UNIVERSITY OF CALIFORNIA
Santa Barbara

Next-generation coarse-grained models for molecular dynamics simulations of fluid phase equilibria and protein biophysics using the relative entropy

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Chemical Engineering

by

Tanmoy Sanyal

Committee in Charge:

Professor M. Scott Shell, Chair

Professor Baron Peters

Professor Glenn H. Fredrickson

Professor Joan-Emma Shea

Professor Linda R. Petzold

December 2018

The Dissertation of
Tanmoy Sanyal is approved:

Professor Baron Peters

Professor Glenn H. Fredrickson

Professor Joan-Emma Shea

Professor Linda R. Petzold

Professor M. Scott Shell, Committee Chair

December 2018

Next-generation coarse-grained models for molecular dynamics simulations of fluid phase equilibria and protein biophysics using the relative entropy

by

Tanmoy Sanyal

Acknowledgements

I acknowledge the constant and incredible support from my fiancée Sampurna. In spite of being 1800 miles and two time zones away for the last five years, not a day has passed by when she has failed to encourage me to strive for success or bear my rants during unproductive periods of my PhD. Being a doctoral student in computational sciences herself, she has also helped me from time to time to workshop my ideas, taking time out from her own research work.

Being an international student, I really did not have an opportunity to visit UCSB and talk to the chemical engineering faculty, before I began my graduate studies here. So, words fail me when I have to describe how lucky I am to have Professor M. Scott Shell as my PhD advisor. Much of the ideas and methods presented in this thesis took shape through continuous, insightful and timely discussions with Professor Shell over the past five years. Beyond core academic advice, you also helped me to imbibe organizational and oratorial skills, that I will carry with me throughout the rest of my life. As I stand at the final frontiers of my graduate student phase of life, I only hope that I have been able to imbibe at least some of the deep analytical outlook that you have towards science, and that I may be able to hone it further and apply it in my future career efforts.

I would like to thank the members of my dissertation committee: Professors Baron Peters, Glenn H. Fredrickson, Joan-Emma Shea and Linda R. Petzold for your continued assessment and valuable advice through annual committee meetings and personal meetings. I would also like to thank Professors Glenn H. Fredrickson and Linda R. Petzold for being incredible course instructors and help me develop a fundamental understanding of essential graduate level mathematics.

Thank you to all the members (past and present) of the Shell and Peters groups: Dave, Jacob, Nick, Nikolai, Kartik, Salman, Craig, Christian and Bryan. Your company and the valuable discussions I had with each of you, have been an inspiration to go to work, and will always remain one of the high points of my PhD experience. I also want to thank Scott P. Carmichael from the Shell group who mentored me when I first joined the group and helped me master much of the programming skills that I have now.

I would be remiss to not mention Sourav da, Ranajay da, Pritam da, Soupitak da and Anish da: a motley crowd of graduate students and postdoctoral scholars who supported me throughout my journey at UCSB. I refer to them not by their full names but with “da” which in my native tongue means brother. Indeed, I found a family in them away from home.

I would like to thank my parents for their constant personal sacrifices throughout my school and college years to ensure that I got the best possible education. I would really not have been able to make it to a PhD program if it had not been for them.

Finally, I am somewhat agnostic, but I do believe there is some unknown form of higher power out there somewhere, that has brought me in connection to all the amazing people I mentioned so far to lead me on this incredible journey.

Curriculum Vitæ

Tanmoy Sanyal

Education

2018	Doctor of Philosophy in Chemical Engineering, University of California, Santa Barbara.
2013	Bachelor of Technology (Hons.) and Master of Technology Dual Degree in Chemical Engineering, Indian Institute of Technology, Kharagpur.

Journal Articles

- T. Sanyal and M.S. Shell, *Coarse-grained models using local-density potentials optimized with the relative entropy: Application to implicit solvation*, J. Chem. Phys. 145 (**3**), 034109 (2016)

- *Invited article for the Ken A. Dill Festschrift*
  T. Sanyal and M.S. Shell, *Transferable coarse-grained models of liquid-liquid equilibrium using local density potentials optimized with the relative entropy*, J. Phys. Chem. B 122 (**21**), 5678-5693 (2018)

- M.P. Howard, W.F. Reinhart, T. Sanyal, M.S. Shell, A. Nikoubashman, A.Z. Panagiotopoulos, *Evaporation-induced assembly of colloidal crystals*, J. Chem. Phys. 149 (**9**), 094901 (2018)

Conference presentations

- 11th Annual Clorox-Amgen Graduate Student Symposium, UC Santa Barbara, CA, **Poster**: *Next generation coarse-grained models of generic heteropeptides for folding and self-assembly* (2018)

- 10th Annual Amgen-Clorox Graduate Student Symposium, UC Santa Barbara, CA, **Talk**: *A new approach to computationally fast coarse-grained models of hydrophobic interactions in fluids* (2018)

- AIChE Annual Meeting, San Francisco, CA, **Talk**: *Improved coarse-grained models of solvation using local density dependent interactions with the relative entropy* (2016)

- Berkeley Mini Stat. Mech. Meeting, Berkeley, CA, **Poster**: *Building coarse-grained models of solvation using local density dependent interactions with the relative entropy*, (2015)

Research Students Mentored

- Jun-Lu (2019, expected): Self-assembly in surface-tetherd dual peptide systems due to polymeric conjugation.

Abstract

Next-generation coarse-grained models for molecular dynamics simulations of fluid phase equilibria and protein biophysics using the relative entropy

Tanmoy Sanyal

Coarse grained models for molecular dynamic simulations of liquid structure and protein folding and self-assembly have been the subject of decades of research efforts. Although such models enable probing into longer length and time scales, they are still limited in accuracy and scale poorly to thermodynamic conditions or chemical entities beyond the ones at which they are developed. In this work, I leverage a powerful coarse graining framework based on an information theoretic metric known as the relative entropy to design novel coarse grained models for equilibrium phase behavior in liquid mixtures that correctly address the relevant manybody physics critically involved in phase behavior and thus, can provide structurally accurate descriptions of macroscopic phase separation across a large range of mixture compositions. I also develop protein models that are "next-generation" in the sense that they are systematically extendable in complexity while depending minimally on experimental data. These protein models offer remarkable predictive accuracy for folding of single proteins and insights into large scale protein self-assembly commonly seen in neurodegenerative pathologies such as Alzheimer's and prion diseases. This dissertation develops and applies powerful coarse-graining techniques to meet longstanding challenges in the field and elevate coarse grained models from being "toy" systems to accurate, "production-level" tools for the development of biotechnologies and advanced materials.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation and Applications

The objective of this thesis is to construct computational algorithms for developing bottom up coarse grained (CG) models for molecular dynamics (MD) simulation, which can be broadly useful in studying two important families of physical phenomena: phase behavior, especially in liquids and protein folding and self-assembly. MD simulations with detailed atomistic resolution were first introduced by Alder and Wainright in the the 1950s, as a simplified classical mechanical description of nature, treating atoms as point particles and evolving their positions in time by integrating Newton's laws of motion.[2] To accurately resolve fast degrees of freedom such as bond vibrations, and to preserve numerical stability, such time integration can only progress very slowly, typically using timescales of 1 fs or less. Owing to the rapid progress in hardware and algorithms, atomistic MD has advanced greatly in terms of system size.[3] However, even with today's resources, MD is still limited to timescales of microseconds and lengthscales of tens of nanometers. Coarse graining MD simulations can tackle this issue by reducing degrees of freedom, so that longer length and time scales can be probed.

CG models have been widely used in studying the driving forces like hydrophobic interactions in liquid phase equlibrium.[4–6] Hydrophobic interactions are ubiqutious in nature. They are one of the most important driving forces for self-assembly in biomolecules like proteins and DNA. At the macroscale, hydrophobic interactions are responsible for phase equilibrium in multicomponent fluid mixtures which in turn is the guiding physical principle for designing a wide variety of everyday consumer products like shampoos, gels, toothpastes, detergents, etc. Closely related to phase equilbrium is the segregation or transition between phases, and for liquid mixtures, this forms the basis of unit operations like liquid-liquid phase segregation which has widespread applications in food processing, organic synthesis, petroleum refineries, renewable energy, nuclear reprocessing, and biotechnology. Water mediated hydrophobic interactions are manifested through water's orientational hydrogen bonds with itself and other solutes.

CG models of water or aqueous systems often coarse grain water into spherical isotropic particles, and in some cases completely remove water to construct what are called "implicit water" models.[7–13] Removing water completely is attractive from a practical standpoint, since it takes up about 70% of the computational effort even for small solutes. Single-site CG or implicit water models essentially remove water's orientational degrees of freedom and its hydrogens, and thus, must necessarily renormalize these interactions suitably to ensure the correct thermophysical behavior of the CG model. This is challenging to do with current CG models,the details of which are explained in section 1.2. Chapters 2 and 3 of this dissertation introduces simple methods to solve both the problems of developing accurate implicit water models and structurally faithful CG descriptions of phase equilibrium in liquid-liquid mixtures.

Protein folding and self-assembly mechanisms form the basis of the cellular machinery necessary for life. They are also key in biotechnological applications such as high throughput drug design.[14–16] CG models for peptides have been mostly useful in uncovering the driving forces responsible for the folding process, but so far, have not achieved structurally accurate predictions of structure. Further, CG peptide models typically embed bioinformatic and experimental data that forestalls the systematic incorporation of synthetic chemical constructs or extension to sequences for which native structures have not been solved. Closely related is the phenomena of protein misfolding and aggregation which is responsible for several neuro-degenerative diseases like Alzheimer's, Parkinson's, ALS and a wide family of prion diseases.[17,18] The lack of an accurate CG peptide model further complicates macrostructure prediction of protein aggregates involved in the aforementioned conditions. In this dissertation, we develop "next-generation" peptide models for structure prediction and apply CG polypeptide models to probe the connection between monomer conformation and the shape and stabilities of fibrillar agglomerates that arise from their anomalous self-assembly. This may serve as a minimal model system for investigating the recently observed conformation dependent differences in aggregation behavior of the tau protein implicated in Alzheimer's and Pick's diseases.[19,20] By next-generation, we mean CG models that can be systematically extended to include sequence specific effects and intermolecular interactions such as with other peptides (or ligands or surfaces) without explicitly requring reparameterization for different sequences.

It should be noted that each of the projects addressed in this thesis illustrates proofs of principle through candidate test systems. Scaling up these methods will require more atomistic simulation data using preferably the most accurate and current AA forcefields to train "production-level" CG models that can be used to not only rationalize the design principles for liquid phase equilibria and protein folding and self-assembly, but also match experimental data as closely as possible. Such scaling-up may also require using some experimental data during the parameterization, which is briefly discussed in chapter 6.

1.2 Bottom-up coarse-graining and the problem of transferability

The most accurate description of reality at small molecular scales is inevitably quantum mechanics, and consequently the most detailed molecular simulation would be one that solves Schrodinger's equation for all the particles involved. However, beyond simple one or two electron systems, fully quantum descriptions embed significant manybody physics and as such is intractable by today's computes. Thus atomistic MD simulations are already a coarse grained description of nature that renormalizes the probabilistic nature of electronic distribution around atomic nucleii and the net nuclear-electron interaction into effective parameters like partial charges, bond lengths and angles, and Lennard Jones interactions between point-particle like atoms. Further up this coarse graining cascade, lie the class of CG models addressed in this dissertation. These models involve clustering groups of atoms into CG “pseudoatoms” or sites. Parameterizing such a CG model typically amounts to “designing” the interactions between the CG sites such that the model can reproduce relevant thermophysical properties of the atomistic system. Coarse graining can be top-down, where CG interactions are paramterized by directly fitting thermophysical observables either from experiment or from ensemble averages from all-atom (AA) trajectories, or a combination of both. In this work, we focus on bottom up strategies where degrees of freedom are systematically removed from a AA system.

The two elements in bottom up coarse graining is a mapping operation $\mathbf{R} = \mathbf{M}(\mathbf{r})$ and a CG forcefield or Hamiltonian $U_{\mathrm{CG}}(\mathbf{R})$. $\mathbf{M}(\mathbf{r})$ is a mapping operator that translates atomistic degrees of freedom $\mathbf{r}$ to a coarse grained configuration $\mathbf{R}$. CG sites are typically placed at the center of mass of the corresponding group of atoms which reduces $\mathbf{M}$ to a matrix. The CG forcefield $U_{\mathrm{CG}}(\mathbf{R})$ is effectively a potential of mean force (PMF) that can be written as:

$$U_{\mathrm{CG}}(\mathbf{R}) = -k_B T \ln \int_V d\mathbf{r} e^{-\beta U_{\mathrm{AA}}(\mathbf{r})} \, \delta[\mathbf{R} - \mathbf{M}(\mathbf{r})] \tag{1.1}$$

where $U_{\mathrm{AA}}(\mathbf{r})$ is the AA forcefield, T is the ambient temperature and k_B is the Boltzmann constant. The Dirac delta projects the atomistic potential energy surface on to the CG degrees of freedom $\mathbf{R}$. The resulting CG energy landscape is much smoother enabling faster equilbrium in a MD simulation. Because of the coarse graining process, CG degrees of freedom are inherently coupled to each other so that U_{CG} is a highly multibody interaction. However, current CG models approximate the CG PMF as a pair potential for computational expediency. Neglecting manybody effects in CG models limits their transferability to thermodynamic states (temperature, density, etc.) beyond which they are parameterized.

State point transferability of CG models not only reduces the tedium of reparameterization at every state, it is necessary for simulations in the isothermal-isobaric (NPT) ensemble. NPT simulations involve constant fluctuations in bulk density, such that an accurate CG model must be transferable across the spectrum of densities sampled by the system. Since the average distance between any two CG particles can be related to the bulk density, CG models that are built entirely from pair potentials embded significant sensitivity to bulk density by neglecting the inherently multidimensional nature of the underlying CG PMF. Fig. 1.1 shows CG pair potentials for a single site CG model of water (parameterized with the relative entropy method, discussed in section 1.3) developed from an AA MD simulation with the TIP4P/2005 forcefield. Three CG pair potentials are constructed by holding the system density fixed at 1.10, 1.17 and 1.25 g/cc. The CG pair potential is extremely sensitive even to small changes in bulk density: the shape of the potential changes significantly from a deep inner core at lower densities to a shoulder at higher densities. Also, the inner well depth decreases by ~ 7.5 kcal/mol from 1.10 g/cc to 1.17 g/cc. As such, none of these potentials alone

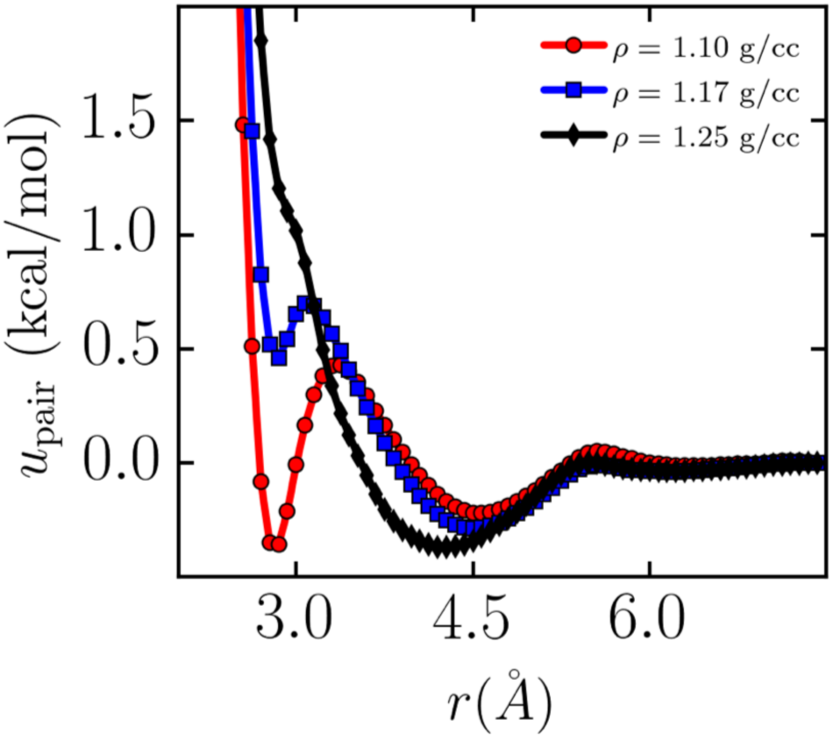


Figure 1.1: CG pair potentials (u_{pair}) for a single site CG model of water optimized (using the relative entropy method) from a atomistic simulation using the TIP4P/2005 water model. The box density is held fixed at 1.1, 1.17 and 1.25 g/cc. Even for small changes in bulk density, CG pair potentials vary considerably in shape (the inner well converts to a shoulder at 1.25 g/cc) and location of the inner well ($\sim$ 7.5 kcal/mol lower at 1.17 g/cc than at 1.10 g/cc)

can be used at a density different from the one at which they were paramterized. Chapters 2 and 3 introduce and use the LD potential as a minimal correction over CG pair potentials to embed multibody correlations in CG models of liquids and make them more transferable across densities.

Further, chemical transferability of peptide CG models may provide a simple way to rationalize relevant interactions in terms of the intrinsic properties of the peptide chain alone, and still achieve correct secondary structure prediction without having to reparameterize for different sequences. In chapter 4, we develop CG models of polypeptides that exclusively stabilize α-helical and β-sheet structures, and subsequently combine them to produce a hybrid CG model that transfers across both α and β regions of the space of backbone dihedrals.

1.3 Relative entropy coarse-graining

Determining an effective CG forcefield from a detailed accurate AA model is a nontrivial inverse design problem. A natural approach is to tune CG interactions to match specific simulation averaged properties or their distributions from the AA system. The literature points to two major families of such efforts: structure based coarse-graining and force-matching. Structure based methods such as inverse Monte-carlo[21] or iterative Boltzmann inversion[22] try to match pair correlations (or radial distribution functions (RDFs)) of the target AA system, and have enjoyed considerable success for modeling simple polymer and lipid bilyaers.[23,24] Force matching aims to match the average interatomic forces between AA and CG models and has been developed by Voth and co-workers through their Multiscale Coarse-Graining (MS-CG) framework.[25,26] The MS-CG approach has been useful in generating CG potentials for phospholipids and biomolecules including peptides and lipid membranes.[27–31]

In this dissertation, we use the relative entropy method pioneered by Shell and co-workers.[32–34] In this approach, instead of searching for candidate thermophysical properties that can be matched, one seeks to directly match the entire microstate probability distribution p_{AA} of the AA system with that of the CG model, p_{CG}. In principle, this ensures an accurate match of all relevant simulation observables (averaged or distributions) from the AA system.[33,35] The relative entropt between a CG model and its reference AA system is given by:

$$S_{\mathrm{rel}} = \int p_{\mathrm{AA}}(\mathbf{r}) \ \ln\left(\frac{p_{\mathrm{AA}}(\mathbf{r})}{p_{\mathrm{CG}}(\mathbf{M}(\mathbf{r}))}\right) d\mathbf{r} + S_{\mathrm{map}} \tag{1.2}$$

where, as mentioned before $\mathbf{M}(\mathbf{r}$ is the mapping matrix that coarse grains groups of atoms to representative CG sites. The integral proceeds over all AA microstates $\mathbf{r}$, though it can be recast using only the CG degrees of freedom $\mathbf{R}$. S_{map} is a mapping entropy that accounts for the degeneracy of the mapping process, i.e., it measures the number of different AA configurations that map to the same CG one. S_{map} is a function of only the mapping operator $\mathbf{M}$ and is independent of the CG forcefield U_{CG}. In the different projects addressed in this thesis, our choice of the mapping is motivated by simple physical arguments and past literature; for instance, removing water in implicit water models to reduce unnecessary computational cost, or coarse graining amino acids into four CG sites, which has been shown to better capture structural properties than two or three site models for peptides. Ultimately, determining the optimal AA to CG mapping is an open problem that has not been completely solved. But in this dissertation, $\mathbf{M}(\mathbf{r})$ is fixed prior to paramterizing the CG model, such that S_{map} is constant.

S_{rel} is an information theoretic metric, known as the Kullback-Liebler divergence in the statistics literature.[36] It measures the overlap between AA and CG microstate probability distributions and therefore quantifies the information lost in reducing the degrees of freedom from AA to CG representations. Minimizing the relative entropy with respect to the CG model parameters therefore provides a natural route to estimating these parameters and consequently U_{CG}. Relative entropy minimization is more general than structure or force matching techniques since the AA and CG models need not be MD simulations necessarily, e.g. either or both of them can be lattice models that utilize monte-carlo moves to sample their corresponding energy landscapes. Importantly however, the relative entropy method requires both AA and CG systems to be in thermodynamic equilibrium, although Espanol and Zuniga have discussed extensions of the relative entropy to dynamic CG simulations by minimizing a "dynamic" relative entropy (involving two time correlation functions) with respect to drift and diffusion terms in a Fokker-Plank like CG model.[37] Further, Sivak and Crooks have shown that the relative entropy is linked to the nonequilibrium work required to convert the CG model back to the correct AA ensemble.[38] The relative entropy approach has been used by Shell and co-workers to develop single-site CG water models that capture several bulk properties and hydrophobic interactions.[39,40] It has also been used to develop extremely accurate CG models of polyalanine which quantitatively reproduce folding behavior,[34] and have been used to explore self-assembly in surface tethered polyalanines conjugated with superhydrophobic polymers.[41]

1.4 Organization of thesis

Chapter 2 of this thesis introduces a simple and computationally efficient, mean-field manybody CG interaction called the "local density potential" that can supplement traditional CG pair interactions, to build back the much needed manybody effects inherent to CG models. We demonstrate the relative improvement over pair-potential-only CG models by paramterizing local density (LD) assisted forcefields for two candidate systems of small hydrophobes: an implicit water model to study folding and collapse of a superhydrophobic alkane-like polymer, and the co-operative aggregation of superhydrophobic methane like particles. Chapter 3 utilizes the LD potential to develop structurally accurate CG models that can predict macroscopic phase separation in binary liquid mixtures, specifically benzene in water, and are transferable across a wide range of mixture compositions. In chapter 4, we extend Carmichael and Shell's four-site CG polypeptide model[41] to construct robust CG backbone models for protein folding simulations. The backbone forcefields are combined with idealized and simplistic sidechain interactions, so called Gō models which depend on the native structure as input, to correctly fold both short peptide fragments and longer globular proteins, as well as very large (≥ 200 residues) sequences that is prohibitively difficult using all-atom MD. In chapter 5, we utilize CG polyvaline models to investigate the role of oligomer conformation on amyloid stability, in templated self-assembly of peptides commonly seen in neuro-degenerative diseases like Alzheimer's, ALS, Parkinson's, prion diseases, etc. Finally, chapter 6 summarizes the main results of this thesis and provides future directions for refining our CG protein forcefield and general numerical improvements to the relative entropy optimziation algorithms.

Chapter 2

Coarse-grained models using local-density potentials optimized with the relative entropy: Application to implicit solvation

2.1 Introduction

Molecular dynamics (MD) simulations with detailed atomic resolution have advanced greatly over the last few decades due to hardware and algorithm improvements, particularly in terms of tractable system size.[3] However, coarse-grained (CG) models have become essential counterparts to all-atom (AA) modeling, bridging their limitations potentially in orders of magnitude further in length and time scales through the elimination of unnecessary details and identification of emergent physical models. In particular, "bottom-up" CG strategies seek to systematically remove degrees of freedom by grouping two or more atoms into a single CG "pseudoatom" or "site." While such methods have seen a flurry of activity in the past two decades,[21–23,25,26,32,42–46]a major outstanding challenge is achieving computationally efficient ways to represent CG interactions faithfully. In principle, by integrating out degrees of freedom in an AA model, the appropriate effective interaction "potential" between the remaining CG sites is a multidimensional potential of mean force (PMF), which is unique to an additive constant,[26,47–50]

$$W(\mathbf{R}) = -k_B T \ln \int_V d\mathbf{r} e^{-\beta u(\mathbf{r})} \, \delta[\mathbf{R} - \mathbf{M}(\mathbf{r})] \tag{2.1}$$

Here, V and T are the system volume and temperature, respectively, $u(\mathbf{r})$ is the inter-atomic potential that is a function of the atomic positions $\mathbf{r}$, and the integral spans all values of r consistent with the system volume V. $\mathbf{M}(\mathbf{r})$ is a "mapping function" that translates an atomic configuration $\mathbf{r}$ to a CG one $\mathbf{R}$. For the special case in which the center of mass of a group of atoms is mapped onto a single CG site, $\mathbf{M}$ becomes a $n \times N$ matrix, where n is the number of atoms in the AA system and N is the number of CG sites. The Dirac delta function within the integral thus serves to project the AA potential energy surface along the reduced degrees of freedom ($\mathbf{R}$) of the CG model.

Bottom-up strategies seek CG forcefields that well-approximate W, but two challenges arise. First, W is by nature a multibody interaction because CG degrees of freedom are highly coupled through the AA ones that were integrated out, and in turn, it is not well-modeled by computationally efficient (and traditionally used) pair nonbonded terms.[21,23,25,26,34] Thus, while many CG models are able to capture pair structure and related properties, higher order correlations and cooperative effects are often poorly represented.[51,52] Second, W is a free energy that has a state dependence because the integral in Eq. (2.1) involves the temperature and volume, and it is not usually clear how this dependence can be captured simply.[53–56] This makes it difficult to transfer the CG potential to states other than the one at which it was parameterized.

In this work, we explore the use of mean-field multibody potentials in the development of CG force fields and characterize the extent to which they improve on the multibody and transferability problems for systems involving the solvation of hydrophobic solutes. Specifically, we use "local density" (LD) potentials that assign an energy to a CG site based on the number of neighboring sites of a given type that lie within a predetermined cutoff distance. Such potentials augment the usual pair nonbonded interactions but remain computationally efficient, scaling similarly in simulation cost. This strategy is inspired by embedded atom and bond order potentials that attempt to capture multibody electronic effects in metal systems.[57–61] Here, however, we generalize the approach in CG systems, describe a systematic parameterization procedure, and test its application to CG implicit solvation models involving idealized alkane-like solutes in aqueous media.

The transferability problem, which refers to a CG forcefield's ability to predict thermophysical properties at states different from those at which it was parameterized, has been characterized by a number of groups.[4, 51–53, 55, 62–64] Proper transferability is important not because it eliminates the need to re-parameterize the CG model at different states, but because it ensures thermodynamic consistency and validity of the CG model in all statistical–mechanical ensembles.[52] Several attempts have addressed the specific issue of transferability in CG models of aqueous solutions of macromolecular solutes. Mullinax and Noid proposed an extended ensemble approach that parameterizes a single CG model by combining information from an "ensemble" of AA reference simulations at different state points, thus fixing a priori the regime of transferability of the CG model in state-point space.[63] Villa, Peter, and van der Vegt used a modified iterative Boltzmann inversion scheme[22] to develop CG models of aqueous solutions of benzene, observing that the use of pair potentials for intermolecular interactions limited the regime of transferability of the models to dilute concentrations.[4] Other recent developments addressing transferability include bulk and local density based corrections to CG forcefields as ways to account for multibody effects not taken into account by CG pair potentials, discussed below.[64–67]

Are there computationally efficient ways to incorporate multibody effects into bottom up CG models, to improve their fidelity and transferability? One approach adds three (or higher) body terms to the CG forcefield. Indeed, Molinero and co-workers used three body Stillinger-Weber[68] potentials to describe hydrogen bonding in a single site CG model of water,[69,70] which has been shown to have excellent accuracy compared to detailed AA water models. While this approach has even been successfully translated to non-bulk scenarios,[71] CG potentials for water might be simple to intuit and implement because of its tetrahedral geometry, whereas general three body potentials for arbitrary interactions may prove conceptually and computationally more difficult to parameterize.[72] Moving towards broader approaches, Voth and co-workers developed a method to systematically parameterize generalized Stillinger-Weber forms within the multiscale coarse-graining force-matching framework, which they also used to create CG models of water.[73] Das and Andersen later proposed an even more general three-body scheme using adaptively applied “multiresolution” functions similar to wavelets to provide a larger basis set for the CG forcefield.[72]

Here, we explore an alternative, mean-field approach to incorporate multibody effects, whereby a particle's energy is modulated by the local density of neighboring CG sites around it. We are inspired by the Embedded Atom Method (EAM) of Daw and Baskes,[74] which was a historical improvement on internuclear CG pair potentials used to model metallic cohesion.[57,74] In this case, the electronic degrees of freedom are coarse-grained out of the picture, but the contribution of the local electronic environment is captured by assigning an effective electron density to each nuclei that gives rise to an "embedding" energy. Such an approach is intrinsically multibody in nature because the embedding energy can depend nonlinearly on the local electron density. At the same time, it is computationally efficient, requiring only two pair interaction loops: one to compute the densities at each nuclei and the other to evaluate energies and forces. Interestingly, Ercolessi and Adams also used similar manybody potentials in their seminal paper that introduced the force-matching algorithm.[42]

We propose to expand the EAM framework to "local density" potentials that can be applied generally to CG models as a simple way to include multibody effects. In this case, a site experiences a local density energy that is modulated by the number of sites of a given type (or types) within a cutoff radius. To parameterize and develop the form of these potentials, we propose to use the relative entropy coarse graining framework,[32,34] which provides a general optimization strategy to minimize the information lost in a CG model, relative to a reference target AA system.

We note several closely related efforts in the literature. Allen and Rutledge proposed global and local density-based corrections to standalone CG pair potentials in constructing implicit solvent models.[65,66] They parameterized the corrections in terms of the excess chemical potential for the transfer of solute groups from a "solvent exposed state to a fully screened environment," since the relative hydrophobicity of a solute depends on the change in chemical potential when passing from a solvent-exposed to a solute-screened state. In addition, Izvekov et al. proposed CG models using pair potentials whose corresponding pair forces vary with the average local density of the two interaction sites;[67] the pair forces and potential at each local density are determined through force-matching on a system at the same bulk density. (We note that, in a very recent work by Moore et al., this CG model was extended to include density dependence through a form of the CG forcefield that conserves energy and is amenable to a dissipative-particle-dynamic treatment.[75]) Finally, Dunn and Noid investigated the effect of global volume dependent corrections to pair CG potentials, determined by matching the ensemble-averaged pressure between the AA and CG representations.[64] Their approach was able to accurately capture the bulk density, compressibility, and pressure in pure liquids like n-heptane and toluene.

It is instructive to compare our local density formalism with the aforementioned flavors of density dependent CG potentials. First, the above methods involve parameterization that happens at the global level (e.g. pair interactions are determined at a fixed bulk density). In the first two approaches, the use of local densities then either corrects[65, 66] or reformulates and applies at the local level[67] these globally determined interactions (e.g. pair potentials are selected by a local density). In contrast, the approach that we consider directly optimizes all potentials from statistics of local densities in the reference ensemble, without the need for multiple systems at varying bulk conditions and the approximations/assumptions needed to translate them to the local level. Second, the use of a mean-field term for the local density potential ensures that, formally, all forces are at most pairwise in computation cost, as shown by Frenkel and Pagonabarraga.[76] In cases where a pair potential is modulated by an average pair local density,[67] instead of using a separate mean-field local density additive term, forces incur three body loops in principle to account for the influence of third-party sites on the local density and hence interactions associated with a given pair. Third, the robustness of the relative entropy minimization framework allows us to simultaneously and generally optimize both the CG pair and the LD potential components of the forcefield, instead of relying on different objective functions for separate construction of the two.[75]

As an application of our approach, we use local density potentials to develop bottom-up implicit aqueous solvation models. Water is a ubiquitous solvent, particularly in biology, and all-atom water in simulations of solvated macromolecules can dominate computational effort,[77–80] making it an attractive target for removal during coarse-graining. In implicit models the effect of solvent is incorporated into the effective solute interatomic interactions, and a wide range of approaches have been proposed.[7–13] The approach here is unique in that we use a bottom-up coarse-graining technique to uncover the form of the potential from reference AA simulations with explicit water. Specifically, we study the hydrophobic collapse of a polymer and the aggregation of small hydrophobes in water, where the water is coarse-grained out entirely.

Water-mediated, and in particular hydrophobic, interactions are critical to many biophysical phenomena, from protein folding to membrane formation,[81–83] and can be highly multibody in nature.[84,85] For example, early studies reported that free energy for trimerization of methanes in water cannot be decomposed into a sum of two-body terms obtained from the dimerization process.[86–89] While there has been some controversy as to whether the three-body effect has a positive[86] or an inhibitive[87,88] (anti-cooperative) effect on methane aggregation, there is evidence that it is non-negligible even for moderately high solute concentration.[89,90] Higher order correlations may also be relevant to electrostatic interactions in water clusters (through underlying polarization effects)[91] and more generally.[92]

Local density potentials are one strategy to capture such multibody effects in implicit solvation models, but they also offer the ability to directly control various particle number fluctuations in the neighborhood of a solute. This point is key as the recent theory[93,94] and simulations[95] suggest that hydrophobic interactions may be understood in terms of fluctuations in the local density of water vicinal to a solute. Indeed, Garde and coworkers have shown that local water density fluctuations are an important signature of molecular hydrophobicity and have theoretical connections[96,97] to compressibility of the solute hydration shell of water molecules. For example, local density fluctuations in water near a solvated polymer directly affect the associated hydration free energy and hence its collapse characteristics.[1,98–100] Below we explore the extent to which local density potentials can improve on CG models of aqueous solvation of hydrophobic solutes.

The remainder of this chapter is organized as follows. Section 2.2 presents the mathematical structure of the local density potential and outlines its theoretical connections to multibody effects. Section 2.3 uses relative entropy minimization to parameterize these CG models for two hydrophobic solute systems and examines their ability to predict structural metrics, including when transferred to systems of different kinds. Finally, Section 2.4 concludes the chapter.

2.2 Methods

2.2.1 Theoretical formulation for local density potentials

For the purposes of illustration, we first consider a CG model consisting of a single type of pseudoatom. In this case, we can measure the local density ρ_i of a site i as the total number of neighboring sites that are within a fixed cutoff r_c, using

$$\rho_i = \sum_{j \neq i} \varphi(r_{ij}), \tag{2.2}$$

where $\varphi(r_{ij})$ is an indicator function that is 1 for neighboring site j when its pair distance r_{ij} is within a cutoff r_c. In this manner, ρ_i measures a local coordination number for neighbors around site i. In turn, the contribution to the total LD energy, for each such atom i, is a (yet unspecified) function of the local density,

$$U_{\mathrm{LD}} = \sum_i f(\rho_i) \tag{2.3}$$

In practice, φ can be a smooth function that quickly but continuously decreases to zero at r_c to ensure the continuity of its first derivative and thus of the interaction forces in a MD simulation. We choose a computationally convenient form that does not require absolute pair distances or square root operations

$$\varphi(r) = \begin{cases} 1, & r \leq r_0 \\ c_0 + c_2 r^2 + c_4 r^4 + c_6 r^6, & r \in (r_0, r_c) \\ 0, & r \geq r_c \end{cases} \tag{2.4}$$

Here, the coefficients $\{c\}$ are determined by requiring continuity of φ and its first derivative at the "outer" cutoff r_c and at an "inner" cutoff r_0 that is slightly smaller. Throughout this chapter we maintain $r_0 = r_c$ - 1.2 Å,

$$\begin{aligned} c_0 &= \frac{1 - 3r_0^2/r_c^2}{(1 - r_0^2/r_c^2)^3}, & c_2 &= \frac{1 - 6r_0^2/r_c^2}{(1 - r_0^2/r_c^2)^3}, \\ c_4 &= \frac{3(1 + r_0^2/r_c^2)}{r_c^4(1 - r_0^2/r_c^2)^3}, & c_6 &= \frac{2}{r_c^6(1 - r_0^2/r_c^2)^3} \end{aligned} \tag{2.5}$$

The indicator function is illustrated in Fig. 2.1

In this approach, the calculation of local densities has a similar computational complexity as pair potentials ($O(n^2)$, ostensibly, but less with neighbor lists). In

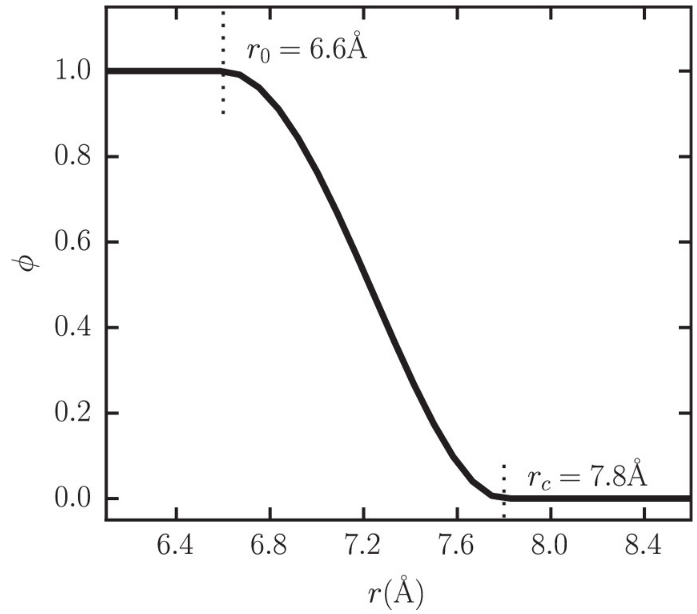


Figure 2.1: Indicator function $\varphi(r)$ for local density potentials, as given in Eq. (2.4) φ goes to zero quickly and continuously between an inner cutoff r_0 and an outer cutoff r_c. Here r_0 = r_c - 1.2 Å has been chosen.

practice, it requires two pair loops: the first calculates the local densities at each site, and the second evaluates the energies and pair forces.

The complete CG forcefield uses U_{LD} as an additive correction to the traditional two-body pair potentials u_{pair} and can be written as

$$U_{\mathrm{CG}} = \sum_{i<j} u_{\mathrm{pair}}(r_{ij}) + \sum_i f(\rho_i) \tag{2.6}$$

In our approach, we do not specify a specific functional form for the LD potential but represent it as a flexible spline whose coefficients (i.e. knot points) are determined through the process of relative entropy optimization with respect to a reference all-atom ensemble.

For CG models with more than one pseudoatom type, there are multiple ways to define a local density, depending on both the central and neighboring types, and hence it is possible to include several distinct LD potentials. A general formulation of the approach includes an arbitrary number of distinct local densities and associated LD potentials, indexed by a variable k,

$$\rho_i^{(k)} = \sum_{j \neq i} b_{\beta(j)}^{(k)} \varphi(r_{ij}) \tag{2.7}$$

and

$$U_{\mathrm{LD}} = \sum_i \sum_k a_{\alpha(i)}^{(k)} f^{(k)}(\rho_i^{(k)}) \tag{2.8}$$

Here, a and b are filters for central and neighbor sites, respectively (and are functions of the site atom types), while α and β denote the types of atoms i and j, respectively: $a_{\alpha(i)}^{(k)}$ is 1 if the central atom i of type α is subject to a LD potential of type k and 0 otherwise. Similarly, $b_{\beta(j)}^{(k)}$ is 1 if the neighboring atom j of type β contributes to the local density of type k around site i; otherwise it is 0.

With these forms, the forces on sites due to the LD potentials extend, as usual, from the potential gradient. For each LD potential k, the expression for the force $\mathbf{f}_i$ on a central atom i, due to its neighbors, is as follows, where we have suppressed the superscript k for clarity:

$$\mathbf{f}_i^{(i\text{ central})} = -\nabla_{\mathbf{r}_i} U = -a_{\alpha(i)} \frac{df(\rho_i)}{d\rho} \sum_j b_{\beta(j)} \frac{d\varphi(r_{ij})}{dr} \frac{\mathbf{r}_i - \mathbf{r}_j}{r_{ij}} \tag{2.9}$$

This has the form of a pairwise force. The force on central atom i due to neighbor j is

$$\mathbf{f}_{ij}^{(i\text{ central})} = -a_{\alpha(i)} b_{\beta(j)} \frac{df(\rho_i)}{d\rho} \frac{d\varphi(r_{ij})}{dr} \frac{\mathbf{r}_i - \mathbf{r}_j}{r_{ij}} = -\mathbf{f}_{ji}^{(i\text{ central})} \tag{2.10}$$

An equal and opposite force applies to atom j. However, a second pair force arises when j is the central atom and i is the neighbor,

$$\mathbf{f}_{ij}^{(j\text{ central})} = -a_{\beta(j)} b_{\alpha(i)} \frac{df(\rho_j)}{d\rho} \frac{d\varphi(r_{ij})}{dr} \frac{\mathbf{r}_i - \mathbf{r}_j}{r_{ij}} = -\mathbf{f}_{ji}^{(j\text{ central})} \tag{2.11}$$

The total pair force that must be added to i and subtracted from j is therefore

$$\mathbf{f}_{ij} = -\left[a_{\alpha(i)} b_{\beta(j)} \frac{df(\rho_i)}{d\rho} + a_{\beta(j)} b_{\alpha(i)} \frac{df(\rho_j)}{d\rho} \right] \frac{d\varphi(r_{ij})}{dr} \frac{\mathbf{r}_i - \mathbf{r}_j}{r_{ij}} \tag{2.12}$$

Note that the local densities must be computed in a separate, earlier pair loop in order for the forces to be determined.

LD potentials are closely related to excess free energy terms that improve the quality of the CG model, as suggested initially by Frenkel and Pagonabarraga.[76,101] Approximating the corrective effect of the LD potential to the pair-potential as a first order perturbation (meaningful when the correction is weak), its addition to the CG model causes a change in free energy equal to $\Delta A \approx \left\langle U_{\mathrm{LD}} \right\rangle_{\mathrm{CG,pair}}$, where the subscript indicates an ensemble average in the pair-only CG case. In this sense, the local density potential captures a multibody contribution to the CG free energy that is in addition to the dominant contribution from pair interactions.

2.2.2 Relative entropy coarse-graining

We parameterize the CG forcefield using the information-theoretic coarse-graining framework introduced by Shell.[32,35] Here, the relative entropy, or Kullback-Liebler divergence,[36] quantifies the quality of a putative CG model in comparison to a detailed, reference AA system; one way of expressing it is

$$S_{\mathrm{rel}} = \int p_{\mathrm{AA}}(\mathbf{r}) \ \ln\left(\frac{p_{\mathrm{AA}}(\mathbf{r})}{p_{\mathrm{CG}}(\mathbf{M}(\mathbf{r}))}\right) d\mathbf{r} + S_{\mathrm{map}} \tag{2.13}$$

where $p_{\mathrm{AA}}(\mathbf{r})$ gives the equilibrium configurational probability in the AA ensemble, while $p_{\mathrm{CG}}(\mathbf{R})$ gives the corresponding CG one. As before, $R = \mathbf{M}(\mathbf{r})$ is the mapping operation that eliminates atoms or replaces groups of them with center-of-mass sites. The integral proceeds over all AA microstates, although it is possible to reformulate Eq. (2.13) as an integral in the CG degrees of freedom.[102] S_{map} is a mapping entropy that measures the number of distinct AA configurations that map to the same CG one; importantly, it is independent of the CG force field U_{CG} such that it plays no role in the scenarios described here.

The relative entropy measures the information “lost” upon moving from the AA to CG ensemble and is thus strictly zero or positive.[32] Its minimization increases overlap between the two systems and suggests a natural systematic strategy for parameterizing CG models from reference AA systems. Shell and co-workers used this approach to construct single-site CG water models that capture a number of bulk properties and hydrophobic interactions.[39, 40, 103] They also showed that the relative entropy is tightly linked to errors incurred upon coarse graining, making it an important measurement for signaling a priori the conditions of validity of CG models.[33]

In the canonical ensemble, the relative entropy can be expressed as

$$S_{\text{rel}} = \beta \langle U_{\text{CG}}(\boldsymbol{\lambda}) - U_{\text{AA}} \rangle_{\text{AA}} - \beta (A_{\text{CG}}(\boldsymbol{\lambda}) - A_{\text{AA}}) + S_{\text{map}} \tag{2.14}$$

where U_X and A_X denote the potential and Helmholtz free energies in ensemble X = AA or CG. The CG quantities are functions of its forcefield parameters $\boldsymbol{\lambda}$, which for the present problem are the unknown coefficients in the cubic spline forms chosen for the pairwise CG potentials u_{pair} and LD potential functions $f(\rho)$. The coarse graining strategy then locates the minimum of S_{rel} in $\boldsymbol{\lambda}$ space, which is achieved through conjugate-gradient minimization.[34] Because all optimized potentials are represented by splines, and thus all parameters appear linearly in the energy, the relative entropy contains a single basin in such a parameter space.[35]

To accelerate the minimization and reduce the number of trial CG MD runs during parameterization, we use the efficient trajectory reweighting and perturbation strategy formulated by Carmichael and Shell that re-uses information from existing CG trajectories.[34]

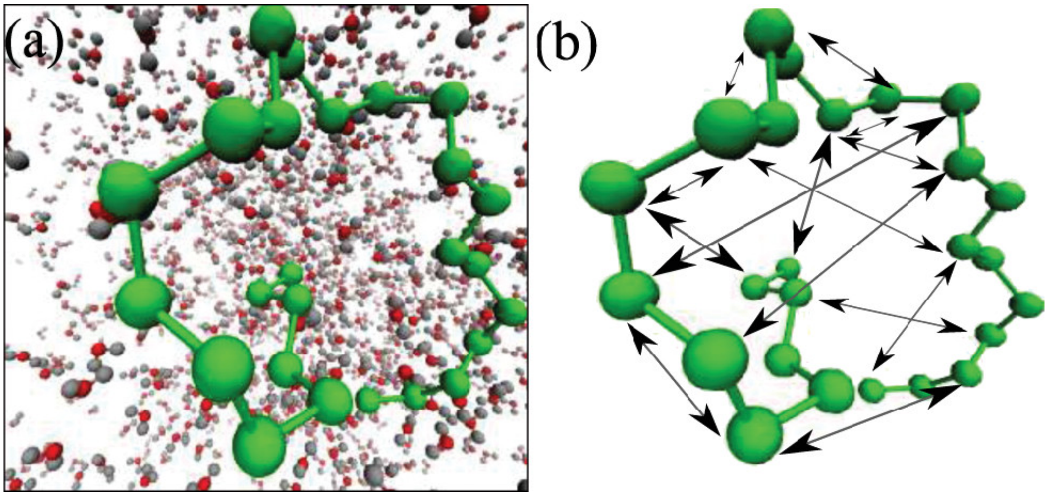


Figure 2.2: (a) Reference AA description of the c-25 polymer with 25 methane sized monomeric beads in 1700 water molecules, at 298 K and 1 atm. (b) Implicit solvent CG model of the same system: the waters are coarse-grained away and the polymer-water interactions are then embedded into effective CG potentials between the monomers.

2.2.3 Test systems and simulation details

Our first test case involves coarse-graining a water-solvated, hydrophobic polymer, as shown in Fig. 2.2 For the reference AA system, we mimic an alkane-like polymer called "c-25" that was studied by Athawale et al.[1] The polymer consists of 25 methane-sized monomers, with diameter 3.73 Å, and equilibrium backbone bond lengths and angles of 1.53 Å and 111°, respectively. Since our primary focus is many-body solvation effects, we study a "superhydrophobic" version of the polymer in which nonbonded monomers interact through Weeks-Chandler-Andersen (WCA)[104] potentials with parameters $\sigma = 3.73$ Å and $\epsilon = 0.14$ kcal/mol, while arithmetic mixing rules are used to evaluate the monomer-water parameters. Harmonic potentials describe bond and angle interactions and are taken from Ref. [1].

We also investigate a second, comparative case in which we delete all polymer bonds, angles, and associated interactions, to assess the effect of backbone connectivity. Thus, this case consists of 25 independent solvated, superhydrophobic methane-like particles that we term a "methane" system although in reality the WCA potential is unrealistic because it lacks attractive interactions. Later we examine the case of full Lennard Jones interactions in both the c-25 and methane systems, in order to study the effect of solute-solvent attractions on the efficacy of the CG models.

Explicit-water AA simulations at 298 K and 1 atm are carried out with the MD engine LAMMPS,[105] using the SPC/E[106] model of water constrained with the SHAKE algorithm.[107] The simulation for a single 25-mer and 1700 water molecules is first equilibrated for 1 ns in a NPT ensemble using the Parniello-

Rahman barostat[108] to determine the equilibrium system volume, followed by a further 2 ns of equilibration under NVT conditions. Trajectory data are then sampled from 26 ns of NVT production time. AA simulations for the methane system involve 2 ns each of NPT and NVT equilibration stages, followed by 30 ns of production.

The CG model for c-25 removes all water molecules, leaving only the single 25-bead polymer, with bond and angle interactions that are retained from the explicit water version. The effect of solvent is then built into effective nonbonded pair and local density potentials that are functions of the inter-monomer pairwise distances and local densities, respectively. Both potentials are represented by flexible cubic splines with unknown coefficients (knot points) that are determined through relative entropy minimization. 40 knot points are used for CG spline pair potentials, while 50 are used for the LD potentials.

Table 2.1: Comparison of and nomenclature for different coarse graining strategies and controls

CG strategy	Bonded interactions	Pair interactions	Local density interactions
SP	Same as AA	Srel optimized spline	None
SPLD	Same as AA	Srel optimized spline	Srel optimized spline
LD	Same as AA	Same as AA	Srel optimized spline

For both the polymer and methane systems, we optimize the CG potentials for three different cases that provide instructive comparisons, as summarized in Table 2.1. The pair-spline-only, or SP, approach renormalizes only the nonbonded intermonomer interactions into an effective pair potential described by a B-spline; this is perhaps the conventional approach to implicit solvent CG interactions. The pair spline and local density (SPLD) approach not only renormalizes the nonbonded interactions but also determines the form of a local density potential based on the number of neighboring monomers surrounding each central one. Finally, the local-density-only case, LD, keeps the WCA pair interactions from the AA system intact as the pair potential and embeds all effects of solvation into a local density potential. SP and SPLD, therefore, serve as negative and positive controls to tease out the relative contribution of the LD potential, while the LD-only case tests its capability as a single mean field representation of the complete implicit solvent force field. Note that the control strategy, LD, should not be confused with the acronym LD used to abbreviate the local density potential throughout the chapter.

2.3 Results and discussion

2.3.1 CG model of c-25

We parameterize the CG c-25 polymer through relative entropy optimization with a LD cutoff of r_c = 7.8 Å, using the three different coarse-graining routes of Table 2.1. Here, the relevant local density is the number of monomers within a range rc from a central monomer. This cutoff is an important free parameter of the CG model, but before discussing its determination, we first provide some illustrative results from the relative entropy minimization procedure.

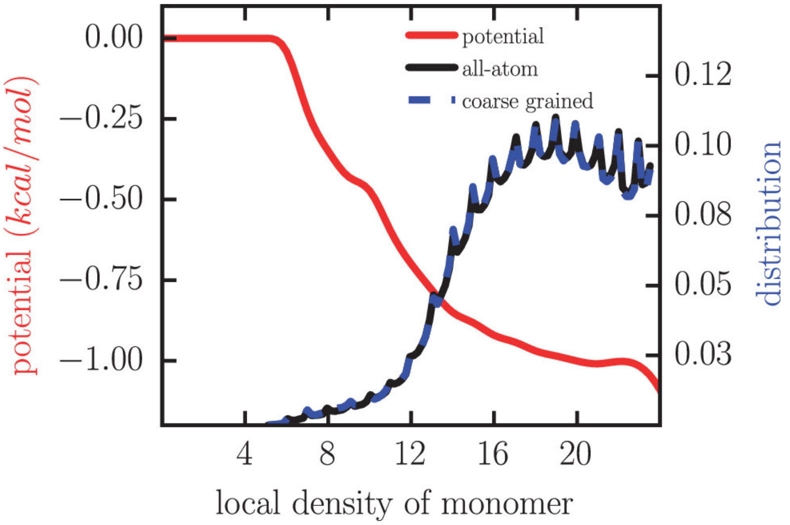


Figure 2.3: Comparison of the local density distribution of the CG polymer for the SPLD case, in which both the LD potential and a CG splined pair potential are optimized with respect to the AA system. The LD potential (shown in red) decreases with greater local crowding around the CG monomers, thus favoring collapse of the polymer chain.

Fig. 2.3 shows the final LD potential and distribution for the SPLD case. First, we note that the LD potential decreases with the local density ($\Delta U_{\mathrm{LD}} \sim$ -1.0 kcal/mol over the full range of the density), thus favoring aggregation of the individual monomers and collapse of the polymer into a compact state, consistent with the idea that the water induces an effective inter-bead attraction of these hydrophobic units. The LD potential is constant for densities less than 5, ultimately due to bond length constraints; each monomer always has a local coordination number of at least 4 due to neighboring bonded monomers and the chosen cutoff length. Above 5, the LD potential sharply decreases, but its effect then weakens beyond a coordination of 13-14. The high-density behavior is expected to flatten, as the influence of additional species diminishes when a full coordination shell around the central particle already exists within the cutoff. Formally, the LD spline is forced to have a zero slope beyond the maximum possible density of 24 (although this is not obvious from the figure), since there is no information in the AA reference to lead to parameterization there.

Interestingly, if we compare the spline pair potentials determined with (SPLD case) and without (SP case) the support of a LD potential (see Figs. 2.A.1, 2.A.2, 2.A.3 in appendix), we find that the former are slightly more repulsive and less attractive; the well-depth decreases from 0.11 to 0.04 kcal/mol for the c-25 CG model. This is to be expected because the LD potential shares the net attractive many-body effect of the monomers which otherwise is sustained entirely by the pair potentials of the SP model. In both cases, however, the inter-monomer pair correlation functions near-exactly reproduce the all-atom one, as is expected from the relative entropy minimization procedure in conjunction with spline potentials.[35]

In addition, we observe that the local density distribution from the reference AA system is exactly reproduced by the LD-corrected CG forcefield. Arguably, it is important to faithfully reproduce this distribution in an aqueous implicit solvent, since the solute local density fluctuations are anti-correlated with the local water density fluctuations, which in turn strongly influence the hydration free energy of hydrophobic units.[93,97,98] Here, the optimization procedure guarantees that the AA and CG local density distributions should match near-exactly at a relative entropy minimum because flexible splines are used for the LD potential.[35] Interestingly, the LD distribution is peaked and skewed around a local density of 20, which suggests that the polymer coils into tightly folded conformations. The jagged peaks throughout the upper contour of the distribution are due to the semi-discrete nature of the indicator function $\varphi(r)$ that is used to determine the local density. If $\varphi(r)$ were a true Heaviside step function, the distribution would involve weighted delta functions at each integer value for the local density; because we use a smooth version (Eq. (2.4)), the distribution is continuous but still shows some of the sharp-peaked character.

