CG pressure equations of state determines the error in the pressure correction,

$$\delta F_{\mu(\gamma)V}(V) = \mathbf{P}_\gamma(V) - \mathbf{P}_{\mu(\gamma)}(V), \tag{3.51}$$

and determines a corresponding correction to $U_{\Gamma V}$, $\delta F_{\Gamma V}(V) = -\mathrm{d}\delta U_{\Gamma V}(V)/\mathrm{d}V$, such that the CG model more accurately reproduces the volume dependence of the PMF. We then determine the pressure equation of the state for the CG model with the modified pressure correction and repeat this procedure until the CG model adequately reproduces the AA pressure equation of state. The appendix demonstrates that this self-consistent pressure-matching approach corresponds to a different variational approach for optimizing $U_{\Gamma V}(V)$ by minimizing a relative entropy in the extended configuration space.

3.4 Computational Details

3.4.1 Atomistic Simulations Details

We employed the GROMACS 4.5.3 simulation package,[103] while using double-precision floating-point arithmetic, to perform atomically detailed simulations for pure heptane, pure toluene, and for the 7 different heptane-toluene mixtures

Table 3.1. Compositions of the simulated AA models, as well as the corresponding probabilities for extended ensemble averages.

Hep:Tol	N_{hep}	N_{tol}	p_γ
0:1	0	642	
1:9	30	267	
1:4	119	476	0.2
2:3	221	321	0.2
1:1	267	267	0.2
3:2	302	208	0.2
4:1	392	98	0.2
9:1	267	30	
1:0	461	0	

that are described in Table 3.1. We adopted the OPLS-AA force field[104] for modeling all atomic interactions and employed the particle mesh Ewald method[105] for treating electrostatic interactions. We employed the same equilibration and simulation protocols that were reported for our previous constant NPT simulations of pure heptane and toluene.[151] In particular, we truncated both short-ranged potentials and also the real space contribution to electrostatic interactions at 1.2 nm. After equilibration, we simulated each system for 100 ns in the constant NPT ensemble, while employing the Parrinello-Rahman barostat[110] and Nosé-Hoover thermostat[108,109] with coupling constants of 0.5 ps and 5 ps, respectively, to maintain a constant external pressure $P_0 = 1.0$ bar and temperature $T = 303$ K. We employed these constant NPT trajectories both for calculating the CG potentials and also for characterizing the equilibrium density fluctuations and pressure equations of state for the AA models. In order to characterize the structural properties of the AA models, we also simulated each system for an additional 50 ns in the constant NVT ensemble at $T = 303$ K. We determined the initial configuration and volume for these constant NVT simulations from the constant NPT simulations by employing the last sampled configuration in which the instantaneous volume was within one standard deviation of the average volume. We sampled configurations every 1 ps and estimated standard errors by assuming that the configurations provided statistically independent samples of the pressure.

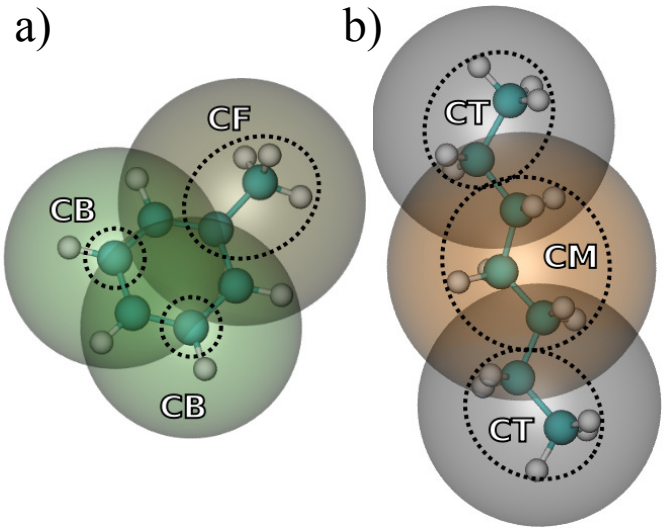

Figure 3.1. CG mapping for toluene (a) and heptane (b). The coordinates for each site are determined by the mass center of the corresponding atomic group, which is indicated by the dashed circle. The transparent spheres indicate the position and size of each site, which is estimated from the corresponding site-site rdf.

3.4.2 CG Representation

We developed CG models for each of the liquid systems described in Table 3.1. In each case, we modeled heptane and toluene with the 3-site representations that are illustrated in Fig. 3.1. The CG mapping associated each site with a corresponding atomic group and determined the site coordinates by the corresponding mass center. The dashed circles in Fig. 3.1 indicate these atomic groups, while the colored spheres indicate the position and size of the CG sites.

3.4.3 Force-matching interaction potentials

The CG interaction potentials included both intramolecular and intermolecular contributions. In order to describe intramolecular geometry and flexibility, we employed bond potentials between adjacent CG sites in each molecule, as well as an angle potential to model the three sites in heptane. We did not employ an angle potential for modeling toluene. We represented these bond and angle potentials with linear spline functions on grids with 0.01 nm and 0.5 degree spacings, respectively. We modeled intermolecular interactions with short ranged pair potentials between

each pair of sites in distinct molecules. We represented these potentials with cubic spline functions on a grid with 0.001 nm spacing, while truncating these potentials at 1.4 nm. The CG models did not include explicit electrostatic interactions or rigid constraints.

We employed the MS-CG force-matching variational principle[25,27,88] in order to optimize these potentials for specific systems. Additionally, we employed the extended ensemble force-matching variational principle[39] in order to optimize a single set of system-independent interaction potentials for heptane-toluene mixtures. As indicated in Table 3.1, the extended ensemble included mixtures with 1:4, 2:3, 1:1, 3:2, and 4:1 heptane:toluene ratios, while assigning an equal weight $p_\gamma = 0.2$ to each mixture. We determined the optimal transferable parameters that minimize χ_1^2 by numerically solving the linear system of equations in Eq. (3.41). We refer to the resulting system-independent potentials as the extended ensemble (xn) interaction potentials.

In all cases, we estimated the relevant ensemble averages with configurations sampled from AA simulations. Despite the relatively extensive simulations, the calculated non-bonded forces demonstrated traces of statistical noise. (See supplementary material.[172]) Since this noise can impact the calculated pressures, we smoothed the tails of these potentials via the same procedure described in our previous work.[151]

3.4.4 Pressure-matching volume potentials

As described above, we determined two interaction potentials, $U_{\Gamma R}$, for each system Γ in the CG extended ensemble: 1) a MS-CG potential optimized for each specific system; and 2) an xn potential optimized for transferability across the entire extended ensemble. Additionally, for each system, Γ, in the CG extended ensemble, as well as for pure heptane and pure toluene, we determined two distinct volume-dependent potentials, $U_{\Gamma V}$, that were optimized independently for the MS-CG and xn interaction potentials, $U_{\Gamma R}$. We optimized both of these volume-dependent potentials via the iterative self-consistent pressure matching approach that is described in Section 3.3.5.2.[151] In brief, given a CG interaction potential, $U_{\Gamma R}$, we first determined a volume-dependent potential, $U_{\Gamma V}$, by minimizing the pressure-matching functional, $\chi_{2\gamma}^2$, for the corresponding AA system. We then iter-

Table 3.2. Average pressure corrections and inverse compressibility corrections, as well as the number of iterations required to converge the pressure correction. The corrections to the pressures and inverse compressibilities are given in units of 10^3 bar. Models with an asterisk (*) for N_{Iter} did not converge within 10 iterations. In these cases, the pressure correction was determined according to the procedure described in the Methods section.

	$\Delta\mathbf{P}$		$\Delta\kappa_T^{-1}$		N_{Iter}	
Hep:Tol	MS-CG	XN	MS-CG	XN	MS-CG	XN
0:1	-2.61	-1.64	-2.66	-0.45	*	*
1:9	-	-1.58	-	-0.74	-	-
1:4	-2.01	-1.52	-1.67	-1.22	2	6
2:3	-1.57	-1.45	-1.71	-1.81	4	2
1:1	-1.55	-1.42	-1.88	-1.51	*	1
3:2	-1.21	-1.39	-1.95	-1.75	2	6
4:1	-0.94	-1.36	-1.33	-2.25	4	4
9:1	-	-1.33	-	-2.79	-	-
1:0	-0.75	-1.31	-2.04	-2.70	1	6

atively refined $U_{\Gamma V}$ until constant NPT simulations with the CG model adequately reproduced the pressure equation of state for the corresponding atomic model.

The volume-dependent potentials are determined by two parameters, $\psi_{\Gamma V1}$ and $\psi_{\Gamma V2}$, which correspond to average corrections for the pressure, $\Delta\mathbf{P}_\Gamma$ and for the (inverse) compressibility, $\Delta\kappa_{T;\Gamma}^{-1}$ respectively. The iterative pressure matching calculations converged quite rapidly in all but 3 of the 14 cases. However, in these 3 cases the calculated parameters oscillated about appropriate values. In these cases, we determined the 2 parameters from different iterations that accurately modeled the density and compressibility, respectively. Table 3.2 presents the optimized parameters.

Finally, having optimized pressure corrections for heptane, toluene, and 5 different mixtures, we developed an approach for predicting pressure corrections for other heptane-toluene mixtures. For a mixture Γ with heptane mole fraction, χ_{hep}, and toluene mole fraction, $\chi_{\text{tol}} = 1 - \chi_{\text{hep}}$, we assume that the average pressure correction may be expressed:

$$\Delta\mathbf{P}_\Gamma \;=\; \chi_{\text{hep}}^2\Delta\mathbf{P}_{\text{hep}} \;+\; \chi_{\text{tol}}^2\Delta\mathbf{P}_{\text{tol}} \;+\; \chi_{\text{hep}}\chi_{\text{tol}}\Delta\mathbf{P}_{\text{mix}}, \tag{3.52}$$

where $\Delta\mathbf{P}_{\text{hep}}$ and $\Delta\mathbf{P}_{\text{tol}}$ are the average pressure corrections for pure heptane and

pure toluene, respectively. We determine the one empirical parameter, $\Delta\mathbf{P}_{\text{mix}}$, according to

$$\Delta\mathbf{P}_{\text{mix}} = \left\langle \frac{1}{\chi_{\text{hep}}\chi_{\text{tol}}} \left\{ \Delta\mathbf{P}_{\mu(\gamma)} - \left(\chi_{\text{hep}}^2 \Delta\mathbf{P}_{\text{hep}} \; + \; \chi_{\text{tol}}^2 \Delta\mathbf{P}_{\text{tol}} \right) \right\} \right\rangle, \tag{3.53}$$

where the angular brackets indicate an extended ensemble average over systems, γ, and $\Delta\mathbf{P}_{\mu(\gamma)}$ is the average optimized pressure correction determined for the corresponding CG model $\Gamma = \mu(\gamma)$. We apply the same scheme to predict a corresponding correction for the inverse compressibility. These corrections then determine a volume-dependent potential for any heptane-toluene mixture according to Eqs. (3.42) - (3.47).

3.4.5 CG Simulation Details

We simulated each CG model in the constant NPT ensemble with external pressure $P_0 = 1.0$ bar and temperature $T = 303$ K in order to characterize the density fluctuations and pressure equations of state for the CG models. We performed these simulations with a modified version[151] of LAMMPS (17Jun13)[114] that incorporates the pressure correction into the Martyna-Tuckerman-Tobias-Klein barostat.[115] These simulations employed the same protocols and parameters that were reported in our previous work.[151] We also performed a series of simulations in the constant NVT ensemble in order to assess the structural accuracy and transferability of the various calculated CG potentials. We performed these simulations with GROMACS 4.5.3, while adopting the same simulation parameters as the constant NPT atomistic simulations, except that the volume did not fluctuate and the CG model did not include explicit electrostatic interactions. We performed each constant NVT CG simulation for 15 ns and employed the last 14 ns for analysis.

3.5 Results

In this section, we investigate the accuracy and transferability of potentials obtained from the extended ensemble framework. We first consider an extended ensemble that includes 5 mixtures with 1:4, 2:3, 1:1, 3:2, and 4:1 heptane:toluene ratios. We employ the MS-CG variational principle[25] to optimize system-specific interaction

potentials for each of these mixtures. Additionally, we employ the extended ensemble framework[39] to determine a single set of interaction potentials that provide optimal transferability across the 5 mixtures. For each system and each potential, we employ the self-consistent pressure-matching approach[151] to optimize a pressure correction, $F_{\Gamma V}(V) = -\mathrm{d}U_{\Gamma V}(V)/\mathrm{d}V$, that reproduces the pressure equation of state for the corresponding AA model. After presenting these potentials, we assess their accuracy for modeling the structure and density fluctuations of the heptane-toluene mixtures in the extended ensemble. We next evaluate their transferability for modeling other heptane-toluene mixtures. Finally, we employ the resulting models to investigate the effects of coarse-graining upon the model thermodynamic properties.

Figure 3.2 compares the system-specific MS-CG potentials with the transferable extended ensemble (xn) potentials. In Fig. 3.2 and in the following, red curves present results for pure toluene and violet curves present results for pure heptane, while brown, orange, green, cyan, and blue curves present results for mixtures with 1:4, 2:3, 1:1, 3:2, and 4:1 heptane:toluene ratios.

Figure 3.2a compares the calculated pair potentials for the interaction between the heptane terminal CT sites and toluene methyl CF sites. (The supplementary material presents all of the calculated pair potentials.[172]) The CT-CF potential is quite representative of the 15 different types of pair potentials, since it demonstrates typical statistical uncertainty and variation with composition. All of the calculated CT-CF potentials demonstrate two distinct attractive wells. The colored curves demonstrate that the MS-CG potentials vary systematically with increasing heptane content as the first minimum deepens and shifts to smaller distances. The black curve presents the CT-CF xn potential, which is most similar to the MS-CG potential optimized for the 1:1 mixture.

Figure 3.2b presents the optimized pressure corrections. For each system, the solid and dashed lines correspond to the pressure corrections that were optimized for the system-specific MS-CG interaction potentials and for the transferable xn interaction potentials, respectively. As indicated in Eqs. (3.45)-(3.47), the midpoint of each line indicates the average density and the average pressure correction for the corresponding CG model, while the slope indicates the corresponding correction to the compressibility. Table 3.2 presents the calculated parameters for each pressure correction.

Figure 3.2b demonstrates that the calculated pressure corrections decrease the

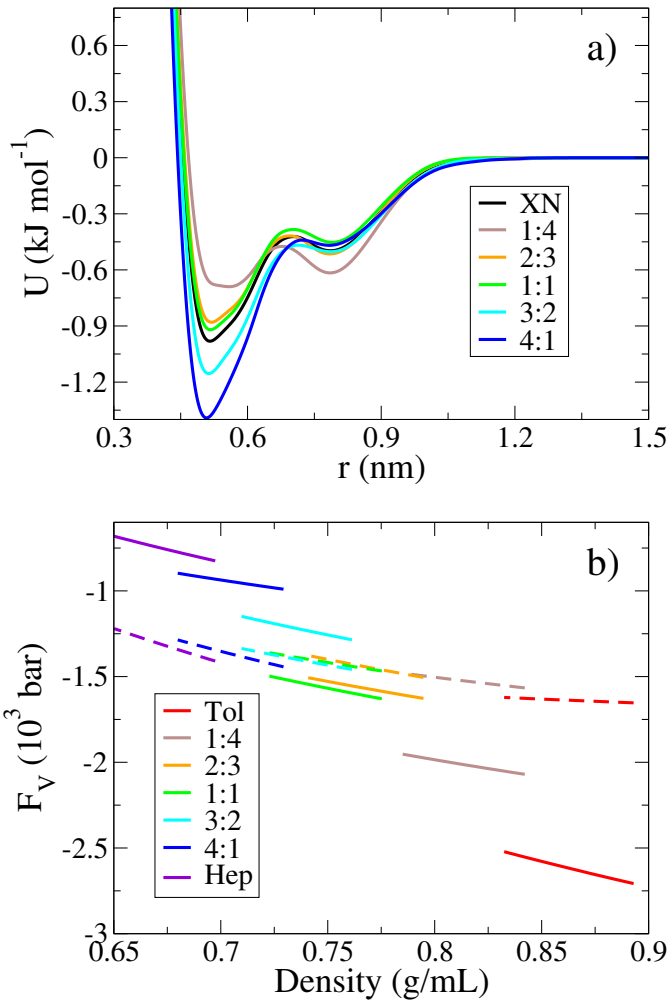

Figure 3.2. Calculated interaction potentials and pressure corrections. Panel a) presents calculated CF-CT pair potentials. In panel a), the colored curves indicate system-specific MS-CG pair potentials for the various mixtures, while the black curve indicates the corresponding transferable extended ensemble potential. Panel b) presents calculated pressure corrections as a function of density. For each system, the solid line indicates the pressure correction that is optimized for the corresponding system-specific MS-CG potentials, while the dashed line indicates the pressure correction that is optimized for the transferable extended ensemble potentials. In this figure, red and purple curves indicate results for pure toluene and pure heptane, while brown, orange, green, cyan, and blue curves indicate results for mixtures with 1:4, 2:3, 1:1, 3:2, and 4:1 heptane:toluene ratio, respectively.

internal pressure and compressibility of each CG model. The pressure corrections generally increase in magnitude with increasing toluene content. In the case of the xn pressure corrections, the density dependence also systematically decreases with increasing toluene content. Since the same pair potentials are employed in each xn model, the corresponding pressure corrections demonstrate relatively little variation. In contrast, since the MS-CG pair potentials vary significantly between systems, the corresponding pressure corrections also demonstrate much greater variance.

For each of the 5 heptane-toluene mixtures in the extended ensemble, we performed constant NVT simulations with each of the 5 system-specific MS-CG potentials and also with the transferable xn potentials. In contrast to other iterative structure-based approaches,[26,28–30] the MS-CG method does not explicitly ensure that the CG models accurately reproduce AA radial distribution functions (rdfs).[101,117] Figure 3.3 assesses the accuracy of the calculated potentials for modeling the structure of these mixtures. As a representative example, Fig. 3.3a presents simulated CT-CB rdfs for the 1:1 heptane:toluene mixture, while the supplementary material presents a comprehensive analysis.[172] As expected, in comparison to MS-CG potentials that are optimized for other mixtures, the MS-CG potential that is optimized for the 1:1 mixture appears to most accurately reproduce the corresponding CT-CB rdf. Notably, the transferable xn potentials appear to provide a similarly, or possibly even more, accurate reproduction of this AA rdf.

Figure 3.3b globally assesses the structural accuracy of each CG potential for each heptane-toluene mixture in the extended ensemble. For each mixture, γ, and each calculated CG potential, U_Γ, we calculated the absolute error in the simulated CG rdfs for each relevant pair of site types $\zeta \in \mu(\gamma)$ at each distance r. We defined the mean absolute error (MAE), $\mathrm{MAE}(\gamma; U_\Gamma)$, by averaging this absolute error over all distances and all relevant site pairs $\zeta \in \mu(\gamma)$:

$$\mathrm{MAE}(\gamma; U_\Gamma) \;=\; \frac{1}{N_{\gamma;\zeta}} \sum_{\zeta \in \mu(\gamma)} \frac{1}{N_{\zeta;r}} \sum_{r} \left| g_{\gamma;\zeta}(r) - g_{\mu(\gamma);\zeta}(r; U_\Gamma) \right|, \tag{3.54}$$

where $g_{\gamma;\zeta}(r)$ is an AA rdf; $g_{\mu(\gamma);\zeta}(r; U_\Gamma)$ is the corresponding rdf obtained from CG simulations with potential U_Γ; $N_{\gamma;\zeta}$ is the number of relevant site pairs; and $N_{\zeta;r} = 2.0\mathrm{nm}/0.002\mathrm{nm} = 1000$ is the number of bins treated in tabulating the rdfs.

Figure 3.3b presents the MAE of each CG potential, U_Γ, for modeling each

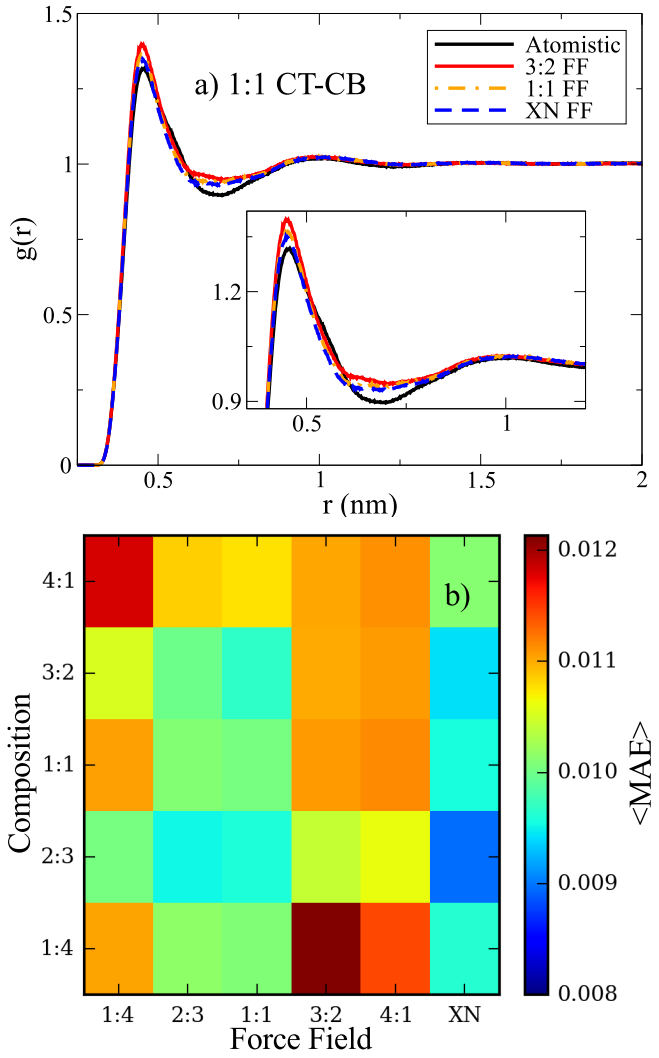

Figure 3.3. Structural accuracy for modeling systems in the extended ensemble. Panel a) presents CT-CB radial distribution functions (rdfs) from simulations of the 1:1 heptane:toluene mixture that employ various potentials. The solid black curve corresponds to AA simulations. The solid red and dashed-dotted orange curves correspond to simulations with system-specific MS-CG potentials that were optimized for the 3:2 and 1:1 mixtures, respectively. The dashed blue curve corresponds to simulations with the transferable xn potentials. The inset highlights the first two peaks of the rdfs. Panel b) presents an intensity map characterizing the mean absolute error (MAE) for modeling the AA rdfs of each mixture with each force field. Warmer colors indicate larger error, while cooler colors indicate smaller error.

mixture, γ. As suggested by Fig. 3.3a, all of the CG potentials perform quite well within the extended ensemble with relatively small errors, i.e., MAE ≈ 0.01. Figure 3.3b also demonstrates two interesting results. First, given the MS-CG variational approach for optimizing CG potentials and the MAE metric for structural accuracy, system-specific MS-CG potentials do not necessarily provide optimal accuracy. Even more interestingly, Fig. 3.3b demonstrates that the xn potential appears not only most transferable, but also most accurate for each of the 5 heptane-toluene mixtures. This surprising accuracy may be partly due to the better statistics obtained from the extended ensemble approach.[175] In particular, the supplementary material[172] demonstrates that the MS-CG potentials for statistically rare interactions can reflect modest statistical uncertainty even after 100 ns of AA simulations with relatively large systems, although this uncertainty could be somewhat mitigated by adopting more sophisticated inference techniques.[176] More fundamentally, though, the discussion section proposes that, by smoothing over complex system-specific many-body correlations, the extended ensemble approach may also address a basic approximation of the MS-CG variational principle.[118]

Figure 3.4 demonstrates that the CG models also accurately reproduce the pressure, density fluctuations, and compressibility of the heptane-toluene mixtures in the extended ensemble. In Figs. 3.4a and 3.4b, the solid curves present results for constant NPT simulations with the OPLS-AA model,[104] while the dashed and dotted curves present results for constant NPT simulations with the system-specific MS-CG potentials and with the transferable xn potentials, respectively. Figure 3.4a demonstrates that the CG models reproduce the AA density fluctuations with nearly quantitative accuracy. In particular, Table 3.3 demonstrates that the CG models accurately model the liquid density and compressibility, which correspond to the average and standard deviation of the volume distributions, respectively. Similarly, Fig. 3.4b demonstrates that the CG models almost quantitatively reproduce the AA pressure equations of state. Of course, the pressure corrections were explicitly parameterized to reproduce the AA pressure equations of state for these 7 liquid systems. Nevertheless, the level of agreement is quite satisfying.

We next assess the transferability of the calculated CG potentials for modeling heptane-toluene mixtures that were not included in the parameterization extended ensemble. Figure 3.5a presents CT-CT rdfs from constant NVT simulations of the 1:9 heptane:toluene mixture, since this rdf nicely distinguishes the performance of

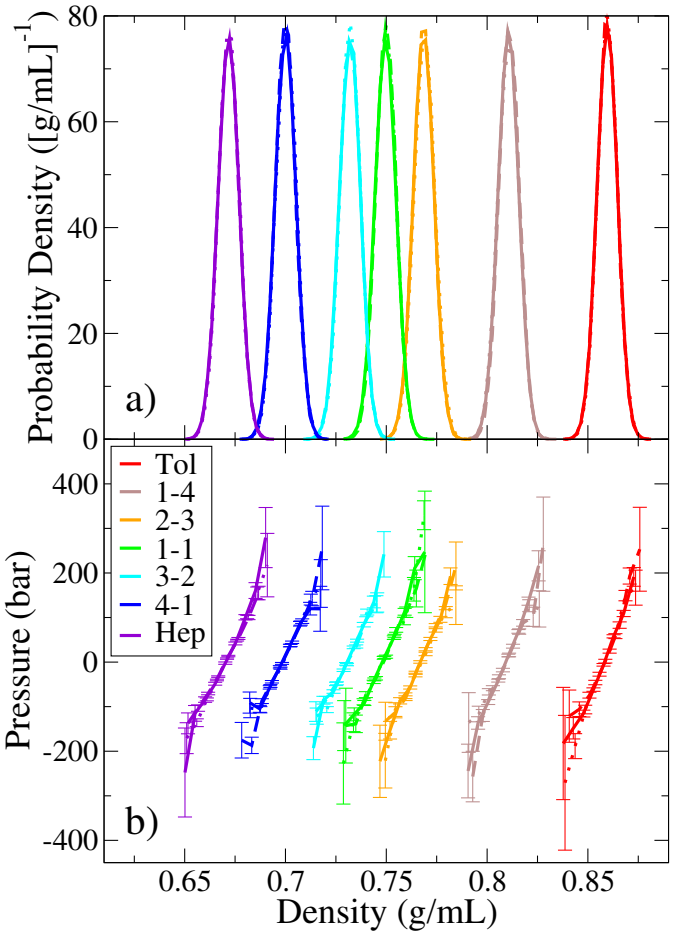

Figure 3.4. Simulated density distributions (a) and pressure equations of state (b) for pure heptane, pure toluene, and the mixtures included in the extended ensemble. The various systems are indicated by the color scheme of Fig. 3.2. In both panels, solid curves indicate results for the OPLS-AA potential, dashed curves indicate results for system-specific MS-CG potentials, and dotted curves indicate results for transferable xn potentials. The bars in panel b) indicate the standard error in the simulated pressure over each volume increment.

the different potentials. The system-specific MS-CG potentials that were optimized for the 3:2 and 1:1 heptane:toluene mixtures underestimate the first two peaks in the AA rdf. In comparison, the transferable xn potential more accurately reproduces the first peak, although it still underestimates the second peak of the AA rdf. The supplementary material presents a more exhaustive analysis of the simulated rdfs.[172]

In analogy to Fig. 3.3b, Fig. 3.5b comprehensively assesses the transferability of these potentials by presenting their MAE for modeling 4 additional liquids:

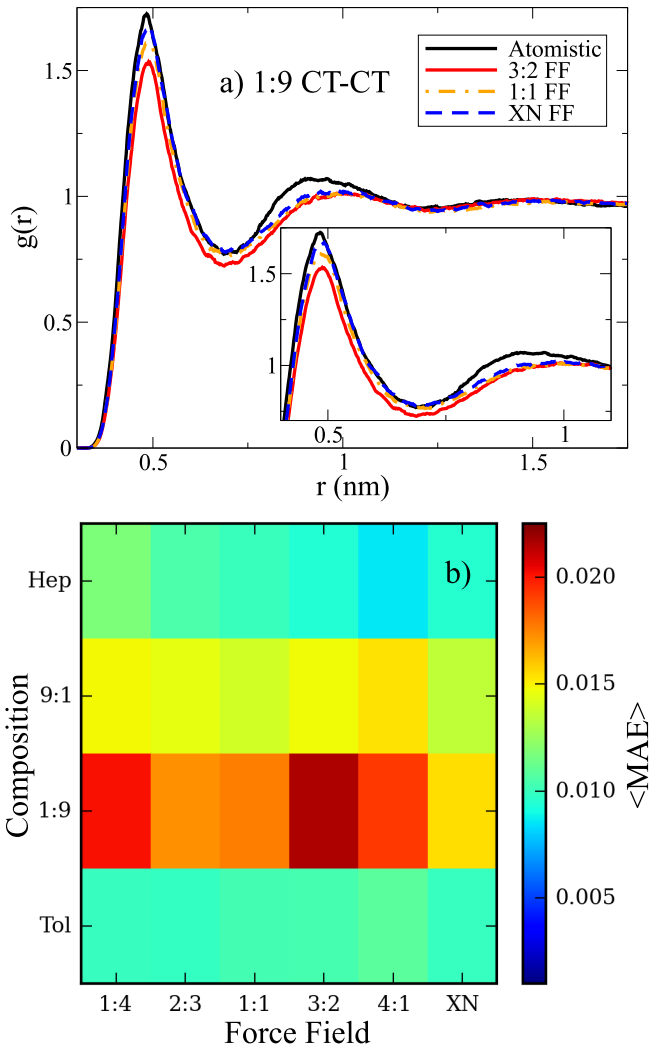

Figure 3.5. Structural accuracy for modeling systems that were not included in the extended ensemble. Panel a) presents CT-CT rdfs from simulations of the 1:9 heptane:toluene mixture that employ various potentials. The solid black curve corresponds to AA simulations. The solid red and dashed-dotted orange curves correspond to simulations with system-specific MS-CG potentials that were optimized for the 3:2 and 1:1 mixtures, respectively. The dashed blue curve corresponds to simulations with the transferable xn potentials. The inset highlights the first two peaks of the rdfs. Panel b) presents an intensity map of the MAE for modeling the AA rdfs of each mixture with each force field. Warmer colors indicate larger error, while cooler colors indicate smaller error.

Table 3.3. Equilibrium densities (g/mL) and compressibilities (10^{-4} bar^{-1}) obtained from constant NPT simulations.

	ρ			κ_T		
Hep:Tol	AA	MS-CG	XN	AA	MS-CG	XN
0:1	0.86	0.86	0.86	0.96	0.93	0.94
1:9	0.83	-	0.84	0.98	-	0.84
1:4	0.81	0.81	0.81	1.11	1.05	1.15
2:3	0.76	0.76	0.76	1.23	1.21	1.25
1:1	0.75	0.75	0.75	1.33	1.24	1.28
3:2	0.72	0.72	0.72	1.38	1.41	1.31
4:1	0.70	0.70	0.70	1.51	1.43	1.45
9:1	0.69	-	0.69	1.49	-	1.52
1:0	0.67	0.67	0.67	1.65	1.67	1.57

pure heptane and pure toluene, as well as 1:9 and 9:1 heptane:toluene mixtures. As suggested by Fig. 3.5a, the potentials provide somewhat reduced accuracy for modeling systems outside of the parameterization extended ensemble, although typical errors remain quite small with MAE ≤ 0.02. More importantly, with the exception of pure heptane, for which the 4:1 MS-CG potential is slightly more accurate, the transferable xn potentials provide the greatest accuracy for modeling these additional 4 liquid mixtures. Thus, among the calculated potentials, the extended ensemble potentials appear to provide the best accuracy and transferability for heptane-toluene mixtures.

While the xn potentials can be applied for modeling the interactions in any heptane-toluene mixture, it remains necessary to predict the appropriate volume-dependent potential, $U_{\Gamma V}$, as a function of composition. This potential is determined by two parameters, $\psi_{\Gamma V1}$ and $\psi_{\Gamma V2}$, that correspond to average corrections for the pressure, $\Delta \mathbf{P}_\Gamma$, and (inverse) compressibility, $\Delta \kappa_{T;\Gamma}^{-1}$, respectively, of the CG model. As described in the methods section, we developed a simple empirical relation for estimating these parameters as a function of the heptane mole fraction, χ. Figure 2.6 compares this empirical relation with the system-specific parameters that were optimized for compatibility with the transferable xn potentials and which correspond to the dashed lines in Fig. 3.2b. The empirical relation provides a reasonably accurate fit for the average pressure corrections, but appears slightly less accurate for corrections to the compressibilities.

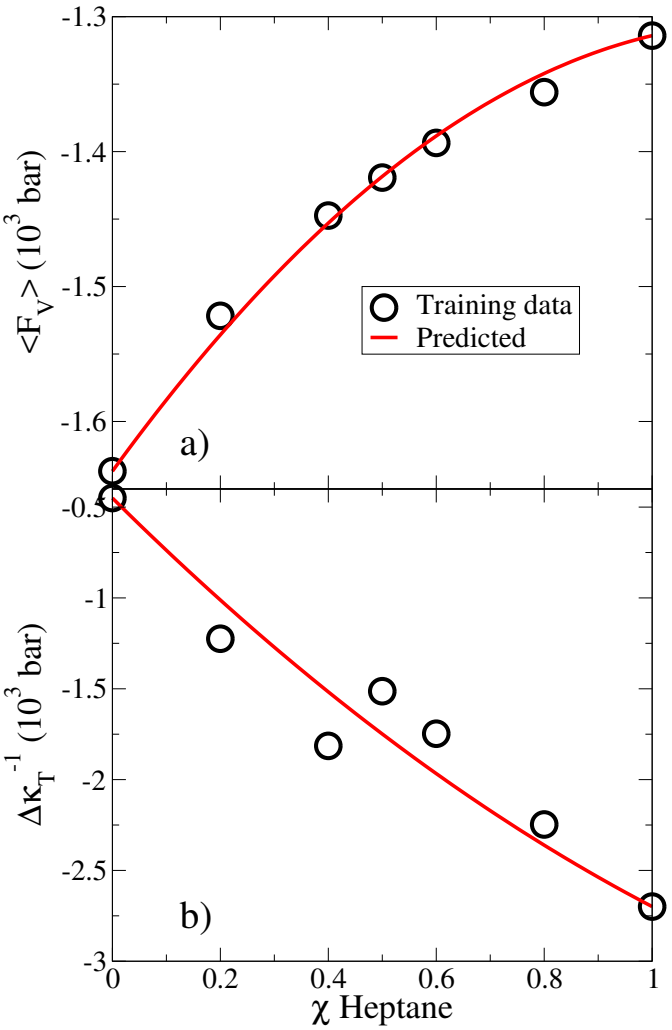

Figure 3.6. Empirical fits for the average pressure correction (a) and the correction to the inverse isothermal compressibility (b). The black circles indicate the calculated corrections to the transferable xn potential for the various systems included in the extended ensemble. The red curves present empirical fits to this data that employ one free parameter for each correction.

We employed this empirical relation to predict pressure corrections for 1:9 and 9:1 heptane-toluene mixtures. We then performed constant NPT simulations of these systems, while using the tranferable xn interaction potentials and the predicted pressure corrections. Figure 3.7 presents the simulated density distributions and pressure equations of state. The solid and dotted lines indicate simulated results for the AA and CG models. Importantly, these AA models were not used for parameterizing either the xn interaction potentials or the pressure corrections.

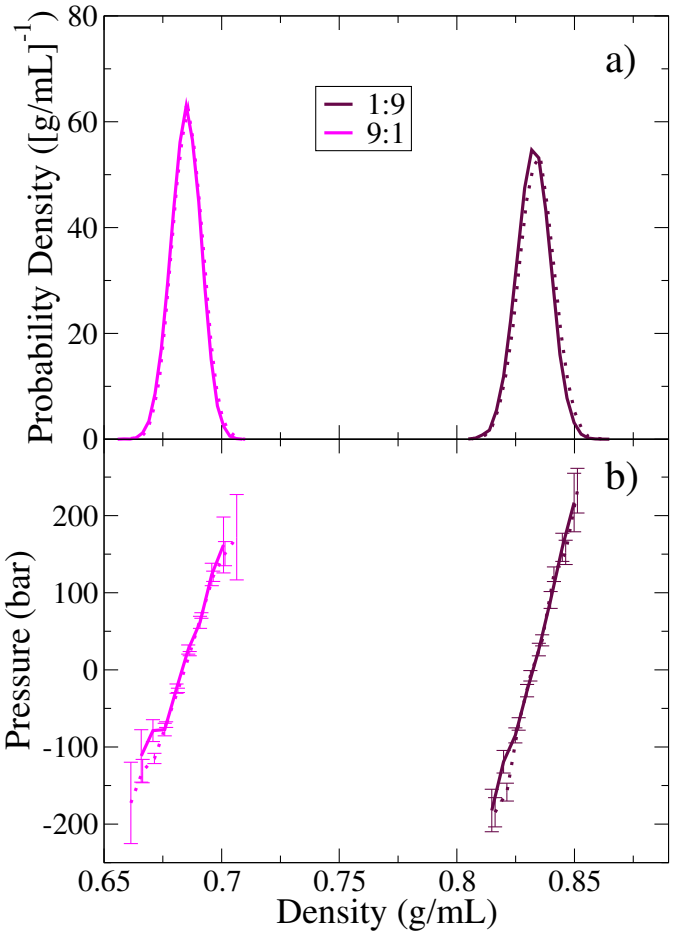

Figure 3.7. Simulated density distributions (a) and pressure equations of state (b) for 1:9 and 9:1 heptane:toluene mixtures. In both panels, solid curves indicate results for the OPLS-AA model, while dotted curves indicate results for CG models that employ the transferable xn interaction potentials with the pressure correction predicted from Fig. 3.6. The bars in panel b) indicate the standard error in the average simulated pressure over each volume increment.

Nevertheless, the CG models reproduce both the density distributions and the pressure equations of state with remarkable accuracy. Consequently, the extended ensemble approach appears capable of developing transferable models that demonstrate predictive accuracy for both structural and thermodynamic properties of heptane-toluene mixtures.

