Perspective: Advances, Challenges, and Insight for Predictive Coarse-grained Models

W. G. Noid*

Department of Chemistry, The Pennsylvania State University, University Park, PA 16802, USA

E-mail: wnoid@chem.psu.edu

Abstract

By averaging over atomic details, coarse-grained (CG) models provide profound computational and conceptual advantages for studying soft materials. In particular, bottom-up approaches develop CG models based upon information obtained from atomically detailed models. At least in principle, a bottom-up model can reproduce all the properties of an atomically detailed model that are observable at the resolution of the CG model. Historically, bottom-up approaches have accurately modeled the structure of liquids, polymers, and other amorphous soft materials, but have provided lower structural fidelity for more complex biomolecular systems. Moreover, they have also been plagued by unpredictable transferability and a poor description of thermodynamic properties. Fortunately, recent studies have reported dramatic advances in addressing these prior limitations. This perspective reviews this remarkable progress, while focusing on its foundation in the basic theory of coarse-graining. In particular, we describe recent insights and advances for treating the CG mapping, for modeling many-body interactions, for addressing the state-point dependence of effective potentials, and even for reproducing atomic observables that are beyond the resolution of the CG model. We also outline outstanding challenges and promising directions in the field. We anticipate that the synthesis of rigorous theory and modern computational tools will result in practical bottom-up methods that are not only accurate and transferable, but also provide predictive insight for complex systems.

1 Introduction

Particle-based coarse-grained (CG) models provide a powerful tool for studying proteins,[1] polymers,[2] and other soft materials.[3] By representing systems in reduced detail, CG models provide the necessary computational efficiency for simulating length- and time-scales that remain far beyond the scope of conventional atomically detailed simulations. For instance, while Bowman and coworkers effectively constructed the first exascale computer to simulate atomically detailed models of key proteins and protein complexes from the SARS-CoV-2 proteome,[4] Voth and coworkers simulated a CG model of the entire coronavirus capsid with far fewer computational resources.[5] At the same time, by retaining essential molecular details, CG models provide insight into conformations, fluctuations, and interactions that are hidden from continuum or mean field theories. Moreover, CG models promote broader participation in computational science by significantly reducing the resources required for simulating compelling phenomena.

Since computational efficiency is a primary motivation, one might suspect that rapid advances in computational methods and resources will soon render low-resolution CG models obsolete. However, CG models will likely remain important computational tools far into the future. As Deserno eloquently discussed, the computational effort necessary for reaching equilibria often grows extremely rapidly with system size and complexity.[6] For instance, the computational effort required for equilibrating simple lipid membranes with a characteristic length, L, scales as L^6.[1] As simulations address systems of ever-increasing scale and complexity, the efficiency of CG models will become increasingly important for minimizing finite size effects, for reaching equilibrium, and for obtaining statistically significant results.[7–9] Furthermore, the efficiency of CG models enables large-scale, high-throughput simulation studies for systematically and exhaustively exploring the influence of various experimental and model parameters. By representing molecules in reduced detail and adopting a "coarse-

[1]The L^6 scaling arises because the size of the simulated system scales as L^2 (i.e., the area of the membrane), while the time scale for equilibrating an undulation of length L decays on a time scale proportional to L^4.[6]

grained periodic table," CG models even reduce the dimension of chemical space, which dramatically reduces the computational effort necessary for exploring this space.[10,11] Consequently, CG models provide an ideal tool for investigating phase behavior, self-assembly, and other emergent mesoscale phenomena,[12,13] for elucidating basic biophysical principles,[14,15] and for establishing functional relations between molecular and material properties.[16] Accordingly, we anticipate that CG models will continue to grow in popularity as long as human imagination outpaces advances in computational capabilities.

More importantly, CG models provide profound conceptual advantages. As famously quipped, one does not need to consider quarks when modeling bulldozers.[17] Similarly, many phenomena in chemistry and physics can be modeled without considering every atomic detail. Just as cartographers employ a reduced description that indicates only the features necessary for navigating a geographic location, in the same way, theorists adopt a reduced representation that considers only the details necessary for predicting and, even more importantly, understanding a particular physical phenomenon.[18–20] While atomic details tend to obscure insight, the very process of coarse-graining liberates researchers to tailor models for addressing specific questions. Coarse-graining empowers researchers to focus their most valuable resources — their time and intellectual horsepower — on identifying and understanding the essential aspects of a phenomenon.

It is perhaps useful to distinguish several complementary, though certainly not mutually exclusive, philosophies for constructing CG models. Generic "toy" models employ minimal detail and exceptionally simple potentials to investigate the general consequences of basic physical principles.[21,22] Conversely, "chemically specific" models are developed to investigate particular molecular systems. Although an over-simplification, it is often convenient to distinguish "top-down" and "bottom-up" approaches for constructing chemically specific CG models.[23,24] While top-down approaches typically adopt relatively simple potentials that are tuned to match macroscopic thermodynamic properties, such as the bulk density or the liquid-vapor surface tension, bottom-up approaches often employ more complex potentials

that are parameterized with information from atomically detailed simulations. Finally, it is perhaps also useful to distinguish between "pragmatic" and "rigorous" approaches for coarse-graining. While pragmatic approaches rely upon physical arguments and chemical intuition when developing CG models, rigorous approaches adhere more closely to the exact statistical mechanical procedure of formally averaging over atomic degrees of freedom. Of course, it should be emphasized that these distinctions are somewhat blurry and that many models reflect, e.g., both top-down and bottom-up aspects. For instance, the popular Martini model exemplifies a pragmatic, hybrid approach: the intermolecular pair potentials have been parameterized in a top-down fashion to match macroscopic thermodynamic properties, while the bonded potentials have been parameterized in a bottom-up manner to match the molecular conformations observed in atomically detailed simulations.[25–27]

From this perspective, rigorous bottom-up approaches provide certain advantages. In particular, rigorous bottom-up approaches benefit from a direct statistical mechanical connection to a high-resolution model for the same system. This multiscale connection provides a rigorous basis for treating the atomic details that are not explicitly present in the CG model when parameterizing, analyzing, and systematically improving bottom-up models.[28] Moreover, it provides a fundamental framework for relating the observable properties of high resolution and CG models.[29,30] Assuming that the high resolution model accurately describes the system of interest,[31] this connection anchors the predictions of the CG model in reality. When this connection is realized, bottom-up CG models become a powerful predictive tool because their simulated structural and thermodynamic properties are consistent with a realistic microscopic model. Such bottom-up models hold unique promise for elucidating the underlying mechanism of mesoscale phenomena because the interactions in the CG model can be directly related to microscopic interactions.

Bottom-up approaches have not yet realized this promise. While they accurately describe the structural properties of relatively amorphous soft materials, such as liquids and polymers, bottom-up approaches have enjoyed less success in modeling biomolecules with complex hi-

erarchical structures.[32] Even more problematically, bottom-up potentials often demonstrate unpredictable transferability: Bottom-up potentials that accurately describe a specific interaction in a particular environment and thermodynamic state point (e.g., the interaction between amino acids that are buried in a folded protein at a given temperature and pressure) may provide a relatively poor description of the same interaction when the thermodynamic state point or environmental context changes.[33,34] Furthermore, and perhaps more surprisingly, bottom-up models that accurately describe the structural properties of soft materials often provide a surprisingly poor description of their thermodynamic properties, such as the internal pressure or cohesive energy.[29,30] Accordingly, it has been humorously suggested that, while bottom-up approaches promise the caviar of CG models, researchers are often limited in practice to Martini soup.[35] (See Fig. 1.)

Fortunately, recent studies have achieved remarkable progress in understanding and addressing these limitations of bottom-up approaches. This perspective attempts to review and synthesize this progress. We attempt to provide a useful, though certainly not exhaustive, survey of recent studies. Rather than artificially distinguishing between "physics" and "data" driven approaches, we instead attempt to integrate these approaches and holistically organize progress around the basic theory for coarse-graining. In so doing, we hope to clarify the fundamental origin of challenges that arise in systematic coarse-graining. Moreover, we hope to identify and highlight promising approaches for rigorously addressing these challenges in practice. For a more comprehensive introduction to bottom-up approaches, the reader is referred to a number of outstanding earlier reviews.[23,28,36–45] More recently, several excellent reviews have discussed methods and theories for bottom-up coarse-graining,[46–48] as well as their application for biomolecular[49–51] and polymeric systems.[52–56] Nevertheless, it is hoped that this perspective will still provide useful insight.

The remainder of this review is organized as follows. We first discuss the choice and consequences of the CG representation. We next discuss methods for determining the conservative potential that describes interactions and governs sampling in the CG model. The

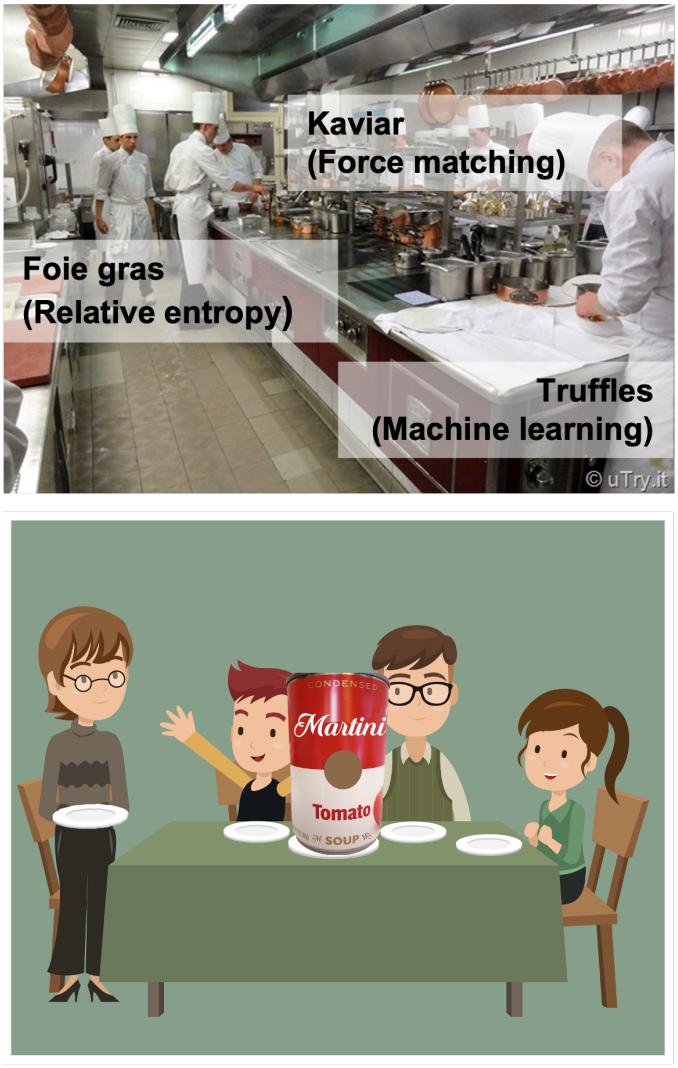


Figure 1: A perspective on coarse-graining kitchens.[35] Top: Rigorous bottom-up models are sometimes perceived as appetizing delicacies that are impractical and inaccessible to non-experts. [Photograph courtesy of Amy Tong uTry.it.] Bottom: Pragmatic top-down models are sometimes perceived as very palatable alternatives that are both affordable and accessible. [Image from https://www.vecteezy.com/vector-art/208237-family-on-dinner-table. Modifications by M. Lesniewski.]

following two sections then discuss the structural fidelity of bottom-up models, as well as their transferability and thermodynamic properties. Due to space and time constraints, we do not discuss the challenges or progress in modeling dynamical properties with bottom-up CG models. We refer readers to several excellent recent reviews on this very interesting topic.[57–59] Finally, we conclude by attempting to concisely summarize the key advances and insights of recent work, as well as highlighting a few emerging challenges and opportunities

in the field.