We determine the LD cutoff r_c, introduced in the local density indicator function (Eq. (2.4)), separately from the spline knots that govern the force field potential energies. To fix r_c, we exploit the property[32,33] that optimal CG force-field parameters should minimize the relative entropy between the CG model and its reference AA system. Starting with an initial choice for r_c, we minimize the relative entropy S_{rel} in the space of the other force field parameters $\boldsymbol{\lambda}$; let their final, optimal values be $\boldsymbol{\lambda}^*$. Then, the variation of the minimized relative entropy with r_c follows:

$$\frac{dS_{\text{rel}}}{dr_c} = \left(\frac{\partial S_{\text{rel}}(\boldsymbol{\lambda}^*, r_c)}{\partial r_c}\right)_{\boldsymbol{\lambda}} + \left(\frac{\partial S_{\text{rel}}(\boldsymbol{\lambda}^*, r_c)}{\partial r_c}\right)_{r_c} \left(\frac{d\boldsymbol{\lambda}}{dr_c}\right) = \left(\frac{\partial S_{\text{rel}}(\boldsymbol{\lambda}^*, r_c)}{\partial r_c}\right) \tag{2.15}$$

where the second term vanishes because $\boldsymbol{\lambda}^*$ is determined by $\frac{\partial S_{\text{rel}}}{d\boldsymbol{\lambda}}\big|_{\boldsymbol{\lambda}=\boldsymbol{\lambda}^*} = 0.11$ We integrate Eq. (2.15) to determine the dependence of S_{rel} on the cutoff by finely discretizing r_c in the range 4-10 Å. The change in relative entropy $\Delta S_{\text{rel},1\to 2}$ from one cutoff ($r_{c,1}$) to the next ($r_{c,2}$) is evaluated by re-casting Eq. (2.13) as

$$\Delta S_{\text{rel},1\to 2} = \beta\big\langle U_{\text{CG}}(\boldsymbol{\lambda}_2^*, r_{c,2}) - U_{\text{CG}}(\boldsymbol{\lambda}_1^*, r_{c,1})\big\rangle_{\text{AA}} - \beta\big(A_{\text{CG}}(\boldsymbol{\lambda}_2^*, r_{c,2}) - A_{\text{CG}}(\boldsymbol{\lambda}_1^*, r_{c,1})\big) \tag{2.16}$$

where $\boldsymbol{\lambda}_k$ refers to the optimal set of forcefield parameters that minimize S_{rel}, which we separately evaluate for each cutoff $r_{c,k}$. The average energy difference $\langle U_{\text{CG}}(\boldsymbol{\lambda}_2^*, r_{c,2}) - U_{\text{CG}}(\boldsymbol{\lambda}_1^*, r_{c,1})\rangle_{\text{AA}}$ is then computed by reprocessing AA trajectories and using the optimized CG parameters, while the free energy difference $A_{\text{CG}}(\boldsymbol{\lambda}_2^*, r_{c,2}) - A_{\text{CG}}(\boldsymbol{\lambda}_1^*, r_{c,1})$ is estimated using MD simulations of the two CG models and the Bennett acceptance ratio method.[109]

Fig. 2.4(a) shows that the (S_{rel}, r_c) space for the polymer-water system is approximately concave and admits two local minima near 6.5 and 7.8 Å, respectively. This S_{rel} landscape does not guarantee a single minimum78 (as discussed earlier in Section 2.2.2), since it depends on r_c which affects the CG forcefield in a non-linear manner. Note that this figure is essentially a projection of the higher dimensional space $S_{\text{rel}}(\boldsymbol{\lambda}, r_c)$ along the r_c coordinate with $\boldsymbol{\lambda} = \boldsymbol{\lambda}^*(r_c)$. Apparent statistical noise in the calculated relative entropy likely stems from the determination of $\boldsymbol{\lambda}^*(r_c)$, which involves a lengthy optimization in the high-dimensional $\boldsymbol{\lambda}$ space separately at each cutoff value. Fig. 2.4(a) shows that low or high LD cutoffs, which stray away from twice the diameter of a monomer, lead to higher relative entropy values and signal a smaller role for the LD potential to improve the CG forcefield. Intuitively, both extremes of cutoffs are undesirable: very low values fail to capture relevant local structures and interactions near a central

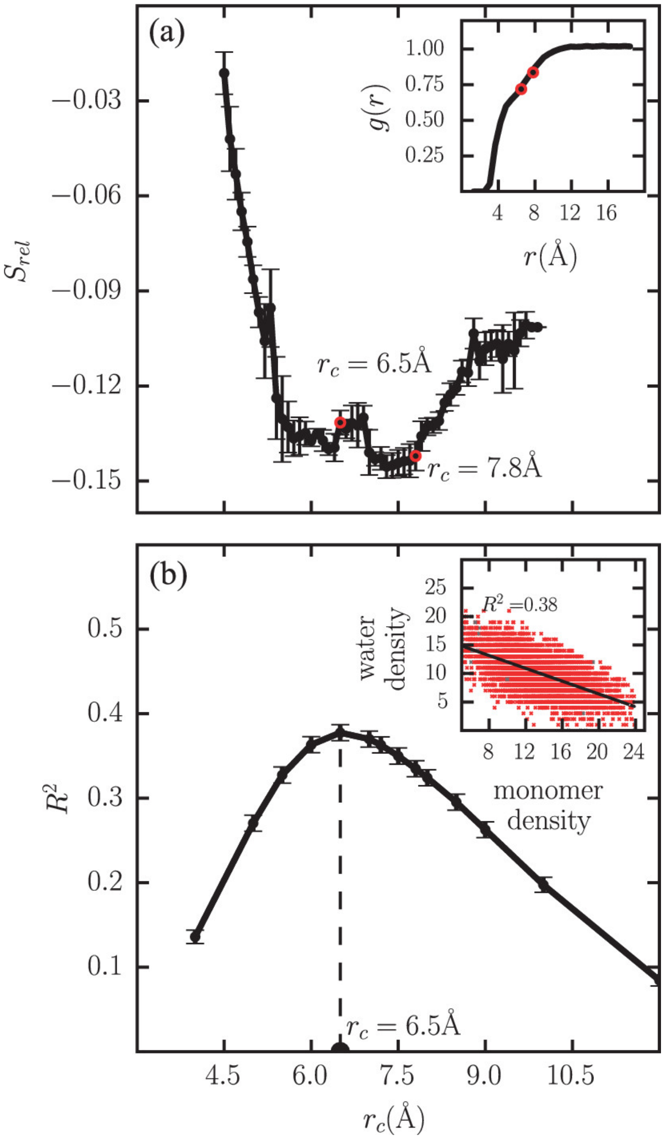


Figure 2.4: (a) Selection of the local density cutoff (r_c) that minimizes the relative entropy Srel. The local minimum near 6.5 Å is related to the first shell of the monomer-water radial distribution function, while the minimum at 7.8 Å includes part of the second shell (inset). Note that the relative entropy is shifted such that its value at the lowest cutoff is zero. (b) The correlation between monomer local density and the number of first shell waters (bivariate scatter plot in inset) also peaks around a cutoff of 6.5 Å.

monomer, while too high values wash out these interactions with irrelevant "bulk" fluctuations that are far away. We pick a cutoff at 7.8 Å which, within the noise of the calculations, seems to be near a global minimum in S_{rel}.

It is instructive to note that the two minima of Fig. 2.4(a) embed information about the hydration shell structure around the monomers. In particular, the minimum at smaller distances captures the first hydration shell and seems to produce local monomer densities that are correlated with the number of hydration shell waters. To quantify this connection, we calculate the correlation coefficient between the instantaneous number of first hydration shell waters (water-monomer distance less than 5.6 Å) and the instantaneous local density of monomers around monomers, on a frame-by-frame basis across the entire AA trajectory. Fig. 2.4(b) shows that the squared correlation coefficient (R^2) has a maximum at $r_c = 6.5$ Å. While the correlation is not extremely high (maximum at R2 = 0.38), it is still encouraging that some information from water density fluctuations is incorporated in the monomer local density distribution, which in turn bodes well for its ability to capture features of hydrophobic interactions.[93]

Admittedly, constructing the entire (S_{rel}, r_c) space requires a significant effort involving many relative entropy optimizations and CG model free energy calculations. On the other hand, evaluating the correlation of monomer local density with vicinal water density is a far simpler calculation using the reference AA trajectory that can be obtained prior to parameterizing the CG model. It also seems intuitive to infer the monomer local density from simple structural metrics of the surrounding water. However, the correlation approach is based on the monomer-water two-body radial distribution functions that may neglect higher order interparticle structural correlations of the type that the local density potential is designed to treat. Therefore, finding the minimum of S_{rel} with respect to r_c is likely to be a more robust approach to selecting the cutoff.

2.3.2 Fidelity of the c-25 CG model in reproducing macromolecular properties

To benchmark the accuracy of the optimized c-25 CG model, we compare distributions of several structural metrics to those of the reference AA system, as shown in Fig. 2.5 In particular, the radius of gyration, R_g, and end-to-end distance, R_{EE}, are important characterizers of polymer dimension and are often related to the total number of monomeric units through simple scaling laws that elucidate both intra polymer and polymer-solvent interactions. Figs. 2.5(a) and 2.5(b) show that including the LD potential in the CG forcefield helps to reproduce the peaks of the distributions (4.2 Å for R_g and 6 Å for R_{EE}), while only renormalizing the CG pair potential without a LD contribution, as in the SP case, leads to a model that deviates more notably.

Fig. 2.5(c) shows that the distribution of per-monomer solvent-accessible-surface-area (SASA), computed with the Shrake-Rupley algorithm,[110] agrees equally well with the all-atom reference for the three CG cases, such that the LD potential has a smaller impact on its behavior. It is worthwhile to note that SASA is a popular proxy[84,111] for the hydrophobic effect in the construction of implicit solvent models, but is expected to break down for very small solutes where hydration free energies scale with volume, not area, and show significant entropic driving forces.[103,112,113] Here the SASA distribution in Fig. 2.5(c) has peaks around 0 and 30 Å^2, and a steady low tail from 50 to 80 Å^2. The two (inner) peaks represent CG monomers that are heavily shielded from the solvent and are suggestive of tightly coiled conformations driven by hydrophobic collapse; the first peak at very low degrees of exposure stems from mid-chain monomers that have a large number of nearby neighbors due to the short bond length (note that the bond length is less than half of the monomer diameter). Thus the presence of this peak is representative of the overall coiled structure of the WCA polymer, while its proximity to zero likely reflects significant burial of monomers due to the disparity in bond length and monomer size. The near-constant values of the distribution

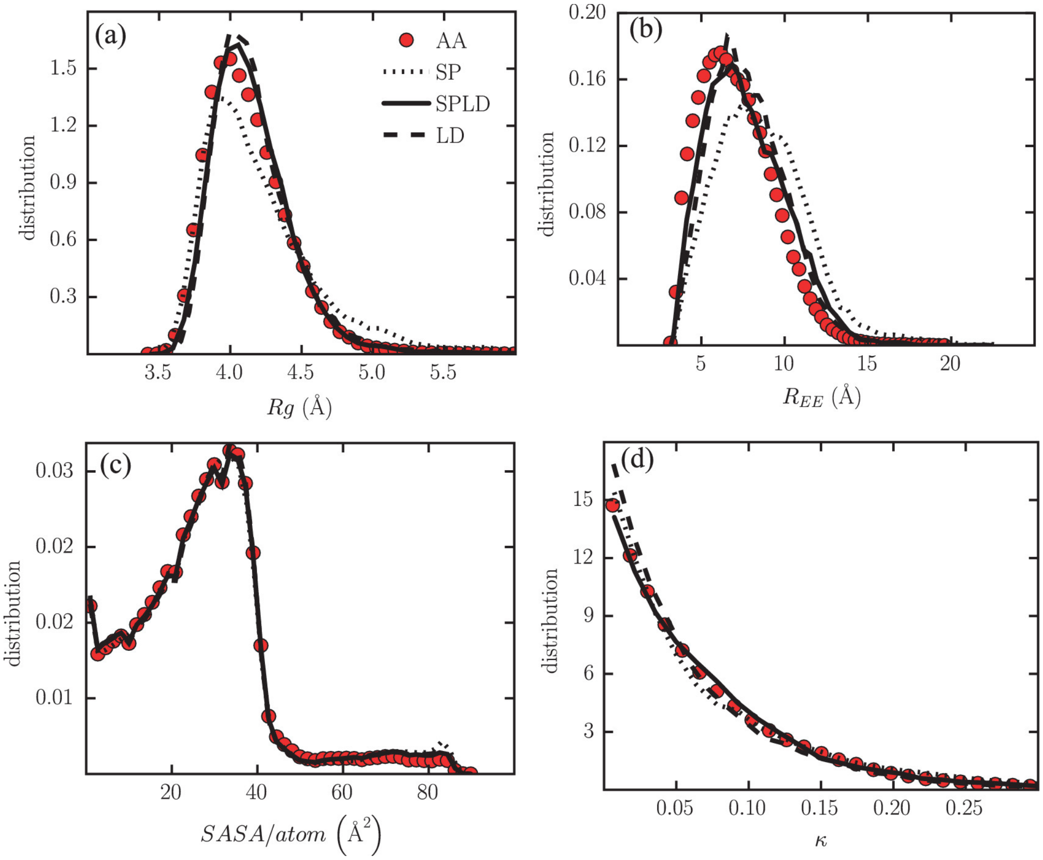


Figure 2.5: Comparison of shape metrics between the AA and the CG polymer systems with different nonbonded CG potentials: CG splined pair potentials (SP), CG splined pair potentials with LD correction (SPLD), and LD potential with the original AA nonbonded interactions (LD). Distributions of (a) radius of gyration R_g and (b) end-to-end distance R_{EE} demonstrate that including the LD potential in the CG forcefield improves representation of the AA system. On the other hand, the distributions of (c) per-atom Solvent-Accessible-Surface-Area (SASA) and (d) coefficient of relative anisotropy κ show less sensitivity to the form of the CG potentials. Using block average analysis, the average relative errors for these distributions (averaged over the range of the shape metrics near the histogram peaks and across the different CG models, and reported as a fraction of the plotted mean value) are (a) 17.6%, (b) 12%, (c) 6%, and (d) 12.6%. The order-parameter averaged uncertainties are relatively similar for the different CG models.

towards higher values of SASA may be attributed to the fact that the polymer occasionally adopts a structure in which one end of the chain is collapsed, while the other projects in an extended form into the solvent. To quantify this behavior, we compute the coefficient of relative shape anisotropy, given by

$$\kappa = \frac{3}{2}\frac{\lambda_x^4 + \lambda_y^4 + \lambda_z^4}{(\lambda_x^2 + \lambda_y^2 + \lambda_z^2)^2} - \frac{1}{2} \tag{2.17}$$

where λ_i's are the mutually orthogonal components of the gyration tensor. $\kappa \in [0, 1]$ and a value closer to 0 implies greater symmetry, or an overall collapsed conformation, while larger values typically allude to more linear chains. The distributions of κ in Fig. 2.5(d) show that, while the predominant structures are globular and spherical in nature, there are still notable fluctuations to asymmetric configurations of the type described above. The addition of local density potentials in the SPLD case shows just slight improvement over the SP scenario in reproducing this distribution. On the other hand, not including renormalized pair interactions (LD case) seems to weaken the quality of the CG model.

Overall, the metrics in Fig. 2.5 show that the SPLD strategy improves the quality of the CG model relative to the SP case. The LD potential alone performs similarly if slightly worse than the combined SPLD case, as it overestimates several distribution peaks. Thus it may be the case that the functionality of an LD potential as a proxy for an implicit solvation energy, which is a mean-field multibody term, is improved when detailed (non-mean-field) pair interactions between monomers are renormalized in the coarse graining process.

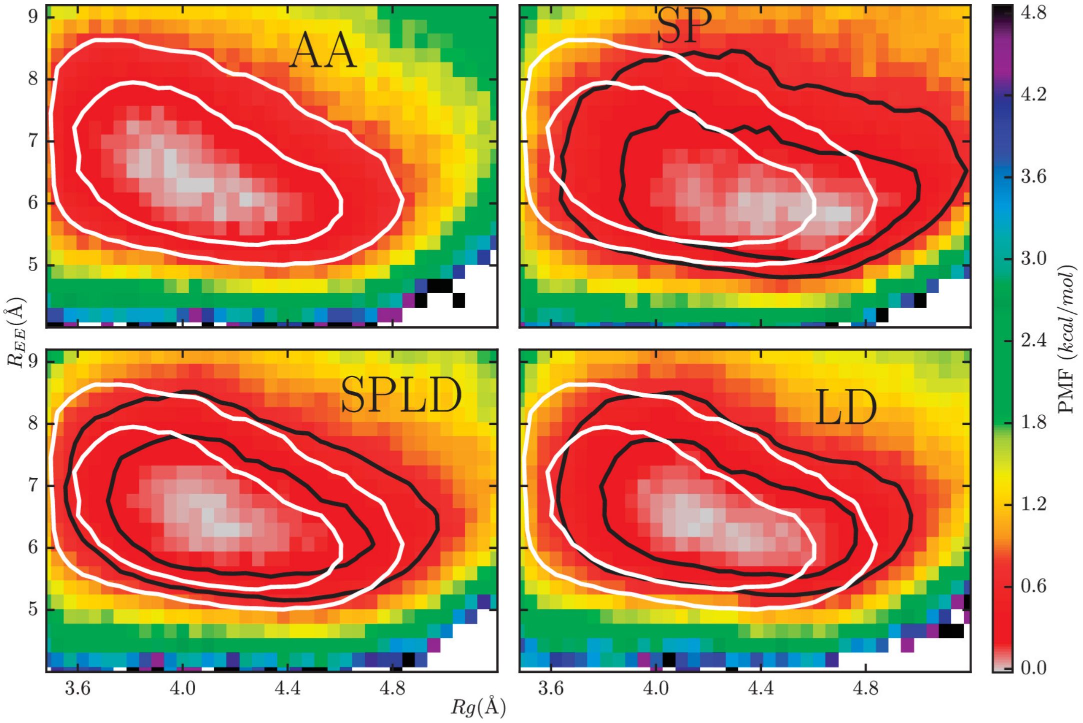


Figure 2.6: Free energy of polymer collapse as a function of the end-to-end distance (R_{EE}) and radius of gyration (R_g) for the AA system compared to the three CG cases: SP, SPLD and LD. The red basins are the regions of most probable conformations, and the contours are shown for 0.5 k_BT and k_BT above the minimum. White contours indicate the AA system and black lines denote the different CG cases. Using block average analysis, the average relative error for the PMFs within the inner contour (as a fraction of the plotted mean value) is 21%. The average errors within the inner contour are relatively similar for the different CG models.

Fig. 2.6 illustrates the coupling between the R_g and R_{EE} distributions in the form of a free energy surface, where a clear minimum is evident that corresponds to a globular, collapsed state. The region within one k_BT extends roughly from 3.5 to 4.8 Å in R_g and 5.5 to 8.5 Å in R_{EE} for the explicit water (AA) reference. The low R_g end of the basin ($\approx$ 3.5 to 3.75 Å), where one might expect more significant multibody effects, is not well-captured by the SP approach, which shifts the entire basin to higher values. Both the SPLD and LD-only forcefields appear more successful in reproducing the extents of the basin as visualized through the contours in Fig. 2.6. However, both cases also slightly extend the basin to higher R_g values.

2.3.3 Transferability of the c-25 CG forcefield

Another measure of the success of a CG forcefield is its transferability to systems beyond those at which the model was parameterized, which indirectly assesses whether or not it captures "true" driving forces in a physically realistic manner or is simply an effective fit to the original AA reference. Here, we explore transferability to different chain lengths, between 10 and 40 monomers (denoted by "c-10" to "c-40"). We perform explicit-water MD simulations for each polymer with the same pressure and temperature conditions as c-25 and then compare structural metrics obtained from these AA trajectories to CG simulations based on the c-25 forcefield.

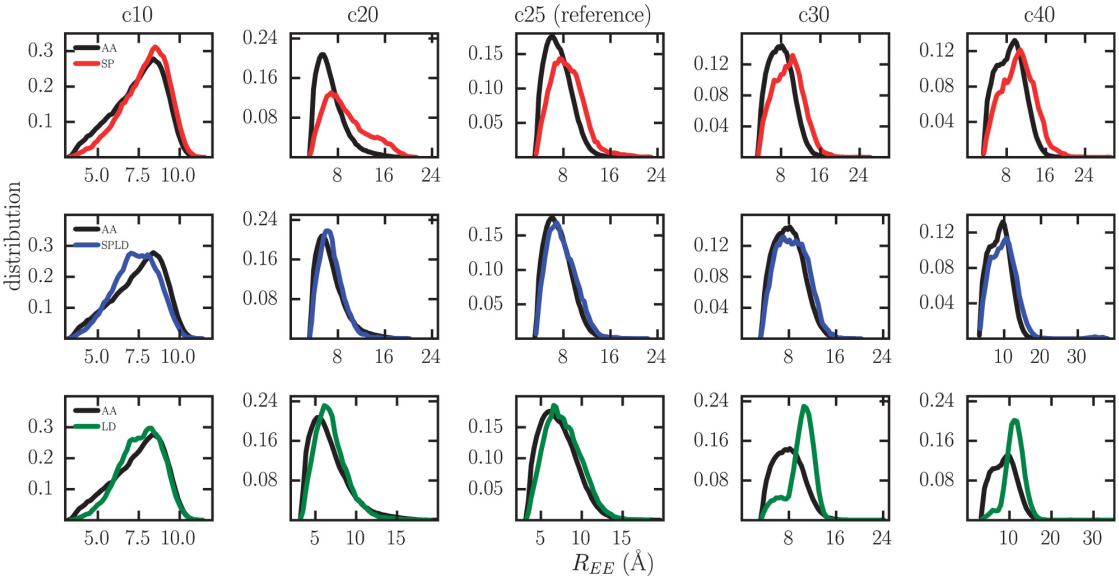


Figure 2.7: End-to-end distance (R_{EE}) transferability of the CG forcefield parameterized from the 25-mer, for the three different CG-ing schemes: SP, SPLD, and LD. Including LD potentials (SPLD) apparently makes the forcefield more robust than pure CG pair potentials (SP) for chain lengths both smaller (10- and 20-mers) and larger (30- and 40-mers) than the reference.

Figs. 2.7 and 2.8 show the transferability of the c-25-parameterized potentials in terms of the R_{EE} and R_g distributions, respectively. Overall, the SPLD potential seems marginally more transferable than the other CG cases. Specifically, it provides a better estimate of the distributions at low and moderately high chain lengths (10, 20, 30) for both the shape metrics, but it does underestimate the peak value of the R_g distribution for the 40-mer. This limit in transferability for R_g at higher chain lengths is likely because the 40-mer explores regimes of local densities that are beyond those sampled by the original system; that is, the c-25 SPLD forcefield embeds local density information only up to 25 monomers within

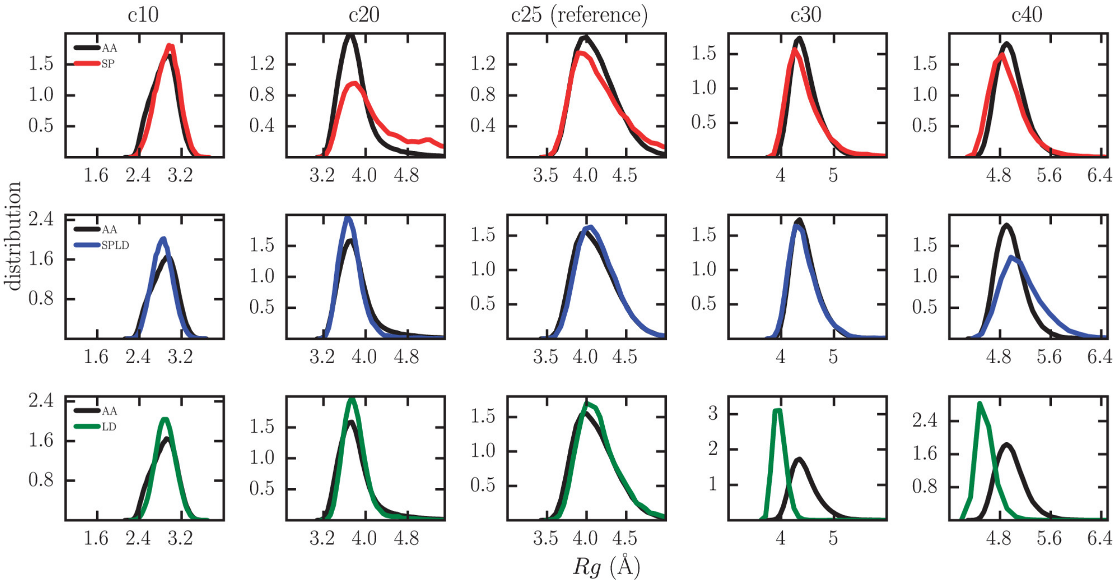


Figure 2.8: Radius of gyration (R_g) transferability of the CG forcefield parameterized from the 25-mer, for the three different CG-ing strategies. Including the LD potential (SPLD) in the CG forcefield reproduces the R_g distribution faithfully for chain lengths smaller than the reference (10- and 20- mer) but is less representative at state points of high chain length (40-mer).

the cutoff radius (Fig. 2.4). Thus, it does not approximate local density interactions for higher chain lengths that are able to sample at or beyond the high local density edge of the original LD potential (right edge of the red curve in Fig. 2.3). We also note that the bare LD potential (third row in Figs. 2.7 and 2.8) shows even more pronounced errors at large chain lengths, which again is likely due to the absence of data for its parametrization in regimes of the higher (than the c-25 case) local densities that are accessible by the longer chain length cases.

2.3.4 Relationship to SASA-based implicit solvation

It is worthwhile to examine the relationship between the SPLD implicit solvation strategy and the longstanding tradition of embedding hydrophobic interactions in SASA-based energy terms. Indeed, phenomenological solvation free energies using an effective molecular surface tension with SASA have been a popular choice for modeling macromolecules.[84, 111, 114–116]

To make the comparison, we examine the parametric relationship between the solvation component of the CG energy and the SASA of individual conformations in the configurational ensemble for c-25. The solvation energy is calculated by subtracting the AA inter-monomer interactions (U_{WCA}) from the total effective energy due to the nonbonded part of the SPLD forcefield ($U_{SP} + U_{\mathrm{LD}}$), since this provides the effect of the implicit solvent relative to the vacuum interactions. Fig. 2.9 shows that the solvation energy and SASA have a strong degree of correlation ($R^2 = 0.86$). The slope of this relationship gives an effective interfacial tension between the polymer-water interface at 32 mN/m, which is in relatively good agreement with a previous estimate of 41.5 mN/m for the c-25 system.73 We perform a similar analysis on other polymer chain lengths, still using the c-25 SPLD force field, and find a 20% variation in the effective interfacial tension;

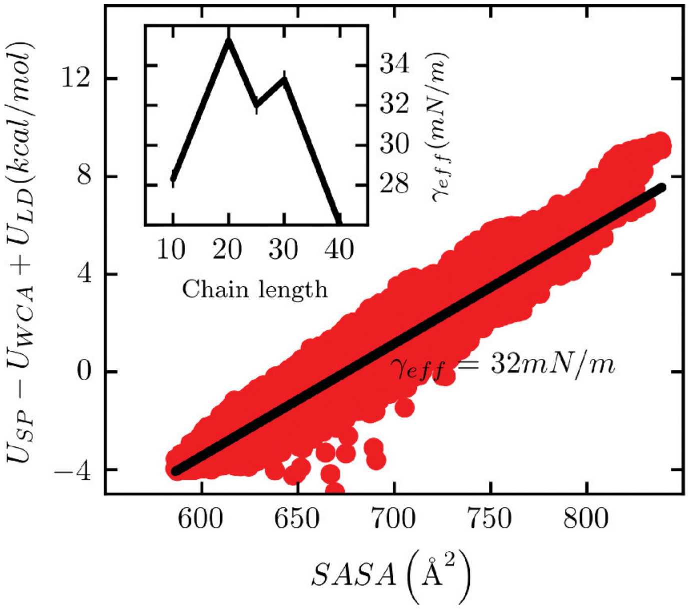


Figure 2.9: Correlation of the total solvation energy in the SPLD CG model with the Solvent Accessible Surface Area (SASA) for the c25 polymer. Solvation energy is calculated as $U_{SP} - U_{WCA} + U_{\mathrm{LD}}$, where U_{SP} and U_{LD} are intrapolymeric contributions of the CG pair and LD potential parts of the CG forcefield, while U_{WCA} is the energy due to the original all-atom WCA pair intermonomer interactions. An effective interfacial tension given by a linear fit is 32 mN/m, in good agreement with 41.5 mN/m reported by Athawale et al.[1] for the c-25 AA system. The inset demonstrates that the effective interfacial tension shows a 20% variation with chain length.

however, it is important to note that these numbers also embed any transferability errors.

2.3.5 CG model of solvated, superhydrophobic methanes

To remove the impact of the polymer architecture and probe "inherent" hydrophobic interactions, we perform comparative simulations in which we delete the bond and angle potentials (Section 2.2.3) and construct a system of 25 solvated, superhydrophobic WCA "methanes" that we coarse grain using the SP and SPLD routes.

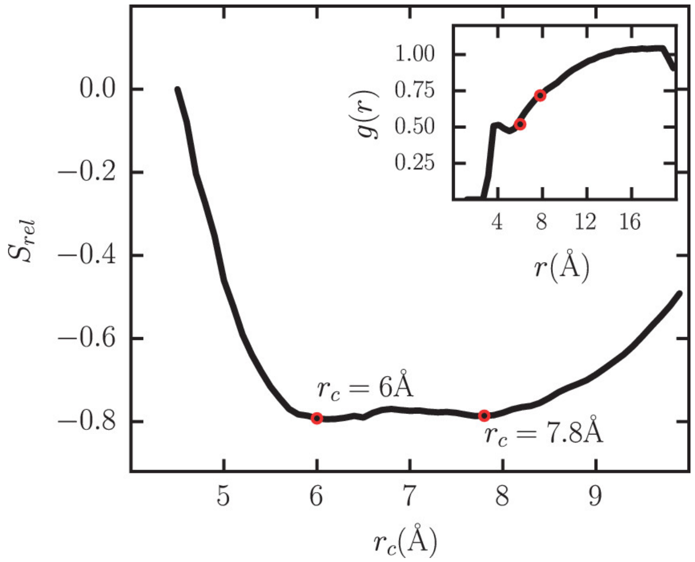


Figure 2.10: LD cutoff selection for the assembly of solvated methanes. Here, it is desirable to minimize relative entropy (S_{rel}) as a function of cutoff (r_c). Both local minima (6.0 and 7.8 Å) seem equally viable candidates, and the first seems to indicate a cutoff for the first hydration shell radius of the methane-water radial distribution function (inset). Note that the relative entropy has been shifted by its value at the lowest cutoff.

Our strategy for selecting an appropriate LD cutoff remains similar to the polymer case. Fig. 2.10 shows the structure of the (S_{rel}, r_c) space for the methanes, in which two minima at 6.0 and 7.8 Å appear to be equally good choices for the cutoff. Further, as seen in the inset, the first minimum has a strong relationship with the first shell radius of the methane-water radial distribution function and hence, qualitatively conveys information about local water structure. To maintain consistency with the c-25 system, we choose $r_c = 7.8$ Å for the CG model of implicitly solvated methanes.

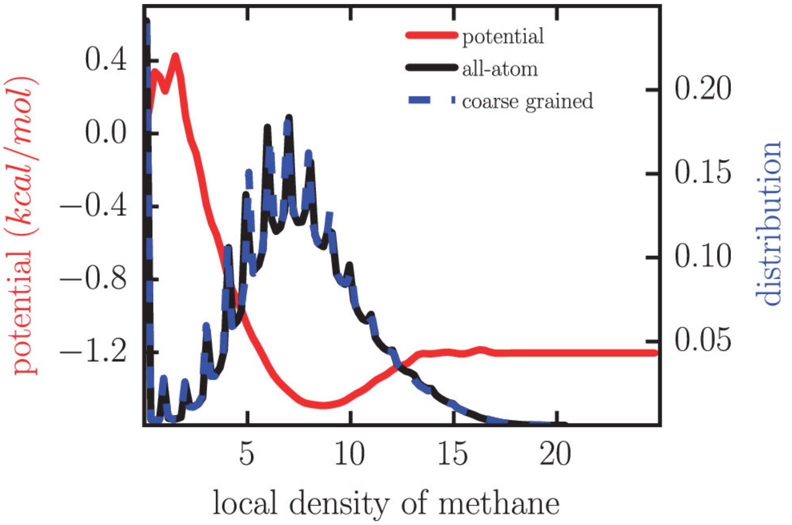


Figure 2.11: Comparison of the local density potential (red) and distribution between the CG and AA methane systems. The potential suggests anti-cooperativity for very low local coordination numbers. We have verified, through multiple optimization runs, that the peak structure in this regime is statistically significant (see Fig. 2.C.1 in the appendix).

Fig. 2.11 shows the LD potential and LD distributions for the SPLD approach for the methane-water system. As with the polymer in Fig. 2.3, the LD potential obtained through relative entropy optimization encourages methane aggregation (with an overall change from 0 to $\sim$ 15 neighbors, and of -1.2 kcal/mol.) However, it shows a slight "repulsive" behavior at very low local densities, where it increases upon the addition of 1-3 neighboring methanes to a lone central one, which is suggestive of anti-cooperativity in low-order assembly. The distribution of local densities from the explicit-water reference is, as expected from relative entropy theory and the use of spline LD potentials, well-matched by the SPLD CG model. The distribution differs from its polymer counterpart (Fig. 2.3) in terms of the location of a peak near zero local density and another around 7, and in the absence of a tail at high local densities. Both features stem from the ability of the methanes to cluster into a singular, dense aggregate due to the absence of backbone rigidity and bond constraints.

To characterize aggregation in the methane system, we examine the distribution of cluster sizes and make comparisons with that of the explicit water reference simulation. In particular, the cluster distribution can signal cooperative assembly, and the potential for multibody physics to play a role, in the form of marked populations of high-number hydrophobic solute assemblies. Fig. 2.12 shows that due to strong hydrophobic interactions, the methanes distribute into two phases: a methane-rich phase described by extensive aggregation (cluster size ≥ 20) and a water rich phase characterized by sparsely distributed monomers (cluster size ~ 1). In a larger system, these species would macroscopically phase-separate, but here the total methane number is small and the simulation volume fixed. In comparing the different CG strategies, the SPLD case best replicates the AA dis-

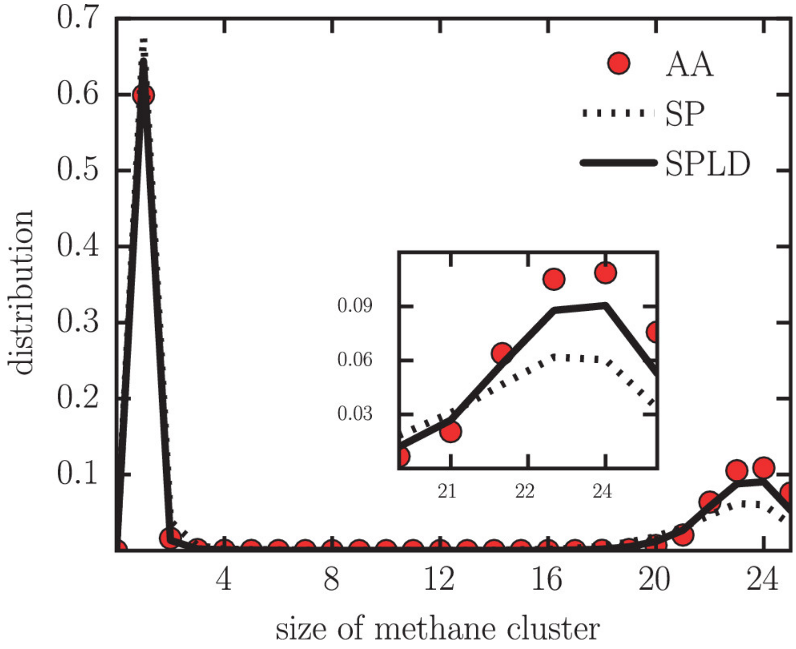


Figure 2.12: Distribution of cluster sizes for methane aggregates in water, for different schemes and the reference AA system of 25 solvated methanes. Large size clusters (zoomed inset) are reproduced well only when the LD potential is included with the renormalized CG splined pair potentials (SPLD case). Using block average analysis, the average relative error at the lower peak (as a fraction of the plotted mean value and averaged over the different CG models) is 17%. The average errors near the distribution peak at high cluster size are relatively close for the different CG models.

tribution and peak at high cluster size showing that the SPLD strategy captures some features of multibody interactions in the aggregate phase, which are not completely described by a pair spline potential alone. It may be noted that the methane cluster observed here is not representative of true methane physics, but rather of the superhydrophobic WCA particles that lack attractive van der Waals interactions.

Fig. 2.13 highlights the transferability of the CG forcefields in the cluster distributions, using different total methane numbers at the same pressure and temperature. Additional explicit water simulations are carried out with 10, 20, and 30 methanes for comparison. Fig. 2.13 shows the high cluster part of the distribution (similar to the zoomed inset of Fig. 2.12) for CG simulations using the SP and SPLD forcefields parameterized from the 25 methane case. The SPLD potential shows much better reproduction of the large-size cluster distribution than the SP potential, while the SP potential appears too weak at smaller methane numbers and too strong at higher ones—a possible signature of the need for a multibody potential.

2.3.6 Effect of attractive hydrophobe interactions

So far, we have investigated solutes that are "superhydrophobic" in nature, described by purely repulsive monomer-monomer and monomer-water WCA interactions, where multibody effects in hydrophobic driving forces should be pronounced. However, it is now well-established that even weak solute-water attractive interactions can produce significant effects and sometimes even qualitatively

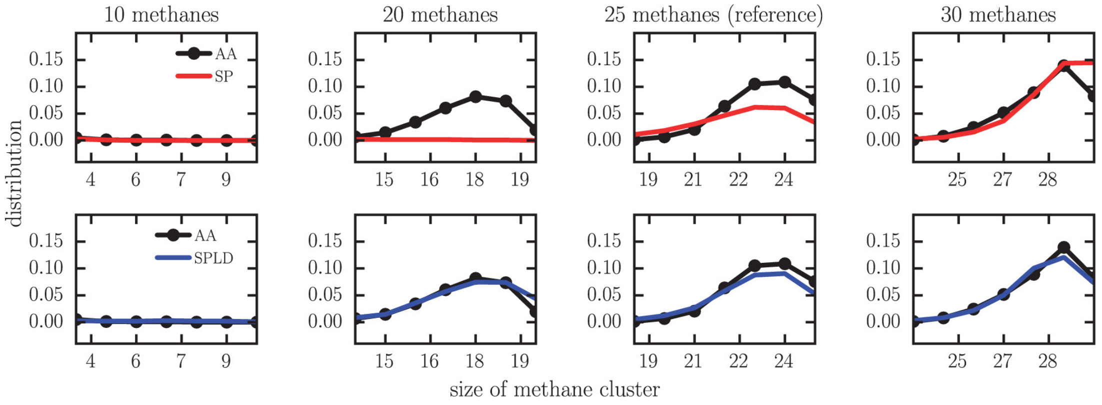


Figure 2.13: Cluster size distribution transferability of the CG forcefield parameterized from the system of 25 solvated superhydrophobic (WCA) methanes. The distributions are zoomed to display only the relevant methane-rich phase with large-size clusters. For low methane numbers such as 10, aggregation does not occur. But with increasingly higher methane concentration (20, 30), large size clusters form and are correctly described by the LD-corrected CG forcefield (SPLD). The average relative errors at the distribution peaks (as a fraction of the plotted mean value and averaged over the different CG models) are 13.5% for 10 methanes, 35.5% for 20 methanes, 17% for 25 methanes, and 24.6% for 30 methanes The average errors across the different CG models are quite close.

distinct behavior.[117–119] For a brief comparison, therefore, we characterize the LD approach for the polymer and methane systems when the AA reference model includes a full Lennard-Jones (LJ) potential, with all other conditions unchanged.

Fig. 2.14 shows the free energy of c-25 as a function of R_g and R_{EE}. The basin within roughly 1 k_BT of the minimum represents the region of most probable polymer conformations and is the focus of our analysis. Note that, compared to the PMF for the superhydrophobic polymer (Fig. 2.6), the basin for the LJ polymer in the AA simulation has a larger R_g ($\sim$ 3.5–5 Å) and smaller R_{EE} ($\approx$

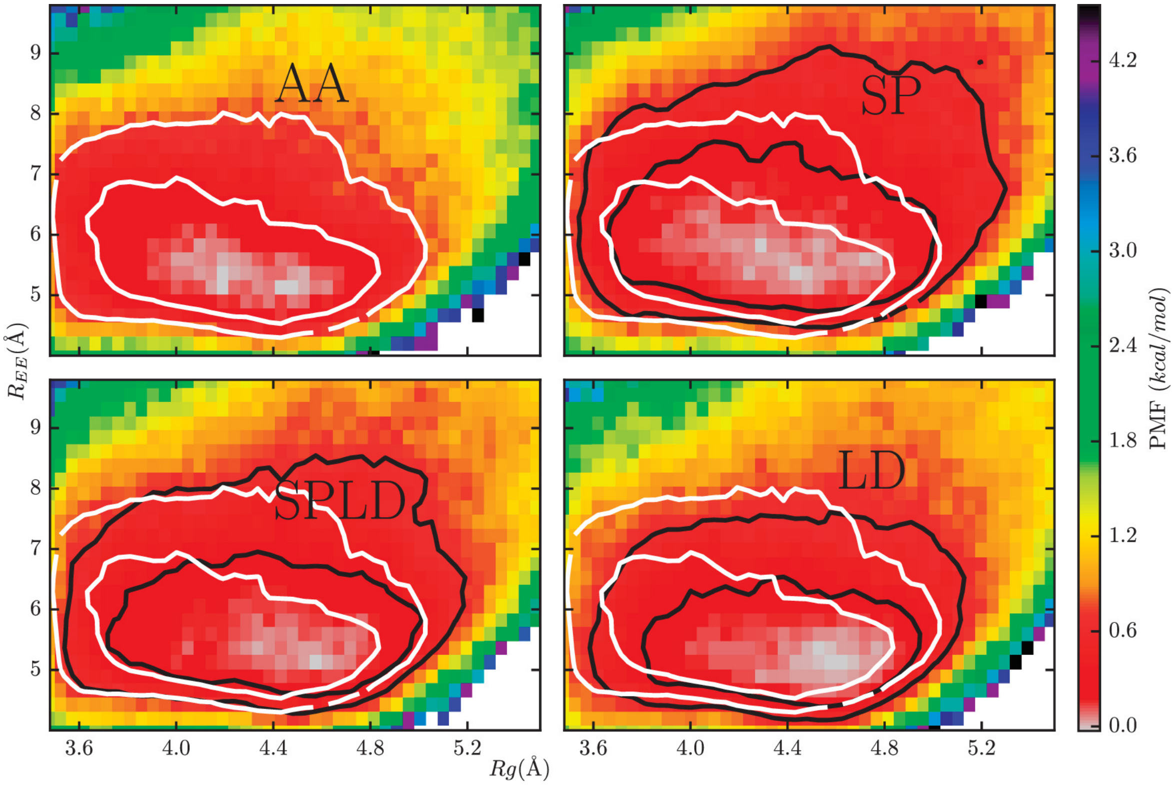


Figure 2.14: Free energy of polymer folding as a function of end-to-end distance (R_{EE}) and radius of gyration (R_g) for a hydrophobic polymer with attractions (LJ rather than WCA potential) using the different CG-ing strategies, SP, SPLD, and LD. The basins represent the regions of most probable conformations (inner and outer contours mark the 0.5 k_BT and k_BT levels, respectively), and all three CG approaches model the inner contour equally well. White contours indicate the AA system and black lines denote the different CG cases. Using block average analysis, the average relative error for the PMFs within the inner contour (as a fraction of the plotted mean value) for the PMFs is 21%. The average errors within the inner contour among the different CG models are comparable.

5-7.5 Å) range. In the CG models, the high end of the basin in R_g is slightly overestimated in all cases. More significant differences exist at the basin's high R_{EE} end, which is overestimated by 13 % for the SP and 10 % for the SPLD cases but underestimated by roughly 3 % for the LD-only case. However, the lowest lying portions of the basin ($\leq 0.5\ k_BT$) seem equally well modeled by each of the CG scenarios, and so the overall differences are not as dramatic as in the superhydrophobic case. It is interesting to note that the LD potential alone, for the attractive polymer, is able to well-describe all of the implicit solvation effects, even without renormalizing the inter-monomer pair potentials.

For the methane AA system modeled with a LJ potential, the distribution of cluster sizes is less interesting than the WCA case, as shown in Fig. 2.15. One only observes a single peak at a cluster size of 1 implying that the presence of weak solute-water attractions leads to a significant weakening of hydrophobicity and cooperative self-assembly such that large clusters do not form at all. Both the SP and SPLD potentials, therefore, capture the single methane peak well.

Taken together, the results above show that the addition of local density potentials to CG force fields has a weaker effect on their performance when hydrophobic driving forces for self-assembly are less pronounced. Presumably, in these cases, multibody interactions are weaker in magnitude and most of the effective solvent-mediated attractions can be subsumed into renormalized pair potentials. Interestingly, a local density potential in the polymer case alone does an excellent job of describing the complete solvation effects; this is likely due to the fact that

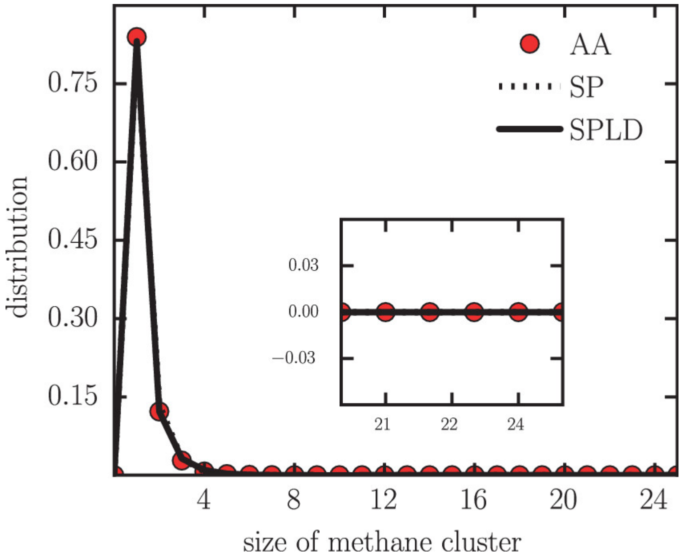


Figure 2.15: Distribution of cluster sizes for solvated methanes described by a LJ rather than WCA potential. Weak van der Waals attractions are sufficient to suppress cooperative assembly and manybody interactions, such that both the SP and SPLD techniques collapse on the same distribution, with only a single peak near dispersed methanes. Using block average analysis, the average relative error near the distribution peak (as fraction of the plotted mean value and averaged over the different CG cases) is 8.9%. The average errors for the different CG models are close in magnitude.

all of the solvent mediated interactions are so weak that they are relative easy to capture with a variety of functional forms (i.e. basis sets) in the CG forcefield.

2.4 Conclusion

In this work, we introduced local density (LD) potentials as a simple and computationally fast mean field approach to capturing multibody effects in coarse-grained (CG) models that may improve transferability. While conventional CG models use effective pair potentials to describe nonbonded interactions, such approaches neglect potentially significant multibody effects that naturally arise during the coarse-graining process. LD potentials thus seek to approximate these interactions through an energetic contribution at each site that depends non-linearly on the number of neighbors (of a given type) within a cutoff distance. We have shown that the relative entropy coarse graining framework of Shell and co-workers[32–34] offers a systematic and transparent way to fully parameterize LD potentials in CG forcefields, given fully atomistic reference simulations, without needing approximations to adapt global density-dependent potentials at a local molecular level.

The present work also examined the utility of LD potentials in the development of implicit solvation models, with a specific focus on hydrophobic interactions. We selected two examples that demonstrate cooperativity in water-mediated interactions: the collapse of a superhydrophobic polymer and the assembly of superhydrophobic methane-sized particles. In both cases, the addition of a LD potential generally improves the ability of the CG models to capture distributions of structural metrics like radius of gyration and cluster size. At the same time, the LD potential improves the transferability of the CG model to related systems with different polymer lengths or concentrations. In many cases, the optimization of a

LD potential alone seemed sufficient to capture the solvent-mediated component of the effective CG force field; however, the optimization of renormalized pair interactions, along with a local density potential, seemed to offer slightly better CG models and transferability. In addition, the superhydrophobic case studies point towards significant multibody interactions that would suggest a role for the LD potential to improve the CG models. However, when the systems are only mildly hydrophobic—accomplished by introducing weak van der Waals solute-water interactions—the improvement afforded by LD potentials is less significant.

It is interesting that such a simple CG potential, based on a single mean-field (the local density), is capable of capturing higher order cooperativity inherent to hydrophobic interactions—without being pre-informed about the microscopic network-forming, tetrahedral, open nature of water structure, or about the unique entropic-enthalpic balance of hydrophobic solvation from a macroscopic point of view. This feature suggests that bottom-up coarse graining techniques using LD and other non-traditional potential forms may offer robust strategies for implicit solvation models that incorporate the hydrophobic effect.[82,85,120] More generally, local density potentials also appear to be a promising tool in the relatively sparse repertoire of methods for developing fast and efficient CG descriptions of phase transitions, where density and concentration-dependent CG interactions are significant. Further, LD potentials can be easily applied to arbitrary CG models of systems, beyond those involving solvation and may have a wide role for improving the ability of CG forcefields to capture multibody effects that emerge as a result of integrating out degrees of freedom.

Appendix

2.A Relative sensitivity of the CG pair potential to inclusion of a local density field

Fig. 2.A.1 (upper panels) shows the sensitivity of the pair potential part of the CG forcefield for the superhydrophobic (WCA) polymer and methanes. The inclusion of the LD potential (SPLD model) makes the pair potential part of the forcefield less attractive, which is expected since the LD potential shares the (net) attractive manybody effect of the monomers that was previously sustained entirely by the pair potentials of the SP model (blue line). The lower panels demonstrate the intra-monomer pair correlation functions for the AA system and the different CG models. The SP and SPLD models capture these correlations equally well for both the polymer and the methane. For the polymer, even the LD model, in which pair potentials are not optimized, is able to reproduce the radial distribution function quantitatively. Note that the LD-only model was not

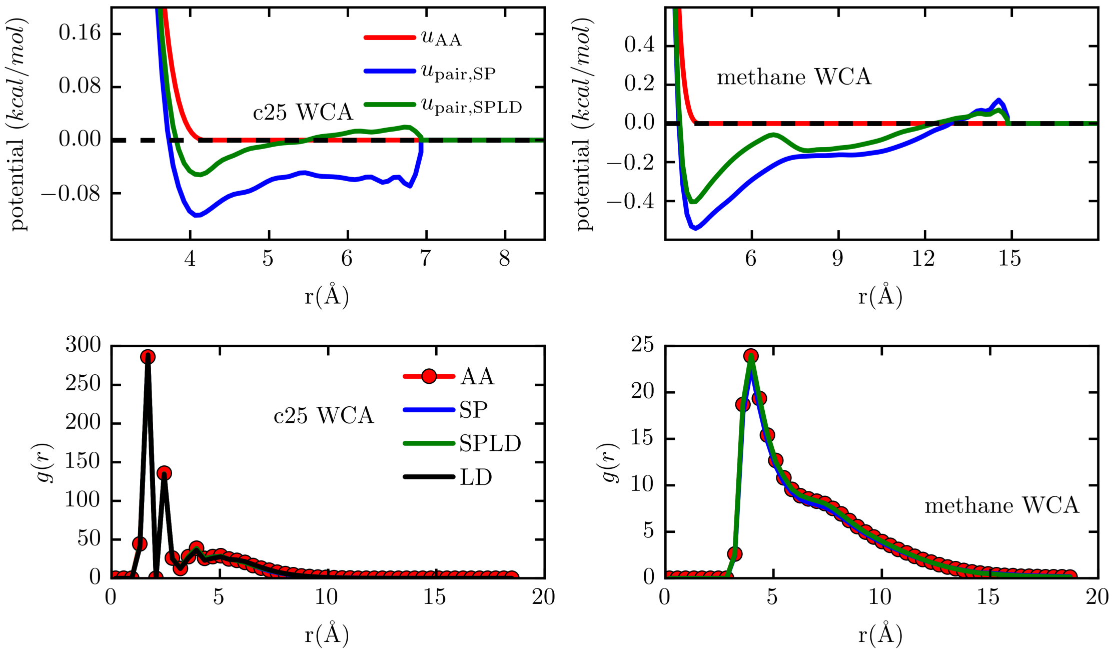


Figure 2.A.1: Upper panels show the pair part of the CG forcefield for the superhydrophobic polymer and free methane systems. For both, the inclusion of the LD potential (SPLD model) makes the CG pair potentials less attractive. Lower panels present a comparison of the intra-monomer radial distribution function between the AA system and the different CG models. For the polymer, even the LD-only model with no CG pair potential accurately reproduces the pair structure. (Note that all radial distributions approach zero because neither the monomers nor the methanes are bulk dispersed, the latter forming a dense cluster.)

optimized for the free methanes, based on our observation that the LD potential functions best only when supported by CG pair potentials.

Fig. 2.A.2 compares the R_{EE} transferability of the SP and SPLD models to the pair component of the SPLD model (labelled SPLD'). Clearly, the pair component alone is incapable of producing good transferability for R_{EE}. Similar results are

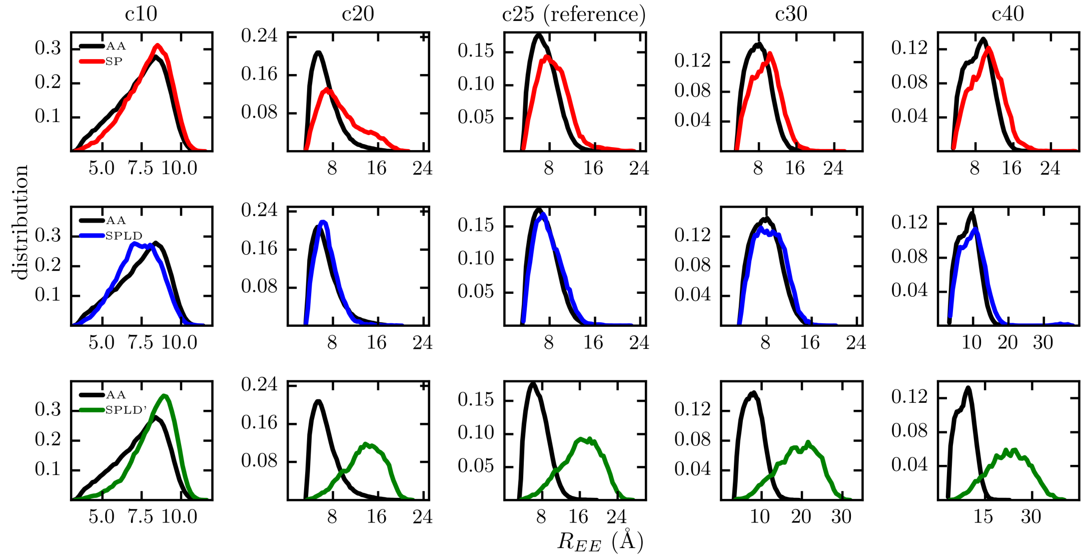


Figure 2.A.2: Comparison of the relative transferabilities of the c25 superhydrophobic WCA polymer end-to-end distance for a special forcefield (labelled SPLD') that considers only the CG pair component of the SPLD forcefield, with that for the SP and the SPLD models. Clearly, the pair component, although different from the SP model, produces poor transferability by mispredicting the distribution peaks.

noted (not shown here) for Rg transferability with this forcefield. It may be noted that though the SP and the CG pair component of the SPLD models in Fig 2.A.1 (top left panel) are similar in shape (other than the difference in well depth) for low and moderate distances, they are remarkably different at distances near the cutoff. This may explain why the pair part of the SPLD model produces vastly different transferability from the SP model. In general, while the sensitivity of the renormalized pair potentials to the local density field may be straightforward to predict (typically), transferability of thermophysical properties derived from different CG models is usually highly nontrivial to anticipate.

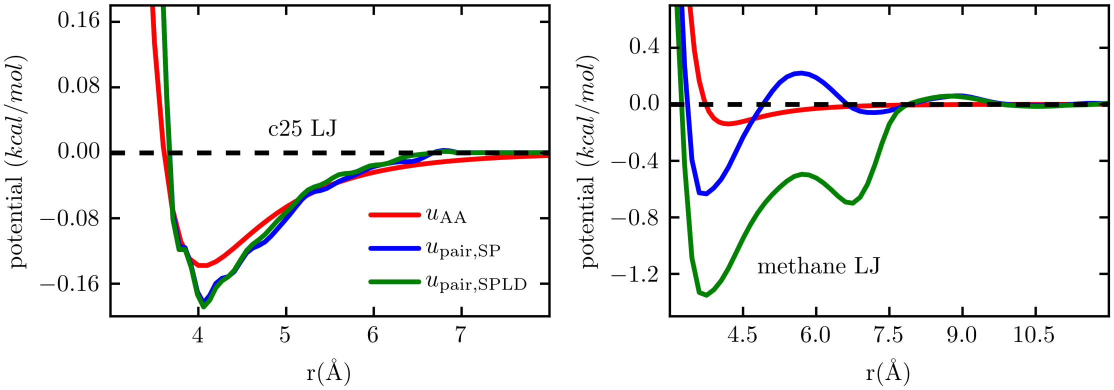


Figure 2.A.3: Sensitivity of the renormalized CG pair potential to the inclusion of the local density field for the LJ polymer (left) and free methanes (right). For the polymer, this sensitivity is low, presumably due to the lack of manybody effects. For the LJ methanes, however, relative entropy minimization produces a relatively repulsive LD potential (not shown here), which leads to more attractive character in the CG pair component of the SPLD model to maintain the same overall level of inter-monomer repulsion as that in the explicit water AA reference.