In our prior study, we considered the impact of resolution upon the thermodynamic properties of 1-, 2-, and 3-site MS-CG models for pure heptane and pure toluene. This study demonstrated that, with decreasing resolution, the

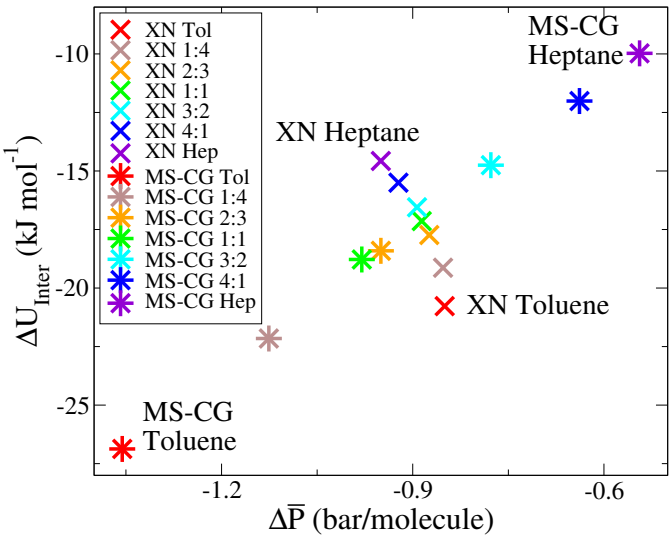

Figure 3.8. Scatter plot correlating the missing cohesive energy ΔU_{Inter} with the average pressure correction $\Delta\mathbf{P}$ required for various CG models. The stars and crosses indicate results for system-specific MS-CG potentials and for transferable xn potentials, respectively. The various systems are indicated by the color scheme of Fig. 3.2.

MS-CG models demonstrated decreasing intermolecular cohesion, i.e., $\Delta U_{\text{Inter}} = \langle u_{\text{Inter}} \rangle_{AA} - \langle U_{\text{Inter}} \rangle_{CG}$ systematically decreased with coarsening. Simultaneously, we observed a significant correlation between the missing cohesion and the required pressure corrections for the MS-CG models. The following calculations further investigate these effects for heptane-toluene mixtures.

Figure 3.8 analyzes the correlation between the missing cohesive energy ΔU_{Inter} (per molecule) and the average pressure correction, $\Delta\mathbf{P}$, for the different CG models. The system-specific MS-CG models for heptane-toluene mixtures, which are indicated by stars, demonstrate very similar correlations to the MS-CG models with varying resolution. The toluene MS-CG model demonstrates the greatest loss in cohesion and requires the largest pressure correction. Conversely, the heptane MS-CG model demonstrates the smallest loss in cohesion and requires the smallest pressure correction. The MS-CG models for heptane-toluene mixtures interpolate nearly linearly between the two pure liquids. In contrast, simulations with the transferable xn force field, which are indicated by crosses, demonstrate very different trends. In particular, the cohesive energy loss and the required pressure corrections demonstrate much smaller variation in simulations with the xn force field. Moreover, in striking contrast to the system-specific MS-CG models, the composition variation

in the pressure correction is actually reversed for the xn models, i.e., the xn force field requires the greatest pressure correction for pure heptane and the smallest pressure correction for pure toluene.

In comparison to the CG models, the AA models for heptane-toluene mixtures demonstrate very similar cohesive density. Consequently, we examined the CG potentials and forces in order to interpret the correlations in Fig. 3.8 and, in particular, the opposite trends observed for the MS-CG and xn force fields. For each system, Γ, and each potential, U_Γ, we defined the average pair potential, $\bar{U}_\Gamma(r)$, as a function of distance, r, by averaging over the relevant pair interactions ζ:

$$\bar{U}_\Gamma(r) = \sum_{\zeta \in \Gamma} N_{\Gamma;\zeta}(r) U_{\Gamma;\zeta}(r) / N_\Gamma(r) \tag{3.55}$$

where $N_{\Gamma;\zeta}(r) = N_{\Gamma;\zeta}\, g_{\Gamma;\zeta}(r)$ is the average number of ζ-type pairs separated by a distance r in corresponding AA simulations and $N_\Gamma(r) = \sum_{\zeta \in \Gamma} N_{\Gamma;\zeta}(r)$. We defined the average pair force from a corresponding average over the CG pair forces.

Figures 3.9b and 3.9d present the average pair potentials and pair forces for the xn models. Since the xn models employ the same interaction potentials for each mixture, the variation in Figs. 3.9b and 3.9d reflects the differences in site composition and pair structure among the mixtures. Thus, in comparison to the average MS-CG potentials, the average xn potentials vary relatively little between different mixtures, which accounts for the relatively narrow variation in cohesion among the xn models in Fig. 3.8. The supplementary material[172] demonstrates that the most attractive xn potential corresponds to the CT-CT interaction between terminal heptane sites, while the least attractive xn potential corresponds to the CB-CB interaction between ring toluene sites. Consequently, Fig. 3.9b demonstrates that the average xn pair potentials become increasingly attractive with increasing heptane content, which accounts for the trend in cohesive energy that is observed in Fig. 3.8. Similarly, the supplementary material[172] demonstrates that the CT heptane site is the largest site in the xn force field, while the CB toluene site is the smallest. Accordingly, Fig. 3.9d demonstrates that the average xn pair force incorporates an increasingly large excluded volume with increasing heptane content. This trend in the excluded volume appears to account for the increasingly large pressure correction that is required for the xn models with increasing heptane content in Fig. 3.8, which is opposite to the trend observed for the MS-CG models.

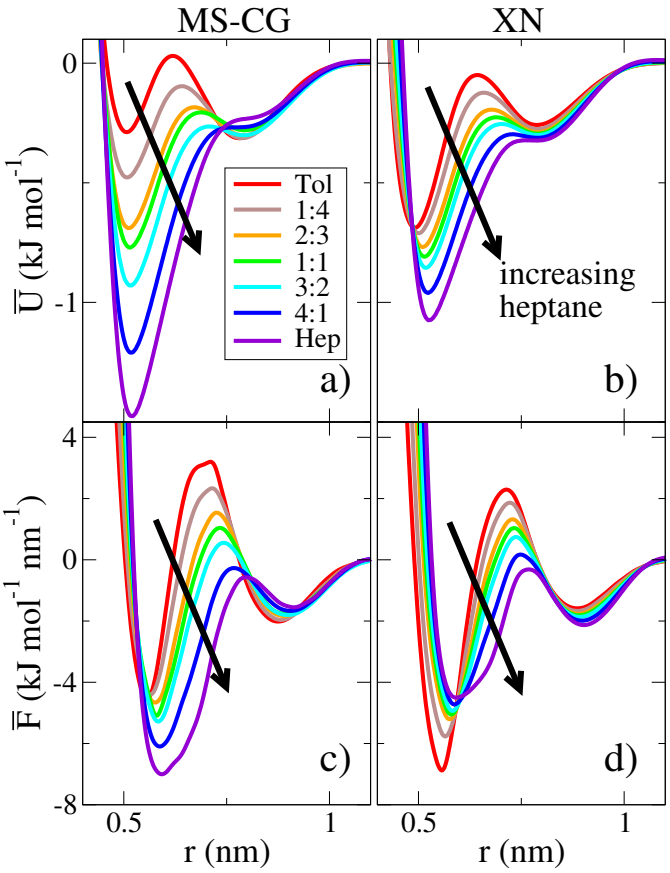

Figure 3.9. Site-averaged pair potentials (top row) and pair forces (bottom row). The left panels (a, c) correspond to system-specific MS-CG force fields, while the right panels (b, d) correspond to the transferable xn force field. The various systems are indicated by the color scheme of Fig. 3.2.

Figures 3.9a and 3.9c present the corresponding average pair potentials and pair forces for the system-specific MS-CG force fields. As expected from Fig. 3.2, the average MS-CG potentials become increasingly attractive with increasing heptane content. Additionally, the average MS-CG potentials demonstrate much greater variation than the average xn pair potentials. These two trends account for the relatively large gain in cohesion observed in Fig. 3.8 for the MS-CG models with increasing heptane content. In comparison to the average xn forces, the average MS-CG forces demonstrate relatively little variation in their excluded volume. However, the average MS-CG pair forces demonstrate increasingly large repulsion at $r \approx 0.7$ nm with increasing toluene content, which generates increasingly large "desolvation barriers" and accounts for the significant differences in the minima of

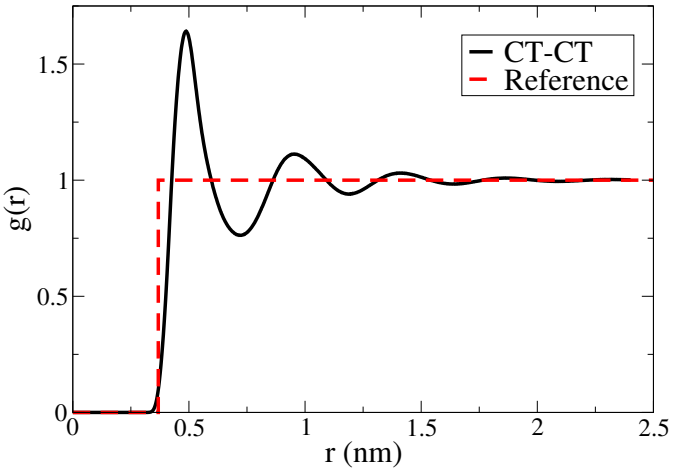

Figure 3.10. Example structure-less rdf (dashed red) used as a reference for quantifying the structure in the CT-CT rdf (black) that is obtained from AA simulations of pure heptane.

the various MS-CG potentials. In turn, this likely accounts for the increasingly large pressure corrections that are observed for the MS-CG models in Fig. 3.8.

Figure 3.9 relates trends in the calculated potentials to the missing cohesion and required pressure corrections for the various CG models. However, these calculations do not provide a microscopic interpretation for the variation in the CG potentials. We hypothesized that the variation in the calculated potentials reflects variation in the "structure" observed in the AA rdfs. In particular, simple Boltzmann inversion suggests that AA rdfs with particularly sharp peaks will result in CG potentials with correspondingly deep potentials.[32] In order to assess this hypothesis, we defined a "structureless" reference rdf for each mapped AA rdf. As illustrated in Fig. 3.10, this reference rdf is a step function with the same integral as the corresponding mapped AA rdf. We then quantify the "structure" in a particular mapped rdf by computing its MAE with respect to the corresponding structureless reference rdf.

Figure 3.11 correlates this structural metric with the minima of the calculated CG potentials. Panels a), b), and c) present results for heptane-heptane, toluene-toluene, and heptane-toluene pair types, respectively. Symbols indicate different pair types, while colors indicate system compositions. The xn structure is characterized by the mean of the MAE for the 5 mixtures included in the extended ensemble. As expected, for each pair type, the pair structure correlates with the well depth of the

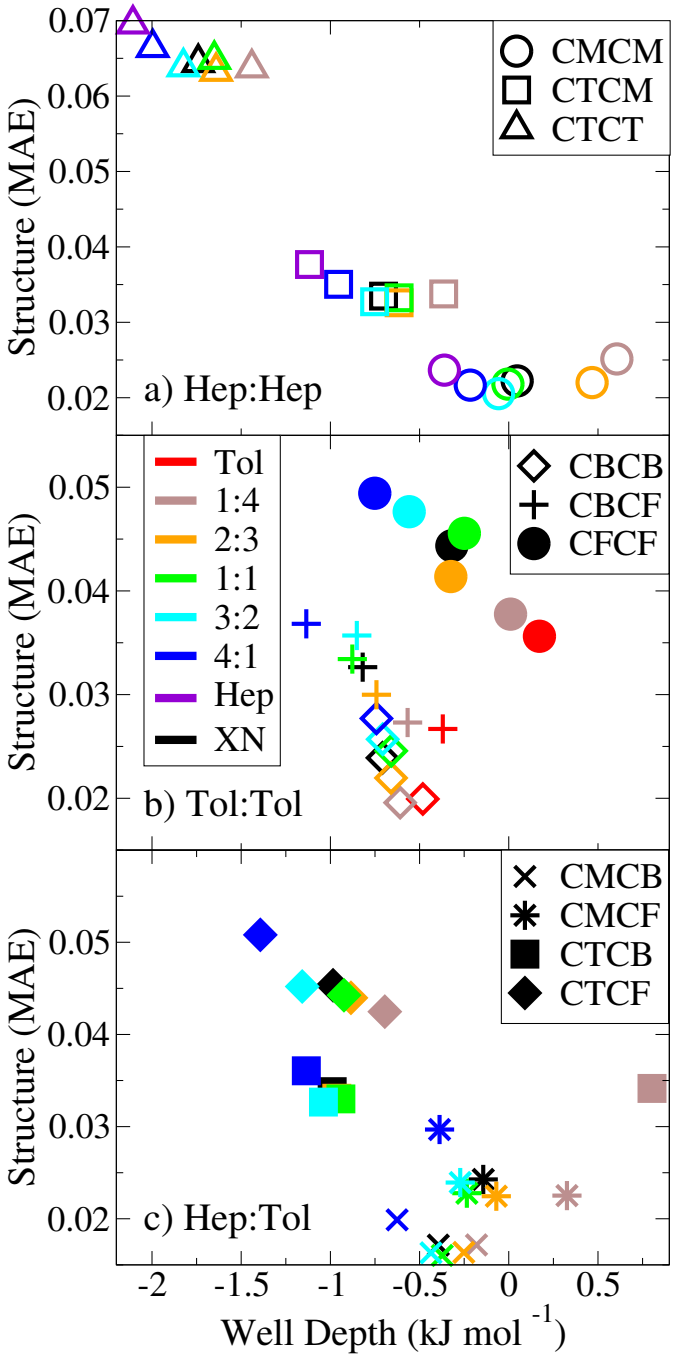

Figure 3.11. Scatter plot correlating the "structure" in AA site-site rdfs, as estimated by the MAE relative to a corresponding structureless rdf, with the well depth of the corresponding calculated system-specific MS-CG pair potential. Panels a, b, and c present results for heptane:heptane, toluene:toluene, and heptane:toluene site pairs. The symbol shapes indicate different pair types. The various systems are indicated by the color scheme of Fig. 3.2. The black symbols indicate results for the transferable xn potentials.

corresponding calculated CG potential. In particular, heptane rich systems generally demonstrate greater structure and deeper potentials. Additionally, the CT-CT pair corresponds to both the most structured rdfs and the deepest potential minima. For each pair type, the xn rdfs demonstrate intermediate structure among the various mixtures, while the xn potentials demonstrate corresponding intermediate minima. Thus, while Fig. 3.9 correlates the observed trends in cohesion and pressure corrections to trends in the calculated pair potentials, Fig. 3.11 relates these trends to the structure of the mapped AA ensembles.

Finally, we employ the current models to further explore the impact of coarse-graining for "simplifying" model properties. Our prior study investigated the impact of coarsening upon R-simplicity, which considers the Pearson correlation (R) between potential and virial fluctuations.[151] In particular, R-simple liquids with $R > 0.9$ reflect an approximate scale invariance in their potential energy surface that results in strikingly simple dynamic and thermodynamic properties,[124,132] which might prove useful for calibrating CG models and for predicting their transferability. However, while our prior study considered the constant NPT ensemble, R-simple behavior is more appropriately related to correlations in the constant NVT ensemble.[177] Accordingly, in the following we compare potential-virial fluctations in both ensembles.

Figure 3.12 presents scatter plots that correlate the intermolecular potential (per molecule) with the Pearson potential-virial correlation. The top row compares AA models with 1-, 2-, and 3-site MS-CG models for heptane and toluene. As discussed in our prior study, coarsening leads to increasingly large potential-virial correlations, although this correlation is somewhat diminished in the constant NVT ensemble. The remaining rows consider the effects of composition upon the R-simplicity of 3-site CG models. These high resolution CG models demonstrate only weak potential-virial correlations. (Note the difference in the scale of the x-axis.) The middle row demonstrates that, with increasing heptane content, the transferable xn potentials generate increasing cohesion, while the potential-virial correlations systematically decrease. In fact, this is opposite to the correlation observed with varying resolution. The bottom row demonstrates no significant trend in the system-specific MS-CG models. Thus, while coarsening generates increasingly simple models as the interactions become increasingly repulsive, the same correlation is not observed between different models at the same resolution.

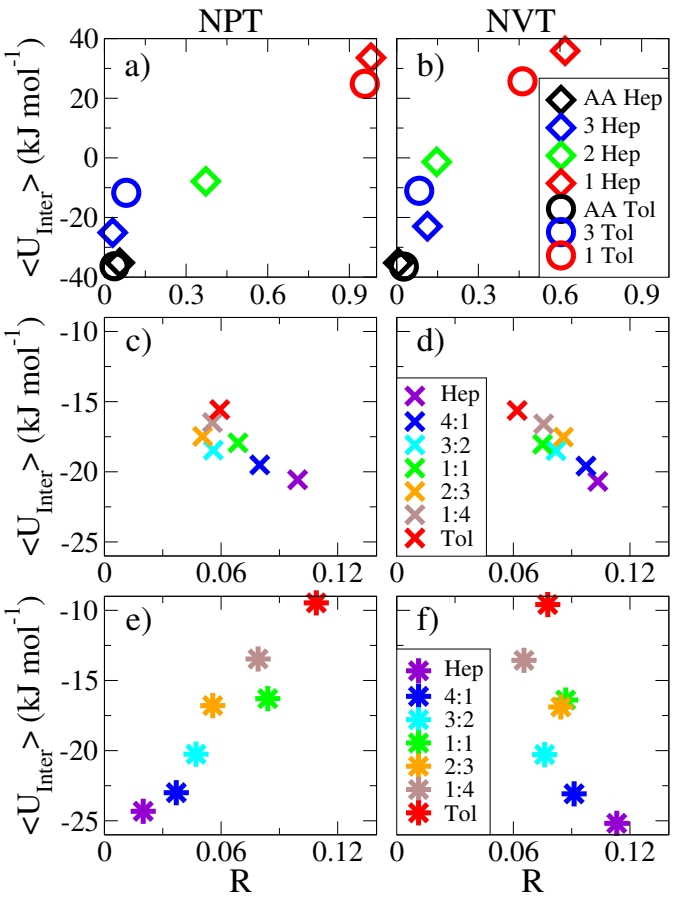

Figure 3.12. Scatter plot correlating the average cohesive energy density, $\langle U_{Inter} \rangle$ (per molecule), with the Pearson correlation coefficient R between the potential and virial. The left and right columns correspond to calculations of this correlation in the constant NPT and constant NVT ensembles, respectively. The top row present results for OPLS-AA models of pure heptane and pure toluene, as well as for corresponding MS-CG models at varying resolution. The bottom two rows present results for the different heptane:toluene mixtures considered in this work, while employing the color scheme of Fig. 3.2. The middle row presents results for the transferable xn force field, while the bottom row presents results for the system-specific MS-CG force fields.

3.6 Discussion

The present work demonstrates the promise of bottom-up approaches for developing transferable CG potentials that accurately model both structural and thermodynamic properties of liquid mixtures. We defined an extended ensemble of 5 heptane-toluene mixtures for parameterizing these potentials. We employed the extended ensemble MS-CG variational principle to determine system-independent,

i.e., transferable, interaction potentials that accurately reproduced the AA site-site rdfs for these 5 mixtures. We employed a self-consistent pressure-matching approach to determine system-specific pressure corrections that accurately reproduced the AA density, compressibility, and pressure equation of state for each of these mixtures. More importantly, the transferable interaction potentials quite accurately reproduced the AA site-site rdfs for an additional 4 heptane-toluene mixtures that were not included in the parameterization. Furthermore, the extended ensemble approach predicted system-specific pressure corrections that very accurately reproduced the pressure equations of state for the additional mixtures. Thus, this bottom-up approach provides predictive accuracy and transferability for both structural and thermodynamic properties of liquid heptane-toluene mixtures.

Intuitively, one expects that the extended ensemble approach will optimize transferability at the cost of reduced accuracy for specific systems. Indeed, we previously demonstrated that AA site-site rdfs for methanol-neopentane mixtures were more accurately modeled by the corresponding system-specific MS-CG potentials than by extended ensemble potentials that were optimized for transferability across a range of mixtures.[39] Consequently, it is somewhat surprising that the transferable extended ensemble potentials appear more accurate than system-specific MS-CG potentials for reproducing the AA site-site rdfs of heptane-toluene mixtures. This surprising accuracy may reflect the increased statistical sampling obtained by averaging over configurations for multiple systems.[175] More significantly, though, we propose that, by averaging over system-specific features, the extended ensemble approach may address a fundamental limitation of the MS-CG method.

In contrast to many other structure-based bottom-up approaches,[26,28–30] the MS-CG approach[27,88] does not iteratively refine the CG potentials in order to accurately reproduce AA rdfs. Instead, the MS-CG approach employs a generalized-Yvon-Born-Green equation[119] as a force balance relation for decomposing AA pair potentials of mean force (or, more precisely, AA pair mean forces) into correlated contributions from the various terms in the CG potential.[117] However, the MS-CG approach employs the many-body cross-correlations in the mapped AA ensemble in order to estimate the cross-correlations that will arise in the simulated CG ensemble. If the MS-CG model reasonably reproduces these cross-correlations, then the MS-CG model will also reproduce the AA pair potentials of mean force and, thus, the AA rdfs. In some cases, though, the simple molecular mechanics form of

the approximate potential precludes the MS-CG model from reproducing the cross-correlations that are observed in the mapped AA ensemble. For instance, molecular mechanics potentials cannot reproduce the bond-angle correlations that arise when mapping hexane from an AA representation to a 3-site CG representation.[122,129] Consequently, 3-site MS-CG models provide a relatively poor description of the hexane angle distribution.[120] Similarly, by manually smoothing over complex many-body cross-correlations for a disordered peptide, we determined MS-CG models that more accurately reproduced the corresponding mapped AA structure ensemble.[118] Accordingly, we propose that extended ensemble averages over multiple systems may actually improve the accuracy of MS-CG models by smoothing over system-specific structural correlations that cannot be accurately reproduced with simple CG potentials. We speculate that this effect was not observed in our prior study of methanol-neopentane mixtures[39] because the system-specific MS-CG models adequately reproduced the structural cross-correlations of these relatively simpler molecules.

The present calculations also contribute additional insight into the thermodynamic properties of CG models. In our previous study of pure heptane and pure toluene, we demonstrated that MS-CG models systematically lose intermolecular cohesion with decreasing resolution.[151] Accordingly, increasingly coarse MS-CG models require increasingly large pressure corrections in order to account for this missing cohesion. The present study demonstrates even stronger correlations between the missing cohesion and the pressure corrections that are required for system-specific MS-CG models of heptane-toluene mixtures. Moreover, the cohesion in these MS-CG models correlates with the "structure" in the mapped AA ensembles. This correlation between structure and cohesion provides a further rationalization for the observed correlations with varying resolution. One expects that increasingly coarse representations will increasingly efface atomic structural details from the mapped ensemble. These increasingly featureless mapped ensembles will determine CG potentials with decreasing cohesion that, in turn, require increasingly large pressure corrections. These correlations for approximate potentials are likely practical manifestations of the more general result that coarsening systematically transfers thermodynamic entropy and information from the configuration space into the many-body PMF.[81,82]

However, the correlations between cohesion and pressure corrections can be

weakened, or even reversed, if the same interaction potentials are applied for modeling different systems. Indeed, CG models that employ the transferable extended ensemble potentials demonstrate the opposite correlation between cohesion and pressure corrections. In particular, the heptane terminal CT site is the largest site, while the CT-CT interaction is the most attractive in the transferable force field. Consequently, CG models that employ these potentials gain cohesion, but also require larger pressure corrections with increasing heptane concentration.

Finally, the present work also builds upon several other results from our prior studies of coarse-graining in the constant NPT ensemble. In particular, the Appendix demonstrates that the self-consistent pressure-matching method corresponds to minimizing a relative entropy with respect to the volume dependence of the approximate CG potential. Additionally, this work provides further evidence that coarse-graining can "simplify" model properties, although three-site CG models demonstrate only slightly greater simplicity than atomically detailed models.

3.7 Conclusion

In closing, we emphasize that this work demonstrates the promise of the extended ensemble approach for developing highly efficient models that provide predictive accuracy and transferability for describing both structural and thermodynamic properties. By combining the MS-CG variational principle with the self-consistent pressure-matching approach, we accurately reproduced the AA site-site rdfs, density fluctuations, and pressure equations of state for five distinct heptane-toluene mixtures within the parameterization extended ensemble. More importantly, the resulting potentials provided similar accuracy for additional mixtures that were not treated in the parameterization. Quite surprisingly, the extended ensemble potentials appear both more transferable and also more accurate than MS-CG potentials that were optimized for individual systems. We speculate that this surprising accuracy results from the extended ensemble averaging, which effaces system-specific many-body correlations that cannot be accurately modeled by the CG potential and, thus, generate inaccuracies in the system-specific MS-CG potentials. Furthermore, we also demonstrated that the required pressure corrections correlate with the intermolecular cohesion of system-specific potentials. This intermolecular cohesion, in turn, correlates with the relative structure within

the corresponding mapped AA ensemble. However, these correlations do not persist when the same transferable interaction potentials are applied to model multiple systems. Finally, the appendix connects the present approach to approaches based upon minimizing a relative entropy.

This work suggests several directions for future study. For instance, the extended ensemble approach may improve the transferability of CG potentials for treating multiple thermodynamic states, e.g., multiple temperatures and pressures. In particular, preliminary studies suggest that this approach may enable transferable CG models for ionomers at various temperatures and various chemical compositions.[178] Additionally, the present approach may also prove useful for addressing other thermodynamic properties, e.g., the surface tension or chemical potential. Finally, we anticipate that the present CG models may prove useful for investigating self-assembly in hydrocarbon solvents.

Appendix: Connection to Relative Entropy

In this work and in our prior study,[151] we have employed a variational pressure-matching approach[44] to parameterize CG models that accurately model AA density fluctuations at constant pressure. The pressure-matching variational principle essentially parallels the MS-CG force-matching variational principle.[25,27,88] Both variational calculations approximate a potential of mean force for a mechanical variable via a least squares fit to fluctuating atomic forces acting on the variable. In the case of force-matching, the mechanical variables are the site coordinates and the fluctuating forces are simply the net atomic forces on the sites. In the case of pressure-matching, the mechanical variable is the volume and the fluctuating force is the instantaneous internal pressure. Significantly, given a simple form for the CG potential, the force- and pressure-matching variational principles do not guarantee that the CG model will reproduce particular observables of the AA model.

In this appendix, we briefly consider the pressure-matching variational principle in the context of a relative entropy,[29] which corresponds to the information theoretic Kullback-Leibler divergence.[179] Previous studies have demonstrated that the relative entropy is minimized with respect to a particular interaction potential, U_ζ, when the CG model reproduces the AA distribution function for the conjugate density variable.[29,81] For instance, the relative entropy is minimized with respect to a pair potential when the CG model reproduces the corresponding AA pair distribution, i.e., the radial distribution function. Below we develop three results that extend these considerations for the isothermal-isobaric ensemble: (1) First, we demonstrate that the many-body PMF minimizes the relevant relative entropy for the isothermal-isobaric ensemble; (2) Secondly, we demonstrate that this relative entropy is minimized with respect to a volume-dependent CG potential, $U_V(V)$, when the CG model matches the equilibrium volume distribution sampled by the AA model. (3) Finally, we demonstrate that this is equivalent to the self-consistent pressure-matching criteria. Consequently, the self-consistent pressure-matching method provides a numerical scheme for minimizing the relative entropy with respect to U_V. We expect that similar analysis follows for other ensembles in which mechanical variables fluctuate subject to constant conjugate intensive parameters.

The extended ensemble analysis appears to follow similarly for both force-matching and relative entropy-based variational principles. In both cases the

relevant correlation functions are generalized to include appropriate averages over γ. Thus, we conduct our analysis for a single system.

1. For an expanded state space with fluctuations in both configuration and volume, we define the relative entropy:

$$S_{\text{rel}} = \int \mathrm{d}V \int_{V^N} \mathrm{d}\mathbf{R}\, p_{RV}(\mathbf{R}, V) \ln \left[p_{RV}(\mathbf{R}, V) \Big/ P_{RV}(\mathbf{R}, V) \right] \tag{3.56}$$

   where $p_{RV}(\mathbf{R}, V)$ is the probability for the AA model to sample a configuration $\mathbf{r}$ that maps to $\mathbf{R}$ at the given volume V. Similarly, $P_{RV}(\mathbf{R}, V)$ is the probability for the CG model to sample the configuration $\mathbf{R}$ and the volume V. By the Gibbs inequality,[32,180] $S_{\text{rel}} \geq 0$ and only vanishes when $p_{RV}(\mathbf{R}, V) = P_{RV}(\mathbf{R}, V)$ for all $\mathbf{R}$ and V, i.e., when the CG model is consistent with the AA model in the isothermal-isobaric ensemble. Thus, minimizing the relative entropy provides a second variational principle for optimizing CG models at constant T and P_0. In particular, the relative entropy is minimized at constant temperature T and external pressure P_0, when the approximate CG potential U equals the many-body PMF W to within a constant that is independent of both $\mathbf{R}$ and V.

2. We next assume the approximate potential adopts the form $U(\mathbf{R}, V) = U_R(\mathbf{R}) + U_V(V)$, where U_R depends upon V only via periodic boundary conditions. Then

$$\frac{\delta S_{\text{rel}}}{\delta U_V(V)} = \beta \left[p_v(V) - P_V(V|U) \right] \tag{3.57}$$

   where p_v is the equilibrium volume distribution for the AA model at the given T and P_0. Similarly, $P_V(V|U)$ is the corresponding volume distribution for the CG model with a potential U. Thus, S_{rel} is minimized with respect to U_V when the CG model reproduces the AA volume distribution:

$$P_V(V|U) = p_v(V). \tag{3.58}$$

3. We define

$$\exp\left[-\beta a(V)\right] \equiv v_0^{-n} \int_{V^n} \mathrm{d}\mathbf{r}\, \exp\left[-\beta u(\mathbf{r}, V)\right] \tag{3.59}$$

$$\exp\left[-\beta A(V|U)\right] \equiv v_0^{-N} \int_{V^N} \mathrm{d}\mathbf{R}\, \exp\left[-\beta U(\mathbf{R}, V)\right], \tag{3.60}$$

as the volume-dependent contributions to the Helmholtz potentials for the AA and CG models. If we consider V as a mechanical variable, then $a(V) + P_0V$ and $A(V|U) + P_0V$ act as corresponding potentials of mean force for V in the AA and CG models, respectively, i.e.,

$$p_v(V) \propto \exp\left[-\beta\left(a(V) + P_0V\right)\right] \tag{3.61}$$

$$P_V(V|U) \propto \exp\left[-\beta\left(A(V|U) + P_0V\right)\right] \tag{3.62}$$

Thus, Eq. (3.58) is equivalent to the criterion that, at the given T and P_0, the CG model should reproduce the volume dependence of the AA Helmholtz potential:

$$A(V|U) = a(V) + \text{const}, \tag{3.63}$$

where const is independent of V. This is equivalent to the self-consistent criterion for iterative pressure-matching:

$$\mathbf{P}_{CG}(V|U) = \mathbf{P}_{AA}(V), \tag{3.64}$$

i.e., iterative pressure-matching provides a numerical procedure for minimizing S_{rel} with respect to U_V. (Note that the original pressure-algorithm does not lead to this criterion, since χ_2^2 is evaluated over the mapped AA ensemble and not the simulated CG ensemble.)

Chapter 4
van der Waals perspective on coarse-graining: Progress towards solving representability and transferability problems

N. J. H. Dunn, T. F. Foley, W. G. Noid, *Acc. Chem. Res.* **2016**, 49 (12), 2832-2840

4.1 Abstract

Low-resolution coarse-grained (CG) models provide the necessary efficiency for simulating phenomena that are inaccessible to more detailed models. However, in order to realize their considerable promise, CG models must accurately describe the relevant physical forces and provide useful predictions. By formally integrating out the unnecessary details from an all-atom (AA) model, "bottom-up" approaches can, at least in principle, quantitatively reproduce the structural and thermodynamic properties of the AA model that are observable at the CG resolution. In practice, though, bottom-up approaches only approximate this "exact coarse-graining" procedure. The resulting models typically reproduce the intermolecular structure of AA models at a single thermodynamic state point, but often describe other state points less accurately and, moreover, tend to provide a poor description of thermodynamic properties. These two limitations have been coined the "transferability" and "representability" problems, respectively. Perhaps, the simplest and most commonly

discussed manifestation of the representability problem regards the tendency of structure-based CG models to dramatically over-estimate the pressure. Furthermore, when these models are adjusted to reproduce the pressure, they provide a poor description of the compressibility. More generally, it is sometimes suggested that CG models are fundamentally incapable of reproducing both structural and thermodynamic properties. After all, there is no such thing as a "free lunch" - any significant gain in computational efficiency should come at the cost of significant model limitations.

At least in the case of structural and thermodynamic properties, though, we optimistically propose that this may be a false dichotomy. Accordingly, we have recently re-examined the "exact coarse-graining" procedure and investigated the intrinsic consequences of representing an AA model in reduced resolution. These studies clarify the origin and inter-relationship of representability and transferability problems. Both arise as consequences of transferring thermodynamic information from the high resolution configuration space and encoding this information into the many-body potential of mean force (PMF), i.e., the potential that emerges from an exact coarse-graining procedure. At least in principle, both representability and transferability problems can be resolved by properly addressing this thermodynamic information. In particular, we have demonstrated that "pressure-matching" provides a practical and rigorous means for addressing the density-dependence of the PMF. The resulting bottom-up models accurately reproduce the structure, equilibrium density, compressibility, and pressure equation of state for AA models of molecular liquids. Additionally, we have extended this approach to develop transferable potentials that provide similar accuracy for heptane-toluene mixtures. Moreover, these potentials provide predictive accuracy for modeling concentrations that were not considered in their parameterization. More generally, this work suggests a "van der Waals" perspective on coarse-graining, in which conventional structure-based methods accurately describe the configuration-dependence of the PMF, while independent variational principles infer the thermodynamic information that is necessary to resolve representability and transferability problems.

4.2 Introduction

Low resolution coarse-grained (CG) models play an important and rapidly growing role in science.[8,74] By eliminating unnecessary details, CG models provide the necessary efficiency for simulating length- and time-scales that remain inaccessible to more detailed models.[7] CG models also empower more systematic investigations of the relevant experimental conditions, while simultaneously providing superior statistical precision in simulated quantities. Furthermore, CG models more effectively harness the intellectual horsepower of researchers by focusing attention on the essential details of a particular phenomenon, which atomically detailed models can easily obscure.[15,16]

Historically, CG models have been extensively employed for investigating the emergent consequences of basic physical principles.[13] More recently, though, many coarse-graining approaches have been developed for modeling specific systems.[75] By formally integrating out unnecessary atomic details, "bottom-up" approaches can, at least in principle, quantitatively reproduce the structural and thermodynamic properties of a high resolution model that can be observed at the resolution of the CG model, although thermodynamic properties require careful consideration.[7,13] Of course, this "exact coarse-graining" procedure cannot be accomplished in practice. Rather, bottom-up models are often parameterized to accurately describe the structure of a high resolution model at a single thermodynamic state point.[32] Unfortunately, the resulting models often prove accurate over a relatively limited range of thermodynamic conditions and, moreover, tend to provide a surprisingly poor description of thermodynamic properties.[75] These difficulties are termed "transferability" and "representability" problems, respectively.

Transferability problems are not surprising. CG potentials are constructed to incorporate the effects of atoms that have been eliminated from the CG model. These effects will certainly vary with thermodynamic state point. Thus, one intuitively expects that CG potentials should depend upon thermodynamic conditions. Indeed, many previous studies have documented the sensitivity of bottom-up potentials to variations in state point.[35,37,38,40–43,152,153,181,182] Consequently, one expects that any approximate potential will accurately describe these effects only over a relatively limited range of thermodynamic conditions.

Representability problems are more subtle. While recent studies introduced the

term to describe thermodynamic inconsistencies in CG models,[33,91] representability problems are related to inconsistencies observed much earlier for effective potentials employed in liquid state theories.[183,184] Perhaps the simplest and most commonly discussed representability problem regards the pressure-volume behavior of structure-based CG models. Indeed, many studies have observed that structure-based CG models generate unrealistically high pressures.[79] For instance, under ambient conditions, structure-based CG models overestimate the internal pressure of liquid water by almost four orders of magnitude.[33,80] Moreover, when these models are modified to reproduce the pressure, they then provide a poor description of the isothermal compressibility.[34] Similarly, previous studies have also demonstrated that CG models poorly describe the energetic and entropic contributions to free energy differences.[185–187] It has been suggested that representability problems reflect fundamental limitations of state-point-dependent effective potentials and that, more simply, CG models cannot accurately describe multiple conflicting observables, such as the pressure and the compressibility.[33,92] Alternatively, it has been proposed that representability problems may be resolved by considering the effects of the missing atomic degrees of freedom upon the CG representation of thermodynamic observables.[75,188] In particular, Guenza and coworkers have demonstrated the importance of these effects for low resolution polymer models.[83,84,86]

This account summarizes our recent studies of representability and transferability challenges.[82,151,189] We have adopted the optimistic hypothesis that both challenges can be resolved by carefully considering exact coarse-graining and the intrinsic consequences of representing a system in reduced detail. Our analysis clarifies both the origin and inter-relation of representability and transferability problems. Moreover, our work demonstrates an extended ensemble pressure-matching approach[39,44] for determining transferable potentials that accurately model the structure, pressure, and compressibility of molecular liquids in practice.

4.3 Exact coarse-graining

4.3.1 Atomic model

We first consider the canonical ensemble for an all-atom (AA) model that represents a system with n atoms labelled $i = 1, \ldots, n$ in a volume, V, at a temperature,

T.[190] We indicate the atomic configuration, $\mathbf{r} = (\mathbf{r}_1, \ldots, \mathbf{r}_n)$. The atoms interact according to a conservative potential, $u(\mathbf{r}; V)$, that generates a force, $\mathbf{f}_i(\mathbf{r}; V)$, on each atom i, as well as a force on the volume, i.e., the fluctuating internal pressure:

$$p_{\text{int}}(\mathbf{r}; V, T) = nk_BT/V - (\partial u(\mathbf{r}; V)/\partial V)_{\hat{\mathbf{r}}}\,. \tag{4.1}$$

The canonical ensemble average of $p_{\text{int}}(\mathbf{r}; V, T)$ equals the thermodynamic internal pressure of the AA model, $p_{\text{int}}(V, T)$.