2 Coarse-grained representation

The first step in constructing a CG model is to determine a CG "representation" for the system of interest. Although several early studies proposed systematic methods for representing systems in CG detail,[60–62] in practice most researchers rely upon their intuition. Consequently, White and coworkers constructed a dataset of annotated "expert" CG representations and trained a graph convolutional neural network to apply this expert intuition in determining representations for new molecules.[63] Similarly, recent studies have developed automated, high-throughput methods for determining CG representations based upon the chemical fragments employed in Martini models.[64,65] In recent years, though, several studies have critically revisited the choice and consequences of the CG representation.

2.1 Mapped ensemble

In systematic bottom-up approaches, the CG representation not only specifies the number and character of the particles that are explicitly modeled, but also defines a (usually) linear mapping, $\mathbf{M} : \mathbf{r} \to \mathbf{R} = \mathbf{M}(\mathbf{r})$, that determines a unique CG configuration, $\mathbf{R}$, for each all-atom (AA) configuration, $\mathbf{r}$.[2] The mapping is critically important because it determines the dynamic, structural, and thermodynamic properties that can be directly observed at the CG resolution.[29] In particular, bottom-up approaches often focus on reproducing the "mapped ensemble" that is obtained by mapping the AA ensemble to the CG representation. Given an AA model with an equilibrium distribution, $p_{\mathrm{r}}(\mathbf{r})$, the "mapped distribution,"

$$p_{\mathrm{R}}(\mathbf{R}) = \int \mathrm{d}\mathbf{r}\; p_{\mathrm{r}}(\mathbf{r}) \delta(\mathbf{M}(\mathbf{r}) - \mathbf{R}), \tag{1}$$

[2]For simplicity, we refer to the high resolution model as an all-atom (AA) model. However, the present analysis applies to more general classical high resolution models.

gives the probability (density) that the AA model assigns to each configuration, **R**, in the mapped ensemble.[3]

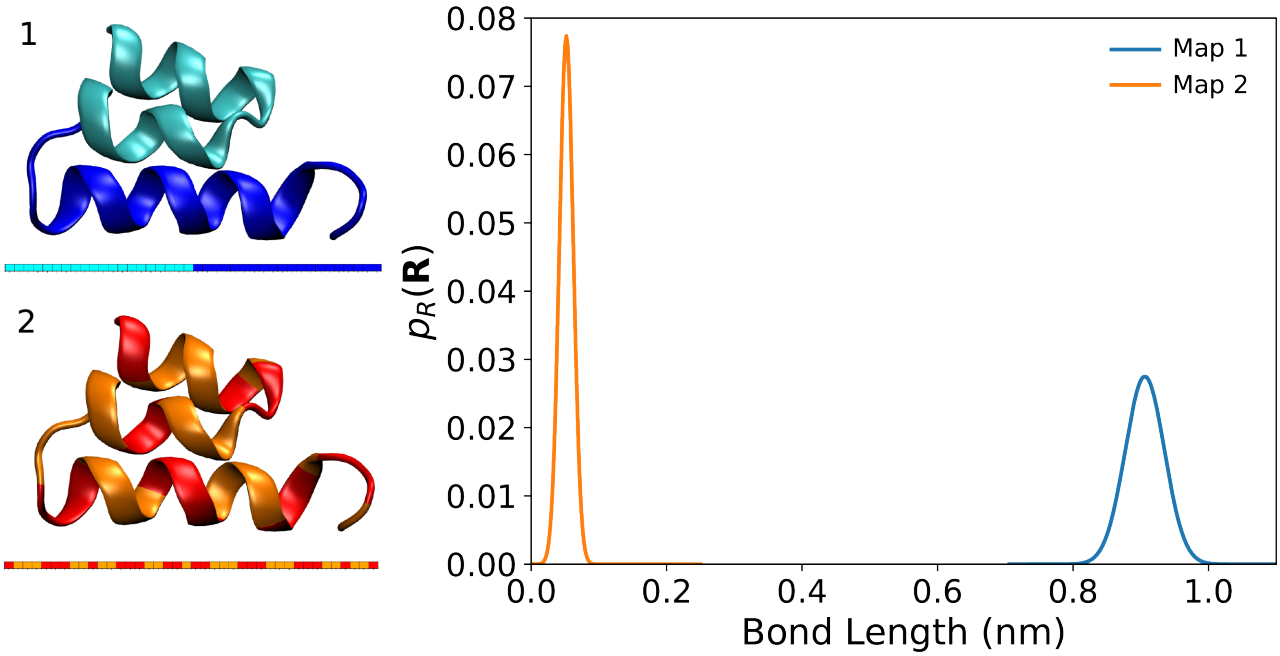


Figure 2: Influence of **M** upon the mapped ensemble. The left panels indicate two different 2-site representations of a simple three helix bundle. Both representations partition the amino acids into disjoint groups and associate a CG site with the mass center of the group. The amino acids that are grouped into the same site are given the same color in both the ribbon cartoon structure and also the 1-D sequence below the structure. The right panel presents the mapped "bond distribution" obtained by mapping the AA ensemble onto each two-site representation. Adapted with permission from Ref. 47. Copyright 2021 Springer Nature.

Figure 2 illustrates the influence of **M** upon the mapped ensemble by comparing two different 2-site CG representations for a small helical protein.[47] Because the first mapping associates the CG sites with distinct structural features that move coherently, the average distance between the two sites is approximately 0.9 nm and the CG "bond" between the two sites samples relatively large fluctuations. In this case, the mapped ensemble nicely preserves large scale motions. Conversely, because the second mapping associates each CG site with an incoherently distributed set of amino acids, the average distance between the two sites is less than 0.1 nm and the corresponding bond distribution is much more narrow. In this

[3]We adopt the notation that AA and CG quantities are represented by lower case and capitalized letters, respectively. Thus, $p_{\mathrm{R}}(\mathbf{R})$ is the probability that the AA model gives to the CG configuration, **R**.

case, the mapped ensemble appears to describe localized, high frequency fluctuations. While these two mappings represent rather extreme possibilities, they clearly demonstrate that the choice of CG mapping can strongly influence the properties of the mapped ensemble. Note that these different mapped ensembles are completely specified by the mapping, $\mathbf{M}$, itself and do not reflect any approximations, e.g., in treating the interaction potential.

2.2 Large scale motions

One intuitively expects that "good" CG mappings should preserve the low frequency, large amplitude motions that are present in AA molecular dynamics (MD) simulations. Indeed, principal component analysis (PCA) demonstrates that the Martini mapping scheme nicely preserves the large amplitude motions that are sampled by lipids in AA simulations.[66] The essential dynamics coarse-graining (ED-CG) method provides a systematic approach for constructing low-resolution representations according to this intuition.[62] The ED-CG method first employs PCA to project an AA trajectory onto the "essential dynamics" subspace of large amplitude motions.[67] The ED-CG method then identifies CG sites with rigid atomic groups that move coherently within this subspace. Recent studies with the ED-CG method have investigated the importance of symmetry for CG representations[68] and developed improved numerical methods for identifying these rigid atomic groups.[69,70]

Two recent studies have applied similar intuition to construct optimal CG representations for atomically-detailed elastic network models (ENMs), which are widely adopted for studying complex biomolecules.[71,72] Both studies employed "decimation" mappings that successively eliminated atoms and then associated each CG site with a specific atom from the high resolution ENM. Inspired by renormalization group approaches,[73,74] Koehl and coworkers constructed a hierarchy of CG ENMs by systematically eliminating high frequency atoms from the AA network.[75] Conversely, Potestio and coworkers[20] determined the mapping for which the effective interactions between the remaining atoms were best described by a CG

ENM.[4] In both cases, the resulting CG models nicely preserved the shape and low frequency fluctuations of the underlying atomic ENM.

Other studies have employed graph-theoretic concepts to determine the CG representation for a molecule based only upon its bonded connectivity. White and coworkers developed a hierarchical graph-based framework for encoding and organizing CG representations that preserve underlying symmetries of the high-resolution model.[76] A specific mapping corresponds to a "slice" through a "tree" of related maps, which can be optimized according to various metrics. de Pablo and coworkers developed an automated graph-based coarse-graining protocol for agglomerating atoms into CG sites by successively contracting edges from the atomically detailed molecular graph.[77] This approach appears quite promising for constructing a hierarchy of CG representations that preserve the low-frequency motions of complex molecules, while requiring only minimal information.

2.3 Information content

Alternatively, rather than focusing on large amplitude motions, one may consider the configurational information that is preserved by the CG mapping. The Kullback-Leibler (KL) divergence,[78] which is also known as the relative entropy, provides a useful metric for quantifying this information.[79] Specifically, given two probability densities, $p_1(x)$ and $p_2(x)$, the KL divergence is defined by

$$\mathrm{D}[p_1||p_2] \equiv \int \mathrm{d}x \, p_1(x) \ln \left[\frac{p_1(x)}{p_2(x)} \right] . \tag{2}$$

By the Gibbs inequality,[80] $\mathrm{D}[p_1||p_2] \geq 0$ and only vanishes when $p_1 = p_2$ (almost) everywhere. Consequently, $\mathrm{D}[p_1||p_2]$ quantifies the difference between p_1 and p_2, although it cannot be

[4] As discussed further below, the process of exactly integrating out a fraction of the microscopic degrees of freedom generally results in a "renormalized" effective potential for the remaining degrees of freedom that does not have the same form as the original microscopic potential.[73] In particular, the effective potential obtained by exactly integrating atoms out of an ENM does not have the same form as the original ENM potential.[20]

considered a formal "distance" metric because it is not symmetric, i.e., $\mathrm{D}[p_1||p_2] \neq \mathrm{D}[p_2||p_1]$.[5] Moreover, because $\ln[p_1(x)/p_2(x)]$ can be interpreted as the information available at x for discriminating between p_1 and p_2,[78] the KL divergence can be interpreted as the average of this information weighted according to p_1. If one considers the uniform distribution, $q_{\mathrm{r}}(\mathbf{r}) = V^{-n}$, for n atoms to contain no information, then one can quantify the information, H_{AA}, present in the AA equilibrium configuration distribution, $p_{\mathrm{r}}(\mathbf{r})$, by

$$H_{\mathrm{AA}} \equiv \mathrm{D}[p_{\mathrm{r}}||q_{\mathrm{r}}] = \int \mathrm{d}\mathbf{r}\, p_{\mathrm{r}}(\mathbf{r}) \ln\left[\frac{p_{\mathrm{r}}(\mathbf{r})}{q_{\mathrm{r}}(\mathbf{r})}\right], \tag{3}$$

which corresponds to the (negative of the) excess configurational entropy of the AA model.