Fig. 2.A.3 compares the pair potential part of the SP and SPLD CG models for the LJ polymer and the LJ free methanes. Due to the absence of significant manybody effects, the SP and SPLD pair potentials for the LJ polymer are nearly the same, i.e. the renormalized pair potential of the SP model (blue line) does not change much after incorporating the local density field. For the case of the free methanes, however, inclusion of the LD potential (green line) makes the renormalized pair part more attractive. This likely stems from the need to maintain a (net) repulsive inter-monomer potential similar to that in the AA system, in the absence of bonds (which provide a natural separation e.g. between the beads in the polymer). The LD potential for the free LJ methanes optimized through relative entropy minimization (not shown here), is largely repulsive at higher local densities and so the CG pair component compensates to control the overall repulsive behavior and keep it close to that of the original methane-water interaction in the AA system.

2.B Local density distribution in c-25 versus c-40

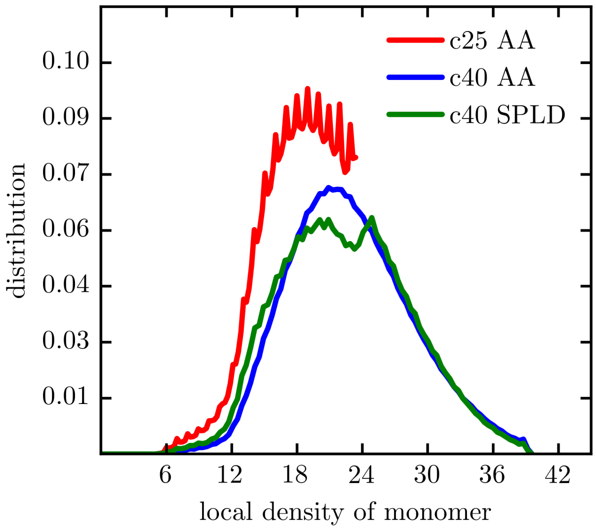


Figure 2.B.1: Comparison of the inter-monomeric local density distribution for the all atom trajectory of the superhydrophobic 25-mer with the all-atom and the SPLD models for a superhydrophobic 40-mer. The 25-mer embeds information about local density only up to 25 monomers which is significantly less than that conveyed by the 40-mer. The SPLD model is designed from the CG forcefield parameterized from the all atom 25-mer and thus fails to capture the peak of the true LD distribution for the 40-mer.

Fig. 2.B.1 compares the distribution of local densities of the monomers for all atom c25 with the all atom and SPLD CG models for c40 (using the same LD cutoff, r_c = 7.8 Å). Although both the c25 and c40 LD distributions peak near 25 monomers, a significant portion of the distribution ($\sim$ 50 %) for c40 contains information about more than 25 monomers. Since the LD potential is multibody, the difference in spread (standard deviation) between the c25 and c40 distributions (in spite of possible agreement in the mean local densities) leads to non-trivial and difficult-to-intuit differences between their folding behavior. The SPLD model for c40, simulated using the forcefield parameterized from the all atom c25 trajectory lacks information beyond 25 monomers, which leads to inaccurate location of the distribution peak (green line).

2.C Statistical uncertainty for the LD potential in the CG model of superhydrophobic methanes

The variation in the LD potential at low density in Fig. 2.11 is significant and does not reflect statistical uncertainty. To confirm, we re-optimized the LD potential for the SPLD model of the superhydrophobic methanes, starting from different initial values for the spline knots that describe the LD potential. Fig. 2.C.1, presents the mean value of the LD potential from these runs, zoomed in on the low local density region. The two peaks are clearly outside the relative

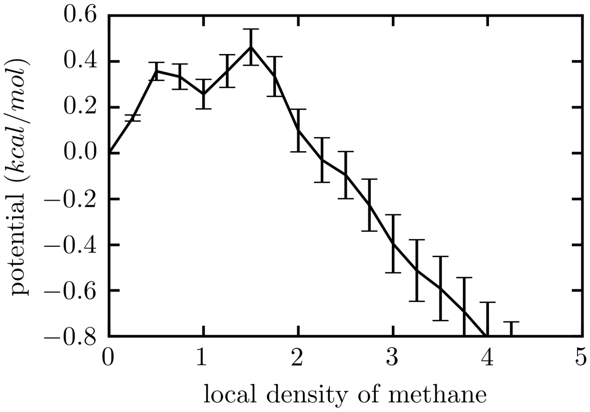


Figure 2.C.1: Statistical uncertainty of the local density potential in the SPLD model of the free methane system. The presence of the peaks at low local densities is not statistical artifact.

statistical uncertainty, which has an average value of 15% (standard deviation normalized by the mean and averaged over the entire range of local densities). This has also been stated in the caption for Fig. 2.11.

Chapter 3

Transferable coarse-grained models of liquid–liquid equilibrium using local density potentials optimized with the relative entropy

3.1 Introduction

Equilibrium fluid phase transitions play a pivotal role in many technologies, ranging from complex fluids in consumer products to separation strategies in large-scale chemical processing. Particularly for liquid mixtures, phase transition forms the basis of liquid-liquid extraction, a unit operation with widespread applications in food processing, organic synthesis, petroleum refineries, renewable energy, nuclear reprocessing, and biotechnology, for example. Molecular simulation methods for predicting phase equilibria for small, relatively rigid molecular species are now well established[121, 122] and typically require clever sampling Monte Carlo (MC) techniques[123, 124] such as Gibbs-ensemble MC,[125] multiple histogram reweighting,[126, 127] or transition-matrix MC.[128] Tackling the phase equilibrium problem for large, flexible, or asymmetrically sized species remains a critical challenge and a major research effort. Coarse-grained (CG) models potentially offer a promising route for complex phase equilibrium calculations through simpler representations that are dramatically easier to sample, particularly for MC algorithms that involve particle insertions but also for direct interfacial simulations using molecular dynamics (MD) that require long equilibration run times.

Indeed, the past decade has seen significant progress in CG model development, with a particular effort directed toward biomacromolecular systems (e.g. polypeptides and polynucleotides) and their folding and self-assembly. CG descriptions of biomolecules have been motivated both by bottom-up methods that parametrize on the basis of small representative all-atom models and by top-down approaches that tune CG interactions to match macroscopic thermophysical properties. A number of these approaches have proven successful in capturing first-order-type structural and thermodynamic transitions. A few examples include morphological phase transitions in membranes and bilayers[24, 129–131] (e.g. using the MARTINI CG model[132]), and studies of folding–unfolding transitions in proteins and peptides.[133–140] Despite the burgeoning success in the biomolecular realm, CG models do not yet seem widely capable of capturing fluid phase equilibria (although there are notable efforts along this direction, as we shortly describe), which has limited their routine application to chemical thermodynamics problems.

The main roadblock in developing CG models of equilibrium fluid phase behavior stems from their lack of transferability, i.e. their limited ability to translate to thermodynamic states (density, composition, etc.) different from the one at which they were parametrized. This prevents accurate realization of phase behavior that inherently spans the range of states encompassing the phases of interest, and is a particular problem for bottom-up CG strategies. One contributor to transferability issues is neglect of the strong coupling between the reduced CG degrees of freedom in the model. In principle, an "ideal" bottom-up CG force field is represented by a highly multidimensional free energy function $W(\mathbf{R})$, constructed by projecting the all-atom (AA) potential $U_{AA}(\mathbf{r})$ on the CG degrees of freedom $\mathbf{R}$:[52]

$$W(\mathbf{R}) = -k_B T \ln \int_V d\mathbf{r} e^{-\beta\, U_{\mathrm{AA}}(\mathbf{r})}\ \delta[\mathbf{R} - \mathbf{M}(\mathbf{r})] \tag{3.1.1}$$

Here, $\mathbf{M}$ is a "mapping function" that translates an AA configuration r to a CG one R, and $\beta = 1/k_B T$. Unfortunately, $W(\mathbf{R})$ is highly multibody and difficult to implement practically; instead, CG force fields typically contain pairwise nonbonded interactions between CG sites (in addition to conventional bonded terms), effectively ignoring true higher order many-body correlations in the underlying potential of mean force.[21,22,26,32,44,141] In turn, this tends to make the CG model sensitive to the thermodynamic state (density and/or composition) at which it is developed, and limits not only the model's capability to scale to other states but also its ability to simultaneously reproduce different properties even at the reference state, as discussed extensively in the works of Louis, Head-Gordon, and co-workers.[51,62]

One resolution that has been particularly successful for CG water models has been to include explicit three-body terms.[69,72,73] However, such approaches may become computationally expensive to parametrize and use in general, particularly if the form and order of the physically relevant interactions are not known. A second approach has been to include information about the local environment around a pair of CG particles to modulate the pair potential between them.[65–67,75] Recently, Voth and co-workers developed a theory of ultra-coarse-graining (UCG) that produces low resolution CG models where CG sites embed discrete internal states that are in a state of local quasi-equilibrium.[142] They used this technique to mix separate CG force fields for different phases in a phase-separated liquid mixture, based on the location of the phase interface as well as a local-density parameter that can discriminate between the two phases. These authors applied the approach to vapor–liquid equilibrium in a Lennard-Jones fluid and the cooperative aggregation of neopentane in methanol. Noid and co-workers also developed transferable CG models of heptane–toluene mixtures by expanding the force field with global volume-based corrections, such that the model reproduces correct NPT fluctuations in density and pressure and consequently the pressure–volume equation of state.[143]

Here, we take a distinct approach and use so-called "local density" potentials to expand the CG force field with mean-field representations of multibody effects that in turn improve transferability. In this case, the CG potential is inspired by mean-field embedded-atom models of metals,[57] in which sites have energies that directly depend on the local density of neighboring CG sites (within a cutoff radius), and these potentials serve as an additive correction over traditional pair interactions.[144] Such potentials are mean-field in nature, which allows them to remain inexpensive, but they account for higher-order interactions beyond pair in a manner modulated by the local environment of a particle, such as the local coordination number or composition. Allen and Rutledge initially explored a similar approach in which they supplemented conventional solute–solvent pair interactions with effective solvent-mediated intersolute interactions.[65,66,145] They expressed these additional interactions in terms of the excess chemical potential associated with transferring a solute from a solvent-exposed to a completely solute-locked state, and parametrized the excess chemical potential as functions of both the global and local solute density.

In the previous chapter, we introduced a general approach to CG local density potentials that we tested in model aqueous solutions, where water was coarse-grained away to an implicit description with only effective intersolute interactions remaining.[144] In that effort, we found that both the fidelity of the CG model and its transferability significantly improved with the addition of local density potentials. Voth and co-workers subsequently employed local density dependent interactions to improve the characterization of vapor–liquid interfaces in methanol and acetronitrile.[5] Noid and co-workers also found that local density dependent potentials improved CG models of methanol, allowing parametrization in the bulk liquid that transfer well to the vapor–liquid coexistence.[6]

Here, we extend our previous work on local density potentials to outline a general strategy for transferable CG models suitable for phase equilibria, using a coarse-graining theoretical framework based on the relative entropy.[32,34] As with conventional bottom-up coarse-graining, we use an underlying AA reference simulation to parametrize CG pair interactions between species, but we also simultaneously parametrize intra- and interspecies local-density potentials. While our previous work showed that local density potentials aid in capturing cooperative folding and association like behaviors for solutes in implicit solvent,[144] the present work explores the generality of local density potentials in mixtures capable of macroscopic phase separation into chemically distinct environments.

As a case study, we investigate macroscopic phase separation in benzene–water mixtures, which provides an excellent and challenging test system because of the water–benzene size asymmetry and the intrinsically multibody nature of water-mediated hydrophobic interactions.[86, 88, 146–149] An earlier study of this system by Villa et al.[4] examined the transferability of single-site CG representations of benzene and water parametrized from very dilute benzene solutions (0.1 M), and using an advanced implementation of the iterative Boltzmann inversion technique. These CG models demonstrated transferability in describing structural and thermodynamic metrics like pair correlations and the chemical potential at low concentrations but produced qualitatively incorrect behavior at high benzene concentrations ($\sim$ 9.5 M) that missed the macroscopic liquid–liquid phase separation, illustrating inherent challenges in capturing this system's composition transferability.

In this chapter, we develop force fields for single-site CG models of benzene–water mixtures that augment CG pair interactions with local density potentials. Here, we add four distinct such potentials that modulate the energy of a CG molecule based on its identity (benzene or water) and its average (mean-field) concentration of either species within a short-range distance cutoff. We parametrize CG models from AA reference systems at several distinct compositions to investigate the effect of AA reference on model quality. Subsequently, we benchmark the structural and thermodynamic transferability of the CG models across composition space, spanning both sides of the phase transition point predicted by the reference AA force field. We discuss both the improvements the local density strategy enables and the limitations that we find.

3.2 Methods

3.2.1 CG model and force field design

The CG model, as illustrated in Fig. 3.2.1, maps benzene and water molecules to single sites and, in that sense, is an explicit water CG model unlike our previous test of local-density CG interactions.[144] In the present case, the local density is

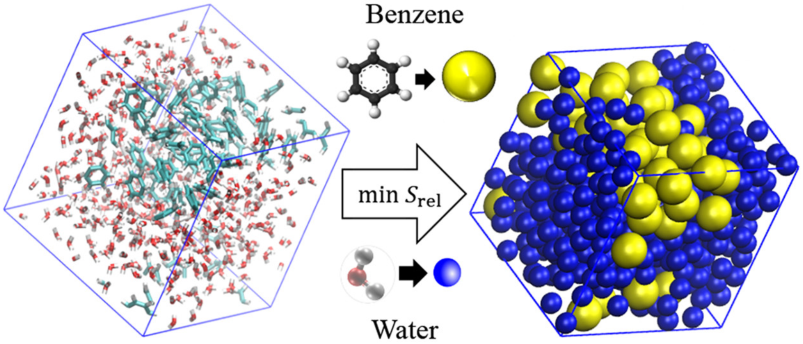


Figure 3.2.1: Atomistic system (left) and single site CG model (right) of a 10% mole fraction benzene / 90% water solution, with a total of 500 molecules. The cubic box length of 27.45 Å is tuned to achieve 1 atm pressure in the AA system at 300 K. In the CG model, molecular orientational degrees of freedom are coarse-grained away and effective intermolecular interactions are determined by relative entropy minimization.

essentially a local coordination number, and is given for a central CG site i of type α with neighboring sites j of type β by

$$\rho_i^{\alpha\beta} = \sum_{\substack{j\neq i \\ j\in\beta}} \varphi(r_{ij}) \tag{3.2.1}$$

where $varphi(r_{ij})$ is an indicator function based on the pair distance r_{ij}, that sharply but smoothly interpolates to zero at a cutoff radius r_c. We use the computationally efficient form

$$\varphi(r) = \begin{cases} 1, & r \leq r_0 \\ c_0 + c_2 r^2 + c_4 r^4 + c_6 r^6, & r \in (r_0, r_c) \\ 0, & r \geq r_c \end{cases} \tag{3.2.2}$$

The coefficients c_0, c_2, c_4, and c_6 are determined? by imposing continuity of φ and its first derivatives at the cutoff r_c and a slightly smaller inner-cutoff r_0, which we fix as $r_c - 1\text{Å}$ for the this work. The LD potential of type $\alpha\beta$ due to all central CG sites of type α and neighbors of type β then follows

$$U_{\mathrm{LD}}^{\alpha\beta} = \sum_{i \in \alpha} f^{\alpha\beta}\left(\rho_i^{\alpha\beta}\right) \tag{3.2.3}$$

where the summation proceeds only over atoms of type α. The functions $f(\alpha\beta)$ are yet unknown and are determined as splines by optimizing the CG model. Although $U_{\mathrm{LD}}^{\alpha\beta}$ is a mean-field many-body potential, it gives rise to a pair-additive force

$$\mathbf{f}_i^{\alpha\beta} = -\sum_{\substack{j \neq i \\ i \in \alpha, j \in \beta}} \left[\frac{df^{\alpha\beta}\left(\rho_i^{\alpha\beta}\right)}{d\rho} + \frac{df^{\beta\alpha}\left(\rho_i^{\beta\alpha}\right)}{d\rho} \right] \frac{d\varphi(r_{ij})}{dr} \, \frac{\mathbf{r}_i - \mathbf{r}_j}{r_{ij}} \tag{3.2.4}$$

The reader is referred to Section 2.2.1 in chapter 2 for further details on the formulation of LD potentials, including expressions for the coefficients in Eq (3.2.4).

We consider several illustrative cases for CG force fields. Each contains three distinct CG pair interactions modeled by splines, in addition to potentially four LD interactions between all possible species and environment types. For convenience, the CG pair potentials are referred to as simply pair-$\alpha\beta$ or $\alpha\beta$ pair potentials, while the LD potentials are annotated as LD-$\alpha\beta$, where α, β = B (benzene) or W

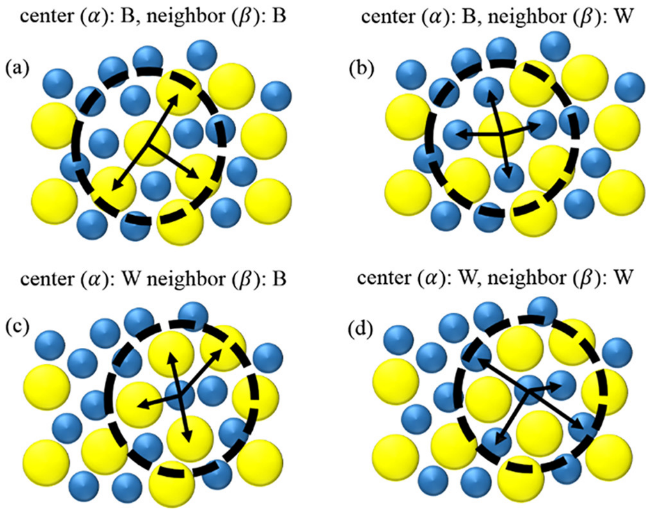

center (α): B, neighbor (β): B
(a)
center (α): B, neighbor (β): W
(b)
center (α): W neighbor (β): B
(c)
center (α): W, neighbor (β): W
(d)

Figure 3.2.2: Schematic representations of central (α) and neighboring (β) species configurations for the different types of LD potentials outlined in Table 3.2.1. Yellow CG sites denote benzene, and blue CG sites are water.

(water). Note that pair-$\alpha\beta$ and pair-$\beta\alpha$ denote the same potential, while LD-$\alpha\beta$ and LD-$\beta\alpha$ do not, given the asymmetric role between central and neighboring site types. Fig. 3.2.2 schematically shows the different types of possible LD-$\alpha\beta$ potentials. If all possible LD potentials are included with the three pair potentials, the overall CG Hamiltonian reads as

$$\begin{aligned} U_{\mathrm{CG}} = &\sum_{\substack{i<j \\ i,j\in \mathrm{B}}} u_{\mathrm{pair}}^{\mathrm{BB}}(r_{ij}) + \sum_{\substack{i<j \\ i,j\in \mathrm{W}}} u_{\mathrm{pair}}^{\mathrm{WW}}(r_{ij}) + \sum_{\substack{i<j \\ i\in \mathrm{B} \\ j\in \mathrm{W}}} u_{\mathrm{pair}}^{\mathrm{BW}}(r_{ij}) \\ &+ \sum_{i\in \mathrm{B}} \left[f^{\mathrm{BB}}\left(\rho_i^{\mathrm{BB}}\right) + f^{\mathrm{BW}}\left(\rho_i^{\mathrm{BW}}\right) \right] \\ &+ \sum_{i\in \mathrm{W}} \left[f^{\mathrm{WW}}\left(\rho_i^{\mathrm{WW}}\right) + f^{\mathrm{WB}}\left(\rho_i^{\mathrm{WB}}\right) \right] \end{aligned} \tag{3.2.5}$$

We study a total of six distinct force fields, which are summarized in Table 3.2.1. This includes the control with only pair potentials, the full potential of Eq. (3.2.1) involving all four local density potentials, and four subsets in which only one of the local density terms is used. The cutoffs for determining local densities for the like LD potentials (LD-BB and LD-WW) are chosen to include the first coordination shell from their first minimum in the respective radial distribution functions; these distances are arithmetically averaged to determine the cutoffs for the unlike potentials (LD-BW and LD-WB).

Admittedly, including all four LD potentials (the pair + LD-all force field) may contain redundant information, since local densities essentially measure coordination numbers and these are subject to geometric constraints in condensed, incompressible fluids. To illustrate this point using simple lattice statistics, consider z^{BB}, z^{WW}, z^{BW}, and z^{WB} as the average nearest-neighbor coordination num-

Table 3.2.1: Case studies for local density potentials examined in this work*

CG forcefield case	Pair potentials	LD potentials
pair-only	pair-BB, pair-WW, pair-WW	None
pair + LD-WW	pair-BB, pair-WW, pair-BW	LD-WW
pair + LD-BW	pair-BB, pair-WW, pair-BW	LD-BW
pair + LD-WB	pair-BB, pair-WW, pair-BW	LD-WB
pair + LD-all	pair-BB, pair-WW, pair-BW	LD-BB, LD-WW, LD-BW, LD-WB

*We consider cases involving zero, one, or all possible (total of four) local density potentials that differ in the central and neighboring molecule types.

bers of the different types, for N total CG particles (benzene and water) with x_{B} benzene mole fraction (and $x_{\mathrm{W}} = 1 - x_{\mathrm{B}}$ water mole fraction) on a lattice with N sites. One might think of $z^{\alpha\beta} = \langle \rho^{\alpha\beta} \rangle$, where $\langle\ \rangle$ denotes ensemble averaging. If each molecule occupies a single lattice site and the total lattice nearest neighbor coordination number is z (e.g. $z = 6$ for a cubic lattice), balances on connections between sites produce $z = z^{\mathrm{BB}} + z^{\mathrm{BW}} = z^{\mathrm{WW}} + z^{\mathrm{WB}}$ and $x_{\mathrm{B}} z^{\mathrm{BW}} = x_{\mathrm{W}} z^{\mathrm{WB}}$. These three constraint equations show that, at most, one of the four possible local densities are independent once the mole fraction is specified. Of course, the off-lattice nature of the actual CG model means that the local coordination shell is not truly constrained to a fixed number of neighbors, and of course, the benzene-water size asymmetry introduces additional complications. Still, this analysis illustrates that the addition of all four local density potentials may not be necessary to cover the functional space or basis that these interactions provide beyond pair potentials, as long as the system remains in a condensed phase. In particular, this analysis also motivates consideration of CG force fields using a single LD potential, namely, the second, third, and fourth rows of Table 3.2.1

In our approach, we represent all CG interactions: both pair and local density—by cubic B-splines. The spline knots form the optimizable parameters of the model and are determined by minimizing the relative entropy between the model and its atomistic reference.[32,34] The relative entropy provides a way to quantitatively compare the quality of a CG model by measuring the loss in information upon coarse-graining the AA system to a reduced set of CG degrees of freedom with a particular CG force field. Minimizing the relative entropy increases the overlap between the ensemble microstate probability distributions of the AA and CG models, and provides a natural strategy for parametrizing CG force fields. In the canonical ensemble, the relative entropy takes the form[32]

$$S_{\mathrm{rel}} = \beta \langle U_{\mathrm{CG}}(\boldsymbol{\lambda}) - U_{\mathrm{AA}} \rangle_{\mathrm{AA}} - \beta (A_{\mathrm{CG}}(\boldsymbol{\lambda}) - A_{\mathrm{AA}}) + S_{\mathrm{map}} \tag{3.2.6}$$

where U_X and A_X are the potential and Helmholtz free energies in ensemble X = AA or CG, which are functions of the CG force field parameters $\boldsymbol{\lambda}$, namely, in this case, the spline knots of all component potentials. Here, S_{map} is a mapping entropy that measures the degeneracy associated with the AA to CG mapping; it is independent of the CG potential and thus does not depend on these parameters.

The derivative of the relative entropy in $\boldsymbol{\lambda}$ space (for the full potential with all LD terns) can be written as

$$\begin{aligned}\frac{\partial S_{\text{rel}}}{\partial \boldsymbol{\lambda}} = \beta &\sum_{\substack{(\alpha,\beta)\\ \in \\ (\text{BB, WW, BW})}} \sum_{i<j} \left\{ \left\langle \frac{\partial u^{\alpha\beta}_{\text{pair}}(r_{ij})}{\partial \boldsymbol{\lambda}} \right\rangle_{\text{AA}} - \left\langle \frac{\partial u^{\alpha\beta}_{\text{pair}}(r_{ij})}{\partial \boldsymbol{\lambda}} \right\rangle_{\text{CG}} \right\} \\ + \ \beta &\sum_{\substack{(\alpha,\beta)\\ \in \\ (\text{BB, WW, BW, WB})}} \sum_{i\in\alpha} \left\{ \left\langle \frac{\partial f^{\alpha\beta}(\rho^{\alpha\beta}_i)}{\partial \boldsymbol{\lambda}} \right\rangle_{\text{AA}} - \left\langle \frac{\partial f^{\alpha\beta}(\rho^{\alpha\beta}_i)}{\partial \boldsymbol{\lambda}} \right\rangle_{\text{CG}} \right\} \end{aligned} \tag{3.2.7}$$

where the set of local density potentials given by atom-type combinations (α, β) would differ across the different cases in Table 3.2.1. The coarse-graining algorithm proceeds by locating the minimum of S_{rel} in $\boldsymbol{\lambda}$ space, i.e. the zeros of Eq. (3.2.1), using a combination of nonlinear conjugate gradient and quasi-Newton methods. Details are provided in earlier work[34] and in Chapter 2.

3.2.2 Simulation details

Atomistic simulations of aqueous benzene solutions are carried out at 300 K and 1 atm pressure, with 500 molecules, spanning a range of benzene mole fractions from $x_{\text{B}} = 10$ to 90%. For these, we use the MD engine GROMACS (version 4.6.5)[150,151] and employ the GROMOS 53a6 force field[152] for benzene and the SPC/E model of water.[106] All bonds are constrained using the LINCS algorithm.(57,58) The system is first equilibrated in an NPT ensemble using the

Parinello–Rahman barostat[108] and Nosé–Hoover thermostat[153,154] for 44 ns to determine the equilibrium bulk density (data from the last 4 ns is used to estimate average equilibrium box volume) followed by a further 60–80 ns of equilibration under NVT conditions, the last 40 ns of which is used to collect trajectory data. It should be mentioned that a previous study inferred that the AA force field overpredicts the solubility of benzene in water by almost an order of magnitude;[4] while the experimental solubility is 0.02 M,[155] benzene and water were still found to be miscible at 0.5 M, which likely stems from the reported underestimation of benzene hydration free energy by the GROMOS 53a6 force field.[156] However, the present study is only interested in the ability of optimized CG systems to recapitulate correct AA properties, no matter how accurate these may be relative to reality; thus, this particular AA reference serves as an instructive model system. It is also worth noting that the AA systems are at concentrations higher than either the experimental or the AA solubility limit, so that the reference solutions are already phase separated. However, because of the small system size, the interfaces encompass a significant fraction of molecules, which allows the references to sample cross-interactions between the demixed species.

To study the effect of reference state point on the transferability, CG models are parametrized from 10, 50, and 90% benzene solutions, respectively. All CG pair potentials and BB, BW, and WB LD potentials are developed using cubic B-splines with 30 knot points, while the LD WW potential uses 60 knots (chosen higher initially to accommodate potentially complex water–water interactions). A cutoff of 10 Å is used for the pair potentials. Outer cutoffs for LD-BB and LD-WW potentials are estimated from the first solvation shell radii of B–B and W–W radial distribution functions, respectively (7.5 Å for B–B and 3.5 Å for W–W), while those for LD-BW and LD-WB potentials are computed as the arithmetic average of the above two (5.5 Å). The inner cutoff for all LD potentials is 1 Å less than these. MD simulations of the CG models are carried out in the NVT ensemble using the LAMMPS MD code,[105] modified to include local density potentials. We find that the LD-augmented CG systems provide at least a 15-fold speedup over the AA simulations; this might be further increased with optimization of the LD code.

3.3 Results and discussion

3.3.1 CG forcefields

We parametrize three versions of the six force fields introduced in Table 3.2.1, involving distinct atomistic references at compositions of 10, 50, and 90% benzene mole fraction, respectively. Fig. 3.3.1 compares the potentials for the pair-only and pair + LD-all force fields, as described in Table 3.2.1. For all of the references, the BB pair potentials with and without the use of LD potenteials have similar forms and are largely repulsive; only for the dilute 10% reference solution do they exhibit a slightly attractive ($\sim$ -0.2 kcal/mol) around 6.5 Å. It is instructive to note that this attractive well is not as significant as the value -0.6 kcal/mol reached for a pair-only CG BB potential reported previously that was developed from an even more dilute benzene solution ($x_{\mathrm{B}} \approx 0.002$).(52) On the other hand, the BB LD potentials are largely all attractive, and the version parametrized from the 50% solution shows the largest decrease of around 2.5 kcal/mol over the entire range of the BB local density.

In contrast to the benzene self-interactions, the WW pair potentials show significant attractive interactions; many have a typical double-well form. This outer wall near 6.5 Å is weakly attractive ($\sim$ -0.4 kcal/mol) and remains similar in magnitude regardless of the presence of LD potentials and parameterization reference. On the other hand, the depth of the inner well increases (to $\sim$ -0.5 kcal/mol) for the 50 and 90% references without LD potentials.When reoptimized with all LD terms, the inner core converts into a repulsive shoulder for the 90% reference but becomes very attractive ($\sim$ -0.8 kcal/mol) for the 10% solution. The LD-WW potential seems to compensate for these changes: it exhibits attractive behavior

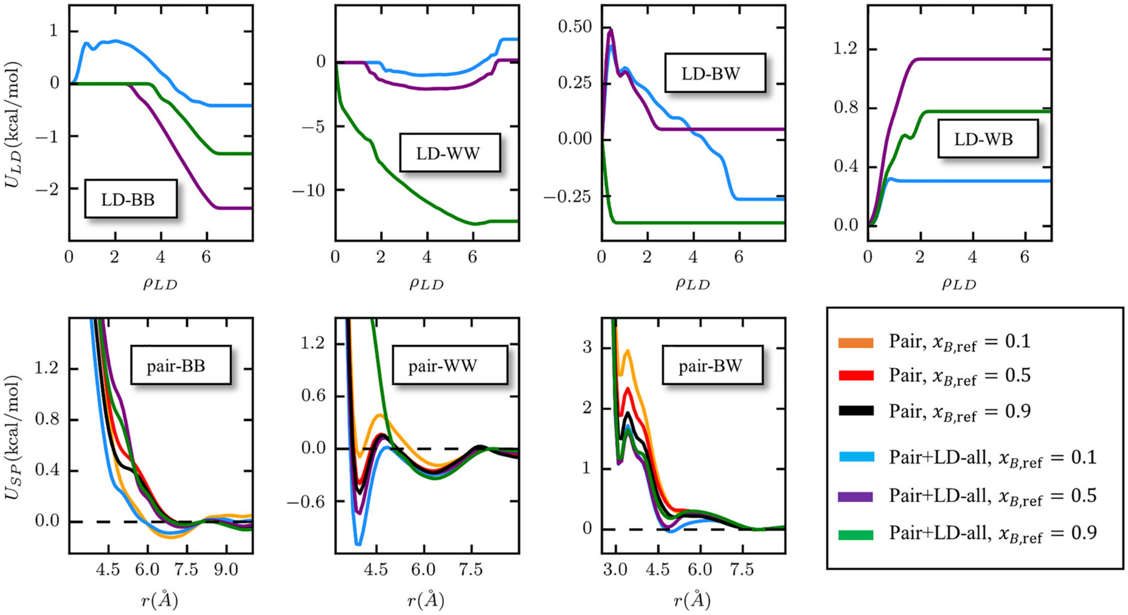


Figure 3.3.1: Intra- and intermolecular local density potentials (top row) and pair potentials (bottom row) optimized with the relative entropy from reference atomistic benzene solutions at 10% (orange and blue), 50% (red and purple), and 90% (black and green) benzene mole fraction. The BB and WW pair potentials show the most significant differences when optimized simultaneously with all the four LD potentials. Unlike LD potentials, potentials of types BW and WB are likely modulated by the small number of intermolecular multibody correlations at the interface, characterized by the low range of local density (~ 2) over which they change.

around a local density of 4 ($\Delta U_{\text{LD-WW}} \approx$ -2 kcal/mol) at 10 and 50% reference compositions and becomes the strongest of all four LD potentials for the 90% solution, decreasing by 12 kcal/mol. It is interesting to note that the minimum in the LD potential of the former two cases occurs near a coordination of four, consistent with the stabilization of water's tetrahedral coordination.

The BW pair potential is strictly positive but nonetheless contains a high-energy inner minimum within the core-repulsive region at 3 Å and a shallow outer minimum or shoulder near 4.5 Å. The pair distances of these features remain constant across all reference solutions and are agnostic to the inclusion of LD potentials. When parametrized without LD potentials, the inner minimum in the BW pair potential decreases (by ~ 0.5 kcal/mol) when moving from the 10 to 90% reference systems, but this variation vanishes when LD potentials are included. The corresponding optimized interspecies LD potentials depend on reference compositions, but both the LD-BW and LD-WB interactions are relatively small in magnitude. The LD-WB potential is always completely repulsive, increasing up to 1.2 kcal/mol over a short range of the W–B local density. The LD-BW potential is overall the weakest, and manifests both attractive and repulsive forms that seem specific to the inclusion of LD potentials and the particular reference composition.

All three pair potentials exhibit either multiple wells or a well-and-shoulder structure, which is commonly seen in CG models that coarse-grain away directional interactions like hydrogen bonds into spherically symmetric interactions.[39, 157, 158] The balance between the inner and outer wells in the WW interaction is well-known for water and, on the basis of the ratio of these characteristic distances, promotes tetrahedrally enriched liquid-phase correlations.[39, 159] The shoulder feature in the BB potentials likely also promotes specific geometries that result from the significant size asymmetry and packing due to benzene's aspect ratio. The addition of LD potentials, which extend only until the first coordination shell of the corresponding types, rather naturally then seems only to modify the inner well or shoulder of the corresponding pair potential.

It is instructive to note the possibility of interaction redundancies between the CG pair and LD potentials, which is distinct from redundancies within the set of LD potentials discussed in section II. In particular, if the LD potential is linear in the coordination number, then it can mimic the behavior of a short-range pair potential (i.e. with an energy that scales with the number of nearest neighbors), leading to the possibility that effective interactions might shift between the LD and pair potentials. We see this in the pair + LD-all CG models from the 90% reference solution case, where the WW interactions involve an entirely repulsive pair potential such that all short-range attractions shift into an almost linear LD WW potential. This may be due to the small number of water molecules in that case and consequently insufficient local crowding and reduced multibody correlations that weaken the role of the LD potential.

It may be noted that each LD potential acts at relatively low local density values, corresponding to small changes in local coordination number, before saturating to constant values for larger variations. The LD-BB and LD-WW potentials saturate near 5–6 benzene and water neighbors, respectively, while the LD-BW and LD-WB potentials show variation only largely between zero and two neighbors (except for the LD-BW potential from the 10% solution). These results may suggest that most of the needed multibody correction needed occurs at low local densities, while the optimized pair potentials are able to capture remaining interactions. Indeed, the calculated local density distributions (discussed shortly) show that the biggest corrections occur at low coordination numbers. Interestingly, the unlike LD potentials (LD-BW and LD-WB) have much smaller magnitudes compared to the like-species potentials, perhaps because B-W and W-B interactions are limited to the thin interfaces in the phase separated AA reference solutions. Ultimately, the same-species local density interactions, i.e. LD BB and LD-WW, seem to capture the most important multibody effects in this system.

It should be mentioned, however, that the choice of knot density in the cubic B-splines that represent the LD potentials can slightly influence the magnitudes and shapes of all of the LD potentials. In principle, a very large number of knots will approach a limiting form for each LD potential; however, because some of them vary significantly over small ranges of local density, even a relatively dense number of knots may not quite reach this limit. In the present investigation, we use 0.3 Å/knot for the pair splines, 0.66 local density per knot for the BB, BW, and WB potentials, and a finer spline of 0.1 local density per knot for the LD-WW interactions. Even with this relatively high knot resolution, we still detect some minor challenges in modeling sharp peaks that appear in the LD distributions at low coordination numbers. However, further increases in resolution make the relative entropy minimization more difficult, and thus, we remain with the present parametrization as a balance of accuracy and efficiency.

3.3.2 Structural transferability

We first evaluate the relative transferabilities of the CG models in Table 3.2.1 by comparing structural properties across the entire composition spectrum, beyond the specific compositions at which they are parametrized. We consider the radial distribution functions (RDFs or $g_{\alpha\beta}(r)$, where $\alpha\beta =$ BB, WW, and BW) as well as the local-density distributions ($P_{\text{LD}}(\rho_{\alpha\beta})$, where $\alpha\beta =$ BB, WW, BW, and WB). For the sake of brevity, two examples of structural correlation functions are presented here; the remainder have qualitatively similar behavior and are given in the Appendix (Figures 3.A.1 - 3.A.5).

Fig. 3.3.2 shows the transferability of B-W pair correlations and local density distributions across benzene compositions for the different CG model. For the RDFs, all CG models show good representability at the composition at which they were developed, as expected from the relative entropy strategy, which should match pair correlations when finely discretized and flexible spline pair potentials are used.[33] At other compositions, the pair-only models fare poorly: the one from 10 and 50% solutions overestimates the RDF peaks at higher compositions, while that developed from the 90% solution underpredicts the RDF peak at 10% composition. The pair + LD-BB and pair + LD-WB models seem equally limited in their ability to capture states beyond the reference composition. On the other hand, the models with LD-WW and LD-BW potentials perform much better and the pair + LD-all CG model is nearly quantitatively transferable to all of the compositions in terms of RDF reproduction. A similar trend emerges in the LD distributions, where the pair-only model shows limited fidelity, even at the reference composition, and grossly overpredicts both the peak and the spread of the distribution at other compositions. Once again, better transferability in the LD distribution results from the inclusion of the LD-WW potential, and the pair + LD-all model scales reasonably well across the entire composition space.

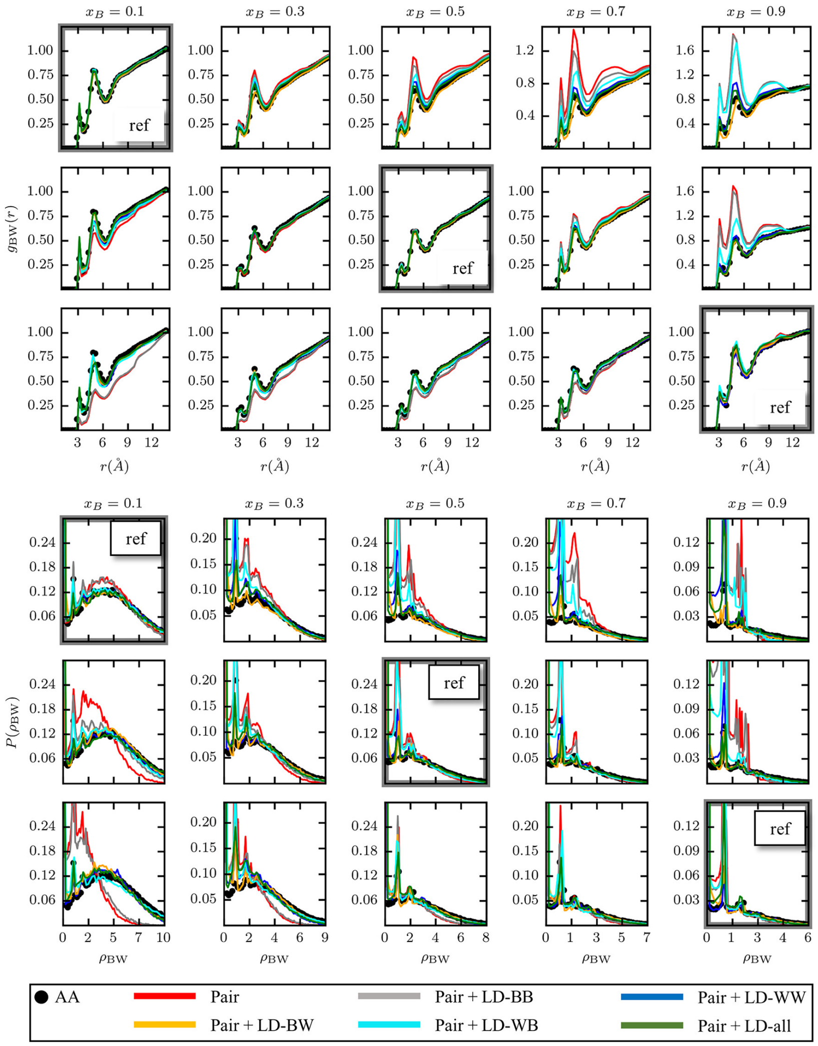


Figure 3.3.2: Transferability of B-W RDF (top panel) and LD distribution (bottom panel) between AA (represented by black dots) and CG models for the different CG force fields in Table 3.2.1. CG models for various combinations of these LD potentials are parametrized from reference compositions of 10, 50, and 90% benzene mole fraction (dark framed plot marked “ref”, along the rows). CG models constructed at a particular AA reference are simulated at compositions spanning 10-90% benzene (along the columns), which are used to calculate and compare RDFs and LD distributions with corresponding AA MD simulations at these compositions. The pair + LD-all CG model is nearly quantitatively transferable for both metrics at every composition.

Fig. 3.3.3 summarizes the overlap of all three RDFs and four LD distributions for the different classes of CG force fields across composition space and for different parametrization references. Here we compute the root-mean-square (RMS) differences between the AA and CG versions of these metrics, averaged over the distinct species combinations and weighed by composition

$$\begin{aligned} \mathrm{RMS}[g(r)] &= \Big[x_{\mathrm{B}}^2 \big\langle \Delta g_{\mathrm{BB}}^2 \big\rangle + x_{\mathrm{W}}^2 \big\langle \Delta g_{\mathrm{WW}}^2 \big\rangle + 2\, x_{\mathrm{B}} x_{\mathrm{W}} \big\langle \Delta g_{\mathrm{BW}}^2 \big\rangle \Big]^{\frac{1}{2}} \\ \mathrm{RMS}[P_{\mathrm{LD}}(\rho)] &= \Big[x_{\mathrm{B}}^2 \big\langle \Delta P_{\mathrm{LD,\ BB}}^2 \big\rangle + x_{\mathrm{W}}^2 \big\langle \Delta P_{\mathrm{LD,\ WW}}^2 \big\rangle \\ &\quad + x_{\mathrm{B}} x_{\mathrm{W}} \Big(\big\langle \Delta P_{\mathrm{LD,\ BW}}^2 + \big\rangle + \big\langle \Delta P_{\mathrm{LD,\ WB}}^2 + \big\rangle \Big) \Big]^{\frac{1}{2}} \end{aligned} \tag{3.3.1}$$

where for $\big\langle \Delta X_{\alpha\beta}^2 \big\rangle = \frac{1}{N} \sum \big(X_{\alpha\beta}^{\mathrm{AA}} - X_{\alpha\beta}^{\mathrm{CG}} \big)^2$, $X_{\alpha\beta} = g_{\alpha\beta}(r)$ or $P_{\mathrm{LD}}(\rho_{\alpha\beta})$ and N numbers the discrete values of the arguments (r or ρ).

Fig. 3.3.3a demonstrates that all of the CG models are accurate in predicting the RDFs at the reference composition at which they are parametrized, but the pair-only and pair + LD-BB force fields worsen significantly at other compositions. For calibration, a RMS error of 1.2 for the pair-only CG model at 10% composition corresponds to an overprediction of the WW RDF peak by 300% and BW peak by 113%, on average. On the other hand, the pair + LD-WW and pair + LD-all CG models have very low RMS errors across all of composition space. Remarkably, the high transferability of these models is not affected by the refer-

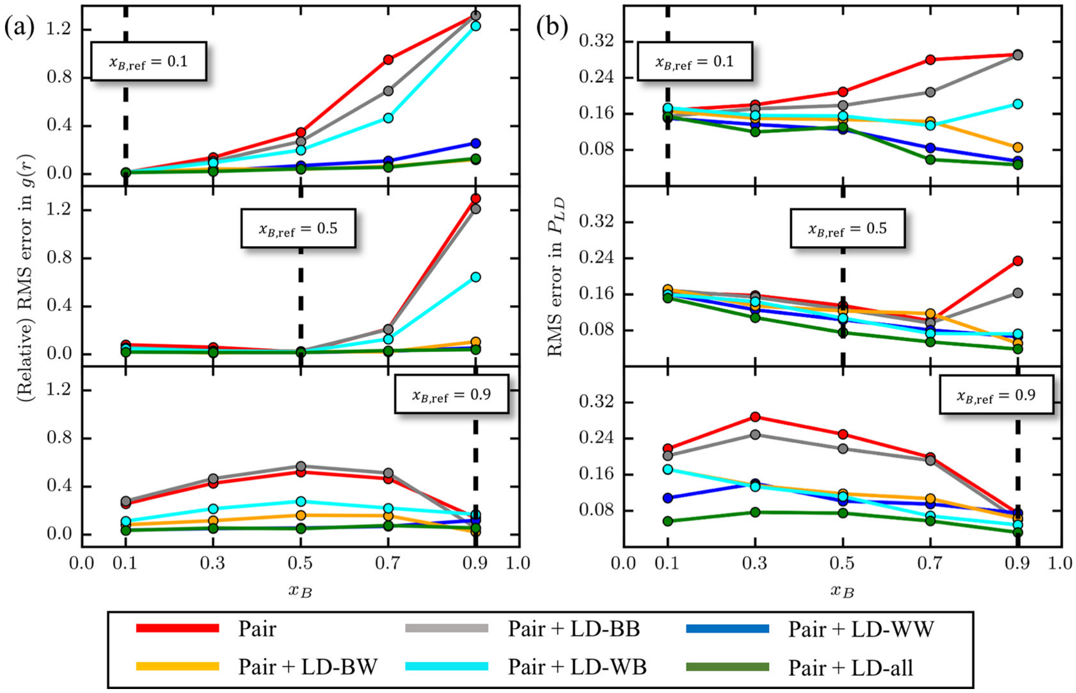


Figure 3.3.3: Composition weighted RMS overlap for (a) pair correlations and (b) local density distributions, between AA systems and CG models parametrized from 10, 50, and 90% reference compositions (along the vertical subpanels). The LD-WW potential is arguably the most important many-body interaction, and CG model transferability is significantly improved when it alone is included. CG models with only pair potentials or with intrabenzene LD potentials have poor transferability. The 50% reference CG model with all LD potentials performs best, being transferable to both more dilute and more concentrated solutions for both structural metrics.

ence composition. Particularly for the pair + LD-all case, models parametrized at either very high or low benzene concentrations perform well. The transferabilities in LD distributions in Figure 3.3.3b have a similar trend, but the RMS error is nonzero everywhere including the reference composition. This is likely because the absolute RMS metric used in Eq. (3.3.2) is sensitive to the large variation in the range of the (normalized) LD distributions across BB, WW, BW, and WB types, and very high-resolution splines for LD interactions would be needed to capture some sharp features of the LD distributions (which pose numerical challenges in the relative entropy minimization). Fig. 3.B.1 in the Appendix presents an alternate version of Fig. 3.3.3 where the RMS errors are normalized relative to the AA distribution for both pair and LD.

The poor performance of the pair + LD-BB force field and the high transferability of CG models that include the LD-WW potential show that, although the BB and WW LD potentials are comparable in magnitude (as observed in Fig. 3.3.1), the LD-WW potential is the more important multibody contribution. Admittedly, the reference solutions used in this study are already phase-separated, unlike previous efforts to parametrize benzene-water CG models from much more dilute conditions.[4] Intra-water local many-body correlations are likely particularly important to hydrophobic-mediated interactions and phase separated conditions, which we believe underlie the success of the LD-WW potential.

It should be mentioned that it is not clear if the LD strategy improves the structural transferability of the CG models across bulk density (or pressure). Figure 3.C.1 in the Appendix shows that LD potentials sometimes improve reproduction of pair correlation functions at slightly higher or lower densities, compared to pair-only models, but the effect is not uniform.

3.3.3 Thermodynamic transferability

We also consider the extent to which the CG models reproduce selected thermodynamic properties in addition to structural correlation functions. As a first example, we calculate the excess chemical potential due to inserting a hard sphere of radius R in benzene solutions of composition 10-90%, $\mu^{\mathrm{ex}}(R)$. Since a hard sphere interacts simply by excluding other particles within a threshold volume, $\mu^{\mathrm{ex}}(R) = -k_B T \ln P_{\mathrm{cav}}$ is calculated from the probability of finding a cavity of radius R in the equilibrium solution within the simulation box. For pure solvents, the hard sphere solvation chemical potential is directly related to the equilibrium fluctuations in local density of the solvent around solutes, which is an important measure of solute hydrophobicity in aqueous solutions.[160] In this case, we introduce hard spheres within the entire aqueous benzene solution and thus the resulting excess chemical potential is related to the species-averaged density fluctuations.

Fig. 3.3.4 compares the fractional error between AA and CG predictions of $\mu^{\mathrm{ex}}(R)$, i.e. $\left(1-\frac{\mu^{\mathrm{ex}}_{\mathrm{CG}}}{\mu^{\mathrm{ex}}_{\mathrm{AA}}}\right)$, for the different CG models. Direct comparison between AA and CG excess chemical potentials can be found in the Appendix (Fig. 3.D.1). The AA and CG predictions match nearly exactly for hard spheres of radius ≤ 1.5Å for all force fields. For $R \geq 1.5$Å, the pair + LD-WW and pair + LD-all models produce $\sim$ 10-15% deviation from the AA predictions at the reference compositions. This error is magnified at compositions away from the reference for the 10 and 50% reference solutions but interestingly decreases to 8% when the pair + LD-all force field from the 90% reference is used at lower compositions. The pair-only, pair + LD-BB, pair + LD-BW, and pair + LD-WB force fields have higher error ($\sim$ 20-30%) regardless of the composition. In most cases, $\mu_{\mathrm{CG}}^{\mathrm{ex}} \leq \mu_{\mathrm{AA}}^{\mathrm{ex}}$, as can be intuitively expected, since coarse-graining multiple atoms into single, isotropic CG point molecules opens more effective free volume in the solution. However, the pair-only model from the 50% reference and several models from the 90% reference predict positive errors, i.e. $\mu_{\mathrm{CG}}^{\mathrm{ex}} \geq \mu_{\mathrm{AA}}^{\mathrm{ex}}$, when applied to lower compositions.

It is well-known from scaled-particle theory that μ^{ex} for a hard sphere solute of radius R is related to the solute contact density[161, 162]

$$\frac{d}{dR}\mu^{\mathrm{ex}}(R) = \frac{4\pi}{\beta} R^2 [\rho_{\mathrm{solvent}} G(R)] \tag{3.3.2}$$

where $G(R)$ is the first-peak value of the hard-particle-solvent RDF. Although this result strictly holds for pure solvents, it remains conceptually similar for liquid mixtures.[161] Therefore, hydrocarbon solvents, with typically lower contact

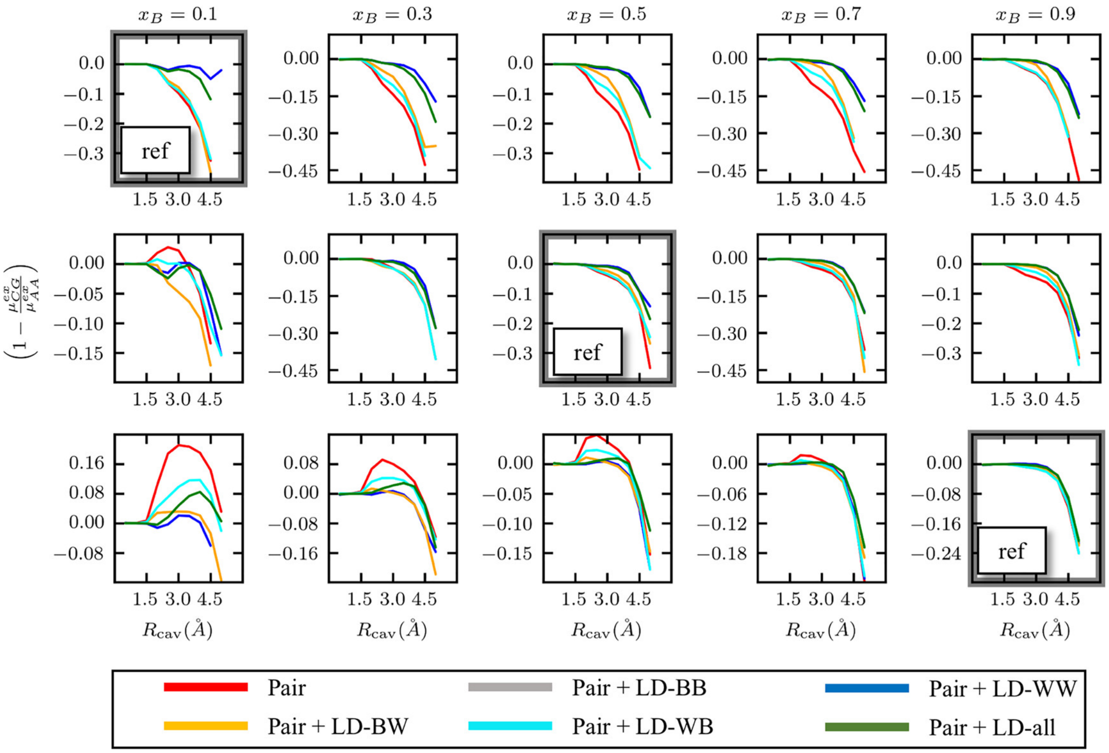


Figure 3.3.4: Fractional errors between AA and CG predictions of excess chemical potential for inserting a hard sphere with diameters ranging from 0 to 5 Å, for the different CG model force fields per Table 3.2.1. CG models are parametrized from reference compositions of 10, 50, and 90% benzene mole fractions (dark framed plot marked "ref" along the rows). Excess chemical potentials are calculated using the Widom test particle insertion method. Including W-W multibody interactions is essential for CG model transferability, as all models without this interaction predict insertion free energies (for $R \geq 1.5$Å) incorrectly modulated by composition.

densities for hard-particle solutes, are expected to exhibit smaller excess chemical potentials.[162] For most of the benzene solutions upward of 50% composition, it is likely that cavity volumes are much more frequent in the benzene phase, thus decreasing the free energy of insertion. CG force fields that do not include W-W multibody interactions (namely, pair-only, pair + LD-BB, pair + LD-BW, and pair + LD-WB) are sensitive to the system composition: when parametrized from the 10% reference and transferred to higher benzene compositions, they predict lower μ^{ex}, and when constructed from the 90% reference and transferred to dilute solutions, they produce systematically higher μ^{ex}. Thus, it appears that capturing W-W self-interactions accurately is crucial to model transferability.

To investigate the macroscopic phase separation between benzene and water and the associated equilibrium interfacial behavior, we perform MD simulations with 380 benzene and 1000 water molecules in a periodic box that is extended ($\sim$ 129 Å) along the z axis. The volume of this system is nearly 5 times larger than those of the parametrization references. The solution composition was chosen such that the equilibrium box volume determined through an initial (40 ns) NPT simulation produces an overall concentration of $\sim$ 7 M, which is close to the highest concentration limit ($\sim$ 9.5 M) at which pair-only CG models studied in ref. 4 fail to capture the macroscopic phase separation. Bulk density profiles are subsequently computed from sampling equilibrated trajectories ($\sim$ 60 ns) in the constant volume NVT ensemble, along the z direction.

Fig. 3.3.5 compares the bulk density profiles of benzene and water for the different CG models from the 10, 50, and 90% reference compositions. The AA system and all of the CG models demonstrate macroscopic phase segregation in which the benzene and water phases move to opposite sides of the box along the longest dimension. The pair-only, pair + LD-BB, pair + LD-BW, and pair + LD-WB models systematically predict up to 10% lower bulk densities for benzene and up to 40% higher ones for water, when parametrized from the 90% solution. Consequently, these force fields also produce a larger size of the benzene phase. While CG models that include the LD-WW potential reproduce the bulk densities nearly correctly, the pair + LD-all model from the 90% reference does not replicate the sharp segregation along the interface seen in the AA system, and leads to a longer benzene phase and a slight nonzero concentration of benzene in the water phase. It is worth noting that this particular model displayed a significantly different character in the water-water pair and LD potentials, as shown in Fig. 3.3.1 It may be difficult for this model to capture the sharp local density gradients across the phase interface.