The first term in Eq. (4.1) describes the kinetic, i.e., ideal, contribution to the pressure. The second term defines the instantaneous excess (xs) pressure, $p_{\text{xs}}(\mathbf{r}; V) = -(\partial u(\mathbf{r}; V)/\partial V)_{\hat{\mathbf{r}}}$, while $\hat{\mathbf{r}} = \left(V^{-1/3}\mathbf{r}_1, \ldots, V^{-1/3}\mathbf{r}_n\right)$ denotes the "scaled configuration." This contribution is often calculated from the virial expression:

$$p_{\text{xs}}(\mathbf{r}; V) = \frac{1}{3V}\sum_i \mathbf{f}_i(\mathbf{r}; V)\cdot\mathbf{r}_i = \frac{1}{3V}\sum_{(i,j)} f_{2;ij}(r_{ij})r_{ij}, \tag{4.2}$$

where the second sum is performed over all intra- and inter-molecular pairs (i, j) that are separated by a distance r_{ij} and interact with a force of magnitude $f_{2;ij}(r_{ij})$. Both expressions for $p_{\text{xs}}(\mathbf{r}; V)$ assume that the atomic potential does not explicitly depend upon the volume, i.e., $(\partial u/\partial V)_{\mathbf{r}} = 0$. The second expression also assumes that the nonbonded potential is pair-additive. While angle-dependent bonded potentials do not contribute to the virial, more complex non-bonded interactions may introduce additional contributions.[151,190]

Finally, we consider the total differential describing variations in the atomic potential, i.e., work:

$$\mathrm{d}u(\mathbf{r}; V) = -\sum_i \mathbf{f}_i(\mathbf{r}; V)\cdot(\mathrm{d}\mathbf{r}_i)_V - p_{\text{xs}}(\mathbf{r}; V)\mathrm{d}V, \tag{4.3}$$

where $(\mathrm{d}\mathbf{r}_i)_V = V^{1/3}\mathrm{d}\hat{\mathbf{r}}_i$. The first term in Eq. (4.3) quantifies changes in potential energy due to configuration changes at constant volume. The second term quantifies changes in potential energy due to isotropic compression or expansion.

4.3.2 Coarse-grained model

We next consider the canonical ensemble (at the same V and T) for a CG model that describes the same system with N "sites" that are labelled $I = 1, \ldots, N$. We

indicate the CG configuration, $\mathbf{R} = (\mathbf{R}_1, \ldots, \mathbf{R}_N)$. The sites interact according to a conservative potential, $U(\mathbf{R}, V)$, that generates a force, $\mathbf{F}_I(\mathbf{R}, V)$, on each site I, as well as a fluctuating internal pressure:

$$P_{\text{int}}(\mathbf{R}, V; T; U) = Nk_BT/V - (\partial U(\mathbf{R}, V)/\partial V)_{\hat{\mathbf{R}}} \,. \tag{4.4}$$

The canonical ensemble average of $P_{\text{int}}(\mathbf{R}, V; T; U)$ equals the thermodynamic internal pressure of the CG model, $P_{\text{int}}(V, T; U)$.

As above, the first term in Eq. (4.4) describes the ideal contribution for the $N \leq n$ remaining CG particles. The second term defines the instantaneous excess pressure of the CG model, $P_{\text{xs}}(\mathbf{R}, V) = -(\partial U(\mathbf{R}, V)/\partial V)_{\hat{\mathbf{R}}}$, while $\hat{\mathbf{R}} = (V^{-1/3}\mathbf{R}_1, \ldots, V^{-1/3}\mathbf{R}_N)$ denotes the scaled CG configuration. Our objective is to parameterize U for accurately describing the structural and thermodynamic properties of the atomic model.

4.3.3 The many-body Potential of Mean Force

In order to relate the AA and CG models, we introduce a mapping, $\mathbf{M}$, that determines the CG configuration as a function of the AA configuration: $\mathbf{R} = \mathbf{M}(\mathbf{r})$.[25] For simplicity, we assume the mapping associates the CG sites with the mass centers of disjoint atomic groups. The central quantity in our analysis is the many-body potential of mean force (PMF), which is the effective potential that results from "exact coarse-graining" in the canonical ensemble:

$$\exp\left[-W(\mathbf{R}; V, T)/k_BT\right] = V^{-(n-N)} \int_V \mathrm{d}\mathbf{r}\, \exp[-u(\mathbf{r}; V)/k_BT]\, \delta(\mathbf{R} - \mathbf{M}(\mathbf{r})), \tag{4.5}$$

where the integral is performed over the volume-dependent configuration space.[20,22,77] The PMF is the appropriate potential for ensuring that the CG model samples configurations according to the probability implied by the atomistic model and the mapping at the given V and T.[25] Moreover, the PMF encodes all information and thermodynamic properties that are observable at the resolution of the CG model. In particular, the PMF ensures that the CG model reproduces the excess free energy of the AA model:

$$\frac{1}{V^N} \int_V \mathrm{d}\mathbf{R}\, \exp[-W(\mathbf{R}; V, T)/k_BT] = \frac{1}{V^n} \int_V \mathrm{d}\mathbf{r}\, \exp[-u(\mathbf{r}; V)/k_BT]. \tag{4.6}$$

4.3.4 Energetic and entropic contributions

The PMF is not a conventional potential, but rather a free energy that depends upon both the configuration and also the thermodynamic state.[13] In collaboration with the Shell group,[82] we have examined the thermodynamic character of the PMF and, in particular, derived its energetic, U_W, and entropic, S_W, components:

$$W(\mathbf{R};V,T) = U_W(\mathbf{R};V,T) - TS_W(\mathbf{R};V,T) \tag{4.7}$$

$$U_W(\mathbf{R};V,T) \equiv \langle u(\mathbf{r};V)\rangle_{\mathbf{R};V,T} \tag{4.8}$$

$$S_W(\mathbf{R};V,T) \equiv \left\langle -k_B \ln\left[\Omega_1 \overline{p}_{r|R}(\mathbf{r}|\mathbf{R};V,T)\right]\right\rangle_{\mathbf{R};V,T}, \tag{4.9}$$

where $\Omega_1 = V^{n-N}$ is the volume element of atomic configurations $\mathbf{r}$ that map to $\mathbf{R}$, $\overline{p}_{r|R}(\mathbf{r}|\mathbf{R};V,T)$ is the conditioned probability density for AA configurations $\mathbf{r}$ satisfying $\mathbf{M}(\mathbf{r}) = \mathbf{R}$, and the subscripted angular brackets indicate corresponding conditioned canonical averages.

The energetic contribution to the PMF, $U_W(\mathbf{R};V,T)$, is simply the conditioned average of the atomic potential for the atomic configurations that map to $\mathbf{R}$. This energetic contribution generates forces that bias the CG model to sample low-energy configurations. The entropic contribution, $S_W(\mathbf{R};V,T)$, quantifies the excess entropy that is stored in the Boltzmann distribution of atomic configurations that map to $\mathbf{R}$. Thus, S_W quantifies the information about the atomic distribution that is "lost" when viewing this distribution at the CG resolution. By the Gibbs inequality,[191] $-TS_W \geq 0$, and only vanishes when $\overline{p}_{r|R} = {\Omega_1}^{-1}$, i.e., when all atomic configurations that map to $\mathbf{R}$ have equal Boltzmann weight. Consequently, $-TS_W$ generates forces that bias the CG model to sample high-entropy configurations, i.e., CG configurations for which the underlying atomic Boltzmann distribution, $\overline{p}_{r|R}$, is more uniform.

This simple decomposition is fundamentally important for representing thermodynamic properties with CG models. Since W is a free energy and incorporates an entropic component, W cannot be directly employed to estimate atomic energies. Similarly, the configurational entropy of the atomic model cannot be directly estimated from the configuration distribution sampled by the CG model.[81] Nevertheless, at least in principle, U_W and S_W can be employed to quantify the excess energy and excess entropy, respectively, of the atomic model, albeit at the resolution

of the CG model.

In order to illustrate these considerations, we analytically derived the exact PMF for the Gaussian Network model (GNM) as a function of the CG resolution.[82] The GNM describes protein fluctuations away from an equilibrium structure with a system of linear springs between nearby residues.[192] For each of 7 proteins, we constructed a high resolution GNM that explicitly represented 120 α carbons of the protein. For each protein, we determined W for a series of N-site CG models in which we mapped each consecutive $120/N$ α carbons to their mass center.

Figure 4.1 illustrates the impact of resolution upon W, U_W, S_W, and s_R, i.e., the excess entropy present in the configuration space. In the absence of coarse-graining, i.e., $N = 120$, the PMF is simply the atomic potential, $W = u$, $S_W = 0$, and the excess entropy is stored in the atomic configurational distribution. Because U_W is simply a conditioned average of the atomic potential, its average magnitude does not vary with coarsening, as indicated by the dashed horizontal line. However, with successive coarsening, configurational entropy and, equivalently, information is eliminated from the atomistic configuration space. The excess entropy is transferred into S_W, which results in a systematic increase in W with coarsening. In the extreme limit of coarse-graining, $N \to 0$, $s_R \to 0$, and the PMF becomes the configuration-independent, excess Helmholtz potential of the atomic model. In this limit, U_W and S_W become the thermodynamic excess energy and excess entropy, respectively.

4.3.5 Variation in the PMF

Further insight into representability and transferability issues can be gleaned from the total differential of the PMF:

$$\mathrm{d}W(\mathbf{R};V,T) = -\sum_I \overline{\mathbf{f}}_I(\mathbf{R};V,T)\cdot(\mathrm{d}\mathbf{R}_I)_V - \overline{p}_{\mathrm{xs}}(\mathbf{R};V,T)\,\mathrm{d}V - S_W(\mathbf{R};V,T)\,\mathrm{d}T \tag{4.10}$$

$$\overline{\mathbf{f}}_I(\mathbf{R};V,T) \equiv \langle \mathbf{f}_I(\mathbf{r};V)\rangle_{\mathbf{R};V,T} \tag{4.11}$$

$$\overline{p}_{\mathrm{xs}}(\mathbf{R};V,T) \equiv \langle p_{\mathrm{xs}}(\mathbf{r};V)\rangle_{\mathbf{R};V,T} \tag{4.12}$$

where $\mathbf{f}_I(\mathbf{r};V)$ is the force on site I and $(\mathrm{d}\mathbf{R}_I)_V = V^{1/3}\mathrm{d}\hat{\mathbf{R}}_I$ indicates changes in the CG configuration at constant volume. While Eq. (4.3) describes variations in an energy, Eq. (4.10) describes variations in a free energy, including both energetic

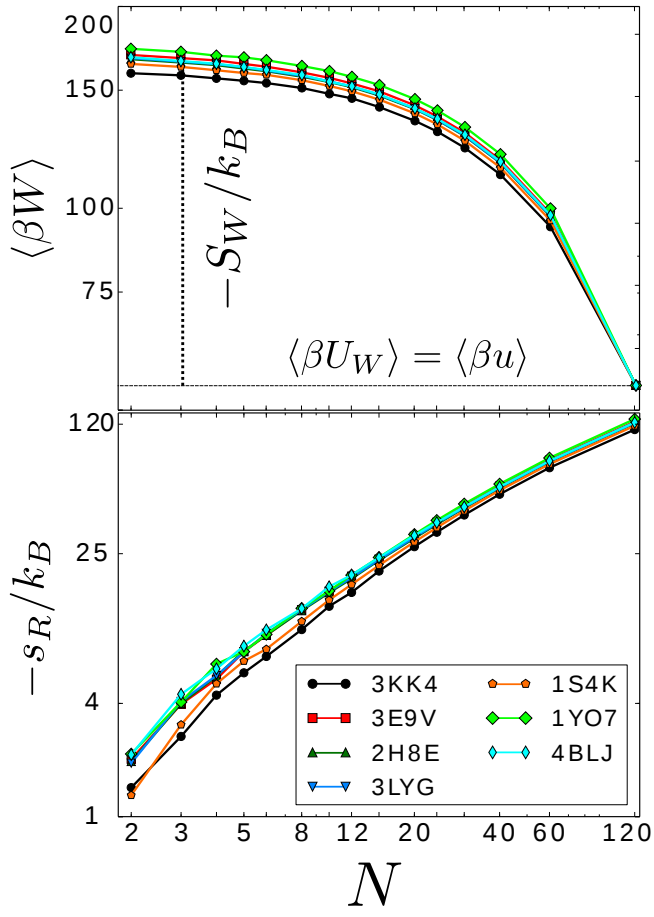

Figure 4.1. Analysis of the PMF (top) and apparent configurational entropy (bottom) as a function of the number, N, of sites considered. The top panel indicates the energetic (horizontal dotted line) and entropic (vertical dotted line) contributions to the average of the dimensionless PMF (solid line), $\langle \beta W \rangle$, for each protein domain. The bottom panel presents the absolute magnitude of the apparent configurational entropy for each protein domain when viewed at the given resolution. Both panels employ a log x - log y scale.

and entropic contributions.

A few points should be noted:

1. Most importantly, Eq. (4.10) equates the configuration-, volume-, and temperature-derivatives of the PMF to conditioned averages of the excess forces, excess pressure, and excess entropy of the atomic model. Consequently, the state-point dependence of the PMF, e.g., with respect to temperature or volume change, is determined by the contributions of the missing atomic degrees of freedom to the conjugate excess thermodynamic quantity, i.e., the excess entropy or excess pressure, respectively. This is the origin of both

transferability and representability problems.

2. The third entropic contribution to Eq. (4.10) is unique to the CG model. Since $S_W \leq 0$, increasing temperature will cause the PMF to increase at each $\mathbf{R}$. In particular, the PMF will vary more rapidly with temperature for CG configurations that correspond to highly structured atomic distributions.

3. Finally, these three contributions are all inter-related via Maxwell-type relations for mixed second derivatives of the PMF. For instance:

$$\left(\partial \bar{\mathbf{f}}_I(\mathbf{R};V,T)\Big/\,\partial T\right)_{\mathbf{R},V} = \left(\partial S_W(\mathbf{R};V,T)/\,\partial \mathbf{R}_I\right)_{T,V}, \tag{4.13}$$

which suggests that the temperature-transferability of CG force fields can be maximized by minimizing the configurational dependence of S_W.

4.3.6 Pressure and the constant NPT ensemble

According to Eq. (4.6), W accounts for the excess, but not the ideal, contributions to the Helmholtz potential from the atoms that have been eliminated from the CG model. Consequently, W does not reproduce the internal pressure and does not provide appropriate Boltzmann weight for sampling different volumes at constant external pressure, $\mathrm{P_{ext}}$. Accordingly, the PMF must be slightly modified in order to model the atomic pressure and the constant NPT ensemble:

$$W_P(\mathbf{R},V;T) = W(\mathbf{R};V,T) - (n-N)k_BT\ln\left(V/V_0\right), \tag{4.14}$$

where V_0 is an arbitrary reference volume. The second term in Eq. (4.14) accounts for the ideal contributions to the free energy from the missing atomic degrees of freedom. Although this term does not impact the configuration-distribution at a given V, it ensures that W_P provides the correct Boltzmann weight for each CG microstate, $(\mathbf{R},V)$:

$$\exp\left[-\beta\left(W_P(\mathbf{R},V;T)+\mathrm{P_{ext}}V\right)\right] = V_0^{-(n-N)}\int_V \mathrm{d}\mathbf{r}\,\exp[-\beta\left(u(\mathbf{r};V)+\mathrm{P_{ext}}V\right)]\,\delta(\mathbf{R}-\mathbf{M}(\mathbf{r})), \tag{4.15}$$

where $\beta = 1/k_B T$.[44,151] Because

$$-\left(\partial W_P(\mathbf{R}, V; T)/\partial V\right)_{\hat{\mathbf{R}};T} = -\left(\partial W(\mathbf{R}; V, T)/\partial V\right)_{\hat{\mathbf{R}},T} + (n-N)k_B T/V, \quad (4.16)$$

according to Eq. (4.4), W_P is the appropriate potential for reproducing the average pressure of the atomic model in each CG microstate,

$$P_{\text{int}}(\mathbf{R}, V; T; W_P) = \langle p_{\text{int}}(\mathbf{r}; V, T)\rangle_{\mathbf{R};V,T} \ , \quad (4.17)$$

and each thermodynamic equilibrium state:

$$P_{\text{int}}(V, T; W_P) = p_{\text{int}}(V, T). \quad (4.18)$$

4.4 Approximate coarse-graining

The preceding analysis not only clarifies their common origin, but also suggests practical computational methods for resolving representability and transferability challenges.

4.4.1 Pressure-matching

In practice, CG models commonly employ a relatively simple effective potential that is independent of both temperature and volume, i.e., $U = U_R(\mathbf{R})$.[32,75] The excess pressure of the CG model is then

$$\begin{aligned} P^0_{\text{xs}}(\mathbf{R}; V) &\equiv -\left(\partial U_R(\mathbf{R})/\partial V\right)_{\hat{\mathbf{R}}} & (4.19)\\ &= \frac{1}{3V}\sum_I \mathbf{F}_I(\mathbf{R})\cdot\mathbf{R}_I = \frac{1}{3V}\sum_{(I,J)} F_{2;IJ}(R_{IJ})R_{IJ}, & (4.20)\end{aligned}$$

where the last expression sums over each pair (I, J), assuming that each nonbonded interaction is modeled with a pair force function $F_{2;IJ}$ as a function of the pair distance, R_{IJ}.

Leading bottom-up methods parameterize U_R to reproduce atomic structural distributions, such as radial distribution functions (RDFs), at a single thermodynamic state point.[32] While this structure-based approach addresses the configuration-dependence of the PMF, it provides little or no insight into the volume- and

temperature-dependence of the PMF. Consequently, there is no reason to expect that the resulting model will accurately describe thermodynamic properties or provide an accurate description at other state points.

In particular, in order for the CG model to accurately describe the pressure of the atomic model, it is necessary that $P^0_{\text{xs}}(\mathbf{R};V)$, given by Eq. (4.20), accurately approximates $-(\partial W_P(\mathbf{R},V;T)/\partial V)_{\hat{\mathbf{R}};T}$. However, while Eq. (4.20) assumes that the CG interactions are pair-additive and do not explicitly depend upon the volume, W_P describes many-body interactions that explicitly depend upon the volume. In fact, one expects that contributions to W_P that vary only weakly with (or are independent of) CG configuration provide cohesion that significantly reduces the pressure.[191,193] These contributions are effectively invisible to structure-based methods that focus on reproducing RDFs, which are primarily determined by short-ranged, rapidly varying repulsive potentials.[78] Thus, it is unsurprising that bottom-up CG models tend to dramatically over-estimate the internal pressure.[34,79]

Following Das and Andersen (DA),[44] we have adopted a more general form for the approximate potential in order to model the volume-dependence of the PMF[151]

$$U(\mathbf{R},V) = U_R(\mathbf{R}) + U_V(V). \tag{4.21}$$

The interaction potential, U_R, is optimized to approximate the configuration dependence of the PMF via standard structure-based methods.[32] Since U_R does not explicitly depend upon V and is (usually) pair-additive, it contributes P^0_{xs} to the pressure according to Eq. (4.20). Conversely, U_V does not impact the equilibrium configuration distribution of the CG model, but directly contributes to the pressure:

$$P_{\text{xs}}(\mathbf{R},V) = P^0_{\text{xs}}(\mathbf{R};V) + F_V(V), \tag{4.22}$$

where $F_V = -\mathrm{d}U_V/\mathrm{d}V$ is a "pressure correction." Similarly, because

$$\left(\frac{\partial P_{\text{xs}}(\mathbf{R},V)}{\partial V}\right)_{\hat{\mathbf{R}}} = \left(\frac{\partial P^0_{\text{xs}}(\mathbf{R};V)}{\partial V}\right)_{\hat{\mathbf{R}}} - \frac{\mathrm{d}^2U_V(V)}{\mathrm{d}V^2}, \tag{4.23}$$

the second derivative of U_V directly contributes to the (inverse) compressibility. Consequently, U_V can be constructed to accurately model the pressure, the compressibility, and more generally the pressure equation of state for the atomic model. Note that employing a potential that "actively" varies with the density can intro-

duce modifications to the chemical potential, which must be considered to reconcile the virial and compressibility routes for calculating the pressure.[93]

Given a fixed interaction potential, U_R, DA proposed optimizing U_V by minimizing a "pressure-matching" functional:

$$\chi_2^2[U] = \left\langle \left| p_{\text{int}}(\mathbf{r}; V, T) - P_{\text{int}}(\mathbf{M}(\mathbf{r}), V; T; U) \right|^2 \right\rangle_{PT}, \tag{4.24}$$

in which the average is evaluated over the constant NPT ensemble for the atomic model. Subsequently, we developed a self-consistent pressure-matching approach that optimizes U_V to quantitatively reproduce the atomic pressure equation of state.[151] This iterative pressure-matching approach corresponds to variationally minimizing a relative entropy[29] with respect to U_V:[189]

$$S_{\text{rel}}[U] = \int \mathrm{d}V \int_V \mathrm{d}\mathbf{R}\, p_{RV}(\mathbf{R}, V) \ln \left[p_{RV}(\mathbf{R}, V) / P_{RV}(\mathbf{R}, V; U) \right], \tag{4.25}$$

where $p_{RV}(\mathbf{R}, V)$ and $P_{RV}(\mathbf{R}, V; U)$ are equilibrium distributions for the atomic and CG models, respectively, at constant P_{ext} and T. It should be noted that W_P minimizes both χ_2^2 and S_{rel}. However, given the approximate potential in Eq. (4.21), minimizing S_{rel} ensures that the CG model reproduces the atomic pressure equation of state, while minimizing χ_2^2 does not ensure such consistency.

4.4.2 Numerical results

Recently, we tested the pressure-matching approach for molecular liquids.[151] Figures 4.2 and 4.3 compare the density fluctuations, pressure equations of state, and pressure-volume fluctuations obtained from constant NPT simulations of the OPLS-AA model for heptane[104] and from simulations of several 3-site CG models. The black points in Fig. 4.3b present a scatter plot of the volume and instantaneous pressure sampled by the OPLS-AA model. The corresponding black curves in Figs. 4.2 and 4.3a present the simulated volume fluctuations and pressure equation of state, respectively.

Given these atomic simulations, we employed the multiscale coarse-graining (MS-CG) variational principle[25,27,88] to determine an interaction potential, U_R, for 3-site CG models. This MS-CG potential quite accurately described the structure of liquid heptane, but dramatically overestimated the pressure of the OPLS-AA

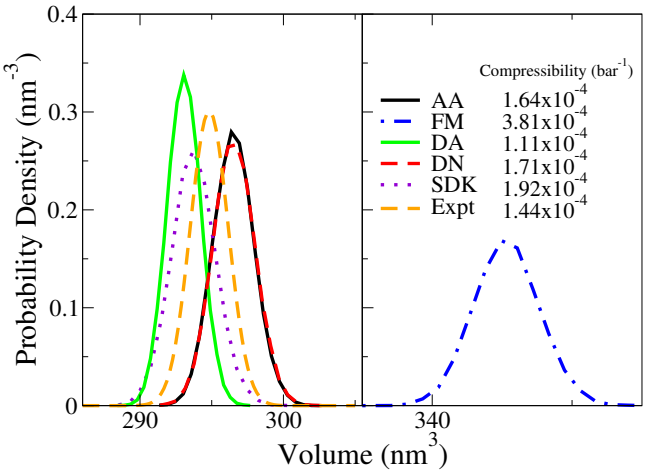

Figure 4.2. Simulated volume distributions for various heptane models. The solid black curve presents the simulated distribution for the OPLS-AA model. The dashed-dotted blue, solid green, dashed red, and dotted purple curves indicate simulated distributions for the MS-CG, DA, DN, and SDK 3-site models, respectively. The dashed orange curve indicates the normal distribution that is constructed from the experimentally known density and compressibility of heptane.

model. The cyan points in Fig. 4.3b present a scatter plot of the instantaneous pressure that is generated by applying the MS-CG interaction potential to the configurations sampled by the OPLS-AA model, i.e., $Nk_BT/V + P^0_{\text{xs}}(\mathbf{M}(\mathbf{r}); V)$. Consequently, constant NPT simulations with the MS-CG interaction potential (without including a pressure correction) overestimated the volume of the OPLS-AA model by more than 10%, as indicated by the blue curve in Fig. 4.2.

Given this MS-CG interaction potential, we then employed the DA pressure-matching variational principle[44] to determine a volume-dependent potential, U_V. As indicated by the green curves in Figs. 4.2 and 4.3, the resulting DA model much more accurately described the OPLS-AA pressure-volume behavior. Finally, we iteratively refined U_V via self-consistent pressure-matching. The red curves in Figs. 4.2 and 4.3 demonstrate that the resulting DN model reproduced the equilibrium density, pressure, and compressibility of the OPLS-AA model with nearly quantitative accuracy.

Figures 4.2 and 4.3 also provide instructive comparisons with experiment and with a top-down model. The orange curves present results inferred from experimental measurements of the equilibrium density and compressibility of heptane.[107] The purple curves present results for a top-down model, which Shinoda, Devane,

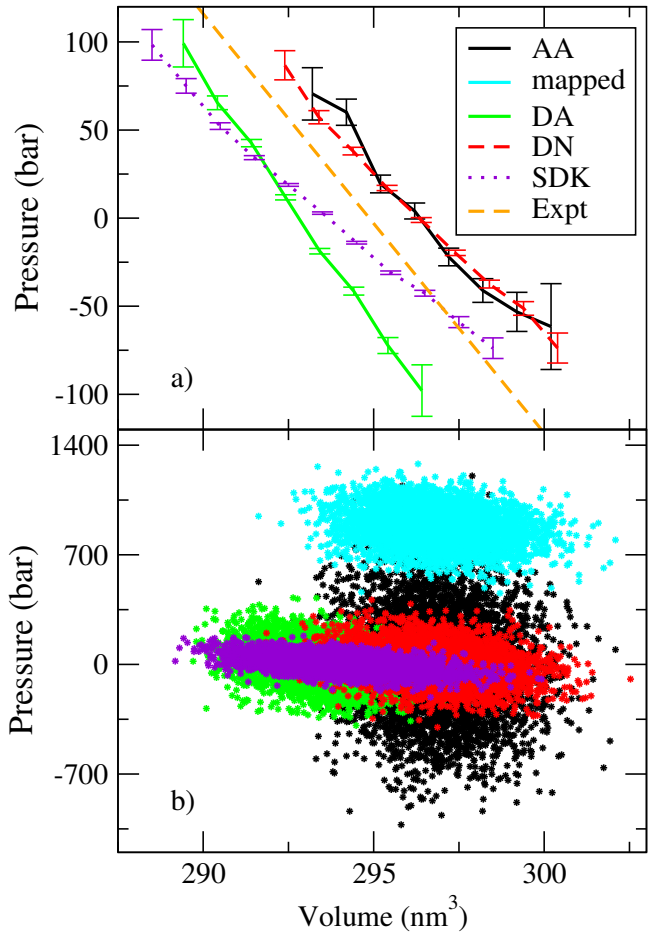

Figure 4.3. Comparison of the pressure-volume behavior for the AA heptane model and for different 3-site CG heptane models. The black, green, red, and purple curves correspond to the models of Fig. 3. Panel a) presents the equation of state for each model, which is estimated from the mean pressure at each volume in the simulated constant NPT ensemble. The error bars indicate the standard error in the simulated means. The orange curve indicates the equation of state that is determined from the experimentally known density and compressibility of heptane. Panel b) presents a scatter plot of the simulated pressure and volume. The cyan points correspond to the pressure, P^0_{CG}, that is determined by applying the MS-CG potential to the mapped ensemble.

and Klein (SDK) parameterized to reproduce the bulk density and liquid-vapor surface tension, but not the compressibility, of heptane.[113] Because it accurately describes the volume-dependence of the PMF, the bottom-up DN model reproduces experimental measurements of the equilibrium density and compressibility with comparable, if not better, accuracy than the top-down SDK model. Thus, Figs. 4.2 and 4.3 demonstrate the promise of bottom-up CG methods for accurately

describing both structural and thermodynamic properties.

We have performed self-consistent pressure-matching for 1-, 2-, and 3-site CG heptane models, for 1- and 3-site toluene models, and for 3-site models of heptane-toluene mixtures. In each case, we reproduced the atomic density, compressibility, and pressure equation of state with nearly quantitative accuracy. Interestingly, the optimized pressure correction always dramatically reduced the internal pressure of the CG model, while the ideal kinetic contribution from the "missing atoms" corresponded to a much smaller increase in pressure. As illustrated in Fig. 4.4a, increasingly large pressure corrections are required with increased coarsening, as the MS-CG interaction potentials systematically lose cohesion, due to the increasingly entropic character of the PMF,[82] as reflected by reduced structure in the site-site RDFs.[189] Figure 4.4b demonstrates the same correlation between cohesion and pressure correction among MS-CG models for heptane-toluene mixtures of varying composition. Thus, the pressure correction appears to compensate for the cohesion that is lost in structure-based potentials, as suggested by the classic van der Waals picture.[191,193]

4.4.3 Transferability for mixtures

The state-point dependence of the PMF limits the transferability of approximate CG potentials. We previously proposed an extended ensemble approach for determining transferable potentials that optimally approximate the PMF across a range of thermodynamic conditions.[39] We have recently combined the extended ensemble and pressure matching approaches to develop predictive CG models for accurately modeling the structure and pressure-volume behavior of heptane-toluene mixtures.[189]

We first employed a global force-matching variational principle to determine a single set of system-independent, i.e., transferable, interaction potentials that accurately approximate the configuration-dependence of the PMF for a range of mixtures. Given this set of transferable interaction potentials, we then performed self-consistent pressure-matching to determine an optimal pressure correction for each mixture. Importantly, this pressure correction can be accurately predicted as a function of the mixture composition. Figure 4.5 demonstrates that the resulting CG models accurately reproduced the pressure-volume behavior not only for the

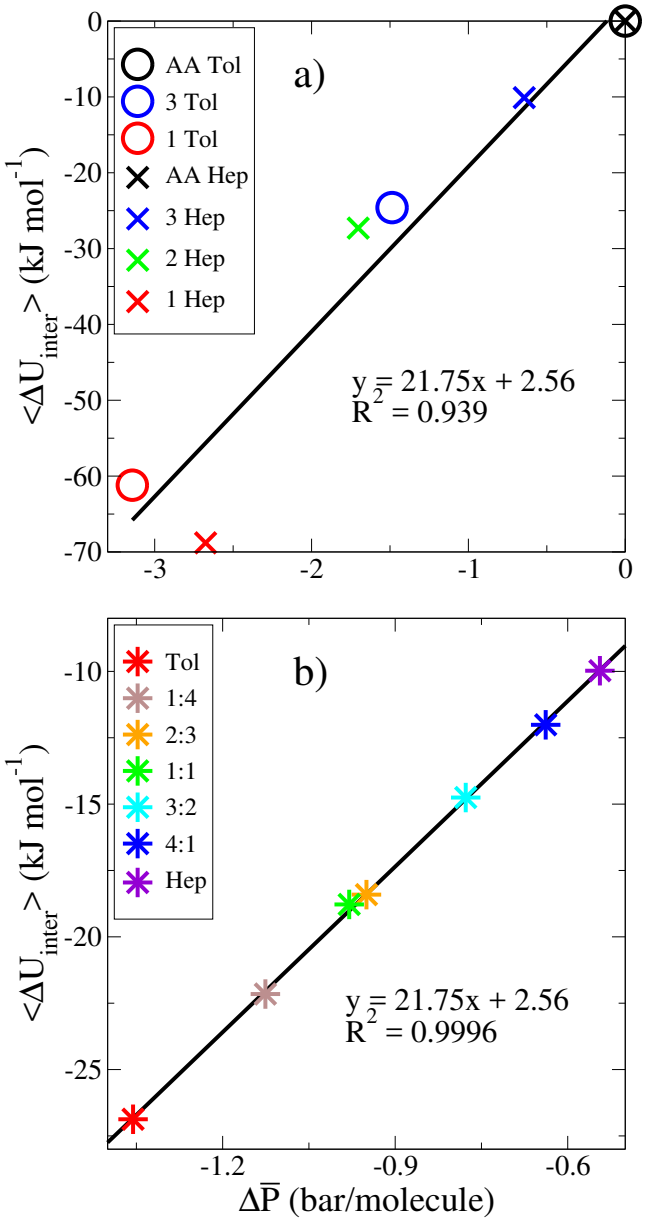

Figure 4.4. Scatter plot correlating the missing cohesive energy density, $\langle \Delta U_{Inter} \rangle$, with the average pressure correction, $\Delta \overline{P}$, required for various models. Panel a presents results for AA (black), as well as 3-site (blue), 2-site (green), and 1-site (red) CG models for heptane (crosses) and for toluene (circles). Panel b presents results for 3-site CG models of liquid mixtures with varying heptane:toluene mole ratio. The slight differences in the two panels for 3-site models of pure heptane and pure toluene reflect finite size effects.[44,151,189]

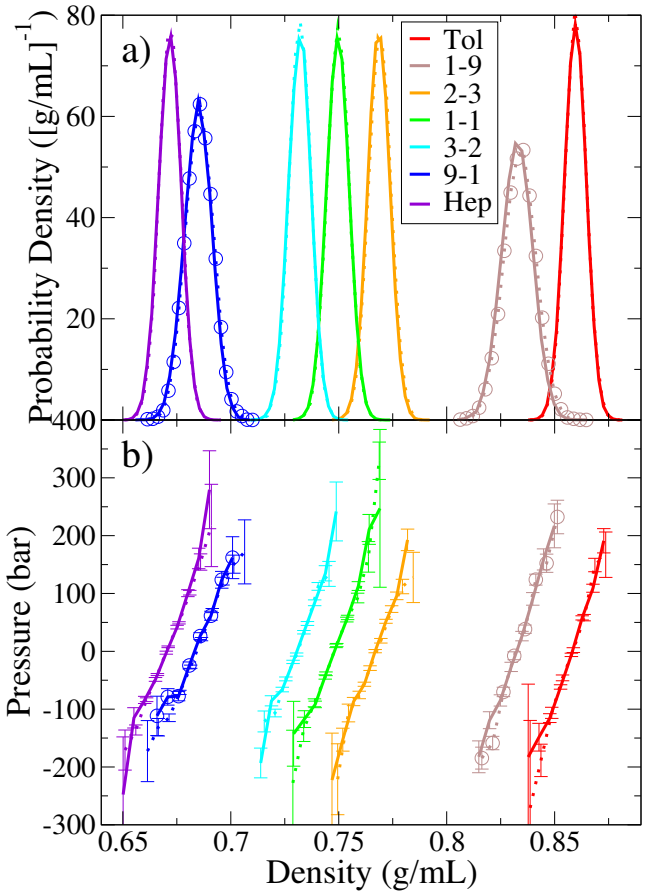

Figure 4.5. Simulated density distributions (top) and pressure-volume equations of state (bottom) for AA (solid) and 3-site CG (dotted) models for various heptane-toluene mixtures. Pure heptane, pure toluene, as well as the 2:3, 1:1, and 3:2 heptane:toluene mixtures were included in parameterizing the CG potentials. The results for the 1:9 and 9:1 heptane:toluene mixtures were not included in the parameterization and reflect predictions of the transferable model.

mixtures employed in the parameterization, but also for two additional mixtures that were not included in the parameterization.

4.5 Conclusion: van der Waals perspective

In closing, we hope that this work helps clarify the origin and inter-relationship of the representability and transferability problems that plague bottom-up coarse-graining approaches. Both arise as a consequence of thermodynamic information that has been extracted from the atomic configuration space and encoded into the many-

body PMF. The key to resolving these problems lies in quantifying this information and then incorporating it into the calculation of thermodynamic properties and the prediction of transferable potentials. In particular, pressure-matching provides a practical and rigorous way for determining the density dependence of the PMF in order to accurately model the pressure equation of state. Thus, bottom-up approaches can develop predictive, transferable potentials that accurately model the structure, density fluctuations, pressure, and compressibility of atomic models.

More generally, this work suggests a "van der Waals" perspective for bottom-up coarse-graining. From this perspective, current bottom-up approaches provide a powerful means for approximating the configuration-dependence of the PMF, such that the resulting models accurately model atomic structure. At the same time, these approaches do not effectively address the thermodynamic information that determines both the state-point dependence of the PMF and also the missing atomic contribution to thermodynamic properties. Complementary variational principles, such as pressure-matching, may provide an effective means for determining this information, both to predict the transferability of approximate potentials and to model thermodynamic properties.

Chapter 5
BOCS: Bottom-up Open-source Coarse-graining Software

N. J. H. Dunn, K. M. Lebold, M. R. Delyser, J. F. Rudzinski W. G. Noid, *J Phys Chem B* **2018**, 122 (13), 3363-3377

5.1 Abstract

We present the BOCS toolkit as a suite of open source software tools for parameterizing bottom-up coarse-grained (CG) models to accurately reproduce structural and thermodynamic properties of high resolution models. The BOCS toolkit complements available software packages by providing robust implementations of both the multiscale coarse-graining (MS-CG) force-matching method and also the generalized-Yvon-Born-Green (g-YBG) method. The g-YBG method allows one to analyze and to calculate MS-CG potentials in terms of structural correlations. Additionally, the BOCS toolkit implements an extended ensemble framework for optimizing the transferability of bottom-up potentials, as well as a self-consistent pressure-matching method for accurately modeling the pressure equation of state for homogeneous systems. We illustrate these capabilities by parameterizing transferable potentials for CG models that accurately model the structure, pressure, and compressibility of liquid alkane systems and by quantifying the role of many-body correlations in determining the calculated pair potential for a one-site CG model of liquid methanol.

5.2 Introduction

By representing systems in reduced detail, coarse-grained (CG) models provide the necessary computational efficiency for investigating length- and time-scales that cannot be effectively addressed with all-atom (AA) models.[7,72] Of course, CG models must be carefully constructed to faithfully describe the relevant physical forces if they are to provide useful predictions and insight. While one can imagine many approaches for constructing CG models, they are often developed via "top-down" or "bottom-up" approaches.[8,73–75]

Top-down approaches commonly parameterize relatively simple interaction potentials to reproduce macroscopic thermodynamic properties. Because top-down approaches often address multiple chemical systems and thermodynamic states, the resulting parameters can be used to define a general purpose force field. For instance, the Martini,[9,194] SDK,,[113] PLUM,[195,196] and OxDna[17,197] force fields each employ a single set of parameters that is quite transferable, i.e., the parameters reasonably describe thermodynamic properties for a fairly broad range of systems and environments.

In contrast, bottom-up approaches commonly parameterize relatively complex interaction potentials to reproduce the structural properties of a high resolution model for a single chemical system in a single thermodynamic state point. Consequently, bottom-up approaches do not usually provide transferable force fields, but rather system-specific potentials that may require re-parameterization for each system and state-point of interest.[73,75] Accordingly, the practical application of bottom-up methods requires appropriate software for parameterizing these potentials. Fortunately, several software packages[100,120,198,199] have been released for implementing bottom-up approaches according to, e.g., iterative Boltzmann Inversion,[28] Inverse Monte Carlo,[26] and the multiscale coarse-graining (MS-CG) methods.[25,27,88,99,200]

Unsurprisingly, bottom-up approaches are currently limited by two common deficiencies. As emphasized above, bottom-up models generally provide limited transferability.[33,35–39,41,42,44,153,182,201] Similarly, because they often focus on reproducing structural properties, bottom-up approaches generally provide a rather poor description of thermodynamic properties, such as the pressure.[33,34,91] Recently, we have examined the fundamental origin and interrelation between these transfer-

ability and representability limitations.[202] Moreover, we have developed rigorous computational methods for addressing these limitations in practice. In particular, the extended ensemble framework provides a principled bottom-up approach for developing potentials that accurately describe multiple chemical systems or thermodynamic states.[39,203] Additionally, self-consistent pressure-matching provides a straight-forward approach for constructing CG models that accurately model the pressure and compressibility of homogeneous systems.[151,189]

In this work, we present the Bottom-up Open-source Coarse-graining Software (BOCS) toolkit to complement the software packages that are currently available for parameterizing bottom-up CG models. The BOCS toolkit includes software written in C, C++, and python for use with the GROMACS[103,204] and LAMMPS[114] simulation packages. The BOCS toolkit includes a robust and stable implementation of the MS-CG force-matching method[25,27] for determining CG potentials directly from atomistic forces. Additionally, the BOCS toolkit implements the generalized Yvon-Born-Green (g-YBG) framework[31,111] for calculating MS-CG potentials directly from structural data. Based upon the g-YBG framework, the BOCS toolkit provides tools for interpreting the physical origin of these potentials in terms of structural correlations generated by the high resolution model.[117] Moreover, the BOCS toolkit implements both the extended ensemble framework[39] and also the self-consistent pressure-matching method.[151,189]

We are releasing the BOCS toolkit as open source software under the GPLv3 license in the hope that the CG modeling community will use and modify these tools according to its needs. Open source software is vital to reproducible computational research, since it facilitates not only the examination of calculations performed with the software, but also of the software itself. The 'many eyes' effect of open source software can help to more quickly identify and correct errors in the software, while also providing opportunity for other researchers to review and improve the underlying algorithms. Finally, open source software lowers the barrier for researchers entering the field of CG modeling, since new researchers can then leverage and build upon prior work, rather than having to start from scratch.