Similarly, one can quantify the information, H_{CG}, present in the mapped ensemble as the KL divergence between the mapped distribution, $p_{\mathrm{R}}(\mathbf{R})$, and the minimally informative, uniform distribution, $q_{\mathrm{R}}(\mathbf{R}) = V^{-N}$, for N CG sites

$$H_{\mathrm{CG}} \equiv \mathrm{D}[p_{\mathrm{R}}||q_{\mathrm{R}}] = \int \mathrm{d}\mathbf{R}\, p_{\mathrm{R}}(\mathbf{R}) \ln\left[\frac{p_{\mathrm{R}}(\mathbf{R})}{q_{\mathrm{R}}(\mathbf{R})}\right] \geq 0, \tag{4}$$

which corresponds to the excess entropy of the AA model when it is observed at the CG resolution. H_{CG} increases as p_{R} becomes less uniform and, thus, "more informative." According to this metric, the mapped ensemble for Map 1 in Fig. 2 is relatively "information poor" because the corresponding mapped distribution is relatively broad and, therefore, contains large uncertainty in the CG bond length. Conversely, the mapped ensemble for Map 2 is more informative because the mapped distribution is much more narrow.

For any CG configuration, $\mathbf{R}$, there exists an entire subensemble of AA configurations that all map to $\mathbf{R}$. This subensemble is characterized by the conditional probability distribution $p_{\mathrm{r}|\mathrm{R}}(\mathbf{r}|\mathbf{R}) = p_{\mathrm{r}}(\mathbf{r})\delta(\mathbf{M}(\mathbf{r}) - \mathbf{R})/p_{\mathrm{R}}(\mathbf{R})$, which is the probability (density) for sampling an AA

[5] Sanov's theorem implies another important property of the KL divergence.[81] Consider $n_s \gg 1$ statistically independent samples $\{x_1, x_2, \ldots, x_{n_s}\}$ drawn from a probability density, $p(x)$. These n_s samples determine an empirical probability density, $\hat{p}(x) \equiv n_s^{-1} \sum_{i=1}^{n_s} \delta(x - x_i)$, that corresponds to the frequency of observing x among the n_s samples. The likelihood, $L[\hat{p}]$, of observing the empirical probability distribution, $\hat{p}(x)$, exponentially decays with n_s at a rate specified by $\mathrm{D}[\hat{p}||p]$, i.e., $L[\hat{p}] \approx \exp[-n_s \mathrm{D}[\hat{p}||p]]$.

configuration $\mathbf{r}$ given the condition that it maps to the CG configuration $\mathbf{R}$. The mapped ensemble eliminates all information about this subensemble, effectively replacing $p_{\mathrm{r}|\mathrm{R}}(\mathbf{r}|\mathbf{R})$ with the uniform distribution, $q_{\mathrm{r}|\mathrm{R}}(\mathbf{r}|\mathbf{R}) = V^{N-n}\delta(\mathbf{M}(\mathbf{r}) - \mathbf{R})$. The information lost from this subensemble can be similarly quantified by the KL divergence between $p_{\mathrm{r}|\mathrm{R}}$ and $q_{\mathrm{r}|\mathrm{R}}$:

$$\mathrm{D}[p_{\mathrm{r}|\mathrm{R}}||q_{\mathrm{r}|\mathrm{R}}](\mathbf{R}) \equiv \int \mathrm{d}\mathbf{r}\; p_{\mathrm{r}|\mathrm{R}}(\mathbf{r}|\mathbf{R}) \ln\left[\frac{p_{\mathrm{r}|\mathrm{R}}(\mathbf{r}|\mathbf{R})}{q_{\mathrm{r}|\mathrm{R}}(\mathbf{r}|\mathbf{R})}\right]. \tag{5}$$

Again by the Gibbs inequality, Eq. (5) is non-negative and attains its minimum (i.e., 0) if and only if $p_{\mathrm{r}|\mathrm{R}}(\mathbf{r}|\mathbf{R}) = q_{\mathrm{r}|\mathrm{R}}(\mathbf{r}|\mathbf{R})$, i.e., when all AA configurations that map to $\mathbf{R}$ have equal probability. Increasing $\mathrm{D}[p_{\mathrm{r}|\mathrm{R}}||q_{\mathrm{r}|\mathrm{R}}](\mathbf{R})$ corresponds to increasing the detail stored in $p_{\mathrm{r}|\mathrm{R}}$ and, thus, reducing the effective degeneracy of AA configurations that map to $\mathbf{R}$. The mapping entropy, H_{map},[6] is defined by averaging Eq. (5) over the mapped ensemble,

$$H_{\mathrm{map}} = \int \mathrm{d}\mathbf{R}\; p_{\mathrm{R}}(\mathbf{R})\mathrm{D}[p_{\mathrm{r}|\mathrm{R}}||q_{\mathrm{r}|\mathrm{R}}](\mathbf{R}) \;\; \geq 0 \tag{6}$$

which quantifies the total information lost from the mapped ensemble.[82–85]

For any CG mapping, $\mathbf{M}$, the total configurational information, H_{AA}, present in the AA distribution can be decomposed according to an entropy chain rule[79,84]

$$H_{\mathrm{AA}} = H_{\mathrm{CG}} + H_{\mathrm{map}}. \tag{7}$$

While H_{CG} and H_{map} both depend upon the mapping, H_{AA} is independent of $\mathbf{M}$. Consequently, Eq. (7) implies a fundamental trade-off regarding the impact of the mapping upon configurational information. Specifically, the fixed information in the AA configuration distribution, p_{r}, is partitioned between the mapped distribution, p_{R}, and the conditioned distribution, $p_{\mathrm{r}|\mathrm{R}}$. Mapping 1 in Fig. 2 and other maps that give rise to relatively uninformative mapped ensembles (i.e., low H_{CG}) correspond to relatively informative conditioned

[6] While early studies[82–84] introduced a negative mapping entropy, $S_{\mathrm{map}} \leq 0$, here we follow the notation of Potestio and coworkers[85] by defining $H_{\mathrm{map}} = -S_{\mathrm{map}} \geq 0$.

distributions (i.e., high H_{map}) and, thus, CG configurations with relatively low degeneracy. Conversely, mapping 2 and other maps that give rise to relatively informative mapped ensembles (i.e., high H_{CG}) correspond to relatively uninformative conditioned distributions (i.e., low H_{map}) and, thus, CG configurations with relatively high degeneracy. While it is generally challenging to quantify entropic quantities, van der Vegt and coworkers recently employed the two-phase thermodynamics (2PT) method[7] to estimate the information contained in the mapped ensemble.[88]

Potestio and coworkers have provided important insight into these considerations for decimation mappings that define each site by a single atom.[85] By employing a second cumulant approximation to estimate $\mathrm{D}[p_{\mathrm{r}|\mathrm{R}}||q_{\mathrm{r}|\mathrm{R}}](\mathbf{R})$ from potential energy fluctuations, they estimated H_{map} based upon statistics from AA protein simulations. Interestingly, even though these simulations considered single proteins, the maps with minimum information loss (i.e., with minimum H_{map}) highlighted atoms that mediate biologically important interactions. Conversely, CG representations that identified sites with α carbons along the protein backbone resulted in relatively high information loss (i.e., high H_{map}). These C-α mappings, which are widely adopted in CG protein models,[89] presumably correspond to highly structured conditioned distributions, $p_{\mathrm{r}|\mathrm{R}}$, since the atomic structure of the protein backbone can be accurately reconstructed from the α carbon coordinates.[90,91] Subsequently, Potestio and coworkers[92,93] related the mapping entropy to "resolution" and "relevance" metrics that have been employed to characterize deep learning.[94] As they have emphasized,[46] fundamental insights into the mapping operator may prove useful not only for determining CG representations, but also for identifying order parameters to analyze and bias molecular simulations, and even much more generally for understanding complex interacting systems, such as economic markets[92] or social networks.[95,96]

Conversely, Gómez-Bombarelli and coworkers have employed variational autoencoders (VAEs) to determine CG mappings.[97,98] This VAE framework simultaneously determines

[7]The 2PT method[86,87] estimates the entropic properties of liquids as a weighted combination of contributions from solid-like vibrations and gas-like diffusion.

both an encoder (i.e., a CG mapping) that transforms an AA configuration into a CG configuration, as well as a decoder (i.e., a "back-mapping") that transforms a CG configuration into an AA configuration. In particular, these studies trained the VAE to learn a CG representation that allowed for optimal reconstruction of given atomic structures. According to the above reasoning, this approach likely determines representations with relatively high H_{map} that reduce the degeneracy of AA configurations mapping to a given CG configuration. Interestingly, regularizing the VAE loss function with the magnitude of the instantaneous mapped forces appeared important for obtaining mappings that were consistent with physical intuition. This regularization appears to favor mappings that give rise to smooth effective potentials, which presumably correspond to relatively unstructured mapped ensembles with relatively low H_{CG}.

2.4 Further considerations

Foley et al. have provided a complementary perspective on the general properties of CG mappings.[96] By adopting the Gaussian network model (GNM)[99,100] as an analytically tractable high resolution model for protein fluctuations, Foley et al. exactly assessed the intrinsic quality of any CG map, $\mathbf{M}$, based upon two metrics that characterize the corresponding mapped ensemble. Specifically, they considered the information content, I, of the mapped ensemble, which roughly corresponds to H_{CG}, and also the spectral quality, $\mathcal{Q}$, which quantifies the large scale motions in the mapped ensemble and roughly corresponds to the metric optimized by the ED-CG method. By employing Monte Carlo methods to sample and characterize the space of CG representations, they estimated a density of states quantifying the number of maps, $\Omega(I, \mathcal{Q})$, with a given information content and spectral quality. The information content and spectral quality generally decreased with decreasing resolution, although $\mathcal{Q}$ displayed considerably greater sensitivity to the particular details of the mapping. The spectral quality correlated quite strongly with the compactness of CG sites and with the modular-

ity[101,102] of the associated clustering.[8] Moreover, $\mathcal{Q}$ and I appeared weakly correlated among high resolution representations, but more strongly anti-correlated at lower resolutions. This is perhaps unsurprising, since $\mathcal{Q}$ and I tend to favor opposite ends of the vibrational density of states. The relatively few low frequency modes generally correspond to large scale motions that are information poor, while the many high frequency modes generally correspond to localized motions that are information rich. Most intriguingly, this study suggested the possibility of a "critical resolution" beyond which a phase transition signifies a qualitative distinction between good and bad CG representations. Since this study considered a restrictive class of CG mappings for a particularly simple microscopic model, future studies should investigate whether these observations generalize to more realistic models. However, Potestio and coworkers recently reported a similar phase transition in the space of CG protein representations and introduced a very promising framework for further exploring this space.[104]

The large majority of these studies have focused on small molecules with relatively little conformational flexibility[76,97] or systems that fluctuate about a well-defined equilibrium conformation.[20,68–70,75,92,93,96] Clementi and coworkers have provided important insights for modeling more complex systems that transition between diverse conformational states.[105] By employing diffusion maps[106] and Markov state methods,[107] they identified coherent domains that persist in microsecond protein simulations with global folding and unfolding transitions.[108,109] In such systems, the optimal mapping may dynamically vary as the system transitions among diverse metastable conformations. Minimal assembly units, which appear similar to "foldons" that are invoked in protein folding theories,[110] may provide a "basis set" for developing dynamically evolving CG representations of such complex systems.