The behavior of the benzene-water interface can be further quantified by calculating the interfacial surface tension (γ), calculated by the Kirkwood–Buff formula(73)

$$\gamma = \int_{-\infty}^{\infty} dz \left(p_{zz} - \frac{p_{xx} + p_{yy}}{2} \right) \tag{3.3.3}$$

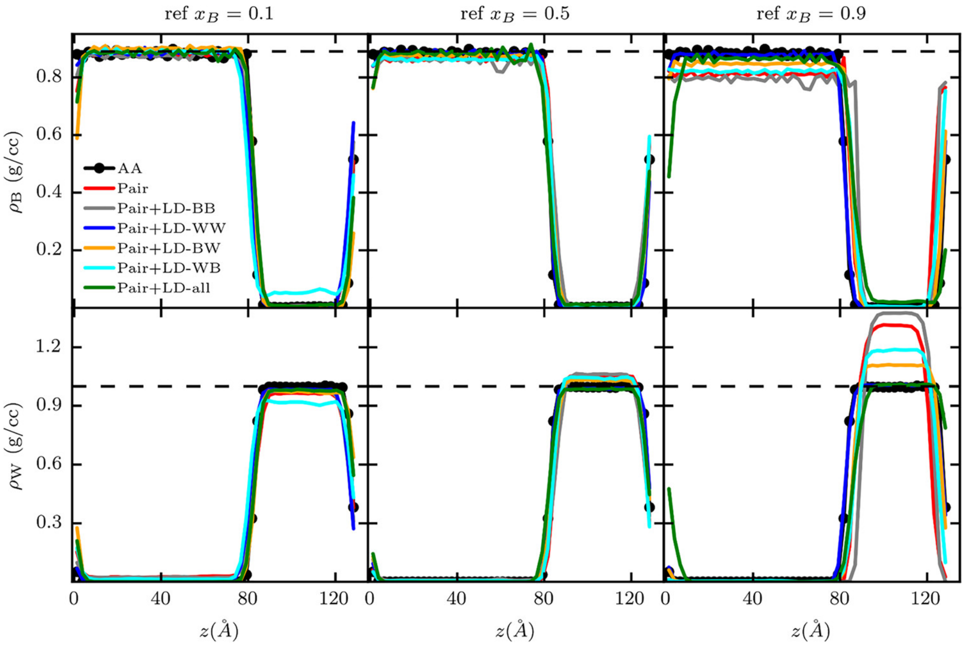


Figure 3.3.5: Macroscopic phase separation in large systems and transferability of density profiles of benzene (top panel) and water (bottom panel) at an overall composition of 27.5% benzene mole fraction in a long rectangular box, using the different CG models parametrized from 10, 50, and 90% reference compositions (across the columns). CG models without LD contribution of water (pair-only, pair + LD-BB) underpredict the equilibrium density of benzene and overestimate that of water, with the deviations increasing when the parameterization reference is more extreme in composition. Dotted lines are the experimental bulk densities of benzene (top panel: ~ 0.89 g/cm^3) and water (bottom panel: ~ 1.0 g/cm^3).

where p_{ii} is the diagonal component of the pressure tensor along the i axis ($i = x, y, z$) and $\pm\infty$ represents the benzene and water bulk phases. Our AA simulations and most of the CG models reveal two nearly sharp planar interfaces (as shown in Fig. 3.3.5) so that the difference between the pressure normal to the interface (p_{zz}) and the total tangential pressure $\left(\frac{p_{xx}+p_{yy}}{2}\right)$ is nearly zero everywhere except the interface, and p_{zz} remains roughly constant along the z axis.

For a finite box of length L_z in the z direction, the discrete version of Eq. (3.3.3) gives the interfacial tension (averaged over two interfaces due to periodic boundary conditions) as[163–165]

$$\gamma = \frac{L_z}{2}\left(\langle p_{zz}\rangle - \frac{\langle p_{xx}\rangle + \langle p_{yy}\rangle}{2}\right) \tag{3.3.4}$$

where $\langle\ \rangle$ denotes spatial averages along the z coordinate. Eqs. (3.3.3) and (3.3.4) can be sensitive to the cutoff for truncating pair potentials.[166] We perform AA simulations with nonbonded pair cutoffs of 10, 11, and 12 *angstrom*, and find that the interfacial tension does not change within statistical certainty between the last two cutoffs, with a value of 43.7 mN/m at 12 Å. Fig. 3.3.6 compares the AA surface tension with that of the different CG models in Table 3.2.1 and from the three parametrization references. The AA surface tension is $\sim$ 43.7 mN/m which overestimates by 36% the experimental value (32 mN/m);[167] this deviation is likely due to the overprediction of the solubility of benzene in water by the GROMOS 53a6 AA force field, as discussed in section 3.2.2. However, we note that our focus here is the ability of the CG models to capture the given reference, regardless of their absolute accuracy.

All CG models in Fig. 3.3.6 embed significant transferability errors such that predicted γ values systematically increase from low to high benzene concentrations. The absolute errors in the interfacial tension are lowest for the 50% refer-

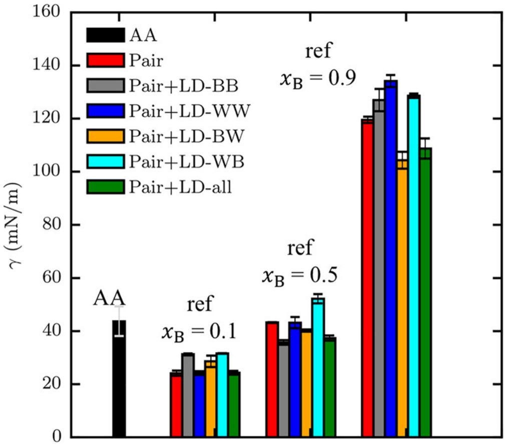


Figure 3.3.6: Comparison of benzene-water interfacial tension between the AA system and the different CG models from 10, 50, and 90% reference solutions, for a 27.5% benzene solution in a long rectangular box. The AA system overpredicts the experimental surface tension (∼ 32 mN/m) of the benzene–water interface,(78) but regardless of its absolute accuracy, all CG models embed different amounts of transferability errors relative to it. The pair-only and pair + LD-WW models parametrized from the 50% solution come closest to reproducing the AA surface tension. CG models from more concentrated reference solutions predict increasingly high surface tensions.

ence solution; under these conditions, the pair-only and pair + LD-WW models come closest to matching the AA value, underestimating it by 1%. CG models from the 10 and 90% references give values off by approximately 37 and 175%, respectively. It should be noted that the pressure tensor components for calculating the surface tension are obtained from constant volume (NVT) CG MD simulations (box dimensions are taken from the equilibrated AA simulation). We verified that the ensemble averaged pressure has no correlation with the surface tension, so that the differences in γ values are due to differences in the CG force fields and not differences in state.

Although the differences in bulk density profiles predicted by the LD-augmented CG models are small, they perform poorly in reproducing the thermodynamic pressure and consequently the interfacial tension. The average CG pressure varies between 4500 and 7000 atm with a 3% average (RMS) fluctuation from the mean. This is not terribly surprising, as bottom-up CG models are well-known to have difficulty capturing correct pressure, as has been explored in a number of papers.[22, 51, 62, 168–171] Das and Andersen[56] and more recently Noid and co-workers[6, 64] have suggested that, to predict the pressure correctly, CG models might incorporate additional volume-dependent terms in the interaction potential; by then sampling AA simulations under constant pressure conditions, where volume fluctuates, the CG volume-dependent energy term can be parametrized in a bottom-up fashion.

As a further investigation of thermodynamic properties of the liquid mixtures, we consider Kirkwood-Buff integrals. The theory of solutions pioneered by Kirkwood and Buff and later worked out in detail by Ben-Naim provides an important link between microscopic structure and macroscopic thermodynamic observables of solutions.[47, 172] The central elements of Kirkwood–Buff (KB) theory are integrals that depend on the RDFs as

$$G_{\alpha\beta}(R) = \int_0^R dr\ 4\pi r^2\ [g_{\alpha\beta}(r) - 1] \tag{3.3.5}$$

where $g_{\alpha\beta}(r)$ is the RDF between species α and β ($\alpha\beta$ = BB, WW, BW) and R is a correlation distance beyond which the RDF becomes flat and the integrand in Eq. (3.3.5) vanishes. The Kirkwood-Buff integral (KBI) is related to the excess coordination number around a particle relative to a flat density profile. Thus, a positive value of $G_{\alpha\beta}$ signifies a higher propensity of molecule type β around α (within a correlation radius of R), whereas a negative value indicates low intermolecular affinity. Solute-solute, solute-solvent, and solvent-solvent KBIs can be additively combined into a preferential solvation parameter

$$\Delta = G_{\mathrm{BB}} + G_{\mathrm{WW}} - 2\ G_{\mathrm{BW}} \tag{3.3.6}$$

which quantifies the mutual affinity between the solute and solvent and provides a gateway between microscopic structure and macroscopic thermodynamic properties through its connection to chemical potentials and activity coefficients. In particular, the composition derivatives of the benzene chemical potential (μ_{B}) and activity coefficient (γ_{B}) at temperature T and pressure P, for a solution with benzene and water concentrations (in M) ρ_{B}, ρ_{W} and mole fractions x_{B}, x_{W}, respectively, follow

$$\left(\frac{\partial \mu_{\mathrm{B}}}{\partial x_{\mathrm{B}}}\right)_{P,T} = \left[\beta x_{\mathrm{B}}\left(1 + \rho_{\mathrm{W}} x_{\mathrm{B}} \Delta_{\mathrm{BW}}\right)\right]^{-1} \tag{3.3.7}$$

and

$$\left(\frac{\partial \ln \gamma_{\mathrm{B}}}{\partial \ln x_{\mathrm{B}}}\right)_{P,T} = -\frac{\rho_{\mathrm{W}} x_{\mathrm{B}} \Delta_{\mathrm{BW}}}{1 + \rho_{\mathrm{W}} x_{\mathrm{B}} \Delta_{\mathrm{BW}}} \tag{3.3.8}$$

It should be noted that, for small values of r and/or small system size, the RDFs are insufficiently sampled at low interparticle distances, which can lead to convergence issues in KBIs and impart oscillatory character to these integrals when considered as a function of R. The convergence properties of KBIs and corrections to adapt them for small system sizes have been of interest in the recent literature and have been studied by van der Vegt and co-workers[173,174] and Schnell and co-workers.[175]

We evaluate the transferability of KBIs calculated using CG models developed from the 50% benzene solution, under extreme composition conditions at 0.2 and 99.8% benzene mole fraction that lie outside of the phase separation envelope. It should be noted that Eq. (3.3.5) is exact only for MD simulations in the grand canonical ensemble, and holds approximately for closed systems in a manner that becomes increasingly accurate with larger system sizes. Therefore, we carry out atomistic simulations with 5000 molecules and larger equilibrium cubic box dimensions ($\approx$ 53 Å for the 0.2% solution and 89 *angstrom* for the 99.8% solution).

Fig. **??** compares the B-B, W-W, and B-W types of KBIs as a function of the correlation length from among the CG models of Table 3.2.1 for $x_{\mathrm{B}} = 0.002$ (left panel) and $x_{\mathrm{B}} = 0.998$ (right panel). In the 0.2% solution, the BB KBI fails to converge at large R for the pair-only CG model, and when LD potentials are used, it converges on values that are different from the AA system (black line). For the WW and BW KBIs, CG models with only pair potentials and/or lacking the LD-WW potential also perform worse. The pair + LD-all force field comes closest to predicting all three KBIs, reproducing the converged values to within 41, 1.8, and 1.6% accuracy, for BB, WW, and BW, respectively. On the other hand, for the 99.8% solution, the BB KBI barely converges and is captured similarly by all of the CG models. Both the WW and BW KBIs do not converge for this case. KBI values for the AA, pair-only, and pair + LD-all models, averaged over the correlation length for the different CG force fields, are compared in Table 3.3.1

The BB KBIs for all of the CG models at 0.2% mole fraction have high positive values pointing toward higher benzene aggregation at very dilute benzene composition than predicted by the AA model. This is likely due to transferability errors stemming from the choice of the parametrization reference of 50% mole fraction, which is significantly more concentrated than the phase separation point of the system ($x_{\mathrm{B}} = 0.0095$, as predicted by the AA force field(55)). It should be kept in

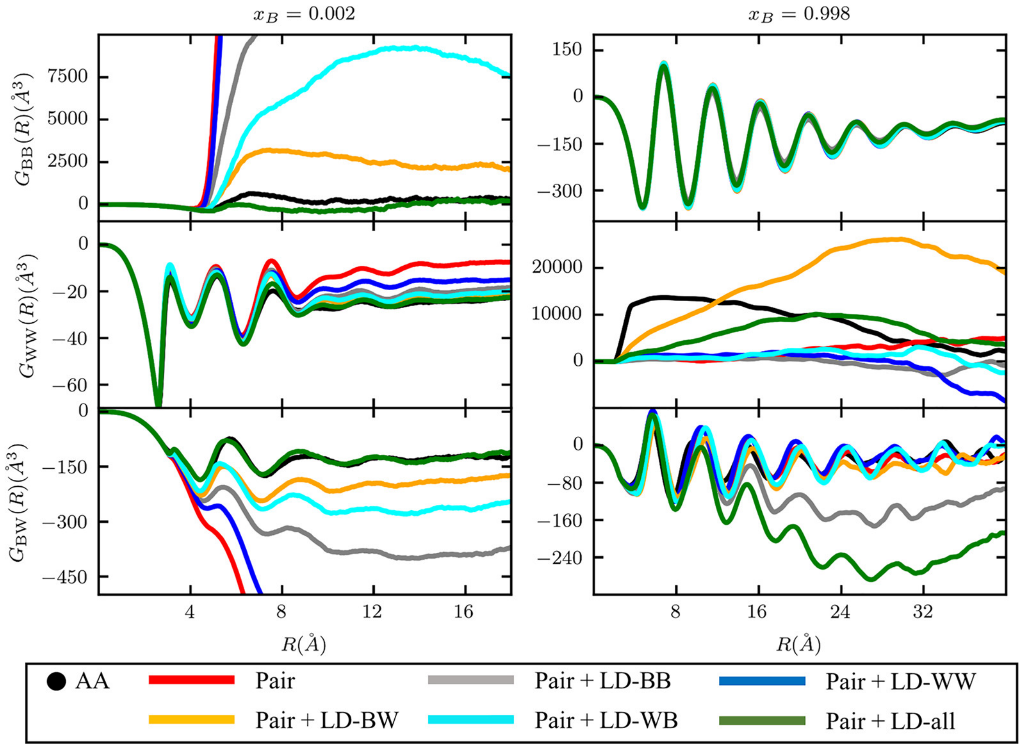


Figure 3.3.7: Comparison of the Kirkwood–Buff integrals (KBIs) of types B-B, W-W, and B-W (along the vertical subpanels) shown as a function of correlation length between AA and CG models for a very dilute solution at 0.2% mole fraction of benzene (left panel) and a superconcentrated solution at 99.8% (right panel). The CG models are parametrized from a 50% reference solution. All CG models have poor transferability for the BB KBI at dilute composition and for the WW KBI at very high composition. CG models with all LD potentials come closest to reproducing the WW and BW KBIs at low composition.

mind that converged KBIs calculated according to Eq. (3.3.5) are highly sensitive to small variations in the RDFs, especially over the correlation lengths that are chosen to report the average converged value. While Fig. reffig3.5 shows that the transferability errors for RDFs using local density-assisted force fields are low, it may be that such small variations are more difficult to capture quantitatively

Table 3.3.1: Comparison of the values of the KBIs (averaged over correlation lengths in the interval 14-18 Å for the 0.2% solution and 35-40 Å for the 99.8% solution) between the AA System, pair-only, and pair + LD-All CG models*

x_B	Force field type	$G_\mathrm{BB}(\mathrm{\AA}^3)$	$G_\mathrm{WW}(\mathrm{\AA}^3)$	$G_\mathrm{BW}(\mathrm{\AA}^3)$
0.002	AA	305 ± 66	-23.96 ± 0.02	-124.58 ± 0.85
0.002	pair-only	102100 ± 2400*	-8.12 ± 0.29	-1310 ± 25
0.002	pair + LD-all	179 ± 14	-23.55 ± 0.03	-122.5 ± 1.1
0.998	AA	-97.55 ± 0.17	2000 ± 1100*	-24.0 ± 3.3*
0.998	pair-only	-93.92 ± 0.17	-4600 ± 1200*	-32.3 ± 2.1*
0.998	pair + LD-all	-86.92 ± 0.26	4050 ± 620*	-209.8 ± 2.7*

*Values marked with an asterisk (*) have not converged.

through the approach used here, and additional constraints may be useful to help optimize CG models to correctly predict the KBIs.[176]

Arguably, this test of transferability is somewhat extreme because it attempts to bridge both sides of the phase transition. Previous studies of this system have sought the opposite route, i.e. constructing CG models at very dilute compositions and observing their transferability to extremely high concentrations.[4] Intra- and interspecies many-body effects are typically absent in dilute solutions, so perhaps mean-field approaches like the LD potential necessitate the use of a reference AA composition that is concentrated enough to demonstrate sufficient multiparticle interactions through macroscopic aggregation and/or consequent phase separation behavior.

3.4 Conclusion

In this chapter, we used local density (LD) potentials as a simple and computationally inexpensive way to develop transferable CG models of molecular liquid solutions and hence their equilibrium phase behavior. Specifically, we used relative entropy minimization to develop CG models of aqueous benzene solutions that map benzene (B) and water (W) molecules to single CG sites, thus coarse-graining away orientational degrees of freedom that heavily mediate the unique geometric and hydrogen bonded interactions governing liquid phase properties. We explored the ability of LD potentials to capture the multibody impact of these interactions and their mediation by the local solution environment. In combination with three CG pair potentials (BB, WW, BW), we tested the relative roles of the four possible LD potentials (BB, WW, BW, WB) by combining them zero, one, and all at a time, and we evaluated the sensitivity of the models to different reference all-atom system compositions spanning low to high benzene mole fractions. Here, the relative entropy minimization strategy provided a straightforward and simple way to parametrize both the conventional pair interactions and the more novel local-density potentials, all modeled by flexible splines.

Our results show that CG models built entirely out of pair potentials have limited transferability and poorly reproduce structural and thermodynamic properties at compositions far from the parametrization conditions. On the other hand, we find that CG models including LD potentials show improved structural transferability, in some cases dramatically so: including all four LD interactions allows quantitative reproduction of radial distribution functions and coordination number distributions across all of the composition space. Of the different LD potentials, the LD-BB interaction appears the least important to improving transferability, while the LD-WW potential has the strongest effect, suggesting that water-water interactions comprise the dominant multibody force in the system. This highlights the role of water's unique structural correlations and its tetrahedral hydrogen bonded network[39, 69, 177] in mediating its interactions with other water molecules as well as hydrophobic interactions with the much larger benzene molecule. Ultimately, the unlike LD potentials have an important role too in characterizing the B-W segregation under phase separated conditions, so that the CG force field including all four LD potentials is always the most transferable in terms of both structure and thermodynamics, regardless of the AA reference.

In terms of other thermodynamic properties, we also found that the LD strategy improved macroscopic simulations of phase separations, allowing individual phases to equilibrate at correct bulk compositions. However, even with the LD strategy, CG models produced more diffuse benzene-water interfaces than expected from all-atom results, likely due to the significant gradients in local composition that are not addressed by LD potentials. All CG models showed difficulty in capturing correct bulk pressures and interfacial tensions, likely connected to the difficulty in resolving the interfacial profile but also to the well-known problem of recapitulating pressure with bottom-up CG models.[22, 51, 62, 168–171] Moreover, we found that the LD strategy generally improved the ability of CG models to capture Kirkwood-Buff integrals at very dilute compositions, although no one model was able to completely reproduce the corresponding AA results.

This chapter illustrates a proof of principle for using LD potentials in bottom-up CG models capable of improving transferability and thus capturing phase separated systems. It is interesting to note that all models were developed from relatively small reference systems, yet translated well to larger-scale simulations. Indeed, here the small system size magnified the effect of interfacial and cross-species interactions, which provided sampling of local composition fluctuations important for model parametrization by relative entropy minimization. The determined forms of the local density interactions were also informative: the self-LD interactions (LD-BB, LD-WW) showed the most significant energy variations and impact. This may suggest the possibility of tabulating self LD interactions for different species independent of any particular mixture, such that these might be combined in arbitrary mixtures to create transferable force fields; we leave this idea for future directions. However, one surprising result was the significant variation in the form of the LD potentials determined at distinct reference compositions, even though each model still showed excellent transferability in structural properties across composition (particularly for the LD-all case). Perhaps even with LD potentials, the CG force field still admits some flexibility that could be adjusted to target reproduction of other properties, for example, those relevant to solvation thermodynamics.

In chapter 2, we used a single type of LD potential between CG monomers of a superhydrophobic polymer to construct implicit-water solvent models that led to a much-improved description of the conformational space sampled by the polymer.[144] Hydrophobic polymer collapse and benzene clustering in water both share common themes of higher-order cooperativity inherent to hydrophobic interactions, which can be modulated by the size asymmetry of the involved species.[149] It seems that, by accounting for local structure, even if simply in terms of coordination numbers without explicitly including orientational degrees of freedom, LD potentials can capture the delicate balance between entropy-driven hydrophobic forces around small solutes and enthalpy-driven hydrophobic interactions around large macromolecules and macroscopic interfaces. Thus, the corrective effect of the LD potential over traditional CG pair interaction may be broadly useful for improving a wide variety of CG models of fluid phase physics characterized by short ranged forces. In turn, the use of CG models may significantly enable the application of rigorous Monte Carlo phase equilibrium calculations[125, 128] that involve insertion and deletion moves. While the minimal set of LD potentials needed for good transferability in a CG model may be difficult to predict a priori, it is useful to note that water is a very common solvent and retaining its local structural correlations in CG models is arguably important for a large class of fluid mixtures.

Appendix

3.A Transferability of pair correlations and local density distributions using different CG models

Fig. 3.A.1-3.A.5 present the comparison between all-atom (AA, black dots) and CG radial distribution functions (RDFs, of types BB and WW) and local density (LD) distributions (of types BB, WW and WB) for the different CG forcefields in Table 3.2.1 in the section 3.2.1. CG models for various combination of these LD potentials are parameterized from reference compositions of 10%, 50% and 90% benzene mole fraction (dark framed plot, along the rows). CG models constructed at a particular atomistic reference are used in MD simulations of mixtures from 10% to 90% composition (along the columns), which are then used to calculate and compare RDFs and LD distributions with corresponding AA MD simulations at these compositions.

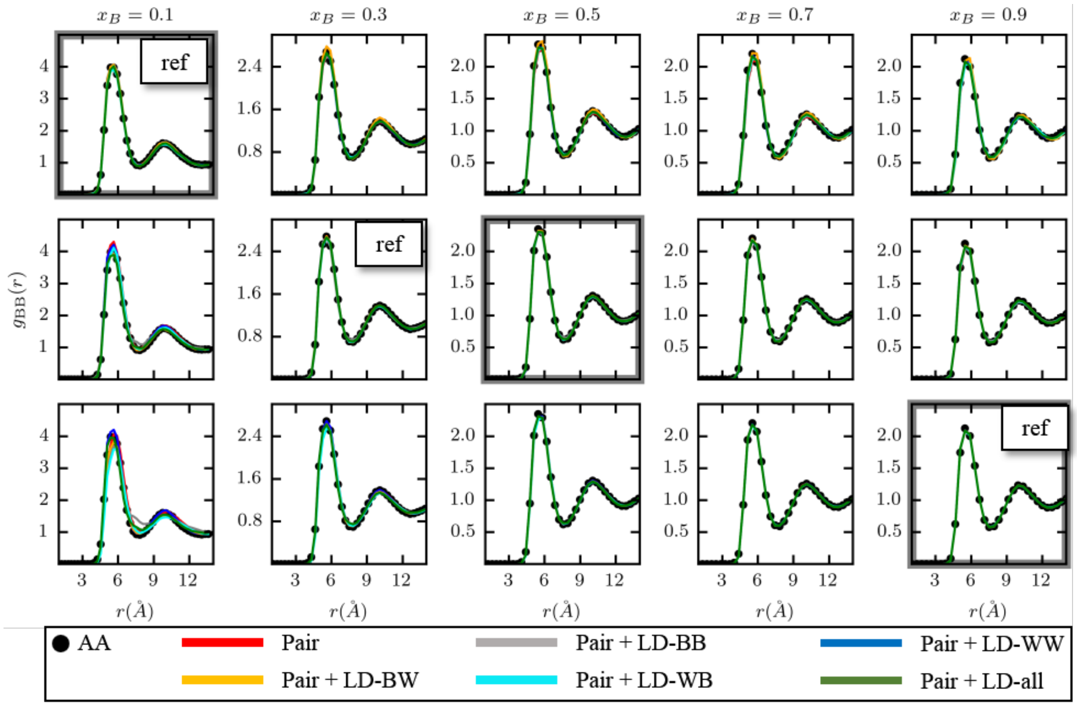


Figure 3.A.1: B-B RDFs

All CG forcefields, even the ones with only pair interactions, demonstrate excellent transferability for the BB pair correlations. This is consistent with the fact that the BB pair potential changes only slightly when re-parameterized in the presence of the LD-BB potential (Fig. 3.3.2).

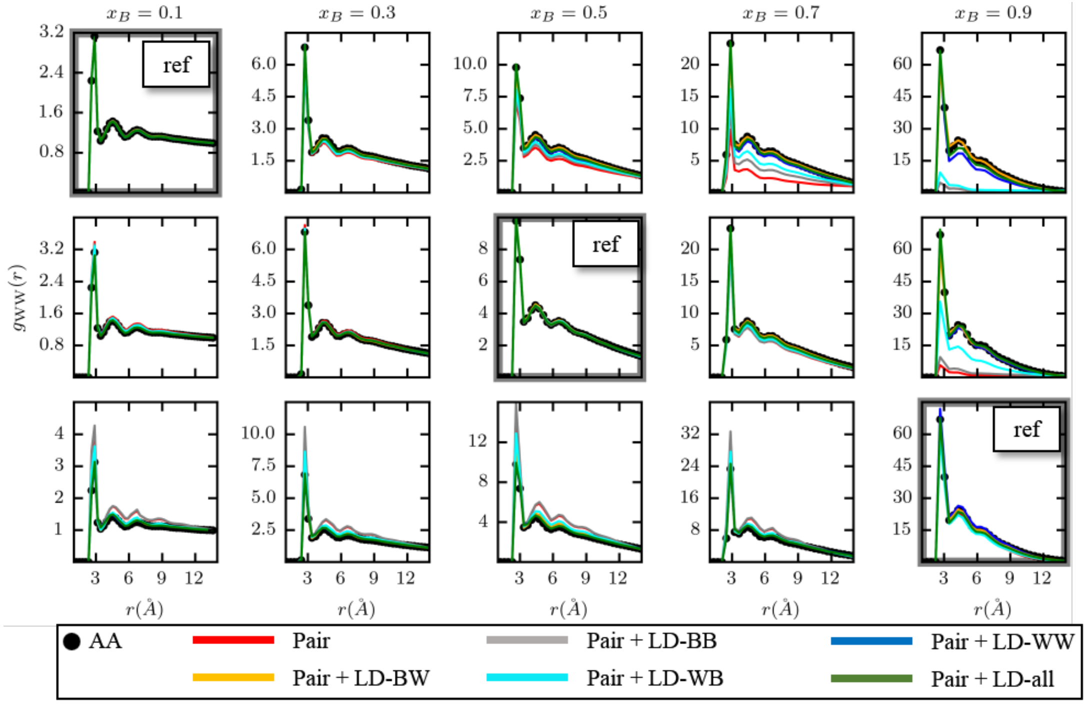


Figure 3.A.2: W-W RDFs

The W-W pair correlations have more pronounced changes in transferability when going from a CG model with only pair potentials to those with LD potentials. The small effect of the B-B many-body interactions is demonstrated by the poor transferability of the pair + LD-BB model, similar to that of the pair-only CG model. Both of these models have good representability at the reference composition but under-predict the significant water-water association at higher compositions. The unusually high magnitudes of the RDF peak at all compositions $\geq 30\%$ reflect the formation of tight water clusters, modulated by the overall bulk density of the system. The LD-WB potential also under-predicts the distribution peaks in dense solutions. The pair + LD-WW and pair + LD-all CG models have the best transferability, quantitatively reproducing the RDFs at all compositions.

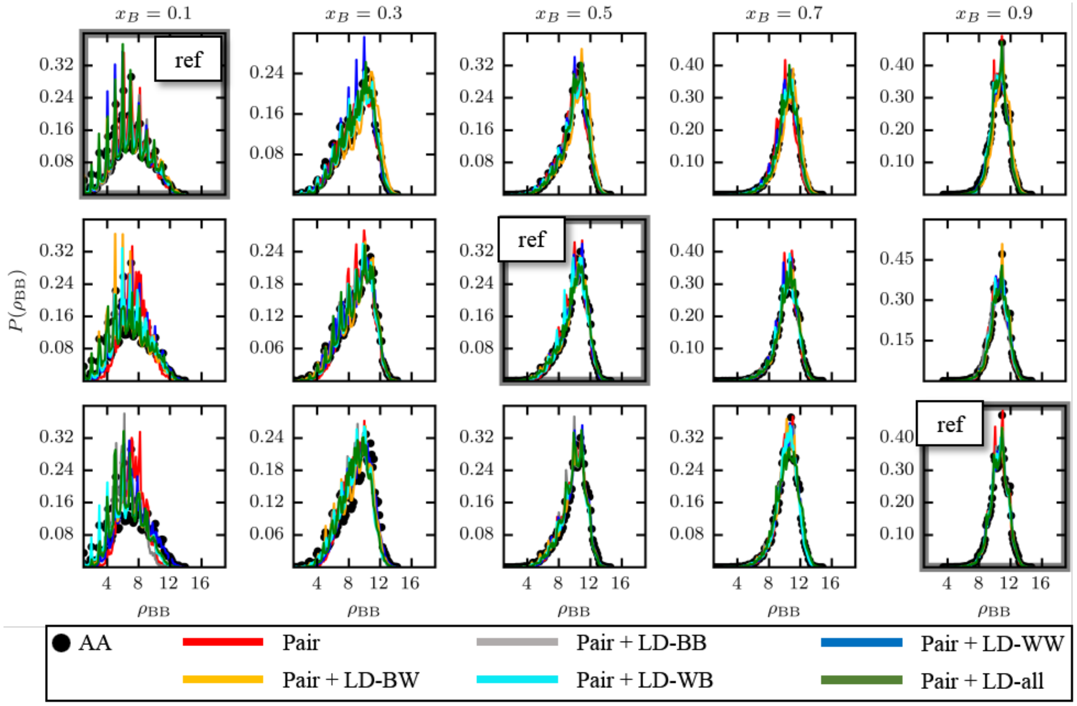


Figure 3.A.3: B-B LD distributions

Transferability for the B-B LD distributions follow similar trends to those for the corresponding pair correlations. However, there are some differences in representability. The pair-only CG model fails to represent the distribution even for the composition at which it was parameterized. All CG models that include one or more LD potentials have similar transferability, for each reference composition. The jagged peaks throughout the upper contour of the distribution at 10% are due to the discrete nature of the indicator function φ (Eq. (3.2.2)) used to determine the local density, as discussed in detail in Chapter 2. Higher benzene compositions add more counts to the LD BB histogram, thus further smoothening out these jagged peaks. The typical number of benzene neighbors around a CG benzene site grows from ~ 6 in the 10% solution to ~ 11 in the 90% solutions. From this and from the B-B RDFs in Fig. 3.A.1, we hypothesize that B-B pair structure in the more concentrated solutions ($\geq 30\%$) closely mimics that of a simple liquid. This might explain why this BB LD distribution or the BB RDF (Fig. 3.A.1) are reproduced reasonably across most of the composition space by the pair-only model alone without the need for multibody interactions to accurately represent the missing orientational degrees of freedom.

The W-W LD distribution is particularly sensitive to the presence of the LD-WW potential. For each parameterization reference, only two CG models, pair + LD-WW and pair + LD-all can reproduce the distribution across the entire range of compositions. All other CG models severely under-predict both the location

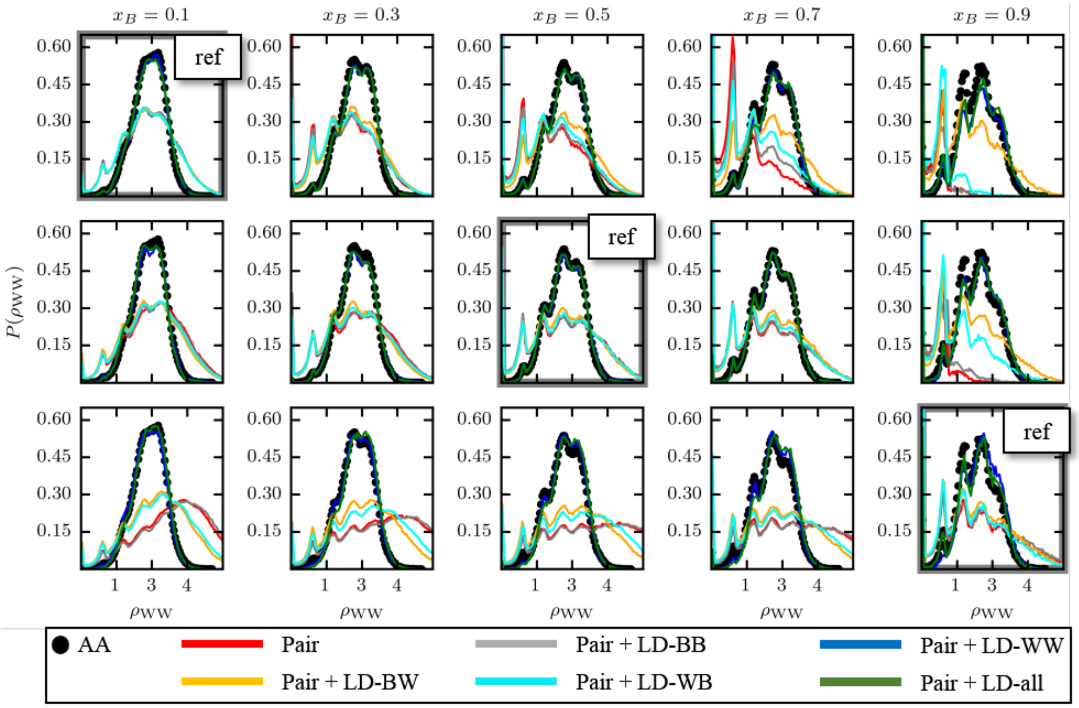


Figure 3.A.4: W-W LD distributions

of the peak and its spread. As discussed throughout this chapter, the LD-WW potential is arguably the most important LD potential to retain in CG models presented in this work. Also notice that the peak values are $\sim$ 1.5-2 times larger than the B-B LD distributions in Fig. 3.A.3, thus pointing to very tight water clusters, compared to benzene association.

Transferability of the W-B LD distribution is best for the pair + LD-all forcefield and reasonably good for the pair + LD-WW. The pair-only and pair + LD-BB forcefields produce distributions that scale very poorly across composition

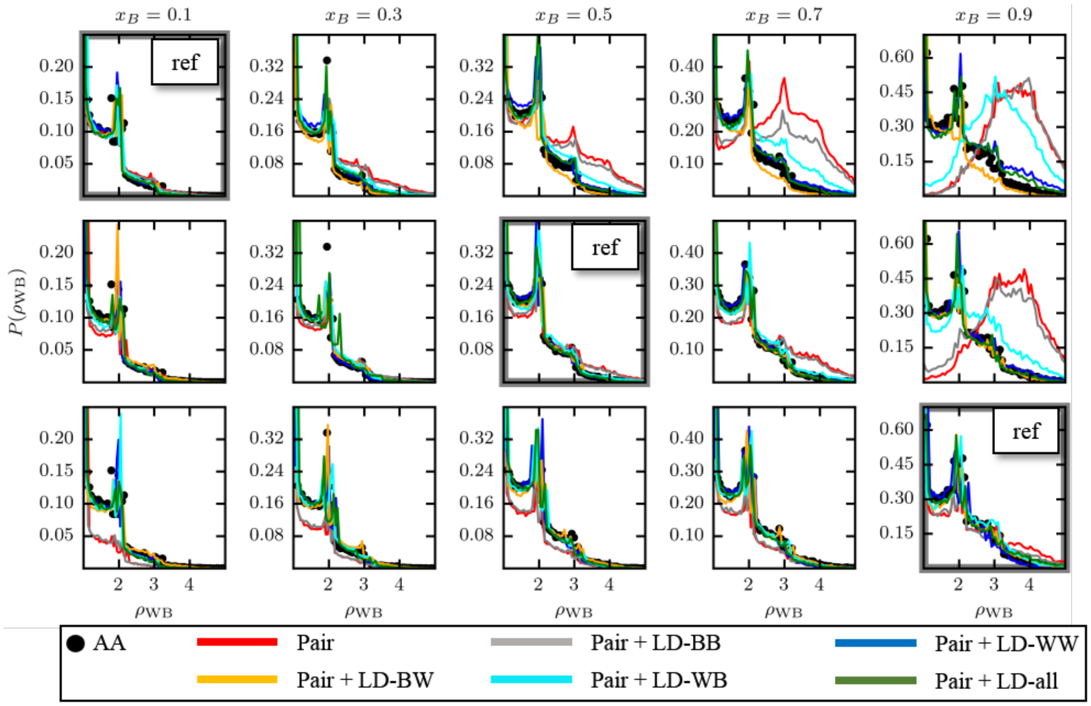


Figure 3.A.5: W-B LD distributions

space. The distributions are also reproduced moderately well by the forcefield that includes only the LD-BW potential, which may be intuitive. It is therefore interesting to see that the same does not hold for the pair + LD-WB model, which transfers poorly, significantly under predicting both the peak and the spread at high concentrations.

3.B RMS errors between AA and CG predictions of pair correlations and local density distributions using a relative deviation metric

Fig. 3.3.3 in section 3.3.2 employs an average RMS error between AA and CG pair correlations and LD distributions characterized by absolute deviations i.e. $\left\langle \Delta X_{\alpha\beta}^2 \right\rangle = \frac{1}{N} \sum \left(X_{\alpha\beta}^{\mathrm{AA}} - X_{\alpha\beta}^{\mathrm{CG}} \right)^2$, $X_{\alpha\beta} = g_{\alpha\beta}(r)$ or $P_{\mathrm{LD}}(\rho_{\alpha\beta})$ and N is the number of discrete histogram bins. Using absolute deviations makes this error metric sensitive to the range of the normalized histogram bin values, especially for the LD distributions. When RMS errors for different LD types (BB, WW, BW and WB) are added, models from 10% and 50% references produce a lower average error at compositions away from the reference, that can be erroneously interpreted as meaning increasingly higher transferability at state points further away from the reference. Further the average RMS errors for P_{LD} are non-zero at the reference compositions, where the AA and CG LD distributions are expected to overlap quantiatively,[35] especially for the pair + LD-WW and pair + LD-all forcefields (Figs. 3.3.2, 3.A.1-3.A.5).

To remove the sensitivity of the absolute RMS error to the histogram peak values, we construct an alternate RMS metric that uses relative mean square deviations normalized by the AA histogram value, namely

$$\begin{aligned} \mathrm{RMS}[g(r)] &= \Big[x_{\mathrm{B}}^2 \big\langle \Delta g_{\mathrm{BB}}^2 \big\rangle + x_{\mathrm{W}}^2 \big\langle \Delta g_{\mathrm{WW}}^2 \big\rangle + 2\; x_{\mathrm{B}} x_{\mathrm{W}} \big\langle \Delta g_{\mathrm{BW}}^2 \big\rangle \Big]^{\frac{1}{2}} \\ \mathrm{RMS}[P_{\mathrm{LD}}(\rho)] &= \Big[x_{\mathrm{B}}^2 \big\langle \Delta P_{\mathrm{LD,\ BB}}^2 \big\rangle + x_{\mathrm{W}}^2 \big\langle \Delta P_{\mathrm{LD,\ WW}}^2 \big\rangle \\ &\quad + x_{\mathrm{B}} x_{\mathrm{W}} \Big(\big\langle \Delta P_{\mathrm{LD,\ BW}}^2 + \big\rangle + \big\langle \Delta P_{\mathrm{LD,\ WB}}^2 + \big\rangle \Big) \Big]^{\frac{1}{2}} \end{aligned} \tag{3.B.1}$$

where

$$\big\langle \Delta X_{\alpha\beta}^2 \big\rangle = \frac{1}{N} \sum \left(\frac{X_{\alpha\beta}^{\mathrm{AA}} - X_{\alpha\beta}^{\mathrm{CG}}}{X_{\alpha\beta}^{\mathrm{AA}}} \right)^2 \tag{3.B.2}$$

and $X_{\alpha\beta} = g_{\alpha\beta}(r)$ or $P_{\mathrm{LD}}\big(\rho_{\alpha\beta}\big)$.

Fig 3.B.1 recapitulates the composition weighted RMS errors, but this time using the relative mean squared deviation in Eq. (3.B.2) above. The Y-axis ranges for the RDFs in panel (a) remain the same as in section 3.3.2, but the ranges of panel (b) increase significantly. Using relative deviations essentially magnifies the errors so that it is apparent at once that forcefields other than pair + LD-WW and pair + LD-all produce very large relative errors. For the LD distributions, these errors decrease by $\sim$ 50% across the range of compositions, regardless of the reference composition. But the pair + LD-WW and pair + LD-all models produce average errors that are 300% lower for the RDFs and 1400% for the LD

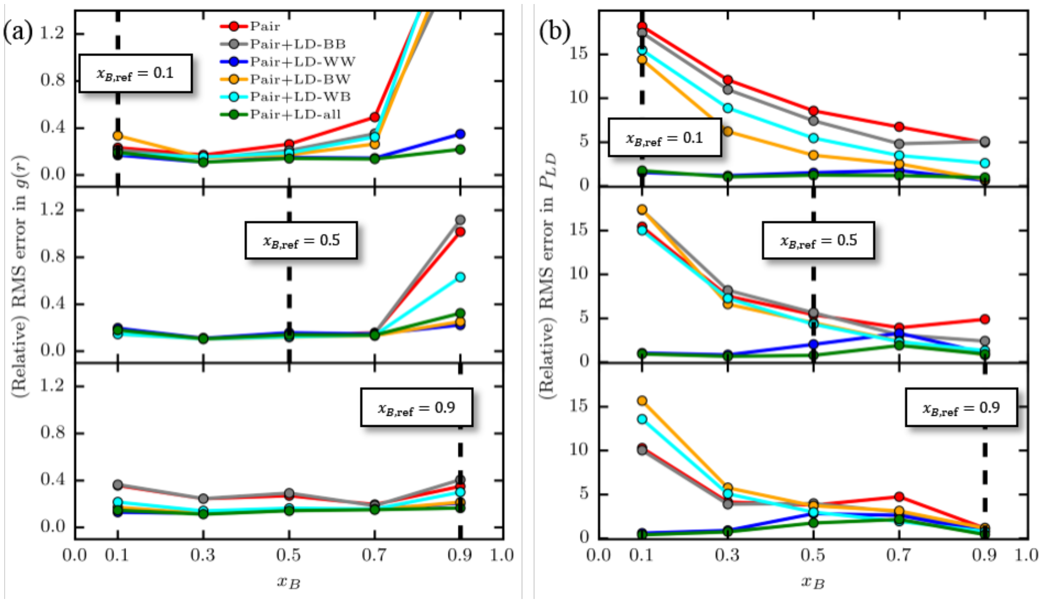


Figure 3.B.1: Composite-weighted relative RMSD overlap for (a) pair correlations and (b) local density distributions, between AA and CG models parameterized from 10%, 50% and 90% reference compositions (along the vertical subpanels). Choosing a relative RMS metric reduces the sensitivity of the errors to the peak values for different types of RDFs and LD distributions and brings out the superior performance of the pair + LD-WW and pair + LD-all forcefields

distributions and change by ∼ 50% (similar to Fig 3.3.3 in section 3.3.2) across the entire range of compositions.

3.C CG model transferability across different bulk densities

To test the transferability of the CG models across bulk density, we perform both AA and CG simulations of a 50% benzene mole fraction mixture using (cubic) boxes that are 10% smaller and larger than the original equilibrium dimension

($L_0 = 35.2\ 5$ Å). Although these simulations are in the NVT ensemble, they provide an idea of the model transferability across different bulk densities and thus also at different average pressures.

Fig. 3.C.1 presents a comparison of the BB, WW and BW radial distribution functions (RDFs) between the AA and CG models at these box sizes, for the 6 different forcefields outlined in Table 3.2.1 in section 3.2.1 parameterized from the 50% mixture reference. The relative RMS errors (RMSE) between the AA and CG RDFs are calculated according to $(RMSE)(X) = \left\langle \left(1 - \frac{X_{\mathrm{CG}}}{X_{\mathrm{AA}}}\right)^2 \right\rangle$, where $X = g_{\mathrm{BB}}$, g_{WW} or g_{BW}. All the CG forcefields perform better at lower bulk densities as evidenced by the lower RMSE for the larger box. Interestingly, the pair + LD-all model retains nearly the same low RMSE ($\sim$ 20-30%) for the BB and BW RDFs across the higher and lower densities but has a very high RMSE for the BW RDF at higher densities.

3.D Hard sphere excess chemical potentials for the different CG models

Fig. 3.D.1 compares the excess chemical potential due for inserting a hard sphere ($\mu^{ex}(R)$) for the AA system and different CG models, at 10%, 50% and 90% references. Fig. 3.3.4 in Section 3.3.3 presented fractional errors between AA and CG predictions of μ^{ex}, while Fig. 3.D.1 presents the absolute AA and CG excess chemical potentials side by side. For $R \geq 1.5$Å, the excess chemical potential decreases (the most apparent change is $\sim$ 2 kcal/mol for the 5 Å sphere) from low to high benzene composition, which is expected from how the hard sphere contact density is modulated by increasing the benzene fraction in the

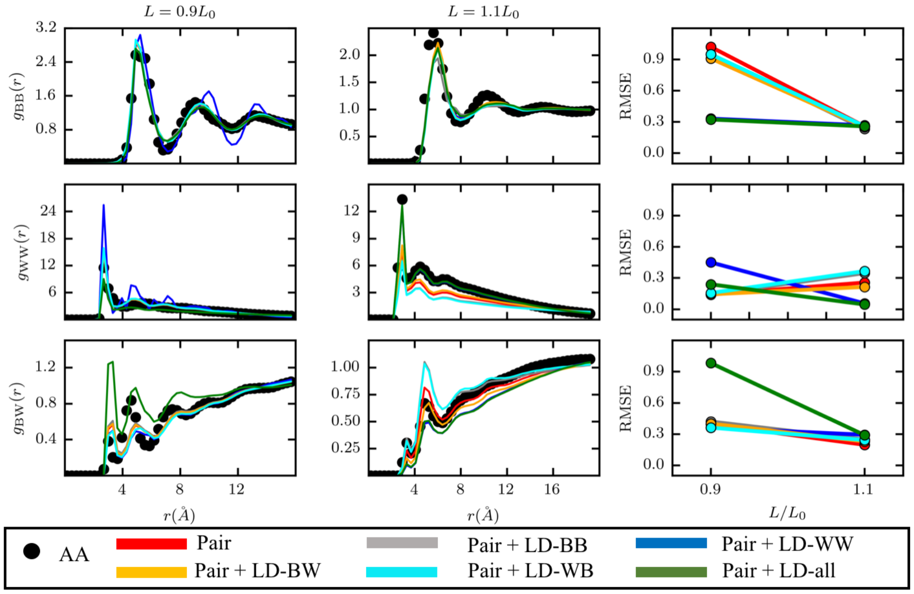


Figure 3.C.1: BB, WW and BW radial distribution functions for the 50% benzene mole fraction at box dimensions 10% smaller (column 1) and larger (column 2) from the one tuned to produce correct equilibrium bulk density of benzene and water, i.e. at different effective pressures. Column 1 and 2 visually compare the different RDFs (each row is a different RDF type) between the AA system and the different CG models of Table 3.2.1 in section 3.2.1, while column 3 quantifies this comparison through a relative RMS error metric. Most CG models seem to be more transferable for the larger box size, i.e. at lower bulk density

liquid mixtures, and is elaborated in the section 3.3.3 (Eq. (3.3.2)). The errors between AA and CG models are difficult to see in this representation and are better represented as fractional deviations presented in Fig. 3.3.4 in section 3.3.3.

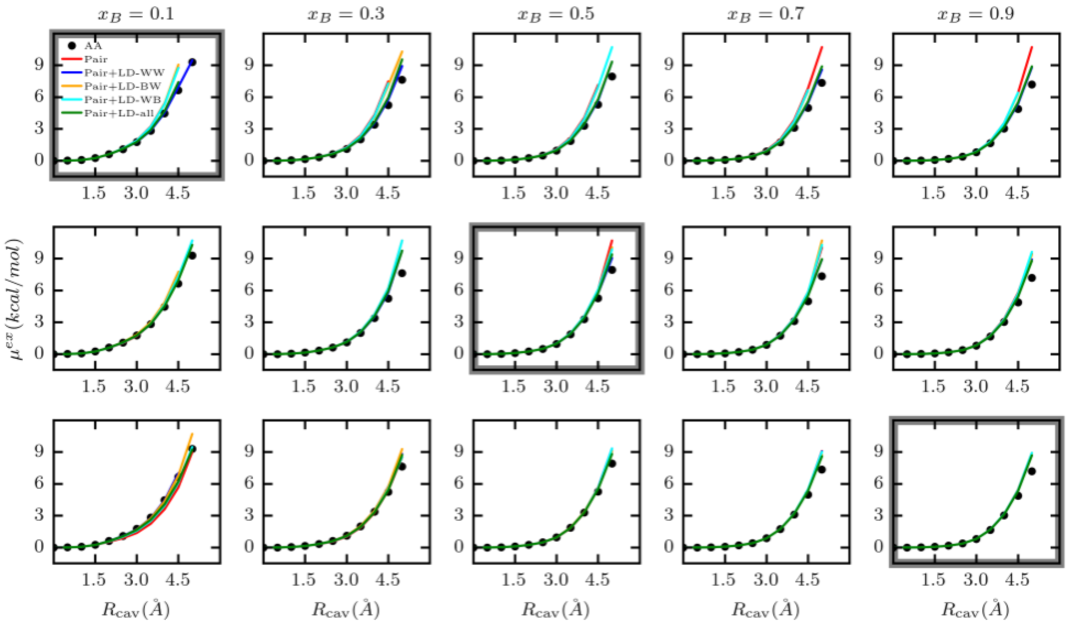


Figure 3.D.1: Excess chemical potentials for inserting a hard sphere with diameters ranging from 0- 5 Å, for the different CG models forcefields in Table 3.2.1 in section 3.2.1, parameterized from reference compositions of 10%, 50% and 90% benzene mole fractions (dark framed plot marked "ref" along the rows). AA values are marked with black dots.

Chapter 4

A bottom-up, structurally-accurate, Gō-like coarse-grained protein model

4.1 Introduction

It was established over half a century ago that folding of protein and other biomolecules form the foundations of structural biology by linking structure to function for molecular-level biological processes.[178,179] With some exceptions, the biological mechanism of a protein is determined by its three dimensional (3D) native structure which is encoded in the particular permutation of amino acid monomers that make up its sequence. Due to the remarkable progress in experimental techniques for protein structure determination over the last two decades, such as crystallographic methods, NMR spectroscopy, cryo-electron microscopy, and other nonlinear optical techniques,,[180–185] we now have an extensive database of more than 130,000 3D protein structures or their complexes.[186] Computational models of protein structure are not only essential for supporting experimental methods of solving native structures from amino acid sequences, they are also crucial for mechanistic investigations into the driving forces for folding and assembly, and in biotechnological applications such as high throughput drug-design.[14–16] Molecular Dynamics (MD) simulations of protein folding using atomistically detailed models can in principle address these applications. However, all-atom (AA) MD is typically limited to time scales of tens of microseconds, while the characteristic folding time for even small ($\sim$ 50 amino acid residues) proteins is on the order of millseconds and upwards. Except in very special cases, using custom-built hardware[187] or globally-distributed computing strategies,[188] so far it has been nearly impossible to break the millisecond barrier. This barrier has motivated the development of coarse-grained (CG) protein models which reduce the number and complexity of the degrees of freedom in MD simulations thus allowing a faster sampling of conformational space.

One of the principal aims of CG peptide models has been to study of the driving forces behind protein folding. The first CG peptide models were developed more than three decades ago by Warshel and Levitt, acheiving a resolution of 6.5 Å backbone RMSD for folding the bovine pancreatic trypsin inhibitor protein PTI,[189] and since then there has been considerable progress in their development.[190–193] Here we mention a few selected examples. Owing to the continued increase in computational resources, CG protein models have steadily become more sophisticated both in resolution as well as complexity of interactions. Early efforts like HP models focused on simple lattice chains described by a binary alphabet of amino acids: hydrophobic (H) and polar (P) to demonstrate that structure is uniquely modulated by sequence.[83,133] However, lattice frameworks restrict orientational degrees of the peptide chain, so that subsequently developed intermediate resolution models containing three or more CG backbone sites describing at least an α carbon and a CG carbonyl group, were much more successful.[34,194–203] E.g., The four-site PRIME CG model by Hall and co-workers was succesfully used to investigate large scale self-assembly behavior or polyalanine and polyglutamine oligomers.[194,204,205] A similar resolution CG model developed by Deserno and co-workers was used to predict structure and kinetics in three-helix bundles and transmembrane helical peptides.[196,206] The extremely popular MARTINI CG model developed by Marrink and co-workers, provides a large set of amino acid specific parameters that can be strung together to create forcefields for different peptide sequences.[132,198] MARTINI has been used extensively to probe protein-lipid interactions.[207]

Most of the approaches enumerated above are top-down in that they utilize experimental data or trajectory-averaged metrics from atomistic simulations to inform CG models. In contrast, bottom-up CG protein models systematically remove degrees of freedom from detailed atomistic simulations and utilize statistical mechanical principles to *inverse* design simpler potentials that renormalize the entire set of atomistic interactions such that the correct folding behavior (as seen in the atomistic system) is emergent in CG model simulations. A notable effort within this class of CG models is the Multi-Scale Coarse-Graining (MS-CG) method by Voth and co-workers, which parameterizes CG interactions by matching forces at CG sites between AA and CG resolutions.[25] The MS-CG approach was used to design four-site CG models of (helical) polyalanine and the (hairpin) sequence V_5PGV_5 which preserved their native states (as seen in the reference AA simulations) within 1 Å backbone RMSD in CG simulations.[28] Rudzinski and Noid used the MS-CG method to parameterize implicit water C-α models from atomistic polyalanine systems, and found that, while order parameters like the radius of gyration and helix propensity were reproduced accurately, the low CG resolution and a simple basis consisting of three types of nonbonded CG potentials were insufficient to capture helix-coil transitions seen in AA polyalanine.[138] Mullinax and Noid used the Extended-Ensemble-Coarse-Graining method[63] (detailed in section 4.2.1 and the appendix) to construct statistical potentials for a three-alphabet CG model (hydrophilic, hydrophobic and neutral CG sites) from a model databank populated with near native conformations of candidate α, β and mixed α/β sequences.[208] Peter and co-workers used iterative Boltzmann inversion to develop an intermediate resolution CG model for the short amphiphilic EALA peptide specifically to investigate its pH induced helix-coil transition.[209]

Despite many successes of both top-down and bottom-up CG peptide models in the literature, a main intent so far, especially for bottom-up models, has been to understand general features of the folding mechanism rather than accurate structure predictions. Although some of the top-down efforts have resulted in structurally robust models, they typically combine physics based forcefields with bioinformatic terms often derived from structure databases, and even then only typically resolve structures within 4-6 Å backbone RMSD as opposed to 1-2 Å that is generally considered correct.[195–197, 202] Thus, the ideal CG model is one that is (a) independent of bioinformatically obtained constraints, (b) bottom-up, and hence leverages the entire underlying folding free energy surface to automatically design CG interactions and, (c) provides high resolution structure prediction. Such a model can also enable the systematic inclusion of non-natural and synthetical chemical constructs for which experimental data would be either difficult or nearly impossible to generate.