The remainder of this manuscript is organized as follows. Section II outlines the theoretical basis for the BOCS toolkit, while Section III describes its computational implementation. Section IV illustrates the capabilities of the BOCS toolkit in the context of parameterizing transferable interaction potentials for CG models

that reasonably describe the structure and pressure-volume thermodynamics of butane, heptane, decane, and a butane-decane mixture. We also present some diagnostic capabilities of the BOCS toolkit using a one-site model of methanol as a representative example. Finally, Section V presents concluding remarks.

5.3 Theory

In this section, we briefly outline the theoretical foundation that is employed by the BOCS software package in parameterizing the potentials for a CG model from a high resolution simulation. The BOCS software package can employ statistics sampled from either the constant NVT or constant NPT ensemble to determine the CG interaction potential. However, CG models will generally require an additional volume-dependent potential to accurately calculate the pressure and to sample the correct density in the constant NPT ensemble.[44,151]

5.3.1 High resolution AA model

We first consider a high resolution model with n particles, $i = 1, \ldots, n$, which we shall refer to as atoms.[190] We indicate the atomic microstate by $(\mathbf{r}, \mathbf{p}, v)$, where the configuration $\mathbf{r} = (\mathbf{r}_1, \ldots, \mathbf{r}_n)$ indicates the Cartesian coordinates of each atom, $\mathbf{p} = (\mathbf{p}_1, \ldots, \mathbf{p}_n)$ indicates the corresponding set of momenta, and v indicates the volume. We assume an atomic Hamiltonian:

$$h(\mathbf{r}, \mathbf{p}, v) = \kappa(\mathbf{p}) + u(\mathbf{r}, v) \tag{5.1}$$

where the kinetic energy $\kappa(\mathbf{p}) = \sum_i \mathbf{p}_i^2/2m_i$, m_i is the mass of atom i, and the potential, $u(\mathbf{r}, v)$, may explicitly depend upon v, e.g., due to long-ranged interactions.[1,205,206] The potential determines a force $\mathbf{f}_i = -\left(\partial u/\partial \mathbf{r}_i\right)_v$ on each atom i and also a force on the wall, i.e., the instantaneous excess pressure, $p_{\text{xs}} = -\left(\partial u/\partial v\right)_{\hat{\mathbf{r}}}$, where the latter partial derivative is performed at constant scaled coordinates, $\hat{\mathbf{r}}$. The fluctuating internal pressure of the AA model is then[102,190]

$$p_{\text{int}}(\mathbf{r}, \mathbf{p}, v) = \frac{2}{3v}\kappa(\mathbf{p}) + p_{\text{xs}}(\mathbf{r}, v). \tag{5.2}$$

5.3.2 Low resolution CG model

We next consider a low resolution model with $N \leq n$ particles, $I = 1, \ldots, N$, which we shall refer to as CG sites. We indicate the CG microstate by $(\mathbf{R}, \mathbf{P}, V)$, where the configuration $\mathbf{R} = (\mathbf{R}_1, \ldots, \mathbf{R}_N)$ indicates the Cartesian coordinates of each site, $\mathbf{P} = (\mathbf{P}_1, \ldots, \mathbf{P}_N)$ indicates the corresponding set of momenta, and V indicates the volume. We assume a CG Hamiltonian:

$$H(\mathbf{R}, \mathbf{P}, V) = \mathscr{K}(\mathbf{P}) + U(\mathbf{R}, V) \tag{5.3}$$

where the kinetic energy $\mathscr{K}(\mathbf{P}) = \sum_I \mathbf{P}_I^2/2M_I$, M_I is the mass of site I, and the potential, $U(\mathbf{R}, V)$, may depend upon both $\mathbf{R}$ and also V, as indicated below.[22,44,151] The potential determines a force $\mathbf{F}_I = -\left(\partial U/\partial \mathbf{R}_I\right)_V$ on each site I and also the instantaneous excess pressure, $P_{\text{xs}} = -\left(\partial U/\partial V\right)_{\hat{\mathbf{R}}}$, where the latter partial derivative is performed at constant scaled CG coordinates, $\hat{\mathbf{R}}$. The fluctuating internal pressure of the CG model is then

$$P_{\text{int}}(\mathbf{R}, \mathbf{P}, V) = \frac{2}{3V}\mathscr{K}(\mathbf{P}) + P_{\text{xs}}(\mathbf{R}, V). \tag{5.4}$$

5.3.3 Mapped Ensemble

We intend for the CG model to reproduce the structural and thermodynamic properties of the AA model that can be observed at the resolution of the CG model. Accordingly, we define a mapped ensemble by mapping each AA microstate $(\mathbf{r}, \mathbf{p}, v)$ to a CG microstate $(\mathbf{R}, \mathbf{P}, V)$. The mapping preserves the volume of the AA microstate, i.e., $V = v$.[44,151] The mapped configuration, $\mathbf{R} = \mathbf{M}(\mathbf{r})$, and momenta, $\mathbf{P} = \mathbf{M}_P(\mathbf{p})$, are specified by determining the Cartesian coordinates, $\mathbf{R}_I$, and momenta, $\mathbf{P}_I$, of each site, I, as a linear combination of atomic coordinates, $\mathbf{r}_i$, and momenta, $\mathbf{p}_i$:

$$\mathbf{R}_I = \mathbf{M}_I(\mathbf{r}) = \sum_i c_{Ii}\mathbf{r}_i \tag{5.5}$$

$$\mathbf{P}_I = \mathbf{M}_{PI}(\mathbf{p}) = M_I \sum_i c_{Ii}\mathbf{p}_i/m_i. \tag{5.6}$$

Note that Eq. (5.6) is equivalent to employing the same linear coefficients for mapping both the coordinates and the velocities.[25]

In principle, the mapping coefficients can be arbitrary positive constants that are appropriately normalized, $\sum_i c_{Ii} = 1$ for each $I = 1, \ldots, N$.[25] This normalization ensures that if each atom is displaced by a constant vector, then each CG site is displaced by the same vector. However, for simplicity, the BOCS package requires that each atom is associated with at most one CG site, i.e., for each atom i, c_{Ii} is non-zero for at most one CG site I. Given this restriction, the mapped atomistic force on each CG site may be expressed

$$\mathbf{f}_I(\mathbf{r}) = \sum_{i \in I} \mathbf{f}_i(\mathbf{r}) \tag{5.7}$$

where the sum is performed over all atoms i that are "involved" in CG site, I, i.e., the atoms i for which $c_{Ii} > 0$.[25]

5.3.4 Consistency and the many-body potential of mean force

It is straightforward to ensure that the CG model samples the mapped momentum distribution. (Of course, this does not imply that the CG model accurately describes any other dynamical property.[207,208]) Because we have assumed that the CG sites correspond to disjoint atomic groups, the mapped CG momenta are statistically independent Gaussian random variables.[98] The CG model will be consistent with this mapped distribution if the site masses are given by

$${M_I}^{-1} = \sum_{i \in I} c_{Ii}^2 {m_i}^{-1}, \tag{5.8}$$

which corresponds to ensuring that the Boltzmann distribution for the CG momenta has the appropriate variance.[25] Note that, if the mapping coefficients c_{Ii} determine the CG coordinates as the mass center of each corresponding atomic group, then $M_I = \sum_{i \in I} m_i$.

In order for the CG model to sample the mapped distribution for the configuration and volume, the Boltzmann weight for each CG configuration, $\mathbf{R}$, must equal the net Boltzmann weight for the atomic configurations, $\mathbf{r}$, that map to $\mathbf{R}$ at the given volume, V. Accordingly, the appropriate potential is the many-body potential of mean force (PMF), W:

$$\exp[-\beta W(\mathbf{R}, V, T)] = V_0^{N-n} \int_{V^n} \mathrm{d}\mathbf{r} \exp[-\beta u(\mathbf{r}, V)] \; \delta(\mathbf{R} - \mathbf{M}(\mathbf{r})), \tag{5.9}$$

where V_0 is an arbitrary reference volume that ensures dimensional consistency.[21,22,44,77,82,151]

The BOCS software package employs two variational principles to determine the potential U for the CG model. The force-[27,173] and pressure-[44,151] matching functionals are defined

$$\chi_1^2[U] = \left\langle \frac{1}{3N} \sum_I |\mathbf{f}_I(\mathbf{r}) - \mathbf{F}_I(\mathbf{M}(\mathbf{r}))|^2 \right\rangle \tag{5.10}$$

$$\chi_2^2[U] = \left\langle \left| p_{\text{int}}(\mathbf{r}, \mathbf{p}, v) - P_{\text{int}}(\mathbf{M}(\mathbf{r}), \mathbf{M}_P(\mathbf{p}), v) \right|^2 \right\rangle, \tag{5.11}$$

where the angular brackets denote an equilibrium ensemble average for the high resolution model. In practice we typically approximate these ensemble averages with configurations sampled from high resolution simulations. By minimizing the functionals χ_1^2 and χ_2^2, the BOCS toolkit determines U to approximate the configuration- and volume-dependence of the PMF, respectively.[25,44,101,151,209]

Das and Andersen (DA) originally proposed weighting each configuration in χ_1^2 by a factor of $v^{2/3}$ in developing the pressure-matching method for systems in which the volume isotropically fluctuates.[44] Accordingly, the BOCS toolkit allows for the option of including this scaling in χ_1^2. However, this factor has no effect at constant V and appears to have little practical significance for condensed phase systems undergoing isotropic volume fluctuations at constant external pressure. Also, we note that the equivalence of Eq. (5.11) to the original pressure-matching functional proposed by Das and Andersen requires that the CG masses are consistently treated according to Eq. (5.8).

5.3.5 Approximate Potentials

We assume the following form for the CG potential:

$$U(\mathbf{R}, V) = U_R(\mathbf{R}) + U_V(V), \tag{5.12}$$

where the interaction potential, U_R, and volume-dependent potential, U_V, are optimized to approximate the configuration- and volume-dependence of the many-body PMF, respectively.[44,151]

5.3.5.1 Interaction potential

The interaction potential, U_R, is expressed as a sum of terms corresponding to different interactions, ζ, involving groups of particles, λ, that depend on scalar functions, ψ_ζ, of the corresponding CG coordinates, $\mathbf{R}_\lambda$:

$$U_R(\mathbf{R}) = \sum_\zeta \sum_\lambda U_\zeta(\psi_{\zeta\lambda}(\mathbf{R})) \tag{5.13}$$

where $\psi_{\zeta\lambda}(\mathbf{R}) = \psi_\zeta(\mathbf{R}_\lambda)$.[99,111] The Appendix illustrates this general potential form for a typical molecular potential. The resulting force on site I is then

$$\mathbf{F}_I(\mathbf{R}) = \sum_\zeta \sum_\lambda F_\zeta(\psi_{\zeta\lambda}(\mathbf{R}))\nabla_I \psi_{\zeta\lambda}(\mathbf{R}), \tag{5.14}$$

where $F_\zeta(x) = -\mathrm{d}U_\zeta(x)/\mathrm{d}x$ and $\nabla_I = \partial/\partial\mathbf{R}_I$. We represent each force function as a linear combination of basis functions, $f_{\zeta d}(x)$, with constant coefficients $\phi_{\zeta d}$:

$$F_\zeta(x) = \sum_d \phi_{\zeta d} f_{\zeta d}(x). \tag{5.15}$$

Given this representation of the force functions, we define force field "basis vectors"[99,111]

$$\mathscr{G}_{I;\zeta d}(\mathbf{R}) = \sum_\lambda f_{\zeta d}(\psi_{\zeta\lambda}(\mathbf{R}))\nabla_I \psi_{\zeta\lambda}(\mathbf{R}) \tag{5.16}$$

such that the force on each site may be expressed:

$$\mathbf{F}_I(\mathbf{R}) = \sum_\zeta \sum_d \phi_{\zeta d}\mathscr{G}_{I;\zeta d}(\mathbf{R}) = \sum_D \phi_D \mathscr{G}_{I;D}(\mathbf{R}) \tag{5.17}$$

where, in the last expression, D is a "super-index" that specifies a combination ζd. Given Eq. (5.17) for the CG forces, χ_1^2 becomes a simple quadratic form in the force field parameters. The parameters that minimize χ_1^2 and, thus, provide an optimal approximation to the configuration-dependence of the PMF are determined by solving the normal system of linear equations[31,99,101]

$$\sum_{D'} G_{DD'}\phi_{D'} = b_D \tag{5.18}$$

where

$$b_D = \left\langle \frac{1}{3N} \sum_I \mathbf{f}_I(\mathbf{r}) \cdot \mathscr{G}_{I;D}(\mathbf{M}(\mathbf{r})) \right\rangle \tag{5.19}$$

$$G_{DD'} = \left\langle \frac{1}{3N} \sum_I \mathscr{G}_{I;D}(\mathbf{M}(\mathbf{r})) \cdot \mathscr{G}_{I;D'}(\mathbf{M}(\mathbf{r})) \right\rangle. \tag{5.20}$$

Equation (5.18) can be interpreted as the projection of either the atomic force field or the many-body PMF (more precisely, the corresponding force field) onto the space of force fields spanned by the basis defined by Eq. (5.17).[25,99,111,209,210] Note that, if χ_1^2 scales each configuration by $v^{2/3}$, then b_D and $G_{DD'}$ both inherit this scaling in Eqs. (5.18)-(5.20). In practice we then divide b_D and $G_{DD'}$ by $\left\langle v^{2/3} \right\rangle$ in order to preserve their original scale and dimensions.

5.3.5.2 Volume-dependent potential

According to Eq. (5.12), the pressure of the CG model may be expressed:

$$P_{\text{int}}(\mathbf{R}, \mathbf{P}, V) = P_{\text{int}}^0(\mathbf{R}, \mathbf{P}, V) + F_V(V) \tag{5.21}$$

where

$$P_{\text{int}}^0(\mathbf{R}, \mathbf{P}, V) = \frac{2}{3V} \mathscr{K}(\mathbf{P}) - \left(\frac{\partial U_R(\mathbf{R})}{\partial V} \right)_{\hat{\mathbf{R}}} \tag{5.22}$$

includes the kinetic and virial contributions to the pressure from U_R, and $F_V(V) = -\mathrm{d}U_V(V)/\mathrm{d}V$ is a "pressure correction" for the CG model. Since U_R is optimized without regard to the pressure, P_{int}^0 will tend to dramatically overestimate the pressure of the underlying atomistic model.[34,44,79,88,151] Consequently, U_V can be adjusted to ensure that the CG model provides appropriate Boltzmann weight for each volume and, equivalently, that it accurately reproduces the pressure of the atomistic model. Importantly, U_V does not impact the configuration distribution at a fixed volume.[93]

Das and Andersen[44] suggested representing the volume-dependent potential as a sum of basis functions:

$$U_V(V) = \sum_d \psi_d u_{Vd}(V). \tag{5.23}$$

where ψ_d act as parameters for U_V, u_{Vd} are basis functions of the form

$$u_{Vd}(V) = \begin{cases} N(V/\bar{v}), & \text{for } d = 1 \\ N(V/\bar{v} - 1)^d, & \text{for } d \geq 2 \end{cases} \tag{5.24}$$

and $\bar{v}$ is the average volume of the reference AA ensemble. The BOCS toolkit can also employ other basis functions for representing U_V. However, in practice Eq. (5.24) is quite convenient, since often only two basis functions are required to accurately model equilibrium density fluctuations at constant external pressure.[44,151] The two coefficients then correspond to corrections for the pressure and the compressibility:

$$\Delta P_{\text{int}} = -N\psi_1/\bar{v} \tag{5.25}$$

$$\Delta \kappa_T^{-1} = 2N\psi_2/\bar{v}. \tag{5.26}$$

Given the interaction potential, U_R, determined from Eq. (5.18), U_V is then determined by minimizing the pressure-matching functional χ_2^2 in Eq. (5.11). Given Eqs. (5.21)-(5.23) for the pressure of the CG model, this pressure-matching variational principle reduces to a linear least squares problem for the parameters ψ_d, which is then solved by a normal system of equations analogous to Eq. (5.18). The resulting U_V significantly reduces the pressure of the CG model and will often provide a qualitatively reasonable description of the AA pressure equation of state.[44,151]

The BOCS toolkit implements an iterative self-consistent pressure-matching method to further refine U_V such that the CG model quantitatively reproduces the AA pressure equation of state, $p_{\text{int}}(V)$.[151] In this method, one first simulates the CG model in the constant NPT ensemble with a fixed interaction potential, U_R, and a trial estimate for U_V. This simulation provides a local estimate of the CG equation of state, $P_{\text{int}}(V)$. The discrepancy between the CG and AA equations of state then determines a correction to $F_V(V)$ in analogy to iterative Boltzmann inversion: $\delta F_V(V) = p_{\text{int}}(V) - P_{\text{int}}(V)$. In practice this procedure often quickly converges such that $p_{\text{int}}(V) \approx P_{\text{int}}(V)$ quite accurately. This procedure corresponds to determining U_V by minimizing a relative entropy[29,151,211] describing the overlap

of AA and CG distributions for the constant NPT ensemble:

$$S_{\mathrm{rel}}[U] = \int \mathrm{d}V \int_V \mathrm{d}\mathbf{R}\, p_{RV}(\mathbf{R}, V) \ln \left[p_{RV}(\mathbf{R}, V) / P_{RV}(\mathbf{R}, V; U) \right] \tag{5.27}$$

where p_{RV} and P_{RV} are the distributions for the mapped ensemble and for the CG model, respectively.

5.3.6 g-YBG formulation

In the canonical ensemble at constant volume, the normal equations for the MS-CG potential parameters are equivalent to a generalization of the Yvon-Born-Green equation from liquid state theory.[119,191] This can be seen by representing the CG force field with a continuous set of basis functions such that Eq. (5.17) can be expressed[31,111]

$$\mathbf{F}_I(\mathbf{R}) = \sum_\zeta \int \mathrm{d}x\, F_\zeta(x)\, \mathscr{G}_{I;\zeta}(\mathbf{R}; x). \tag{5.28}$$

The normal MS-CG equations may then be expressed:

$$b_\zeta(x) = \bar{g}_\zeta(x)\, F_\zeta(x) + \sum_{\zeta'} \int \mathrm{d}x'\, \bar{G}_{\zeta\zeta'}(x, x')\, F_{\zeta'}(x') \tag{5.29}$$

where

$$b_\zeta(x) = \frac{1}{3N} \left\langle \sum_\lambda \left(\sum_I \mathbf{f}_I(\mathbf{r}) \cdot \nabla_I \psi_{\zeta\lambda}(\mathbf{M}(\mathbf{r})) \right) \delta\left(\psi_{\zeta\lambda}(\mathbf{M}(\mathbf{r})) - x\right) \right\rangle \tag{5.30}$$

may be interpreted as an average atomic force along the ψ_ζ order parameter, while

$$\bar{g}_\zeta(x) = \frac{1}{3N} \left\langle \sum_\lambda \left(\sum_I \left| \nabla_I \psi_{\zeta\lambda}(\mathbf{M}(\mathbf{r})) \right|^2 \right) \delta\left(\psi_{\zeta\lambda}(\mathbf{M}(\mathbf{r})) - x\right) \right\rangle \tag{5.31}$$

$$\begin{aligned} \bar{G}_{\zeta\zeta'}(x, x') = \frac{1}{3N} \Bigg\langle & \sum_{\lambda \neq \lambda'} \left(\sum_I \nabla_I \psi_{\zeta\lambda}(\mathbf{M}(\mathbf{r})) \cdot \nabla_I \psi_{\zeta'\lambda'}(\mathbf{M}(\mathbf{r})) \right) \\ & \times \delta\left(\psi_{\zeta\lambda}(\mathbf{M}(\mathbf{r})) - x\right) \delta\left(\psi_{\zeta'\lambda'}(\mathbf{M}(\mathbf{r})) - x'\right) \Bigg\rangle \end{aligned} \tag{5.32}$$

are ensemble averages describing equilibrium structural correlations. Eq. (5.29) can provide insight into the physical origin of the calculated potential, $U_\zeta(x)$, since it decomposes $b_\zeta(x)$ into a direct contribution from $U_\zeta(x)$ and correlated indirect

contributions from every other interaction in the system.[117,119]

Moreover, we have previously demonstrated that $b_\zeta(x)$ can be directly calculated from structures[31,111]

$$b_\zeta(x) = k_B T \left[\mathrm{d}\bar{g}_\zeta(x)/\mathrm{d}x - L_\zeta(x) \right] \tag{5.33}$$

where

$$L_\zeta(x) = \frac{1}{3N} \left\langle \sum_\lambda \left(\sum_I \nabla_I^2 \psi_{\zeta\lambda}(\mathbf{M}(\mathbf{r})) \right) \delta \left(\psi_{\zeta\lambda}(\mathbf{M}(\mathbf{r})) - x \right) \right\rangle. \tag{5.34}$$

In particular, if $U_\zeta(x)$ is a central pair potential, then

$$b_\zeta(r) = -(2r^2/c_\zeta) w'_\zeta(r) g_\zeta(r) \tag{5.35}$$

where $g_\zeta(r)$ is the radial distribution function, $w_\zeta(r) = -k_B T \ln g_\zeta(r)$ is the corresponding pair potential of mean force,[212] and c_ζ is a dimensioned normalization constant. In this simple case, Eq. (5.29) may be re-expressed

$$-\mathrm{d}w_\zeta(r)/\mathrm{d}r = F_\zeta(r) + \sum_{\zeta'} \int \mathrm{d}x \; \bar{g}_\zeta^{-1}(r) \bar{G}_{\zeta\zeta'}(r,x) F_{\zeta'}(x), \tag{5.36}$$

in direct analogy to the YBG equation.[119]

These results also hold in the constant NPT ensemble as long as b_ζ and $G_{\zeta\zeta'}$ are defined according to Eqs. (5.19) and (5.20), respectively. However, if b_ζ and $G_{\zeta\zeta'}$ include the $v^{2/3}$ rescaling proposed in Ref. 44, then this analysis only approximately holds in the constant NPT ensemble.

5.3.7 Extended Ensemble Formulation

The extended ensemble approach provides a simple framework for determining interaction potentials that are transferable to multiple systems.[39] An extended ensemble is defined as a collection of multiple conventional ensembles that may differ in chemical identity or in thermodynamic conditions. We assign a label, γ, and a probability, p_γ, for each ensemble. We define extended ensemble averages

$$\langle a_\gamma(\mathbf{r}_\gamma) \rangle = \sum_\gamma p_\gamma \left\langle a_\gamma(\mathbf{r}_\gamma) \right\rangle_\gamma \tag{5.37}$$

where $\mathbf{r}_\gamma$ indicates a configuration for ensemble γ and $\langle \cdots \rangle_\gamma$ indicates the corresponding conventional equilibrium ensemble average. In practice, we simply assign equal weight to each γ included in the extended ensemble. For each ensemble, γ, we define a CG representation, $\Gamma = \mu(\gamma)$, and a corresponding configuration mapping: $\mathbf{R}_\Gamma = \mathbf{M}_\gamma(\mathbf{r}_\gamma)$. This mapping then determines a weight, $p_\Gamma = \sum_\gamma p_\gamma \delta_{\gamma,\mu(\gamma)}$, and also a many-body potential of mean force, W_Γ, for each Γ. In practice, the CG representation typically provides a one-to-one relationship between the atomistic and CG ensembles, i.e., each CG ensemble Γ corresponds to a single atomistic ensemble γ_Γ and $p_\Gamma = p_{\gamma_\Gamma}$

We seek to determine potentials U_Γ that provide an optimal approximation to W_Γ for each Γ. The MS-CG force-matching variational principle can be readily extended for this purpose by simply interpreting Eqs. (5.10) and (5.18)-(5.20) in terms of extended ensemble averages. If the potentials U_Γ are treated independently for each Γ, then the extended ensemble approach determines independent MS-CG models for each Γ. However, if the potentials share transferable parameters, then these parameters are determined to provide an optimal approximation across the entire extended ensemble.

5.4 Computational Methods

The BOCS toolkit provides software tools for parameterizing the potential, $U(\mathbf{R}, V)$, for a CG model based upon information from an AA trajectory. Table 5.1 summarizes the primary input and output for these tools. Figure 5.1 outlines the workflow for determining the interaction potential, U_R, while Fig. 5.2 outlines the workflow for determining the volume-dependent potential, U_V.

The cgmap tool generates a mapped ensemble as the CG representation of an AA ensemble. The cgmap tool requires an AA trajectory file that contains atomically detailed configurations and, optionally, the corresponding velocities and forces. The cgmap tool also requires 1) a plain text file that determines the mapping coefficients, $\{c_{Ii}\}$, by specifying the CG representation for each type of molecule in the system; and 2) a CG topology file that specifies the type and connectivity of the sites in the CG model. Based upon the specified mapping coefficients, the cgmap tool determines the CG representation of each AA configuration according to Eq. (5.5). If the AA trajectory file includes velocity and force information, the

Table 5.1. Tools included in the BOCS toolkit with their primary inputs and outputs

Tool	Purpose	Input	Output
cgmap	Maps AA trajectory	AA trajectory, CG and map topologies	Mapped CG trajectory
cgff	Determines interaction potential, U_R	Mapped CG trajectory, CG potential definition	Interaction potential parameters, ϕ_D
tables	Converts CG potentials to GROMACS format	CG potential parameters	GROMACS table files
translate_table.py	Converts CG potentials to LAMMPS format	GROMACS table files	LAMMPS table files
pmatch	Determines volume potential, U_V	AA, CG pressures and volumes	Volume potential parameters, ψ_d
lmp_pmatch	Simulates CG model with $U = U_R + U_V$	LAMMPS table files, pressure correction	Simulated CG trajectory

cgmap tool determines the mapped velocities and forces according to Eqs. (5.6) and (5.7). The cgmap tool then provides a mapped CG trajectory file that can be analyzed using standard GROMACS tools.

The cgff tool calculates the parameters for U_R from the mapped CG trajectory file. The cgff tool requires a plain text input file to specify the types of potentials, U_ζ, included in U_R and also the basis functions, $f_{\zeta d}$, employed to represent each U_ζ. The cgff tool also requires a CG topology file to specify the instances, λ, of each interaction. Assuming that the mapped CG trajectory contains explicit force information, the cgff tool calculates the force correlation function, b_D, and structural correlation function, $G_{DD'}$, according to Eqs. (5.19) and (5.20), respectively, for each pair of basis functions $D = \zeta d$ and $D' = \zeta' d'$. If forces are not present in the mapped trajectory, the cgff tool calculates b_D directly from structural information according to Eq. (5.33). Although force-based calculations (i.e., via Eq. (5.19))

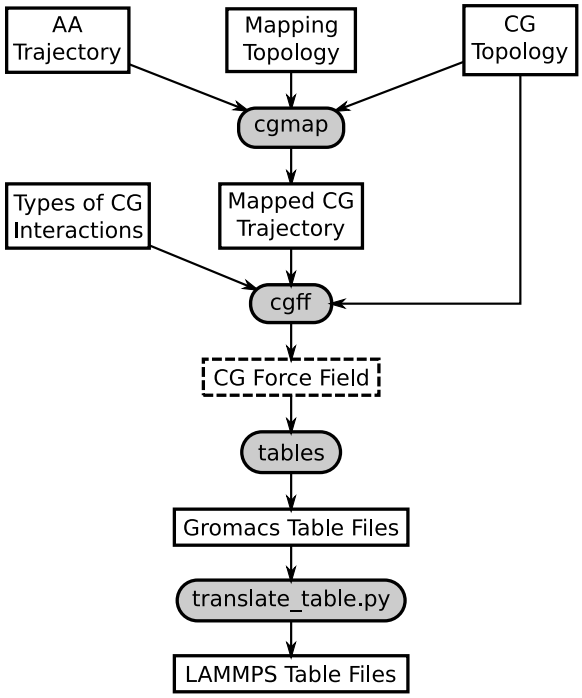

Figure 5.1. Workflow for the force-matching/g-YBG component of the BOCS toolkit. Boxes with sharp corners denote files, while boxes with rounded corners indicate operations performed on these files. Boxes filled with gray represent software tools provided in the BOCS toolkit. The dashed box indicates the major output of this workflow: the CG interaction potential, U_R.

require less sampling to accurately determine b_D, structure-based calculations (i.e., via Eq. (5.33)) yield equivalent results for sufficiently well sampled systems[31,39,171] and have proven quite useful for several applications.[178,203] The cgff tool then solves the normal system of linear equations, Eq. (5.18), for the potential parameters, ϕ_D. Finally, the cgff tool outputs these parameters, as well as, b_D, $G_{DD'}$, and additional supplemental files that characterize the system and provide diagnostic information about the calculation.

The cgff tool treats a fairly wide range of CG potentials that can be represented according to Eq. (5.13), i.e., bond-stretch potentials, bond-angle potentials, dihedral potentials, and short-ranged pair potentials. The cgff tool can represent each of these potentials with either piecewise constant functions, piecewise linear functions, or B-spline functions. The cgff tool also implements several standard analytic functional forms, including harmonic bond-stretch or bond-angle potentials, Fourier-series dihedral potentials, and Lennard-Jones-type pair potentials. Additionally, the cgff

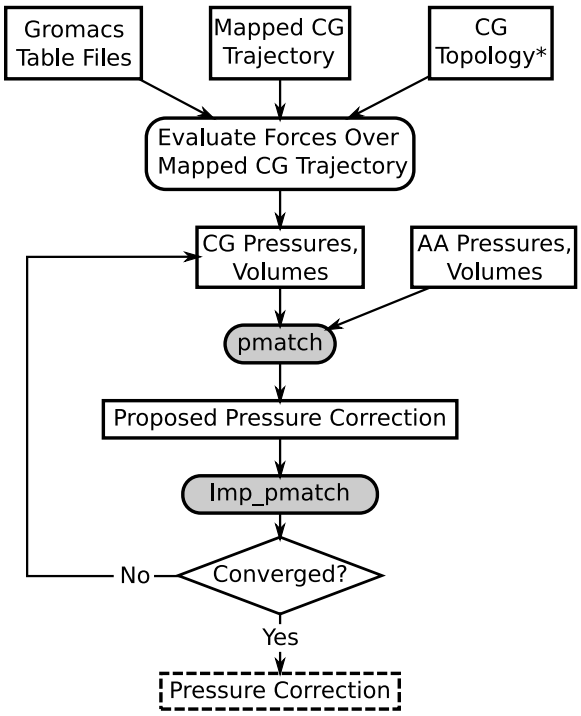

Figure 5.2. Workflow for the pressure-matching component of the the BOCS toolkit. See legend of Figure 5.1 for the meaning of the box shapes and outlines. The dashed box indicates the major output of this workflow: the CG volume potential, U_V.

tool allows for fixed "reference potentials," U_R^{Ref}, that are specified by the user and can be of arbitrary complexity.[99,213] In this case, the user must supply an additional trajectory file specifying the resulting reference force, $\mathbf{F}_I^{\text{Ref}}$, on each CG site in each mapped AA configuration. The cgff tool computes a corresponding contribution to each force projection, b_D^{Ref}, from the reference potential, i.e., using $\mathbf{F}_I^{\text{Ref}}$ in the place of $\mathbf{f}_I$ in Eq. (5.19). The cgff tool then optimizes the remaining terms in U_R to match the remainder of each force projection, $\delta b_D = b_D - b_D^{\text{Ref}}$. In particular, if Coulombic or other long-ranged potentials are defined as reference potentials, then the cgff tool will determine the short-ranged potentials that, when combined with the specified long-ranged potentials, provide an optimal approximation to the many-body PMF.

The cgff tool provides several additional options for the calculation of ϕ_D and the resulting output. The cgff tool can precondition the normal equations, Eq. (5.18), by normalizing the $G_{DD'}$ matrix according to the norm of each column, the max of each row, the total variance in each column, or the variance in b_D. The cgff tool can solve

these normal equations via single value, Cholesky, UU, or LU decomposition.[214] The cgff tool can regularize these methods according to Bayesian inference[176] or a simpler uncertainty estimation.[122] The cgff tool also provides several options for specialized diagnostic output, including error estimates, eigendecomposition of $\bar{G}_{DD'}$, and also decomposition of b_D into contributions from different interactions according to the g-YBG theory, i.e., Eq. (5.29).

Because it quantifies many-body structural correlations, the calculation of $G_{DD'}$ can be quite time-consuming for large systems with many interacting CG sites. As indicated by Eq. (5.32), the cgff tool calculates the correlation between the forces generated on each site, I, from each pair, λ and λ', of nonbonded interactions. The cgff tool performs this calculation by looping over all triples of interacting particles. For a CG model with N sites, this calculation scales as $O(N^3)$. We have expedited this calculation by exploiting the symmetry of this loop and by employing the OpenMPI framework to distribute the frames of the mapped trajectory over multiple processors. This parallelization scales perfectly because each frame is treated independently in calculating $G_{DD'}$ and because this nested triple loop typically dominates the time required for calculating ϕ_D.

We note that the MSCGFM code[100] implements the normal equations, Eq. (5.18), as well as several other numerical methods for minimizing χ_1^2 to determine the MS-CG force field. Lu et al. have provided an excellent discussion of various numerical methods for minimizing χ_1^2, including methods for solving an over-determined system of linear equations with a block-averaging approximation.[100] In comparison to this block-averaging approach, the normal system of equations is more time consuming, due to the nested triple loop discussed above, and also requires the numerical inversion of a matrix with a relatively high condition number. Nevertheless, we find that, with proper choices of solution method, preconditioning, and regularization, our implementation performs well and, in test cases that we can rigorously test, accurately determines the MS-CG potential. Additionally, because they correspond to a g-YBG integral equation that is explicitly expressed in terms of equilibrium ensemble averages, the normal equations facilitate molecular insight into the system and the resulting CG potentials.[117,119] Moreover, the normal equations allow for the calculation of these potentials directly from structural information.[31,111,171,203]

The cgff tool separates the calculated potential parameters, $\{\phi_D\}$, into files corresponding to different interactions. For interactions represented with simple

functional forms, such as bond-stretch interactions represented with harmonic potentials, the resulting parameters can be immediately employed as input for CG simulations. However, for potentials represented with more flexible functional forms, such as non-bonded interactions represented with spline functions, the calculated parameters may require additional processing. The tables tool performs the necessary smoothing, extrapolation, and interpolation to generate input files for use in GROMACS simulations.[103,204] The lammps_tables.py script converts these files for use in LAMMPS simulations.[114]

The cgff tool also implements the extended ensemble framework[39] to determine transferable potentials that provide an optimal approximation to the many-body PMF's for multiple mapped ensembles, Γ, that correspond to distinct chemical systems or distinct thermodynamic state points. In this case, the cgff tool requires a mapped CG trajectory file for each AA ensemble, as well as plain text and CG topology files that specify the contributions to the interaction potential, U_Γ, for modeling each Γ. The cgff tool also requires the user specify the weight, p_γ, for each AA ensemble, γ, included in the extended ensemble. Given this input, the cgff tool calculates b_D and $G_{DD'}$ as extended ensemble averages and determines the optimal potential parameters, ϕ_D, from Eq. (5.18), as in the case of a single system.

The cgmap and cgff tools have been historically developed for use with GROMACS and currently employ several functions and data structures from the GROMACS libraries.[103,204] In particular, we currently employ GROMACS functionality to read and write GROMACS trajectory and topology files, as well as for some aspects of the user interface employed by the cgmap tool. In order to buffer these tools from the GROMACS source code and in order to facilitate future compatibility, we developed an interface that wraps all references to GROMACS functions and addresses changes to relevant GROMACS libraries and files. The BOCS toolkit is currently natively compatible with GROMACS 4.5.x, 4.6.x, 5.0.x, and 5.1.x.

The BOCS toolkit also provides tools for determining U_V in order to simulate CG models that sample isotropic volume fluctuations under constant external pressure. The first step in this process is to estimate U_V via pressure-matching.[44,151] This calculation requires a fixed CG interaction potential, U_R, and a mapped CG trajectory file containing the mapped configuration, $\mathbf{M}(\mathbf{r}_t)$, mapped momentum, $\mathbf{M}_P(\mathbf{p}_t)$, and volume, V_t, for each time t. We then evaluate, for each t, the pressure,

P^0_{int}, that is defined by Eq. (5.22) and accounts for the kinetic and interaction contributions to the instantaneous pressure of the CG model. In practice, this can be done by post-processing the mapped CG trajectory file using the '-rerun' option with the standard GROMACS mdrun tool. (Note that, if the CG potential includes table files, then these files must be specified in the topology files for this post-processing calculation and for subsequent CG simulations with GROMACS, as indicated by * in Fig. 5.2.) Given the resulting set of CG pressures, $\{P^0_{\text{int}}(t)\}$, as well as the corresponding AA pressures and volumes, $\{p_{\text{int}}(t), V_t\}$, the pmatch tool then determines U_V to minimize χ^2_2.

The resulting CG potential, $U(\mathbf{R}, V) = U_R(\mathbf{R}) + U_V(V)$, can then be simulated with lmp_pmatch, which is a modification of the LAMMPS distribution[114] from 17 June 2013 that includes the contributions from U_V in the barostat equation of motion. These simulations determine an estimate for the pressure-volume equation of state, $P_{\text{int}}(V)$, for the CG model. In practice, this CG model does not perfectly reproduce the pressure equation of state, $p_{\text{int}}(V)$, of the AA model.[44,151,189] This discrepancy presumably arises due to differences between the mapped and simulated configurational distributions at each V. Consequently, if necessary, we perform iterative self-consistent pressure-matching in order to refine U_V.[151,189] The CG and AA pressure equations of state are provided as input to the pmatch tool, which then estimates the necessary correction for $U_V(V)$. This process can be iterated until the CG model adequately reproduces the AA pressure equation of state. In practice, this usually requires fewer than 10 iterations.[151,189]

There is no special workflow for determining U_V for transferable potentials obtained via the extended ensemble approach. In practice, we perform self-consistent pressure-matching to determine a separate potential $U_{V\Gamma}$ for each mapped ensemble, Γ. In principle, it may be possible to generalize the extended ensemble approach to determine a transferable pressure correction for modeling multiple state points or chemically distinct systems with similar interaction potentials. However, we have not yet tested this possibility.