[8] In the case that each atom is associated with a single site, the CG mapping corresponds to clustering nodes in an atomic graph. The modularity is a common metric for assessing the "strength" of communities in complex networks.[103] Specifically, the modularity compares the number of edges within a cluster to the number expected for a random graph.[101,102]

2.5 Back-mapping

Recent studies have also developed new approaches and insights for the closely related problem of back-mapping CG configurations to AA configurations. Back-mapping approaches are practically important not only for high resolution analysis of CG simulations and structures, but also for serial and parallel multiscale schemes[111] that simulate AA models to investigate details that are below the resolution of the CG model.[112] For instance, accurate back-mapping can significantly simplify the coupling between the low- and high-resolution regions[113] that are concurrently simulated in the Adaptive Resolution Scheme (AdResS) method.[114,115]

In comparison to the forward-mapping, $\mathbf{M}$, which determines a unique CG configuration, $\mathbf{R} = \mathbf{M}(\mathbf{r})$, for each AA configuration, $\mathbf{r}$, back-mapping schemes are necessarily more complex and more ambiguous. As already discussed, a single CG configuration, $\mathbf{R}$, corresponds to an entire subensemble of AA configurations, $\mathbf{M}^{-1}(\mathbf{R}) = \{\mathbf{r}|\mathbf{M}(\mathbf{r}) = \mathbf{R}\}$, which is described by the conditional probability distribution, $p_{\mathrm{r}|\mathrm{R}}(\mathbf{r}|\mathbf{R};\mathbf{M})$. Most back-mapping schemes do not explicitly model the degeneracy of AA configurations that map to $\mathbf{R}$. Instead they typically seek a representative AA configuration, $\mathbf{r}^*_{\mathbf{R}} = \mathbf{M}^+(\mathbf{R})$, that (ideally) maximizes $p_{\mathrm{r}}(\mathbf{r}|\mathbf{R})$ for the given $\mathbf{R}$.[9] In practice, back-mapping schemes often first introduce atomic detail by geometrically interpolating atoms between CG sites or by inserting fragments from libraries of atomistic structures.[116–119] The resulting AA configuration is then typically relaxed via energy minimization and possibly short MD simulations that are often restrained by the initial CG configuration.

The forward-mapping, $\mathbf{M}$, impacts both the practical difficulty of finding this representative configuration, $\mathbf{r}^*_{\mathbf{R}}$, as well as the fundamental significance of $\mathbf{r}^*_{\mathbf{R}}$. For instance, it is likely more challenging to apply back-mapping for lower resolution representations. More fundamentally, if the conditional probability distribution, $p_{\mathrm{r}|\mathrm{R}}(\mathbf{r}|\mathbf{R};\mathbf{M})$, is multimodal, then

[9]The notation $\mathbf{M}^+(\mathbf{R})$ indicates that a deterministic back-mapping operator is analogous to a "pseudo-inverse" for $\mathbf{M}(\mathbf{r})$.

$\mathbf{R}$ no longer determines a single representative configuration. In this case, many diverse configurations may make equally important contributions to this subensemble.

Recent studies have reported several interesting back-mapping approaches for determining this representative configuration. For instance, Bussi and coworkers employed steered molecular dynamics[120] to determine a high resolution RNA structure that was consistent with the nucleobase arrangement predicted by their knowledge-based CG model.[121,122] Samaey and coworkers employed the AdResS method to gradually convert CG configurations for complex polymer networks into AA configurations.[123,124] Kremer and coworkers obtained equilibrated high resolution configurations of polystyrene melts by employing a hierarchy of restrained simulations to back-map from a generic soft-sphere blob model to a relatively high resolution, chemically specific bottom-up model and then to an accurate united atom model.[125] Zheng and coworkers developed a Bayesian framework that may prove useful not only for determining $\mathbf{r}_{\mathbf{R}}^{*}$, but also for more generally sampling AA configurations according to $p_{\mathrm{r}}(\mathbf{r}|\mathbf{R})$.[126]

Machine learning (ML) approaches have provided a new toolbox for determining a unique back-mapped configuration, $\mathbf{r}_{\mathbf{R}}^{*}$. In particular, An and Deshmukh investigated the use of artificial neural networks and Gaussian process regression for back-mapping gas phase configurations of hexane and several aromatic ring systems.[127] Laughton and coworkers developed a rather general GLIMPS method for learning a deterministic back-mapping via linear regression and principal component analysis.[128] Doxastakis and coworkers trained a generative adversarial network (GAN) to perform deterministic back-mapping of polymer configurations in analogy to super-resolution image reconstruction,[129] in which the CG and AA configurations correspond to low- and high-resolution images, respectively.[130] Similarly, Harmandaris and coworkers recently trained a convolutional neural network to predict AA configurations for polymer chains by predicting atomic bond vectors conditioned upon the CG coordinates and chemistry of the corresponding monomer.[131] After relaxing local intermolecular interactions, the resulting polymer melt quite accurately matched the equilibrium structural and

thermodynamic properties of the AA polymer model. The fore-mentioned study by Wang and Gómez-Bombarelli employed a VAE framework to simultaneously optimize both the mapping, $\mathbf{M}$, and also a deterministic back-mapping, $\mathbf{M}^{+}$.[97] As already noted, this approach likely determines a CG mapping that corresponds to a sharply peaked conditional distribution, $p_{\mathrm{r}|\mathrm{R}}$.

Perhaps the most exciting advances in back-mapping stem from employing ML approaches to (approximately) sample AA configurations according to the conditioned probability distribution, $p_{\mathrm{r}|\mathrm{R}}$, for a given CG configuration. For instance, Bereau and coworkers[132] developed a GAN framework for sampling $p_{\mathrm{r}|\mathrm{R}}$.[10] This deepBackmap approach introduced each successive atom in a manner conditioned upon the pre-existing local atomic environment and then employed Gibbs sampling to account for non-bonded interactions. The resulting model generated AA configurations that, without additional simulations, described the local packing and many-body structure of polystyrene melts with remarkable fidelity. Subsequently, they demonstrated that the use of local chemical environments and physics-based priors enabled the deepBackmap approach to be transferable between molecular liquids and polymer melts.[134] More recently, Gómez-Bombarelli and coworkers have employed a VAE framework to model $p_{\mathrm{r}|\mathrm{R}}$ for short peptides in implicit solvent.[135] Moreover, Shmilovich and coworkers have employed a VAE to sample $p_{\mathrm{r}|\mathrm{R}}$ in a Markovian fashion that generates a "temporally coherent" back-mapped trajectory that is consistent with not only the structure, but also the energetics and dynamics of the AA model.[136]

3 Interaction potentials

Given an AA model and a CG mapping, the next step is to determine a potential for modeling interactions in the CG model. The ideal interaction potential is the many-body potential of

[10]The basic notion of the GAN framework[133] is to simultaneously train two adversarial networks - a generator, G, and a critic, C. In the context of back-mapping, G attempts to sample atomic configurations according to $p_{\mathrm{r}|\mathrm{R}}$, while C attempts to determine whether a given atomic configuration was generated by G or by the underlying AA model. If successfully trained, G generates a conditioned atomic distribution that cannot be distinguished from $p_{\mathrm{r}|\mathrm{R}}$.

mean force (PMF), W, which may be defined[36,137–139]

$$\exp[-\beta W(\mathbf{R})] = V^{-(n-N)} \int_{V^n} \mathrm{d}\mathbf{r} \exp[-\beta u(\mathbf{r})] \delta(\mathbf{M}(\mathbf{r}) - \mathbf{R}). \tag{8}$$

Here $\beta = 1/k_B T$ is the inverse temperature, V is the volume, n and N are the number of particles in the AA and CG models, respectively, and u is the potential function of the AA model. As analyzed further below, W is the excess Helmholtz free energy[11] associated with the subensemble of AA configurations that map to $\mathbf{R}$. If $W(\mathbf{R})$ is known at a single state point, then canonical simulations with this potential will perfectly reproduce the mapped distribution at that state point. Unfortunately, $W(\mathbf{R})$ is generally a complex many-body function that cannot be exactly determined.

In practice, bottom-up approaches typically approximate W with a relatively simple interaction potential, U. This approximate potential can often be expressed as a sum of terms, U_ζ, each of which governs a particular interaction or bonded degree of freedom:

$$U(\mathbf{R}) = \sum_\zeta \sum_\lambda U_\zeta(\psi_{\zeta\lambda}(\mathbf{R})). \tag{9}$$

Here ζ is a label indicating a particular type of interaction, λ identifies a particular instance of this interaction, and $\psi_{\zeta\lambda}$ is the mechanical degree of freedom describing this interaction.[28] For instance, if ζ is a pair non-bonded interaction, then $U_\zeta = U_2$ is a pair potential, $\lambda = (I, J)$ is a pair of interacting sites, and $\psi_{\zeta\lambda} = R_{IJ}$ is the distance between the pair. Most commonly, U describes non-bonded interactions with central pair potentials, while describing the intramolecular geometry of bonded CG sites with additive bond, angle, and torsion potentials. ML approaches also often adopt a similar form for the approximate potential. In this case, though, the scalar variable, $\psi_{\zeta\lambda}$, may be replaced with a multidimensional

[11] The canonical partition function, q, of a classical model may be expressed $q = c \int \mathrm{d}\mathbf{p} \exp[-\sum_i \mathbf{p}_i^2/2m_i k_B T] z$ where c is a constant, while $z = \int \mathrm{d}\mathbf{r} \exp[-\beta u(\mathbf{r})]$ is the configuration integral. By defining an ideal reference state $u_{\mathrm{id}}(\mathbf{r}) \equiv 0$, the ideal configuration integral is $z_{\mathrm{id}} = V^n$ and the excess contribution to the AA free energy, a_{xs}, is given by $\exp[-\beta\, a_{\mathrm{xs}}] = z/z_{\mathrm{id}} = V^{-n} \int \mathrm{d}\mathbf{r} \exp[-\beta u(\mathbf{r})]$.

feature vector that contains more detailed many-body information about the configuration. Moreover, the sum over interaction potentials, U_ζ, may be replaced by a sum over one-body potentials, U_I, for each site I.

3.1 Structure-based optimization

Researchers continue to invest considerable effort in developing and refining bottom-up approaches for determining interaction potentials that accurately reproduce structural properties of the mapped ensemble.[140] The venerable Inverse Monte Carlo (IMC)[141] and Iterative Boltzmann inversion (IBI)[142] methods remain two of the most useful approaches for solving this inverse problem. For instance, Nordenskiöld and coworkers recently employed the IMC method to parameterize a CG model that realistically describes the structure and interactions of nucleosomal core particles.[143] Both IBI and IMC systematically refine each interaction potential, $U_\zeta(x)$, until simulations with the CG model reproduce the mapped probability distribution, $p_\zeta(x)$, for the corresponding degree of freedom, ψ_ζ, i.e., $P_\zeta(x;U) = p_\zeta(x)$.[12] While IMC explicitly treats the correlations between interactions when updating these potentials, IBI does not account for these correlations, which can lead to practical difficulties in converging the myriad potentials that are necessary for modeling complex systems with many distinct site types.[144–146] Because the resulting pair potentials tend to dramatically over-estimate the internal pressure, they are often modified with a linear pressure correction[13] that can be tuned to match the AA internal pressure while minimally impacting the structural fidelity of the CG model.[142] Recent studies indicate that integral equation methods can improve the efficiency and convergence properties of IBI and IMC.[146,148]

While ML methods have proven useful for parameterizing top-down models to reproduce thermodynamic properties,[149–153] they have also recently been harnessed to parameterize

[12] As noted earlier, we adopt the convention that probabilities determined by the AA model are represented by lower case letters, e.g., p_ζ or p_R, while probabilities determined by a CG model are represented by upper case letters, e.g., $P_\zeta(U)$ or $P_\mathrm{R}(U)$, where U is the interaction potential for the CG model.