This work explores the ability of a recent bottom-up coarse-graining strategy to develop a generic CG protein model that folds arbitrary protein structures to high accuracy at 1-2 Å resolution. Our investigation is motivated by the earlier effort of Carmichael and Shell that developed a bottom-up four-site CG model for polyalanine from a reference atomistic polyalanine simulation.[41] This earlier CG model demonstrated surprising accuracy and transferability, whereby both the free energy landscape and temperature-dependent folding behavior were quantitatively captured relative to atomistic simulations. Interstingly, though this model was developed from a mostly helical single polypeptide reference simulation, it also produced β-rich amyloid structures in self-assembly simulations, with correct sheet alignment, packing and twist. Here, we investigate whether it is possible to extend Carmichael and Shell's approach to capture native structures of arbitrary globular proteins. Ideally, one would seek a sequence-flexible model with distinct sidechain parameters for all twenty amino acids, but as a necessary first step towards that grand goal, here we consider the creation of native-structure-informed Gō type models in a bottom-up fashion. Thus, the CG model we develop includes favorable sidechain interactions between amino acids known to be in contact in the native structure. In this sense, this effort is a test of whether or not bottom-up methods can produce accurate backbone interactions and secondary structures even when "ideal" sidechain interactions are present, exploring their ability to capture wide range of protein folds. Similar to the Carmichael and Shell model, we develop CG polypeptide forcefields from reference AA polypeptide simulations, but subsequently parameterize sidechain potentials to reproduce native contact interactions. While a twenty-alphabet model with distinct sidechain parameters for all twenty amino acids is our eventual goal, we break this task into several incremental stages to validate and refine the strategy. In this chapter, we limit ourselves to the first stage which involves native-sentient Gō-like models.

Long studied in the protein folding community since their introduction,[210] Gō models modulate sidechain interactions between residues depending on whether or not they are in contact (i.e., within a threshold proximity) in the experimental native structure; contact residues are given an attractive interaction while all other non-native residue pairs merely experience excluded volume interactions. Such models are motivated by the principle of minimum frustration, which postulates that proteins were evolved towards sequences with the ability to fold into low-energy conformations, while actively selecting against local minima traps on the energy landscape by avoiding misfolding and non-native contacts.[211] Gō models take this idea a step further by treating native contacts as the sole driving force for the folding process. Although such models are not *de-novo*, (i.e., they cannot be used to predict structure from sequence alone), they have proven useful for understanding functional protein conformational fluctuations and generic aspects of folding pathways, and may also rapidly provide structural ensembles consistent with experimentally-determined native contact information (e.g. via NMR) to improve structure prediction efforts.[212–216] Over the last decade, several flavors of Gō-like and other native-structure based CG models have been proposed which utilize various physical and bioinformatic data like the distribution of hydrogen-bonds, homologous sequences or solvation transfer free energies of amino acids, in addition to the native structure.[217–220] In contrast, our approach here does not include any experimental data beyond the native structure. Specifically, we parameterize both generic (i.e., amino acid unspecific) backbone interactions and Gō-like native interactions in a bottom-up fashion directly from small-scale atomistic simulations. The Gō-like interactions pursued here are meant to serve as ideal sidechain potentials that enforce a minimal set of restraints (native contacts) on the folding process that bring out the intrinsic capability of the backbone in sampling the correct conformational space.

Our premise is based on the observation that backbone interactions are key to the landscape of folds: even simple sequences that do not fold to unique structures (e.g., homopolymeric peptides) can explore a wide range of realistic folds, involving both α and β secondary structure elements. An early study by Head-Gordon and Stillinger et al. found that polyalanine can mimic the secondary structures of several globular proteins such as PTI, crambin, ribonuclease A and superoxide dismutase. Specifically, the authors observed that the potential energy landscapes of alanine polymers of different lengths have local minima close to the global free energy minima at the native structures of different proteins of equivalent sequence length.[221] More recent work by Laio and co-workers found that an atomistic 60-mer of polyvaline was able to produce practically *all* compact folds ($\sim$ 300) observed in nature for proteins of comparable length, including α, β and mixed-content structures.[222] A similar conclusion was reached through CG polyvaline simulations by Vendruscolo and co-workers, where they recovered 135 folds from the ensemble of structures generated by the valine 60-mer, out of the 265 folds reported in the CATH database[223] for proteins containing between 40 and 75 residues.[224] These studies suggest a picture whereby backbone interactions such as hydrogen bonding, dihedral energetics and excluded volume interactions, together with the inherent geometrical constraints of the backbone, define the conformational landscape for a protein chain, while the side chains then adjust the energetic favorability of different structures so as to select a single fold.[225–228] Thus a fundamental challenge that we test in this chapter is the ability of a bottom-up strategy to generate computationally fast, generic CG backbone interactions that correctly reproduce the structural landscape. Gō-like interactions provide a natural context to establish an upper bound on the quality of CG backbones developed.

4.2 Methods

We use a four-site CG model motivated by the Carmichael and Shell[41] four-site polyalanine model and by early work by Carmichael and Shell which showed that a higher resolution polyalanine CG model with three CG sites per residue was more representable and transferable among other one- and two-site variants.[34,229] This is also consistent with earlier four-site model development efforts by the Hall and[194] and Deserno groups.[196] As shown in Fig. 4.2.1(A) with an example polyleucine 15-mer, the CG mapping ignores hydrogens and discretizes each amino-acid into CG sites corresponding to four heavy atom centers: nitrogen (N), α-carbon (C), carbonyl carbon and oxygen (O) and sidechain (S). Admittedly, larger amino acids, especially those with one or more aromatic rings (HIS, PHE, TYR, TRP, etc.) may need a higher sidechain resolution, but we shall not pursue that here and leave it for a future endeavor.

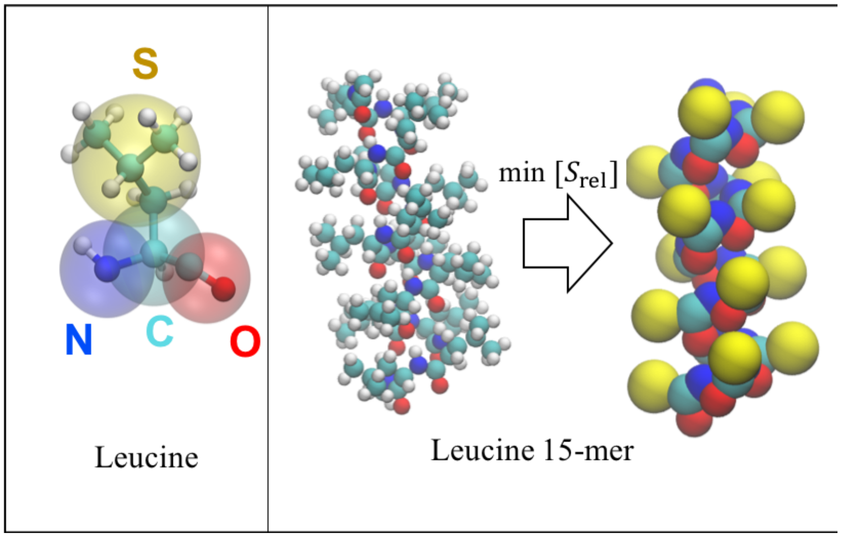


Figure 4.2.1: Left: leucine is mapped to four heavy atom centers N, C, O, S that sit at the centers-of-mass of the amino, α-carbon, carbonyl carbon and oxygen, and the side-chain groups respectively. Right: The CG model of a 15-mer leucine polypeptide is parameterized by minimizing the relative entropy from a reference atomistic simulation.

In this chapter, we develop CG peptide models by combining CG backbone forcefields with Gō-like interactions. Backbone forcefields and Gō potentials are both developed in a bottom-up manner from atomistic reference simulations of candidate peptides, by using the relative entropy minimization method.[32] The relative entropy between an AA system and its corresponding CG model is an information theoretic measure that encodes the overlap between AA and CG microstate probability distributions.[32] Minimizing the relative entropy with respect to forcefield parameters of the CG model guarantees a maximal overlap of AA and CG microstate probability distributions, such that the CG model may be able to recapitulate important thermophysical properties (such as secondary structure) of the atomistic system.[35] This technique requires atomistic trajectory data to calculate the relative entropy between an AA and a proposed CG model, and iteratively minimizes it by searching for the zeroes of its derivative in the space of CG forcefield parameters. The relative entropy between an AA polypeptide reference and its corresponding CG model is given by:

$$S_{\mathrm{rel}} = \int p_{\mathrm{AA}}(\mathbf{r}) \; \ln\left(\frac{p_{\mathrm{AA}}(\mathbf{r})}{p_{\mathrm{CG}}(\mathbf{M}(\mathbf{r}))}\right) d\mathbf{r} \, + \, S_{\mathrm{map}} \tag{4.2.1}$$

where, $\mathbf{r}$ and $\mathbf{R}$ represent AA and CG configurations respectively, and $\mathbf{M}(\mathbf{r})$ is a mapping operator (typically a matrix) that replaces groups of atoms in the AA representation with center-of-mass sites in the CG model. p_{X} gives the equilibrium conformation probabilities for the ensembles X = AA or CG. The integral in Eq. (4.2.2) proceeds over all AA conformations, although it is possible to cast it in terms of the CG conformations.[102] S_{map} is a "mapping entropy" that accounts for the degeneracy in the AA → CG mapping, i.e., it measures the number of distinct AA configurations that map to the same CG one. However, it is independent of the CG forcefield, and because we fix the CG resolution to four sites per amino acid, S_{map} can be treated as a constant that plays no role in this work. In the canonical ensemble, the derivative of the relative entropy for the CG polypeptide model with respect to the CG forcefield parameters $\boldsymbol{\lambda}$ can be written as:

$$\frac{\partial S_{\text{rel}}}{\partial \boldsymbol{\lambda}} = \beta \left[\left\langle \frac{\partial U_{\text{CG}}^{\text{poly}}}{\partial \boldsymbol{\lambda}} \right\rangle_{\text{AA}} - \left\langle \frac{\partial U_{\text{CG}}^{\text{poly}}}{\partial \boldsymbol{\lambda}} \right\rangle_{\text{CG}} \right] \tag{4.2.2}$$

where $U_{\text{CG}}^{\text{poly}}$ is given by Eq. (4.2.3), β represents the inverse temperature $(1/k_BT)$ and $\langle\,\rangle_{\text{X}}$ represents averaging over all conformations in the model X = AA or CG. The parameterization of the CG model proceeds by minimizing S_{rel}, i.e., iteratively solving the equation $\partial S_{\text{rel}}/\partial \boldsymbol{\lambda} = 0$. In practice, we use a combination of perturbation theory and conjugate gradient minimization to preform the minimization efficienty as described in Ref. 34

4.2.1 CG backbone forcefields

Similar to Carmichael and Shell's previous work,[41] we parameterize CG models from atomistic polypeptides and extract "backbone" forcefields from these models as the sum of intra-backbone and backbone-sidechain interactions. CG polypeptide forcefields ($U_{\text{CG}}^{\text{poly}}$) are represented using bonded (U_{b}), angular (U_θ), torsional ($U_{\varphi,\psi}$) and non-bonded pair potentials (U_{pair}) which are applied to intra-backbone (BB), backbone-sidechain (BS) and inter-sidechain (SS) CG sites. The bond potentials are harmonic while all other potentials are represented as cubic B-splines whose knot points are optimizable parameters of the CG model. The non-bonded pair potentials are cut off at 10 Å. Angle, torsion and pair potentials use 40 knot points each with densities of 1.4 °/knot, 0.22 °/knot and 0.25 Å/knot respectively. The full CG polypeptide model for a reference AA peptide is constructed through relative entropy minimization: subsequently the *backbone* forcefield ($U_{\mathrm{CG}}^{\mathrm{BB}}$) for the CG model simply amounts to taking all CG interactions except for the inter-sidechain pair potentials.

$$\begin{aligned} U_{\mathrm{CG}}^{\mathrm{poly}} &= \left(U_{\mathrm{b}}^{\mathrm{BB}} + U_{\theta}^{\mathrm{BB}} + U_{\varphi,\psi}^{\mathrm{BB}} + U_{\mathrm{pair}}^{\mathrm{BB}}\right) + \left(U_{\mathrm{b}}^{\mathrm{BS}} + U_{\theta}^{\mathrm{BS}} + U_{\varphi,\psi}^{\mathrm{BS}} + U_{\mathrm{pair}}^{\mathrm{BS}}\right) + uU\mathrm{pair}^{\mathrm{SS}} \\ U_{\mathrm{CG}}^{\mathrm{BB}} &= \left(U_{\mathrm{b}}^{\mathrm{BB}} + U_{\theta}^{\mathrm{BB}} + U_{\varphi,\psi}^{\mathrm{BB}} + U_{\mathrm{pair}}^{\mathrm{BB}}\right) + \left(U_{\mathrm{b}}^{\mathrm{BS}} + U_{\theta}^{\mathrm{BS}} + U_{\varphi,\psi}^{\mathrm{BS}} + U_{\mathrm{pair}}^{\mathrm{BS}}\right) \end{aligned} \tag{4.2.3}$$

An important design criterion for a *transferable* CG peptide backbone forcefield is an appropriate balance between α helical, β-sheet and extended conformations, such additional protein- or sequence-specific sidechain interactions will "steer" the peptide chain towards the correct fold. As mentioned in section 4.1, proteins achieve this balance through geometrical constraints, steric repulsions, hydrogen bonds, electrostatic interactions and hydrophobic effects.[225–228] CG backbones developed in this effort do not encode such interactions explicitly, but instead, critically depend on the sampling of conformation space by the reference AA polypeptides to "learn" effective coarse interactions within functional form of the potentials described in Eq. (4.2.3). Further, to retain computational efficiency, we do not incorporate either explicit solvent or explicit intra-peptide and peptide-solvent electrostatics in our CG models. Thus, we focus on moderate-length polypeptides generated from hydrophobic amino acids alone, as putative AA references.

A critical choice that influences the secondary structure balance is the reference AA system from which to develop the CG backbone model. We examine four such references to explore and identify cases with a good balance of α, β and flexible conformational propensities, as summarized in Table 4.2.1. It is non-trivial to quantify such a balance a-priori and ultimately we characterize the success of a reference in terms of its later performance in folding known structures. While an equal sampling of primary dihedral conformations in a Ramachandran plot[230] may seem attractive, it is not clear that this is realistic or necessary for correct folding. Thus, in a bottom-up CG approach, relevant secondary structure information necessarily comes from the choice of AA reference. We first develop CG backbones from two homogeneous polypeptides which capture either α and β conformations, and subsequently discuss two strategies to design hybrid references that encode both these secondary structure elements. In all cases, we parameterize the CG models at the folding temperatures of the AA references to sample both folded and extended conformations.

Among the hydrophobic amino acids, alanine has been suggested as a useful reference for building polymers that can approximate heteropeptide chains, due to its small size and the ability of alanine polymers to mimic the correct tertiary structure of several proteins of equivalent length.[221] However, our preliminary tests (not reported here) revealed that a polyalanine 15-mer is not sufficiently stable in either of α and β folds. It has a large degree of unstructured conformations and for a variety of force fields (including implicit and explicit solvent atomistic simulations), less than 4% exist in fully α-helical states at physiological temperatures. Gō-like models later built out of these polyalanine-derived backbone forcefields were unable to fold small mini-proteins and did not well capture α and β structure.

Instead we use polyleucine and polyvaline peptides as our first two refernce AA simulations. Both have sidechains that are larger than alanine but still smaller than other hydrophobic amino acids and devoid of aromatic groups. These systems provide excellent contrasts because leucine has a reasonably high helix propensity,[231] while valine has very high stability for β-sheets.[232] We parameterize CG models from atomistic polyleucine (LEU_{15}, Fig 4.2.1) and from polyvaline (VAL_{15}) 15-mers, at their folding temperatures, using relative-entropy-minimization. We expect the LEU_{15} CG model to preferentially stabilize only α helices and the VAL_{15} model to be more biased towards β conformations.

In addition, we examine two additional reference cases that are designed to hybridize secondary structure preferences from the leucine and valine systems. The first strategy investigates AA leucine-valine copolymers. We tested four randomly generated 15-mer sequences with $\sim$ 30-60% Leucine (L) and $\sim$ 70-40% valine (V) content and found that the sequence LVVVVVVVLLLVVLL (LEU_6VAL_9) has the most balanced helical (cluster fraction $\sim$ 36% at 270 K) and hairpin (cluster fraction $\sim$ 42% at 270 K) conformations in AA simulations. Clearly there may be other sequence combinations that may produce even closer α and β populations, but we proceed with this particular case. The second hybrid strategy and the fourth reference AA case, involves directly combining information from multiple atomistic reference simulations into a single CG model. This approachs is motivated by the extended-ensemble coarse-graining concept of Mullinax and Noid,[63] in which a CG model can be parameterized simultaneously against an ensemble of AA references. While originally developed in the context of force-matching, we adapt that approach to the relative entropy framework. Here, an extended ensemble can be formulated as follows: the optimal CG model that minimizes information loss simultaneously from N AA references also minimizes the sum of *individual* relative entropies between the model and each AA member of the ensemble, i.e.,

$$S_{\text{rel}} = \sum_{i}^{N} S_{\text{rel}}^{(i)} \tag{4.2.4}$$

where $S_{\text{rel}}^{(i)}$ is the relative entropy between the CG model and the i^{th} AA reference. Motivation for the form of Eq. (4.2.4) can be found in the appendix. The extended-ensemble strategy was originally introduced for developing models that are independent of the thermodynamic state point at parameterization and hence transferable across a range of states,[63] but more generally, the ensemble AA members can also be chemically distinct,[63, 233] such as different peptide sequences.[208] In this chapter, we leverage the extended-ensemble relative entropy method by parameterizing a hybrid CG polypeptide model ($\text{LEU}_{15} + \text{VAL}_{15}$) simultaneously from the polyleucine and polyvaline AA simulations. It is unclear how an overall folding temperature should be defined for the extended AA ensemble, so we maintain the hybrid CG model at room temperature by parameterizing it now from polyleucine and polyvaline references at room temperature . Table 4.2.1 outlines the four different CG backbone forcefields studied in this work.

Table 4.2.1: Nomenclature for different CG peptide backbone forcefields*

Forcefield	Reference sequence/(s)	Parameterization temperature
LEU_{15}	L_{15}	407 K (folding temperature)
VAL_{15}	V_{15}	367 K (folding temperature)
$\mathrm{LEU}_6\mathrm{VAL}_9$	LVVVVVVVLLLVVLL	360 K (folding temperature)
$\mathrm{LEU}_{15} + \mathrm{VAL}_{15}$	L_{15}, V_{15}	300 K (room temperature)

*Gō models derived from these backbone forcefields are also named similarly.

4.2.2 CG sidechain forcefields (Gō models)

The final CG strategy pursued here utilizes a Gō model for the sidechains, which requires the native structure of the target protein sequence as input. Residues that are in contact (i.e., within a certain threshold proximity) in the native structure are supplied attractive (or "native") interactions while others are provided with excluded volume (or "nonnative") interactions. In this work a cutoff of 8 Å between the residue centers-of-mass is used to determine contacts. Residues within a contact order of 3 along the sequence are considered chain-adjacent and native potentials are not applied to them. The non-bonded native interaction $\left(u_{\text{pair}}^{\text{SS, native}}\right)$ is represented using a spline (with a cutoff of 10 Å). Non-bonded non-native interactions $\left(u_{\text{WCA}}^{\text{SS, non-native}}\right)$ are represented using a constant Weeks-Chandler-Andersen (WCA)[104] potential with $\epsilon = 4.2\ k_B T$ and $\sigma = 3.8$ Å, which mimics the inner repulsive core region of the optimized native potentials (see section 4.3.2). The overall CG forcefield is given by:

$$U_{\mathrm{CG}} = \underbrace{\left(U_{\mathrm{b}}^{\mathrm{BB}} + U_{\theta}^{\mathrm{BB}} + U_{\varphi,\psi}^{\mathrm{BB}} + U_{\mathrm{pair}}^{\mathrm{BB}}\right) + \left(U_{\mathrm{b}}^{\mathrm{BS}} + U_{\theta}^{\mathrm{BS}} + U_{\varphi,\psi}^{\mathrm{BS}} + U_{\mathrm{pair}}^{\mathrm{BS}}\right)}_{\text{min. } \mathrm{S}_{\mathrm{rel}} \text{ from AA polypeptide}} \\ + \underbrace{U_{\mathrm{pair}}^{\mathrm{SS,\ native}}}_{\text{min. } \mathrm{S}_{\mathrm{rel}} \text{ from AA peptide}} + U_{\mathrm{WCA}}^{\mathrm{SS,\ non\text{-}native}} \tag{4.2.5}$$

We parameterize the native interactions by minimizing the relative entropy between an AA simulation of a candidate peptide and the CG model represented by the forcefield U_{CG} in Eq. (4.2.5). Note that this constitutes a *second* round of relative entropy optimization, post CG backbone parameterization. In this case, the backbone potentials are held fixed at their optimized values obtained from the reference AA polypeptide, and only the spline knots in $U_{\mathrm{pair}}^{\mathrm{SS,\ native}}$ are allowed to vary. Fig. 4.2.2 provides a flowchart for the overall two-step parameterization of the CG forcefields.

Specifically, we use an AA trajectory of the 20-residue trp-cage miniprotein (PDB code: 1L2Y) to parameterize the native interactions. 1L2Y is a synthetic protein that folds rapidly to a globular structure representative of larger proteins. Its small size (the smallest protein like construct) and fast folding times make it an ideal model system for experimental and computational studies of protein folding mechanisms.[234] 1L2Y folds extremely fast (~ 4.1 μs in-vitro[235]) such that short nanosecond implicit solvent MD trajectories centered around the native structure

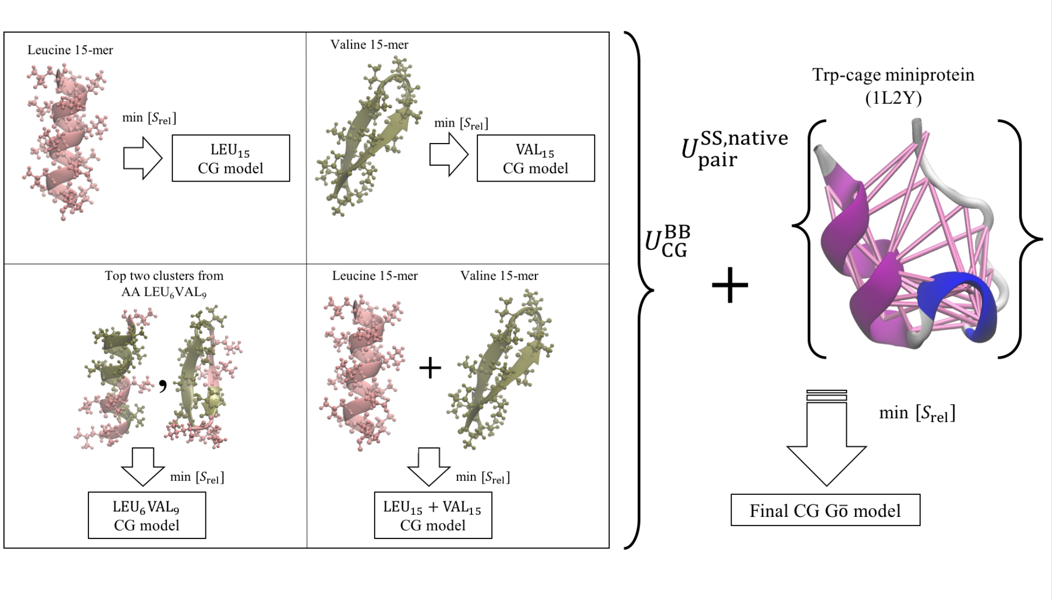

Leucine 15-mer
min [S_{rel}]
LEU_{15} CG model
Valine 15-mer
min [S_{rel}]
VAL_{15} CG model
Top two clusters from AA $\mathrm{LEU}_6\mathrm{VAL}_9$
min [S_{rel}]
$\mathrm{LEU}_6\mathrm{VAL}_9$ CG model
Leucine 15-mer
Valine 15-mer
min [S_{rel}]
LEU_{15} + VAL_{15} CG model
$U_{\mathrm{CG}}^{\mathrm{BB}}$
+
$U_{\mathrm{pair}}^{\mathrm{SS,native}}$
Trp-cage miniprotein (1L2Y)
min [S_{rel}]
Final CG Gō model

Figure 4.2.2: CG backbone potentials ($U_{\mathrm{CG}}^{\mathrm{BB}}$) are first extracted from CG polypeptide models parameterized by minimizing relative entropy from AA polyleucine (red), polyvaline (dark yellow) and a mixed leucine-valine sequence LVVVVVVVLLLVVLL. A Hybrid CG backbone embedding both α helical (from polyleucine) and β-sheet (from polyvaline) characteristics is also parameterized by simultaneously minimizing the total relative entropy from both AA leucine and valine polymer references. Gō models are developed for each of these backbone forcefields, by parameterizing native interactions $U_{\mathrm{pair}}^{\mathrm{SS,\ native}}$ through a *second round* of relative entropy minimization with a AA simulation of the trp-cage miniprotein (1L2Y). Lines across 1L2Y connect the native contacts.

are sufficient for calculating the relative entropy during the iterative optimization procedure. The trp-cage protein has well defined helical and flexible regions but no β sheets. Thus, in our Gō model, the task of capturing β-sheets is largely shouldered by the CG backbone forcefield, consistent with our hypothesis that the backbone interactions define the a landscape of putative folds. Alternatively, one might develop native interactions from either a more $\alpha - \beta$ balanced protein (e.g. protein G) or using an extended ensemble of two (or more) short peptides that selectively stabilize α and β conformations, but due to added computational complexity, we leave these directions for future work. It is instructive to note that, beyond the native structure, we do not incorporate any other bioinformatic-type information in the Gō model. This contrasts with other Gō models in the literature which include e.g., additional hydrogen bond potential functions and parameterize them using secondary structure databases or fragment libraries.[197,212,214]

In the developed CG models, glycine must be treated specially because it lacks a sidechain site for native-contact interactions. We tested a preliminary Gō model (not reported here) that omitted native-contact interactions for glycine, and found poor folding performance for 1L2Y ($\geq$ 5 Å) in the flexible regions. Instead, we supply glycine residues in the CG model with a pseudo sidechain that permits native contact interactions with other amino acids. Such a treatment is reasonable since hydrogen bond lengths (between glycine and other residues along the backbone) are typically less than the 8 Å cutoff used to determine native contacts. Glycine has two equivalent hydrogens on its α carbon, either of which can be mapped to the CG pseudo sidechain. We select the prochiral-L hydrogen, whose substitution with a higher group can result in a chiral L-center (further details can be found in the appendix). It is also worthwhile noting that glycine's lack of a sidechain should also affect the backbone dihedral and secondary structure preferences.[236] Proline would also expected to have distinct preferences. Thus a key assumption in our models is that a single, amino-acid-unspecific set of backbone paramters is sufficient for the success of the CG model.

4.2.3 Simulation details

All simulations employ Replica Exchange Molecular Dynamics (REMD) for enhanced sampling of conformational space.[237] Atomistic polypeptide simulations are carried out using an in-house python wrapper for the Amber-16 MD engine.[238] In each simulation 16 replicas spaced exponentially in temperature between 270-500 K run for 60 ns each with 5 swap attemps between neighboring repliacs every 20 ps. Most simulations employ the Amber ff96 forcefield[239] with a modified version of the igb5 implicit solvent model,[240] which we have found to correctly fold a variety of helical and sheet-like secondary structures.[241, 242] Atomistic simulations of 1L2Y (60 ns) at 300 K are taken from a previous study using the Amber ff14SBonlysc forcefield[243] with a modified version of the igb5 implicit water model,[240] which was reported to produce the best top-cluster structure prediction among other Amber forcefields tested.[244] Data from the last 20 ns is used to generate statistics.

CG models are simulated using the Lammps MD package.[105] CG simulations of polypeptides during relative entropy minimization and subsequent structure prediction runs using Gō models are performed for 500 ns / replica and the last 100 ns are used to collect statistics. Relative entropy runs use 8 replicas and structure prediction runs are simulated with 10 replicas, both types using temperatures distributed exponentially in the range 270-500 K swapping every 10 ps. Data from replicas is reweighted using the Multi-state Bennett Acceptance Ratio (MBAR) as implemented using the pymbar package.[245] Top-cluster structures from both atomistic and CG simulations are determined using a hierarchical K-Means like clustering algorithm based on the RMSD[241] and aligned with native structures using the well-known Kabsch algorithm.[246] Cartoon representations of protein secondary structures are rendered using both Pymol[247] and VMD.[248]

4.3 Results and discussion

4.3.1 Folding behavior of CG polypeptide models

Fig. 4.3.1 presents a comparison of the folding curves between AA and CG simulations of polyleucine and polyvaline. These curves are calculated from the fractions of temperature reweighted trajectories that fold to within 3 Å RMSD of the atomistic top cluster structures at 270 K (polyleucine: 97% top cluster

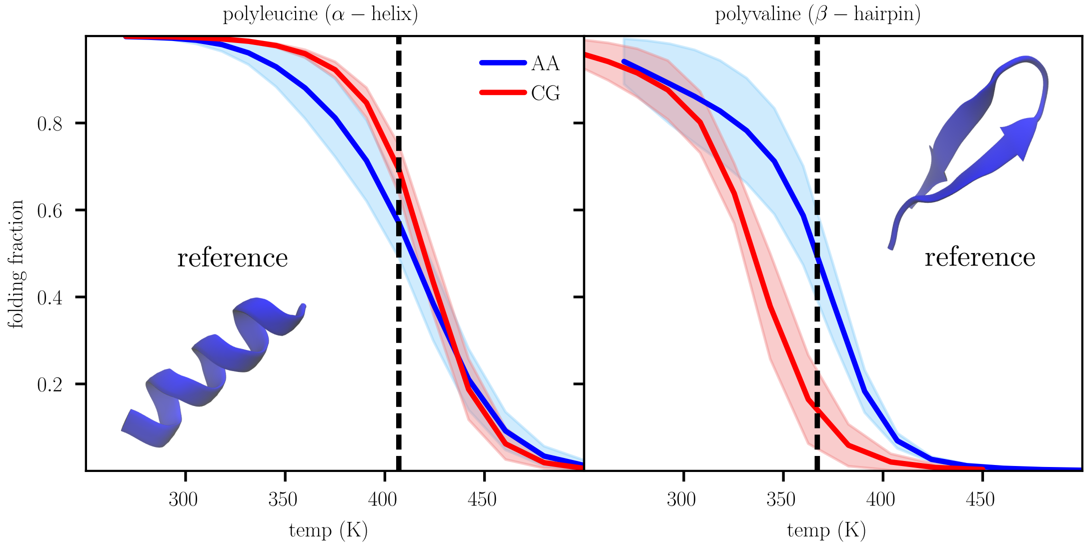


Figure 4.3.1: Comparison of folding curves between atomistic (blue) and CG (red) simulations of 15-mers of leucine (left panel) and valine (right). The folding fraction at a particular temperature is calculated as the fraction of the (reweighted) trajectory that is within 3 Å RMSD of the reference atomistic top-cluster structure at 270 K (helix for polyleucine and hairpin for polyvaline). The CG LEU_{15} model has nearly similar (within 5 K error) folding temperature as its atomistic counterpart while the VAL_{15} model has a $\sim$ 27 K error in the folding temperature, underestimating it.

fraction for a near-ideal helix, polyvaline: 59%: top cluster fraction for a hairpin with a slightly twisted loop). The AA folding temperatures for both polypeptides are quite high (407 K for polyleucine and 367 K for polyvaline) which indicates the high stability of their corresponding folds. However, this is expected since the reference atomistic simulations use implicit solvent. CG polyleucine reproduces the temperature dependent folding behavior for helix formation of its AA reference reasonably well with a $\sim$ 5 K deviation from the AA folding temperature, while CG polyvaline has $\sim$ 50% lesser folding fraction than its AA counterpart and incurs a ∼ 27 K error in the folding temperature.

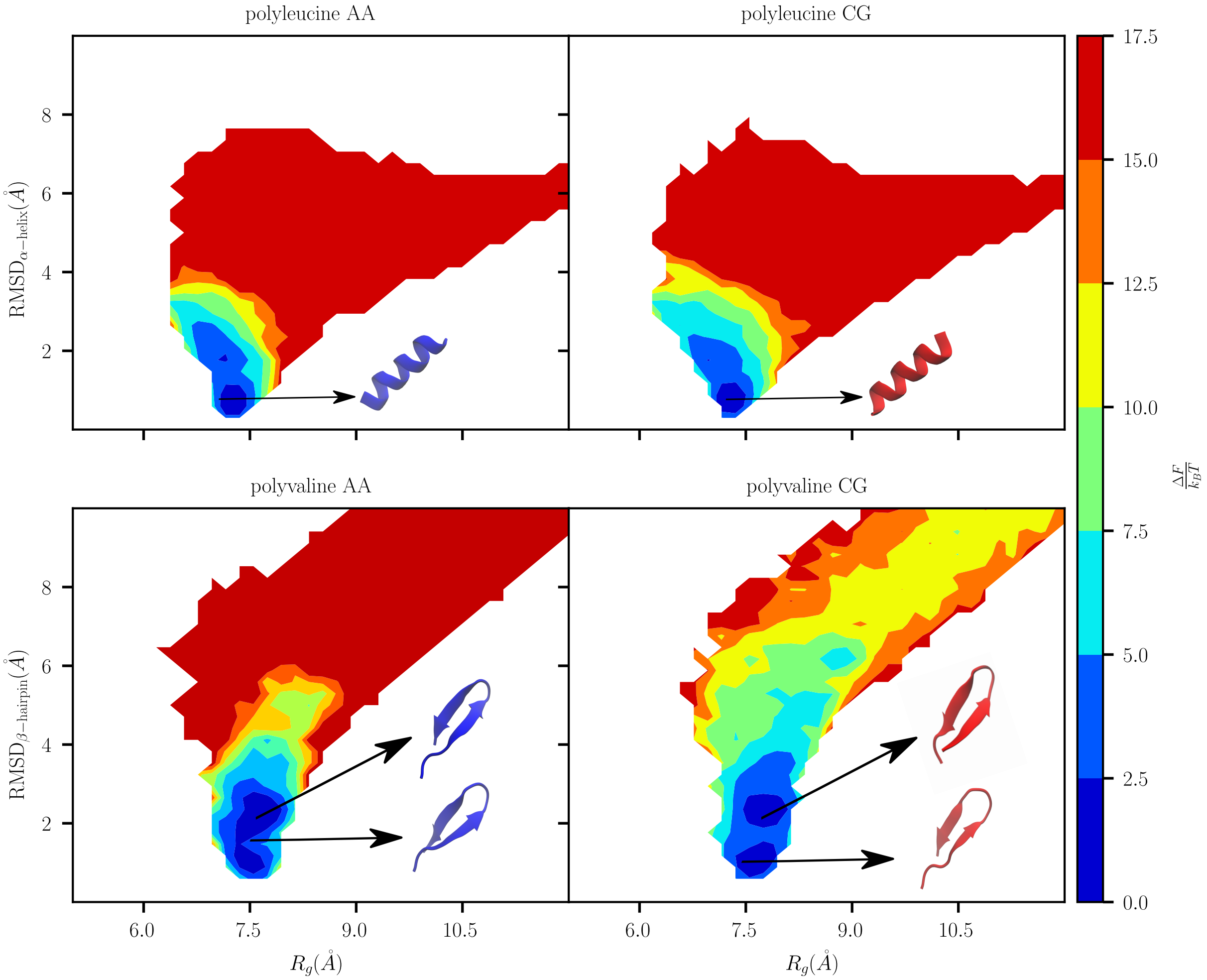


Figure 4.3.2: Comparison between AA and CG simulations of folding free energy surfaces (ΔF) at 280 K, as functions of radius of gyration (R_g) and RMSD from the atomistic top-cluster structure at 270 K, for LEU_{15} (top panel) and VAL_{15} (bottom panel) CG models. While the LEU_{15} CG model exclusively stabilizes an ideal helix similar to its AA reference, the VAL_{15} AA and CG models have significant populations of two closely similar hairpins that are register-shifted from each other. In either case, the top cluster structures are reproduced nearly quantitatively (RMSD $\leq$ 1 Å). AA structures are shown in blue, while CG structures are colored red.

Fig. 4.3.2 compares the free energy of folding at 280 K between the AA and CG models for polyleucine snd polyvaline, as a function of radius of gyration (R_g) and RMSD from the atomistic top-cluster structures. Both AA and CG polyleucine models exclusively stabilize a near-ideal helix (see inset in Fig. 4.3.1), while those for polyvaline are centered around two closely similar hairpin ($\Delta F \leq 2.5\ k_BT$) conformations that are register shifted from one another. Interestingly, both the helix and the hairpin have a similar $R_g \sim$ 7.2-7.5 Å. Both the folding curves in Fig. 4.3.1 and the free energy surfaces in Fig. 4.3.2 show that the polvaline hairpin is harder to stabilize than the polyleucine helix. However, the free energy surfaces are very similar for both leucine and valine polymers and the top cluster conformations are captured nearly quantitatively in the CG model. The only major difference is that CG polyvaline has a broader sampling of large R_g (upto 9 Å) and RMSD (upto 6 Å) structures relative to its AA counterpart, although still these are $\sim 7\ k_BT$ above the minimum and hence rarely visited. Further, it should be noted that classifying folded vs. unfolded states in the CG valine polymer is based on examining alignment with the particular 270 K atomistic hairpin. If one were to look solely at the β-content, the second and third clusters with relative populations of $\sim$ 16%, and 6.5% respectively, are all hairpins that are register-shifted from the AA reference.

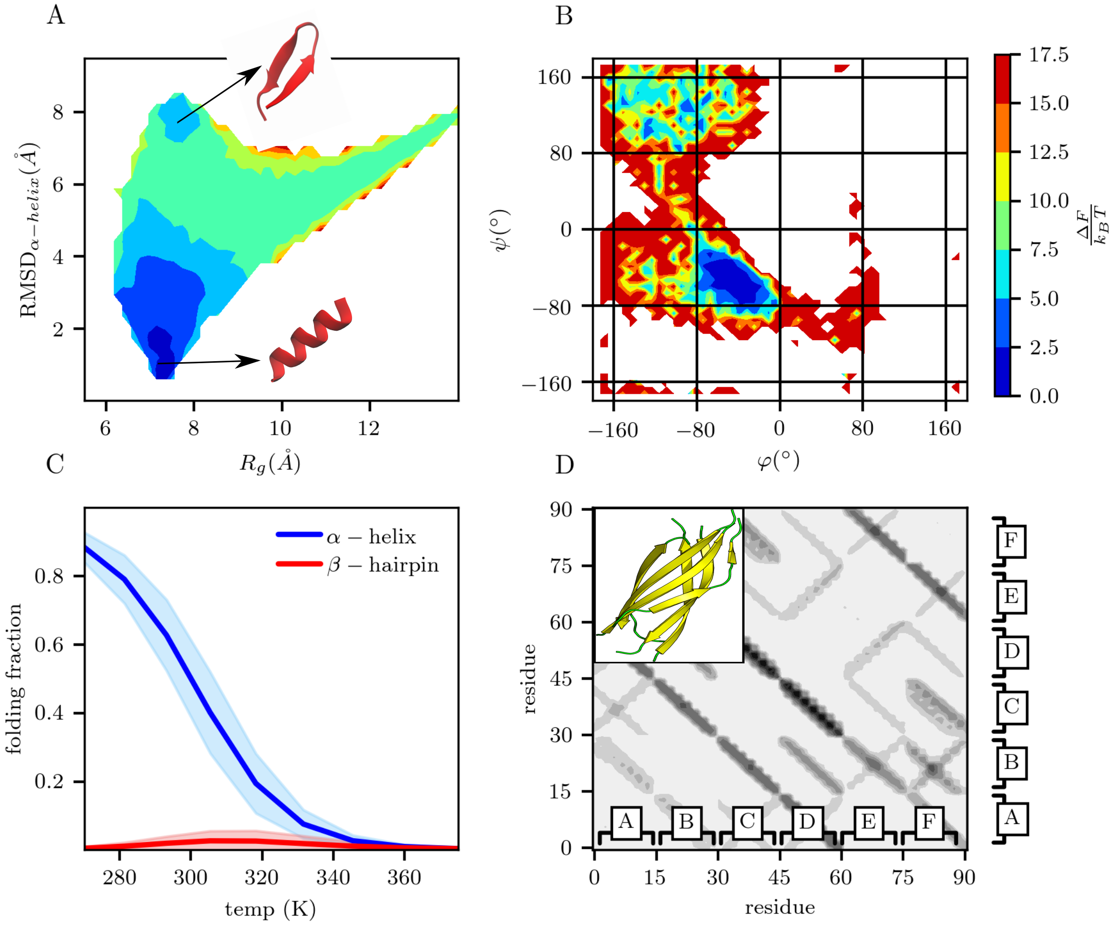


Figure 4.3.3: The $\mathrm{LEU}_{15} + \mathrm{VAL}_{15}$ CG polypeptide model constructed using the extended-ensemble method, by combining data from AA leucine and valine 15-mers: Both the (A) free energy surface as a function of R_g and RMSD from the polyleucine top cluster structure (near perfect helix), and (B) Ramachandran plot at 280 K, reveal the presence of basins dominated by both helices and hairpins separated by a $\sim 7.5 k_B T$ barrier. (C) Folding curves (constructed by considering trajectory fractions within 3 Å of the top clusters of AA polyleucine and polyvaline) show that β-fractions remain consistently lower than 5%. (D) However, when used in a self-assembly simulation of six polypeptide chains, the $\mathrm{LEU}_{15} + \mathrm{VAL}_{15}$ forcefield produces an expected antiparallel zipper structure (in the inset) with $\sim$ 80% β-content, attested by the dominant off-diagonal patterns on an inter-residue contact map (A-F refers to the 6 peptide chains).

As mentioned in section 4.2.2, we investigated two approaches to design a hybrid leucine-valine CG backbone: by parameterizing a CG polypeptide directly from a leucine-valine copolymer reference, and by parameterizing simultaneously from separate polyleucine and polyvaline references, so as to minimize the sum of relative entropies between the model and each reference. Fig. 4.3.3 presents the folding behavior of the extended ensemble CG polypeptide model $\mathrm{LEU}_{15} + \mathrm{VAL}_{15}$. Panel (A) shows the free energy surface (at 280 K) as a function of R_g and RMSD from the AA polyleucine helical top cluster structure at 270 K. The free energy surface has two minima separated by a $7.5 - 10k_BT$ barrier. Both minima have $R_g \sim 7.5\,\text{Å}$; the low RMSD minima represents helical states while the other one represents hairpin conformations. A Ramachandran plot in panel (B) shows the free-energy landscape in the space of dihedral angles (φ, ψ) along the polymer backbone and shows the relatively higher stability of α-helical over β states in this CG model. Interestingly, there is a low fraction of highly unstructured conformations ($\geq 15k_BT$) in the $\varphi \in [60°, 90°], \psi \in [-90°, -60°]$ region which is likely contributed by the more flexible polyvaline. Folding curves in panel (C) relative to the AA top-cluster structures of polyleucine and polyvaline (similar to the procedure used in Fig. 4.3.1) reinforce the relative stability of helical states in this mixed CG model as the β folding fraction is consistently less than 5% across all temperatures.

Although it seems like the LEU_{15} + VAL_{15} CG model prefers α-helices over β sheets, we re-iterate that folding behavior in Fig. 4.3.3 (as also in Figs. 4.3.1 and 4.3.2) is characterized with reference to a single conformation and may not necessarily reflect true β content. Panel (D) presents a more stringent test of the β-like behavior of this CG polypeptide model by using it in a self-assembly simulation of six polypeptide chains. Panel (D) demonstrates that the trajectory averaged inter-residue "global" contact map across all polymer chains contains exclusively parallel and anti-parallel β sheets characterized by contact patterns that are off-diagonal and orthogonal to the main diagonal respectively. Peptide assemblies have enhanced inter-strand hydrogen bonded interactions relative to a single hairpin, and as such are expected to typically stabilize β-rich structures, unless the underlying forcefield has an inordinate amount of helical or flexible character. In fact, a clustering analysis reveals that the dominant self-assembled oligomer is an antiparallel steric zipper, which is known to be one of the most stable motifs that can result from an assembly of short peptide fragments.[249] The self-assembly behavior of this CG polypeptide thus provides good evidence that it contains $\alpha - \beta$ balance which when aided by suitable side chain interactions, may fold heteropeptides.

4.3.2 Native and non-native Gō interactions

Fig. 4.3.4 demonstrates the native Gō interactions optimized from the atomistic trp-cage miniprotein (1L2Y), after fixing the backbone parameters, for the different forcefields summarized in Table 4.2.1. All native potentials have nearly similar forms, with a minima near ~ 3.8 Å (except the LEU_{15} model, which has a minimum at ~ 4.5 Å), followed by an approximately linear slope until the cutoff of 10 Å. The $\mathrm{LEU}_{15}+\mathrm{VAL}_{15}$ native interaction optimized from the extended leucine-valine ensemble has the largest inner-core depth $\sim 9.5 k_B T$ and is also

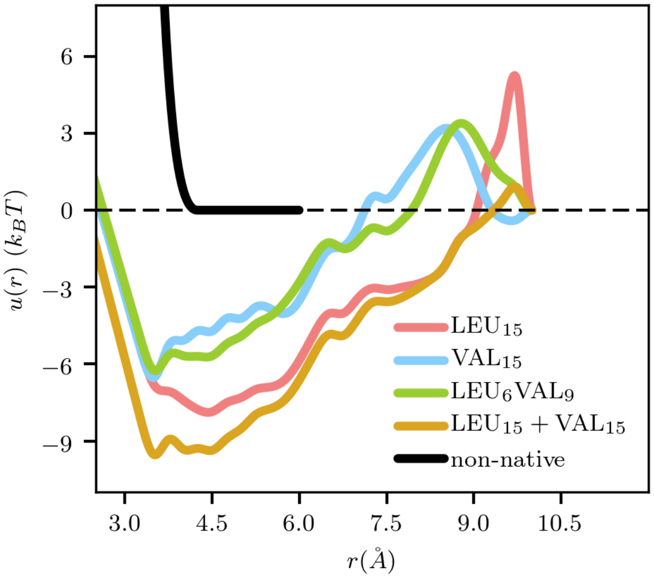


Figure 4.3.4: Native interactions optimized from the atomistic trp-cage miniprotein (1L2Y) at 300 K, for the different backbone forcefields in Table 4.2.1. All native potentials have an inner repulsive core near ~ 3.8 Å and are cut off at 10 Å. The non-native interaction is fixed as a WCA potential with $\sigma = 3.8$Å and $\epsilon = 4.2 k_B T$

somewhat flat between 4-5 Å. The well-depth of this potential is interesting because it exceeds that of both the LEU_{15} and VAL_{15} derived potentials, and since the $\mathrm{LEU}_{15} + \mathrm{VAL}_{15}$ backbone forcefield was parameterizd from both polyleucine and polyvaline AA references. Near the cutoff of 10 Å, all the potentials exhibit a small repulsive bump instead of flat-lining. To assess if the jump results from cutting off the potential prematurely, we re-optimized the $\mathrm{LEU}_{15} + \mathrm{VAL}_{15}$ native interaction using larger cutoffs of 15 and 20 Å (not reported here), but still observed this behavior. Perhaps, the optimal native interaction (i.e., at minimum relative entropy) is indeed repulsive near the cutoff, such that the overall shape produces a highly confining effect for native contacts, reminiscent of harmonic restraints commonly used in elastic network models.[215]

For the repulsive non-native interactions, we use a WCA potential. Based on the location of the inner core for the native potentials, we set the σ for this potential to 3.8 Å. The ϵ for the non-native WCA potential is set to 4.2 k_BT to match the ϵ of pure Lennard Jones type native interactions ($\sim$ 4 - 4.4 k_BT) which we optimized in a preliminary study (not reported here) by following a relative entropy minimization using the 1L2Y reference, similar to the development in this chapter. Unlike the native interactions, the non-native potential is not parameterized from the 1L2Y reference. However, its exact form is likely unimportant in the Gō model, since it is applied to residues that are not in contact.

4.3.3 Gō model performance

We use the Gō models from Table 4.2.1) in folding experiments on two sets of peptides / proteins, starting all simulations from fully extended conformations (all dihedral angles set to 180°). The first set is a balanced collection of short peptide fragments ($\sim$ 11-20 residues) with both helical and hairpin structures used in the study in Ref. 244. The second set is a collection of moderately large globular proteins (26-73 residues) taken from Ref 250. Table 4.3.1 provides more details on these sequences.

Fig. 4.3.5 shows the fraction of sequences whose structures are predicted within a given threshold for α, β and $\alpha + \beta$ sequences outlined in Table 4.3.1, in REMD simulations of the CG Gō models of Table 4.2.1. The LEU$_{15}$ + VAL$_{15}$ model is most consistent in producing faithful native structure alignments. It predicts 80% of the pure helical and pure β sequences and 50% (one of the two) mixed $\alpha + \beta$ sequences within 2 Å RMSD (details follow in Figs. 4.3.6 and 4.3.7). The LEU$_6$VAL$_9$ copolymer model is a close second best with 60% helical, 40% β and 40% mixed sequences predicted within 2 Å RMSD. Models derived from pure

Table 4.3.1: Target peptide sequences used for validating the Go-models in this study (shorter sequences are tabulated above the dividing line)

Name	PDB code	Length	Structure*	CG folding temp.**
1CB3	1CB3	11	α	n.a.
1L2Y	1L2Y	20	α	n.a.
2I9M	2I9M	17	α	n.a.
C Peptide	†	13	α	n.a.
EK Peptide	†	12	α	n.a.
15-β	†	15	β	n.a.
1GB1	1GB1	16	β	419 K
1E0Q	1E0Q	17	β	419 K
1LE1	1LE1	12	β	n.a.
1J4M	1J4M	14	β	$\in$ [380 K, 419 K]
Protein A	1BDD	46	α	n.a.
Albumin binding domain	2FS1	49	α	n.a.
α3D	2A3D	73	α	n.a.
YJQ8 WW (res 7-31)	1E0N	27	β	n.a.
FBP28 WW (res 6-31)	1E0L	26	β	n.a.
Ubiquitin (res 1-35)	1UBQ	35	$\alpha + \beta$	n.a.
Protein G	1PGB	56	$\alpha + \beta$	n.a.
α-spectrin SH3	1SHG	57	β	$\in$ [419 K, 457 K]
src-SH3	1SRL	56	β	$\in$ [450 K, 500 K]
bacterial Flavodoxin	1FUE	163	$\alpha + \beta$	$\in$ [331 K, 345 K]
TIM barrel monomer	8TIM	247	α/β	$\in$ [305 K, 318 K]

*α and β are simplified structure classifications that neglect finer details like 3-10 helices, β bulges, turns, coils and loops.

**Using the $\text{LEU}_{15} + \text{VAL}_{15}$ CG model

† Details of the native structure can be found in Ref. 244

n.a. : Sequences exhibit > 50% folding fraction across all temperatures in REMD simulations.

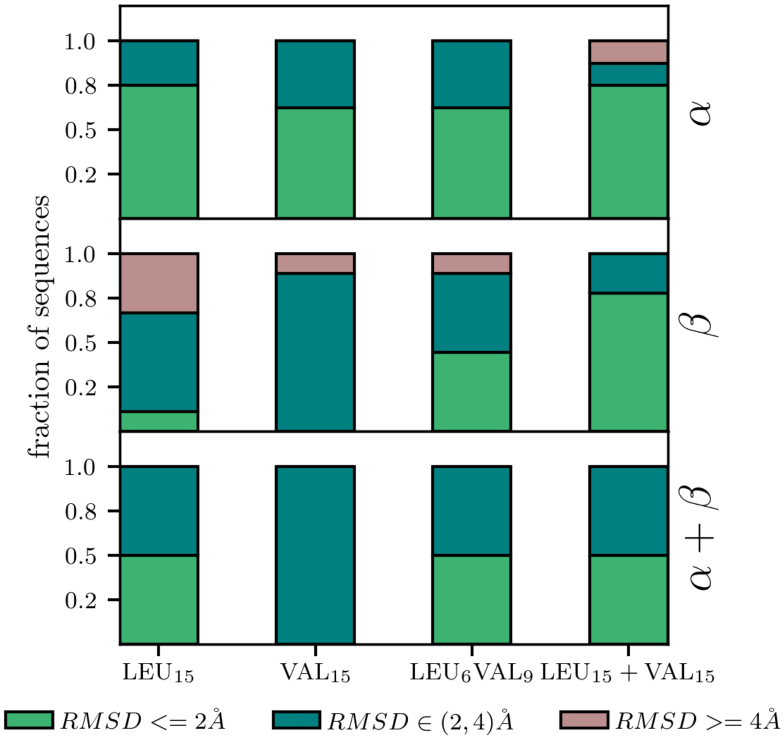
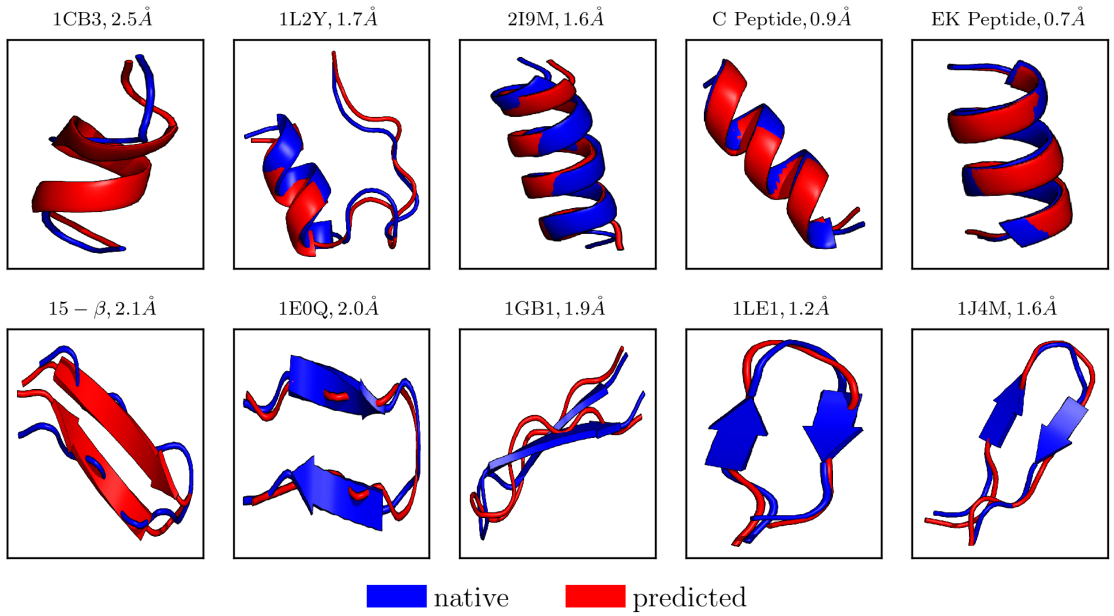


Figure 4.3.5: Assessment of the accuracy of Gō-like models developed from the different backbone forcefields (along the X axis) in Table 4.2.1, for predicting the structure of α-helical (top panel), β-rich (middle panel) and $\alpha + \beta$ (bottom panel) sequences (detailed in Table 4.3.1. In each panel, stacked bar charts show the fraction of sequences (along the Y axis) that fold to within 2 Å (green), 2-4 Å (teal), and $>$ 4 Å (brown) RMSD from the native structure. The LEU$_{15}$ + VAL$_{15}$ CG backbone at 300 K, derived from an extended ensemble of polyleucine and polyvaline AA references, has relatively better prediction rates across all sequences. All RMSDs are ensemble-averaged values from trajectories at 290 K.

polyleucine and pure polyvaline references, unsuprisingly, are biased towards α and β sequences respectively: LEU$_{15}$ resolves 80% of helical sequences within 2 Å, and VAL$_{15}$ captures 90% of β sequences within 2-4 Å. Interestingly, the polyleucine model also allows for some β character since it correctly predicts one of the two mixed sequences (protein G) barely within 2 Å, while the polyvaline model has the worst overall accuracy, consistently straying from the desired $\geq$ 2 Å RMSD alignment even for β-rich sequences. However, it is remarkable that most of the models produce reasonably close alignment for the shorter peptides. Native interactions in Gō models are essentially a set of restraints dependent on native contact information, so the ability to accurately fold shorter peptides with fewer contacts, is arguably an important performance metric for such models. Success with the shorter peptides is thus directly linked to the efficiency of the backbone/(s) in exploring the relevant dihedral space and reflects considerable promise for the first stage of the bottom-up peptide model.