5.5 Results and Discussion

In this section we illustrate the capabilities of the BOCS toolkit for parameterizing bottom-up CG models. In particular, we determine system-specific MS-CG poten-

tials that accurately describe the structure of butane, heptane, and decane. We employ the extended ensemble (XN) approach to determine a single set of transferable XN potentials for modeling the structure of all three liquids. Additionally, we determine volume potentials, U_V, for accurately modeling the pressure-volume behavior of each alkane system. Finally, we also employ the BOCS toolkit to characterize many-body correlations in liquid methanol and to investigate their contribution to the pair potential of mean force.

We performed atomistic MD simulations of three alkane systems with 267 butane, heptane, or decane molecules in order to parameterize three corresponding system-specific MS-CG potentials as well as a single set of transferable XN potentials. We also performed an atomistic MD simulation of a mixture with 134 butane molecules and 134 decane molecules in order to assess the predictive capability of the XN potential. We performed these simulations according to the procedures described in Ref. 189, which we briefly summarize in the following. We performed all atomistic simulations with GROMACS 4.5.3,[103] while using double-precision and a 1.0 fs timestep. We employed the OPLS-AA force field[104] to describe all interactions and employed the particle mesh Ewald method with a grid spacing of 0.08 nm to model electrostatic interactions.[105] In order to equilibrate these systems, we first heated each system to 1000 K and then cooled the system back to room temperature at constant volume. We next equilibrated each system at constant pressure, while employing the Berendsen thermostat and barostat.[106] Finally, we simulated each system at 1.0 bar pressure and an external temperature of 300 K, using the Parrinello-Rahman barostat[110] and the stochastic dynamics thermostat[215] with an inverse friction constant of 0.1 ps. The production runs of the pure systems were 45 ns in duration, while the production run of the mixture system was 70 ns. We note that, although we performed these simulations in double precision, BOCS can parameterize CG potentials from either single- or double-precision simulations.

We first employed the cgmap tool to map these AA trajectories to their CG representation. Figures 5.3a, 5.3b, and 5.3c present the CG representations for butane, heptane, and decane molecules, respectively. In each case, we represented terminal CH_2CH_3 groups with 'CT' sites and internal $CH_2CH_2CH_2$ groups with 'CM' sites. We employed a standard molecular mechanics CG potential to model each system. The intramolecular potentials included bond-stretch and bond-angle potentials between each pair and triple, respectively, of consecutive sites in the

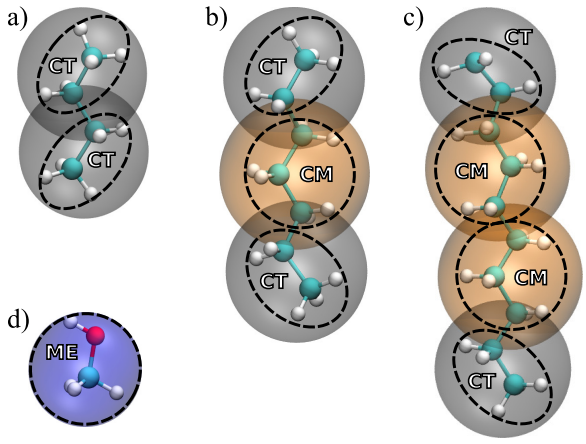

Figure 5.3. Mapping schemes for CG models superimposed upon the corresponding all-atom models, which are indicated in ball-and-stick representation. The CG sites (transparent spheres) are associated with the mass centers for the corresponding atomic groups, which are enclosed by the dashed circles. The size of the CG spheres indicates the distance at which the corresponding site-site radial distribution function vanishes, providing an estimate of the excluded volume for each site.

same molecule. The intermolecular potentials included short-ranged pair potentials between each pair of sites in distinct molecules. Table 5.2 lists the interactions included in the CG models for each liquid. The interactions that are highlighted in bold font were described by transferable potentials in the XN models, i.e., the XN models employed the same potential function for modeling these interactions in each alkane system. Note that the CG sites were not charged and that the intramolecular potential for the CG model of decane did not include a dihedral potential.

We next employed the cgff tool to determine system-specific MS-CG potentials[25,27] for each pure alkane system. Additionally, we also defined a parameterization extended ensemble by assigning a weight $p_\gamma = 1/3$ to each pure alkane system. We then employed the cgff tool to determine a single set of transferable XN potentials for optimally approximating the many-body PMF for all three systems. We note that we employed the $v^{2/3}$ rescaling in these calculations, although this appears to have minimal impact upon the resulting potentials. The Supporting Information section presents the calculated intramolecular potentials.

Figure 5.4 presents the calculated nonbonded pair potentials for CT-CT, CT-CM, CM-CM pairs in panels a, b, and c, respectively. The red, blue, and green solid curves in Fig. 5.4 indicate the system-specific MS-CG pair potentials for butane,

Table 5.2. Contributions included in the interaction potential for each alkane system. Highlighted interactions correspond to XN potential functions that are employed in multiple alkane systems.

Molecule	Bonds	Angles	Nonbonded
Butane	CT-CT	-	**CT-CT**
Heptane	**CT-CM**	CT-CM-CT	**CT-CT**
	-	-	**CT-CM**
	-	-	**CM-CM**
Decane	**CT-CM**	CT-CM-CM	**CT-CT**
	CM-CM	-	**CT-CM**
	-	-	**CM-CM**

heptane, and decane, respectively. Each MS-CG potential reflects two characteristic distances of approximately 0.5 nm and 0.8 nm. The CM-CM and CT-CM MS-CG potentials demonstrate relatively weak attraction and are quite similar for heptane and decane. The XN potentials are quite similar to the MS-CG potentials for these interactions. In comparison to the CM-CM and CM-CT potentials, the CT-CT potentials tend to be much more attractive and demonstrate much greater variation between different liquids. In particular, the CT-CT MS-CG potentials for butane and heptane are much more attractive than the CT-CM or CM-CM MS-CG potentials. The XN CT-CT potential is most similar to the corresponding MS-CG potential for butane.

We then employed the pmatch tool to determine the volume potential, U_V, via pressure-matching.[44,151] In particular, for each of the three pure liquid alkane systems, we determined two distinct volume potentials for compatibility with the system-specific MS-CG potential and the transferable XN potential. In each case, we represented U_V according to Eq. (5.24) with two basis functions that correspond to corrections for the mean pressure and the compressibility according to Eq. (5.25) and (5.26). The resulting potentials, U_V, provided a qualitative, but not quantitative description of the AA pressure-volume fluctuations. Consequently, we employed the self-consistent pressure-matching approach described in Section III to iteratively refine U_V.[151,189] Table 5.3 expresses the final parameters for U_V in terms of corrections to the mean pressure and compressibility. Table 5.3 also presents the number of iterations required to optimize U_V for each potential and

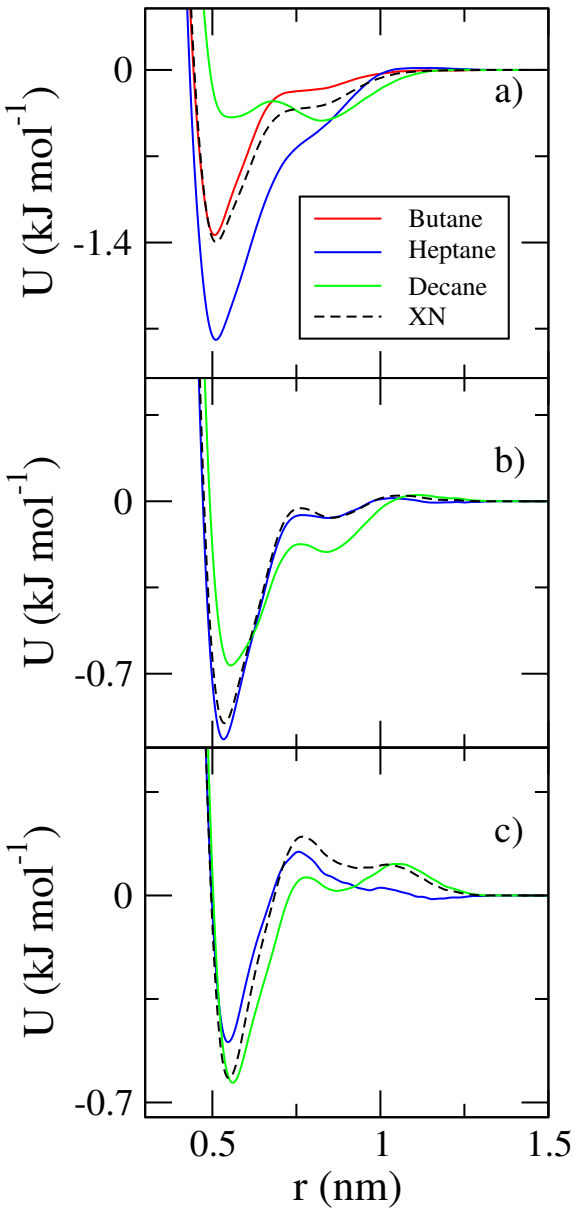

Figure 5.4. Calculated nonbonded potentials for a) CT-CT, b) CT-CM, c) CM-CM pair interactions. The solid red, blue, and green curves present MS-CG potentials calculated for butane, heptane, decane, respectively. The dashed black curves present the transferable XN potentials.

each system. In almost all cases, self-consistent pressure-matching converged within 6 iterations.

However, butane required special treatment during this pressure matching procedure. Because the CG model adopts a particularly high resolution for butane, the

Table 5.3. Average corrections for the pressure and inverse compressibility, as well as the number of iterations required by self-consistent pressure-matching. Pressures and inverse compressibilities are given in units of 10^3 bar. The asterisk (*) indicates that the pressure correction did not converge within 10 iterations and was manually determined according to the procedure described in Section 5.5.

	$\langle F_V \rangle$		$\Delta\kappa_T^{-1}$		N_{Iter}	
System	MS-CG	XN	MS-CG	XN	MS-CG	XN
Butane	-0.36	0.033	-0.86	-0.67	*	1
Heptane	-0.77	-1.59	-1.57	-2.93	6	6
Decane	-3.15	-2.46	-6.23	-6.23	4	6
But/Dec Mix	-	-1.55	-	-3.50	-	3

necessary pressure correction is quite small and requires special care. In particular, the first 10 iterations of self-consistent pressure-matching did not converge upon a pressure correction for the MS-CG butane model that simultaneously reproduced both the mean pressure and the compressibility of the AA model. Consequently, we selected the ψ_1 and ψ_2 coefficients from two different iterations that accurately modeled the mean pressure and the compressibility, respectively. Because the XN potential for butane is more attractive than the corresponding MS-CG potential, the XN butane model requires an even smaller pressure correction. Indeed, given the XN interaction potential, the volume potential that minimized χ_2^2 resulted in the XN butane model vaporizing. Consequently, in order to accurately reproduce the AA pressure-volume behavior with the XN butane model, we discarded the parameters $\{\psi_1, \psi_2\}$ obtained directly from pressure matching and performed iterative pressure matching starting from the trial potential $U_V = 0$. Starting from this trial potential, the iterative pressure-matching determined a satisfactory pressure correction with a single iteration.

All simulations of CG models were performed with the lmp_pmatch program included in the BOCS toolkit. These CG simulations employed the MTTK barostat[205,216] and Nose-Hoover chain thermostat[116] with the default chain length of 3. Otherwise, these simulations employed equivalent parameters to the AA simulations. Figures 5.5-5.6 quantify the equilibrium structure and pressure-volume behavior of the CG models for the pure alkane liquids. The Supporting Information more exhaustively compares the AA and CG models.

The system-specific MS-CG and transferable XN potentials reasonably describe

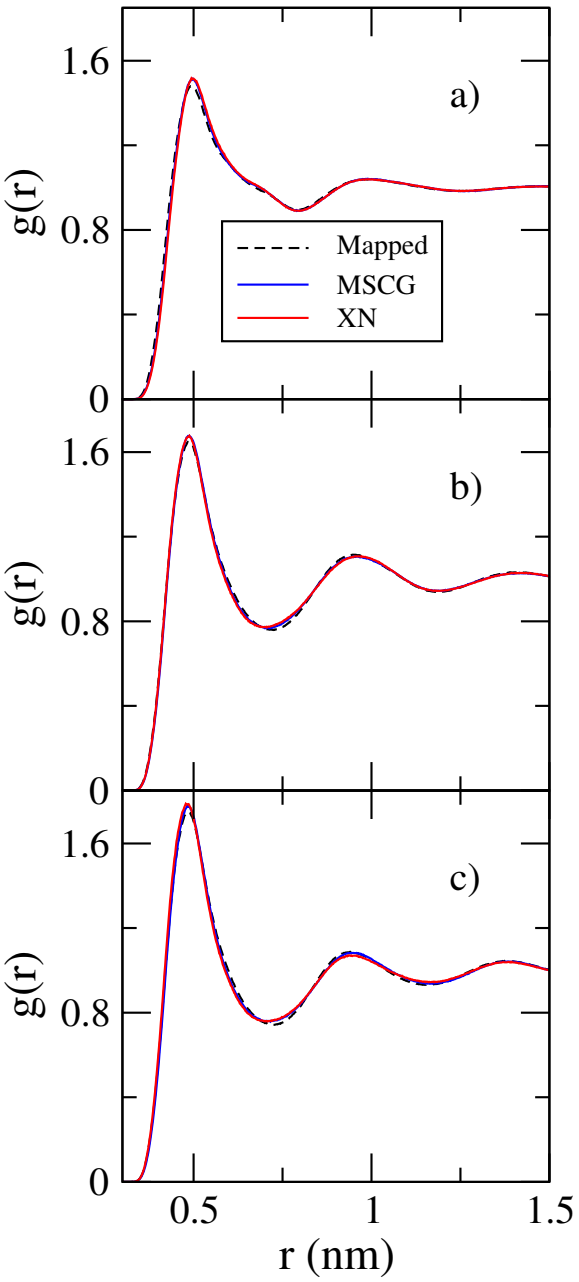

Figure 5.5. Radial distribution functions for the CT-CT pair interactions in a) butane, b) heptane, and c) decane. The dashed black, solid blue, and solid red curves present results for the mapped atomistic ensemble, the system-specific MS-CG model, and the transferable XN model, respectively.

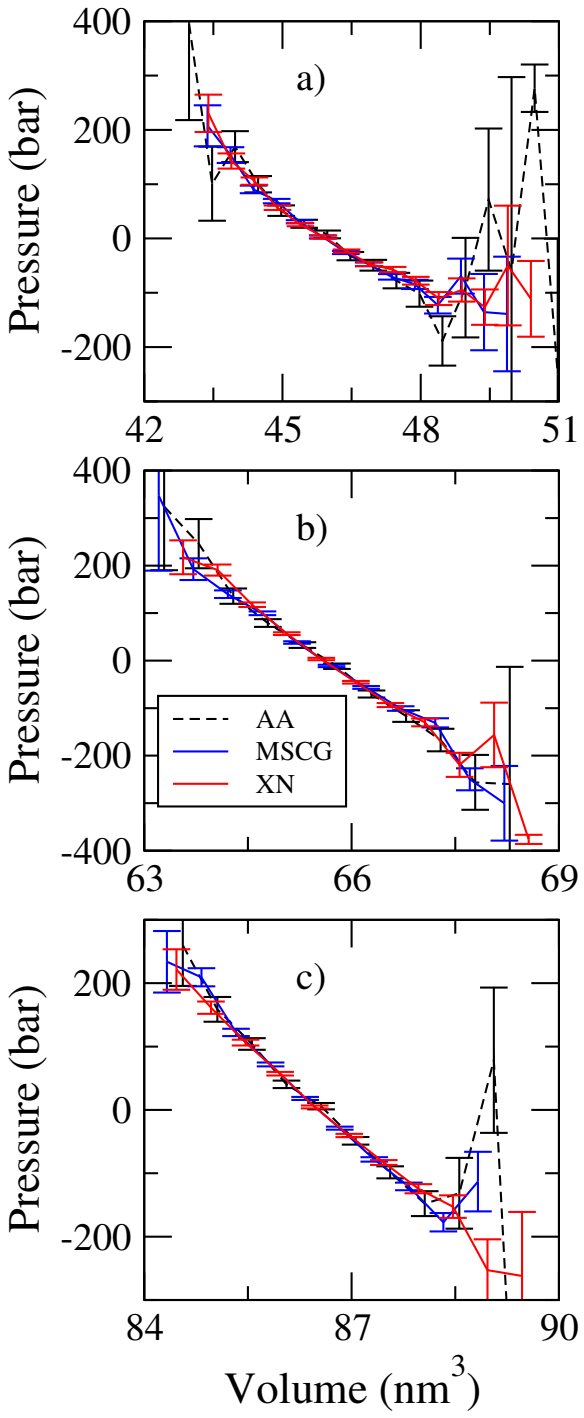

Figure 5.6. Simulated pressure-volume equations of state for a) butane, b) heptane, and c) decane. The error bars indicate the standard error of the corresponding bin. The dashed black, solid blue, and solid red curves present results for the mapped atomistic ensemble, the system-specific MS-CG model, and the transferable XN model, respectively.

the equilibrium structure for each pure liquid. Panels a, b, and c of Fig. 5.5 present the CT-CT nonbonded radial distribution functions for butane, heptane, and decane, respectively. In each panel, the dashed line presents results for the mapped AA ensemble, while the solid lines present results for the CG models. The MS-CG models reproduce the AA CT-CT rdfs with nearly quantitative accuracy. In particular, the MS-CG models describe the asymmetry in the first peak of the AA butane rdf and also accurately reproduce the increasing structure in the rdf that is observed with increasing chain length. Importantly, although the XN models employ the same transferable potentials for modeling each liquid, the XN models also reproduce the AA CT-CT rdfs with nearly quantitative accuracy. The Supporting Information demonstrates that the MS-CG and XN models provide a slightly less accurate, although still very satisfactory, description of the AA rdfs for CM-CM and CM-CT pairs.

Figure 5.7 compares simulated distributions of the radius of gyration, R_{G}, for each pure alkane system. In each case, the MS-CG and XN models generate almost identical distributions. In the case of butane, R_{G} corresponds to the bond between CG sites, which is accurately described by the CG models. In the cases of heptane and decane, the CG models reasonably reproduce the overall shape of the AA distributions and, moreover, reproduce the average R_{G} of the AA models to within approximately 1% error. However, the CG models fail to reproduce the fine details of the AA distributions. In particular, the AA distributions are multimodal with relatively sharp peaks at large R_{G}, which correspond to all atomic torsions sampling trans conformations, and long tails toward more compact conformations. In contrast, the CG distributions are simpler unimodal distributions and, in particular, fail to reproduce the sharp peaks of extended conformations. This discrepancy reflects the tendency of the CG models to sample smaller angles (between triples of bonded sites) than the AA models, as seen in Supporting Figures 7b and 8c. Ultimately, this error reflects the inability of the simple molecular mechanics potential to capture correlations between the bond-stretch and bond-angle in the mapped ensemble.[121,122] Interestingly, as the alkane chains become progressively longer, one expects that the AA distribution will become increasingly simple as more dihedral angles contribute to R_{G} and, consequently, more similar to the CG distribution.

Figure 5.6 presents the average internal pressure of each model as a function of the volume. As a consequence of the iterative self-consistent pressure-matching ap-

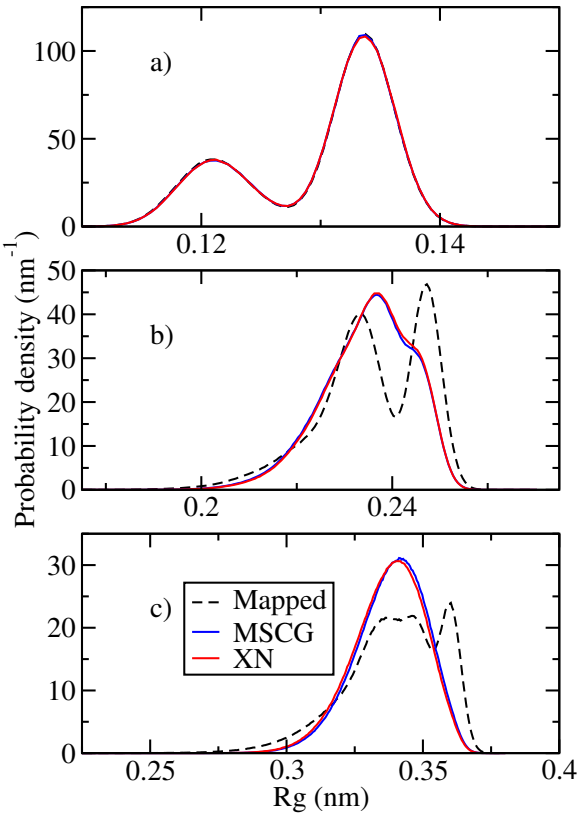

Figure 5.7. Probability distributions for the radius of gyration in a) butane, b) heptane, and c) decane. The dashed black, solid blue, and solid red curves present results for the mapped atomistic ensemble, the system-specific MS-CG model, and the transferable XN model, respectively.

proach, the CG models quantitatively reproduce the AA pressure-volume relations.

We briefly assessed the predictive power of the XN approach by considering a 50:50 butane:decane mixture, which was not considered in parameterizing the XN potential. Figure 5.8 presents the intermolecular CT-CT rdfs obtained from AA simulations and from CG simulations with the XN potential as the dashed black and solid red curves, respectively. Panels a, b, and c of Fig. 5.8 correspond to CT sites from butane-butane pairs, from butane-decane pairs, and from decane-decane pairs. Although the XN potential was parameterized without information about the interactions or packing in butane-decane mixtures, the XN model describes the structure of this mixture quite accurately. The XN model overestimates the AA CT-CT rdf for butane-butane pairs, but almost quantitatively reproduces the AA CT-CT rdfs for butane-decane and decane-decane pairs. The Supporting

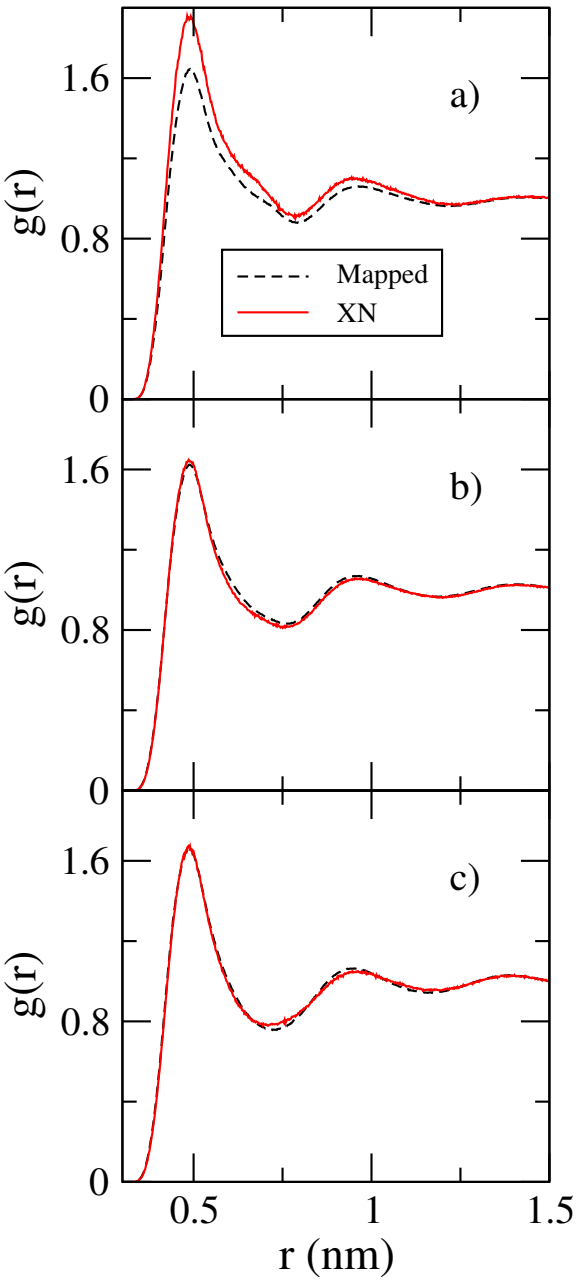

Figure 5.8. CT-CT radial distribution functions in the 50:50 butane-decane mixture for CT sites in a) butane-butane, b) butane-decane, and c) decane-decane pairs. The dashed black and solid red curves present results for the atomistic model and for the extended ensemble CG model, respectively.

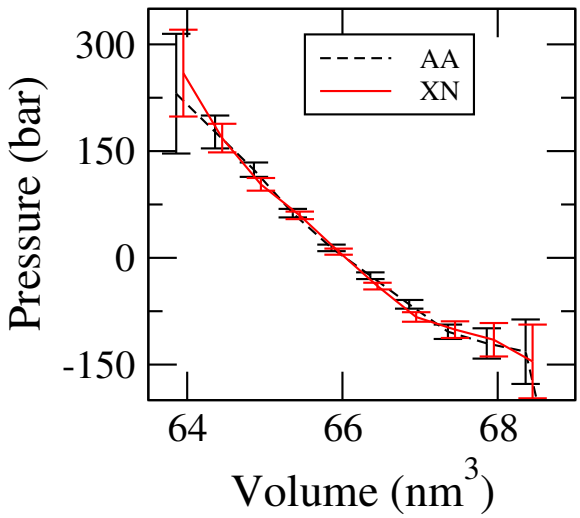

Figure 5.9. Pressure-volume equations of state for 50:50 butane-decane mixture. The error bars indicate the standard error of the corresponding bin. The dashed black and solid red curves present results for the atomistic model and for the extended ensemble CG model, respectively.

Information demonstrates that the XN model also accurately reproduces the AA CT-CM rdf for butane-decane and decane-decane pairs, as well as the CM-CM rdf for decane-decane pairs. Figure 5.9 presents the results of self-consistent pressure matching for this mixture. The CG model accurately reproduces the pressure-volume behavior of the AA model by construction.

In addition to determining the interaction potential, U_R, the cgff tool also characterizes many-body correlations in the mapped AA ensemble and quantifies their contribution to U_R. In order to illustrate these features, we consider a system of 968 methanol molecules. As illustrated in Fig. 5.3d, we represent each methanol with a single site that corresponds to its mass center. We choose this smaller molecule and simpler representation for convenience, since the many-body correlations in the mapped ensemble are then simpler to analyze and interpret. We performed AA simulations for the methanol system in the same manner as described above for the alkane systems, except that the AA production simulation lasted only 5 ns. We did not simulate the resulting CG potential, although previous studies have demonstrated that the MS-CG 1-site model quite accurately describes the structure of liquid methanol.[88]

Panel a of Fig. 5.10 employs Eq. (5.36) to decompose the pair mean force, $-w'(r)$, between methanol molecules into the direct force, $F(r) = \phi(r)$, between

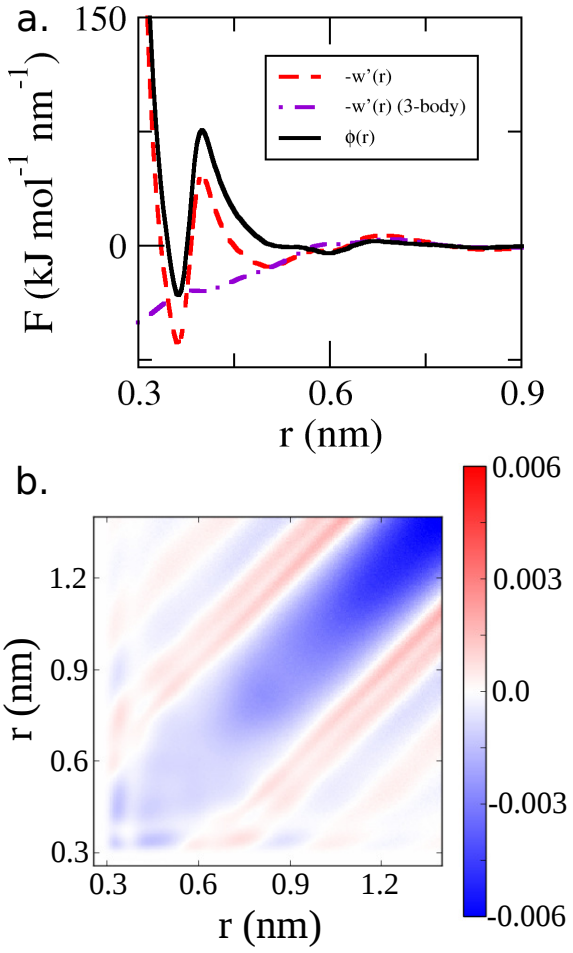

Figure 5.10. a) Contributions to the nonbonded ME-ME pair mean force for methanol. The solid black curve presents the MS-CG pair force, $F(r) = \phi(r)$, that minimizes χ^2_1. The dashed red curve presents the corresponding pair mean force, $-w'(r)$. The dashed-dotted purple curve presents the 3-body (indirect) contributions to the pair mean force. Panel b) presents the 3-body contributions to the metric tensor, $\bar{G}(r, r')$. Red and blue regions indicate positive and negative values, respectively.

the pair and an indirect "three-body" contribution from correlated interactions with other particles in the environment. The pair mean force can be directly calculated from the pair potential of mean force, $w(r) = -k_B T \ln g(r)$, while the direct force is determined via force-matching. The cgff tool uses Eq. (5.36) to decompose the indirect contribution to the pair mean force into contributions from every other type of interaction in the system.[119] Note that the 3-body contribution is attractive at short ranges, indicating that the environment forces particles closer together once the pair approaches 0.6 nm of one another. It is also interesting that the

2-body MS-CG force function includes a relatively large repulsion corresponding to a desolvation barrier near 0.4 nm that is not so pronounced in the pair mean force. This desolvation barrier in the 2-body force function is partially offset by the contributions of correlated interactions from the environment, as described by the the metric tensor, $\bar{G}_{\zeta\zeta'}(r, z)$. We note that, although we included the $v^{2/3}$ rescaling in calculating b_ζ and $\bar{G}_{\zeta\zeta'}$, we find that Eq. (5.35) remains valid to within the numerical precision of the calculations.

Because the one-site CG methanol model considers only one type of interaction, the metric tensor reduces to a single block matrix that depends upon the distances, r and r', of a pair of sites from a single central site. Panel b of Fig. 5.10 presents an intensity plot of this metric tensor, $\bar{G}(r, r')$. As defined by Eq. (5.32), $\bar{G}(r, r')$ describes the contribution to the pair mean force at r from correlations with particles a distance r' away. In particular, $\bar{G}(r, r')$ corresponds to the average cosine of the angle formed between such triplets of particles.[117] Red and blue regions of this intensity plot indicate positive and negative elements of $\bar{G}$, which in turn correspond to acute and obtuse angles between triplets, respectively. As previously described,[117] the negative blue band along the diagonal $r' \approx r$ indicates the tendency of equidistant particles to form obtuse angles due to their excluded volume. The positive red off-diagonal stripes along $r' \approx r \pm \sigma$ correspond to correlated forces arising from molecules in adjacent solvation shells about a central molecule, where σ characterizes the size of the molecules. The alternating red and blue bands moving out from the diagonal reflect the successive solvation shells of methanol molecules.

5.6 Conclusions

We are releasing the BOCS toolkit as open source software for parameterizing bottom-up CG models. As we illustrated for alkane mixtures, the BOCS toolkit provides a robust implementation of both the MS-CG and g-YBG methods for determining interaction potentials. In principle, the g-YBG approach may be used for determining potentials directly from experimentally determined structure ensembles.[171] In this context, the g-YBG framework may prove useful for interpreting and possibly improving the reference states employed in knowledge based potentials that are empirically inferred from known protein structures.[119,217,218] Moreover, the

BOCS toolkit implements an extended ensemble approach for optimizing the transferability of these potentials and also a self-consistent pressure-matching method for accurately modeling isotropic volume fluctuations at constant external pressure. We have recently demonstrated that the resulting volume potential can also be adapted[219] as a function of the local density[220–223] in order to model inhomogeneous systems. Finally, the BOCS toolkit provides unique capabilities for interpreting CG potentials and their relation to many-body correlations in condensed phases.

At the same time, it is worth noting several limitations of the BOCS toolkit. First and most fundamentally, in contrast to iterative methods, such as Iterative Boltzmann Inversion,[28] the Inverse Monte Carlo method,[26] or relative entropy minimization,[29,211] the MS-CG[25,27,88,99,200] and g-YBG methods[31,111] do not guarantee that the CG interaction potential will necessarily reproduce any particular structural features of the underlying mapped ensemble.[32] In practice, the MS-CG and g-YBG models often provide a very good description of intermolecular structure, as illustrated in this work. More generally, though, the structural fidelity of MS-CG and g-YBG models depends upon the adequacy of the approximate potential to account for the relevant many-body correlations in the mapped ensemble.[119,121,122] Consequently, we intend in future work to implement more complex potentials into the BOCS toolkit and also develop more predictive tools for identifying appropriate CG representations. Furthermore, it may be fruitful to develop an iterative wrapper for the cgff tool in order to take advantage of iterative versions of the MS-CG/g-YBG method that can provide improved accuracy for modeling complex structure ensembles.[118,122,128,129] Similarly, while the BOCS toolkit is currently useful for accurately modeling the pressure equation of state for homogeneous systems, we anticipate developing tools for modeling other thermodynamic properties. Additionally, the BOCS toolkit is currently limited by the requirement for simple CG representations in which sites correspond to disjoint atomic groups and by the restriction to systems that are either at constant volume or that sample isotropic volume fluctuations. These limitations clearly motivate future work to further develop the BOCS toolkit. Moreover, in future work we envision implementing more efficient methods for calculating the $G_{DD'}$ matrix, as well as checkpointing methods for saving the results of partial calculations.

Finally, we note that the current version of the BOCS toolkit is incompatible with the most recent versions of GROMACS and LAMMPS, as well as with the

trajectory formats of other MD engines. However, we are currently developing the next version of the BOCS toolkit, which will eliminate all GROMACS dependencies from the cgmap and cgff codebase. Instead, these tools will employ a simpler topology file format and be compatible with both plain text and binary trajectory file formats. These formats can then be readily translated for use with Gromacs2016 or other MD engines. Moreover, we are also developing software for employing barostats with CG pressure corrections in current and future distributions of the LAMMPS package. These developments should significantly extend the utility of the BOCS toolkit.

Nevertheless, despite the aforementioned limitations, we hope that the BOCS toolkit will provide a useful complement to the software already available for developing bottom-up CG models. The source code, as well as documentation and tutorials, for the BOCS toolkit are available for download at `https://github.com/noid-group/BOCS` under the terms of the GPLv3 license.

Appendix

The theory section employs rather abstract notation in order to address a correspondingly general class of interaction potentials, U_R. This appendix provides a more concrete and explicit treatment of U_R for a common molecular mechanics potential with contributions from bond-stretch, bond-angle, dihedral, and pair potentials. The potential in configuration $\mathbf{R}$ may be expressed

$$\begin{aligned} U_R(\mathbf{R}) &= \sum_{\zeta}\sum_{\lambda} U_\zeta(\psi_{\zeta\lambda}(\mathbf{R})) \\ &= \sum_{\alpha}^{\text{bonds}} U^{(b)}_{t_b(\alpha)}(b_\alpha) + \sum_{\alpha}^{\text{angles}} U^{(\theta)}_{t_\theta(\alpha)}(\theta_\alpha) + \sum_{\alpha}^{\text{dihedrals}} U^{(\psi)}_{t_\psi(\alpha)}(\psi_\alpha) + \sum_{(I,J)}^{\text{pairs}} U^{(2)}_{t_2(I,J)}(R_{IJ}). \end{aligned} \tag{5.38}$$

The first term in Eq. (5.38) describes all contributions from bond-stretch interactions. In this first term, α is a label indexing each bond, the sum ranges over all bonds, $t_b(\alpha)$ indicates the type of bond α, $U^{(b)}_{t_b(\alpha)}$ is the bond-stretch potential governing all bonds of type $t_b(\alpha)$, and b_α indicates the length of bond α in configuration $\mathbf{R}$. The second and third sums in Eq. (5.38) describe similar contributions from bond-angles and dihedral angles with α indexing the bond-angles and dihedral

angles, respectively. Finally, the fourth term describes all non-bonded contributions from pair potentials. In this fourth term, (I, J) indicates a particular pair of sites, the sum is performed over all non-bonded pairs, $t_2(I, J)$ specifies the particular non-bonded potential, $U^{(2)}_{t_2(I,J)}$, describing the interaction between the pair, and R_{IJ} is the distance between the pair in configuration $\mathbf{R}$. Given this potential, the force on each site K may be expressed

$$\begin{aligned}
\mathbf{F}_K(\mathbf{R}) &= \sum_\zeta \sum_\lambda F_\zeta(\psi_{\zeta\lambda}(\mathbf{R})) \frac{\partial \psi_{\zeta\lambda}(\mathbf{R})}{\partial \mathbf{R}_K} \\
&= \sum_\alpha^{\text{bonds}} F^{(b)}_{t_b(\alpha)}(b_\alpha) \frac{\partial b_\alpha}{\partial \mathbf{R}_K} + \sum_\alpha^{\text{angles}} F^{(\theta)}_{t_\theta(\alpha)}(\theta_\alpha) \frac{\partial \theta_\alpha}{\partial \mathbf{R}_K} \\
&+ \sum_\alpha^{\text{dihedrals}} F^{(\psi)}_{t_\psi(\alpha)}(\psi_\alpha) \frac{\partial \psi_\alpha}{\partial \mathbf{R}_K} + \sum_{(I,J)}^{\text{pairs}} F^{(2)}_{t_2(I,J)}(R_{IJ}) \frac{\partial R_{IJ}}{\partial \mathbf{R}_K}, \qquad (5.39)
\end{aligned}$$

where $F^{(b)}_{t_b(\alpha)}(x) = -\mathrm{d}U^{(b)}_{t_b(\alpha)}(x)/\mathrm{d}x$ is the bond-force function, while $F^{(\theta)}_{t_\theta(\alpha)}(x)$, $F^{(\psi)}_{t_\psi(\alpha)}(x)$, and $F^{(2)}_{t_2(I,J)}(x)$ are corresponding force functions governing angles, dihedrals, and pair non-bonded interactions, respectively. Each of these force functions is represented by a linear combination of basis functions. For instance, if t_b specifies a particular type of bond governed by the potential function $U^{(b)}_{t_b}$, then the corresponding bond-force function is represented

$$F^{(b)}_{t_b}(x) = \sum_d \phi^{(b)}_{t_b d} f^{(b)}_{t_b d}(x), \qquad (5.40)$$

where d indexes parameters, $\phi^{(b)}_{t_b d}$, that describe the bond force function $F^{(b)}_{t_b}(x)$, while $f^{(b)}_{t_b d}(x)$ indicates the corresponding basis function of a single variable. Similar expansions are adopted for the angle, dihedral, and non-bonded force functions.