[13] This correction modifies the pair potential, $U_2(r) \to U_2(r) + A(1 - r/r_\mathrm{cut})$, where r_cut is the cut-off of U_2 and A is an adjustable parameter that directly contributes to the pressure.[142,147]

physics-based potentials in a bottom-up fashion. For instance, Hajizadeh and coworkers employed a genetic algorithm to parameterize top-down non-bonded potentials that matched temperature-dependent density measurements, while employing an artificial neural network (ANN) to parameterize bottom-up bonded potentials based upon information from united atom polymer simulations.[154] In this work, iterative simulations of trial CG models were avoided by training independent ANNs to predict the conformational and thermodynamic properties of the CG model as a function of the potential parameters. Similarly, Pavan and coworkers combined top-down nonbonded potentials from the Martini model with bottom-up bonded potentials that were optimized via particle swarm optimization (PSO).[155] Subsequently, they employed PSO to optimize a CG lipid model according to a multi-objective function that included both experimentally measured thermodynamic properties (i.e., bilayer thickness and area per lipid) and also structural metrics that were determined from AA simulations.[156] Interestingly, they employed the Wasserstein distance[14] for comparing the structural distributions sampled by the AA and CG models. By combining ML tools for automatic differentiation along with statistical reweighting methods, Differentiable Trajectory Reweighting provides another promising framework for extending structure-based bottom-up methods to incorporate additional information about higher-order structural observables or thermodynamic properties.[158] Conversely, several recent studies have employed ML approaches to determine bottom-up potentials that distinguish the mapped AA ensemble from a "fake" noise distribution.[159–161] In this vein, Jumper et al. parameterized the Upside CG protein model with a contrastive divergence approach[162] that employs Newton's method in a manner reminiscent of the IMC method.[163]

The preceding approaches often rely upon a series of CG simulations to iteratively refine CG potentials. In contrast, several other approaches determine these potentials directly (i.e., noniteratively) from AA simulations or mapped structural distributions. For instance, approaches based upon the Ornstein-Zernicke integral equation[164] can deduce interaction

[14]The Wasserstein distance is commonly employed in modeling mass transport as a quantitative measure of the effort required to transform one mass (or probability) distribution into another.[157]

potentials for molecular and polymeric liquids directly from structural correlations within the mapped ensemble.[148,165–167] By employing approximate closures to account for many-body correlations, these approaches obtain approximate, analytic relationships between pair potentials and the corresponding equilibrium pair correlations for homogeneous, isotropic liquids.[164,167] ANNs have also been employed to directly deduce effective pair potentials from radial distribution functions (rdfs).[168,169] In this case, ANNs are trained to determine the exact, nonlinear relationship between pair potentials and corresponding rdfs. In contrast, the effective force coarse-graining (EFCG) method determines each pair potential by simply averaging the force between the corresponding atomic groups in condensed phase simulations.[170] The conditioned reversible work (CRW) method adopts a similar, but even simpler approach by employing constrained gas phase AA simulations of molecular fragments.[171,172] While the simplicity and computational efficiency of the CRW method are very appealing, practical problems can arise when these gas phase AA simulations do not sample the relevant conformations or, more generally, do not properly describe the local environment that is relevant for condensed phase interactions.[172,173]

3.2 Variational approaches

Two variational principles provide a central foundation for developing and relating a wide range of bottom-up approaches. In the limit that the approximate CG potential, U, is sufficiently flexible, both variational principles achieve their global solution when U equals the exact PMF, W, to within an additive, configuration-independent constant. More generally, these variational principles provide a rigorous framework for calculating, understanding, and systematically improving approximate potentials for bottom-up models.

3.2.1 Relative entropy

The relative entropy (RE) variational principle[45,82] determines the approximate potential, U, by minimizing the KL divergence[78] between the reference mapped distribution, $p_{\mathrm{R}}(\mathbf{R})$,

and the equilibrium distribution for U, $P_{\mathrm{R}}(\mathbf{R};U)$, i.e., by minimizing

$$S_{\mathrm{rel}}[U] \equiv \mathrm{D}[p_{\mathrm{R}}||P_{\mathrm{R}}(U)] = \int_{V^N} \mathrm{d}\mathbf{R}\; p_{\mathrm{R}}(\mathbf{R}) \ln\left[\frac{p_{\mathrm{R}}(\mathbf{R})}{P_{\mathrm{R}}(\mathbf{R};U)}\right] \geq S_{\mathrm{rel}}[W] = 0. \tag{10}$$

As in Eq. (4), $S_{\mathrm{rel}}[U]$ is always nonnegative and vanishes only if $P_{\mathrm{R}}(\mathbf{R};U) = p_{\mathrm{R}}(\mathbf{R})$ for all $\mathbf{R}$, in which case U equals W to within an additive constant. According to footnote 5, minimizing S_{rel} corresponds to maximizing the likelihood that the CG model will reproduce the mapped distribution, $p_{\mathrm{R}}(\mathbf{R})$.[28,82,83] Moreover, if U adopts the simple form of Eq. (9), then minimizing S_{rel} with respect to the interaction potential, U_ζ, leads to the "self-consistent" condition:

$$P_\zeta(x;U) = p_\zeta(x), \tag{11}$$

which ensures that simulations with the CG model will reproduce the distribution along the corresponding degree of freedom in the mapped ensemble.[28,82,83,174]

In practice, S_{rel} is minimized by performing a series of CG simulations with trial approximate potentials and successively revising, e.g., U_2 until simulations with the CG model accurately reproduce the mapped AA rdf.[175,176] Thus, the RE variational principle provides a unifying formalism for iterative structure-based methods that target specific structural correlations.[174,177] Although the Henderson uniqueness theorem[178] and its generalizations[83,179,180] indicate that the resulting pair potentials should be unique, in many cases a wide range of pair potentials can accurately reproduce a target mapped AA rdf.[181,182] Consequently, Shen et al. suggested employing the Fisher information,[79] $K(r,r';U) \equiv \delta^2 S_{\mathrm{rel}}[U]/\delta U_2(r)\delta U_2(r')$, to identify more informative ensembles for determining the CG potential.[183] This is an intuitively appealing idea because K is closely related to a susceptibility matrix, $\delta g_2(r;U)/\delta U_2(r')$, that quantifies the sensitivity of the CG rdf to U_2, and moreover plays a central role in the IMC method.[28,141,177]

Recent studies have developed several interesting extensions of the RE approach. For instance, Aluru and coworkers introduced thermodynamic constraints into the RE variational

principle to reproduce the thermodynamic pressure.[184] Katsoulakis and coworkers developed a path-space RE formalism to optimize CG models for describing equilibrium and non-equilibrium dynamics.[185] Subsequently, they developed an intriguing "predictive" framework that treats the CG coordinates as latent variables in order to model AA properties that are beyond the resolution of the CG model, i.e., properties that cannot be expressed as a function of CG coordinates.[186] Moreover, Pretti and Shell extended the RE formalism to approximate the mapped joint configuration-energy probability distribution, $p_{\mathrm{RE}}(\mathbf{R}, E)$, in an elegant microcanonical formalism that is discussed further below.[187]

3.2.2 Force-matching

The multiscale coarse-graining (MS-CG) force-matching (FM) variational principle[40,188–190] determines U by minimizing the difference between the instantaneous AA force, $\mathbf{f}_I(\mathbf{r})$, acting on each site[15] and the force, $\mathbf{F}_I(\mathbf{R}) = -\partial U/\partial \mathbf{R}_I$, determined by U in the corresponding mapped configuration, i.e., by minimizing

$$\chi^2[U] \equiv \left\langle \frac{1}{3N} \sum_{I=1}^{N} \left| \mathbf{f}_I(\mathbf{r}) - \mathbf{F}_I(\mathbf{M}(\mathbf{r})) \right|^2 \right\rangle \tag{12}$$

$$= \chi^2[W] + \int_{V^N} \mathrm{d}\mathbf{R}\, p_{\mathrm{R}}(\mathbf{R}) \frac{1}{3N} \sum_{I=1}^{N} \left| \bar{\mathbf{f}}_I(\mathbf{R}) - \mathbf{F}_I(\mathbf{R}) \right|^2 \geq \chi^2[W]. \tag{13}$$

Here the angular brackets denote an average over the AA canonical ensemble and $\bar{\mathbf{f}}_I(\mathbf{R}) \equiv \langle \mathbf{f}_I(\mathbf{r}) \rangle_{\mathbf{R}} = -\partial W(\mathbf{R})/\partial \mathbf{R}_I$ is the conditioned mean AA force, which equals the (negative) gradient of the exact PMF.[36,137–139] Since the first term in Eq. (13) is independent of U and the second term compares the gradients of W and U, $\chi^2[U]$ attains its global minimum when U differs from W only by an additive constant.[190] In the more general case that U adopts the simple form of Eq. (9), then the condition for minimizing χ^2 may be expressed as a simple

[15] In the common case that the mapping is linear and each atom contributes to at most one CG site, $\mathbf{f}_I$ is simply the net force on the atoms contributing to site I, i.e., $\mathbf{f}_I(\mathbf{r}) = \sum_{i \in I} \mathbf{f}_i(\mathbf{r})$.[190] More generally, one can define[191–194] - and even optimize[195] - a mapping from AA forces to CG forces.

system of linear equations for the MS-CG force functions, $F_\zeta(x) = -\mathrm{d}U_\zeta(x)/\mathrm{d}x$:

$$b_\zeta(x) = \sum_{\zeta'} \int \mathrm{d}x'\, G_{\zeta\zeta'}(x,x') F_\zeta(x'), \tag{14}$$

where $b_\zeta(x)$ and $G_{\zeta\zeta'}(x,x')$ are quantities that can be directly computed from the mapped ensemble. Equation (14) may be considered a generalization of the Yvon-Born-Green (YBG) integral equation theory[164] for arbitrarily complex potentials.[196–198] The generalized-YBG (g-YBG) framework[199] interprets $b_\zeta(x)$ as an average force along the ψ_ζ degree of freedom in the mapped ensemble when $\psi_\zeta(\mathbf{R}) = x$.[16] Similarly, $G_{\zeta\zeta'}(x,x')$ describes structural correlations between the ψ_ζ and $\psi_{\zeta'}$ degrees of freedom in the mapped ensemble when $\psi_\zeta(\mathbf{R}) = x$ and $\psi_{\zeta'}(\mathbf{R}) = x'$. The condition for minimizing χ^2 then corresponds to a force-balance relation between the AA and CG models. Specifically, when the MS-CG potentials are applied to the mapped ensemble, they reproduce the average AA force, $b_\zeta(x)$, along each relevant degree of freedom, ψ_ζ.[17]

It is important to emphasize that the MS-CG variational principle does not seek to reproduce fluctuating atomic forces or dynamical properties. Rather these fluctuating atomic forces are used as noisy samples for estimating the conditioned mean force, $\overline{\mathbf{f}}_I$, which gives the gradient of the PMF. Indeed, least squares variational principles of fluctuating AA properties provide a rather general framework for approximating conditioned averages at the resolution of the CG model.[201,202]

Several recent studies have provided insight into the numerical properties of the MS-CG variational principle. The first term in Eq. (13), which can be interpreted as statistical noise, quantifies the fluctuations in the AA forces about their conditioned mean.[203] This term is somewhat analogous to the mapping entropy, H_{map}, which quantifies the information

[16] In particular, if $U_\zeta(x) = U_2(R_{IJ})$ is a pair potential, then b_ζ is closely related to the pair mean force, $-w_2'(r)$, that is determined by the pair potential of mean force, $w_2(r) = -k_B T \ln g(r)$.