Figs. 4.3.6 and 4.3.7 show the structures of target sequences, predicted with the $\mathrm{LEU}_{15} + \mathrm{VAL}_{15}$ Gō model, superposed on the corresponding native structures of these sequences. All CG REMD simulations were initialized from fully extended states. It is remarkable that such a simple CG model parameterized from *single* polypeptide references is able to correctly predict 74% (14 out of 19) of all the target sequences within 2 Å, attesting to an overall high prediction accuracy. The reported RMSDs are all ensemble averaged from the 290 K trajectory of the corresponding REMD simulation. Note however, that the folding temperatures reported in Table 4.3.1 for the CG models are very high and in some cases above



1CB3, 2.5Å

1L2Y, 1.7Å

2I9M, 1.6Å

C Peptide, 0.9Å

EK Peptide, 0.7Å

$15-\beta$, 2.1Å

1E0Q, 2.0Å

1GB1, 1.9Å

1LE1, 1.2Å

1J4M, 1.6Å

Figure 4.3.6: Top cluster structures for short sequences (11-20 residues) predicted by the Gō model derived from the extended-ensemble LEU_{15} + VAL_{15} backbone. Native structures are in blue while simulated ones are in red. The RMSD (averaged from the trajectory at 290 K) from the native structure is reported beside the sequence name. The average standard error (standard deviation / mean) in calculating the RMSDs is $\sim$ 6.5%

the maximum replica temperature, preventing their exact calculation. While this certainly disagrees with experiment or atomistic simulations, it shows that the bottom-up procedure with a Gō strategy embeds highly stabilizing native contact interactions that are biased towards the native state. Typically a Gō model provides faster access to the native structure by smoothening amd exaggerating the energy landscape.[215] The disagreement may also partially result from the use of implicit water reference polypeptide simulations for parameterizing the backbone CG model, which have high folding temperatures (refer to Fig. 4.3.1). There

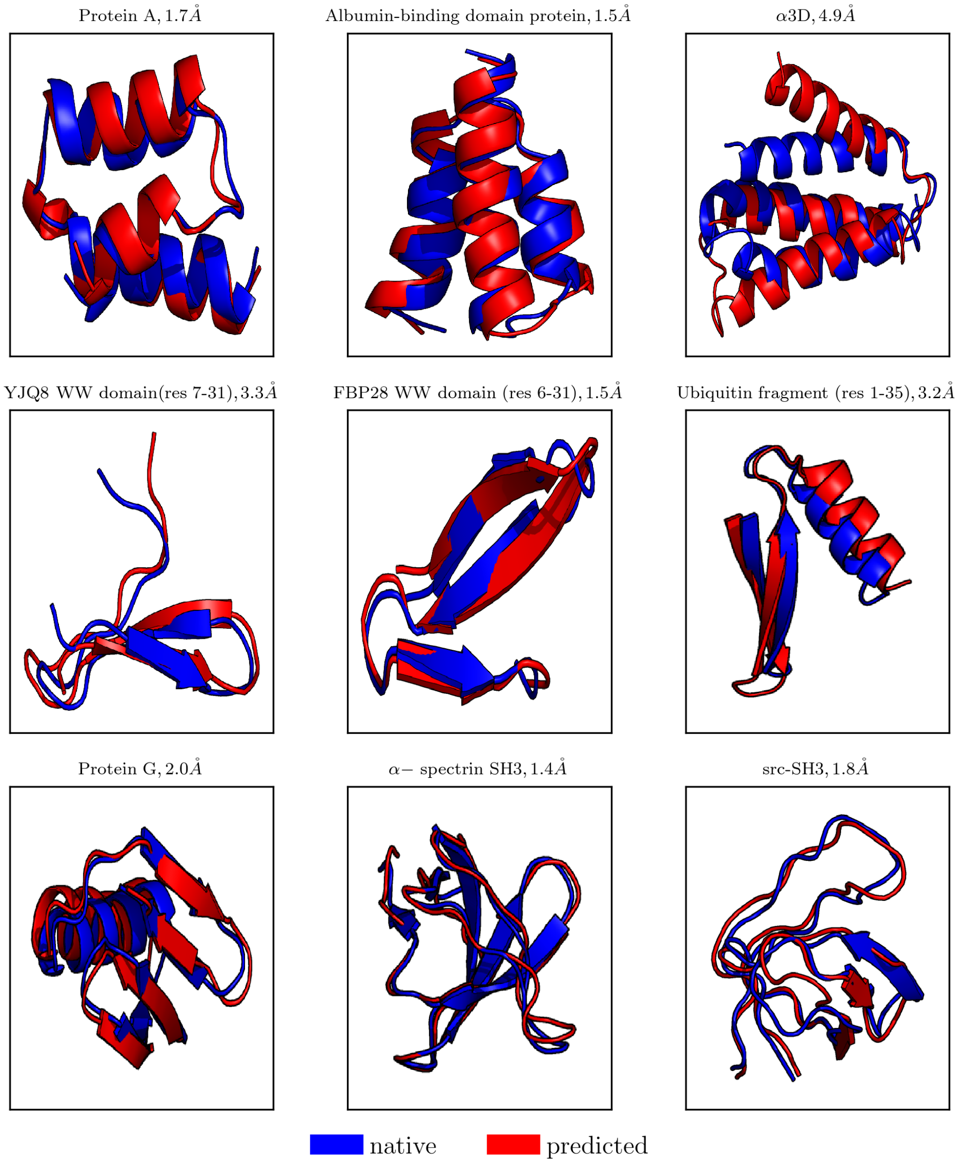

Protein A, 1.7Å
Albumin-binding domain protein, 1.5Å
α3D, 4.9Å
YJQ8 WW domain(res 7-31), 3.3Å
FBP28 WW domain (res 6-31), 1.5Å
Ubiquitin fragment (res 1-35), 3.2Å
Protein G, 2.0Å
α− spectrin SH3, 1.4Å
src-SH3, 1.8Å
native
predicted

Figure 4.3.7: Top cluster structures for longer sequences (26-73 residues) predicted by the Gō model derived from the extended-ensemble $\mathrm{LEU}_{15}+\mathrm{VAL}_{15}$ backbone. Native structures are in blue while simulated ones are in red. The RMSD (averaged from the trajectory at 290 K) from the native structure is reported beside the sequence name. The average standard error (standard deviation / mean) in calculating the RMSDs is $\sim 6.5\%$

are, however, native-centric models in the literature that use experimentally observed folding temperatures to fit the native interactions and by design, provide quantitative agreement with experimental folding temperature.[212] One can envision a constrained relative-entropy-minimization protocol where experimentally obtained folding temperatures can be included as constraints during backbone or sidechain parameterization to provide a closer match to experimental folding behavior. Alternatively, the attractive wells in the currently derived CG native contact potentials might be scaled systematically to obtain better agreement for folding temperatures. However, we leave these directions to future work.

Differences from native structures in Figs. 4.3.6 and 4.3.7 result mostly from mis-predicting flexible regions (turns and loops) with the exception of 1L2Y, α-spectrin SH3 and src-SH3. It is no surprise that the flexible region for 1L2Y is captured entirely, since the native interactions are optimized from a 1L2Y reference. The only sequence that does not fold well with this Gō model is α3D, where large mis-alignment in the interconnecting turn regions (residues 20-26, 46-53) leads to different major axis vectors for two of the helices (residues 1-19, 54-73). The fraction of native contacts satisfied for this sequence is 87% (calculated from the top-cluster) as opposed to 97 % (averaged over all targets) for the other sequences. Other major mis-alignments in turn regions include the 35-mer fragment of Ubiquitin (residues 8-11), 1GB1 (residues 7-10) and 15-β (residues 7-10). The frequency distribution of amino acids in these four sequences reveals 23% glycine and 11% proline in the hard-to-capture flexible regions as opposed to 3% glycine and no proline in regions with well defined secondary structures. This indicates that a more accurate CG model should incoroprate special backbone interactions (angular and dihedral) for glycine and proline residues. We note that these parameters may be parameterized from an extended ensemble of AA references that contain glycine and proline polymers.

The CG REMD simulations of the target sequences are ostensibly ~ 30 times faster than their corresponding atomistic counterparts for the short fragments in Fig. 4.3.6, while the longer sequences in Fig. 4.3.7 take ~ 10 hours of total CPU time, starting from fully extended states. While not an entirely fair comparison, it is worth noting that implicit water MD simulations of (NuG2 variant of) atomistic protein G have been reported to take ~ 54 μs of simulation time and approximately 54 hours of real time with the Amber14 MD software on GPGPUs[251] while explicit water MD simulations of protein G can take as long as 1154 μs of simulation time on the Anton supercomputer designed specifically for fast MD simulations.[252] While the efficiency in our simulations is expected since Go forcefields are known fast, unfrustrated folders,[215] it nevertheless provides an important upper bound on the expected speed-up over AA forcefields using these four site CG models.

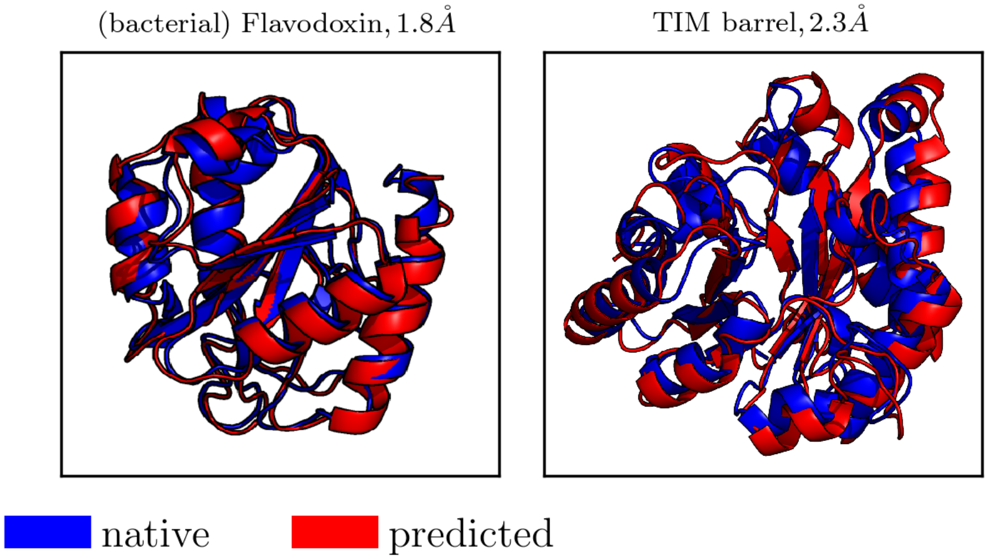


Figure 4.3.8: Top cluster structures for bacterial flavodoxin (163 residues) and the TIM barrel protein (247 residues) predicted by the Gō model derived from the extended-ensemble LEU_{15} + VAL_{15} backbone. Native structures are in blue while simulated ones are in red. The RMSD (averaged from the trajectory at 290 K) from the native structure is reported beside the sequence name. The average standard error (standard deviation / mean) in calculating the RMSDs is $\sim$ 6%

As a final set of structure prediction tests, we use the $\mathrm{LEU}_{15}+\mathrm{VAL}_{15}$ Gō model to fold two much longer sequences, namely a 163 residue flavodoxin encoded in the *H.pylori* genome (PDB code: 1FUE) and a 247 residue Triose-Phosphate-Isomerase (TIM) barrel monomer (PDB code for the dimer: 8TIM). Flavodoxin is characterized by a five-stranded parallel β sheet core sandwiched by α helices, while the TIM barrel is one of the most common folds found in nature, consisting of eight overlapping, alternating $\beta - \alpha - \beta$ supersecondary structures. Starting with fully extended conformations, the Gō model achieves remarkable prediction accuracies in each case, scored by trajectory averaged RMSDs of 1.8 Å for flavodoxin and 2.3 Å for the TIM barrel, after 500 ns of REMD simulation taking $\sim$ 8 days of total CPU time. To the best of our knowledge, CG Gō models have not been used in structurally-accurate folding experiments for such large proteins. This shows promise for the $\text{LEU}_{15} + \text{VAL}_{15}$ Gō model in reproducing local conformational fluctuations of large macromolecules. So, for instance, this Gō model may be useful for efficiently capturing the fluctuations of large proteins in docking simulations instead of keeping them rigid, which may increase the accuracy of the computational screening process without significantly increasing the computation overhead.

4.3.4 Fault tolerance of the $\text{LEU}_{15} + \text{VAL}_{15}$ Go model

As mentioned previously, a Gō model essentially encodes a set of experimentally determined restraints on the positions of native contact residues. This information can entail error depending on the experimental / atomistic simulation method used to solve the structure. To simulate the effect of such error, we randomly delete a fraction of the native contacts and apply the Gō potentials (Fig. 4.3.4) to the reduced set of contacts. Fig 4.3.9 shows the prediction quality for protein G with an increasing fraction of deleted contacts using the $\mathrm{LEU}_{15} + \mathrm{VAL}_{15}$ model. The prediction quality deteriorates from 2 Å with complete native structure information to 5.6 Å when 20% of native contacts are removed. The fraction of the reduced set of native contacts satisfied, fluctuates between 76-80% (98% when all contacts are retained). However, the prediction quality does not decrease monotonically since the contacts are removed at random, and the ones that are relatively more important, such as those closer to the folding core may be better retained in some cases even when a larger overall number of contacts are removed. It is interesting to note however, that the simulated structures retain the helix in all cases with relatively high native alignment, while the β character fluctuates. Moreover, even when 20% of the contacts are removed, the Gō model still gives

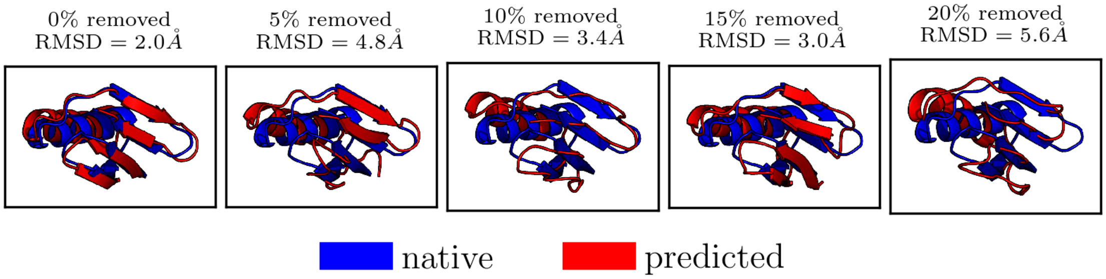


Figure 4.3.9: A test of robustness for the Gō model derived from the extended ensemble $\mathrm{LEU}_{15} + \mathrm{VAL}_{15}$ backbone forcefield. This model is used in CG REMD simulations of protein G while reducing the available native-contact information by deleting zero to 20% of the native contacts. The RMSD (ensemble averaged from the 290 K trajectory) with the native structure varies between 2-5.6 Å and has a standard error (standard deviation / mean) of $\sim$ 6.5 %. The prediction quality does not decrease monotonically since contacts are removed randomly. Native and predicted structures are colored blue and red respectively.

a structural collapse that achieves the correct spatial proximity of the helical and hairpin regions. We will not probe further into which contacts are more important to retain over others, or what the critical fraction of retained contacts needs to be for a desired accuracy, but such directions may be interesting to study. Nevertheless, Fig. 4.3.9 shows that at least for protein G, the extended ensemble CG backbone produces Gō models that can reasonably tolerate low to moderate levels of missing information.

4.3.5 Conclusions

In this chapter, we developed four-site CG polypeptide models through relative entropy minimization using atomistic single peptide references. Supplemented with suitable sidechain interactions, these models serve as CG backbone forcefields for folding short fragments as well as large globular proteins. Specifically, we parameterized polypeptide backbones separately from atomistic references of leucine and valine 15-mers and then simultaneously from both, by minimizing the relative entropy with an extended ensemble containing both these references. We augmented the backbone interactions with simple Gō-like sidechain potentials that are modulated by the proximity of residues in the native structure. We optimized the functional form of these native interactions by further minimizing relative entropy from a short native-centric AA simulation of the trp-cage miniprotein

(1L2Y). Gō-like sidechain interactions require the native structure as an input and thus, do not produce de-novo structure predictions. However, they furnish an excellent upper bound for the capability of CG backbone interactions in sampling the relevant dihedral space, which is the focus of this work. We used the extended-ensemble backbone in conjunction with inter-sidechain Gō potentials to perform folding experiments on both short fragments (11-20 residues), globular proteins (26-73 residues) and two examples that would be intractable with AA forcefields, with sequences that contained α, β and $\alpha - \beta$ mixed structures.

Our results show that CG models derived from pure polyleucine and polyvaline are biased towards the secondary structure propensities of the references. Thus, the LEU_{15} model did not predict the correct native structures for sequences with high β content, while the VAL_{15} backbone failed to produce high resolution predictions for both α and β sequences. While combining leucine and valine chemistries by using a 40% leucine - 60% valine coopolymer improved structure prediction to some extent, a better strategy proved to be an extended-ensemble approach to parameterize a CG polypeptide forcefield simultaneously to two reference simulations: leucine and valine. Our final model is the $\text{LEU}_{15} + \text{VAL}_{15}$ extended-ensemble model, which could ultimately resolve simulated structures within 2 Å of the native conformations, when used in conjunction with Gō-like sidechain interactions.

It is remarkable that such a simple bottom-up CG backbone without any bioinformatic assistance, is transferable enough to be able to resolve up to ~ 247 residue globular proteins within 2.5 Å. This chapter underscores the importance of the backbone in defining the fold landscape, and of the sidechain interactions in filtering for specific conformations.[225–228] Our results also show that sidechain interactions can be perhaps described at a lower level of detail than that required for the backbone. For instance, we maintained a uniform resolution of a single bead for all sidechains in the Gō model and ignored differences in sidechain size, which is arguably important to consider for de-novo folding. Nevertheless, it is quite compelling that the backbone interactions play such a significant role in sampling the relevant dihedral space. Thus, relative-entropy-optimized, bottom-up, four-site CG models of peptides hold considerable promise as putative CG protein forcefields that can offer predictive insight into folding and oligomerization processes. Moreover, while native-aware sidechain potentials do not offer de-novo capabilities, we believe that the Gō model presented in this chapter is at a level of sophistication to be readily applicable to cases where one needs a quick enumeration of conformations with the correct local fluctuations around the native state. For instance, this Gō model can be employed as a very rigorous, physics-based scoring function for a a preliminary screening of allowed conformations of flexible peptide ligands docking on to large macromolecules. The Gō model can be especially useful to structure prediction if contact and /or secondary structure constraints are available, e.g. through experiments or bioinformatics. This could be extremely helpful to quickly narrow down NMR-compatible ensembles of structures, given NOE constraints.

The Gō model presented in this chapter has a total of 808 parameters, most in the form of spline knots and of which 40 knots correspond to the splined native-contact potential. This set might be reduced by using two-parameter Lennard Jones functions for the native potentials. Other immediate areas of improvement include (a) incorporating a richer variety of secondary structural elements in the training set for (extended-ensemble) optimization of both backbone and/or Gō-like native interactions, (b) accounting for glycine and proline through special backbone potentials, and (c) including experimental data such as folding temperatures through generic constraints in the relative entropy minimization algorithm, perhaps by using Bayesian inference based approaches.[192, 253]

This work presents a proof-of-principle of CG peptide models derived directly from the underlying folding free energy landscapes, such that most of the macroscopic structural properties including self-assembly behavior are emergent and don't need to be fit separately. The results presented in this chapter serves as the starting point for more refined, sequence-chemistry based peptide models that will need minimal or no additional bioinformatics and be truly predictive, not only for the now-well-studied, protein folding problem, but also for complex self-assembly phenomena such as those in cancerous or neuro-degenerative disease pathophysiology.

Appendix

4.A Reformulation of the extended-ensemble CG algorithm within the relative-entropy framework

Consider N atomistic reference simulations, such that simulation k has n_k number of snapshots. While parameterizing a CG model from any one of these at a time, the likelihood of the model, given the reference is simply: $\prod_i P(\mathbf{R}_i \mid \boldsymbol{\lambda})$, where $\mathbf{R}_i$ is the set of atomistic degrees of freedom from snapshot i projected or mapped on to the reduced set of CG degrees of freedom, and $\{\boldsymbol{\lambda}\}$ is the set of parameters describing the CG forcefield. Since the N simulations are independent, the likelihood for the model given the *extended-ensemble* of all N references is:

$$L = \prod_{k=1}^{N} \prod_{i=1}^{n_k} P(\mathbf{R}_i \mid \boldsymbol{\lambda})$$

Maximizing the log-likelihood, one gets

$$\arg\max_{\boldsymbol{\lambda}} \log L = \arg\max_{\boldsymbol{\lambda}} \sum_{k=1}^{N}\sum_{i=1}^{n_k} \log P(\mathbf{R}_i \mid \boldsymbol{\lambda})$$

The inner summation can be manipulated as:

$$\begin{aligned}\sum_i \log P(\mathbf{R}_i \mid \boldsymbol{\lambda}) &= n_k \left[\frac{1}{n_k}\sum_i \log P(\mathbf{R}_i \mid \boldsymbol{\lambda})\right] \\ &\approx n_k \big\langle \log P(\mathbf{R} \mid \boldsymbol{\lambda})\big\rangle_k \\ &= -n_k \left\langle \log \frac{P_k(\mathbf{R})}{P(\mathbf{R} \mid \boldsymbol{\lambda})}\right\rangle_k + n_k \big\langle \log P_k(\mathbf{R})\big\rangle_k\end{aligned}$$

Here, in the second step, we take the thermodynamic limit $n_k \to \infty$, i.e., we have a large number of snapshots from all the ensemble members, which allows us to replace the sum with a expectation over the (atomistic) microstate probability distribution of system k, denoted by $\langle\,\rangle_k$. Note that we can avoid such approximation by using a more precise multinomial expression to compute the likelihood as demonstrated in Ref. 32 However, the present approach still provides a reasonably correct proof. The next step involves multiplication and division by $P_k(\mathbf{R})$, which is the atomistic probability distribution function of system k. Thus, the sum splits up into two parts, the first of which can be identified as the negative of the relative entropy ($S_{\text{rel}}^{(k)}$) between the CG model and reference k, while the second term is the Gibbs entropy of reference k and does not depend on the CG model parameters. This results in,

$$\begin{aligned}\arg\max_{\boldsymbol{\lambda}} \log L &= \sum_{k=1}^{N} \arg\min_{\boldsymbol{\lambda}} \left\langle \log \frac{P_k(\mathbf{R})}{P(\mathbf{R} \mid \boldsymbol{\lambda})} \right\rangle_k \\ &= \arg\min_{\boldsymbol{\lambda}} \sum_{k=1}^{N} S_{\text{rel}}^{(k)}\end{aligned}$$

which completes the proof. Thus, simultaneous relative entropy minimization from multiple reference simulations guarantees maximal overlap of microstate probabilities irrespective of thermodynamic state or chemical dissimilarities between the references.

4.B Selecting a pseudo-side chain for glycine

Glycine has two equivalent hydrogens which make it achiral. The Gō models presented in this chapter assume one of these hydrogens as a pseudo sidechain for glycine to prevent missing any native contact pairs where glycine participates. The location of the pseudo sidechain is determined as the hydrogen whose substitution with a higher group converts the stereochemistry to the L form which is the typical conformation for all naturally found amio acids. Such a hydrogen is known as a prochiral-S hydrogen (S equivalent to L here) and can be determined from the Cahn Ingold Prelog rules as the one that points away from the outward normal to the peptide bond plane.

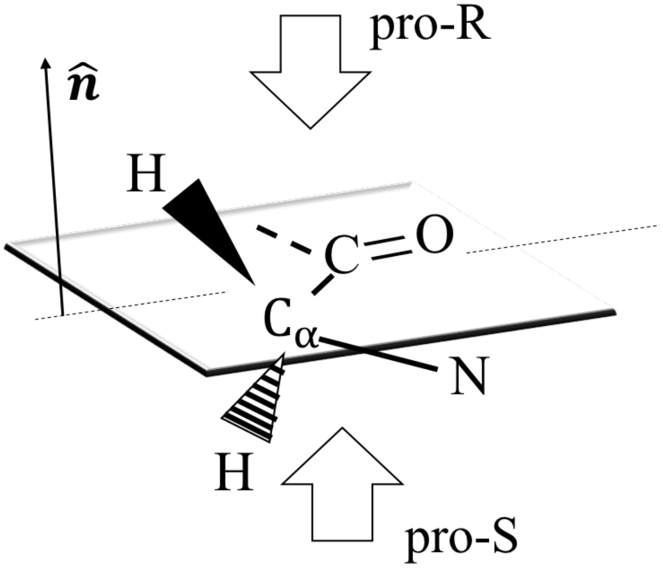


Figure 4.B.1: The geometry of atomistic glycine in the reference plane of the sp^2 carbonyl carbon. Glycine has two equivalent hydrogens pointing along the outward (marked $\hat{\mathbf{n}}$) and inward normals to this plane. The Cahn-Ingold-Prelog stereochemical rules reveal the prochiral-S hydrogen as the one whose projection along the inward normal has a positive component.

4.C The $LEU_{15} + VAL_{15}$ CG backbone and Gō-like sidechain forcefields

The Go-like native interactions for the various backbones have been presented in section 4.3.2. Here, we show the results of the relative entropy optimized CG polypeptide forcefield parameterized from the joint ensemble of Leucine and Valine 15-mer atomistic references. Fig. 4.B.1 shows that (and this is theoretically guaranteed[35]) at a minimum of relative entropy, the atomistic and CG distribution functions for arguments of the different potential functions (i.e., bond lengths and angles, dihedral angles and pair distances) match quantitatively. The only exception is the α carbon - sidechain or CS bond potential (2$^{\text{nd}}$ row, 3$^{\text{rd}}$ column) where the optimal CG bond-length distribution lies in between the bimodal atomistic counterpart. This is expected since it is not possible to capture a non-parabolic bond-length distribution using a harmonic bonded potential, and minimizing relative entropy guides the optimal CS bond length to an average between the two atomistic modes that represent the Leucine and Valine bond CS bond lengths. Similarly, all interactions (bond, angle and nonbonded pair) involving side-chains admit several peaks in their corresponding distributions due to contribution from both Leucine and Valine polymers.

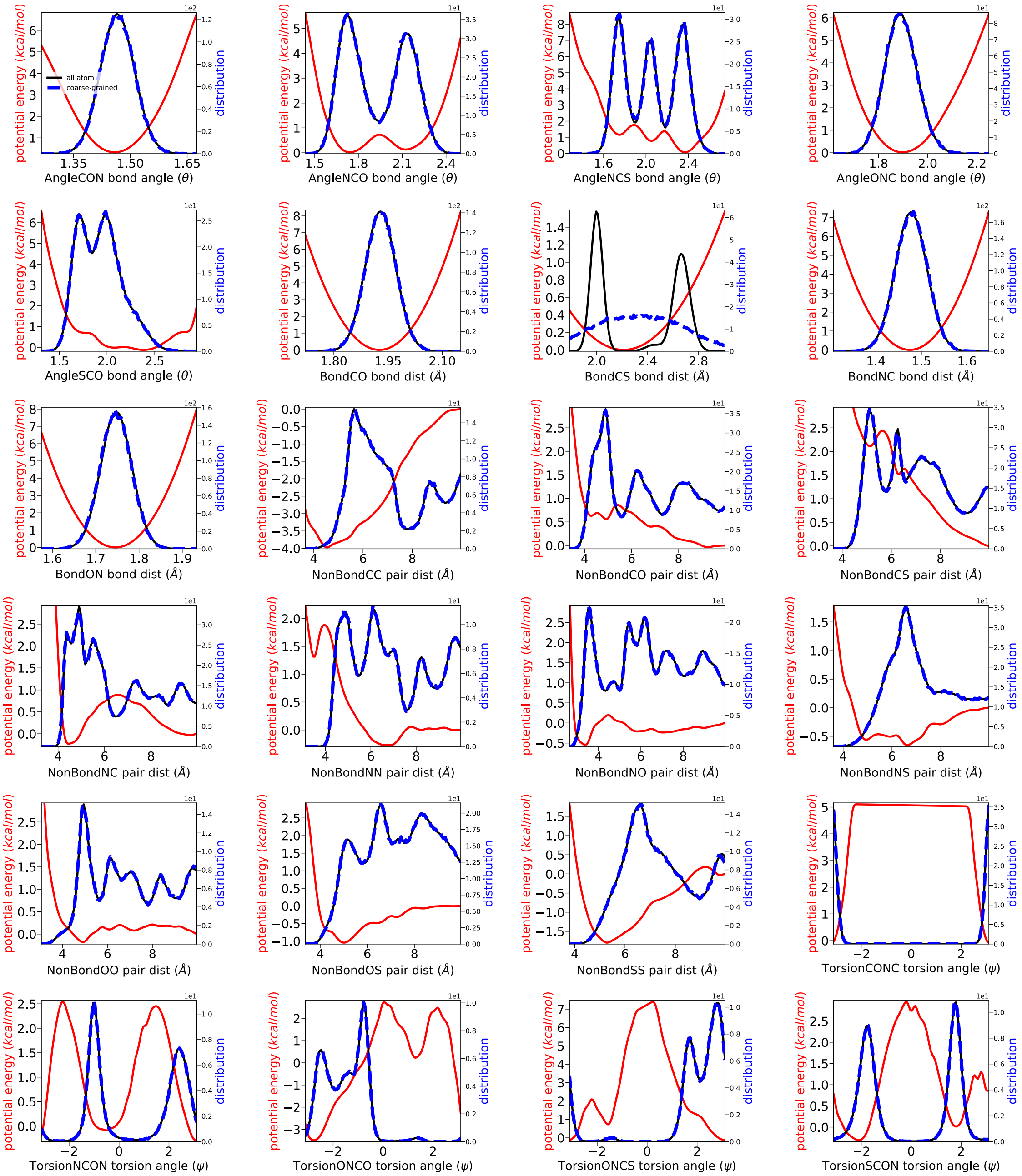


Figure 4.C.1: Optimized interactions (intra-backbone, backbone-sidechain and inter-sidechain) for the $\mathrm{LEU}_{15}+\mathrm{VAL}_{15}$ extended-ensemble CG polypeptide model. Potential functions are in red, while black (atomistic) and blue (CG) lines show that at a relative entropy minimum, the atomistic and CG distribution functions for the argument of the potential function (bond lengths, bond angles, dihedral angles, pair distances) match quantitatively. Potentials are reported in kcal/mol.

Chapter 5

Self-propagating propensities of polypeptide oligomers in templated assembly

5.1 Introduction

Aggregation of partially folded or misfolded proteins into fibrillar super-structures has been identified as the underlying mechanism for a host of neuro-degenerative diseas such as Alzhiemer, Parkinson, Huntington, ALS, prion diseases like CJD, and others like type II diabetes. The fibrillar assemblies formed under such pathogenic conditions are generically called amyloids that manifest macroscopically as plaques on tissues and organs in the body.[17,18] The process of aggregation was first illustrated with prion proteins over three decades ago. The "prion hypothesis" posits that fibril formation is a protein-based replication process that propagates relentlessly, whether initiated by an infectious inoculum or through the spontaenous development of pathogenic aggregates.[254, 255] It is now widely accepted that amyloid formation is essentially a nucleation-polymerization process, where the rate limiting step is the formation of small metastable aggregates which then grow rapidly through subsequent monomeric attachment.[256–258] Thus, small soluble peptide aggregates are amyloidogenic precursors, contributing to cellular toxicity.[257, 259] In this chapter, we focus on such small oligomers in the context of their ability to seed or template amyloid formation. Specifically, we ask: are certain templates inherently more capable of inducing spontaneous aggregation than others? There is experimental evidence that amyloid aggregates can be polymorphic to the extent that there may be significant correlations between fibril topology and variations in disease development.[260] Very recent investigations into the structure of the tau protein have taken this idea one step further, to reveal that the filamentous shapes of tau protein *monomers* implicated in Pick's disease and Alzheimer's are significantly different.[19, 20] This suggests that there may be a unique relationship between different tau conformers and the neurodegenerative pathologies they lead to through continuous fibrillarization. In this chapter, we present a minimal test of this hypothesis by investigating how differences in the shape of *starting templates* (monomers and short protofibrils) may give rise to polymorphism in the corresponding self-assembled fibrils. While peptide folds are now well classified,[223, 261] similar efforts for amyloidogenic superstructures are only at their nascent stage.[249, 262] Thus, this chapter may form a first step towards mapping the space of oligomeric folds prone to self-propagation.

Experimental techniques like X-ray scattering, solid-state NMR, infra-red spectroscopy, fluoresence spectroscopy, electron microscopy, atomic force microscopy, and computational studies typically utilizing molecular dynamics (MD) have shown that amyloids are enriched in β-sheet content with a typical cross-β architecture.[249, 263–266] In spite of these developments, high resolution structure detection through experiments remains somewhat challenging for amyloids due to their non-crystallinity and high insolubility, such that MD simulations can be used to provide important structural, thermodynamic and kinetic insights into amyloid formation.[204, 267–270] For instance, early atomistic simulations by Ma, Nussinov and co-workers investigated the structural stabilities of amyloidogenic fragments such as the AGAAAAGA region of the Syrian hamster prion protein, the GNNQQNY region of the yeast prion sup-35, various regions from the β-amyloid (Aβ) protein associated with Alzheimer's and the DFKNF region (residues 15-19) from the human calcitonin hormone.[271–274] They found that a functional aggregation seed or nucleus typically requires 8 monomers (from the AGAAAGA study)[271] but can also be as small as 3-4 monomers (from the GNNQQNY study)[272] . Further, the authors also found that the stabilities of different β-sheet motifs can be sequence-dependent: antiparallel conformations were most stable for Aβ_{16-22} oligomers[273] while parallel strands were more ordered and resistant to high-temperature dissociation for the DFKNF fragment.[274] However, these studies probed only the early stages of fibril formation because detailed atomistic MD simulations are limited to length and time-scales of nanometers and tens of microseconds respectively. While this is already less than the characteristic millisecond (and greater) timescales for protein folding, amyloid formation is slower. MD simulations of amyloid self-assembly starting from random monomeric configurations amount to a blind conformational search on energy landscapes much more complicated than folding funnels for single proteins. The energy landscape for protein aggregation is not fully understood, but as such it should account for both soluble monomers and oligomers as well as the competition between native-contact-enabled folding and intermolecular non-native driving forces in crowded environments containing these species.[259,275] This necessitates coarse-grained (CG) peptide models that can reduce degrees of freedom and help probe deeper into long-timescale processes.[190,193,276]

CG studies of aggregation typically utilize polypeptides as model systems for self-assembly, since it is well established that amyloid formation is a very general phenomenon independent of sequence or even fold-type of the proteins involved across different proteopathies, such that fibril assembly may be governed by simple physiochemical properties of the peptide chain like charge distribution, permutations of hydrophobic and polar residues and β-sheet propensity.[277,278] For instance, Hall and co-workers developed and used the PRIME CG model to characterize the concentration and temperature dependendent amyloid formation in polyalanine monomers.[204] Dobson, Vendruscolo and co-workers utilized a tube-like CG representation of the peptide backbone supplemented by explicit hydrophobic, hydrogen-bond and excluded volume interactions to illustrate the transition from initial aggregation into disordered globular states to reordering into ordered fibrils in model polypeptides.[279] Caflisch and co-workers constructed an intermediate resolution CG model which modulated backbone dipoles through partial charges to describe a disordered ("amyloid-protected") and a β-sheet state ("amyloid-competent") and used it in simulations of model amphiphatic polypeptides to find that changes in β-sheet propensity can modulate the heterogeneity of aggregation pathways: amyloid protected monomers progressed through the formation of proto-fibrillar intermediates to give rise to less ordered assemblies while amyloid competent monomers could bypass intermediate aggregates and produce highly ordered amyloids.[280] In a similar study, Bellesia and Shea used a two-site CG model with an explicit dihedral term in the Hamiltonian for controlling the β-sheet propensity of monomers, to characterize the phase space of fibril polymorphs ranging from double and triple layered steric zippers to barrel-like formations as a function of temperature and the β-tendency. They also examined the order-disorder transition during amyloid formation using liquid structure theories of isotropic-nematic phase shifts.[281] Statistical mechanical theories using analytical expressions of the aggregation free energy have also been used to model order-disorder transition in protein aggregation.[282–284]

In this chapter, we characterize oligomeric templates built from valine polypeptides on the basis of their ability to self-propagate into amyloids, using a CG polyvaline model developed earlier in chapter 4. We paramterized the model in a bottom up fashion directly from an all-atom (AA) implicit water polyvaline MD simulation by minimzing the relative entropy. The relative entropy measures the likelihood that the of reproducing the AA microstate probability distribution using the CG model.[32] Minimizing the relative entropy automatically guarantees a maximum in this likelihood and consequently an optimal recapitulation of relevant thermophysical properties of the atomistic system.[35] Our choice of valine is motivated by its very high stability for β-sheets.[232] Further (atomistic and CG) polyvaline was found to exhibit considerable sampling of the space of backbone dihedral angles in folding simulations, by producing a large number of stable, compact folds that could be identified with native structures of known proteins of equivalent sequence length.[222,224] We study the abilities of oligomeric polyvaline templates (monomers and dimers) to form amyloids by inducing fibril formation in unstructured peptides and characterize the fibril shapes and stabilities in terms of the template shape.

We note a recent and closely similar study by Laio and co-workers, where they characterized a complex nucleation pathway for amyloid formation from a disordered aggregate using a model system of atomistic polyvaline monomers.[268] The authors performed self-assembly simulations directly from the dispersed state for polyvalines in different parallel and antiparallel β-sheet conformations and extracted an aggregation free-energy landscape that revealed that aggregation is initially dominated by antiparallel motifs but the eventual oligomer conformation rests on a balance between parallel and antiparallel structures. Another notable effort by Hall and co-workers investigated the amyloid core structures in different prion strains, specifically in terms of the parallel β sheet content.[270] Our work differs from these examples and others mentioned earlier in that we focus on the *connection between the template shape and the corresponding fibril conformation and stability*, and illustrate it with candidate shapes such as hairpins, steric zippers and zippers with three fold symmetry. Other computational efforts have not yet probed the relative self-replicating abilities of different template motifs beyond steric zippers and in some-cases have looked at the stabilities of a specific fibril conformation such as β solenoids.[285] To the best of our knowledge, an investigation of amyloidogenic signatures in template conformations and the consequent polymorphism in the emergent fibrils have not yet been undertaken. Thus, our work may provide more insight into the dependence of pathogenic conditions on monomer conformation such as those observed for the tau protein mentioned earlier.

5.2 Candidate Backbone Templates

In this work, we focus on four families of β sheets: steric zippers (Z), β hairpins (H), hairpin-like zippers (H') and β-amyloids with three-fold symmetry (BA).

Steric zippers are one of the most common motifs seen in large amyloid plaques, and so forms a natural choice for a template. In Fig. 5.2.1, the steric zippers in panel (a) are taken from structures reported through atomic resolution crystallography by Eisenberg and co-workers.[249] Z1, Z4 and Z7 are respectively the "Class 1", "Class 4" and "Class 7" structures, out of eight classes of zipper motifs reported in this study. Z1 is a tetramer of the GNNQQNY region of the yeast prion sup-35, with parallel in-register strands within a sheet with their same sides facing each other and both sheet edges facing in the same direction. This causes

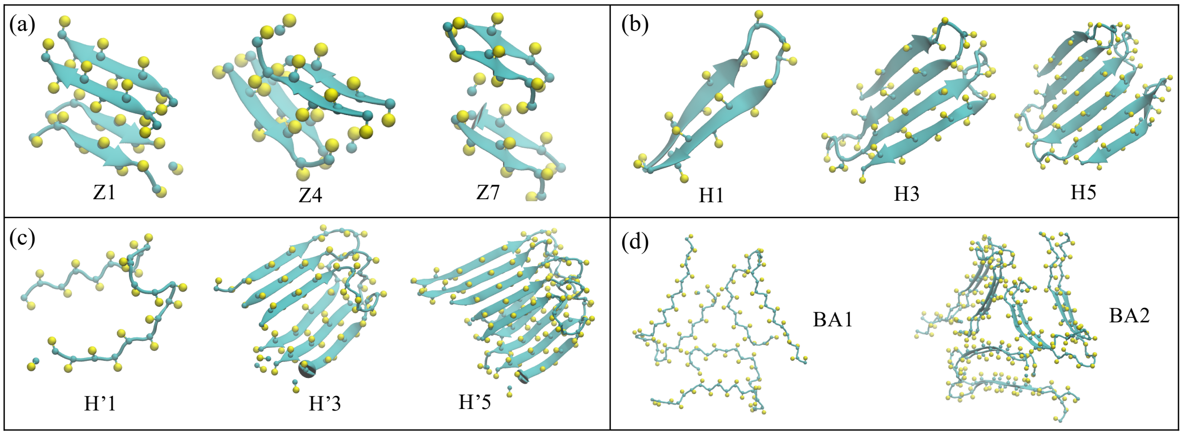


Figure 5.2.1: Backbone templates used in this work. Panel (a) shows steric zippers (Z) that are parallel (Z1, Z4) or antiparallel (Z7) within the same β sheet. Panel (b) shows hairpins H1, H3 and H5 with 1, 3 and 5 turns respectively. Templates in panel (c) are called "hairpin-like zippers" (H'1, H'2, H'3) in this work; they are basically zippers with a strand-arch-strand shape and (1, 3 and 5) flexible regions between the sheets. Panel (d) shows single (BA1) and double (BA2) sheet units with three-fold symmetry, from the Aβ_{1-40} protein involved in Alzheimer's disease. Monomers in each template have been "mapped" to polyvalines of equivalent length. CG valine sidechains (shown in yellow) interdigitate between the sheets.

each sheet to be antiparallel to its complementary sheet such that the sidechains between the sheets are staggered or interdigitated. The authors found Z1 to be the most abundant type of motif, covering $\sim 46\%$ of the total motifs isolated. Z4 is a tetramer of the GGVVIA fragment of the β-amyloid protein, with parallel inter-sheet strands that pack with opposite facing sides. Z7 corresponds to the tetramer from the VEALYL fragment of human insulin and consists of inter-sheet antiparallel β strands in register that pack with similar sides facing each other.

Hairpins H1, H3 and H5 in panel (b) form meanders and are named according to the number of hairpin turns between β strands. The motifs are taken from residues 480-500, 480-526 and 480-548 of the outer-surface protein A which is an immunogenic lipoprotein associated with Lyme disease causing bacteria.[286] Our choice of hairpins is motivated by the fact that top-cluster structures for both AA and CG polyvaline 15-mers examined in chapter 4 were hairpins. Shea and co-workers found that β-hairpins have a dual role in amyloid formation. They stabilize the sheet through inter-strand hydrogen bonds, but also destabilize through addition to a growing fibril just enough to present an unstructured backbone fragment that can serve as a latching point for solvated monomers to promote further self-assembly.[287] This destabilization effect is arguably important for fibril growth such that spatially restrained hairpins such as the templates H1, H3 and H5 may not directly spawn amyloid fibrils from a terminal strand.[287,288] Still, it is interesting to see whether these restrained hairpin templates can induce at least a single hairpin from the free monomers which can then carry on the amyloid growth process as described above.

Aggregates formed from hairpins have been typically observed to be not exactly hairpins themselves but rather adopt a strand-arch-strand architecture.[287] held together by hydrogen bonds between β strands in *distinct* sheets instead of laterally between strands within the same sheet. Further, formation of the flexible turn region in a hairpin is well known to be the rate limiting step in assembling the entire structure.[289,290] Motivated by these observations, we study a third family of templates (panel (c)) that we call hairpin-like zippers (H') as a control against true hairpins of the H family. These have the strand-arch-strand motif, with flexible U-shaped region between intra-sheet strands. H'1, H'3 and H'5 are respectively monomeric, trimeric and pentameric zippers (1, 3 and 5 flexible regions) corresponding to residues 6-28, 6-28 + 38-61 + 69-93, and 6-28 + 38-61 + 69-93 + 99-124 + 131-156 respectively of the Aβ_{1-40} fibril structure reported by Tycko et al.[291]

Finally, in panel (d), we consider templates with three-fold symmetry such that looking down the fibril axis presents a triangular cross-section. BA1 and BA2 represent one- and two-sheet zippers with three-fold symmetry taken from the amyloid structure of the full 40-residue Aβ protein, resolved through solid state NMR and electron microscopy techniques by Tycko and co-workers.[291] The authors found that solvated monomeric Aβ_{1-40} under quiescent conditions produce striated fibrils reminscent of simple steric zippers of the Z family in panel (a), while agitating the solution influences nucleation and fragmentation rates to stabilize a twisted zipper with a triangular cross section. Highly symmetric fibrils are arguably not governed by simple β sheet forming hydrogen bonds alone, but also includes intricate intra-sheet interactions necesary to stabilize the quarternary structure. So, it may be instructive to see if such complicated interactions can be self-propagated even in simple polypeptide assemblies, such that they may be rationalized in terms of fundamental peptide chain properties.

The net concentration of free peptides is maintained at 3 mM which is close to the in-vitro concentrations observed for Aβ_{1-42} *aggregates*.[292,293] While in-vivo and in-vitro critical aggregation concentrations of soluble amyloidogenic proteins like Aβ_{1-42} are typically in the order of nM,[294,295] such low concentrations would be prohibitively expensive for MD simulations even with our CG model. Further, the exact quantitative effects of concentration on aggregation behavior and fibril shape is not the focus of this study. Table 5.2.1 provides details of the different templates studied in this chapter. N_x and n_x are the number of monomers and number of valine residues per monomer, respectively, where $x = t$ for template or $x = f$ for free monomer.

Table 5.2.1: Nomenclature of templates and details of the simulation setup

Template	PDB code	Conc.	BoxL	N_t^*	n_t^{**}	N_f^*	n_f^{**}
Z1	2OMM	3 mM	164.20 Å	4	7	8	7
Z4	2ONV	3 mM	164.20 Å	4	7	6	7
Z7	2OMQ	3 mM	164.20 Å	4	7	6	7
H1	1OSP (res 480-500)	3 mM	164.20 Å	1	20	8	24
H3	1OSP (res 480-526)	3 mM	164.20 Å	1	46	8	24
H5	1OSP (res 480-548)	3 mM	164.20 Å	1	68	8	24
H'1	2LMP (res 6-28)	3 mM	164.20 Å	1	~ 24	8	24
H'3	2LMP (res 6-28, 38-61, 69-93)	3 mM	164.20 Å	3	~ 24	8	24
H'5	2LMP (res 6-28, 38-61, 69-93, 99-124, 131-156)	3 mM	164.20 Å	5	~ 24	8	24
BA1	2LMP (chains A, G, M)	3 mM	188.00 Å	3	40	12	30
BA2	2LMP (chains A, G, M and B, H, N)	3 mM	188.00 Å	6	40	12	30

5.3 Methods

In this work we focus on the ability of the template to induce conformation changes in free peptides. Thus, we restrain the template to a fixed point in space, while mobile polypeptide chains are dispersed around it to represent a disordered, solution-like state for the entire system. The objective is to see which templates can "recruit" the dispersed peptide chains from solution to coalesce into a strongly ordered fibrillar structure, and then characterize the emergent fibril shapes. Restraining the template effectively removes the possibility of changes in its conformation induced by the self-assembly process, although the restraints are designed to allow $\sim$ 1 Å fluctuation in template atomic positions. For the remainder of this chapter, we will refer to the entire assembly including the template and the mobile peptide chains as an oligomer, and individual polypeptide chains as monomers.

Each atomistic template is first "mutated" to an atomistic polyvaline structure with the same number of monomers (chains: N) and sequence length (number of valine residues: n) per monomer, by aligning polyvaline monomers of equivalent length to template monomers using the well known Kabsch algorithm,[246] followed by an additional rotation of all backbone dihedral angles along each monomer to match those in the template. The template is then relaxed through a short energy minimization. This energy minimized atomistic polyvaline structure represents the final template that is incorporated as a seed in the self-assembly simulation. Post energy minimization, the template is centered within a cubic simulation box with periodic boundary conditions, and all atoms on the template are tethered to their current positions through harmonic restraints. Free polyvaline chains are distributed randomly on the surface of a sphere inscribed within the simulation box, and centered on the template center-of-mass. As summarized later in Table 5.2.1, the number of free peptides and their initial structures as well as the box length are adjusted to be commensurate with the desired concentration, and to prevent unwanted interactions between peptide chains and their periodic images.

Finally, the entire assembled initial topology is mapped to a four site CG representation and MD simulations are launched using a CG forcefield described in the following section.

5.3.1 CG model

The CG polypeptide model we use here, was developed in chapter 4 from a reference AA system of a valine 15-mer simulated using the ff96 Amber forcefield[239] with a modified version of the igb5 implicit solvent model.[240] As shown in Fig. 5.3.1, this CG model reduces an atomistic amino acid to four CG sites based on the heavy atoms: a nitrogen (N) site, an α-carbon (C) site, a sidechain (S) site, and an oxygen (O) site which lumps together the carbonyl carbon and oxygen. The CG forcefield is represented using bonded (U_{b}), angular (U_θ), torsional ($U_{\varphi,\psi}$) and nonbonded pair-wise interactions (U_{pair}) that are intra-backbone (BB), inter-sidechain (SS) and inter-backbone-sidechain (BS) in nature:

$$U_{\mathrm{CG}} = \left(U_{\mathrm{b}}^{\mathrm{BB}} + U_{\theta}^{\mathrm{BB}} + U_{\varphi,\psi}^{\mathrm{BB}} + U_{\mathrm{pair}}^{\mathrm{BB}}\right) + \left(U_{\mathrm{b}}^{\mathrm{BS}} + U_{\theta}^{\mathrm{BS}} + U_{\varphi,\psi}^{\mathrm{BS}} + U_{\mathrm{pair}}^{\mathrm{BS}}\right) + U_{\mathrm{pair}}^{\mathrm{SS}} \quad (5.3.1)$$

Bond potentials are harmonic in nature while all other potentials are represented using cubic B-splines. All pair potentials are cut off at 10 Å. This forcefield consists of ~ 800 parameters, which are optimized by minimizing the relative entropy between the CG model and the reference atomistic polyvaline simulation. This

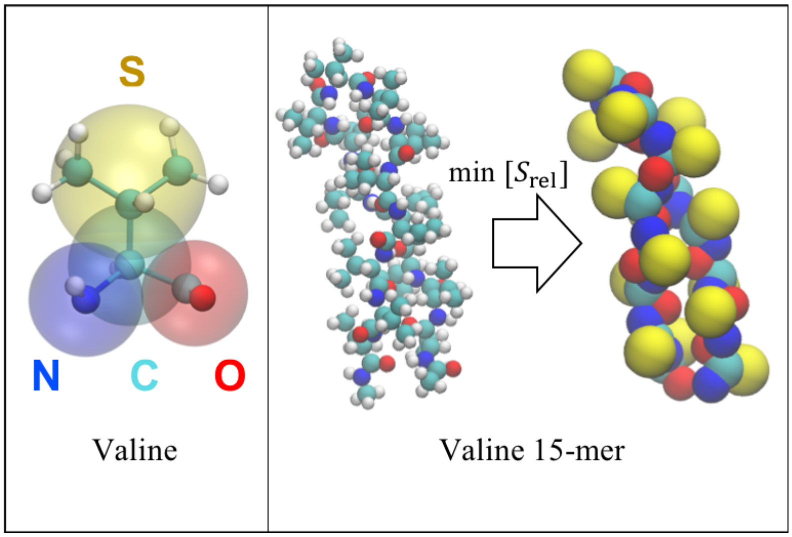


Figure 5.3.1: Left: valine is mapped to four heavy atom centers N, C, O, S that sit at the centers-of-mass of the amino, α-carbon, carbonyl carbon and oxygen, and the side-chain groups respectively. Right: The CG model of a 15-mer valine polypeptide is parameterized by minimizing the relative entropy from a reference atomistic simulation.

technique minimizes the information loss upon coarse graining and guarantees a maximal overlap between AA and CG microstate ensemble probability distributions, such that structural correlations (such as secondary structure propensities) may be adequately replicated in the CG model.[32,34] Further details can be found in chapter 4 and Refs. 34, 41. In chapter 4, AA and CG polyvaline models demonstrated high β-hairpin character which was stable across a wide range of temperatures from 270 K to at least 350 K.

5.3.2 Order parameters for assessing template stability

While the RMSD to a reference structure is a very useful metric for assessing structure prediction for the folding of single proteins, we do not have such references available for self-assembling oligomers examined in this chapter. Even if we can determine a putative reference, the RMSD may embed significant errors for homogeneous oligomers built completely from smilar polypeptides. A single homoegeneous polypeptide chain does not have directional preference and looks the same from both the N-terminal or the C-terminal end. While the Kabsch algorithm[246] for determining RMSD accounts for a bi-directional symmetry for single strands, a homogeneous oligomer may admit multidirectional degeneracy in shape, i.e., it may look visually similar from many different angles. Thus oligomers with similar fibril shape can have very different RMSDs from a single reference structure. Hence, in this chapter we employ different order parameters to assess oligomeric stabilty.

We calculate the "β content" (f_β) of an oligomer as the fraction of residues belonging to β sheets (both intra- and inter-strand). f_β encodes information about both inter-residue contacts as well as the propensity of backbone dihedral angles to lie in typical β rich regions on the Ramachandran plot. In general, determining if a CG residue contributes to a particular type of secondary structure is a non-trivial classification problem that may require searching for known secondary structure patterns in inter-residue contact maps and Ramachandran plots.[296] However, our four-site CG model has the advantage of being sufficiently close to an atomistic backbone resolution, such that the CG O site can be reverse-mapped to approximately predict the carbonyl carbon and oxygen. The presence of an amino center (N) and a carbonyl oxygen atom is sufficient for available bioinformatic algorithms to infer the strength of inter-residue hydrogen bonds along the backbone as well as dihedral angles, and accordingly classify residues into particular secondary structures. Again, while sophisticated algorithms exist for backmapping CG peptide structures of arbitrary resolution to full atomistic detail,[297–300] owing to the near-atomistic backbone resolution of our CG model, we use simple geometrical constraints, namely the planarity of the sp^2 hybridized carbonyl carbon, and the average bond length and bond angle around the peptide linkage. Further details can be found in the appendix.

To characterize fibrillar shapes, we examine an orientational order parameter that has been succesfully used in the literature for describing anisotropy of CG protein agglomerates.[281] The orientation order S is defined as the largest eigenvalue of the diagonalizable, second rank tensor:

$$Q_{ij} = \frac{1}{2N} \sum_{k=1}^{N} \left(3u_i^k u_j^k - \delta_{ij}\right) \tag{5.3.2}$$

where, $i, j = x, y, z$ are Cartesian axes, δ_{ij} is the Kronecker delta, and N is the total number of polypeptide monomers (including the template). $\mathbf{u}^k$ is the principal axis vector of the k^{th} monomer, given by $\mathbf{u}^k = \mathbf{r}_1^k - \mathbf{r}_n^k$, where $\mathbf{r}_m^k$ denotes the center of mass of the m^{th} residue within the k^{th} monomer and n is the sequence length of this monomer. Q_{ij} is commonly used as a measure of isotropic to nematic transitions in liquid crystal systems and can easily detect uniaxial order in a system.[301] S varies between zero for isotropic globular aggregates, to one for fibrillar oligomers with all monomers perfectly aligned along a preferred spatial direction. Trajectory-averaged values of S can be used to detect the temperature dependent order-disorder transition from compact fibrils to dispersed monomeric assemblies.[281] However, in this chapter, we use the joint distributions of f_β and S to cluster the ensemble of emergent oligomers, and identify the dominant structures.