Given this expansion the total force on site K may be expressed

$$\begin{aligned}
\mathbf{F}_K(\mathbf{R}) &= \sum_D \phi_D \mathscr{G}_{K;D}(\mathbf{R}) \\
&= \sum_{t_b}^{b-\text{types}} \sum_d \phi^{(b)}_{t_b d} \mathscr{G}_{K;t_b d}(\mathbf{R}) + \sum_{t_\theta}^{\theta-\text{types}} \sum_d \phi^{(\theta)}_{t_\theta d} \mathscr{G}_{K;t_\theta d}(\mathbf{R}) \\
&+ \sum_{t_\psi}^{\psi-\text{types}} \sum_d \phi^{(\psi)}_{t_\psi d} \mathscr{G}_{K;t_\psi d}(\mathbf{R}) + \sum_{t_2}^{\text{pair}-\text{types}} \sum_d \phi^{(2)}_{t_2 d} \mathscr{G}_{K;t_2 d}(\mathbf{R}). \qquad (5.41)
\end{aligned}$$

In Eq. 5.41, the first double sum describes contributions from bond-stretch forces. In this term, the first sum is over all types, t_b, of bonds, while the second sum ranges over the parameters $\phi^{(b)}_{t_b d}$ describing the potential for bonds of type t_b. The corresponding force field basis vectors may be expressed

$$\mathscr{G}_{K;t_b d}(\mathbf{R}) = \sum_{\alpha \in t_b} f^{(b)}_{t_b d}(b_\alpha) \frac{\partial b_\alpha}{\partial \mathbf{R}_K}, \tag{5.42}$$

where the sum is performed over all bonds α of type t_b. The remaining terms represent corresponding contributions from bond-angle, dihedral, and pair potentials.

Chapter 6
Effect of solvent and structure on asphaltene nanoscale aggregation

N. J. H. Dunn, B. Gutama, W. G. Noid, *In-progress manuscript*

6.1 Abstract

We examined the aggregation behavior of model asphaltene compounds under varied solvent conditions via coarse-grained molecular dynamics simulation. The model asphaltenes studied spanned a variety of molecular structures, varying the core flexibility and the aromatic:aliphatic ratio of the molecules. We observed the formation of one-dimensional, rod-like nanoaggregates for those model asphaltenes with a rigid core under solvents that promoted aromatic cohesion. These aggregates were not observed for those model asphaltenes with a flexible core. Further, these rod-like aggregates were observed to form more readily in mixed solvents that promoted both aromatic and aliphatic cohesion than in solvents that only promoted aromatic cohesion. Both core types were observed to form large, disordered aggregates in solvents that promoted aliphatic cohesion, with those molecules with longer tails aggregating more readily under these conditions. These results support the Yen-Mullins model for asphaltene nanoaggregation behavior.

6.2 Introduction

Asphaltenes, known as the 'cholesterol of petroleum,'[50] are a problematic class of molecules found in crude oil that have a propensity for aggregating and coating the surface of oil processing and transport equipment. These molecules are thought to form colloidal suspensions in oil that flocculate into viscoelastic masses when destabilized by a change in the processing conditions.[45,61] These floccs are an expensive issue for the petroleum industry due to their ability to clog pipes and foul heat exchanging equipment.[47–49] It is estimated that asphaltene flocculation costs the petroleum industry billions of dollars a year.[50–52]

The precise molecular properties of asphaltenes have been a topic of extensive debate.[46,47,58–60] Asphaltenes are defined as the fraction of crude oil that is soluble in toluene, and insoluble in *n*-heptane.[47,50,58,224] Due to this definition as a solubility class, there are a wide variety of molecular species represented in any given asphaltene sample, making it difficult to determine specific information about the molecular structures that are present.[45,53–57] Further complicating matters, the high propensity of asphaltenes to aggregate means that it can be difficult to differentiate between the properties of single molecules and those of aggregates.[225] Two main views have emerged for describing the average asphaltene molecule:[45,226] continental (or island) asphaltenes are proposed to have a single aromatic core surrounded by alkyl tails, while archipelago-type asphaltenes are proposed to have several aromatic cores linked by alkyl chains. Recently, the continental model has grown in popularity among asphaltene researchers,[227–231] although there is significant evidence to suggest that both archipelago and continental structures are represented in the asphaltene fraction.[232,233]

The Yen-Mullins (or modified Yen) model has emerged as a leading theory that describes both the average asphaltene molecule and asphaltene aggregation.[57,61] The prototypical asphaltene molecule proposed by this model is an approximately 750 g/mol continental-type structure composed of an aromatic core surrounded by alkyl chains.[61] These molecules are proposed to aggregate via a hierarchical mechanism, where 7-10 molecules form a nanoaggregate, and these nanoaggregates then cluster together into larger particles.[57,61] The nanoaggregates proposed by the Yen-Mullins model are formed by stacking the aromatic cores against each other to form short rod-like structures, which then cluster together into a fractal

arrangement.[57,61,234,235] Nanoaggregation is thought to begin around 100 mg/L of asphaltenes in oil,[62–67] while clustering begins around 2-5 g/L.[68–70] This model is supported by experimental studies employing such varied techniques as two-step laser desorption ionization,[227] NMR,[65,228,229] alternating and direct current conductivity,[62–64] and centrifugation.[66,67]

Due to the difficulty in using experimental techniques to characterize molecule-level details of asphaltene behavior, molecular simulation is often used to study the behavior of proposed model asphaltenes. Most commonly, all-atom (AA) molecular dynamics is the methodology of choice for such simulation studies. These AA simulation studies have explored the nanoaggregation behavior[236,237] of various model asphaltenes under different solvents,[49,54,238,239] at interfaces,[47,240–243] and under different temperatures and pressures.[56,237] However, these studies are necessarily limited in size and in scope by the computational complexity of simulating an AA system of solvated asphaltenes at a realistically low concentration, although recent advances in molecular dynamics on GPUs are extending these limits.[244] At these low concentrations, the majority of the computation time is spent on simulating the solvent, as the solvent molecules greatly outnumber the asphaltenes.

Coarse-grained (CG) modeling is one approach for extending the accessible scale of asphaltene simulation. CG models represent a target system at a reduced resolution by mapping out atoms from an AA model.[245–247] This can provide a significant simulation speedup relative to an AA model, enabling larger scale simulation studies that can explore physically larger systems and/or multiple state points. Such CG models have been used in the context of asphaltene-like molecules to study nano-aggregation,[248] mesoscale clustering behavior,[249,250] the effect of solvent dependence,[249,251] and to extend simulations to the microsecond time scale.[48,249–252]

In this work, we use an implicit solvent CG toy model to explore the nanoaggregation behavior of several model asphaltene compounds. This model captures the quality of the solvent for the model asphaltenes by scaling the attractive contribution to the nonbonded pair potentials. The simplicity of this model allowed us to simulate two replicates each of 106 distinct systems, for a total of over 80 μs of simulation without accounting for dynamical speedup from coarsening. This wide sweep of solvent conditions for several molecular structures allows us to charac-

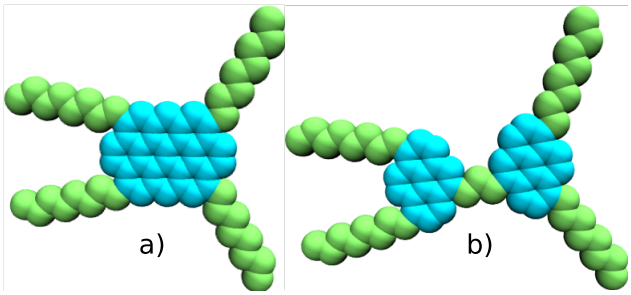

Figure 6.1. Structures of the model asphaltene compounds. The green and cyan sites correspond to tail and core sites, respectively. a) has an ovalene-type core, while b) has a bipyrene-type core. Both molecules shown have tail lengths of eight sites.

terize the impact of molecular structure on the solubility profile of these model asphaltenes, and the effect of the solvent quality on the structure of the resulting nanoaggregates.

The remainder of the paper is organized as follows. Section 6.3 provides the key details of our simulation methods and characterization metrics. Section 6.4 presents an analysis of the resulting aggregation behavior. Section 6.5 discusses these results in the context of the Yen-Mullins model and other recent studies. Finally, Section 6.6 summarizes the main conclusions of our work and suggests possible future work in this area.

6.3 Methods

6.3.1 Coarse-Grained Model

We used toy CG models of several model asphaltene compounds to simulate asphaltene aggregation phenomena. In particular, we used six distinct model asphaltene structures - three with ovalene-type cores, and three with bipyrene-type cores, as seen in Figure 6.1. For each of these two core types, we used variant molecules with tail lengths of five, eight, and eleven sites in order to vary the aromatic to aliphatic ratio for these molecules. These structures are indicated with the notation ovalene-N or bipyrene-N, where N denotes the length of the tails. This representation places a single CG site at the location of each carbon atom in a corresponding AA molecule, and omits the hydrogen atoms. Tail (green) sites correspond to aliphatic sp3 carbons, while core (teal) sites correspond to aromatic

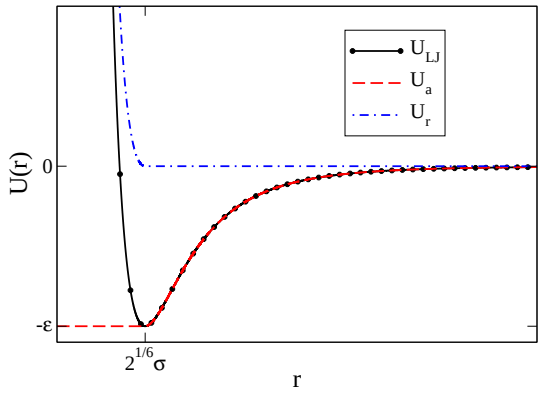

Figure 6.2. Weeks-Chandler-Anderson contributions to the reduced nonbonded potential. The form of the potential is given in black, while the attractive and repulsive components are shown in red dashed and blue dashed-dotted curves, respectively.

sp2 carbons. All bond, angle, and dihedral interactions are maintained from the corresponding atom types in the OPLS-AA force field. This makes the cores of the model asphaltenes rigid and planar, while their tails remain flexible. These two core types were chosen in order to study the impact of internal flexibility of the cores of model asphaltenes on their propensity to form Yen-Mullins-type nanoaggregates.

Importantly, these CG models do not explicitly include any solvent. Instead, the nonbonded pair potentials are modified in order to capture the effect of solvent quality on the different site types. The functional form of the pair potential used is the Weeks-Chandler-Anderson (WCA) decomposition[253] of the standard Lennard-Jones 12-6 potential

$$U_{LJ}(r) = 4\varepsilon \left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6} \right]. \tag{6.1}$$

The WCA decomposition splits U_{LJ} into purely attractive and repulsive components, as shown in Figure 6.2. The attractive contribution is given by

$$U_a(r) = \begin{cases} -\varepsilon & \text{if } r < 2^{1/6}\sigma \\ U_{LJ}(r) & \text{if } r \geq 2^{1/6}\sigma \end{cases} \tag{6.2}$$

while the repulsive contribution is given by

$$U_r(r) = \begin{cases} U_{LJ}(r) + \varepsilon & \text{if } r < 2^{1/6}\sigma \\ 0 & \text{if } r \geq 2^{1/6}\sigma. \end{cases} \tag{6.3}$$

We defined distinct values of ε for core-core (ε_R) and tail-tail (ε_L) interactions, while core-tail interactions were determined using the mixing rule $\varepsilon_{RL} = \sqrt{\varepsilon_R \varepsilon}_L$. We kept the value of σ constant at 0.3525 nm for both core and tail sites. This nonbonded potential was also applied to intramolecular nonbonded interactions between sites separated by three or more bonds. Under this framework, a solvent is defined by a pair of values ε_R,ε_L, indicating the quality of the solvent for aromatic (core) and aliphatic (tail) components of the model asphaltenes. Each molecule has a set of possible solvents (solvent space) defined by the possible combinations of ε_R,ε_L. A poor solvent has $\varepsilon > 0$ and encourages aggregation through attractive nonbonded interactions, while an ideal solvent has $\varepsilon = 0$, resulting in sites that have purely repulsive nonbonded interactions. For ease of notation, we introduce the scaled unitless $\bar{\varepsilon}_R = \varepsilon_R / k_B T$ and $\bar{\varepsilon}_L = \varepsilon_L / k_B T$.

6.3.2 Coarse-Grained Simulations

We simulated all CG models in the NVT ensemble under full periodic boundary conditions with temperature T=303 K, enforced by a stochastic dynamics thermostat[254] with a coupling constant of 0.5 ps. We selected constant volumes for each model compound such that the resulting system would have a fixed 1.0 wt% (6.8 g/L) asphaltenes, assuming the implicit solvent was pure heptane. Each simulation contained 25 model asphaltene molecules of a single type. We performed these simulations using GROMACS 4.5.3[255] compiled with double floating point precision. Each simulation used a 2 fs time step and ran for up to 400 ns or until it had reached a stable state lasting at least 50 ns. We implemented the nobonded potentials as GROMACS table files that separated the attractive and repulsive contributions into columns that could be independently parameterized according to the solvent quality of the simulation. These table files were constructed such that the force goes smoothly to zero at the cutoff of 1.2 nm. None of the CG models included electrostatic interactions.

We selected the solvent conditions to simulate for each molecule by qualitatively

exploring solvent space and observing the aggregation behavior at each state point. We selected new state points in order to determine the boundaries of different behavior regimes. We performed a total of 212 simulations, covering 106 state points. For each of these state points we ran a pair of simulations, one with a starting configuration where the molecules were placed on a regular grid within the simulation box, and one where the starting configuration was taken from the end of a short simulation with purely repulsive nonbonded potentials. Unless otherwise specified, all metrics reported for a system are averaged between these replicate simulations.

Simulations performed with these models enjoy a dynamical speedup relative to a corresponding AA model due to a relative smoothing of the free energy landscape.[9] While there is no AA simulation that corresponds exactly to these CG simulations due to our treatment of the solvent, comparing the self-diffusion constant of our model asphaltenes to AA simulations of similar molecules diffusing through heptane and toluene[249] estimates that our model has a speedup factor between 1.7 and 4.3. Durations reported in the text of this paper are not scaled by this speedup factor and reflect only the GROMACS timesteps taken.

6.3.3 Aggregate Analysis

In order to characterize the aggregation behavior for these model asphaltenes, we adopted the three distance criteria defined by Wang et al.[249] for aggregate membership. These metrics are briefly summarized below.

6.3.3.1 Minimum Distance

The minimum distance criteria considers two molecules A and B to be part of the same aggregate if any site on A is within $r_{min} \leq$0.49 nm of any site on B. This cutoff was selected as $1.25*(2^{1/6}\sigma)$, based on the position of the attractive well in the total nonbonded potential.

6.3.3.2 Minimum Core Distance

The minimum core distance criteria considers two molecules A and B to be part of the same aggregate if any core site on A is within $r_{core} \leq$0.49 nm of a core site on B. This cutoff is again based on the position of the attractive well in the total

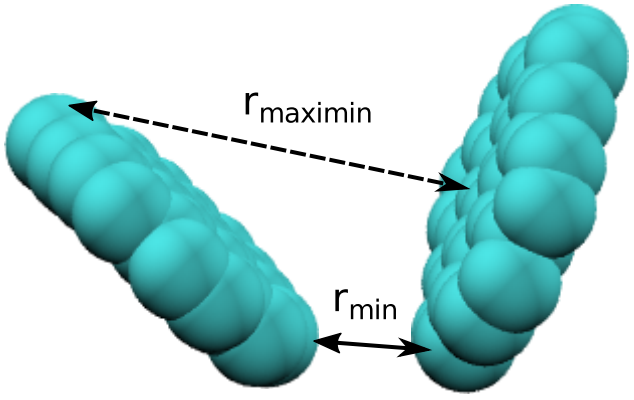

Figure 6.3. Demonstration of the difference between r_{min} (solid black) and $r_{maxmimin}$ (dashed black) distances for a pair of ovalene cores. Tail sites are hidden for clarity.

nonbonded potential. Tail sites are ignored for the purposes of this criterion. The minimum core distance metric is used in conjunction with r_{min} to help distinguish between aggregates that are driven by core-core attraction and those that are driven by tail-tail or core-tail attraction.

For bipyrene-type molecules that have two cores each, there are four values of r_{core} between the two molecules. In this case, the minimum of these values is used to determine whether the molecules belong to the same aggregate.

6.3.3.3 Maximin Core Distance

The maximin core distance is a metric that captures an element of the core-core orientation between a pair of molecules. For a pair of molecules A and B with core sites indexed by a and b, this distance can be defined as

$$r_{maximin} = \max\left[\left(\max_{a\in A}(\min_{b\in B}\ r_{ab})\right)\left(\max_{b\in B}(\min_{a\in A}\ r_{ab})\right)\right]. \tag{6.4}$$

Algorithmically, this metric can be calculated by 1) finding the minimum distance between each core site a to any core site b, 2) taking the maximum of those distances to be $r_{maximin,1}$, 3) finding the minimum distance between each core site b to any core site a, 4) taking the maximum of those distances to be $r_{maximin,2}$, and finally 5) take $r_{maximin} = \max\left(r_{maximin,1}, r_{maximin,2}\right)$.

This distance is illustrated in Figure 6.3. When $r_{maximin}$ is small, the cores are roughly parallel to each other and aligned face-to-face, indicating the presence of ordered core-core stacking within an aggregate. Systems that count aggregates

using this as the cutoff metric have contiguous, aligned stacks of cores, similar to the Yen-Mullins-type nanoaggregate. Nanoaggregates that have a skewed angle between cores or are stacked in an offset arrangement will not be counted under this metric. As in the case of r_{core}, there are four values of $r_{maximin}$ between a pair of bipyrene-type molecules, and the minimum of these values is used to determine aggregation.

6.3.3.4 Extent of Aggregation

We used the mass-averaged aggregation number ${g_2}^{256}$ as an indicator for the extent of aggregation in each system, where

$$\langle g_2(t) \rangle = \left\langle \frac{\sum_i n_i(t) i^2}{\sum_i n_i(t) i} \right\rangle . \tag{6.5}$$

Here t indicates a timestep, i is the size of an aggregate, and $n_i(t)$ is the number of aggregates of size i at time t. The angular brackets indicate an average over the set of equilibrated frames from both replicates of the system. For our systems, this metric can vary from 1 for a completely non-aggregated system to 25 for a system with a single aggregate containing all 25 molecules.

The value of g_2 varies depending on the distance metric used to determine aggregation. The different g_2 metrics are labeled as G_{min}, G_{core}, and $G_{maximin}$ for the minimum distance, minimum core distance, and maximin aggregation criteria.

6.3.3.5 Intra-Aggregate Alignment

We used the second Legendre polynomial of the cosine of the angle between cores to further characterize the structure of aggregates:

$$\langle P_2(cos\theta) \rangle = \left\langle \frac{1}{2} \left(3cos^2\theta - 1 \right) \right\rangle_{intra} \tag{6.6}$$

Here, θ is the angle between the normal vectors of the planes formed by the molecule cores. The subscripted angular brackets indicate an average over all intra-aggregate pairs for all equilibrated time steps for both replicates of the system. This metric will range from -0.5 for systems that prefer a perpendicular core-core alignment to 1.0 for systems that prefer parallel (or antiparallel) core stacking. A system with no strong preference for the orientation between cores will show 0 for this metric.

Each segment of the core of a bipyrene molecule contributes independently to this average, giving a total of four contributing core orientations. Further, the intramolecular alignment between the core segments in a single bipyrene molecule also contributes to this average, but only when that molecule is part of an aggregate containing at least two distinct molecules.

As in the case of g_2, the different distance metrics lead to different values of $\langle P_2(cos\theta)\rangle$. The same subscripts are used with this metric as for g_2 to indicate which distance metric was used to determine aggregation.

6.4 Results

In the following, we examine the aggregation behavior of the CG model asphaltenes under a range of implicit solvent conditions. We first examine the aggregation behavior of the ovalene-8 and bipyrene-8 model compounds in a set of representative solvents. Then, we focus on the behavior of ovalene-8 to introduce key features of the solvent phase diagrams. Finally, we expand our scope to consider the behavior of all six model compounds to see the impact of both the core flexibility and the ratio of aromatic to aliphatic sites on aggregation behavior.

Figure 6.4 shows examples of what representative ovalene-8 (left column) and bipyrene-8 (right column) aggregates look like under selected solvent conditions. The top row corresponds to a solvent that strongly promotes aromatic cohesion, the middle row to a solvent that promotes both aliphatic and aromatic cohesion, and the bottom row to a solvent that strongly promotes aliphatic cohesion. Considering the left-hand (ovalene-8) column first, we see strong columnar stacking in both the strong aromatic cohesion case (Fig. 6.4a) and the mixed attraction case (Fig. 6.4c). Both of these aggregates closely resemble the prototypical Yen-Mullins nanoaggregate,[57,61] with an ordered, columnar core composed of stacked aromatic segments surrounded by a layer of aliphatic tails. The aggregate formed under strong aliphatic cohesion shown in Figure 6.4e has a markedly different structure, with no apparent preference for stacking the aromatic core segments of the molecules. Now considering the right-hand (bipyrene-8) column, the representative aggregates shown in Figures 6.4b and 6.4d are smaller than the corresponding aggregates in the ovalene-type systems and demonstrate no clear core stacking. In contrast, the aggregates formed under strong aliphatic cohesion in Figure 6.4f are similar to

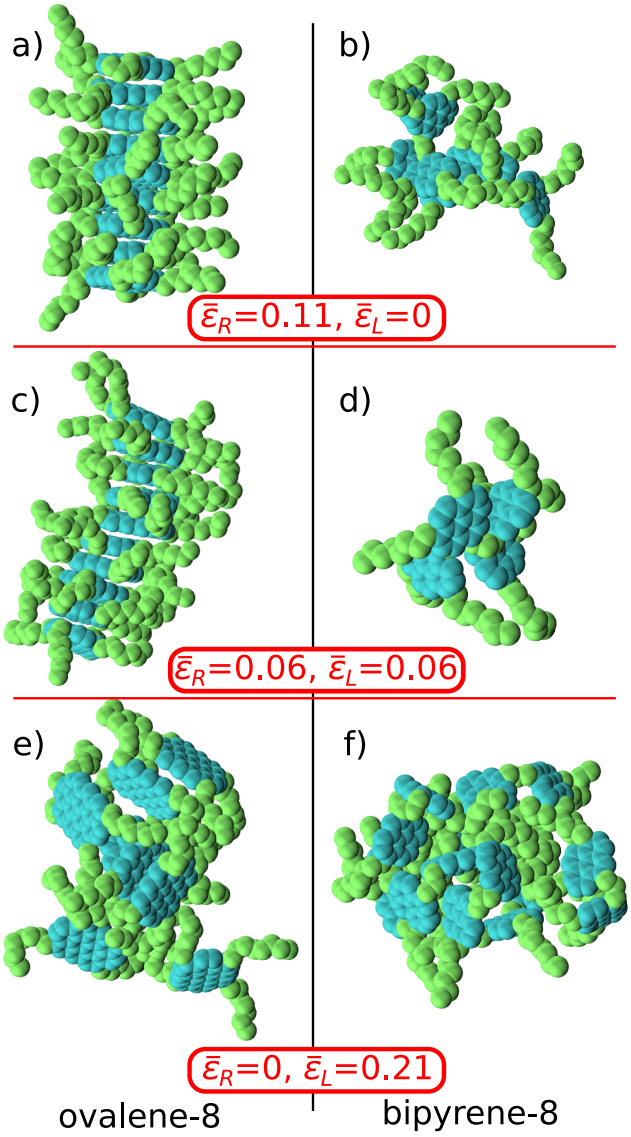

Figure 6.4. Representative aggregates from selected solvent conditions. The solvents shown are a+b) $\bar{\varepsilon}_R = 0.11$, $\bar{\varepsilon}_L = 0$, c+d) $\bar{\varepsilon}_R = 0.06$, $\bar{\varepsilon}_L = 0.06$ and e+f) $\bar{\varepsilon}_R = 0$, $\bar{\varepsilon}_L = 0.21$. Ovalene-8 and bipyrene-8 molecules are shown in the left and right columns, respectively.

the corresponding ovalene-type system. This difference in aggregate structure in aromatic-driven and mixed solvents is evidently due to the internal flexibility of the bipyrene-type cores that allows these molecules to take on different conformations that may not be conducive to contact between the core sites. In contrast, this core flexibility has little effect on the structure of aggregates driven by cohesive interactions between the aliphatic tails as seen in Figures 6.4e and 6.4f.

Figure 6.5 plots the probability that a molecule will belong to an aggregate

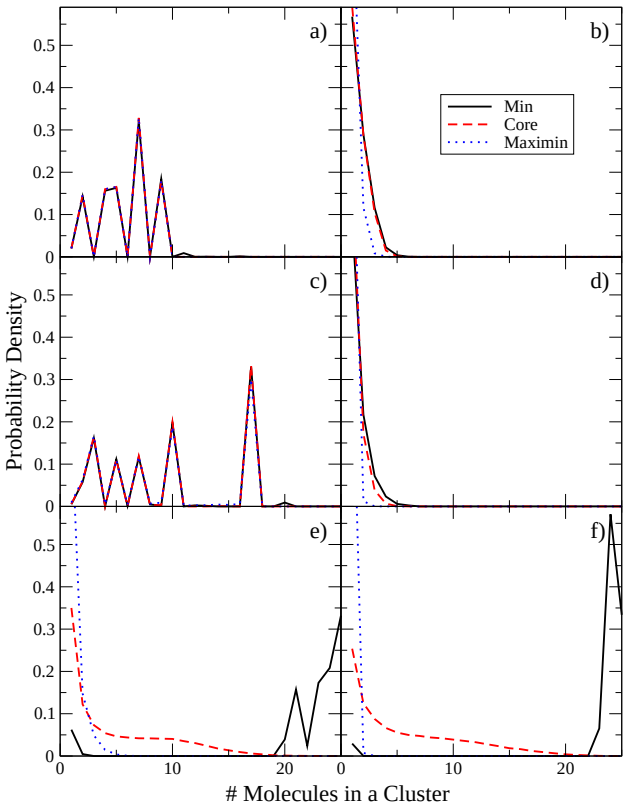

Figure 6.5. Distributions for the probability that a molecule will belong to an aggregate of size N for representative ovalene-8 and bipyrene-8 systems. The panels are laid out to correspond to the organization of molecule types and solvents in Fig 6.4.

of size N for each of the systems represented in Figure 6.4. The structure and solvent represented in each panel in Figure 6.5 are the same as in the corresponding panel in Figure 6.4. The different curves in each panel correspond to the different aggregation metrics. Figure 6.5a shows this probability distribution for ovalene-8 in a solvent that promotes core-driven aggregation. The jagged nature of the curves indicate that this system exists in a state where monomers are not able to freely join and leave a set of aggregates of roughly constant size. We emphasize here that these curves are averaged between two independent replicate simulations, and that these curves qualitatively agree between the replicates (see SI 1). Further, the flat nature of the time traces of each aggregation metric G (see SI 2) demonstrates that the simulations have reached a state of local stability. This suggests that these panels reflect metastable states that are a result of the molecular structure and the solvent conditions, rather than occurring by chance.

The low probability of belonging to an aggregate of size one indicates that monomer dissociation and addition is rare compared to the formation of small aggregates that later group into larger ones. This may be due to the small number of monomers (N=25) in the system compared to the cohesive strength between molecules. This contrasts with bipyrene-8 in the same solvent (Fig 6.5b), which shows a relatively high probability of membership in an aggregate of size one that smoothly decays to zero with increasing aggregate size. This smooth curve indicates that there is sufficiently weak cohesion between molecules that it is possible for monomers to join and leave an aggregate, and that the system is at equilibrium with respect to its aggregation state. The probability distributions for a solvent that promotes mixed cohesion (Figs 6.5c and 6.5d) are qualitatively similar to the core-driven case, except that larger aggregates form on average in the mixed case. This is likely due to the addition of attractive interactions between the tail sites in the mixed case allowing aggregates to coalesce after forming, while the tail sites act as purely repulsive barriers in the core-driven case.

Moreover, the probability curves for each aggregation criterion align completely ($r_{min} \approx r_{core} \approx r_{maximin}$) for the ovalene cases in Figures 6.5a and 6.5c. This corresponds to aggregate structures where the core sites of different molecules are tightly packed together, and the cores are roughly parallel. Similarly, the bipyrene systems in Figures 6.5b and 6.5d have min and core curves that closely align, indicating that aggregates are often joined by contact between core molecules. However, the curve corresponding to aggregation by the maximin distance shows a consistently lower probability of aggregate membership, indicating a lack of strong core-core alignment in these systems. It is important to note that the metrics reported for bipyrene-type molecules reflect the minimum distance from the four pairs of cores.

The solvents promoting tail-driven aggregation in Figures 6.5e (ovalene-8) and 6.5f (bipyrene-8) are distinct from the other systems, but similar to each other. In both cases the probability curves defined by core-driven metrics demonstrate a high probability of membership in a monomer and a smooth decay to zero, while the probability curve defined by the global minimum distance demonstrates a peak for large aggregates and a relatively low probability of membership in a monomer. This suggests that the aggregates exist in metastable states, and that the aggregates are held together by a network of aliphatic sites with only incidental contact between

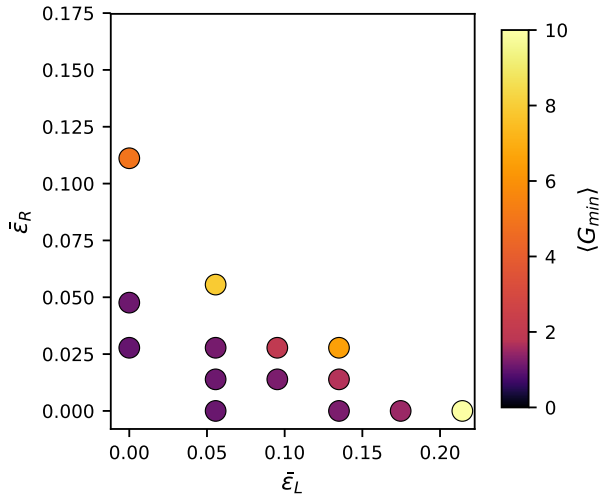

Figure 6.6. Phase diagram of $\langle G_{min} \rangle$ for the ovalene-8 model as a function of core site (horizontal) and tail site (vertical) affinity. This affinity is given by the depth of the site-site attractive potential ε scaled by $k_B T$ to give unitless axes. The colors of the circles indicate the value of $\langle G_{min} \rangle$ for the solvent conditions corresponding to the location of the center of the circle. Brighter colors indicate a higher value of $\langle G_{min} \rangle$.

core sites. Given sufficient time, it seems likely that molecules in these systems would coalesce into a single, large aggregate. Further, the relatively low probability of membership in aggregates of size $N > 1$ by the maximin metric indicates that there is no strong tendency towards aligning or stacking the aromatic cores in these systems. This is consistent with the representative aggregates shown in Figure 6.4.

We use phase diagrams across solvent space to demonstrate the impact of the solvent conditions on the aggregation behavior of the model asphaltenes. Figure 6.6 is such a phase diagram for ovalene-8, plotting the average extent of aggregation ($\langle G_{min} \rangle$) as the point color against the quality of the solvent for aliphatic sites (x-axis) and for aromatic sites (y-axis). The origin of the diagram is an ideal solvent where the site-site potentials have an attractive well depth $\varepsilon = 0$, and solvent quality decreases with distance from the origin. Darker points correspond to a low extent of aggregation, while brighter points correspond to a higher extent of aggregation. Points near the origin display low extents of aggregation as expected for the ideal and nearly ideal solvents in this region of the phase diagram. We observe an increased propensity for aggregation as the solvent quality decreases away from the origin.

Looking at the set of solvents along $\bar{\varepsilon}_L = 0$, we observe the first solvent promoting

significant aggregation at $\bar{\varepsilon}_R = 0.11$. In contrast, along $\bar{\varepsilon}_R = 0$ the first solvent promoting significant aggregation is at $\bar{\varepsilon}_L = 0.21$. From this, we observe that ovalene-8 aggregates readily for relatively small values of $\bar{\varepsilon}_R$ compared to $\bar{\varepsilon}_L$. This demonstrates that for this model asphaltene, attraction between the aromatic sites is a stronger driving force towards aggregation than attraction between the aliphatic sites. The different arrangement of the two site types is expected to cause this difference. In particular, the rigid, planar core geometry allows for all of the core sites to simultaneously interact without incurring much of an entropic penalty, but the flexibility and linear geometry of the tails means that there is a strong entropic penalty for the tails to align. Additionally, slightly increasing either $\bar{\varepsilon}_R$ or $\bar{\varepsilon}_L$ decreases the aggregation threshold for the other, suggesting that aromatic-aliphatic interactions may also be important for promoting aggregation formation.

We can investigate the impact of different core types as well as different tail lengths by comparing such phase diagrams between the different model molecules. Figure 6.7 shows the extent of aggregation by plotting phase diagrams of $\langle G_{min} \rangle$ over different solvent conditions for each molecule type. Here, ovalene-type cores are on the left, while bipyrene-type cores are on the right. From top to bottom, the rows correspond to tails of length 5, 8, and 11 sites. The colors correspond to the value of $\langle G_{min} \rangle$, with brighter colors corresponding to larger values and more highly aggregated systems. Two overall trends emerge from this view. First, consider the set of solvents along $\bar{\varepsilon}_L = 0$ that promote aromatic cohesion for each molecule type. In each case, aggregation onsets at a lower value of $\bar{\varepsilon}_R$ for the ovalene-type molecules than for the bipyrene-type molecules. This demonstrates that the flexible bipyrene-type cores do not aggregate as readily in solvents that promote aromatic cohesion. Second, consider the solvent at ($\bar{\varepsilon}_L = 0.21$, $\bar{\varepsilon}_R = 0$) for each system. For those molecules with tail lengths of 5 (panels 6.7a and 6.7b), there is no appreciable aggregation in this solvent. However, as the tail length increases to 8 and 11, there is significant aggregation observed in this solvent. Intuitively, increasing the amount of aliphatic sites in a molecule type enhances the propensity of that molecule type to aggregate in solvents that promote aliphatic-driven aggregation. Interestingly, for $\bar{\varepsilon}_R = 0$, the bipyrene- and ovalene-type molecules behave similarly. This reflects the fact that in both cases the tails are flexible and evenly arranged about a central core, and the internal flexibility of the bipyrene core does not strongly impact the

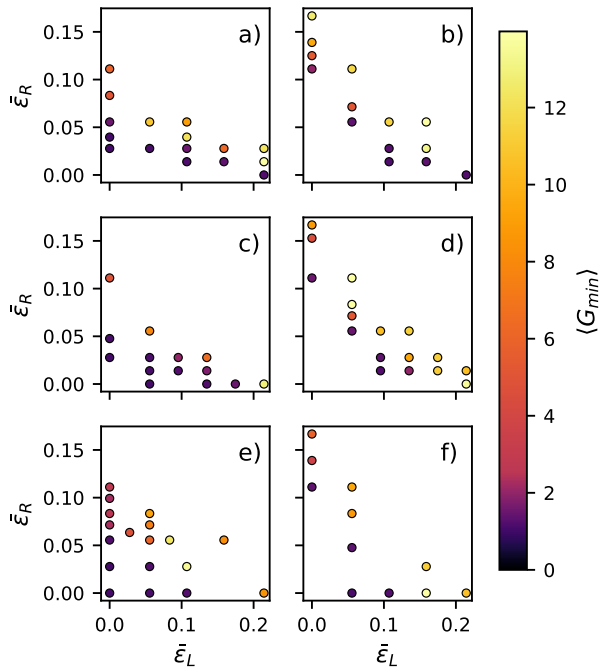

Figure 6.7. Phase diagrams of $\langle G_{min} \rangle$ for all model asphaltenes. The panels correspond to a) ovalene-5, b) bipyrene-5, c) ovalene-8, d) bipyrene-8, e) ovalene-11, and f) bipyrene-11. The color scheme is the same as in Fig 6.6. Note, however, the different maxima for the color bars.

relative positions of the tails.

Differences in aggregate structure can also be observed from this type of comparison. Figure 6.8 shows phase diagrams of the core alignment metric $\langle P_2(cos\theta) \rangle_{min}$. The panel layout is the same as in Figure 6.7. Here, brighter colors indicate a larger value of $\langle P_2(cos\theta) \rangle_{min}$ and a higher degree of core alignment within an aggregate, while darker colors indicate more disordered aggregates. The most general trends to emerge are that all molecules form more ordered aggregates in solvents that promote core-driven or mixed aggregation, and that this trend is stronger for the ovalene-type molecules than for the bipyrene-type molecules. Indeed, there is only a weak core alignment effect present for the bipyrene-type cores under any solvent conditions. Interestingly, there is an onset of core alignment for ovalenes with $\bar{\varepsilon}_R > 0.5$, and for smaller $\bar{\varepsilon}_R$ there is little or no alignment observed. Further, this threshold appears to be slightly smaller for molecules with shorter tails, reflecting an entropic repulsion effect from the tails that increases with their length.

Figure 6.9 shows phase diagrams of the extent of core-driven aggregation via

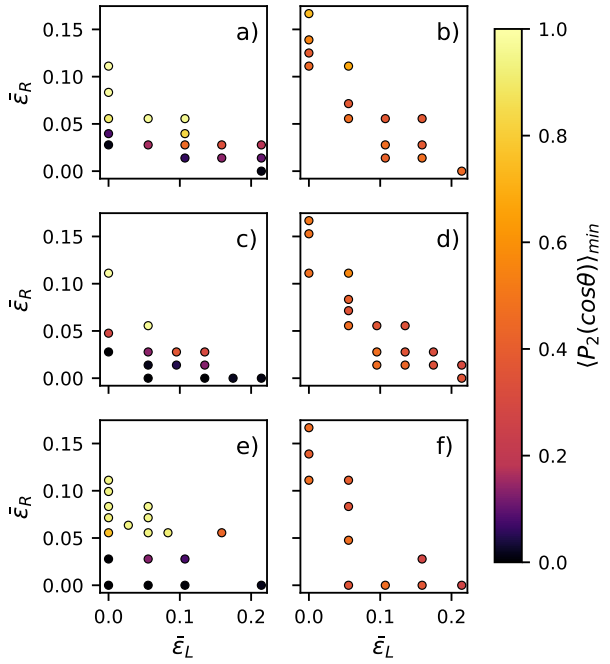

Figure 6.8. Phase diagrams of $\langle P_2(cos\theta)\rangle_{min}$ for all model asphaltenes. The panels are laid out to correspond to the organization of molecule types in Fig 6.7. Brighter colors indicate a higher value of $\langle P_2(cos\theta)\rangle_{min}$. Note, however, the different maxima for the color bars.