[17] This condition is not self-consistent because the right hand side of the force-balance equation., Eq. (14), applies the MS-CG potentials to the mapped ensemble and not to the ensemble generated by simulations of the MS-CG model.[200]

lost when viewing the AA ensemble at the CG resolution. Recent studies suggest that this noise term can dominate χ^2 and may become increasingly dominant with coarsening.[204,205] Conversely, the second term in Eq. (13), can be interpreted as the error in the CG potential. This error can be decomposed into bias and variance contributions, which can be used to avoid over-fitting when constructing complex many-body potentials.[194]

Recent studies have also extended the MS-CG formalism in several interesting directions. For instance, Kalligiannaki et al. extended the FM variational principle for non-linear CG mappings.[193] Voth and coworkers generalized the MS-CG formalism to parameterize interactions involving virtual sites that do not correspond to explicit atomic groups,[160,206,207] as well as sites that switch between distinct states in the "ultra" coarse-graining approach.[192,208,209] Dequidt and coworkers introduced a Bayesian trajectory-matching approach that performs force-matching with time-averaged, rather than instantaneous, forces.[210–212] Moreover, Nguyen and Huang extended the MS-CG formalism to optimize orientation-dependent interactions between anisotropic CG sites via both force- and torque-matching variational principles.[213]

3.2.3 Further considerations

Although they initially appear rather distinct, the RE and MS-CG FM variational principles share many similarities. Since the minimizing condition for χ^2 corresponds to a generalization of the YBG integral equation,[196–199] both variational principles can be applied from structural information. Moreover, the FM variational principle can also be given an information theoretic interpretation[18] that is closely related to the RE.[83,97] Both formalisms have been applied to parameterize polymer field theories[214,215] and both have been extended to parameterize CG models for the constant NPT ensemble.[184,216–218] In particular, the pressure-matching approach of Das and Andersen is a natural extension of FM that treats the pressure as the force on the system volume.[216] Similarly, minimizing S_{rel} with respect

[18] As noted when discussing the KL divergence in Eq. (2), $\Phi(\mathbf{R}) = \ln[p_{\text{R}}(\mathbf{R})/P_{\text{R}}(\mathbf{R};U)]$ quantifies the information available in CG configuration $\mathbf{R}$ for distinguishing the mapped ensemble from the equilibrium ensemble for the approximate CG model.[78] Then S_{rel} is the average of $\Phi(\mathbf{R})$, while χ^2 is essentially the average of $|\nabla\Phi(\mathbf{R})|^2$.[83]

to a configuration-independent volume potential, $U_V(V)$, leads to a self-consistent condition ensuring that the CG model reproduces the AA pressure-volume equation of state.[217,218] Furthermore, recent numerical calculations suggest structure-based potentials that minimize S_{rel} are also nearly optimal with respect to χ^2.[204] The similarities are perhaps most striking when considering dynamics, since dynamic generalizations of the RE variational principle result in force-matching conditions.[185,210,219]

There are also important distinctions between the two approaches. In particular, the condition for minimizing S_{rel} is a self-consistent criterion ensuring that simulations with the CG model reproduce specific mapped distribution functions. In contrast, the minimizing condition for χ^2 is not self-consistent and only employs information that is present in the mapped ensemble.[199] Consequently, one can determine the MS-CG potential directly from the mapped ensemble without performing any CG simulations, which is particularly convenient for optimizing the complex many-body ML potentials discussed later. However, due to this lack of self-consistency, the MS-CG potential is not guaranteed to reproduce mapped rdfs or any other structural features present in the mapped ensemble.[199,220] Thus, high structural fidelity is a requisite part of calibrating the CG potential in the RE approach, but provides an independent assessment of the CG potential in the FM approach.

The g-YBG formalism indicates that if the MS-CG model can reproduce the mapped cross-correlations between the ψ_ζ and $\psi_{\zeta'}$ degrees of freedom in the mapped ensemble, i.e., $G_{\zeta\zeta'}(x, x')$, then the MS-CG model should also reproduce the corresponding mapped distributions $p_\zeta(x)$ and $p_{\zeta'}(x)$. Therefore, one expects that errors in the MS-CG model may result from the failure of the approximate potential, U, to describe the many-body cross-correlations present in the mapped ensemble.[200,221,222] In this case, the RE approach ensures that the CG model will reproduce lower order distribution functions, but may do so at the expense of distorting these higher order cross-correlations.[199] Conversely, when the CG potential is sufficiently flexible or when many-body cross-correlations are not very significant, then the RE and FM approaches can both provide very high structural fidelity.

Recent studies have also provided important insight into the practical application of the RE and FM variational principles for modeling systems with complex free energy landscapes based upon imperfect sampling of the mapped ensemble. As already discussed, the FM variational principle attempts to reproduce the local gradients of the PMF (i.e., the mean forces $\bar{\mathbf{f}}_I = -\partial W/\partial \mathbf{R}_I$), based upon information present in the mapped ensemble.[83,223–225] Consequently, one expects that the MS-CG potential will accurately describe the shape (i.e., the gradients) of the PMF in basins that are well sampled by the AA simulation. However, in order to accurately reproduce the relative depths of these minima in the PMF, the MS-CG model must accurately reproduce the gradient of the PMF in the barrier region connecting these basins.[225] In practice, one expects that the mapped ensemble may provide little information about barrier regions that are rarely sampled. If the CG interactions in these barrier regions are significantly different from the interactions in the well sampled basins, then one expects that the resulting MS-CG potential may not accurately reproduce the stability of each basin.[226] In contrast, the RE variational principle employs explicit simulations with the CG potential, U, to ensure that the resulting CG probability distribution, $P(\mathbf{R};U)$, optimally matches the mapped probability distribution, $p_{\mathrm{R}}(\mathbf{R})$. This in and of itself ensures that the CG model properly weights each well sampled basin, while also ensuring that barrier regions have comparably low statistical weight.

This reasoning suggests an important trade-off between data efficiency and computational efficiency when applying the FM and RE variational principles for modeling complex systems in practice.[223–225] Specifically, the FM variational principle is computationally efficient to apply because it can be directly applied from the mapped ensemble and does not require iterative simulations with successive trial CG potentials. However, the FM variational principle has relatively large data requirements and, in particular, may require extensive sampling of the AA model to accurately determine the mean force in rarely sampled barrier regions. Conversely, the RE variational principle has comparatively modest data requirements and, in particular, only requires that the mapped ensemble determine the proper statistical weight

for each basin. At the same time, the RE variational principle may be relatively computationally expensive to apply, as it may require many simulations with trial CG potentials in order to accurately reproduce this mapped ensemble. Another perspective on this trade-off is that FM approaches may need to invest more computational resources in simulating the AA model in order to obtain mapped ensembles with sufficient information about rarely sampled barrier regions, while RE approaches may need to invest more computational resources in simulating the CG model in order to obtain sufficiently accurate CG potentials.

This reasoning also suggests several additional considerations for modeling systems that are characterized by distinct substates separated by large free energy barriers. In particular, FM may prove particularly useful for optimizing potentials that are specific to distinct conformational states and then modulated as the system moves between regions of configuration space.[192,208,209] Moreover, it may prove beneficial to further optimize FM potentials, e.g., either by iterative FM approaches[227–229] or by employing the FM potential as starting point for minimizing the relative entropy.[225] Another possibility is that, given inadequate sampling of barrier regions, FM approaches may benefit from employing more complex potentials that can infer the physics governing these barrier regions based upon well-sampled basins. Future studies should certainly further explore these considerations.

4 Structural fidelity

Ideally, the CG model should perfectly reproduce the configuration distribution that is determined by the AA model and given mapping:

$$P_{\mathrm{R}}(\mathbf{R};U) = p_{\mathrm{R}}(\mathbf{R}). \tag{15}$$

Andersen, Voth, and coworkers originally defined consistency in configuration space by this criterion, as part of a more general criterion for consistency in phase space.[190] Subsequently, Eq. (15) has sometimes been referred to as a "thermodynamic consistency" criterion since

it reflects equilibrium Boltzmann statistics. (However, it is important to recognize that Eq. (15) does not ensure that the CG model will reproduce particular thermodynamic properties.) Durumeric and Voth have elegantly extended this consistency condition to consider virtual sites that cannot be explicitly defined by a mapping from atomic degrees of freedom.[160] Moreover, Rotskoff and coworkers recently introduced the interesting notion of "weak" consistency, which assesses the ability of a CG model and a back-mapping procedure to reproduce averages of AA observables.[230]

As already mentioned, the CG model will achieve the configurational consistency criterion of Eq. (15) when the CG potential, U, equals the exact PMF, W, given by Eq. (8) to within a configuration-independent constant.[19] However, in practice the exact PMF cannot be exactly calculated and must be approximated. Accordingly, researchers continue to investigate and improve the structural fidelity of bottom-up models with approximate potentials. In particular, several recent studies have investigated the impact of the mapping upon the structural fidelity of bottom-up models with simple pair-additive potentials. The limitations of pair-additive potentials have motivated more sophisticated physics-based potentials for describing many-body structural correlations. In turn, this naturally leads to the use of flexible ML architectures for more accurately modeling the many-body PMF.

4.1 Mapping

As discussed above, given an AA model, the CG mapping completely determines the mapped ensemble and, thus, also the PMF. Because CG models typically approximate the PMF with relatively simple potentials, one expects that their structural fidelity may be quite sensitive to the choice of mapping. These considerations are particularly clear when considering intramolecular degrees of freedom. Most studies employ independent potentials to govern each bond, angle, and dihedral degree of freedom in the CG model. Consequently, these models

[19] Because it is independent of configuration, this "constant" is irrelevant for sampling configurations in a given thermodynamic state. However, because this constant can depend upon thermodynamic variables, such as volume, it can make significant contributions to the conjugate thermodynamic properties, such as the internal pressure.[30,36]

are incapable of reproducing complex cross-correlations that may exist between these degrees of freedom in the mapped ensemble.[200,221,231] Early studies by Kremer and coworkers clearly demonstrated that CG models can more accurately describe (mapped) polymer conformations when the CG mapping simplifies the cross-correlations between bonded degrees of freedom.[232,233]

Several recent studies have provided insight into these considerations for modeling intermolecular interactions. For instance, van der Vegt and coworkers demonstrated that the CRW method provides greatest structural fidelity for small symmetric molecules when the mapping preserves this symmetry.[88] Conversely, the CRW approach appears less accurate for lower resolution representations and, moreover, may suffer from sampling difficulties when the CG sites correspond to atomic groups with significant internal flexibility.[173]

Recent studies have also investigated the impact of the mapping upon the structural fidelity of CG models that have been parameterized via the MS-CG FM variational principle.[200,204,234–236] In particular, White and coworkers demonstrated that the structural fidelity of MS-CG models does not necessarily improve when increasing either the resolution or the symmetry of the CG mapping.[236] Savoie and coworkers demonstrated that 1-site[20] MS-CG models accurately describe the intermolecular pair structure for liquids of small alcohols.[204] In contrast, higher resolution MS-CG models provided relatively poor structural fidelity when the mapping associated CG sites with isolated hydroxy groups.[204] Similarly, Voth and coworkers demonstrated that 2-site MS-CG models accurately reproduce the site-site correlation functions for liquids of small carboxylic acids.[235] However, conventional pair-additive MS-CG potentials provided significantly lower structural fidelity for higher resolution 4-site representations that associated separate sites with the carbonyl oxygen and the hydroxyl parts of the carboxylic group.[235] These studies echo early observations that the structural fidelity of MS-CG models for methanol decreases with increasing resolution.[189,234]

These studies suggest that bottom-up methods will generally provide greater structural

[20] In general, m-site models represent each molecule with m sites.