5.3.3 Simulation details

Initial AA topologies are assembled using PACKMOL.[302] Angle, torsion and pair splined potentials in the CG polyvaline model use 40 knot points each, with densities of 1.4 °/knot, 0.22 °/knot and 0.25 Å/knot respectively. CG model simulations are run in the canonical ensemble with the LAMMPS MD engine,[105] using periodic boundary conditions. Particle positions are evolved in time using a Langevin dynamics integrator with a timestep of 1 fs and a damping coefficient of 10 ps^{-1}. We perform Replica Exchange Molecular Dynamics (REMD) which allows greater sampling of the conformation space,[237] using 10 replicas with an exponential temperature schedule between 260-400 K. Each replica is simulated for 1 μs, and data from the last 400 ns is used for calculating statistics. Replica swaps are attempted every 10 ps. Templates are spatially locked using harmonic restraints on each template CG site with a force constant of 1 $k_B T$ and mean distance of 1 Å. Secondary structures of reverse-mapped CG structures are calculated using the STRIDE algorithm.[303] Data from the different replicas is reweighted using the Multi-state Bennett Acceptance Ratio (MBAR) algorithm, and implemented using the pymbar package.[245] Top-cluster oligomers are determined at 280 K, using the K-Means clustering algorithm. Cartoon representations of secondary structures are rendered in VMD.[248]

5.4 Preliminary results and future objectives

Fig. 5.4.1 shows the free energy landscape of fibril formation AT 280 k, as a function of the orientational order parameter S and the β content f_β, for the antiparallel steric zipper Z1, hairpins H1, H3 and H5, and the single Aβ_{1-40} sheet with three-fold symmetry. The dominant structures obtained by clustering along S and f_β are shown in the bottom panels. Z1 induces exactly self-similar antiparallel zippers with a $\sim$ 70% β content and orientational order between 40-85%, although some unstructured agglomerates with a $\sim 6\ k_BT$ barrier are found occasionally. Hairpins H1, H3 and H5 template fibrils with β content progressively decreasing from 85 to 75%, while sampling lower orientational orders of 30% for H5 as opposed to 50-70% for H1. Interestingly the fibrils all pack like steric zippers, remaining antiparallel to the hairpin strands, such that the average orientational order for H motifs (50%) is somewhat close to that for Z1 (62%). It is likely that lower values of S for hairpin templates with more turns results from the twist in the induced zipper. For instance, fibrils H1 to H5 increasingly twist away from the plane of the template, lowering preference for the axis parallel to the hairpin plane and thus decreasing S. Fibrils induced by templates with higher turn numbers may have barrel forming propensities, but this needs to be verified with further simulations. The BA1 template exhibits two minima on the aggregation

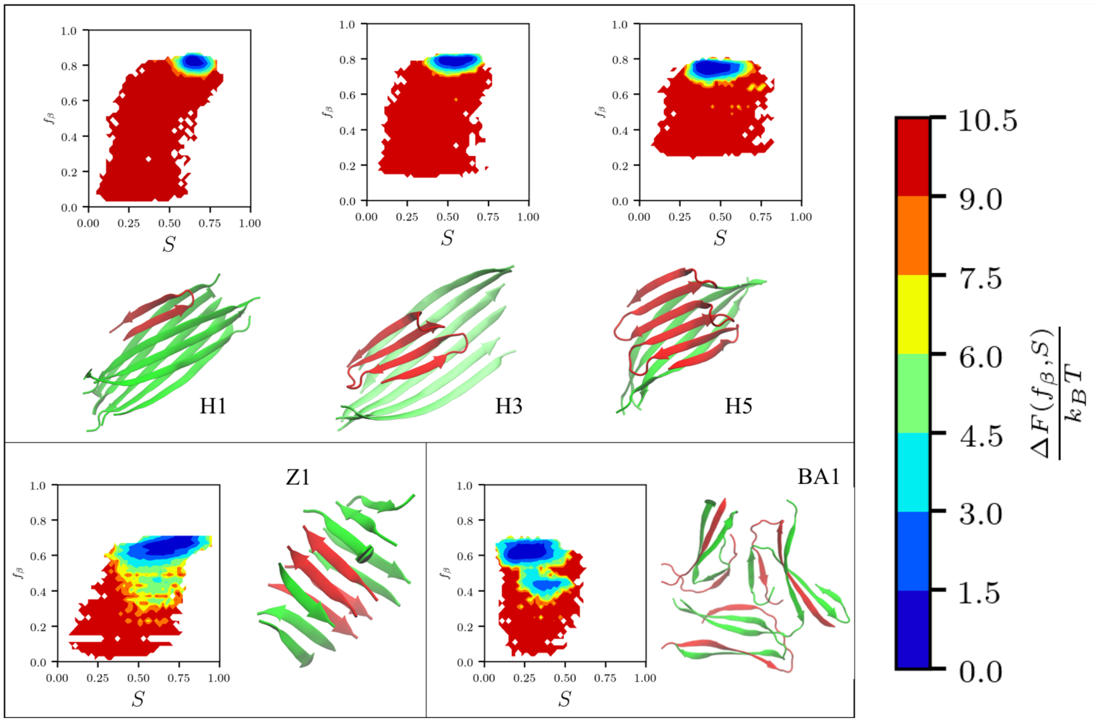


Figure 5.4.1: Aggregation free energy landscapes as a function of the orientational order parameter S (along the X axis) and the β content f_β along the Y axis for templates Z1, H1, H3 and H5, and BA1 as notated in Table 5.2.1. These two metrics are used to cluster the aggregates observed in the last 200 ns from 700 ns/replica of REMD simulation, which are shown with the free energy surfaces. Templates are colored in red and free peptides in green. The antiparallel steric zipper is seen to be very stable, since even hairpin templates lead to Z-like topologies.

free energy surface, characterized by f_β values of 60% and 40% respectively, and S ranging between 20-50% for the basin at higher f_β, and 30-55% for the one at lower f_β. Low orientation order is to be expected for a system with three-fold symmetry but it is unclear at the moment if this symmetry can enforce three states with distinct S values or if two of those states are nearly similar in free energy to each other. Further simulations are required to verify this. Note that BA1 enforces nearly self-similar assembly where the free peptides line up to form antiparallel zippers at each arm of the triangle. However, in this preliminary simulation, BA1 was locked spatially using a low harmonic restraint with a force constant of 1 $k_B T/\text{Å}^2$, so that the sheets in each arm have been somewhat "pulled out" of their β conformations. We need to rerun this simulation using stricter restraints.

Fig. 5.4.2 compares the folding curves by plotting the (replica reweighted) β content with temperature for the templates Z1, H1, H3 and H5, and BA1. The

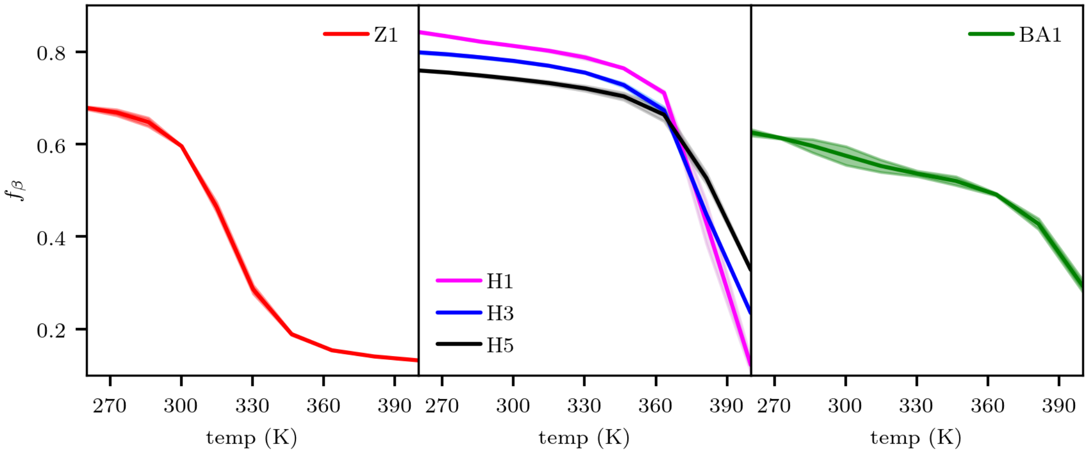


Figure 5.4.2: Folding curves illustrating the temperature dependence of the β content, f_β (along the Y axis) for templates Z1 (left), H1, H3 and H5 (middle) and BA1 (right). Hairpin induced fibrils appear to be the most stable with folding temperatures close to 380 K.

steric zipper Z1 has the lowest folding temperature ∼ 320 K, among all the templates studied so far, while BA1 has a moderately higher folding temperature of

345 K. All three hairpins however induce zippers that are tremendously stable with a folding temperature of $\sim$ 380 K for all three cases H1, H3 and H5. Interstingly, the folding curves for all three hairpins indicate a sharp cusp between 360-380 K, and the location of the cusp shifts to higher values while going from H1 to H5. Is it possible that the aggregation process for these templates can be represented with simple two-state models? One way to find out would be to fit a simple van't Hoff like relation derived from mass-action kinetic considerations.

Appendix

5.A Reverse mapping the CG O site to the AA carbonyl group

As discussed in section 5.3.2, the CG O site needs to be reverse mapped to the full AA carbonyl group, so that STRIDE can infer secondary structures from this higher resolution representation. We do this using simple geometry arguments. Consider the peptide bond between any two residues in semi-atomistic resolution, shown in Fig. 5.A.1 N and C_{alpha} are the usual CG N and C sites (of adjacent residues) respectively, while $\tilde{C}$ and $\tilde{O}$ are carbonyl group atoms which were coarse grained into the CG O site, shown here schematically with a bounding ellipse. Let $\mathbf{r}_{C_\alpha}, \mathbf{r}_O, \mathbf{r}_N$ be the position vectors of the CG sites which *we know* and $\mathbf{r}_{\tilde{C}}, \mathbf{r}_{\tilde{O}}$

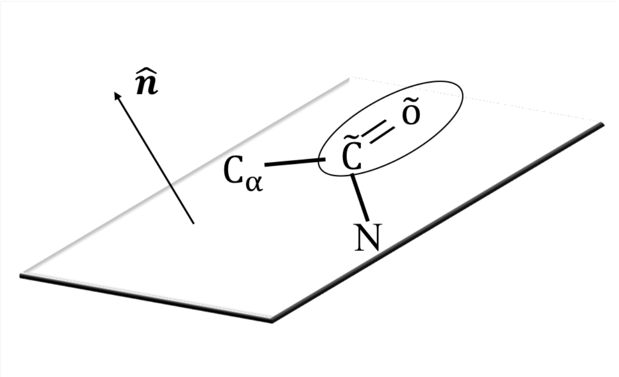


Figure 5.A.1: Peptide bond plane in semi-atomistic resolution. N and C_α are the CG N and C sites (of adjacent residues) respectively, while $\tilde{C}$ and $\tilde{O}$ are reverse mapped atoms of the carbonyl group. The CG O site sits at the center of mass of $\tilde{C}$ and $\tilde{O}$. $\hat{\mathbf{n}}$ represents the outward normal from the plane.

be those of the carbonyl group atoms that we have to determine. Center of mass coarse graining gives us the relation:

$$\mathbf{r}_O = \frac{(m_C\,\mathbf{r}_{\tilde{C}} + m_O\,\mathbf{r}_{\tilde{O}})}{m_C + m_O} \tag{5.A.1}$$

where m_C and m_O are the masses of carbon and oxygen respectively. Since $\mathbf{r}_O$ is known, the only remaining task is to determine either of $\mathbf{r}_{\tilde{C}}$ or $\mathbf{r}_{\tilde{O}}$. Here we evaulate $\mathbf{r}_{\tilde{C}}$. If $d_{C_\alpha \tilde{C}}$ represents the bond length between the alpha-carbon and the carbonyl group, then our task reduces to simply finding the vector $\mathbf{r}_{C_\alpha \tilde{C}}$, so that:

$$\mathbf{r}_{\tilde{C}} = \mathbf{r}_{C_\alpha} + d_{C_\alpha \tilde{C}}\,\frac{\mathbf{r}_{C_\alpha \tilde{C}}}{||\mathbf{r}_{C_\alpha \tilde{C}}||} \tag{5.A.2}$$

The outward normal to the peptide bond plane in terms of known vectors can be written using the cross-product as:

$$\hat{\mathbf{n}} = \frac{(\mathbf{r}_{C_\alpha} - \mathbf{r}_O) \times (\mathbf{r}_N - \mathbf{r}_O)}{||\ (\mathbf{r}_{C_\alpha} - \mathbf{r}_O) \times (\mathbf{r}_N - \mathbf{r}_O)\ ||}$$

while, $\angle NC_\alpha\tilde{C}$ can be evaulated usng the sine rule as:

$$\angle NC_\alpha\tilde{C} = \sin^{-1}\left(\frac{d_{\tilde{C}N}}{||\mathbf{r}_N - \mathbf{r}_{C_\alpha}||}\ \sin\angle C_\alpha\tilde{C}N\right)$$

Here, $d_{\tilde{C}N}$ and $\angle C_\alpha\tilde{C}N$ are the bond length and angle respectively around the peptide bond which are typically conserved for protein chains (plus our CG model is already near AA resolution, so we don't expect these quantities to change significantly) and hence known quantities, but can also be determined from their AA distributions. For the polyvaline CG model developed in chapter 4, we find $d_{\tilde{C}N} = 1.32$Å and $\angle C_\alpha\tilde{C}N = 114°$. Finally, then evaulating $\mathbf{r}_{C_\alpha\tilde{C}}$ amounts to rotating the vector $\mathbf{r}_N - \mathbf{r}_{C_\alpha}$ in the peptide bond plane by $\angle NC_\alpha\tilde{C}$about the normal axis $\hat{\mathbf{n}}$, which can be achieved using the well-known Rodrigues' rotation formula:

$$\begin{aligned}\mathbf{r}_{C_\alpha\tilde{C}} = (\mathbf{r}_N - \mathbf{r}_{C_\alpha})\ \cos\angle NC_\alpha\tilde{C}\ \ +\ \ [(\mathbf{r}_N - \mathbf{r}_{C_\alpha}) \times \hat{\mathbf{n}}]\ \sin\angle NC_\alpha\tilde{C} \\ +\ \hat{\mathbf{n}}\ [\hat{\mathbf{n}} \cdot (\mathbf{r}_N - \mathbf{r}_{C_\alpha})]\ (1 - \cos\angle NC_\alpha\tilde{C})\end{aligned}$$

Once, $\mathbf{r}_{C_\alpha\tilde{C}}$ is determined, it is straightforward to apply Eqs. (5.A.1) and (5.A.2) to fully determine the carbonyl group at a near-atomistic resolution.

Chapter 6

Conclusion

6.1 Implications

In chapter 2, we introduced local density (LD) potentials as a simple and computationally fast strategy to incorporate manybody effects into CG models, which may improve transferability. Traditionally CG models have been built in the image of atomistic systems. This approach has traded off the inherent multibody nature of CG interactions due to coupling between the CG degrees of freedom, for computational speed, by approximating the manybody CG potential of mean force with pairwise nonbonded potentials. Inspired from mean-field electronic theories of metals,[57,74] LD potentials aim to build back such missing manybody interactions through mean-field potentials that account for the energetic contribution to a CG site from its neighboring sites within a threshold distance. We demonstrated that Shell's relative entropy framework[32,34] enables a systematic parameterization of LD interactions from reference atomistic simulations. Subsequently, we investigated the utility of LD potentials in the development of implicit water models for two candidate systems characterized by hydrophobic interactions: hydrophobic collapse of an alkane-like superhydrophobic polymer and assembly of superhydrophobic methane sized partices. Our results in chapter 2 that in both of these test systems, augmenting traditional CG pair interactions with LD potentials generally improved the ability of water-free CG models to capture the co-operativity of water-mediated interactions in the corresponding explicit-water atomistic simulations. Thus CG models assisted by LD potentials could replicate polymer folding or (superhydrophobic) methane clustering nearly quantitatively, as opposed to CG models built entirely from pair interactions. More importantly, these improved CG models were transferable across a wide range of polymer lengths and methane bulk densities.

Our results from chapter 2 revealed that LD potentials may be promising tools for constructing computationally fast and accurate CG models of phase transitions, which necessitate proper transferability to span across phases with widely different local structure. Accordingly, in chapter 3 we extended the pevious work on LD potentials to develop CG models of liquid mixtures which are transferable in the space of compositions. Specifically, we used relative entropy minimization to parameterize LD-interactions-assited CG models of benzene-water mixtures, mapping benzene and water to single CG sites. In combination with the usual intra- and inter-species pair potentials between benzene and water we tested the relative capabilities of intra- and inter-species LD potentials to capture the orientational degrees of freedom afforded by benzene-water and water-water hydrogen bonds which are primarily responsible for the macroscopic benzene-water phase split. Our CG model that combined all four possible LD potentials between benzene and water with the pair interactions showed improved structural transferability (pair correlations and local co-ordination numbers) over pair-only models across a wide range of benzene mole fractions and was relatively agnostic to the benzene composition at which it was parameterized. Even when parameterized from relatively small system sizes, this CG model exhibited macroscopic phase segregation in larger systems while quantitatively predicting the location of the interface. We also learnt that the intra-water LD potential was the dominant multibody interaction in the system, thus highlighting the role of water's distinct structural correlations and tetrahedral network in mediating self and cross interactions with much larger asymmetric species like benzene which has been challenging to capture in CG models so far.

Our investigations in chapters 2 and 3, revealed an important requirement for the role of LD potentials to be meaningful, namely sufficiently high multi-body correlations in the physics of the system under consideration. E.g. in chapter 2, when the polymer and/or the methane-like particles were made somewhat less hydrophobic by using attractive intra-particle Lennard Jones interactions commensurate with those of true methane, the LD-augmented CG models presented no significant improvement over pure pair interaction based models. Similarly, in chapter 3, we noted the relative dominance of intra-water LD interactions over intra-benzene and even benzene-water interactions. Benzene does not hydrogen bond appreciably with itself thereby decreasing the overall manybody character of self interactions. Thus CG models both with and without LD potentials reproduce the bulk benzene radial distribution functions and co-ordination number distribution equally well. Based on these results, we recommend a prior-assessment of the relative strength of multibody interactions in the physics of a system before using parameterizing CG models for such systems that include LD potentials.

In chapter 4, we employed relative entropy minimization to construct four-site CG models of polypeptides (four CG sites for amino, α-carbon, carbonyl and side chain groups) that can serve as putative backbone forcefields, and subsequently we added native-centric sidechain interactions to construct Gōmodels for MD simulations of protein folding. This work extended the Shell and Carmichael's previous parameterization of very accurate CG poly-alanine models[41] to heteropeptides with arbitrary sequence complexity. Specifically, we developed standalone CG models from polymers of leucine and valine by minimizing relative entropy from atomistic polyleucine and polyvaline simulation. Later, we parameterized a CG model simultaneously from both of these atomistic references by minimizing the sum of relative entropies with each reference simulation, following the extended-ensemble protocol of Noid and co-workers.[63] The extended ensemble approach was able to combine the α-helix and β-hairpin propensities of polyleucine and polyvaline respectively, into a hybrid model that correctly addressed the balance between α and β behavior. Simplistic native-centric sidechain potentials added to the hybrid backbone resulted in a Gō model that could resolve native structures of both short sequences and globular proteins to within 2 Å. While Gō models require the native structure of the target protein as input and are therefore not entirely *predictive*, the role of the Gō-like sidechain interactions in this work was somewhat passive, serving as a test for the quality of the CG backbone forcefield. The results in chapter 4 demonstrate that four site CG polypeptide models optimized with the relative entropy are promising tools for building high quality protein backbones, which can be ultimately refined using chemistry-based sidechain interactions for different amino acids, and employed in folding and self-assembly simulations.

In chapter 5, we performed CG MD simulations of templated protein aggregation commonly seen in neuro-degenerative diseases, to elucidate the connections between template topology and self-propagating ability during oligomerization. Specifically, we used a β-rich CG polyvaline model developed in chapter **??** to simulate the assembly of polyvaline strands around a given template or *seed* structure (also built entirely from polyvaline units). Our preliminary results show that antiparallel steric-zipper like motifs are exceptionally stable across a wide range of temperatures and can be induced even by non-zipper like motifs such as hairpins. Further work is necessary to study a broader range of template conformations to distinguish templates that can replicate nearly self-similarly from those that lead to ordered but self-dissimlar aggregates.

The methods developed in this thesis are broadly useful for developing new and powerful CG models for studying important phenomena in chemical engineering like hydrophobic interactions and phase transitions in liquids on one hand, and protein folding and self-assembly on the other. The accuracy afforded by CG models developed in the various chapters go beyond simple academic investiagations into the driving forces and may be implemented (with appropriate computational resources) on large-scale systems to make structurally and thermodynamically accurate *predictions*, which can supplement and guide experimental efforts.

6.2 Improvements in numerical algorithms

A major outcome from this thesis has been a robust relative entropy based coarse-graining algorithm with the ability to handle relatively large numbers of force field parameters and, importantly, to include arbitrary force field types. This lets us throw a large variety of CG interactions into the mix, develop parallel families of force fields for the same system, and gauge the relative significance of one or more of these interactions over the others. The relative-entropy optimization package is hosted on a private GitHub account and can be accessed through requests to the author of this thesis or Prof. M. Scott Shell at University of California Santa Barbara. The key features incorporated into the group-software as a result of projects addressed in this thesis are:

6.2.1 LAMMPS LD potential

The LD potential was added as a separate manybody potential to the MD software LAMMPS,[105] and we are currently in the process of submitting it to the LAMMPS repository. This not only enables fast CG MD simulations with the LD potential, but also allows the construction of CG models from large systems, where the most computationally expensive steps are on-the-fly CG MD simulations with the LD potential, that are necessary for calculating successive iterates of the relative entropy during the course of minimization.

6.2.2 Optimization convergence

We briefly discuss two approaches that were necessary to ensure a smoother convergence of the relative-entropy optimization, especially when working with non-standard function spaces like the local density, or handling very large parameter sets such as in the CG peptide models.

Treatment of inner-core regions in nonbonded potentials

We found that the treatment of the inner core region of spline-based pair potentials, where the pair distance approaches zero, requires careful conditioning to optimize. In this regime, the potential should grow to a large value to prevent core overlaps, but at the same time there is little to no sampling of the corresponding distances in the reference atomistic trajectory and so there is sparse data to dictate "how large" the potential should be and what form. If we did not treat this issue at all, potentials in our algorithm varied wildly in this regime, and sometimes manifested unphysical minima at $r = 0$. Our initial treatment was to simply constrain the inner core region to have a fixed negative slope, but this led to artifacts for complex systems. Fig. 6.2.1 gives an example with benzene-benzene and benzene-water CG pair potentials. These are preliminary (and suboptimal) versions of the potentials presented in section 3.3.1. Over-constraining the inner-core part of the potential by fixing the slope lead to artifacts in the first shell distribution / correlation function, which are arguably important to reproduce

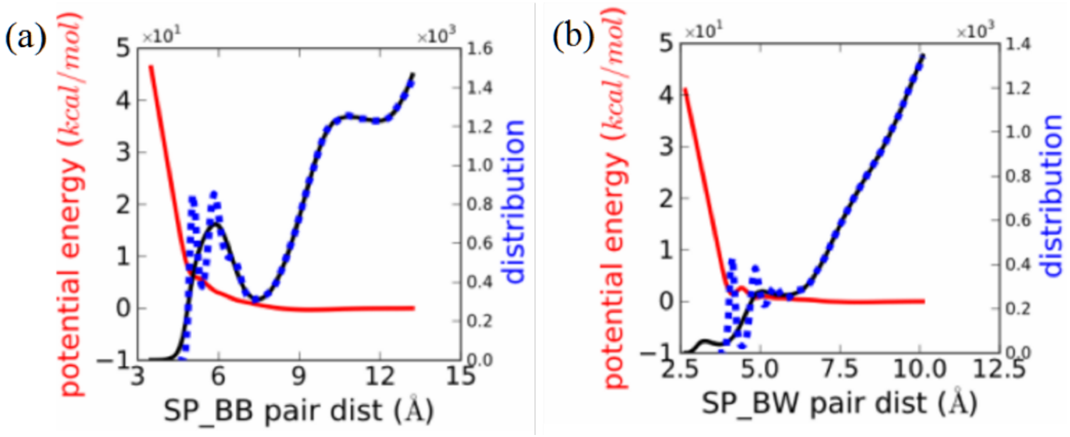


Figure 6.2.1: Artifacts for our early approach to determining pair spline potential "inner core" regimes where the pair distance approaches zero. Pair potentials are shown in red, the corresponding all-atom distance distribution is in black, and the CG distribution is dotted blue. Spline pair potentials that constrain the slope in the inner core region lead to artifacts near the first distribution peak, as shown in both cases. The solution has been a staged optimization procedure.

quantitatively due to the influence of local structure on thermodynamic properties. Our solution, which now works reliably, was to stage the optimization in several successive rounds: fixed inner slope, followed by slope variation during the minimization, followed by full unconstrained relaxation.

Incorporating memory effects

The relative entropy minimization algorithm requires an estimate of the relative entropy at every optimization step. This is achieved by running MD simulations with current estimates of the CG forcefield and exploiting Eq. (**??**) to calculate the relative entropy from simulation observables. Trial MD simulations are the indeed most computationally intensive part of the overall optimization algorithm. Carmichael and Shell improved on the overall computational efficiency and reduced inherent stochasticity in the algorithm by reusing old trial CG simulations through reweighting techniques.[34] Specifically, they ran a trial MD simulation with guess CG model parameters $\boldsymbol{\lambda}^0$, following which, instead of directly minimizing the relative entropy S_{rel}, they addressed the quantity:

$$\begin{aligned}\Delta S_{\mathrm{rel}} &= S_{\mathrm{rel}}(\boldsymbol{\lambda}) - S_{\mathrm{rel}}(\boldsymbol{\lambda}^0)\\ &= \beta\big\langle \Delta U_{\mathrm{CG}}\big\rangle_{\mathrm{AA}} + \log\big\langle \exp\left(\beta U_{\mathrm{CG}}\right)\big\rangle_{\mathrm{CG},\boldsymbol{\lambda}^0}\end{aligned} \tag{6.2.1}$$

where, $\Delta U_{\text{CG}} = U_{\text{CG},\boldsymbol{\lambda}^0} - U_{\text{CG},\boldsymbol{\lambda}}$ is the difference in potential energies between the trial MD trajectory and its reweighted version at new CG forcefield parameters $\boldsymbol{\lambda}$. Eq. (6.1) employs the classic Zwanzig free energy perturbation[304] to reweight snapshots from a trial MD simulation with CG forcefield parameters $\boldsymbol{\lambda}$ to updated parameters $\boldsymbol{\lambda}$. Minimizing ΔS_{rel} directly leads to fewer MD simulations and thus reduces stochasticity. However, free energy perturbation is valid only when the ensemble probability distributions of the target state ($U_{\text{CG},\boldsymbol{\lambda}}$) and the reference state ($U_{\text{CG},\boldsymbol{\lambda}^0}$) have sufficient overlap. Shell and Carmichael developed a simple metric to monitor how far the target strays from the reference and launch a new trial MD simulation with the CG forcefield $U_{\text{CG},\boldsymbol{\lambda}}$ when the target has moved too far. They monitored the effective fractional number of frames from the reference trajectory that contribute to reweighting.[34]

Launching a new trial MD simulation at any point during the optimization effectively destroys the memory of the path taken so far on the relative entropy hypersurface. For instance, prior to the trial simulation, we could be at a local minima. Using observables (potential energy, free energy differences) solely from the new trajectory will not "remember" this information and consequently provide an effective reset to the algorithm, which may cause it to spend more time in the local minima instead of moving downhill faster. To alleviate this issue and make the convergence process free of resets, we combine both the old and new trajectories into a hybrid timeseries that may retain long-term memory of the iteration process hitherto. This combination is acheived by calculating the effective cross configurational weights between the two trajectories using Bennett's algorithm[109] and gluing together frames from each trajectory with probability proportional to its weight.

6.3 Future directions

In this section, we make a note of future goals, both general objectives for bottom-up coarse graining with a focus on the role of the relative entropy, as well as specific future deliverables for the projects discussed in chapters 2-5.

6.3.1 Assessing transferability *a-priori*

This thesis demonstrates two important techniques to improve transferability of CG models across thermodynamic state points (density, temperature, etc.): the LD potential to account for manybody effects inherent in CG models (chapters 2 and 3) and the extended ensemble method for combining information from reference atomistic simulations at different state points or even with different chemistries[63] (chapter 4). In spite of these efforts, the general question of determing the regime of transferability for a given CG model *a-priori* i.e. assess transferability without the need to run CG MD simulations at different states and compare with their atomistic counterparts, still remains an open problem. One suggestion to tacke this problem through the relative entropy formalism is to include the thermodynamic state point directly into the CG Hamiltonian. Consider a CG Hamiltonian $U_{\mathrm{CG}}(\boldsymbol{\lambda}, \mathbf{p})$ as a function of forcefield parameters $\boldsymbol{\lambda}$ and other parameters $\mathbf{p}$ which encode state point information. This can be the thermodynamic state or even other relevant CG model parameters. The derivative of the relative entropy in this parameter space can be written as:

$$\left.\frac{dS_{\mathrm{rel}}}{d\mathbf{p}}\right|_{\boldsymbol{\lambda}^*} = \left.\frac{\partial S_{\mathrm{rel}}}{\partial \boldsymbol{\lambda}}\right|_{\boldsymbol{\lambda}^*} \frac{\partial \boldsymbol{\lambda}}{\partial \mathbf{p}} + \frac{\partial S_{\mathrm{rel}}}{\partial \mathbf{p}} = \frac{\partial S_{\mathrm{rel}}}{\partial \mathbf{p}} \tag{6.3.1}$$

Here, we project the multidimensional S_{rel} surface along the $\mathbf{p}$ co-ordinate, and $\boldsymbol{\lambda}^* = \boldsymbol{\lambda}^*(\mathbf{p})$ describes the optimal CG forcefield at a given state $\mathbf{p}$. This allows us to write $\left(\partial S_{\mathrm{rel}}/d\boldsymbol{\lambda}\right)_{\boldsymbol{\lambda}^*} = 0$ in the second step of Eq. (6.3.1). Specializing to the canonical ensemble gives us:

$$\begin{aligned} \left.\frac{dS_{\mathrm{rel}}}{d\mathbf{p}}\right|_{\boldsymbol{\lambda}^*} &= \beta\left\langle \frac{\partial U_{\mathrm{CG}}}{\partial \mathbf{p}} \right\rangle_{\mathrm{AA}} - \beta\left\langle \frac{\partial U_{\mathrm{CG}}}{\partial \mathbf{p}} \right\rangle_{\mathrm{CG}} \\ \text{or, } \Delta S_{\mathrm{rel}} &= \int_{\mathbf{p_0}}^{\mathbf{p}} d\mathbf{p}\; \beta \left[\left\langle \frac{\partial U_{\mathrm{CG}}}{\partial \mathbf{p}} \right\rangle_{\mathrm{AA}} - \left\langle \frac{\partial U_{\mathrm{CG}}}{\partial \mathbf{p}} \right\rangle_{\mathrm{CG}} \right] \end{aligned} \tag{6.3.2}$$

where $\langle \bullet \rangle_{\mathrm{AA}}$ and $\langle \bullet \rangle_{\mathrm{CG}}$ denotes ensemble averages for the AA or atomistic and CG models and β represents the inverse temperature. ΔS_{rel} is the change in relative entropy when transfering from a model parameterized at state $\mathbf{p_0}$ to state $\mathbf{p}$, while moving along the path $\boldsymbol{\lambda}^* = \boldsymbol{\lambda}^*(\mathbf{p})$. Thus, if the CG Hamiltonian is separable in $\boldsymbol{\lambda}$ and $\mathbf{p}$, Eq. (6.3.1) can be used to readily calculate the relative entropy change while transferring across state space. The states with high ΔS_{rel} can be screened as states at which the CG model would have limited or no transferability.

6.3.2 Including experimental information

In chapters 2 - 4, we have seen that thermodynamic properties like interfacial tension or folding temperature can be reproduced only with limited accuracy, not only at state points or system chemistries different from the parameterization reference, but also at the reference conditions. Structural properties like pair correlations, co-ordination number distribution, radii of gyration and end-to-end distances along polymer chains, solvent accessible surface, etc. are captured much more accurately. We suggest including thermodynamic metrics from experiments or ensemble averages from AA simulations in the CG model *prior* to parameterization, to capture such properties better. Voth and co-workers introduced the Experiment-Directed-Simulation (EDS) scheme to include experimental information in CG models by designing a correction function to the CG Hamiltonian $f(\mathbf{r})$ such that its trajectory average from CG MD simulations $\langle f(\mathbf{r}) \rangle$ is direct experimental data or available from atomistic MD simulations.[305] We suggest a somewhat similar approach in which corrections based on available data can be built directly into the CG optimization objective, i.e. the relative entropy, for instance:

$$O(\boldsymbol{\xi}) = S_{\text{rel}} + \sum_i c_i \left(\xi_i(\mathbf{r}) - \xi_i^0\right)^2 \tag{6.3.3}$$

Here, the overall optimization objective $O(\boldsymbol{\xi})$ corrects the usual relative entropy S_{rel} with quadratic constraints for each of the desired metrics ξ_i using a set of coefficients c_i. These constraints represent penalties due to deviation in these properties from their known experimental (or AA average) values ξ_i^0. Using quadratic constraints helps in avoiding non-convexity of the objective function surface. Optimizing the CG forcefield parameters and the penalty coefficients c_i simultaneously, in principle, should produce a CG model that faithfully recapitulates at least the metrics $\{\boldsymbol{\xi}\}$. Thus, one should be careful in choosing these metrics and ensure that they have little to no correlation between themselves, to prevent overfitting.

6.3.3 Determining the optimal AA → CG mapping

An open problem in bottom-up coarse graining is to design the most efficient AA to CG mapping that not only captures all the relevant physics of the phenomena under study, but also provides the simplest possible description of the AA system. An immediately obvious way to approach this problem is through the mapping entropy S_{map} (Eq. (2.13)) which is basically an AA ensemble average of the degeneracy of the AA to CG mapping, i.e. an average estimate of the number of

AA configurations that map to a single CG configuration. S_{map} is independent of the CG forcefield and depends only on the mapping function $\mathbf{M}(\mathbf{r})$ that translates the co-ordinates $\mathbf{r}$ of a group of AA particles to a CG particle position. Therefore, minimizing S_{map} with respect to $\mathbf{M}(\mathbf{r})$ seems like a promising approach to systematically estimate the most efficient AA to CG mapping. Foley, Noid and Shell estimated a modified form of the mapping entropy as function of CG resolution by coarse graining a set of single-domain proteins to Gaussian Network Models (GNMs).[306] Their work revealed a rather counter-intuitive result: the information retained per site in translating from the AA to the CG, had a *maximum* when plotted as a function of the number of CG sites, i.e. the CG resolution. This work suggests that there may be an optimal CG resolution for a particular atomistic system that is different and possibly lesser from the maximum possible resolution, i.e. a 1:1 mapping. Of course, a putative mapping function still needs to incorporate some chemical information and as such will always be system-specific. But at least, we can systematically estimate optimal mapping functions for families of compounds, namely alkanes, aromatics, polar groups, proteins, etc, by targeting the mapping entropy. One important problem to solve before we reach that stage, is to calculate or numerically approximate the mapping entropy for CG models with arbitrarily complex mapping functions beyond simplistic models like the GNM, where the mapping entropy can be calculated analytically.

6.3.4 A functional form for the LD potential

The LD potentials in chapters 2 and 3 are practically represented as cubic B-splines, whose knot points are parameterized through relative entropy minimization. While the robustness of the minimization algorithm enables highly flexible splines with several knot points, it is still desirable to develop a (semi) analytical expression for the LD potential function to gain deeper physical insight into how it embeds manybody correlations in CG models. More importantly, such an expression may enable us to connect the LD potential with the bulk density and ultimately with the extent of manybody correlations that the system admits. This may provide *a-priori* estimates of efficiency for the LD potential. As mentioned earlier, the LD potential is inspired from the theory of metallic bond-order potentials.[57,74] This theory accounts for mean electronic density $\varphi(\mathbf{r})$ around metallic nucleii through an "embedding function" $F(\varphi)$, similar to how the LD potential function accounts for local density ρ of neighboring sites around a central CG site through the LD potential function $f(\rho)$ (see section 2.2.1). So, it may be suggestive to follow Finnis and Sinclair's work where they examined the special case of bond-order potentials for solid transition metals with regular lattice spacings to approximate the embedding function F as:[60]

$$F(\varphi) \approx \sqrt{\varphi} \tag{6.3.4}$$

Approximations for the LD potential functions for liquids in terms of the local density might be constructed by translating Finnis and Sinclair's procedure to liquid structure and closely examining two-body or three-body correlations, in particular.

6.3.5 Refinements to the CG peptide model

The Gō models developed in chapter 4 represent the first step in developing CG peptide models for protein folding and self-assembly. Gōmodels require the native structure of the target sequence as an input and as such serve as a test for the quality of the CG peptide backbones developed in chapter 4 rather than demonstrating structure predictive abilities for arbitrary sequences with unknown / un-determined native structure. The next steps in this line of CG model development is introducing sidechain interactions based on sequence chemistry, perhaps a four alphabet model that classifies residues as hydrophilic, hydrophobic, cationic and anionic. The ultimate goal in this cascade is a full alphabet model with separate sidechain interaction parameters for all twenty amino acids. Further, it is necessary to account for glycine and proline through special bond-angle and dihedral interaction parameters in the backbone model, and eventually parameterize such backbones from a carefully designed extended ensemble of reference simulations of relevant polypeptides. E.g. special backbone parameters for glycine and proline can be readily parameterized by including polyglycine and polyproline atomistic data in the reference extended ensemble.

6.3.6 Secondary structure classification for CG peptides

In chapter 5, we used the STRIDE algorithm[303] to calculate oligomeric β content. STRIDE needs a fully resolved carbonyl group to estimate the possibility of a hydrogen bond linking the amino and carbonyl groups. This required backmapping the CG oxygen site in our four-site CG peptide models to predict approximate locations of the carbonyl carbon and oxygen. The reverse mapping was performed using simple geometrical arguments such as planarity of the sp^2 carbonyl carbon and mean carbon-oxygen double bond length. While STRIDE is very accurate, the reverse-mapping step introduces approximation errors that may increase with CG mapping resolutions lower than four-site (in spite of sophisticated reverse-mapping algorithms[297–300]). This necessitates a secondary structure classification algorithm for CG peptide residues, perhaps based on combined information from contact maps and Ramachandran-like plots of dihedral angle distributions.

One possible classification algorithm is to mine contact maps for patterns. As demonstrated in Fig. 6.3.1, two or more contact pairs (i.e. residue pairs in contact) can be classified as nearest neighbors along particular "paths" or

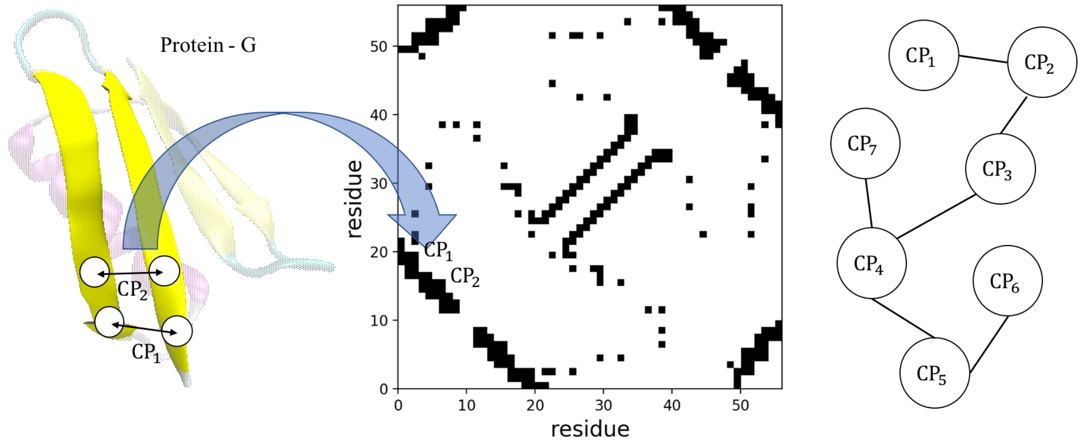


Figure 6.3.1: Schematic for β sheet classification in CG peptides using protein-G as an example. Contact pairs $\mathrm{CP}_1, \mathrm{CP}_2$, which lie along a β sheet can be enumerated using simple contact-map based conditions such as continuous paths away from the main diagonal (either parallel or orthogonal to the main diagonal). Nearest neighbors in the space of contact pairs can be ascertained from the Euclidean distance between contact pair co-ordinates in the contact map. Nearest neighbor contact pairs can be represented as connected vertices of a graph. Classifying a residue as part of a β sheet then reduces to computing the longest connected components (path $1 \rightarrow 2 \rightarrow 3 \rightarrow 4 \rightarrow 5 \rightarrow 6$ in this example) of this induced graph.

patters such as parallel to the main diagonal (helices), parallel but off diagonal (parallel β sheets) or orthogonal to the main diagonal (antiparallel β sheets). Consider two contact pairs $\mathrm{CP}_1 \equiv (i_1, j_1)$ and $\mathrm{CP}_2 \equiv (i_2, j_2)$ formed from residue pairs numbered (i_1, j_1) and (i_2, j_2). To determine if these contact pairs lie along a pattern corresponding to a (parallel or antiparallel) β sheet, we can simply examine if they are away from the main diagonal. To assess if they are nearest neighbors along such a path, we need to consider the (Euclidean) distance between the contact pairs in the contact map, i.e. not in physical space but in the sequence space. Eq. (6.3.6) expresses both these criteria:

$$\begin{aligned} &(\mathrm{CP}_1, \mathrm{CP}_2) \in \beta, \text{ iff} \\ &|i_1 - j_1| \geq d_{\mathrm{co}}, \;\; |i_2 - j_2| \geq d_{\mathrm{co}} \\ &\sqrt{(i_1 - j_1)^2 + (i_2 - j_2)^2} \leq d_{\mathrm{nn}} \end{aligned} \tag{6.3.5}$$

where the first condition ensures sufficient distance from the main diagonal through the contact order parameter d_{co} and the second condition ensures nearest neighborhood within a cutoff contact order d_{nn}. Once two contact pairs satisfy the conditions in Eq. (6.3.6), they can be considered as connected nodes in a graph of contact pairs, as illustrated in Fig. 6.3.1. The problem of searching for a continuous path through candidate CPs then reduces to finding connected components in this graph, which can be achieved through a depth-first-search or a breadth-first-search.[307] However, a true β sheet needs at least four or more continuous inter-strand hydrogen bonds, so that only paths above a critical length say d_{max} can qualify as being part of a β-sheet. The final step that remains is to extract the unique set of residues from the union of residues belonging to all paths found through the graph search. However, there are three free parameters d_{co}, d_{nn}

and $d_{\max}$ in this algorithm which either need to be assigned meaningful values or estimated systematically. It is instructive to optimize / learn these parameters from predicting secondary structures for known structures, perhaps across a large subsection of sequences in the Protein Data Bank.[296]

Bibliography

[1] Manoj V. Athawale, Gaurav Goel, Tuhin Ghosh, Thomas M. Truskett, and Shekhar Garde. Effects of lengthscales and attractions on the collapse of hydrophobic polymers in water. *Proceedings of the National Academy of Sciences of the United States of America*, 104(3):733–738, 2007.

[2] B.J. Alder and T.E. Wainwright. Studies in molecular dynamics. i. general method. *Journal of Chemical Physics*, 31:459, 1959.

[3] Jeremy C. Palmer and Pablo G. Debenedetti. Recent advances in molecular simulation: A chemical engineering perspective. *AIChE Journal*, 61(2):370–383, 2015.

[4] Alessandra Villa, Christine Peter, and Nico F. A. van der Vegt. Transferability of Nonbonded Interaction Potentials for Coarse-Grained Simulations: Benzene in Water. *Journal of Chemical Theory and Computation*, 6(8):2434–2444, 2010.

[5] Jacob W. Wagner, Thomas Dannenhoffer-Lafage, Jaehyeok Jin, and Gregory A. Voth. Extending the range and physical accuracy of coarse-grained models: Order parameter dependent interactions. *The Journal of Chemical Physics*, 147(4):044113, 2017.

[6] Michael R. DeLyser and William G. Noid. Extending pressure-matching to inhomogeneous systems via local-density potentials. *The Journal of Chemical Physics*, 147(13):134111, 2017.

[7] Beno?t Roux and Thomas Simonson. Implicit solvent models. *Biophysical Chemistry*, 78(1?2):1–20, 1999.

[8] Themis Lazaridis and Martin Karplus. Effective energy function for proteins in solution. *Proteins: Structure, Function, and Bioinformatics*, 35(2):133–152, 1999.

[9] Nathan A. Baker, David Sept, Simpson Joseph, Michael J. Holst, and J. Andrew McCammon. Electrostatics of nanosystems: Application to microtubules and the ribosome. *Proceedings of the National Academy of Sciences*, 98(18):10037–10041, 2001.

[10] Alexey Onufriev, David A. Case, and Donald Bashford. Effective Born radii in the generalized Born approximation: The importance of being perfect. *Journal of Computational Chemistry*, 23(14):1297–1304, 2002.

[11] Christopher J. Fennell, Charlie Kehoe, and Ken A. Dill. Oil/Water Transfer Is Partly Driven by Molecular Shape, Not Just Size. *Journal of the American Chemical Society*, 132(1):234–240, 2010.

[12] Christopher J. Fennell, Charles W. Kehoe, and Ken A. Dill. Modeling aqueous solvation with semi-explicit assembly. *Proceedings of the National Academy of Sciences*, 108(8):3234–3239, 2011.

[13] Yaohong Wang, Jon Karl Sigurdsson, Erik Brandt, and Paul J. Atzberger. Dynamic implicit-solvent coarse-grained models of lipid bilayer membranes: Fluctuating hydrodynamics thermostat. *Physical Review E*, 88(2):023301, 2013.

[14] Torsten Schwede, Andrej Sali, Barry Honig, Michael Levitt, Helen M. Berman, David Jones, Steven E. Brenner, Stephen K. Burley, Rhiju Das, Nikolay V. Dokholyan, Roland L. Dunbrack, Krzysztof Fidelis, Andras Fiser, Adam Godzik, Yuanpeng Janet Huang, Christine Humblet, Matthew P. Jacobson, Andrzej Joachimiak, Stanley R. Krystek, Tanja Kortemme, Andriy Kryshtafovych, Gaetano T. Montelione, John Moult, Diana Murray, Roberto Sanchez, Tobin R. Sosnick, Daron M. Standley, Terry Stouch, Sandor Vajda, Max Vasquez, John D. Westbrook, and Ian A. Wilson. Outcome of a Workshop on Applications of Protein Models in Biomedical Research. *Structure*, 17(2):151–159, February 2009.

[15] George A. Khoury, James Smadbeck, Chris A. Kieslich, and Christodoulos A. Floudas. Protein folding and de novo protein design for biotechnological applications. *Trends in Biotechnology*, 32(2):99–109, February 2014.

[16] Paweł Sledz and Amedeo Caflisch. Protein structure-based drug design: from docking to molecular dynamics. *Current Opinion in Structural Biology*, 48:93–102, February 2018.

[17] Fabrizio Chiti and Christopher M. Dobson. Protein misfolding, functional amyloid and human disease. *Annual Review of Biochemistry*, 75:333–366, 2006.

[18] Mathias Jucker and Lary C. Walker. Self-propagation of pathogenic protein aggregates in neurodegenerative diseases. *Nature*, 501(7465):45–51, 2013.

[19] Benjamin Falcon, Wenjuan Zhang, Alexey G. Murzin, Garib Murshudov, Holly J. Garringer, Ruben Vidal, R. Anthony Crowther, Bernadio Ghetti, Sjors H.W. Scheres, and Michel Goedert. Structure of filaments from pick's disease reveal a novel tau protein fold. *Nature*, 561:137–140, 2018.

[20] Hilda Mirbaha, Dailu Chen, Olga A. Morazova, Kiersten M. Ruff, Apurwa M. Sharma, Xiaohua Liu, Mohammad Goodarzi, Rohit V. Pappu, David W. Colby, Hamid Mirzaei, Lukasz A. Joachimiak, and Marc I. Diamond. Inert and seed-competent tau monomers suggest structural origins of aggregation. *eLife*, 7:e36584, 2018.

[21] Alexander P. Lyubartsev and Aatto Laaksonen. Calculation of effective interaction potentials from radial distribution functions: A reverse Monte Carlo approach. *Physical Review E*, 52(4):3730–3737, 1995.

[22] Dirk Reith, Mathias Putz, and Florian Muller-Plathe. Deriving effective mesoscale potentials from atomistic simulations. *Journal of Computational Chemistry*, 24(13):1624–1636, 2003.

[23] Florian Muller-Plathe. Coarse-Graining in Polymer Simulation: From the Atomistic to the Mesoscopic Scale and Back. *ChemPhysChem*, 3(9):754–769, 2002.

[24] Alexander P. Lyubartsev and Alexander L. Rabinovich. Recent development in computer simulations of lipid bilayers. *Soft Matter*, 7(1):25–39, 2010.

[25] Sergei Izvekov and Gregory A. Voth. A Multiscale Coarse-Graining Method for Biomolecular Systems. *The Journal of Physical Chemistry B*, 109(7):2469–2473, 2005.

[26] W. G. Noid, Jhih-Wei Chu, Gary S. Ayton, Vinod Krishna, Sergei Izvekov, Gregory A. Voth, Avisek Das, and Hans C. Andersen. The multiscale coarse-graining method. I. A rigorous bridge between atomistic and coarse-grained models. *The Journal of Chemical Physics*, 128(24):244114, 2008.

[27] Qiang Shi, Sergei Izvekov, and Gregory A. Voth. Mixed atomistic and coarse-grained molecular dynamics: Simulation of a membrane bound ion channel. *The Journal of Physical Chemistry B*, 110(31):15045–15048, August 2006.

[28] Jian Zhou, Ian F. Thorpe, Sergey Izvekov, and Gregory A. Voth. Coarse-Grained Peptide Modeling Using a Systematic Multiscale Approach. *Biophysical Journal*, 92(12):4289–4303, June 2007.

[29] Ian F. Thorpe, Jian Zhou, and Gregory A. Voth. Peptide Folding Using Multiscale Coarse-Grained Models. *The Journal of Physical Chemistry B*, 112(41):13079–13090, October 2008.

[30] Gregory A. Voth. *Coarse-Graining of Condensed Phase and Biomolecular Systems*. CRC Press, September 2008.

[31] Ian F. Thorpe, David P. Goldenberg, and Gregory A. Voth. Exploration of Transferability in Multiscale Coarse-Grained Peptide Models. *The Journal of Physical Chemistry B*, 115(41):11911–11926, October 2011.

[32] M. Scott Shell. The relative entropy is fundamental to multiscale and inverse thermodynamic problems. *The Journal of Chemical Physics*, 129(14):144108, 2008.

[33] Aviel Chaimovich and M. Scott Shell. Relative entropy as a universal metric for multiscale errors. *Physical Review E*, 81(6):060104, 2010.

[34] Scott P. Carmichael and M. Scott Shell. A New Multiscale Algorithm and Its Application to Coarse-Grained Peptide Models for Self-Assembly. *The Journal of Physical Chemistry B*, 116(29):8383–8393, 2012.

[35] Aviel Chaimovich and M. Scott Shell. Coarse-graining errors and numerical optimization using a relative entropy framework. *The Journal of Chemical Physics*, 134(9):094112, 2011.

[36] S. Kullback and R. A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[37] Pep Espanol and Ignacio Zuniga. Obtaining fully dynamic coarse-grained models from md. *Physical Chemistry Chemical Physics*, 13:10538–10545, 2011.

[38] David A. Sivak and Gavin E. Crooks. Near-equilibrium measurements of nonequilibrium free energy. *Phys. Rev. Lett.*, 108:150601, Apr 2012.

[39] Aviel Chaimovich and M. Scott Shell. Anomalous waterlike behavior in spherically-symmetric water models optimized with the relative entropy. *Physical Chemistry Chemical Physics*, 11(12):1901–1915, 2009.

[40] Malte U. Hammer, Travers H. Anderson, Aviel Chaimovich, M. Scott Shell, and Jacob Israelachvili. The search for the hydrophobic force law. *Faraday Discussions*, 146(0):299–308, 2010.

[41] Scott P. Carmichael and M. Scott Shell. Entropic (de)stabilization of surface-bound peptides conjugated with polymers. *The Journal of chemical physics*, 143 24:243103, 2015.

[42] F. Ercolessi and J. B. Adams. Interatomic Potentials from First-Principles Calculations: The Force-Matching Method. *EPL (Europhysics Letters)*, 26(8):583, 1994.

[43] Sergei Izvekov and Gregory A. Voth. Multiscale coarse graining of liquid-state systems. *The Journal of Chemical Physics*, 123(13):134105, 2005.

[44] J. F. Rudzinski and W. G. Noid. A generalized-Yvon-Born-Green method for coarse-grained modeling. *The European Physical Journal Special Topics*, 224(12):2193–2216, 2015.

[45] James F. Dama, Anton V. Sinitskiy, Martin McCullagh, Jonathan Weare, Benoit Roux, Aaron R. Dinner, and Gregory A. Voth. The Theory of Ultra-Coarse-Graining. 1. General Principles. *Journal of Chemical Theory and Computation*, 9(5):2466–2480, 2013.

[46] Victor Ruhle, Christoph Junghans, Alexander Lukyanov, Kurt Kremer, and Denis Andrienko. Versatile Object-Oriented Toolkit for Coarse-Graining Applications. *Journal of Chemical Theory and Computation*, 5(12):3211–3223, 2009.

[47] John G. Kirkwood. Statistical Mechanics of Fluid Mixtures. *The Journal of Chemical Physics*, 3(5):300–313, 1935.

[48] Frank H. Stillinger and Thomas A. Weber. Hidden structure in liquids. *Physical Review A*, 25(2):978–989, 1982.

[49] A. Liwo, R. Kazmierkiewicz, C. Czaplewski, M. Groth, S. Oldziej, R. J. Wawak, S. Rackovsky, M. R. Pincus, and H. A. Scheraga. United-residue force field for off-lattice protein-structure simulations: III. Origin of backbone hydrogen-bonding cooperativity in united-residue potentials. *Journal of Computational Chemistry*, 19(3):259–276, 1998.

[50] Christos N. Likos. Effective interactions in soft condensed matter physics. *Physics Reports*, 348(4?5):267–439, 2001.

[51] Margaret E. Johnson, Teresa Head-Gordon, and Ard A. Louis. Representability problems for coarse-grained water potentials. *The Journal of Chemical Physics*, 126(14):144509, 2007.

[52] W. G. Noid. Perspective: Coarse-grained models for biomolecular systems. *The Journal of Chemical Physics*, 139(9):090901, 2013.

[53] Jayeeta Ghosh and Roland Faller. State point dependence of systematically coarse-grained potentials. *Molecular Simulation*, 33(9-10):759–767, 2007.

[54] Emil Sobolewski, Mariusz Makowski, Stanislaw Oldziej, Cezary Czaplewski, Adam Liwo, and Harold A. Scheraga. Towards temperature-dependent coarse-grained potentials of side-chain interactions for protein folding simulations. I: Molecular dynamics study of a pair of methane molecules in water at various temperatures. *Protein Engineering Design and Selection*, 22(9):547–552, 2009.

[55] Vinod Krishna, Will G. Noid, and Gregory A. Voth. The multiscale coarse-graining method. IV. Transferring coarse-grained potentials between temperatures. *The Journal of Chemical Physics*, 131(2):024103, 2009.

[56] Avisek Das and Hans C. Andersen. The multiscale coarse-graining method. V. Isothermal-isobaric ensemble. *The Journal of Chemical Physics*, 132(16):164106, 2010.

[57] Murray S. Daw, Stephen M. Foiles, and Michael I. Baskes. The embedded-atom method: a review of theory and applications. *Materials Science Reports*, 9(7-8):251–310, 1993.

[58] J. Tersoff. New empirical approach for the structure and energy of covalent systems. *Physical Review B*, 37(12):6991–7000, 1988.

[59] Donald W. Brenner. Empirical potential for hydrocarbons for use in simulating the chemical vapor deposition of diamond films. *Physical Review B*, 42(15):9458–9471, 1990.

[60] M. W. Finnis and J. E. Sinclair. A simple empirical N-body potential for transition metals. *Philosophical Magazine A*, 50(1):45–55, 1984.

[61] Joao F. Justo, Martin Z. Bazant, Efthimios Kaxiras, V. V. Bulatov, and Sidney Yip. Interatomic potential for silicon defects and disordered phases. *Physical Review B*, 58(5):2539–2550, 1998.

[62] A. A. Louis. Beware of density dependent pair potentials. *Journal of Physics: Condensed Matter*, 14(40):9187, 2002.

[63] J. W. Mullinax and W. G. Noid. Extended ensemble approach for deriving transferable coarse-grained potentials. *The Journal of Chemical Physics*, 131(10):104110, 2009.

[64] Nicholas J. H. Dunn and W. G. Noid. Bottom-up coarse-grained models that accurately describe the structure, pressure, and compressibility of molecular liquids. *The Journal of Chemical Physics*, 143(24):243148, 2015.

[65] Erik C. Allen and Gregory C. Rutledge. A novel algorithm for creating coarse-grained, density dependent implicit solvent models. *The Journal of Chemical Physics*, 128(15):154115, 2008.

[66] Erik C. Allen and Gregory C. Rutledge. Evaluating the transferability of coarse-grained, density-dependent implicit solvent models to mixtures and chains. *The Journal of Chemical Physics*, 130(3):034904, 2009.

[67] Sergei Izvekov, Peter W. Chung, and Betsy M. Rice. Particle-based multiscale coarse graining with density-dependent potentials: Application to molecular crystals (hexahydro-1,3,5-trinitro-s-triazine). *The Journal of Chemical Physics*, 135(4):044112, 2011.

[68] Frank H. Stillinger and Thomas A. Weber. Computer simulation of local order in condensed phases of silicon. *Physical Review B*, 31(8):5262–5271, 1985.