$\langle G_{maximin}\rangle$. As in the other aggregation phase diagrams, warmer colors indicate a higher extent of aggregation. This figure uses the maximin metric to highlight solvent regions that promote the formation of the ordered aggregates of size 7-10 predicted by the Yen-Mullins model.[57,61] The panels are laid out in the same configuration as in Figure 6.7. Note that none of the bipyrene-type molecules demonstrate significant aggregation under this metric for any solvent conditions. However, the ovalene-type molecules demonstrate significant ordered aggregation in solvents promoting strong aromatic cohesion, as well as in mixed solvents. Mixed solvents promote this type of aggregate at relatively lower values of $\bar{\varepsilon}_R$ compared to those promoting only aromatic cohesion, likely due to the tails' role as purely repulsive barriers in the aromatic-promoting solvent. This suggests that Yen-Mullins-type aggregation is most favorable when the solvent is moderately poor for both aromatic and aliphatic components of the model asphaltenes, or when the solvent is extremely poor for aromatic components. It is notable that stacks larger than 7-10 molecules are only rarely observed for the solvent conditions sampled here.

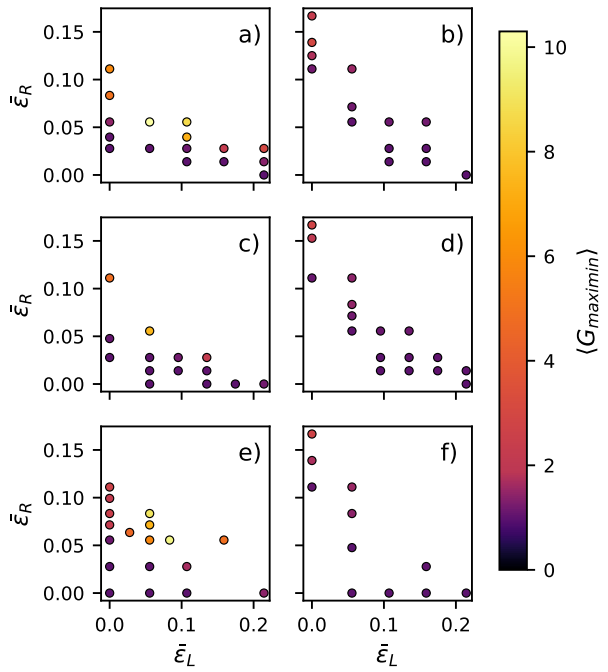

Figure 6.9. Phase diagrams of $\langle G_{maximin} \rangle$ for all model asphaltenes. The panels are laid out to correspond to the organization of molecule types in Fig 6.7. The color scheme is the same as in Fig 6.6. Note, however, the different maxima for the color bars.

This may be due to finite size effects - there are only 25 molecules in each simulation at a relatively low concentration, so once a few small aggregates have formed there are no longer any monomers available to continue their growth. However, this may also be due to the entropic repulsion of the tails effectively capping the aggregate stacks once a critical amount of tail sites are present in an aggregate.

6.5 Discussion

In this work we used CG toy models of model asphaltene compounds to investigate the impact of solvent quality and molecular structure on asphaltene aggregation behavior. We simulated a set of model asphaltenes under a variety of solvent conditions implemented as nonbonded pair potentials between asphaltene sites that varied in their cohesive strength. The simplicity of this model allowed us to simulate a wide variety of solvent conditions for the selected model asphaltenes. We then characterized the aggregation propensity of each system, as well as the

structure of the resulting aggregates.

Our work demonstrates that for ovalene-type molecules, the rod-like Yen-Mullins nanoaggregates form most readily in solvents that promote cohesion between both the core and tail sites, as highlighted in Figure 6.9. In particular, solvents that have approximately the same cohesive strength between all of the sites display the most distinctive Yen-Mullins-type structures for the nanoaggregates. This is consistent with expectations for a real crude oil system, where the crude oil mixture would not be an ideal solvent for any component of the asphaltene molecule. This is also consistent with observations of Yen-Mullins-type aggregates forming in previous AA simulation studies on model asphaltenes in heptane and toluene,[238,257] which would both qualify as mixed solvents under our toy model.

Smaller versions of these aggregates also form under solvents that strongly promote cohesion between the cores but are ideal solvents for the tails. In these cases, it is expected that the tails interfere with the core-driven cohesion by physically blocking the binding faces on an aggregate. The aggregate-limiting behavior of longer alkyl tails has also been shown in a previous AA simulation study.[257] This may provide a mechanism by which Yen-Mullins nanoaggregates would stop growing when they contain 7-10 molecules, if the crude oil mixture is a worse solvent for the cores than for the tails.

The internally flexible bipyrene-type molecules display little to no Yen-Mullins-like nanoaggregation behavior under any solvent regime investigated, and aggregate much less readily in the core-driven and mixed cohesion regimes. So while asphaltenes with flexible cores are expected to be present in real crude oil, based on these results we would not expect them to be the basis of Yen-Mullins-type noaggregate formation in these systems. They may however, be involved in the formation of clusters and networks of aggregates, or could form stacked aggregates under solvent conditions not explored here. A recent study by Wang et al. explores the aggregation behavior of similar archipelago-type asphaltenes with 2-3 aromatic cores joined by aliphatic chains.[258] Their results showed Yen-Mullins-type aggregation for these model asphaltenes up to concentrations of 10 wt%, with those asphaltenes possessing three cores aggregating more readily than those with two. This discrepancy between our results and those of Wang et al. may be due to the difference in the molecular structure of the modeled asphaltenes. Their model compounds had larger aromatic cores and longer aliphatic links than our

bipyrene-type molecules, which could offer different aggregation pathways to the two molecule types. At archipelago asphaltene concentrations higher than 10 wt%, Wang et al. observed aggregation that bypassed the Yen-Mullins aggregation hierarchy and immediately began forming distributed networks as a result of the flexible aggregation pathways available to this type of molecule.[258] It would be interesting to test our bipyrene-type molecules in this concentration regime to see if this structure also favors network formation under these conditions. Future work could also explore the role of internally flexible model asphaltenes in systems containing a variety of model asphaltene structures, as they may play a role in transitioning from nanoaggregation to clustering, or in the termination of nanoaggregate formation.

In the tail-driven cohesion regime, both ovalene-type and bipyrene-type molecules aggregate readily into large, loosely bound, disordered clusters. This regime corresponds to a highly aromatic solvent that poorly solvates the asphaltene tails but is a good solvent for the cores. Previous simulation studies have also shown that aromatic solvents such as toluene lead to less ordered aggregates by associating with the asphaltene cores.[238,257] In our simulations, these clusters grew to contain all of the asphaltene molecules present in the simulation box, and demonstrated limited monomer association and dissociation behavior throughout the course of the simulation. These clusters had little to no core alignment or other obvious ordering, although it is expected that those model asphaltenes with longer tails could eventually form micelle-like structures in systems with more molecules available. This solvent regime does not correspond to the expected composition of a real crude oil system, which would be unlikely to be comprised of solely aromatic compounds.

Compared to other nanoaggregation studies, the structures we observed for ovalene-type molecules under solvents promoting aromatic cohesion are relatively uniform and simple. We observed stacks of ovalene-type molecules of varying sizes in these solvents, with the cores aligned in a face-on position. This is in contrast to other simulation studies that have shown more variety in the structure of nanoaggregates of model asphaltenes and asphaltene-like molecules. Nanoaggregate archetypes not observed in this study include branched aggregates,[237] curved aggregates,[259] helical aggregates,[248] and aggregates containing t-shaped interfaces between the aromatics cores.[250,260] The lack of variety in our observed nanoaggregate structures may result from the simplicity and symmetry of the model asphaltene

compounds that we selected for this study. In particular, the symmetric placement of the tails around the cores of the model asphaltenes we used precludes the possibility of t-shaped interfaces between the cores, and discourages off-center core stacking that might lead to branched or helical aggregates. Future studies with this type of model could use model asphaltene molecules with varied and asymmetric structures to observe the impact of these structural details on the resulting nanoaggregates.

A recent study by Wang et al.[250] used an explicit solvent CG model to study the mesoscale clustering behavior of model asphaltenes under varying solvent composition, temperature, and pressure. They identified regions of the resulting phase space where nanoaggregation, clustering, and network formation occurred, and characterized the resulting aggregate structure. In the region of their phase space relevant to this study (300 K, 1 bar), they observed nanoaggregation to onset in a 25% toluene / 75% heptane solvent, and clustering to onset in a 100% heptane solvent. There is no direct correspondence between their solvent composition and our solvent space, and their asphaltene concentrations were higher (20 wt% vs 1 wt%). However, their results suggest that there may be regions of solvent space sampled in our study where clustering would occur, given a sufficient supply of asphaltene molecules. Depending on the solvent, a concentration of 1 wt% is within the regime where the onset of clustering has been experimentally observed. This is consistent with our observation that metastable nanoaggregates form under relatively poor solvents, as evidenced by Figures 6.5a and 6.5c. The solvent in these systems is likely poor enough to promote clustering (or even network formation) in a system with more asphaltene molecules available. Future work with our model could focus on these regions of solvent space and simulate larger systems to observe the clustering and network-formation behvaior of these model asphaltenes.

It should be noted that our phase diagrams across solvent space use temperature-scaled coordinates to express the cohesion of the nonbonded potential. In this type of simple toy model the cohesive interactions scale directly with temperature, so changing the temperature effectively scales the solvent quality. Interestingly, this property of the toy model is consistent with the more detailed CG simulations of model asphaltenes by Wang et al., where the authors demonstrate that the heptane:toluene solvent composition acts like an effective temperature for their model.[250] This supports the idea that the type of toy model we used in this study

approximately captures the effect of solvent quality and its impact on asphaltene nanoaggregation.

In this work, we have compared our toy models with more detailed simulations, and with real crude oil systems. It is important to note, however, that there is only a loose correspondence between the solvent quality as represented in this toy model and real systems. In particular, there is no a priori correspondence between the solvent conditions in our model and the composition of an actual solvent. Therefore the trends observed here can can only be suggestive rather than prescriptive in identifying specific mechanisms or conditions.

6.6 Conclusions

In conclusion, our use of toy model asphaltenes in implicit solvent has demonstrated the effect of molecular structure and solvent conditions on the aggregation of model asphaltene compounds. This work supports the Yen-Mullins model of nanoaggregation for island-type asphaltenes in a crude oil mixture, where short rod-like structures are formed by stacking the asphaltene cores. Further, the increased propensity for this Yen-Mullins-type aggregation seen in mixed solvents compared to solvents promoting only aromatic cohesion suggests that the asphaltene tails may play a role in promoting aggregate formation. Moreover, this work demonstrates the utility of CG toy models for qualitatively probing asphaltene aggregation behavior. The range of solvents and structures examined here would not have been practical to simulate at AA resolution, or even at CG resolution with an explicit solvent.

Finally, this work presents several directions for future studies in this area. For example, it would be interesting to select a few solvent conditions to simulate with hundreds or thousands of asphaltenes using this type of toy model to study mesoscale clustering behavior. Further, it could also be informative to use such a model to study the behavior of a mixture of molecular structures to investigate how rigid and flexible core model asphaltenes interact.

Chapter 7
Conclusions and Outlook

7.1 Overview

In this work we have presented several aspects of CG modeling of petrochemical systems, as well as of CG modeling in general. In particular, we demonstrated a general pressure-matching approach for quantitatively correcting the pressure of bottom-up CG models. We applied this approach to CG models of heptane and toluene both because these molecules are sufficiently complex to test the method, and because they are solvents of interest in the context of asphaltene simulations. We then extended this approach to an extended-ensemble across system compositions with varying heptane:toluene ratios and demonstrated a transferable CG force field and pressure correction across this range of mixtures. This demonstrates the importance of considering the state-point dependence of the PMF in creating thermodynamically accurate and transferable models. Next, we discussed bottom-up CG modeling in the context of a 'van der Waals' perspective where thermodynamic information from the underlying AA model is encoded into the many-body PMF. This perspective suggests that thermodynamically accurate and transferable bottom-up CG models can be parameterized given the appropriate degrees of freedom. We then present the BOCS software package used for force matching and pressure matching as an open-source project for use by the research community at large. In particular, this provides an open-source implementation of pressure-matching and extended-ensemble force matching, features which were not previously available to the wider research community. Finally, we use toy model asphaltenes to explore the aggregation behavior of several model asphaltenes

across a range of solvent conditions. This study highlighted the importance of core-core cohesion in forming Yen-Mullins-type nanoaggregates, and suggests ranges of solvent quality where this behavior might be observed.

7.2 Future Work

7.2.1 Transferable Bottom-Up Models

In order for a bottom-up CG model to save computational effort compared to an AA model, the CG model should enable studies that would be intractable with the underlying AA model. In practice, computational cost of the AA sampling and force matching calculation required to parameterize a CG model of a single system are nearly as limiting as simply extending the AA simulation. A CG model derived in this way may be productively put to use in the context of a larger scale or longer simulation, but behavior not sampled during the reference AA simulation may be poorly represented in the CG model.

In service of computational efficiency, transferable CG models are more appealing as they can be parameterized once and used many times across their range of relevance, saving significant computational effort relative to a purely AA study. Traditional bottom-up models are not transferable beyond the conditions of their parameterization, and so are less appealing than the more transferable top-down models for researchers with limited computational resources. In this work we have demonstrated pressure-matching and extended-ensemble approaches that can be used to produce transferable bottom-up models, and considered the 'van der Waals' perspective on coarse-graining that suggests a similar approach may be useful in the context of other thermodynamic properties. A direct extension of this method could include treatment of anisotropic systems by considering the correction to each component of the pressure tensor independently. This approach could then be used for parameterizing CG models simulated under a constant surface tension ensemble against a solid interface.

While the pressure-matching approach presented here is limited to bulk systems without interfaces, recent work on bottom-up models with local density dependent potentials provides an even more transferable approach towards accurately modeling CG pressure and density that can be used in interfacial systems.[219,223] This would

allow, for instance, the simulation of bottom-up CG asphaltenes at an oil-water, oil-rock, or oil-metal interface, where the clustering and network formation behavior may differ from that seen in solution. These interfacial systems are of particular interest to the petrochemical industry as they are present in several stages of petroleum extraction and transportation.

7.2.2 BOCS Software Development

Maintenance and development of the BOCS software package will be an ongoing project as new features are added and new edge cases with bugs are identified. The project's home on GitHub provides a forum for users to report issues and offer fixes and features of their own. Engaging the community on these issues and additions will be a vital part of maintaining BOCS as an effective open-source project moving forward.

7.2.3 Asphaltene Modeling

The efficiency of the toy-modeling approach demonstrated in Chapter 6 allows for future work in several directions. First, mesoscale simulation studies with hundreds or even thousands of asphaltene molecules are feasible at any concentration using this model. This opens the door for studies examining mesoscale clustering behavior, and even the onset of viscoelastic network formation. These studies would be interesting in the context of single model asphaltene structures to investigate the clustering and network formation behavior of specific molecular structures. It would also be informative to study this regime with a range of molecular structures present, and to track the role of the different structural archetypes in cluster and network formation. Further, simulating a larger system containing either a single type of structure or a mixture of structures under different solvent regimes could prove instructive as to the behavior of asphaltenes at different stages during oil processing. Another research avenue suggested by this work would be to extend the phase diagrams presented in Chapter 6 into another order parameter such as concentration or aromatic core size. This would provide additional insight into the state-point- and structure-dependence of the observed asphaltene aggregation behavior.

Finally, this type of implicit-solvent toy model is not limited to the study of

asphaltene aggregation. Simulation studies using similar models could be used to study the aggregation and self-assembly behavior of a broad variety of molecules. If the CG sites are not constrained to represent single atoms or small groups of atoms, then this type of model could also be used to study self-assembly in nanoparticles of varying shapes and dimensions.

7.3 CG Modeling and GPUs

Bottom-up CG models currently lag behind AA models in their ability to take advantage of emerging computing technologies. At the time of writing features required for simulating the bottom-up models presented in this work are not supported in any major MD package when running on graphics processing units (GPUs). In particular, the tabulated interactions required to represent the CG potentials are not supported when running on a GPU.

MD simulations run on GPU can see performance enhancements of 3-10x increase in ns/day sampling efficiency,[2,261,262] comparable to the increase in sampling efficiency seen for the heptane and toluene models presented in Chapters 2 and 3. Further, while bottom-up CG models require an involved parameterization process and access to high-performance computing resources, simulating AA models on a GPU is becoming increasingly user-friendly and can be performed on a commodity desktop with a $1000 graphics card. As a result, for a large-scale simulation study it is currently more cost-effective to run AA simulations on GPUs than it is to parameterize a CG model for this purpose.

Currently, some MD software packages such as GROMACS[204] support GPU simulations with CG models with simple functional forms for their interactions. This has enabled large scale CG simulations with top-down forcefields such as the Martini model[194] and derivatives that would not be feasible running on CPUs alone. Once implemented, the addition of GPU support for the tabulated nonbonded interactions required for flexible CG models will open up a new regime of simulation scale for bottom-up CG models. In the context of asphaltene simulation, this would enable the use of bottom-up models to simulate cluster formation and mesoscale network formation in an explicit CG solvent. Such studies would provide an important complement to the current top-down approaches by more accurately capturing the structural details of the nanoaggregates, which may impact the

aggregation propensity and mechanism of the model asphaltenes.

7.4 Outlook

In conclusion, this work suggests that bottom-up CG models are poised to become more widely applicable. The addition of thermodynamic degrees of freedom for bottom-up CG models presents an interesting approach for continuing to improve the accuracy and transferability of these models through the addition of the appropriate degrees of freedom to the CG models. Local-density-dependent CG force fields have already extended this idea to reproduce both pressure and interfacial tension of reference AA systems.

The open-source BOCS software released as part of this work provides tools for researchers to reproduce and extend our work on transferable bottom-up CG models, and the software is still being actively developed and updated for broader compatibility with MD codes and additional feature sets. It is our hope that the research community will find BOCS useful as a tool for both investigating and developing CG models in the coming years.

7.5 Supporting Information

The Supporting Information for this work provides additional description of the methods and results described here, as well as some additional analysis of this data. The Supporting Information is available online free of charge at `https://pubs.acs.org` and `https://aip.scitation.org`.

Bibliography

[1] Allen, M. P. and D. P. Tildesley (1987) Computer Simulation of Liquids, Oxford Press, New York, NY USA.

[2] Kutzner, C., S. PÃąll, M. Fechner, A. Esztermann, B. L. de Groot, and H. GrubmÃijller (2015) "Best bang for your buck: GPU nodes for GROMACS biomolecular simulations," Journal of Computational Chemistry, **36**(26), pp. 1990–2008, https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.24030.
URL https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.24030

[3] Eastman, P., J. Swails, J. D. Chodera, R. T. McGibbon, Y. Zhao, K. A. Beauchamp, L.-P. Wang, A. C. Simmonett, M. P. Harrigan, C. D. Stern, R. P. Wiewiora, B. R. Brooks, and V. S. Pande (2017) "OpenMM 7: Rapid development of high performance algorithms for molecular dynamics," PLOS Computational Biology, **13**(7), pp. 1–17.
URL https://doi.org/10.1371/journal.pcbi.1005659

[4] Adcock, S. A. and J. A. MacCammon (2006) "Molecular dynamics: survey of methods for simulating the activity of proteins," Chem. Rev., **106**, pp. 1589–1615.

[5] Guenza, M. G. (2008) "Theoretical models for bridging timescales in polymer dynamics," J. Phys.: Condens. Matter, **20**(3), p. 033101.

[6] 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**, pp. 1869–92.

[7] Peter, C. and K. Kremer (2010) "Multiscale simulation of soft matter systems," Faraday Disc., **144**, pp. 9–24.

[8] Saunders, M. G. and G. A. Voth (2013) "Coarse-Graining Methods for Computational Biology," Annu. Rev. Biophys., **42**, pp. 73–93.

[9] Marrink, S. J. and D. P. Tieleman (2013) "Perspective on the Martini model," Chem. Soc. Rev., **42**, pp. 6801–6822.

[10] Chebaro, Y., S. Pasquali, and P. Derreumaux (2012) "The coarse-grained OPEP force field for non-amyloid and amyloid proteins," J. Phys. Chem. B, **116**(30), pp. 8741–52.

[11] Bereau, T., Z.-J. Wang, and M. Deserno (2014) "More than the sum of its parts: Coarse-grained peptide-lipid interactions from a simple cross-parametrization," The Journal of Chemical Physics, **140**(11), p. 115101, https://doi.org/10.1063/1.4867465.
URL https://doi.org/10.1063/1.4867465

[12] Kar, P., S. M. Gopal, Y.-M. Cheng, A. Predeus, and M. Feig (2013) "PRIMO: A Transferable Coarse-Grained Force Field for Proteins," Journal of Chemical Theory and Computation, **9**(8), pp. 3769–3788, pMID: 23997693, https://doi.org/10.1021/ct400230y.
URL https://doi.org/10.1021/ct400230y

[13] Müller, M., K. Katsov, and M. Shick (2006) "Biological and synthetic membranes: What can be learned from a coarse-grained description?" Phys. Rep., **434**, pp. 113–176.

[14] Schmid, F. (2009) "Toy amphiphiles on the computer: What can we learn from generic models?" Macromol. Rapid Comm., **30**(9-10), pp. 741–751.

[15] Deserno, M. (2009) "Mesoscopic Membrane Physics: Concepts, Simulations, and Selected Applications," Macromol. Rapid Comm., **30**(9-10), pp. 752–771.

[16] Hyeon, C. and D. Thirumalai (2011) "Capturing the essence of folding and functions of biomolecules using coarse-grained models," Nat. Commun., **2**, p. 487.

[17] Doye, J. P. K., T. E. Ouldridge, A. A. Louis, F. Romano, P. Sulc, C. Matek, B. E. K. Snodin, L. Rovigatti, J. S. Schreck, R. M. Harrison, and W. P. J. Smith (2013) "Coarse-graining DNA for simulations of DNA nanotechnology," Phys. Chem. Chem. Phys., **15**, pp. 20395–20414.

[18] 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), pp. 2115–2124.

[19] 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), pp. 2531–2543.

[20] Kirkwood, J. G. (1935) "Statistical mechanics of fluid mixtures," J. Chem. Phys., **3**(5), pp. 300–313.

[21] 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**, pp. 849–73.

[22] Likos, C. N. (2001) "Effective interactions in soft condensed matter physics," Phys. Rep., **348**(4–5), pp. 267 – 439.

[23] Akkermans, R. L. C. and W. J. Briels (2001) "Coarse-grained interactions in polymer melts: a variational approach," J. Chem. Phys., **115**(13), pp. 6210–6219.

[24] Hansen, J.-P., C. I. Addison, and A. A. Louis (2005) "Polymer solutions: from hard monomers to soft polymers," J. Phys.: Condens. Matter, **17**(45), p. S3185.
URL `http://stacks.iop.org/0953-8984/17/i=45/a=001`

[25] 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**, p. 244114.

[26] 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**, pp. 3730–37.

[27] Izvekov, S. and G. A. Voth (2005) "A multiscale coarse-graining method for biomolecular systems," J. Phys. Chem. B, **109**, pp. 2469 – 2473.

[28] Müller-Plathe, F. (2002) "Coarse-graining in polymer simulation: From the atomistic to the mesoscopic scale and back," ChemPhysChem, **3**, pp. 754 – 769.

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

[30] 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**, pp. 7785–93.

[31] 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**, pp. 5661–74.

[32] Noid, W. G. (2013) "Systematic methods for structurally consistent coarse-grained models," Methods Mol Biol, **924**, pp. 487–531.

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

[34] 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), pp. 221–229.

[35] 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**, pp. 2522–5.

[36] Murtola, T., E. Falck, M. Karttunen, and I. Vattulainen (2007) "Coarse-grained model for phospholipid/cholesterol bilayer employing inverse Monte Carlo with thermodynamic constraints," J. Chem. Phys., **121**, p. 075101.

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

[38] 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), p. 024103.

[39] Mullinax, J. W. and W. G. Noid (2009) "Extended Ensemble approach for deriving transferable Coarse-grained potentials," J. Chem. Phys., **131**, p. 104110.

[40] Chaimovich, A. and M. S. Shell (2009) "Anomalous waterlike behavior in spherically-symmetric water models optimized with the relative entropy," Phys. Chem. Chem. Phys., **11**(12), pp. 1901–1915.

[41] 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), p. 224107.

[42] Izvekov, S. (2011) "Towards an understanding of many-particle effects in hydrophobic association in methane solutions," J. Chem. Phys., **134**(3), p. 034104.

[43] Farah, K., A. C. Fogarty, M. C. Böhm, and F. Müller-Plathe (2011) "Temperature dependence of coarse-grained potentials for liquid hexane," Phys. Chem. Chem. Phys., **13**(7), pp. 2894–902.

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

[45] Boek, E. S., D. S. Yakovlev, and T. F. Headen (2009) "Quantitative Molecular Representation of Asphaltenes and Molecular Dynamics Simulation of Their Aggregation," Energy & Fuels, **23**(3), pp. 1209–1219, https://doi.org/10.1021/ef800876b.
URL https://doi.org/10.1021/ef800876b

[46] Mullins, O. C. (2009) "Rebuttal to Strausz et al. Regarding Time-Resolved Fluorescence Depolarization of Asphaltenes," Energy & Fuels, **23**(5), pp. 2845–2854, https://doi.org/10.1021/ef801067v.
URL https://doi.org/10.1021/ef801067v

[47] Kuznicki, T., J. H. Masliyah, and S. Bhattacharjee (2009) "Aggregation and Partitioning of Model Asphaltenes at Toluene-Water Interfaces: Molecular Dynamics Simulations," Energy & Fuels, **23**(10), pp. 5027–5035, https://doi.org/10.1021/ef9004576.
URL https://doi.org/10.1021/ef9004576

[48] Boek, E. S., T. F. Headen, and J. T. Padding (2010) "Multi-scale simulation of asphaltene aggregation and deposition in capillary flow," Faraday discussions, **144**, pp. 271–284.

[49] Ortega-Rodriguez, A., S. A. Cruz, A. Gil-Villegas, F. Guevara-Rodriguez, and C. Lira-Galeana (2003) "Molecular View of the Asphaltene Aggregation Behavior in Asphaltene-Resin Mixtures," Energy & Fuels, **17**(4), pp. 1100–1108, https://doi.org/10.1021/ef030005s.
URL https://doi.org/10.1021/ef030005s

[50] Creek, J. L. (2005) "Freedom of Action in the State of Asphaltenes: Escape from Conventional Wisdom," Energy & Fuels, **19**(4), pp. 1212–1224, https://doi.org/10.1021/ef049778m.
URL https://doi.org/10.1021/ef049778m

[51] Rogel, E., C. Ovalles, and M. Moir (2010) "Asphaltene Stability in Crude Oils and Petroleum Materials by Solubility Profile Analysis," Energy & Fuels, **24**(8), pp. 4369–4374, https://doi.org/10.1021/ef100478y.
URL https://doi.org/10.1021/ef100478y

[52] BuenrostroâĂŘGonzalez, E., C. LiraâĂŘGaleana, A. GilâĂŘVillegas, and J. Wu (2004) "Asphaltene precipitation in crude oils: Theory and experiments," AIChE Journal, **50**(10), pp. 2552–2570, https://onlinelibrary.wiley.com/doi/pdf/10.1002/aic.10243.
URL https://onlinelibrary.wiley.com/doi/abs/10.1002/aic.10243

[53] Hoepfner, M. P., V. Limsakoune, V. Chuenmeechao, T. Maqbool, and H. S. Fogler (2013) "A Fundamental Study of Asphaltene Deposition,"

Energy & Fuels, **27**(2), pp. 725–735, https://doi.org/10.1021/ef3017392.
URL https://doi.org/10.1021/ef3017392

[54] Li, D. D. and M. L. Greenfield (2014) "Chemical compositions of improved model asphalt systems for molecular simulations," Fuel, **115**, pp. 347 – 356.
URL http://www.sciencedirect.com/science/article/pii/S001623611300611X

[55] Wiehe, I. and K. Liang (1996) "Asphaltenes, resins, and other petroleum macromolecules," Fluid Phase Equilibria, **117**(1), pp. 201 – 210, proceedings of the Seventh International Conference on Fluid Properties and Phase Equilibria for Chemical Process Design.
URL http://www.sciencedirect.com/science/article/pii/0378381295029540

[56] Hansen, J. S., C. A. Lemarchand, E. Nielsen, J. C. Dyre, and T. SchrÃŸder (2013) "Four-component united-atom model of bitumen," The Journal of Chemical Physics, **138**(9), p. 094508, https://doi.org/10.1063/1.4792045.
URL https://doi.org/10.1063/1.4792045

[57] Mullins, O. C., H. Sabbah, J. Eyssautier, A. E. Pomerantz, L. BarrÃŁ, A. B. Andrews, Y. Ruiz-Morales, F. Mostowfi, R. McFarlane, L. Goual, R. Lepkowicz, T. Cooper, J. Orbulescu, R. M. Leblanc, J. Edwards, and R. N. Zare (2012) "Advances in Asphaltene Science and the YenâĂŞMullins Model," Energy & Fuels, **26**(7), pp. 3986–4003, https://doi.org/10.1021/ef300185p.
URL https://doi.org/10.1021/ef300185p

[58] Wargadalam, V., K. Norinaga, and M. Iino (2002) "Size and shape of a coal asphaltene studied by viscosity and diffusion coefficient measurements," Fuel, **81**(11-12), pp. 1403–1407.

[59] Groenzin, H. and O. C. Mullins (2000) "Molecular Size and Structure of Asphaltenes from Various Sources," Energy & Fuels, **14**(3), pp. 677–684, https://doi.org/10.1021/ef990225z.
URL https://doi.org/10.1021/ef990225z

[60] Mullins, O. C., B. MartÃĐnez-Haya, and A. G. Marshall (2008) "Contrasting Perspective on Asphaltene Molecular Weight. This Comment vs the Overview of A. A. Herod, K. D. Bartle, and R. Kandiyoti," Energy & Fuels, **22**(3), pp. 1765–1773, https://doi.org/10.1021/ef700714z.
URL https://doi.org/10.1021/ef700714z

[61] Mullins, O. C. (2010) "The Modified Yen Model," Energy & Fuels, **24**(4), pp. 2179–2207, https://doi.org/10.1021/ef900975e.
URL https://doi.org/10.1021/ef900975e

[62] Sheu, E., Y. Long, and H. Hamza (2007) "Asphaltene self-association and precipitation in solventsâĂŤAC conductivity measurements," in Asphaltenes, Heavy Oils, and Petroleomics, Springer, pp. 259–277.

[63] Zeng, H., Y.-Q. Song, D. L. Johnson, and O. C. Mullins (2009) "Critical Nanoaggregate Concentration of Asphaltenes by Direct-Current (DC) Electrical Conductivity," Energy & Fuels, **23**(3), pp. 1201–1208, https://doi.org/10.1021/ef800781a.
URL https://doi.org/10.1021/ef800781a

[64] Goual, L. (2009) "Impedance Spectroscopy of Petroleum Fluids at Low Frequency," Energy & Fuels, **23**(4), pp. 2090–2094, https://doi.org/10.1021/ef800860x.
URL https://doi.org/10.1021/ef800860x

[65] Lisitza, N. V., D. E. Freed, P. N. Sen, and Y.-Q. Song (2009) "Study of Asphaltene Nanoaggregation by Nuclear Magnetic Resonance (NMR)," Energy & Fuels, **23**(3), pp. 1189–1193, https://doi.org/10.1021/ef800631a.
URL https://doi.org/10.1021/ef800631a

[66] Indo, K., J. Ratulowski, B. Dindoruk, J. Gao, J. Zuo, and O. C. Mullins (2009) "Asphaltene Nanoaggregates Measured in a Live Crude Oil by Centrifugation," Energy & Fuels, **23**(9), pp. 4460–4469, https://doi.org/10.1021/ef900369r.
URL https://doi.org/10.1021/ef900369r

[67] Mostowfi, F., K. Indo, O. C. Mullins, and R. McFarlane (2009) "Asphaltene Nanoaggregates Studied by Centrifugation," Energy & Fuels, **23**(3), pp. 1194–1200, https://doi.org/10.1021/ef8006273.
URL https://doi.org/10.1021/ef8006273

[68] Goual, L., M. Sedghi, H. Zeng, F. Mostowfi, R. McFarlane, and O. C. Mullins (2011) "On the formation and properties of asphaltene nanoaggregates and clusters by DC-conductivity and centrifugation," Fuel, **90**(7), pp. 2480 – 2490.
URL http://www.sciencedirect.com/science/article/pii/S0016236111000950

[69] Anisimov, M. A., I. K. Yudin, V. Nikitin, G. Nikolaenko, A. Chernoutsan, H. Toulhoat, D. Frot, and Y. Briolant (1995) "Asphaltene Aggregation in Hydrocarbon Solutions Studied by Photon Correlation Spectroscopy," The Journal of Physical Chemistry, **99**(23), pp. 9576–9580, https:

//doi.org/10.1021/j100023a040.
URL https://doi.org/10.1021/j100023a040

[70] Yudin, I. K. and M. A. Anisimov (2007) Dynamic Light Scattering Monitoring of Asphaltene Aggregation in Crude Oils and Hydrocarbon Solutions, chap. 17, Springer New York, New York, NY, pp. 439–468.
URL https://doi.org/10.1007/0-387-68903-6_17

[71] 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), pp. 191–228.

[72] Klein, M. L. and W. Shinoda (2008) "Large-scale molecular dynamics simulations of self-assembling systems," Science, **321**(5890), pp. 798–800.

[73] 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), pp. 12423–30.

[74] 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**, pp. 2108–2119.
URL http://dx.doi.org/10.1039/C2SM27201F

[75] Noid, W. G. (2013) "Perspective: Coarse-grained models for biomolecular systems," J. Chem. Phys., **139**(9), 090901.
URL http://scitation.aip.org/content/aip/journal/jcp/139/9/10.1063/1.4818908

[76] 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), pp. 113–176.

[77] Akkermans, R. L. C. and W. J. Briels (2001) "A structure-based coarse-grained model for polymer melts," J. Chem. Phys., **114**(2), pp. 1020–1031.
URL http://link.aip.org/link/?JCP/114/1020/1

[78] 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**, p. 105.

[79] Guenza, M. (2015) "Thermodynamic consistency and other challenges in coarse-graining models," Eur. Phys. J. ST, **224**, pp. 2177–2191.

[80] 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**, pp. 43–56.

[81] Rudzinski, J. F. and W. G. Noid (2011) "Coarse-graining entropy, forces, and structures," J. Chem. Phys., **135**(21), p. 214101.

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

[83] McCarty, J., A. J. Clark, I. Y. Lyubimov, and M. G. Guenza (2012) "Thermodynamic Consistency between Analytic Integral Equation Theory and Coarse-Grained Molecular Dynamics Simulations of Homopolymer Melts," Macromolecules, **45**(20), pp. 8482–8493.

[84] Clark, A. J., J. McCarty, I. Y. Lyubimov, and M. G. Guenza (2012) "Thermodynamic Consistency in Variable-Level Coarse Graining of Polymeric Liquids," Phys. Rev. Lett., **109**, p. 168301.
URL http://link.aps.org/doi/10.1103/PhysRevLett.109.168301

[85] Clark, A. J., J. McCarty, and M. G. Guenza (2013) "Effective potentials for representing polymers in melts as chains of interacting soft particles," J. Chem. Phys., **139**(12), 124906.

[86] McCarty, J., A. J. Clark, J. Copperman, and M. G. Guenza (2014) "An analytical coarse-graining method which preserves the free energy, structural correlations, and thermodynamic state of polymer melts from the atomistic to the mesoscale," J. Chem. Phys., **140**(20), 204913.

[87] Fu, C.-C., P. M. Kulkarni, M. S. Shell, and L. G. Leal (2012) "A test of systematic coarse-graining of molecular dynamics simulations: Thermodynamic properties," J. Chem. Phys., **137**(16), 164106.
URL http://link.aip.org/link/?JCP/137/164106/1

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

[89] 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**, p. 064109.

[90] Trofimov, S. Y., E. L. F. Nies, and M. A. J. Michels (2002) "Thermodynamic consistency in dissipative particle dynamics simulations of strongly nonideal liquids and liquid mixtures," J. Chem. Phys., **117**(20), pp. 9383–9394.

[91] Louis, A. A. (2002) "Beware of density dependent pair potentials," J. Phys.: Condens. Matter, **14**, pp. 9187–206.

[92] D'Adamo, G., A. Pelissetto, and C. Pierleoni (2013) "Predicting the thermodynamics by using state-dependent interactions," J. Chem. Phys., **138**(23), p. 234107.

[93] Stillinger, F. H., H. Sakai, and S. Torquato (2002) "Statistical mechanical models with effective potentials: Definitions, applications, and thermodynamic consequences," J. Chem. Phys., **117**(1), pp. 288–296.
URL http://link.aip.org/link/?JCP/117/288/1

[94] Tuckerman, M. E. (2010) Statistical mechanics: Theory and molecular simulation, Oxford.

[95] Ciccotti, G. and J. P. Ryckaert (1986) "Molecular Dynamics Simulation of Rigid Molecules," Comp. Phys. Rep., **4**, pp. 345–92.

[96] Martyna, G. J., D. J. Tobias, and M. L. Klein (1994) "Constant pressure molecular dynamics algorithms," J. Chem. Phys., **101**(5), pp. 4177–4189.

[97] Bekker, H., H. J. C. Berendsen, and W. F. van Gunsteren (1995) "Force and virial of torsional-angle-dependent potentials," J. Comp. Chem., **16**(5), pp. 527–533.

[98] Wang, M. C. and G. E. Uhlenbeck (1945) "On the Theory of the Brownian Motion II," Rev. Mod. Phys., **17**, pp. 323–342.

[99] 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**, p. 244115.

[100] 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), pp. 954–965.

[101] 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**, pp. 4116–27.

[102] Andersen, H. C. (1980) "Molecular dynamics simulations at constant pressure and/or temperature," J. Chem. Phys., **72**, pp. 2384–93.

[103] Pronk, S., S. PÃąll, R. Schulz, P. Larsson, P. Bjelkmar, R. Apostolov, M. R. Shirts, J. C. Smith, P. M. Kasson, D. van der Spoel, B. Hess, and E. Lindahl (2013) "GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit," Bioinformatics, **29**(7), pp. 845–854.

[104] 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**, pp. 11225–36.

[105] 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**, pp. 8345–48.

[106] 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**, pp. 3684–90.

[107] Lide, D. R. (ed.) (2009) CRC Handbook of Chemistry and Physics, 90th Edition, 90 ed., CRC Press.

[108] Nose, S. (1984) "A molecular dynamics method for simulations in the canonical ensemble," Mol. Phys., **52**, pp. 255–68.

[109] Hoover, W. G. (1985) "Canonical dynamics: Equilibrium phase-space distributions," Phys. Rev. A, **31**, pp. 1695–7.

[110] Parrinello, M. and A. Rahman (1980) "Crystal Structure and Pair Potentials: A Molecular-Dynamics Study," Phys. Rev. Lett., **45**(14), pp. 1196–1199.

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

[112] See supplemental material at [URL will be inserted by AIP] for additional plots that describe the CG potentials, the resulting equilibrium structure, and the pressure-matching method.

[113] 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**, pp. 27–36.