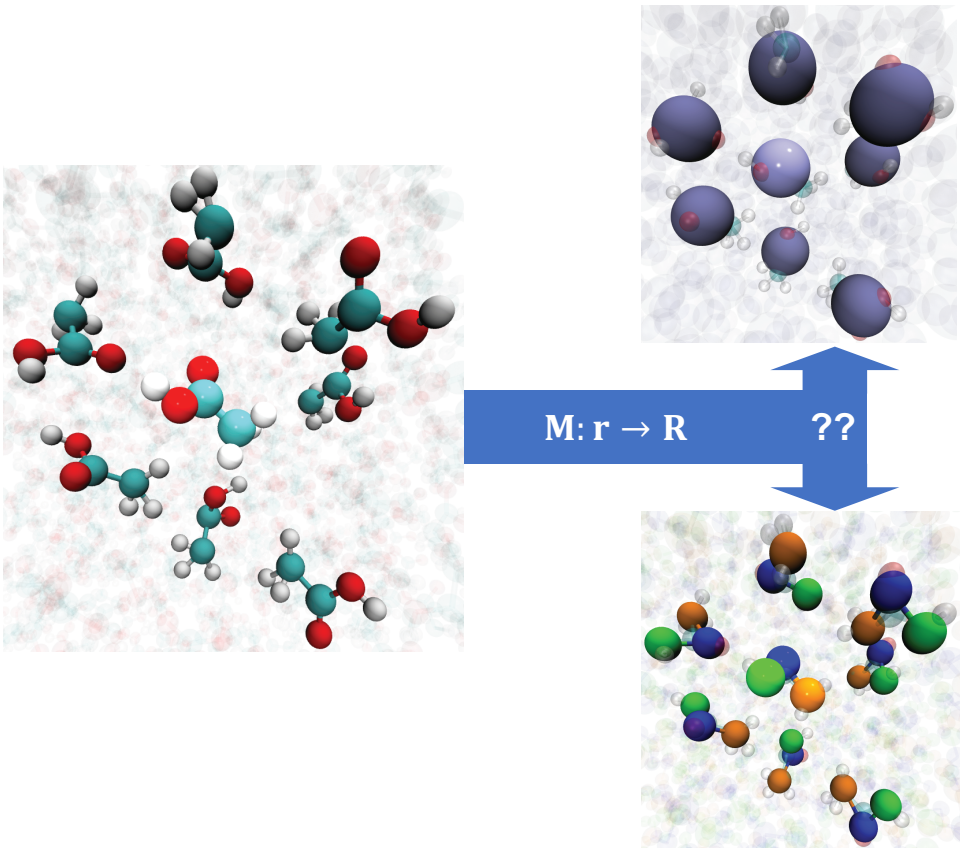


Figure 3: The influence of the mapping upon the complexity of CG configurations. Increasing the resolution of the CG model necessarily increases the density of sites, which likely increases both the magnitude and the the complexity of many-body correlations in the mapped ensemble. Left: An AA configuration of acetic acid highlighting a central molecule and its nearest neighbors. Right: 1-site (top) and 3-site (bottom) CG representations of the same AA configuration.

fidelity when the CG representation simplifies the mapped ensemble. As illustrated in Fig. 3, low resolution representations result in mapped ensembles with a lower density of CG sites and generally weaker many-body correlations. One expects that pair additive potentials can accurately describe such simple mapped ensembles. Conversely, high resolution representations that associate CG sites with small functional groups may result in complex mapped ensembles with strong many-body correlations. In this latter case, pair additive potentials are unlikely to accurately describe the mapped ensemble. Moreover, in this case, iterative structure-based methods, such as IBI and IMC, may determine pair potentials that reproduce mapped site-site rdfs at the expense of distorting these higher order correlations.[199,200]

Indeed, Lyubartsev and Laaksonen proved that such pair-additive CG models have maximum entropy and, thus, the simplest configuration distribution that is consistent with the target mapped rdfs.[237] These considerations suggest that it may be beneficial to develop more general frameworks that simultaneously optimize the CG mapping and approximate potential for self-consistency with each other. The Bayesian approach of Chen and Habeck may represent a first step towards this goal.[238]

4.2 Advanced physics-based potentials

Given these limitations of pair-additive potentials, many studies have investigated new classes of physically motivated potentials for improving the structural fidelity of bottom-up models. For instance, a growing number of studies have revisited the use of anisotropic sites[138,239,240] to model polymer backbones, conjugated organic groups, nucleic acids, and even anisotropic small molecules.[241–250] The interactions between anisotropic sites are often modeled by extensions of the venerable Gay-Berne potential.[251] Importantly, Nguyen and Huang[213] employed both force- and torque-matching variational principles to optimize a considerably more general class of orientation-dependent potentials[252,253] for modeling interactions between anisotropic sites.

Motivated by the success of the mW water model,[254] several studies have adopted Stillinger-Weber-type potentials[255] for modeling orientation-dependent hydrogen-bonding and 3-body interactions.[256–258] Voth and coworkers developed a general framework for incorporating these 3-body interactions into effective pair potentials, which they employed to parameterize an accurate 1-site water model.[205,259] Conversely, Jayaraman and coworkers modeled hydrogen-bonding interactions in cellulose by introducing additional internal sites[260] in a manner similar to the patchy particles that are employed in top-down models of self-assembly[261] or the base-pairing sites in the OxDNA model.[262]

Alternatively, some bottom-up models have retained the simplicity and computational efficiency of spherical sites that interact via pair-additive potentials, but then modulated

these pair potentials as a function of the local environment. For instance, if the conformational space of a protein can be partitioned into distinct conformational substates, one may reasonably hope that a relatively simple potential can accurately describe each substate, although the potentials for distinct substates may be quite different.[229] In particular, one expects that protein side chains may interact very differently in aqueous and hydrophobic environments. This intuition suggests labelling CG sites with dynamic internal state variables that reflect their local environment and then modulating the site-site potentials as a function of this internal state.[177,263] Voth and coworkers pioneered a powerful "ultra" coarse-graining framework for treating these internal states.[192,208,209,264,265] In recent years, they have applied this approach to model hydrogen-bonding liquids,[235] liquid interfaces,[266] and even protein self-assembly.[267] Subsequently, Bereau and Rudzinski[268,269] developed a similar approach in which the CG model "hops" between different effective potentials in analogy to surface hopping approaches for modeling quantum dynamics.[270]

Another interesting approach for improving the structural fidelity of bottom-up pair potentials is to introduce virtual sites that do not directly correspond to any particular atomic groups. While they have been extensively employed in Martini CG models,[271] a rigorous bottom-up framework for parameterizing interactions between virtual sites was lacking until quite recently.[160] When properly defined and parameterized, virtual sites provide a computationally efficient means for incorporating information about the local environment that can improve the description of complex anisotropic interactions and many-body effects. For instance, instead of representing benzene with one spherical site that corresponds to its mass center or with three spherical sites that correspond to specific atoms, the anistropy of the planar ring can be efficiently captured by representing benzene with two virtual sites that define the direction perpendicular to the ring.[206] Voth and coworkers have also employed virtual sites to describe the local solvation environment in implicit solvent models of lipid bilayers.[207,272] Along similar lines, Lafond and Izvekov developed a bottom-up framework for modeling electrostatic interactions with virtual sites that describe effective polar-

izabilities.[273,274] Moreover, the intriguing Upside protein model describes protein sidechains with virtual sites that interact via a complex many-body function of the backbone coordinates.[162,275]

These approaches generally rely upon physical intuition to identify additional terms that should be introduced into the approximate CG potential. In contrast, Liwo and coworkers recently revisited a cluster-cumulant approach[276] to systematically determine higher order interactions in the effective potential for the United Resolution CG protein model.[277–279] More generally, it would be highly desirable to develop automated, systematic methods for identifying new classes of potentials that depend upon simple, computationally efficient local order parameters. Ideally, these order parameters should quantify important structural features that are not properly described by central pair potentials between spherically symmetric sites.[280] One anticipates that ML tools may prove particularly useful for identifying such potentials.[281] Indeed, Dijkstra and coworkers employed linear regression to identify structure-property relations that informed the construction of physics-based potentials for reproducing system-specific, many-body properties.[282] Once these potentials have been identified, they can be optimized either in an ad hoc fashion or by existing variational approaches.[283]

4.3 ML potentials

ML approaches also hold considerable promise for representing, parameterizing, and simulating complex many-body potentials that more accurately approximate the configuration-dependence of the exact many-body PMF. ANNs and other universal function approximators can, at least in principle, quantitatively describe the PMF and, thus, perfectly reproduce the mapped ensemble.

While ML approaches accurately reproduce quantum mechanical energetics with classical atomic potentials,[284] new challenges arise when coarse-graining. For instance, the Born-Oppenheimer approximation directly relates the energy of a classical AA configuration to a quantum mechanical calculation for the corresponding nuclear coordinates. In contrast,

the many-body PMF is not an observable that can be directly calculated from a single AA configuration. Rather, the PMF is a free energy incorporating the statistical weight of all the AA configurations that map to a given CG configuration. Consequently, many studies have adopted the MS-CG FM variational principle because it provides two major advantages for parameterizing ML potentials. In particular, the FM variational principle employs instantaneous atomic forces that can be directly calculated from an AA configuration. Furthermore, FM determines the ML potential directly from the mapped ensemble without requiring simulations with a series of trial potentials. However, ML approaches that rely upon FM may require large databases of AA configurations in order to overcome the statistical noise in the atomic forces and, moreover, to estimate the mean forces in rarely sampled barrier regions.

John and Csányi first demonstrated the promise of kernel-based methods for developing CG models of molecular liquids, such as methanol, water, and benzene.[285] They employed Gaussian process regression to predict the many-body mean force[21] in a given CG configuration based upon its similarity with a library of CG configurations. Importantly, they numerically demonstrated that the mean forces for such molecular liquids are quite local, i.e., the mean force, $\bar{\mathbf{f}}_I$, acting on each CG site is dominated by contributions from nearby sites. Consequently, they approximated the PMF with a molecular cluster expansion, in which the m-body term was an arbitrary function of all pair distances between the m-molecules. Moreover, the locality of the mean force allowed them to model these m-body terms based upon libraries of representative monomer, dimer, and trimer configurations. The resulting molecular 2-body potential provided much higher structural fidelity than conventional pair-additive site-site potentials.

Subsequent studies have reduced the computational cost of parameterizing and simulating kernel-based potentials.[286] For instance, Scherer et al. transformed a kernel-based potential into a more efficient tabulated potential that very accurately described the 2- and 3-body correlations of liquid water.[287] Conversely, Wang et al employed ensemble learning to train

[21] Because the mean forces are gradients of the PMF, canonical simulations with these mean forces will reproduce the mapped ensemble.

a kernel-based potential for a 6-site implicit solvent CG model of alanine dipeptide.[288] They first trained an ensemble of kernel-based potentials to model fluctuating atomic forces. By averaging over this ensemble of atomic force predictors, they constructed a library of mean forces as a function of CG configuration, which was then employed to train a final predictor for the mean force. However, since each atomic predictor required training over the entire configuration space, it may be challenging to extend this approach to more complex systems. More generally, the libraries of known configurations that are employed in kernel-based methods are intriguingly reminiscent of the "memories" that have long been employed in the associative-memory, water-mediated, structure and energy model (AWSEM) CG model for proteins.[289–291]

Recent bottom-up models have also employed ANNs to approximate the many-body PMF for liquids and amorphous polymers. For instance, E and coworkers developed a very accurate DeePCG 1-site model for water[292] based upon their DeePMD AA model,[293] which was parameterized from density functional theory (DFT) simulations. The DeePCG potential approximated the PMF with a sum of one-body terms, each of which employed a detailed description of the local environment surrounding the molecule. This DeePCG model very accurately reproduced the structure of the original DFT model. Similarly, Gómez-Bombarelli and coworkers developed ANN potentials for molecular and polymeric liquids.[97,294] Interestingly, they found it necessary to develop separate networks for treating intra- and intermolecular interactions in ionic liquids.[294] Very recently, Jaakkola and coworkers employed graph neural networks to learn time-averaged forces for simulating CG polymer models with very large time steps.[295] It would be interesting to relate this approach to the Bayesian trajectory-matching method,[210–212] which also approximates time-averaged forces instead of the PMF.