[69] Valeria Molinero and Emily B. Moore. Water Modeled As an Intermediate Element between Carbon and Silicon. *The Journal of Physical Chemistry B*, 113(13):4008–4016, 2009.

[70] Jibao Lu, Yuqing Qiu, Riccardo Baron, and Valeria Molinero. Coarse-Graining of TIP4p/2005, TIP4p-Ew, SPC/E, and TIP3p to Monatomic Anisotropic Water Models Using Relative Entropy Minimization. *Journal of Chemical Theory and Computation*, 10(9):4104–4120, 2014.

[71] Brandon C. Knott, Valeria Molinero, Michael F. Doherty, and Baron Peters. Homogeneous Nucleation of Methane Hydrates: Unrealistic under Realistic Conditions. *Journal of the American Chemical Society*, 134(48):19544–19547, 2012.

[72] Avisek Das and Hans C. Andersen. The multiscale coarse-graining method. IX. A general method for construction of three body coarse-grained force fields. *The Journal of Chemical Physics*, 136(19):194114, 2012.

[73] Luca Larini, Lanyuan Lu, and Gregory A. Voth. The multiscale coarse-graining method. VI. Implementation of three-body coarse-grained potentials. *The Journal of Chemical Physics*, 132(16):164107, 2010.

[74] Murray S. Daw. Model of metallic cohesion: The embedded-atom method. *Physical Review B*, 39(11):7441–7452, 1989.

[75] Joshua D. Moore, Brian C. Barnes, Sergei Izvekov, Martin Lisal, Michael S. Sellers, DeCarlos E. Taylor, and John K. Brennan. A coarse-grain force field for RDX: Density dependent and energy conserving. *The Journal of Chemical Physics*, 144(10):104501, 2016.

[76] I. Pagonabarraga and D. Frenkel. Non-Ideal DPD Fluids. *Molecular Simulation*, 25(3-4):167–175, 2000.

[77] Evan J. Arthur and Charles L. Brooks. Parallelization and improvements of the generalized born model with a simple sWitching function for modern graphics processors. *Journal of Computational Chemistry*, 37(10):927–939, 2016.

[78] John T. King, Evan J. Arthur, Charles L. Brooks, and Kevin J. Kubarych. Crowding Induced Collective Hydration of Biological Macromolecules over Extended Distances. *Journal of the American Chemical Society*, 136(1):188–194, 2014.

[79] B. R. Brooks, C. L. Brooks, A. D. Mackerell, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, and M. Karplus. CHARMM: The biomolecular simulation program. *Journal of Computational Chemistry*, 30(10):1545–1614, 2009.

[80] T. E. III Cheatham and P. A. Kollma. Observation of theA-DNA toB-DNA Transition During Unrestrained Molecular Dynamics in Aqueous Solution. *Journal of Molecular Biology*, 259(3):434–444, 1996.

[81] W. Kauzmann. Some factors in the interpretation of protein denaturation. *Advances in Protein Chemistry*, 14:1–63, 1959.

[82] Charles Tanford. *The hydrophobic effect: formation of micelles and biological membranes.* Wiley, New York, 1973.

[83] Ken A. Dill, S. Banu Ozkan, M. Scott Shell, and Thomas R. Weikl. The Protein Folding Problem. *Annual Review of Biophysics*, 37(1):289–316, 2008.

[84] G. Hummer. Hydrophobic Force Field as a Molecular Alternative to Surface-Area Models. *Journal of the American Chemical Society*, 121(26):6299–6305, 1999.

[85] G. Hummer, S. Garde, A. E. Garc?a, and L. R. Pratt. New perspectives on hydrophobic effects. *Chemical Physics*, 258(2?3):349–370, 2000.

[86] Cezary Czaplewski, Sylwia Rodziewicz-Motowidlo, Adam Liwo, Daniel R. Ripoll, Ryszard J. Wawak, and Harold A. Scheraga. Molecular simulation study of cooperativity in hydrophobic association. *PRS*, 9(06):1235?1245, 2000.

[87] Seishi Shimizu and Hue Sun Chan. Anti-cooperativity in hydrophobic interactions: A simulation study of spatial dependence of three-body effects and beyond. *The Journal of Chemical Physics*, 115(3):1414–1421, 2001.

[88] Seishi Shimizu and Hue Sun Chan. Anti-cooperativity and cooperativity in hydrophobic interactions: Three-body free energy landscapes and comparison with implicit-solvent potential functions for proteins. *Proteins: Structure, Function, and Bioinformatics*, 48(1):15–30, 2002.

[89] Lingle Wang, Richard A. Friesner, and B. J. Berne. Hydrophobic interactions in model enclosures from small to large length scales: non-additivity in explicit and implicit solvent models. *Faraday Discussions*, 146:247–262; discussion 283–298, 395–401, 2010.

[90] Sergei Izvekov. Towards an understanding of many-particle effects in hydrophobic association in methane solutions. *The Journal of Chemical Physics*, 134(3):034104, 2011.

[91] Helena W. Qi, Hannah R. Leverentz, and Donald G. Truhlar. Water 16-mers and Hexamers: Assessment of the Three-Body and Electrostatically Embedded Many-Body Approximations of the Correlation Energy or the Nonlocal Energy As Ways to Include Cooperative Effects. *The Journal of Physical Chemistry A*, 117(21):4486–4499, 2013.

[92] A. Subha Mahadevi and G. Narahari Sastry. Cooperativity in Noncovalent Interactions. *Chemical Reviews*, 116(5):2775–2825, 2016.

[93] Sumanth N. Jamadagni, Rahul Godawat, and Shekhar Garde. Hydrophobicity of Proteins and Interfaces: Insights from Density Fluctuations. *Annual Review of Chemical and Biomolecular Engineering*, 2(1):147–171, 2011.

[94] G. Hummer, S. Garde, A. E. Garcia, A. Pohorille, and L. R. Pratt. An information theory model of hydrophobic interactions. *Proceedings of the National Academy of Sciences*, 93(17):8951–8955, 1996.

[95] Rahul Godawat, Sumanth N. Jamadagni, and Shekhar Garde. Characterizing hydrophobicity of interfaces by using cavity formation, solute binding, and water correlations. *Proceedings of the National Academy of Sciences*, 106(36):15119–15124, 2009.

[96] Donald Allan McQuarrie. *Statistical Mechanics*. University Science Books, 2000.

[97] Sapna Sarupria and Shekhar Garde. Quantifying Water Density Fluctuations and Compressibility of Hydration Shells of Hydrophobic Solutes and Proteins. *Physical Review Letters*, 103(3):037803, 2009.

[98] Sumanth N. Jamadagni, Rahul Godawat, Jonathan S. Dordick, and Shekhar Garde. How Interfaces Affect Hydrophobically Driven Polymer Folding. *The Journal of Physical Chemistry B*, 113(13):4093–4101, 2009.

[99] Sumanth N. Jamadagni, Rahul Godawat, and Shekhar Garde. How surface wettability affects the binding, folding, and dynamics of hydrophobic polymers at interfaces. *Langmuir: the ACS journal of surfaces and colloids*, 25(22):13092–13099, 2009.

[100] H. S. Ashbaugh, S. Garde, G. Hummer, E. W. Kaler, and M. E. Paulaitis. Conformational equilibria of alkanes in aqueous solution: relationship to water structure near hydrophobic solutes. *Biophysical Journal*, 77(2):645–654, 1999.

[101] I. Pagonabarraga and D. Frenkel. Dissipative particle dynamics for interacting systems. *The Journal of Chemical Physics*, 115(11):5015–5026, 2001.

[102] Joseph F. Rudzinski and W. G. Noid. Coarse-graining entropy, forces, and structures. *The Journal of Chemical Physics*, 135(21):214101, 2011.

[103] Aviel Chaimovich and M. Scott Shell. Length-scale crossover of the hydrophobic interaction in a coarse-grained water model. *Physical Review E*, 88(5):052313, 2013.

[104] John D. Weeks, David Chandler, and Hans C. Andersen. Role of Repulsive Forces in Determining the Equilibrium Structure of Simple Liquids. *The Journal of Chemical Physics*, 54(12):5237–5247, 1971.

[105] Steve Plimpton. Fast Parallel Algorithms for Short-Range Molecular Dynamics. *Journal of Computational Physics*, 117(1):1–19, 1995.

[106] H. J. C. Berendsen, J. R. Grigera, and T. P. Straatsma. The missing term in effective pair potentials. *The Journal of Physical Chemistry*, 91(24):6269–6271, 1987.

[107] Jean-Paul Ryckaert, Giovanni Ciccotti, and Herman J. C Berendsen. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics*, 23(3):327–341, 1977.

[108] M. Parrinello and A. Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics*, 52(12):7182–7190, 1981.

[109] Charles H Bennett. Efficient estimation of free energy differences from Monte Carlo data. *Journal of Computational Physics*, 22(2):245–268, 1976.

[110] A. Shrake and J. A. Rupley. Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *Journal of Molecular Biology*, 79(2):351–371, 1973.

[111] K. A. Sharp, A. Nicholls, R. F. Fine, and B. Honig. Reconciling the magnitude of the microscopic and macroscopic hydrophobic effects. *Science (New York, N.Y.)*, 252(5002):106–109, 1991.

[112] Lawrence R. Pratt and David Chandler. Theory of the hydrophobic effect. *The Journal of Chemical Physics*, 67(8):3683–3704, 1977.

[113] Ka Lum, David Chandler, and John D. Weeks. Hydrophobicity at Small and Large Length Scales. *The Journal of Physical Chemistry B*, 103(22):4570–4577, 1999.

[114] F. M. Richards. Areas, Volumes, Packing, and Protein Structure. *Annual Review of Biophysics and Bioengineering*, 6(1):151–176, 1977.

[115] T. Ooi, M. Oobatake, G. Nemethy, and H. A. Scheraga. Accessible surface areas as a measure of the thermodynamic parameters of hydration of peptides. *Proceedings of the National Academy of Sciences*, 84(10):3086–3090, 1987.

[116] Jens Kleinjung and Franca Fraternali. Design and application of implicit solvent models in biomolecular simulations. *Current Opinion in Structural Biology*, 25(100):126–134, 2014.

[117] Jed W. Pitera and Wilfred F. van Gunsteren. The Importance of Solute-Solvent van der Waals Interactions with Interior Atoms of Biopolymers. *Journal of the American Chemical Society*, 123(13):3163–3164, 2001.

[118] David M. Huang and David Chandler. The Hydrophobic Effect and the Influence of Solute-Solvent Attractions. *The Journal of Physical Chemistry B*, 106(8):2047–2053, 2002.

[119] Richard C. Remsing and John D. Weeks. Dissecting Hydrophobic Hydration and Association. *The Journal of Physical Chemistry B*, 117(49):15479–15491, 2013.

[120] Michael E Paulaitis, Shekhar Garde, and Henry S Ashbaugh. The hydrophobic effect. *Current Opinion in Colloid & Interface Science*, 1(3):376–383, 1996.

[121] Keith E. Gubbins. Molecular Simulation: Phase equilibria and confined systems. In *Scientific Computing in Chemical Engineering II*, pages 2–11. Springer, Berlin, Heidelberg, 1999.

[122] J. J. de Pablo, Q. Yan, and F. A. Escobedo. Simulation of phase transitions in fluids. *Annual Review of Physical Chemistry*, 50:377–411, 1999.

[123] C. Chipot and A. Pohorille. *Free Energy Calculations - Theory and Applications in Chemistry and Biology.* Springer, Berlin, Heidelberg, 2007.

[124] Fabio Pietrucci. Strategies for the exploration of free energy landscapes: Unity in diversity and challenges ahead. *Reviews in Physics*, 2:32–45, 2017.

[125] Athanassios Z. Panagiotopoulos. Direct determination of phase coexistence properties of fluids by Monte Carlo simulation in a new ensemble. *Molecular Physics*, 61(4):813–826, 1987.

[126] Alan M. Ferrenberg and Robert H. Swendsen. New Monte Carlo technique for studying phase transitions. *Physical Review Letters*, 61(23):2635–2638, 1988.

[127] Alan M. Ferrenberg and Robert H. Swendsen. Optimized Monte Carlo data analysis. *Physical Review Letters*, 63(12):1195–1198, 1989.

[128] Jeffrey R. Errington. Direct calculation of liquid?vapor phase equilibria from transition matrix Monte Carlo simulation. *The Journal of Chemical Physics*, 118(22):9915–9925, 2003.

[129] John C. Shelley, Mee Y. Shelley, Robert C. Reeder, Sanjoy Bandyopadhyay, and Michael L. Klein. A Coarse Grain Model for Phospholipid Simulations. *The Journal of Physical Chemistry B*, 105(19):4464–4470, 2001.

[130] Amy Y. Shih, Peter L. Freddolino, Anton Arkhipov, and Klaus Schulten. Assembly of lipoprotein particles revealed by coarse-grained molecular dynamics simulations. *Journal of Structural Biology*, 157(3):579–592, 2007.

[131] Jocelyn M. Rodgers, Jesper Sorensen, Frederick J.-M. de Meyer, Birgit Schiott, and Berend Smit. Understanding the Phase Behavior of Coarse-Grained Model Lipid Bilayers through Computational Calorimetry. *The Journal of Physical Chemistry B*, 116(5):1551–1569, 2012.

[132] Siewert J. Marrink, Alex H. de Vries, and Alan E. Mark. Coarse Grained Model for Semiquantitative Lipid Simulations. *The Journal of Physical Chemistry B*, 108(2):750–760, 2004.

[133] Ken A. Dill, Sarina Bromberg, Kaizhi Yue, Hue Sun Chan, Klaus M. Ftebig, David P. Yee, and Paul D. Thomas. Principles of protein folding : A perspective from simple exact models. *Protein Science*, 4(4):561–602, 1995.

[134] Eivind Tostesen, Shi-Jie Chen, and Ken A. Dill. RNA Folding Transitions and Cooperativity. *The Journal of Physical Chemistry B*, 105(8):1618–1630, 2001.

[135] Vincent K. Shen, Jason K. Cheung, Jeffrey R. Errington, and Thomas M. Truskett. Coarse-Grained Strategy for Modeling Protein Stability in Concentrated Solutions. II: Phase Behavior. *Biophysical Journal*, 90(6):1949–1960, 2006.

[136] Yanting Wang and Gregory A. Voth. Molecular Dynamics Simulations of Polyglutamine Aggregation Using Solvent-Free Multiscale Coarse-Grained Models. *The Journal of Physical Chemistry B*, 114(26):8735–8743, 2010.

[137] Kiersten M. Ruff, Siddique J. Khan, and Rohit V. Pappu. A Coarse-Grained Model for Polyglutamine Aggregation Modulated by Amphipathic Flanking Sequences. *Biophysical Journal*, 107(5):1226–1235, 2014.

[138] Joseph F. Rudzinski and William G. Noid. Bottom-Up Coarse-Graining of Peptide Ensembles and Helix-Coil Transitions. *Journal of Chemical Theory and Computation*, 11(3):1278–1291, 2015.

[139] Tyler S. Harmon, Alex S. Holehouse, Michael K. Rosen, and Rohit V. Pappu. Intrinsically disordered linkers determine the interplay between phase separation and gelation in multivalent proteins. *bioRxiv*, page 164301, 2017.

[140] Zachary Monahan, Veronica H Ryan, Abigail M Janke, Kathleen A Burke, Shannon N Rhoads, Gul H Zerze, Robert O'Meally, Gregory L Dignon, Alexander E Conicella, Wenwei Zheng, Robert B Best, Robert N Cole, Jeetain Mittal, Frank Shewmaker, and Nicolas L Fawzi. Phosphorylation of the FUS low?complexity domain disrupts phase separation, aggregation, and toxicity. *The EMBO Journal*, 36(20):2951–2967, 2017.

[141] Emiliano Brini, Elena A.Algaer, Pritam Ganguly, Chunli Li, Francisco Rodriguez-Ropero, and van der Nico F. A. Vegt. Systematic coarse-graining methods for soft matter simulations - a review. *Soft Matter*, 9(7):2108–2119, 2013.

[142] James F. Dama, Jaehyeok Jin, and Gregory A. Voth. The Theory of Ultra-Coarse-Graining. 3. Coarse-Grained Sites with Rapid Local Equilibrium of Internal States. *Journal of Chemical Theory and Computation*, 13(3):1010–1022, 2017.

[143] Nicholas J. H. Dunn and W. G. Noid. Bottom-up coarse-grained models with predictive accuracy and transferability for both structural and thermodynamic properties of heptane-toluene mixtures. *The Journal of Chemical Physics*, 144(20):204124, 2016.

[144] Tanmoy Sanyal and M. Scott Shell. Coarse-grained models using local-density potentials optimized with the relative entropy: Application to implicit solvation. *The Journal of Chemical Physics*, 145(3):034109, 2016.

[145] Erik C. Allen and Gregory C. Rutledge. Coarse-grained, density dependent implicit solvent model reliably reproduces behavior of a model surfactant system. *The Journal of Chemical Physics*, 130(20):204903, 2009.

[146] Aatto Laaksonen, Peter Stilbs, and Roderick E. Wasylishen. Molecular motion and solvation of benzene in water, carbon tetrachloride, carbon disulfide and benzene: A combined molecular dynamics simulation and nuclear magnetic resonance study. *The Journal of Chemical Physics*, 108(2):455–468, 1998.

[147] Per Linse. Molecular dynamics simulation of a dilute aqueous solution of benzene. *Journal of the American Chemical Society*, 112(5):1744–1750, 1990.

[148] Liem X. Dang and David Feller. Molecular Dynamics Study of Water-Benzene Interactions at the Liquid/Vapor Interface of Water. *The Journal of Physical Chemistry B*, 104(18):4403–4407, 2000.

[149] David Chandler. Interfaces and the driving force of hydrophobic assembly. *Nature*, 437(7059):640–647, 2005.

[150] H. J. C. Berendsen, D. van der Spoel, and R. van Drunen. GROMACS: A message-passing parallel molecular dynamics implementation. *Computer Physics Communications*, 91(1):43–56, 1995.

[151] Berk Hess, Carsten Kutzner, David van der Spoel, and Erik Lindahl. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *Journal of Chemical Theory and Computation*, 4(3):435–447, 2008.

[152] Chris Oostenbrink, Alessandra Villa, Alan E. Mark, and Wilfred F. Van Gunsteren. A biomolecular force field based on the free enthalpy of hydration and solvation: The GROMOS force-field parameter sets 53a5 and 53a6. *Journal of Computational Chemistry*, 25(13):1656–1676, 2004.

[153] Shuichi Nose. A unified formulation of the constant temperature molecular dynamics methods. *The Journal of Chemical Physics*, 81(1):511–519, 1984.

[154] William G. Hoover. Canonical dynamics: Equilibrium phase-space distributions. *Physical Review A*, 31(3):1695–1697, 1985.

[155] David R. Lide. *CRC handbook of chemistry and physics, 89th edition, edited by David R. Lide*. 2008.

[156] Pim Schravendijk and Nico F. A. van der Vegt. From Hydrophobic to Hydrophilic Solvation: An Application to Hydration of Benzene. *Journal of Chemical Theory and Computation*, 1(4):643–652, 2005.

[157] Frank H. Stillinger and Teresa Head-Gordon. Perturbational view of inherent structures in water. *Physical Review E*, 47(4):2484–2490, 1993.

[158] Zhenyu Yan, Sergey V. Buldyrev, Pradeep Kumar, Nicolas Giovambattista, Pablo G. Debenedetti, and H. Eugene Stanley. Structure of the first- and second-neighbor shells of simulated water: Quantitative relation to translational and orientational order. *Physical Review E*, 76(5):051201, 2007.

[159] Aviel Chaimovich and M. Scott Shell. Tetrahedrality and structural order for hydrophobic interactions in a coarse-grained water model. *Physical Review E*, 89(2):022140, 2014.

[160] G. Hummer, S. Garde, A. E. Garcia, M. E. Paulaitis, and L. R. Pratt. Hydrophobic Effects on a Molecular Scale. *The Journal of Physical Chemistry B*, 102(51):10469–10482, 1998.

[161] Andrew Pohorille and Lawrence R. Pratt. Cavities in molecular liquids and the theory of hydrophobic solubilities. *Journal of the American Chemical Society*, 112(13):5066–5074, 1990.

[162] Lawrence R. Pratt and Andrew Pohorille. Theory of hydrophobicity: Transient cavities in molecular liquids. *Proceedings of the National Academy of Sciences*, 89(7):2995–2999, 1992.

[163] Jonathan G. Harris. Liquid-vapor interfaces of alkane oligomers: structure and thermodynamics from molecular dynamics simulations of chemically realistic models. *The Journal of Physical Chemistry*, 96(12):5077–5086, 1992.

[164] Jose Alejandre, Dominic J. Tildesley, and Gustavo A. Chapela. Molecular dynamics simulation of the orthobaric densities and surface tension of water. *The Journal of Chemical Physics*, 102(11):4574–4583, 1995.

[165] Guy J. Gloor, George Jackson, Felipe J. Blas, and Enrique de Miguel. Test-area simulation method for the direct determination of the interfacial tension of systems with continuous or discontinuous potentials. *The Journal of Chemical Physics*, 123(13):134703, 2005.

[166] Aziz Ghoufi, Patrice Malfreyt, and Dominic J. Tildesley. Computer modelling of the surface tension of the gas?liquid and liquid?liquid interface. *Chemical Society Reviews*, 45(5):1387–1409, 2016.

[167] H. L. Cupples. Interfacial Tension by the Ring Method: The Benzene-Water Interface. *The Journal of Physical and Colloid Chemistry*, 51(6):1341–1345, 1947.

[168] Han Wang, Christoph Junghans, and Kurt Kremer. Comparative atomistic and coarse-grained study of water: What do we lose by coarse-graining? *The European Physical Journal E*, 28(2):221–229, 2009.

[169] Sergei Izvekov, Peter W. Chung, and Betsy M. Rice. The multiscale coarse-graining method: Assessing its accuracy and introducing density dependent coarse-grain potentials. *The Journal of Chemical Physics*, 133(6):064109, 2010.

[170] Chia-Chun Fu, Pandurang M. Kulkarni, M. Scott Shell, and L. Gary Leal. A test of systematic coarse-graining of molecular dynamics simulations: Thermodynamic properties. *The Journal of Chemical Physics*, 137(16):164106, 2012.

[171] Vipin Agrawal, Pedro Peralta, Yiyang Li, and Jay Oswald. A pressure-transferable coarse-grained potential for modeling the shock Hugoniot of polyethylene. *The Journal of Chemical Physics*, 145(10):104903, 2016.

[172] Arieh Ben-Naim. *Molecular Theory of Solutions*. Oxford University Press, Oxford, New York, 2006.

[173] Pritam Ganguly and Nico F. A. van der Vegt. Representability and Transferability of Kirkwood-Buff Iterative Boltzmann Inversion Models for Multicomponent Aqueous Systems. *Journal of Chemical Theory and Computation*, 9(12):5247–5256, 2013.

[174] Pritam Ganguly and Nico F. A. van der Vegt. Convergence of Sampling Kirkwood-Buff Integrals of Aqueous Solutions with Molecular Dynamics Simulations. *Journal of Chemical Theory and Computation*, 9(3):1347–1355, 2013.

[175] Peter Kruger, Sondre K. Schnell, Dick Bedeaux, Signe Kjelstrup, Thijs J. H. Vlugt, and Jean-Marc Simon. Kirkwood-Buff Integrals for Finite Volumes. *The Journal of Physical Chemistry Letters*, 4(2):235–238, 2013.

[176] Pritam Ganguly, Debashish Mukherji, Christoph Junghans, and Nico F. A. van der Vegt. Kirkwood-Buff Coarse-Grained Force Fields for Aqueous Solutions. *Journal of Chemical Theory and Computation*, 8(5):1802–1807, 2012.

[177] Emiliano Brini, Christopher J. Fennell, Marivi Fernandez-Serra, Barbara Hribar-Lee, Miha Luksic, and Ken A. Dill. How Water?s Properties Are Encoded in Its Molecular Structure and Energies. *Chemical Reviews*, 117(19):12385–12414, 2017.

[178] J. C. Kendrew, G. Bodo, H. M. Dintzis, R. G. Parrish, H. Wyckoff, and D. C. Phillips. A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis. *Nature*, 181(4610):662–666, March 1958.

[179] M. F. Perutz, M. G. Rossmann, Ann F. Cullis, Hilary Muirhead, Georg Will, and A. C. T. North. Structure of Hæmoglobin: A Three-Dimensional Fourier Synthesis at 5.5-Å. Resolution, Obtained by X-Ray Analysis. *Nature*, 185(4711):416–422, February 1960.

[180] José A. Brito and Margarida Archer. Chapter 9 - X-ray Crystallography. In Robert R. Crichton and Ricardo O. Louro, editors, *Practical Approaches to Biological Inorganic Chemistry*, pages 217–255. Elsevier, Oxford, January 2013.

[181] Christoph Nitsche and Gottfried Otting. NMR studies of ligand binding. *Current Opinion in Structural Biology*, 48:16–22, February 2018.

[182] Caitlin M Davis, Martin Gruebele, and Shahar Sukenik. How does solvation in the cell affect protein folding and binding? *Current Opinion in Structural Biology*, 48:23–29, February 2018.

[183] Kazuyoshi Murata and Matthias Wolf. Cryo-electron microscopy for structural analysis of dynamic biological macromolecules. *Biochimica et Biophysica Acta (BBA) - General Subjects*, 1862(2):324–334, February 2018.

[184] Ben Moree, Katelyn Connell, Richard B. Mortensen, C. Tony Liu, Stephen J. Benkovic, and Joshua Salafsky. Protein Conformational Changes Are Detected and Resolved Site Specifically by Second-Harmonic Generation. *Biophysical Journal*, 109(4):806–815, August 2015.

[185] Irina Sorokina and Arcady Mushegian. Modeling protein folding in vivo. *Biology Direct*, 13(1):13, July 2018.

[186] Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, January 2000.

[187] David E. Shaw, Ron O. Dror, John K. Salmon, J. P. Grossman, Kenneth M. Mackenzie, Joseph A. Bank, Cliff Young, Martin M. Deneroff, Brannon Batson, Kevin J. Bowers, Edmond Chow, Michael P. Eastwood, Doug Ierardi, John L. Klepeis, Jeffrey Kuskin, Richard H. Larson, Kresten Lindorff-Larsen, Paul Maragakis, Mark A. Moraes, Stefano Piana, Yibing Shan, and Brian Towles. Millisecond-scale molecular dynamics simulations on Anton. *Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis*, pages 1–11, 2009.

[188] Michael Shirts and Vijay S. Pande. Screen Savers of the World Unite! *Science*, 290(5498):1903–1904, December 2000.

[189] Michael Levitt and Arieh Warshel. Computer simulation of protein folding. *Nature*, 253(5494):694–698, February 1975.

[190] Valentina Tozzini. Coarse-grained models for proteins. *Current Opinion in Structural Biology*, 15(2):144–150, April 2005.

[191] Alex Morriss-Andrews and Joan-Emma Shea. Simulations of Protein Aggregation: Insights from Atomistic and Coarse-Grained Models. *The Journal of Physical Chemistry Letters*, 5(11):1899–1908, June 2014.

[192] Benjamin Webb, Keren Lasker, Javier Velázquez-Muriel, Dina Schneidman-Duhovny, Riccardo Pellarin, Massimiliano Bonomi, Charles Greenberg, Barak Raveh, Elina Tjioe, Daniel Russel, and Andrej Sali. Modeling of Proteins and Their Assemblies with the Integrative Modeling Platform. In Yu Wai Chen, editor, *Structural Genomics: General Applications*, Methods in Molecular Biology, pages 277–295. Humana Press, Totowa, NJ, 2014.

[193] Sebastian Kmiecik, Dominik Gront, Michal Kolinski, Lukasz Wieteska, Aleksandra Elzbieta Dawid, and Andrzej Kolinski. Coarse-grained protein models and their applications. *Chemical Reviews*, 116(14):7898–7936, 2016.

[194] Anne Voegler Smith and Carol K. Hall. α-helix formation: Discontinuous molecular dynamics on an intermediate-resolution protein model. *Proteins: Structure, Function, and Bioinformatics*, 44(3):344–360, 2001.

[195] Andrzej Kolinski. Protein modeling and structure prediction with a reduced representation. *Acta Biochimica Polonica*, 51(2):349–371, 2004.

[196] Tristan Bereau and Markus Deserno. Generic coarse-grained model for protein folding and aggregation. *The Journal of Chemical Physics*, 130(23):235106, 2009.

[197] Aram Davtyan, Nicholas P. Schafer, Weihua Zheng, Cecilia Clementi, Peter G. Wolynes, and Garegin A. Papoian. AWSEM-MD: Protein Structure Prediction Using Coarse-Grained Physical Potentials and Bioinformatically Based Local Structure Biasing. *The Journal of Physical Chemistry B*, 116(29):8494–8503, July 2012.

[198] Luca Monticelli, Senthil K. Kandasamy, Xavier Periole, Ronald G. Larson, D. Peter Tieleman, and Siewert-Jan Marrink. The MARTINI Coarse-Grained Force Field: Extension to Proteins. *Journal of Chemical Theory and Computation*, 4(5):819–834, May 2008.

[199] Marco Pasi, Richard Lavery, and Nicoletta Ceres. PaLaCe: A Coarse-Grain Protein Model for Studying Mechanical Properties. *Journal of Chemical Theory and Computation*, 9(1):785–793, January 2013.

[200] Parimal Kar, Srinivasa Murthy Gopal, Yi-Ming Cheng, Alexander Predeus, and Michael Feig. PRIMO: A Transferable Coarse-Grained Force Field for Proteins. *Journal of Chemical Theory and Computation*, 9(8):3769–3788, August 2013.

[201] Nathalie Basdevant, Daniel Borgis, and Tap Ha-Duong. Modeling Protein–Protein Recognition in Solution Using the Coarse-Grained Force Field SCORPION. *Journal of Chemical Theory and Computation*, 9(1):803–813, January 2013.

[202] Fabio Sterpone, Simone Melchionna, Pierre Tuffery, Samuela Pasquali, Normand Mousseau, Tristan Cragnolini, Yassmine Chebaro, Jean-Francois St-Pierre, Maria Kalimeri, Alessandro Barducci, Yoann Laurin, Alex Tek, Marc Baaden, Phuong Hoang Nguyen, and Philippe Derreumaux. The OPEP protein model: from single molecules, amyloid formation, crowding and hydrodynamics to DNA/RNA systems. *Chemical Society Reviews*, 43(13):4871–4893, June 2014.

[203] Adam Liwo, Maciej Baranowski, Cezary Czaplewski, Ewa Gołaś, Yi He, Dawid Jagieła, Paweł Krupa, Maciej Maciejczyk, Mariusz Makowski, Magdalena A. Mozolewska, Andrei Niadzvedtski, Stanisław Ołdziej, Harold A. Scheraga, Adam K. Sieradzan, Rafal Ślusarz, Tomasz Wirecki, Yanping Yin, and Bartłomiej Zaborowski. A unified coarse-grained model of biological macromolecules based on mean-field multipole–multipole interactions. *Journal of Molecular Modeling*, 20(8):2306, July 2014.

[204] Hung D. Nguyen and Carol K. Hall. Spontaneous fibril formation by polyalanine; discontinuous molecular dynamics simulations. *Journal of The American Chemical Society*, 128(6):1890–1901, 2006.

[205] Alexander J. Marchut and Carol K. Hall. Side-chain interactions determine amyloid formation by model polyglutamine peptides in molecular dynamics simulations. *Biophysical Journal*, 90(12):4574 – 4584, 2006.

[206] Tristan Bereau and Markus Deserno. Enhanced sampling of coarse-grained transmembrane-peptide structure formation from hydrogen-bond replica exchange. *The Journal of Membrane Biology*, 248(3):395–405, 2015.

[207] Siewert J. Marrink and D. Peter Tieleman. Perspective on the martini model. *Chemical Society Reviews*, 42(16):6801, 2013.

[208] J. W. Mullinax and W. G. Noid. Recovering physical potentials from a model protein databank. *Proceedings of the National Academy of Sciences*, 107(46):19867–19872, 2010.

[209] Cahit Dalgicdir, Christoph Globisch, Mehmet Sayar, and Christine Peter. Representing environment-induced helix-coil transitions in a coarse grained peptide model. *The European Physical Journal Special Topics*, 225(8):1463–1481, October 2016.

[210] Hiroshi Taketomi, Yuzo Ueda, and Nobuhiro Go. Studies on Protein Folding, Unfolding and Fluctuations by Computer Simulation. *International Journal of Peptide and Protein Research*, 7(6):445–459, November 1975.

[211] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes. Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins*, 21(3):167–195, March 1995.

[212] D. K. Klimov and D. Thirumalai. Mechanisms and kinetics of β-hairpin formation. *Proceedings of the National Academy of Sciences*, 97(6):2544–2549, March 2000.

[213] Cecilia Clementi, Hugh Nymeyer, and José Nelson Onuchic. Topological and energetic factors: what determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? an investigation for small globular proteins11edited by F. E. Cohen. *Journal of Molecular Biology*, 298(5):937–953, May 2000.

[214] John Karanicolas and Charles L. Brooks. The origins of asymmetry in the folding transition states of protein L and protein G. *Protein Science*, 11(10):2351–2361, October 2002.

[215] Ronald Hills, Charles Brooks, Ronald D. Hills, and Charles L. Brooks. Insights from Coarse-Grained Gō Models for Protein Folding and Dynamics. *International Journal of Molecular Sciences*, 10(3):889–905, March 2009.

[216] Pengfei Tian and Robert B. Best. Structural Determinants of Misfolding in Multidomain Proteins. *PLOS Computational Biology*, 12(5):e1004933, May 2016.

[217] Jeffrey K. Noel, Paul C. Whitford, Karissa Y. Sanbonmatsu, and Jose N. Onuchic. SMOG@ctbp: simplified deployment of structure-based models in GROMACS. *Nucleic Acids Research*, 38(suppl_2):W657–W661, July 2010.

[218] Ke Chen, John Eargle, Jonathan Lai, Hajin Kim, Sanjaya Abeysirigunawardena, Megan Mayerle, Sarah Woodson, Taekjip Ha, and Zaida Luthey-Schulten. Assembly of the five-way junction in the ribosomal small subunit using hybrid md - go simulations. *The Journal of Physical Chemistry B*, 116(23):6819–6831, June 2012.

[219] Benjamin Lutz, Claude Sinner, Geertje Heuermann, Abhinav Verma, and Alexander Schug. eSBMTools 1.0: enhanced native structure-based modeling tools. *Bioinformatics*, 29(21):2795–2796, November 2013.

[220] Joseph F. Rudzinski and Tristan Bereau. Structural-kinetic-thermodynamic relationships identified from physics-based molecular simulation models. *The Journal of Chemical Physics*, 148(20):204111, May 2018.

[221] T. Head-Gordon, F. H. Stillinger, M. H. Wright, and D. M. Gay. Poly(L-alanine) as a universal reference material for understanding protein energies and structures. *Proceedings of the National Academy of Sciences of the United States of America*, 89(23):11513–11517, December 1992.

[222] Pilar Cossio, Antonio Trovato, Fabio Pietrucci, Flavio Seno, Amos Maritan, and Alessandro Laio. Exploring the Universe of Protein Structures beyond the Protein Data Bank. *PLOS Computational Biology*, 6(11):e1000957, November 2010.

[223] Natalie L. Dawson, Tony E. Lewis, Sayoni Das, Jonathan G. Lees, David Lee, Paul Ashford, Christine A. Orengo, and Ian Sillitoe. Cath: an expanded

resource to predict protein function through structure and sequence. *Nuclei Acids Research*, 45(D1):D289–D295, 2017.

[224] Predrag Kukic, Arvind Kannan, Maurits J. J. Dijkstra, Sanne Abeln, Carlo Camilloni, and Michele Vendruscolo. Mapping the Protein Fold Universe Using the CamTube Force Field in Molecular Dynamics Simulations. *PLOS Computational Biology*, 11(10):e1004435, October 2015.

[225] Peter G. Wolynes. Symmetry and the energy landscapes of biomolecules. *Proceedings of the National Academy of Sciences*, 93(25):14249–14255, December 1996.

[226] D. Baker. A surprising simplicity to protein folding. *Nature*, 405(6782):39–42, May 2000.

[227] Boris Fain and Michael Levitt. Funnel sculpting for in silico assembly of secondary structure elements of proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 100(19):10700–10705, September 2003.

[228] Jayanth R. Banavar and Amos Maritan. Physics of Proteins. *Annual Review of Biophysics and Biomolecular Structure*, 36(1):261–280, 2007.

[229] M. Scott Shell. Coarse graining with the relative entropy. In *Advances in Chemical Physics*, pages 395–441. John Wiley & Sons, Ltd, 2016.

[230] G.N. Ramachandran, C. Ramakrishnan, and V. Sasisekharan. Stereochemistry of polypeptide chain configurations. *Journal of Molecular Biology*, 7(1):95 – 99, 1963.

[231] C.N. Pace and J.M. Scholtz. A helix propensity scale based on experimental studies of peptides and proteins. *Biophysical Journal*, 75(1):422–427, 1998.

[232] James S. Nowick and Shabana Insaf. The propensities of amino acids to form parallel β-sheets. *Journal of the American Chemical Society*, 119(45):10903–10908, 1997.

[233] Joseph F. Rudzinski, Keran Lu, Scott T. Milner, Janna K. Maranas, and William G. Noid. Extended ensemble approach to transferable potentials for low-resolution coarse-grained models of ionomers. *Journal of Chemical Theory and Computation*, 13(5):2185–2201, 2017.

[234] Jonathan W. Neidigh, R. Matthew Fesinmeyer, and Niels H. Andersen. Designing a 20-residue protein. *Nature Structural and Molecular Biology*, 9:425, 2002.

[235] Linlin Qiu, Suzette A. Pabit, Adrian E. Roitberg, and Stephen J. Hagen. Smaller and faster: The 20-residue trp-cage protein folds in 4 μs. *Journal of the American Chemical Society*, 124(44):12952–12953, 2002.

[236] Bosco K. Ho and Robert Brasseur. The ramachandran plots of glycine and pre-proline. *BMC Structural Biology*, 5(1):14, 2005.

[237] Yuji Sugita and Yuko Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters*, 314(1):141 – 151, 1999.

[238] D.A. Case, R.M. Betz, D.S. Cerutti, T.E. Cheatham III, T.A. Darden, R.E. Duke, T.J. Giese, H. Gohlke, A.W. Goetz, N. Homeyer, S. Izadi, P. Janowski, J. Kaus, A. Kovalenko, T.S. Lee, S. LeGrand, P. Li, C. Lin, T. Luchko, R. Luo, B. Madej, D. Mermelstein, K.M. Merz, G. Monard, H. Nguyen, H.T. Nguyen, I. Omelyan, A. Onufriev, D.R. Roe, A. Roitberg, C. Sagui, C.L. Simmerling, W.M. Botello-Smith, J. Swails, R.C. Walker, J. Wang, R.M. Wolf, X. Wu, L. Xiao, and P.A. Kollman. *AMBER 2016*. University of California, San Francisco, 2016.

[239] P. A. Kollman, R. Dixon, W. Cornell, T. Fox, C. Chipot, and A. Pohorille. The development/application of a “minimalist” organic/biochemical molecular mechanic force field using a combination of ab initio calculations and experimental data. In A. Wilkinson, P. Weiner, and W. F. van Gunsteren, editors, *Computer Simulation of Biomolecular Systems*, volume 3, pages 83–96. Elsevier, 1997.

[240] Raphael Geney, Melinda Layten, Roberto Gomperts, Viktor Hornak, and Carlos Simmerling. Investigation of salt bridge stability in a generalized born solvent model. *Journal of Chemical Theory and Computation*, 2(1):115–127, 2006.

[241] M. Scott Shell, Ryan Ritterson, and Ken A. Dill. A test on peptide stability of amber force fields with implicit solvation. *The Journal of Physical Chemistry B*, 112(22):6878–6886, 2008.

[242] Edmund Lin and M. Scott Shell. Convergence and heterogeneity in peptide folding with replica exchange molecular dynamics. *Journal of Chemical Theory and Computation*, 5(8):2062–2073, 2009.

[243] James A. Maier, Carmenza Martinez, Koushik Kasavajhala, Lauren Wickstrom, Kevin E. Hauser, and Carlos Simmerling. ff14sb: Improving the accuracy of protein side chain and backbone parameters from ff99sb. *Journal of Chemical Theory and Computation*, 11(8):3696–3713, 2015.

[244] Melina K. Robinson, Jacob I. Monroe, and M. Scott Shell. Are amber force fields and implicit solvation models additive? a folding study with a balanced peptide test set. *Journal of Chemical Theory and Computation*, 12(11):5631–5642, 2016.

[245] Michael R. Shirts and John D. Chodera. Statistically optimal analysis of samples from multiple equilibrium states. *The Journal of Chemical Physics*, 129(12):124105, 2008.

[246] W. Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, 32(5):922–923, Sep 1976.

[247] Schrödinger, LLC. The PyMOL molecular graphics system, version 1.8. November 2015.

[248] William Humphrey, Andrew Dalke, and Klaus Schulten. VMD – Visual Molecular Dynamics. *Journal of Molecular Graphics*, 14:33–38, 1996.

[249] Michael R. Sawaya, Shilpa Sambashivan, Rebecca Nelson, Magdalena I. Ivanova, Stuart A. Sievers, Marcin I. Apostol, Michael J. Thompson, Melinda Balbirnie, Jed J. W. Wiltzius, Heather T. McFarlane, Anders O Madsen, Christian Riekel, and David Eisenberg. Atomic structures of amyloid cross-β spines reveal varied steric zippers. *Nature*, 447(7143):453–457, 2007.

[250] S. Banu Ozkan, G. Albert Wu, John D. Chodera, and Ken A. Dill. Protein folding by zipping and assembly. *Proceedings of the National Academy of Sciences*, 104(29):11987–11992, 2007.

[251] Hai Nguyen, James Maier, He Huang, Victoria Perrone, and Carlos Simmerling. Folding simulations for proteins with diverse topologies are accessible in days with a physics-based force field and implicit solvent. *Journal of the American Chemical Society*, 136(40):13959–13962, 2014.

[252] Kresten Lindorff-Larsen, Stefano Piana, Ron O. Dror, and David E. Shaw. How fast-folding proteins fold. *Science*, 332(6055):517–520, 2011.

[253] Alberto Perez, Justin L. MacCallum, and Ken A. Dill. Accelerating molecular simulations of proteins using bayesian inference on weak information. *Proceedings of the National Academy of Sciences of the United States of America*, 112(38):11846–11851, 2015.

[254] Stanley B. Prusiner. Novel proteinaceous infectious particles cause scrapie. *Science*, 216(4542):136–44, 1982.

[255] Stanley B. Prusiner. Prions. *Proceedings of the National Academy of Sciences of the United States of America*, 95(23):13363–13383, 1998.

[256] Tuomas P. J. Knowles, Christopher A. Waudby, Glyn L. Devlin, Samuel I. A. Cohen, Adriano Aguzzi, Michele Vendruscolo, Eugene M. Terentjev, Mark E. Welland, and Christopher M. Dobson. An Analytical Solution to the Kinetics of Breakable Filament Assembly. *Science*, 326(5959):1533–1537, December 2009.

[257] Theodoros K. Karamanos, Arnout P. Kalverda, Gary S. Thompson, and Sheena E. Radford. Mechanisms of amyloid formation revealed by solution NMR. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 88-89:86–104, August 2015.

[258] Ana Rojas, Nika Maisuradze, Khatuna Kachlishvili, Harold A. Scheraga, and Gia G. Maisuradze. Elucidating Important Sites and the Mechanism for Amyloid Fibril Formation by Coarse-Grained Molecular Dynamics. *ACS Chemical Neuroscience*, 8(1):201–209, January 2017.

[259] Thomas R. Jahn and Sheena E. Radford. Folding *versus* aggregation: Polypeptide conformations on competiting pathways. *Archives of Biochemistry and Biophysics*, 469(1):100–17, 2008.

[260] Robert Tycko. Amyloid polymorphism: Structural basis and neurobiological relevance. *Neuron*, 86(3):632–645, 2015.

[261] Loredana Lo Conte, Bart Ailey, Tim J.P. Hubbard, Steven E. Brenner, Alexey G. Murzin, and Cyrus Clothia. Scop: a structural classification of proteins database. *Nucleic Acids Research*, 28(1):257–259, 2000.

[262] William Close, Matthias Neumann, Andreas Schmidt, Manuel Hora, Karthikeyan Annamalai, Matthias Schmidt, Bernd Reif, Volker Schmidt, Nikolaus Grigorieff, and Marcus Fandrich. Physical basis of amyloid fibril polymorphism. *Nature Communications*, 9(1):699, February 2018.

[263] John J. Balbach, Yoshitaka Ishii, Oleg N. Antzutkin, Richard D. Leapman, Nancy W. Rizzo, Fred Dyda, Jennifer Reed, and Robert Tycko. Amyloid Fibril Formation by Aβ16-22, a Seven-Residue Fragment of the Alzheimer's β-Amyloid Peptide, and Structural Characterization by Solid State NMR. *Biochemistry*, 39(45):13748–13759, November 2000.

[264] Carolin Seuring, Kartik Ayyer, Eleftheria Filippaki, Miriam Barthelmess, Jean-Nicolas Longchamp, Philippe Ringler, Tommaso Pardini, David H. Wojtas, Matthew A. Coleman, Katerina Dörner, Silje Fuglerud, Greger Hammarin, Birgit Habenstein, Annette E. Langkilde, Antoine Loquet, Alke Meents, Roland Riek, Henning Stahlberg, Sébastien Boutet, Mark S. Hunter, Jason Koglin, Mengning Liang, Helen M. Ginn, Rick P. Millane, Matthias Frank, Anton Barty, and Henry N. Chapman. Femtosecond X-ray coherent diffraction of aligned amyloid fibrils on low background graphene. *Nature Communications*, 9(1):1836, May 2018.

[265] Daniel Miguel Angel Villalobos Acosta, Brenda Chimal Vega, Jose Correa Basurto, Leticia Guadalupe Fragoso Morales, and Martha Cecilia Rosales Hernández. Recent Advances by In Silico and In Vitro Studies of Amyloid-β 1-42 Fibril Depicted a S-Shape Conformation. *International Journal of Molecular Sciences*, 19(8), August 2018.

[266] Theint Theint, Yongjie Xia, Philippe S. Nadaud, Dwaipayan Mukhopadhyay, Charles D. Schwieters, Krystyna Surewicz, Witold K. Surewicz, and Christopher P. Jaroniec. Structural Studies of Amyloid Fibrils by Paramagnetic Solid-State Nuclear Magnetic Resonance Spectroscopy. *Journal of the American Chemical Society*, 140(41):13161–13166, October 2018.

[267] Jie Zheng, Hyunbum Jang, Buyong Ma, Chung-Jun Tsai, and Ruth Nussinov. Modeling the Alzheimer Aβ17-42 Fibril Architecture: Tight Intermolecular Sheet-Sheet Association and Intramolecular Hydrated Cavities. *Biophysical Journal*, 93(9):3046–3057, November 2007.

[268] Fahimeh Baftizadeh, Xevi Biarnes, Fabio Pietrucci, Fabio Affinito, and Alessandro Laio. Multidimensional View of Amyloid Fibril Nucleation in Atomistic Detail. *Journal of the American Chemical Society*, 134(8):3886–3894, February 2012.

[269] Joan-Emma Shea and Zachary A. Levine. Studying the Early Stages of Protein Aggregation Using Replica Exchange Molecular Dynamics Simulations.

In David Eliezer, editor, *Protein Amyloid Aggregation: Methods and Protocols*, Methods in Molecular Biology, pages 225–250. Springer New York, New York, NY, 2016.

[270] Yiming Wang, Qing Shao, and Carol K. Hall. N-terminal prion protein peptides (PrP(120–144)) form parallel in-register β-sheets via multiple nucleation-dependent pathways. *Journal of Biological Chemistry*, 292(50):20655–20655, December 2017.

[271] Buyong Ma and Ruth Nussinov. Molecular dynamics simulations of alanine rich β-sheet oligomers: Insight into amyloid formation. *Protein Science*, 11(10):2335–2350, 2009.

[272] Jie Zheng, Buyong Ma, Chung-Jung Tsai, and Ruth Nussinov. Structural stability and dynamics of an amyloid forming peptide gnnqqny from the yeast prion sup-35. *Biophysical Journal*, 91(3):824–833, 2006.

[273] Buyong Ma and Ruth Nussinov. Stabilities and conformations of alzheimer's β-amyloid peptide oligomers (aβ_{16-22}, aβ_{16-35}, aβ_{10-35}): Sequence effects. *Proceedings of the National Academy of Sciences of the United States of America*, 99(22):14126–14131, 2002.

[274] Nurit Haspel, David Zanuy, Buyong Ma, Hail Wolfson, and Ruth Nussinov. A comparative study of amyloid fibril formation by residues 15-19 of the human calcitonin hormone: A single β-sheet model with a small hydrophobic core. *Journal of Molecular Biology*, 345(4):1213–1227, 2005.

[275] Weihua Zheng, Min-Yeh Tsai, Mingchen Chen, and Wolynes Peter G. Exploring the aggregation free energy landscape of the amyloid-β protein (1-40). *Proceedings of the National Academy of Sciences of the United States of America*, 113(42):11835–11840, 2016.

[276] Chun Wu and Joan-Emma Shea. Coarse-grained models for protein aggregation. *Current Opinion in Structural Biology*, 21:209–220, 2011.

[277] Christopher M. Dobson. Protein folding and misfolding. *Nature*, 426:884–890, December 2003.

[278] D Thirumalai, DK Klimov, and RI Dima. Emerging ideas on the molecular basis of protein and peptide aggregation. *Current Opinion in Structural Biology*, 13(2):146–159, April 2003.

[279] Stefan Auer, Christopher M. Dobson, Michele Vendruscolo, and Amos Maritan. Self-templated nucleation in peptide and protein aggregation. *Physical Review Letters*, 101:258101, 2008.

[280] Ricardo Pellarin, Enrico Guarnera, and Amedeo Caflisch. Pathways and intermediates of amyloid fibril formation. *Journal of Molecular Biology*, 374(4):917–924, 2007.

[281] Giovanni Bellesia and Joan-Emma Shea. Effect of β-sheet propensity on peptide aggregation. *The Journal of Chemical Physics*, 130(14):145103, 2009.

[282] Fumio Oosawa and Sho Asakura. *Thermodynamics of the polymerization of protein*. Academic Press, London; New York, 1975.

[283] Maarten G. Wolf, Jeroen van Gestel, and Simon W. de Leeuw. Modeling Amyloid Fibril Formation. In Ehud Gazit and Ruth Nussinov, editors, *Nanostructure Design: Methods and Protocols*, Methods in Molecular Biology™, pages 153–179. Humana Press, Totowa, NJ, 2008.

[284] Jeremy D. Schmit, Kingshuk Ghosh, and Ken Dill. What Drives Amyloid Molecules To Assemble into Oligomers and Fibrils? *Biophysical Journal*, 100(2):450–458, January 2011.

[285] Binwu Zhao, Martien A. Cohen Stuart, and Carol K. Hall. Navigating in foldonia: Using accelerated molecular dynamics to explore stability, unfolding and self-healing of the β-solenoid structure formed by a silk-like polypeptide. *PLOS Computational Biology*, 13(3):e1005446, 2017.

[286] Hong Li, John J. Dunn, Benjamin J. Luft, and Catherine L. Lawson. Crystal structure of lyme disease antigen outer surface protein a complexed with an fab. *Proceedings of the National Academy of Sciences of the United States of America*, 94(8):3584–3589, 1997.

[287] Luca Larini and Joan-Emma Shea. Role of β-hairpin formation in aggregation: The self-assembly of the amyloid-β(25-35) peptide. *Biophysical Journal*, 103(3), 2012.

[288] Christian J. Pike, Andrea J. Walencewicz-Wasserman, Joseph Kosmoski, David H. Cribbs, Charles G. Glabe, and Carl W. Cotman. Structure-activity analyses of β-amyloid peptides: Contributions of the β25-35 region to aggregation and neurotoxicity. *Journal of Neurochemistry*, 64(1):253–265, 1995.

[289] Victor Munoz, Peggy A. Thompson, James Hofrichter, and William A. Eaton. Folding dynamics and mechanism of β-hairpin formation. *Nature*, 390(6656):196–199, November 1997.

[290] Gul H. Zerze, Bilge Uz, and Jeetain Mittal. Folding thermodynamics of β-hairpins studied by replica-exchange molecular dynamics simulations. *Proteins: Structure, Function, and Bioinformatics*, 83(7):1307–1315, July 2015.

[291] Ananat K. Paravastu, Richard D. Leapman, Wai-Ming Yau, and Robert Tycko. Molecular structural basis for polymorphism in alzheimer's β-amyloid fibrils. *Proceedings of the National Academy of Sciences of the United States of America*, 105(47):19349–18354, 2008.

[292] Winnie Yong, Aleksey Lomakin, Marina D. Kirkitadze, David B. Teplow, Sow-Hsin Chen, and George B. Benedek. Structure determination of micelle-like intermediates in amyloid β-protein fibril assembly by using small angle neutron scattering. *Proceedings of the National Academy of Sciences*, 99(1):150–154, 2002.

[293] Xiaoyan Hu, Scott L. Crick, Guojun Bu, Carl Frieden, Rohit V. Pappu, and Jin-Moo Lee. Amyloid seeds formed by cellular uptake, concentration, and aggregation of the amyloid-beta peptide. *Proceedings of the National Academy of Sciences*, 106(48):20324–20329, 2009.

[294] D. Strozyk, K. Blennow, L. R. White, and L. J. Launer. Csf aβ 42 levels correlate with amyloid-neuropathology in a population-based autopsy study. *Neurology*, 60(4):652–656, 2003.

[295] Mercedes Novo, Sonia Freire, and Wajih Al-Soufi. Critical aggregation concentration for the formation of early Amyloid-β (1-42) oligomers. *Scientific Reports*, 8(1):1783, January 2018.

[296] Barak Raveh, Ofer Rahat, Ronen Basri, and Gideon Schreiber. Rediscovering secondary structures as network motifs—an unsupervised learning approach. *Bioinformatics*, 23(2):e163–e169, 2007.

[297] Azadeh Ghanbari, Michael C. Bohm, and Florian Muller-Plathe. A simple reverse mapping procedure for coarse-grained polymer models with rigid side groups. *Macromolecules*, 44(13):5520–5526, 2011.

[298] Tsjerk A. Wassenaar, Kristyna Pluhackova, Rainer A. Böckmann, Siewert J. Marrink, and D. Peter Tieleman. Going backward: A flexible geometric

approach to reverse transformation from coarse grained to atomistic models. *Journal of Chemical Theory and Computation*, 10(2):676–690, 2014.

[299] Tristan Bereau and Kurt Kremer. Protein-backbone thermodynamics across the membrane interface. *The Journal of Physical Chemistry B*, 120(26):6391–6400, 2016.

[300] Leandro E. Lombardi, Marcelo A. Martí, and Luciana Capece. Cg2aa: backmapping protein coarse-grained structures. *Bioinformatics*, 32(8):1235–1237, 2016.

[301] P.G. de Gennes and J. Prost. *The Physics of Liquid Crystals*, volume 83 of *International Series of Monographs on Physics*. Clarendon, Oxford, second edition, 1993.

[302] L. Martinez, R. Andrade, E. G. Birgin, and J. M. Martinez. Packmol: A package for building initial configurations for molecular dynamics simulations. *Journal of Computational Chemistry*, 30(13):2157–2164, 2009.

[303] Dmitrij Frishman and Patrick Argos. Knowledge-based protein secondary structure assignment. *Proteins: Structure, Function, and Bioinformatics*, 23(4):566–579, 1995.

[304] Robert W. Zwanzig. High-temperature equation of state by a perturbation method. i. nonpolar gases. *The Journal of Chemical Physics*, 22(8):1420–1426, 1954.

[305] Thomas Dannenhoffer-Lafage, Andrew D. White, and Gregory A. Voth. A direct method for incorporating experimental data into multiscale coarse-grained models. *Journal of Chemical Theory and Computation*, 12(5):2144–2153, 2016.

[306] Thomas T. Foley, M. Scott Shell, and W. G. Noid. The impact of resolution upon entropy and information in coarse-grained models. *The Journal of Chemical Physics*, 143(24):243104, 2015.

[307] John Hopcroft and Robert Tarjan. Algorithm 447: Efficient algorithms for graph manipulation. *Commun. ACM*, 16(6):372–378, June 1973.