[114] Plimpton, S. (1995) "FAST PARALLEL ALGORITHMS FOR SHORT-RANGE MOLECULAR-DYNAMICS," J. Comp. Phys., **117**, pp. 1 – 19.

[115] Martyna, G. J., M. E. Tuckerman, D. J. Tobias, and M. L. Klein (1996) "Explicit reversible integrators for extended systems dynamics," Mol. Phys., **87**(5), pp. 1117–1157.

[116] Martyna, G. J., M. L. Klein, and M. Tuckerman (1992) "NosÃĬâĂŞHoover chains: The canonical ensemble via continuous dynamics," J. Chem. Phys., **97**(4), pp. 2635–2643.

[117] 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), pp. 8621–35.

[118] ——— (2015) "Bottom-Up Coarse-Graining of Peptide Ensembles and HelixâĂŞCoil Transitions," J. Chem. Theor. Comp., **11**(3), pp. 1278–1291.

[119] ——— (2015) "A generalized-Yvon-Born-Green method for coarse-grained modeling," Eur. Phys. J. ST, **224**, pp. 2193–2216.

[120] Ruhle, 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), pp. 3211–3223.

[121] 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.
URL http://link.aip.org/link/?JCP/136/194115/1

[122] 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), pp. 8295–8312.

[123] Ellis, C. R., J. F. Rudzinski, and W. G. Noid (2011) "generalized-Yvon-Born-Green model for toluene," Macromol. Theory Sim., **20**, pp. 478–95.

[124] Dyre, J. C. (2014) "Hidden scale invariance in condensed matter," J. Phys. Chem. B, **118**, pp. 10007–10024.

[125] Papini, J. J., T. B. Schrøder, and J. C. Dyre (2013) "Do all liquids become strongly correlating at high pressure?" ArXiv e-prints, 1103.4954v2.

[126] Jover, J. F., E. A. MÃijller, A. J. Haslam, A. Galindo, G. Jackson, H. Toulhoat, and C. Nieto-Draghi (2015) "Aspects of Asphaltene Aggregation Obtained from Coarse-Grained Molecular Modeling," Energy & Fuels, **29**(2), pp. 556–566.

[127] Soper, A. K. (1996) "Empirical potential Monte Carlo simulation of fluid structure," Chem. Phys., **202**(2-3), pp. 295–306.

[128] 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), p. 134107.

[129] 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.
URL http://link.aip.org/link/?JCP/139/121906/1

[130] 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), pp. 8383–93.

[131] Carbone, P. and C. AvendaÃśo (2014) "Coarse-grained methods for polymeric materials: enthalpy- and entropy-driven models," Wiley Interdiscip. Rev. Comput. Mol. Sci., **4**(1), pp. 62–70.
URL http://dx.doi.org/10.1002/wcms.1149

[132] Gnan, N., T. B. SchrÃŸder, U. R. Pedersen, N. P. Bailey, and J. C. Dyre (2009) "Pressure-energy correlations in liquids. IV. âĂIJIsomorphsâĂİ in liquid phase diagrams," J. Chem. Phys., **131**(23), 234504.
URL http://scitation.aip.org/content/aip/journal/jcp/131/23/10.1063/1.3265957

[133] Wagner, J. W., J. F. Dama, and G. A. Voth (2015) "Predicting the Sensitivity of Multiscale Coarse-Grained Models to their Underlying Fine-Grained Model Parameters," J. Chem. Theor. Comp., **11**(8), pp. 3547–3560.

[134] Wittmer, J. P., H. Xu, P. PoliåĎska, F. Weysser, and J. Baschnagel (2013) "Communication: Pressure fluctuations in isotropic solids and fluids," J. Chem. Phys., **138**(19), 191101.

[135] ——— (2013) "Shear modulus of simulated glass-forming model systems: Effects of boundary condition, temperature, and sampling time," J. Chem. Phys., **138**(12), p. 12A533.

[136] Green, M. S. (1954) "Markoff Random Processes and the Statistical mechanics of Time-Dependent Phenomena. II. Irreversible Processes in Fluids," J. Chem. Phys., **22**(3), pp. 398–413.

[137] Kubo, R. (1957) "Statistical-Mechanical Theory of Irreversible Processes. I. General Theory and Simple Applications to Magnetic and Conduction Problems," J. Phys. Soc. Jpn., **12**(6), pp. 570–586.

[138] Mori, H. (1958) "STATISTICAL-MECHANICAL THEORY OF TRANSPORT IN FLUIDS," Phys. Rev., **112**(6), pp. 1829–1842.

[139] Zwanzig, R. (1961) "Memory Effects in Irreversible Thermodynamics," Phys. Rev., **124**, pp. 983–992.
URL http://link.aps.org/doi/10.1103/PhysRev.124.983

[140] Hijon, C., P. Espanol, E. Vanden-Eijnden, and R. Delgado-Buscalioni (2010) "Mori-Zwanzig formalism as a practical computational tool," Faraday Disc., **144**, pp. 301–322.

[141] Izvekov, S. and G. A. Voth (2006) "Modeling real dynamics in the coarse-grained representation of condensed phase systems," J. Chem. Phys., **125**, p. 151101.

[142] Junghans, C., M. Praprotnik, and K. Kremer (2008) "Transport properties controlled by a thermostat: An extended dissipative particle dynamics thermostat," Soft Matter, **4**, pp. 156–161.

[143] Davtyan, A., J. F. Dama, G. A. Voth, and H. C. Andersen (2015) "Dynamic force matching: A method for constructing dynamical coarse-grained models with realistic time dependence," J. Chem. Phys., **142**(15), 154104.

[144] Frenkel, D. and B. Smit (2002) Understanding Molecular Simulation: From Algorithms to Applications, second ed., Academic Press, San Diego, CA USA.

[145] Lindorff-Larsen, K., P. Maragakis, S. Piana, M. P. Eastwood, R. O. Dror, and D. E. Shaw (2012) "Systematic validation of protein force fields against experimental data," PLoS One, **7**(2), p. e32131.

[146] Caleman, C., P. J. van Maaren, M. Hong, J. S. Hub, L. T. Costa, and D. van der Spoel (2012) "Force field benchmark of organic liquids: Density, enthalpy of vaporization, heat capacities, surface tension, isothermal compressibility, volumetric expansion coefficient, and dielectric constant," J. Chem. Theor. Comp., **8**, pp. 61–74.
URL `https://pubs.acs.org/doi/abs/10.1021/ct200731v`

[147] Dill, K. A. and S. Bromberg (2011) Molecular driving forces: Statistical thermodynamics in chemistry, 2 ed., Garland Science.

[148] Callen, H. B. (1985) Thermodynamics and an Introduction to Thermostatistics, Wiley.

[149] Ottinger, H. C. (2005) Beyond Thermodynamics, Wiley Interscience.

[150] 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), pp. 3452–3465.

[151] Dunn, N. J. H. and W. G. Noid (2015) "Bottom-up coarse-grained models that accurately describe the structure, pressure, and compressibility of molecular liquids," J. Chem. Phys., **143**, p. 243148.
URL `http://scitation.aip.org/content/aip/journal/jcp/143/24/10.1063/1.4937383`

[152] Vettorel, T. and H. Meyer (2006) "Coarse graining of short polyethylene chains for studying polymer crystallization," J. Chem. Theor. Comp., **2**, pp. 616–629.

[153] Ghosh, J. and R. Faller (2007) "State point dependence of systematically coarse-grained potentials," Mol. Sim., **33**, pp. 759–767.

[154] Liwo, A., M. Khalili, C. Czaplewski, S. Kalinowski, S. Ołdziej, K. Wachucik, and H. A. Scheraga (2007) "Modification and optimization of the united-residue (UNRES) potential energy function for canonical simulations. I. Temperature dependence of the effective energy function and tests of the optimization method with single training proteins," J. Phys. Chem. B, **111**(1), pp. 260–85.

[155] Mognetti, B. M., L. Yelash, P. Virnau, W. Paul, K. Binder, M. MÃijller, 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," The Journal of Chemical Physics, **128**(10), p. 104501, https://doi.org/10.1063/1.2837291.
URL https://doi.org/10.1063/1.2837291

[156] 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), pp. 9919–29.

[157] 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**, p. 034904.

[158] Enciso, M., C. Schutte, and L. Delle Site (2013) "A pH-dependent coarse-grained model for peptides," Soft Matter, **9**, pp. 6118–6127.
URL http://dx.doi.org/10.1039/C3SM27893J

[159] Hsu, D. D., W. Xia, S. G. Arturo, and S. Keten (2015) "Thermomechanically Consistent and Temperature Transferable Coarse-Graining of Atactic Polystyrene," Macromolecules, **48**(9), pp. 3057–3068.

[160] Snodin, B. E. K., F. Randisi, M. Mosayebi, P. ÅĂulc, J. S. Schreck, F. Romano, T. E. Ouldridge, R. Tsukanov, E. Nir, A. A. Louis, and J. P. K. Doye (2015) "Introducing improved structural properties and salt dependence into a coarse-grained model of DNA," The Journal of Chemical Physics, **142**(23), p. 234901, https://doi.org/10.1063/1.4921957.
URL https://doi.org/10.1063/1.4921957

[161] Cao, F. and H. Sun (2015) "Transferability and Nonbond Functional Form of Coarse Grained Force Field âĂŞ Tested on Linear Alkanes," Journal of Chemical Theory and Computation, **11**(10), pp. 4760–4769, pMID: 26574265,

https://doi.org/10.1021/acs.jctc.5b00573.
URL https://doi.org/10.1021/acs.jctc.5b00573

[162] Betancourt, M. R. and S. J. Omovie (2009) "Pairwise energies for polypeptide coarse-grained models derived from atomic force fields," J. Chem. Phys., **130**(19), p. 195103.

[163] Hills, R. D., L. Y. Lu, and G. A. Voth (2010) "Multiscale Coarse-Graining of the Protein Energy Landscape," PLoS Comput. Biol., **6**(6), p. e1000827.

[164] 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), pp. 11911–26.

[165] Engin, O., A. Villa, C. Peter, and M. Sayar (2011) "A Challenge for Peptide Coarse Graining: Transferability of Fragment-Based Models," Macromol. Theory Sim., **20**(7), pp. 451–465.
URL http://dx.doi.org/10.1002/mats.201100005

[166] Dalgicdir, C., O. Sensoy, C. Peter, and M. Sayar (2013) "A transferable coarse-grained model for diphenylalanine: How to represent an environment driven conformational transition," The Journal of Chemical Physics, **139**(23), p. 234115, https://doi.org/10.1063/1.4848675.
URL https://doi.org/10.1063/1.4848675

[167] Zhang, J. and H. Guo (2014) "Transferability of Coarse-Grained Force Field for nCB Liquid Crystal Systems," The Journal of Physical Chemistry B, **118**(17), pp. 4647–4660, pMID: 24712306, https://doi.org/10.1021/jp411615f.
URL https://doi.org/10.1021/jp411615f

[168] Wang, Y. T., W. G. Noid, P. Liu, and G. A. Voth (2009) "Effective force coarse-graining," Phys. Chem. Chem. Phys., **11**(12), pp. 2002–2015.

[169] Brini, E., V. Marcon, and N. F. A. van der Vegt (2011) "Conditional reversible work method for molecular coarse graining applications," Phys. Chem. Chem. Phys., **13**(22), pp. 10468–74.

[170] Brini, E., C. R. Herbers, G. Deichmann, and N. F. A. van der Vegt (2012) "Thermodynamic transferability of coarse-grained potentials for polymer-additive systems," Phys. Chem. Chem. Phys., **14**, pp. 11896–11903.
URL http://dx.doi.org/10.1039/C2CP40735C

[171] Mullinax, J. W. and W. G. Noid (2010) "Recovering physical potentials from a model protein databank," Proc. Natl. Acad. Sci. USA, **107**, pp. 19867–72.

[172] See supplemental material at http://aip.scitation.org/doi/suppl/10.1063/1.4952422 for additional plots that describe the CG potentials, the resulting equilibrium structure, and the pressure-matching method.

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

[174] 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**, pp. 10896 – 10913.

[175] Moore, T. C., C. R. Iacovella, and C. McCabe (2014) "Derivation of coarse-grained potentials via multistate iterative Boltzmann inversion," The Journal of Chemical Physics, **140**(22), p. 224104, https://doi.org/10.1063/1.4880555.
URL https://doi.org/10.1063/1.4880555

[176] Liu, P., Q. Shi, H. Daume, and G. A. Voth (2008) "A Bayesian statistics approach to multiscale coarse graining," J. Chem. Phys., **129**, p. 214114.

[177] SchrÃŸder, T. B., N. P. Bailey, U. R. Pedersen, N. Gnan, and J. C. Dyre (2009) "Pressure-energy correlations in liquids. III. Statistical mechanics and thermodynamics of liquids with hidden scale invariance," The Journal of Chemical Physics, **131**(23), p. 234503, https://doi.org/10.1063/1.3265955.
URL https://doi.org/10.1063/1.3265955

[178] 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**, pp. 978–989.
URL http://dx.doi.org/10.1039/C3SM52671B

[179] Kullback, S. and R. A. Leibler (1951) "ON INFORMATION AND SUFFICIENCY," Ann. Math. Stat., **22**(1), pp. 79–86.

[180] Isihara, A. (1968) "GIBBS-BOGOLIUBOV INEQUALITY," J. Phys. A: Math., Nucl., Gen., **1**(5), pp. 539–548.

[181] Lyubartsev, A. P. and A. Laaksonen (1997) "Osmotic and activity coefficients from effective potentials for hydrated ions," Phys. Rev. E, **55**, pp. 5689–5696.
URL http://link.aps.org/doi/10.1103/PhysRevE.55.5689

[182] Mirzoev, A. and A. P. Lyubartsev (2011) "Effective solvent mediated potentials of Na+ and Cl- ions in aqueous solution: temperature dependence," Phys. Chem. Chem. Phys., **13**, pp. 5722–5727.
URL http://dx.doi.org/10.1039/C0CP02397C

[183] Barker, J., D. Henderson, and W. Smith (1969) "Pair and triplet interactions in argon," Molecular Physics, **17**(6), pp. 579–592, https://doi.org/10.1080/00268976900101451.
URL https://doi.org/10.1080/00268976900101451

[184] van der Hoef, M. A. and P. A. Madden (1999) "Three-body dispersion contributions to the thermodynamic properties and effective pair interactions in liquid argon," The Journal of Chemical Physics, **111**(4), pp. 1520–1526, https://doi.org/10.1063/1.479390.
URL https://doi.org/10.1063/1.479390

[185] Baron, R., A. H. de Vries, P. H. Hünenberger, and W. F. van Gunsteren (2006) "Configurational Entropies of Lipids in Pure and Mixed Bilayers from Atomic-Level and Coarse-Grained Molecular Dynamics Simulations," J. Phys. Chem. B, **110**(31), pp. 15602–15614.

[186] Baron, R. and V. Molinero (2012) "Water-Driven CavityâĂŞLigand Binding: Comparison of Thermodynamic Signatures from Coarse-Grained and Atomic-Level Simulations," J. Chem. Theor. Comp., **8**(10), pp. 3696–3704.

[187] Lu, J., Y. Qiu, R. Baron, and V. Molinero (2014) "Coarse-Graining of TIP4P/2005, TIP4P-Ew, SPC/E, and TIP3P to Monatomic Anisotropic Water Models Using Relative Entropy Minimization," J. Chem. Theor. Comp., **10**(9), pp. 4104–4120.

[188] Wagner, J. W., J. F. Dama, A. E. P. Durumeric, and G. A. Voth (2016) "On the representability problem and the physical meaning of coarse-grained models," The Journal of Chemical Physics, **145**(4), p. 044108, https://doi.org/10.1063/1.4959168.
URL https://doi.org/10.1063/1.4959168

[189] Dunn, N. J. H. and W. G. Noid (2016) "Bottom-up coarse-grained models with predictive accuracy and transferability for both structural and thermodynamic properties of heptane-toluene mixtures." J. Chem. Phys., **144**, p. 204124.
URL http://aip.scitation.org/doi/full/10.1063/1.4952422

[190] Tuckerman, M. (2013) Statistical Mechanics: Theory And Molecular Simulation, Oxford University Press: Oxford, Great Britain.

[191] Hansen, J.-P. and I. R. McDonald (1990) Theory of Simple Liquids, 2 ed., Academic Press, San Diego, CA USA.

[192] Bahar, I., T. R. Lezon, A. Bakan, and I. H. Shrivastava (2010) "Normal Mode Analysis of Biomolecular Structures: Functional Mechanisms of Membrane Proteins," Chemical Reviews, **110**(3), pp. 1463–1497.

[193] Weeks, J. D. (2002) "CONNECTING LOCAL STRUCTURE TO INTERFACE FORMATION: A Molecular Scale van der Waals Theory of Nonuniform Liquids," Annual Review of Physical Chemistry, **53**(1), pp. 533–562, pMID: 11972018, https://doi.org/10.1146/annurev.physchem.53.100201.133929.
URL https://doi.org/10.1146/annurev.physchem.53.100201.133929

[194] 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**, pp. 7812–7824.

[195] Bereau, T. and M. Deserno (2009) "Generic coarse-grained model for protein folding and aggregation," J. Chem. Phys., **130**, p. 235106.

[196] 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), pp. 11207–11220.

[197] Ouldridge, T. E., A. A. Louis, and J. P. K. Doye (2011) "Structural, mechanical, and thermodynamic properties of a coarse-grained DNA model," J. Chem. Phys., **134**(8), 085101.
URL http://link.aip.org/link/?JCP/134/085101/1

[198] Karimi-Varzaneh, H. A., H.-J. Qian, X. Chen, P. Carbone, and F. Müller-Plathe (2011) "IBIsCO: a molecular dynamics simulation package for coarse-grained simulation," J Comput Chem, **32**(7), pp. 1475–87.

[199] Mirzoev, A. and A. P. Lyubartsev (2013) "MagiC: Software Package for Multiscale Modeling," J. Chem. Theor. Comp., **9**(3), pp. 1512–1520, http://pubs.acs.org/doi/pdf/10.1021/ct301019v.
URL http://pubs.acs.org/doi/abs/10.1021/ct301019v

[200] Lu, L. and G. A. Voth (2012) "The Multiscale Coarse-Graining Method," Adv. Chem. Phys., **149**, pp. 47–81.
URL http://dx.doi.org/10.1002/9781118180396.ch2

[201] Chaimovich, A. and M. S. Shell (2010) "Relative entropy as a universal metric for multiscale errors," Phys. Rev. E, **81**(6).

[202] Dunn, N. J. H., T. T. Foley, and W. G. Noid (2016) "Van der Waals perspective on coarse-graining: progress toward solving representability and

transferability problems." Acc. Chem. Res., **49**, pp. 2832–2840.
URL https://pubs.acs.org/doi/abs/10.1021/acs.accounts.6b00498

[203] Rudzinski, J. F., K. Lu, S. T. Milner, J. K. Maranas, and W. G. Noid (2017) "Extended Ensemble Approach to Transferable Potentials for Low-Resolution Coarse-Grained Models of Ionomers," Journal of Chemical Theory and Computation, **13**(5), pp. 2185–2201, pMID: 28399373, https://doi.org/10.1021/acs.jctc.6b01160.
URL https://doi.org/10.1021/acs.jctc.6b01160

[204] Abraham, M. J., T. Murtola, R. Schulz, S. PÃąll, J. C. Smith, B. Hess, and E. Lindahl (2015) "GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers," SoftwareX, **1-2**, pp. 19 – 25.
URL http://www.sciencedirect.com/science/article/pii/S2352711015000059

[205] Martyna, G. J., D. J. Tobias, and M. L. Klein (1994) "Constant pressure molecular dynamics algorithms," The Journal of Chemical Physics, **101**(5), pp. 4177–4189, https://doi.org/10.1063/1.467468.
URL https://doi.org/10.1063/1.467468

[206] Hummer, G., N. Gro/nbech-Jensen, and M. Neumann (1998) "Pressure calculation in polar and charged systems using Ewald summation: Results for the extended simple point charge model of water," The Journal of Chemical Physics, **109**(7), pp. 2791–2797, https://doi.org/10.1063/1.476834.
URL https://doi.org/10.1063/1.476834

[207] Espanol, P. and I. Zuniga (2011) "Obtaining fully dynamic coarse-grained models from MD," Phys. Chem. Chem. Phys., **13**, pp. 10538–10545.

[208] Davtyan, A., J. F. Dama, G. A. Voth, and H. C. Andersen (2015) "Dynamic force matching: A method for constructing dynamical coarse-grained models with realistic time dependence," The Journal of Chemical Physics, **142**(15), p. 154104, https://doi.org/10.1063/1.4917454.
URL https://doi.org/10.1063/1.4917454

[209] Chorin, A. J. (2003) "Conditional expectations and renormalization," Multiscale Model. Simul., **1**, pp. 105–18.

[210] Kalligiannaki, E., V. Harmandaris, M. A. Katsoulakis, and P. PlechÃąÄĎ (2015) "The geometry of generalized force matching and related information metrics in coarse-graining of molecular systems," The Journal of Chemical Physics, **143**(8), p. 084105, https://doi.org/10.1063/1.4928857.
URL https://doi.org/10.1063/1.4928857

[211] Shell, M. S. (2016) Adv. Chem. Phys., John Wiley & Sons, Inc.

[212] Hill, T. L. (1997) An introduction to statistical thermodynamics, Addison Wesley Publishing Company.

[213] Mullinax, J. W. and W. G. Noid (2010) "Reference state for the generalized Yvon-Born-Green theory: Application for a coarse-grained model of hydrophobic hydration," J. Chem. Phys., **133**, p. 124107.

[214] 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.

[215] Bussi, G., D. Donadio, and M. Parrinello (2007) "Canonical sampling through velocity rescaling," The Journal of Chemical Physics, **126**(1), p. 014101, https://doi.org/10.1063/1.2408420.
URL https://doi.org/10.1063/1.2408420

[216] Martyna, G. J., M. E. Tuckerman, D. J. Tobias, and M. L. Klein (1996) "Explicit reversible integrators for extended systems dynamics," Molecular Physics, **87**(5), pp. 1117–1157, https://doi.org/10.1080/00268979600100761.
URL https://doi.org/10.1080/00268979600100761

[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**, pp. 859–83.

[218] Skolnick, J. (2006) "In Quest of an Empirical Potential for Protein Structure Prediction," Curr. Opin. Struc. Biol., **16**, pp. 166–171.

[219] DeLyser, M. R. and W. G. Noid (2017) "Extending pressure-matching to inhomogeneous systems via local-density potentials," The Journal of Chemical Physics, **147**(13), p. 134111, https://doi.org/10.1063/1.4999633.
URL https://doi.org/10.1063/1.4999633

[220] Pagonabarraga, I. and D. Frenkel (2001) "Dissipative particle dynamics for interacting systems," The Journal of Chemical Physics, **115**(11), pp. 5015–5026, https://doi.org/10.1063/1.1396848.
URL https://doi.org/10.1063/1.1396848

[221] Moore, J. D., B. C. Barnes, S. Izvekov, M. LÃnsal, M. S. Sellers, D. E. Taylor, and J. K. Brennan (2016) "A coarse-grain force field for RDX: Density dependent and energy conserving," The Journal of Chemical Physics, **144**(10), p. 104501, https://doi.org/10.1063/1.4942520.
URL https://doi.org/10.1063/1.4942520

[222] Sanyal, T. and M. S. Shell (2016) "Coarse-grained models using local-density potentials optimized with the relative entropy: Application to implicit solvation," The Journal of Chemical Physics, **145**(3), p. 034109, https://doi.org/10.1063/1.4958629.
URL https://doi.org/10.1063/1.4958629

[223] Wagner, J. W., T. Dannenhoffer-Lafage, J. Jin, and G. A. Voth (2017) "Extending the range and physical accuracy of coarse-grained models: Order parameter dependent interactions," The Journal of Chemical Physics, **147**(4), p. 044113, https://doi.org/10.1063/1.4995946.
URL https://doi.org/10.1063/1.4995946

[224] Sheu, E. Y. (2002) "Petroleum Asphaltene Properties, Characterization, and Issues," Energy & Fuels, **16**(1), pp. 74–82, https://doi.org/10.1021/ef010160b.
URL https://doi.org/10.1021/ef010160b

[225] Behrouzi, M. and P. F. Luckham (2008) "Limitations of Size-Exclusion Chromatography in Analyzing Petroleum Asphaltenes: A Proof by Atomic Force Microscopy," Energy & Fuels, **22**(3), pp. 1792–1798, https://doi.org/10.1021/ef800064q.
URL https://doi.org/10.1021/ef800064q

[226] Murgich, J. (2003) "Molecular Simulation and the Aggregation of the Heavy Fractions in Crude Oils," Molecular Simulation, **29**(6-7), pp. 451–461, https://doi.org/10.1080/0892702031000148762.
URL https://doi.org/10.1080/0892702031000148762

[227] Pomerantz, A. E., M. R. Hammond, A. L. Morrow, O. C. Mullins, and R. N. Zare (2009) "Asphaltene Molecular-Mass Distribution Determined by Two-Step Laser Mass Spectrometry," Energy & Fuels, **23**(3), pp. 1162–1168, https://doi.org/10.1021/ef8006239.
URL https://doi.org/10.1021/ef8006239

[228] Klee, T., T. Masterson, B. Miller, E. Barrasso, J. Bell, R. Lepkowicz, J. West, J. E. Haley, D. L. Schmitt, J. L. Flikkema, T. M. Cooper, Y. Ruiz-Morales, and O. C. Mullins (2011) "Triplet Electronic Spin States of Crude Oils and Asphaltenes," Energy & Fuels, **25**(5), pp. 2065–2075, https://doi.org/10.1021/ef101549k.
URL https://doi.org/10.1021/ef101549k

[229] Andrews, A. B., J. C. Edwards, A. E. Pomerantz, O. C. Mullins, D. Nordlund, and K. Norinaga (2011) "Comparison of Coal-Derived and Petroleum Asphaltenes by 13C Nuclear Magnetic Resonance, DEPT, and XRS,"

Energy & Fuels, **25**(7), pp. 3068–3076, https://doi.org/10.1021/ef2003443.
URL https://doi.org/10.1021/ef2003443

[230] Sabbah, H., A. L. Morrow, A. E. Pomerantz, and R. N. Zare (2011) "Evidence for Island Structures as the Dominant Architecture of Asphaltenes," Energy & Fuels, **25**(4), pp. 1597–1604, https://doi.org/10.1021/ef101522w.
URL https://doi.org/10.1021/ef101522w

[231] Schuler, B., G. Meyer, D. PeÃśa, O. C. Mullins, and L. Gross (2015) "Unraveling the Molecular Structures of Asphaltenes by Atomic Force Microscopy," Journal of the American Chemical Society, **137**(31), pp. 9870–9876, pMID: 26170086, https://doi.org/10.1021/jacs.5b04056.
URL https://doi.org/10.1021/jacs.5b04056

[232] ChacÃşn-PatiÃśo, M. L., S. M. Rowland, and R. P. Rodgers (2017) "Advances in Asphaltene Petroleomics. Part 1: Asphaltenes Are Composed of Abundant Island and Archipelago Structural Motifs," Energy & Fuels, **31**(12), pp. 13509–13518, https://doi.org/10.1021/acs.energyfuels.7b02873.
URL https://doi.org/10.1021/acs.energyfuels.7b02873

[233] ——— (2018) "Advances in Asphaltene Petroleomics. Part 2: Selective Separation Method That Reveals Fractions Enriched in Island and Archipelago Structural Motifs by Mass Spectrometry," Energy & Fuels, **32**(1), pp. 314–328, https://doi.org/10.1021/acs.energyfuels.7b03281.
URL https://doi.org/10.1021/acs.energyfuels.7b03281

[234] Eyssautier, J., D. Frot, and L. BarrÃł (2012) "Structure and Dynamic Properties of Colloidal Asphaltene Aggregates," Langmuir, **28**(33), pp. 11997–12004, pMID: 22827858, https://doi.org/10.1021/la301707h.
URL https://doi.org/10.1021/la301707h

[235] Tanaka, R., E. Sato, J. E. Hunt, R. E. Winans, S. Sato, and T. Takanohashi (2004) "Characterization of Asphaltene Aggregates Using X-ray Diffraction and Small-Angle X-ray Scattering," Energy & Fuels, **18**(4), pp. 1118–1125, https://doi.org/10.1021/ef034082z.
URL https://doi.org/10.1021/ef034082z

[236] Teklebrhan, R. B., L. Ge, S. Bhattacharjee, Z. Xu, and J. SjÃűblom (2012) "Probing StructureâĂŞNanoaggregation Relations of Polyaromatic Surfactants: A Molecular Dynamics Simulation and Dynamic Light Scattering Study," The Journal of Physical Chemistry B, **116**(20), pp. 5907–5918, pMID: 22512276, https://doi.org/10.1021/jp3010184.
URL https://doi.org/10.1021/jp3010184

[237] Lemarchand, C. A. and J. S. Hansen (2015) "Simple Statistical Model for Branched Aggregates: Application to Cooee Bitumen," The Journal of Physical Chemistry B, **119**(44), pp. 14323–14331, pMID: 26458140, https://doi.org/10.1021/acs.jpcb.5b08320.
URL https://doi.org/10.1021/acs.jpcb.5b08320

[238] Headen, T. F., E. S. Boek, and N. T. Skipper (2009) "Evidence for Asphaltene Nanoaggregation in Toluene and Heptane from Molecular Dynamics Simulations," Energy & Fuels, **23**(3), pp. 1220–1229, https://doi.org/10.1021/ef800872g.
URL https://doi.org/10.1021/ef800872g

[239] Lemarchand, C. A., M. L. Greenfield, and J. S. Hansen (2016) "Dynamics and Structure of BitumenâĂŞWater Mixtures," The Journal of Physical Chemistry B, **120**(24), pp. 5470–5480, pMID: 27248331, https://doi.org/10.1021/acs.jpcb.6b01451.
URL https://doi.org/10.1021/acs.jpcb.6b01451

[240] Teklebrhan, R. B., C. Jian, P. Choi, Z. Xu, and J. SjÃűblom (2016) "Competitive Adsorption of Naphthenic Acids and Polyaromatic Molecules at a TolueneâĂŞWater Interface," The Journal of Physical Chemistry B, **120**(50), pp. 12901–12910, pMID: 27959570, https://doi.org/10.1021/acs.jpcb.6b07938.
URL https://doi.org/10.1021/acs.jpcb.6b07938

[241] ——— (2016) "Competitive Adsorption of Naphthenic Acids and Polyaromatic Molecules at a TolueneâĂŞWater Interface," The Journal of Physical Chemistry B, **120**(50), pp. 12901–12910, pMID: 27959570, https://doi.org/10.1021/acs.jpcb.6b07938.
URL https://doi.org/10.1021/acs.jpcb.6b07938

[242] Liu, J., Y. Zhao, and S. Ren (2015) "Molecular Dynamics Simulation of Self-Aggregation of Asphaltenes at an Oil/Water Interface: Formation and Destruction of the Asphaltene Protective Film," Energy & Fuels, **29**(2), pp. 1233–1242, https://doi.org/10.1021/ef5019737.
URL https://doi.org/10.1021/ef5019737

[243] Lv, G., F. Gao, G. Liu, and S. Yuan (2017) "The properties of asphaltene at the oil-water interface: A molecular dynamics simulation," Colloids and Surfaces A: Physicochemical and Engineering Aspects, **515**, pp. 34 – 40.
URL http://www.sciencedirect.com/science/article/pii/S0927775716310214

[244] Wang, S., J. Xu, and H. Wen (2014) "The aggregation and diffusion of asphaltenes studied by GPU-accelerated dissipative particle dynamics,"

Computer Physics Communications, **185**(12), pp. 3069 – 3078.
URL http://www.sciencedirect.com/science/article/pii/S0010465514002604

[245] Voth, G. A. (ed.) (2008) Coarse-graining of condensed phase and biomolecular systems, CRC Press, Boca Raton, FL USA.

[246] Peter, C. and K. Kremer (2010) "Multiscale simulation of soft matter systems," Faraday Disc., **144**, pp. 9–24.

[247] ——— (2009) "Multiscale simulation of soft matter systems–from the atomistic to the coarse-grained level and back," Soft Matter, **5**(22), pp. 4357–4366.

[248] Hernandez-Rojas, J., F. Calvo, and D. J. Wales (2016) "Coarse-graining the structure of polycyclic aromatic hydrocarbons clusters," Phys. Chem. Chem. Phys., **18**, pp. 13736–13740.
URL http://dx.doi.org/10.1039/C6CP00592F

[249] Wang, J. and A. L. Ferguson (2016) "Mesoscale Simulation of Asphaltene Aggregation," The Journal of Physical Chemistry B, **120**(32), pp. 8016–8035, pMID: 27455391, https://doi.org/10.1021/acs.jpcb.6b05925.
URL https://doi.org/10.1021/acs.jpcb.6b05925

[250] Wang, J., M. A. Gayatri, and A. L. Ferguson (2017) "Mesoscale Simulation and Machine Learning of Asphaltene Aggregation Phase Behavior and Molecular Assembly Landscapes," The Journal of Physical Chemistry B, **121**(18), pp. 4923–4944, pMID: 28418682, https://doi.org/10.1021/acs.jpcb.7b02574.
URL https://doi.org/10.1021/acs.jpcb.7b02574

[251] Aguilera-Mercado, B., C. Herdes, J. Murgich, and E. A. MÃijller (2006) "Mesoscopic Simulation of Aggregation of Asphaltene and Resin Molecules in Crude Oils," Energy & Fuels, **20**(1), pp. 327–338, https://doi.org/10.1021/ef050272t.
URL https://doi.org/10.1021/ef050272t

[252] Zhang, S.-F., L. Sun, J.-B. Xu, H. Wu, and H. Wen (2010) "Aggregate Structure in Heavy Crude Oil: Using a Dissipative Particle Dynamics Based Mesoscale Platform," Energy & Fuels, **24**(8), pp. 4312–4326, https://doi.org/10.1021/ef1003446.
URL https://doi.org/10.1021/ef1003446

[253] Weeks, J. D., D. Chandler, and H. C. Andersen (1971) "Role of Repulsive Forces in Determining the Equilibrium Structure of Simple Liquids," The Journal of Chemical Physics, **54**(12), pp. 5237–5247, https://doi.org/10.1063/1.1674820.
URL https://doi.org/10.1063/1.1674820

[254] Bussi, G., D. Donadio, and M. Parrinello (2007) "Canonical sampling through velocity rescaling," The Journal of Chemical Physics, **126**(1), p. 014101, https://doi.org/10.1063/1.2408420.
URL https://doi.org/10.1063/1.2408420

[255] Pronk, S., S. PÃąll, R. Schulz, P. Larsson, P. Bjelkmar, R. Apostolov, M. R. Shirts, J. C. Smith, P. M. Kasson, D. van der Spoel, B. Hess, and E. Lindahl (2013) "GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit," Bioinformatics, **29**(7), pp. 845–854, http://dx.doi.org/10.1093/bioinformatics/btt055.
URL http://dx.doi.org/10.1093/bioinformatics/btt055

[256] Sedghi, M., L. Goual, W. Welch, and J. Kubelka (2013) "Effect of Asphaltene Structure on Association and Aggregation Using Molecular Dynamics," The Journal of Physical Chemistry B, **117**(18), pp. 5765–5776, pMID: 23581711, https://doi.org/10.1021/jp401584u.
URL https://doi.org/10.1021/jp401584u

[257] Aray, Y., R. HernÃąndez-Bravo, J. G. Parra, J. RodrÃĳguez, and D. S. Coll (2011) "Exploring the StructureâĂŞSolubility Relationship of Asphaltene Models in Toluene, Heptane, and Amphiphiles Using a Molecular Dynamic Atomistic Methodology," The Journal of Physical Chemistry A, **115**(42), pp. 11495–11507, pMID: 21905686, https://doi.org/10.1021/jp204319n.
URL https://doi.org/10.1021/jp204319n

[258] Wang, J., M. Gayatri, and A. L. Ferguson (2018) "Coarse-Grained Molecular Simulation and Nonlinear Manifold Learning of Archipelago Asphaltene Aggregation and Folding," The Journal of Physical Chemistry B, **122**(25), pp. 6627–6647, pMID: 29856608, https://doi.org/10.1021/acs.jpcb.8b01634.
URL https://doi.org/10.1021/acs.jpcb.8b01634

[259] Jian, C. and T. Tang (2014) "One-Dimensional Self-Assembly of Polyaromatic Compounds Revealed by Molecular Dynamics Simulations," The Journal of Physical Chemistry B, **118**(44), pp. 12772–12780, pMID: 25302404, https://doi.org/10.1021/jp506381z.
URL https://doi.org/10.1021/jp506381z

[260] Pacheco-SÃąnchez, J. H., F. ÃĄlvarez RamÃŋrez, and J. M. MartÃŋnez-MagadÃąn (2004) "Morphology of Aggregated Asphaltene Structural Models," Energy & Fuels, **18**(6), pp. 1676–1686, https://doi.org/10.1021/ef049911a.
URL https://doi.org/10.1021/ef049911a

[261] Glaser, J., T. D. Nguyen, J. A. Anderson, P. Lui, F. Spiga, J. A. Millan, D. C. Morse, and S. C. Glotzer (2015) "Strong scaling of

general-purpose molecular dynamics simulations on GPUs," Computer Physics Communications, **192**, pp. 97 – 107.
URL http://www.sciencedirect.com/science/article/pii/S0010465515000867

[262] Anderson, J. A., C. D. Lorenz, and A. Travesset (2008) "General purpose molecular dynamics simulations fully implemented on graphics processing units," Journal of Computational Physics, **227**(10), pp. 5342 – 5359.
URL http://www.sciencedirect.com/science/article/pii/S0021999108000818

Vita

Nicholas J. H. Dunn

Nicholas J. H. Dunn was born in Portland, ME and raised in Bangor, ME. He received his Bachelors of Science degree in Chemistry at Union College in Schenectady, NY in 2011. During his time at Union, Nicholas worked in the multidisciplinary Aerogel Lab under the supervision Mary Carroll. While working in the Aerogel Lab, Nicholas focused on the synthesis of catalytically active nickel-alumina aerogel materials for exhaust processing.

Nicholas started his graduate studies in the fall of 2011 at Pennsylvania State University where he joined Will Noid's theoretical and computational chemistry research group. During his first years at Penn State, Nicholas took additional coursework in mathematics and physics to fill in gaps in his theoretical chemistry background. As an extension of his teaching assistant duties, Nicholas developed and wrote a polymer identification laboratory exercise for an introductory materials chemistry course. Nicholas was fortunate to have the opportunity to mentor Besha Gutama during her time as an REU summer student in the Noid lab. As he became more experienced in systems administration, Nicholas took over management of the group's local computing resources and high-performance computing software environment.

Nicholas is currently living in Saint Paul, MN with his wife of seven years, Emily Dunn. Nicholas and Emily moved to Minnesota in the fall of 2016 for Nicholas' new position as a Scientific Computing Consultant at the Minnesota Supercomputing Institute in Minneapolis, MN.