Clementi, Noé, and coworkers have extensively investigated ANN potentials for implicit solvent CG models of miniproteins. Their initial study considered the 10-residue peptide chignolin and represented each amino acid with a single CG site located at the α carbon.[194]

They trained a dense ANN potential to approximate the PMF as a function of hand-selected features, including the pair distances among the amino acids, as well as bond, angle, and torsion degrees of freedom. The resulting CGNet model quite accurately described the free energy landscape for chignolin, including its folded, unfolded, and misfolded basins. Subsequently, they employed a SchNet[296] to determine optimal features for the CGNet potential.[297] The SchNet representation not only improved the accuracy and robustness of the CG potential, but may also open the path for developing transferable ANN protein potentials because SchNet features can be applied to describe different protein sequences. Interestingly, though, the accuracy of this CGSchNet potential did not significantly improve when the resolution of the CG model was increased to include both α and β carbons.[298] This group also investigated the importance of many-body interactions for CG protein models by systematically training a series of m-body ANN potentials.[299] While the 2-body ANN potential was similar to a conventional molecular mechanics potential, the higher order m-body ANN potentials were constructed as dense networks that depended upon all combinations of m pair distances. In the case of chignolin, 5-body interactions were necessary to satisfactorily reproduce the underlying FES, which suggests it may be challenging to intuit simple transferable potentials for CG protein models. Most recently, this group proposed an interesting "flow-matching" methodology for optimizing ANN potentials.[224] This approach first employs a relative entropy variational principle to parameterize an invertible normalizing flow[22] for modeling the mapped probability distribution, p_{R}, and for estimating mean forces from the gradients of p_{R}. These mean forces are then employed to parameterize the final ANN potential via the FM variational principle.

The early work of Lemke and Peter introduced a rather distinct approach for parameterizing ANN potentials based upon a discriminative classification method.[159] Lemke and Peter interpreted the PMF as an excess free energy describing the difference between the

[22]In essence, normalizing flows are a generative ML framework that determines a 1-to-1 mapping between a complex probability distribution (e.g., the mapped density, p_{R}, for $3N$ CG coordiantes) and a much simpler probability distribution (e.g., the distribution for $3N$ independent Gaussian variables).[300,301] This mapping allows one to use samples from the simple distribution in order to model the complex distribution.

mapped ensemble and a uniform distribution of CG configurations. They then trained the ANN potential to distinguish between the mapped configurations determined from AA simulations and "fake" CG configurations sampled from this uniform distribution. The resulting model quite accurately reproduced the configuration distribution of oligopeptides. This discriminative approach appears quite similar to the relative entropy formalism[28,45,82] and to several contrastive approaches that have been subsequently employed to parameterize simpler physics-based potentials.[161,162] Durumeric and Voth have related these discriminative approaches to a variational classification framework[302] that lies at the heart of generative adversarial networks.[160,303] In the context of bottom-up coarse-graining, the objective is to parameterize a CG potential, U, such that a critic can no longer distinguish whether a CG configuration, $\mathbf{R}$, has been sampled from the mapped ensemble, $p_{\mathrm{R}}(\mathbf{R})$, or instead from $P_{\mathrm{R}}(\mathbf{R};U)$, i.e., from simulations with the CG potential, U.[160,303] The variational bound is achieved when the two distributions, p_{R} and P_{R}, are no longer distinguishable, which corresponds to the configurational consistency criterion of Eq. (15).

Finally, the very recent work of Zavadlav and coworkers provided considerable insight into the application of FM and RE variational principles for parameterizing ANN potentials.[225] Specifically, they developed ANN potentials for a 1-site model of water and also an implicit solvent CG model of alanine dipeptide that treated all heavy atoms. While both variational approaches provided very accurate models for water, the relative entropy approach provided a much more accurate model for alanine dipeptide. As discussed above, Zavadlav and coworkers elegantly related the inaccuracies in the FM model to a lack of information in the mapped ensemble for determining the mean force in barrier regions that separate conformational substates. Interestingly, the authors also observed that RE optimization allowed for CG simulations with larger simulation time steps. While FM optimizes the CG potential to reproduce the actual gradients of the PMF, the RE approach effectively optimizes a shadow Hamiltonian that reproduces the mapped ensemble in CG simulations. Consequently, the resulting RE potential depends upon the time step employed in the CG

simulations that were used to optimize the potential. Future studies should further explore these considerations.

4.4 Further considerations

These studies demonstrate that ML potentials can significantly improve the structural fidelity of CG models. For instance, the DeePCG model for water[292] and the kernel-based model for benzene[285] both described the many-body structure of molecular liquids with remarkable fidelity. Similarly, the CGNets C-α model quite accurately reproduced the AA free energy surface for chignolin.[194] However, SchNet potentials for molecular and ionic liquids have appeared somewhat less accurate,[294,304] while the inclusion of β-carbons in the more detailed CGNets model for chignolin did not significantly improve its structural fidelity.[298] The reason for this variation in accuracy is currently unclear, but it may possibly reflect the training of the ML potential or the choice of certain hyperparameters.

It is striking that, with the notable exceptions of the early work by Lemke and Peter[159] and the recent study by Zavadlav and coworkers,[225] almost all of these studies employed a FM method to parameterize the ML potential. Intriguingly, the FM variational principle is closely related to a "score-matching" method that is widely adopted for optimizing ML models in high dimensional spaces.[305–308] A direct comparison of the model, $P_R(\mathbf{R};U) = Z^{-1}[U]\exp[-\beta U(\mathbf{R})]$, and target, $p_{\mathrm{R}}(\mathbf{R})$, probability densities requires knowledge of the model normalizing constant, $Z[U]$, which cannot be efficiently calculated.[23] Score-matching avoids calculating $Z[U]$ by instead parameterizing the model to match the gradients of the (logarithm of the) target probability density. Specifically, the scores for the model density are defined $\psi_{\mathrm{M}I}(\mathbf{R};U) \equiv \nabla_{\mathbf{R}_I} \log P_R(\mathbf{R};U) = \beta \mathbf{F}_I(\mathbf{R};U)$, which correspond to the approximate CG forces. The scores for the target density are defined $\psi_{\mathrm{T}I}(\mathbf{R}) \equiv \nabla_{\mathbf{R}_I} \log p_{\mathrm{R}}(\mathbf{R}) = \beta \overline{\mathbf{f}}_I(\mathbf{R})$, which correspond to the exact many-body mean forces. The target score, ψ_{T}, cannot be directly determined in many ML contexts. However, in the context of coarse-graining, ψ_{T}

[23] It is for this reason that RE minimization (and many other structure-based methods) rely upon iterative simulations in order to estimate $\delta \log Z[U]/\delta U$ when optimizing CG potentials, as discussed earlier.

can be estimated from noisy atomic forces. The score-matching loss function,[305] $J(U) \equiv \frac{1}{2}\int \mathrm{d}\mathbf{R}\, p_{\mathrm{R}}(\mathbf{R})\, ||\psi_{\mathrm{M}}(\mathbf{R};U) - \psi_{\mathrm{T}}(\mathbf{R})||^2$, corresponds to the Fisher divergence[306,309] and to the second term in Eq. (13) for the FM functional, $\chi^2[U]$, which quantifies the difference between U and W.

As already discussed, the FM variational principle is not guaranteed to reproduce any particular structural correlation functions. The accuracy of the SchNet model for ionic liquids[294] and Cα + Cβ model for chignolin are both reminiscent of the errors that sometimes arise in MS-CG models when the interaction potential is not sufficiently flexible to reproduce the relevant many-body correlations present in the mapped ensemble.[199,200] As indicated in the work of Potestio and coworkers,[85] the Cα representation likely corresponds to a relatively simple mapped ensemble. The inclusion of Cβ coordinates may significantly increase the complexity of the mapped ensemble. As discussed above, several studies indicate that this increased complexity can reduce the structural fidelity of MS-CG potentials.[200,204,234–236] Because they likely describe many-body environments more effectively than pair distances, it may be useful to include local densities, which are discussed in the next section, as features in ANN potentials. Indeed, an early precursor of the AWSEM protein model employed local density potentials to describe many-body solvation effects.[310]

While the structural fidelity of physics-based potentials is usually most limited by the flexibility of the assumed functional forms (e.g., additive central pair potentials), ML potentials are more likely limited by the available data in the mapped ensemble.[225,285] As discussed earlier, because FM variational principles determine potentials to match gradients of the PMF, they rely upon information in rarely sampled barrier regions to determine a CG potential that provides appropriate weight to different basins. Consequently, ML potentials that are parameterized via FM variational principles may well be limited by the lack of sampling in these barrier regions. Conversely, the RE variational principle or other discriminative approaches that are more data efficient may prove useful for parameterizing ML potentials to reproduce global aspects of the mapped ensemble.[159,161,162,225] From this

perspective, the flow-matching approach appears particularly promising for capitalizing upon both the data efficiency of the RE variational principle and also the computational efficiency of the FM variational principle.[224] More generally, because the accuracy of ML approaches likely varies with the CG mapping, it may be beneficial to use ML methods to simultaneously optimize the CG mapping, potential input features, and potential parameters to ensure self-consistency between the mapped ensemble and the approximate potential. The recent work of Rotskoff and coworkers appears a promising step in this direction.[230]

5 Transferability and Thermodynamics

Once a CG model has been parameterized to reproduce structural properties at a single thermodynamic state point, one hopes that the model will be "transferable," i.e., that it will provide similar accuracy for modeling a wide range of state points. Moreover, one hopes that the CG model will also accurately describe thermodynamic properties. Consequently, many recent studies have investigated the transferability and thermodynamic properties of bottom-up models.

The most straight-forward approach is to treat the approximate interaction potential, U, as analogous to a conventional AA potential energy that does not explicitly depend upon thermodynamic conditions. One can then compute thermodynamic properties, such as the energy or pressure, using conventional textbook expressions.[311] In some cases, this approach works quite well.[211,312,313] For instance, Guo and coworkers determined IBI potentials with pressure corrections for modeling poly-imides at 800K and 1 bar pressure.[312] Quite remarkably, these potentials accurately modeled the polymer pair structure, thermal expansion, and bulk modulus down to 300K, although the CG model over-estimated the compressibility by a factor of 5-10.

Unfortunately, though, conventional bottom-up models often provide unpredictable transferability and a rather poor description of thermodynamic properties. For instance, bottom-

up potentials often systematically vary with temperature,[314–320] which can result in poor structural fidelity away from the reference state point[259,321] and especially across phase boundaries.[33,322,323] Bottom-up potentials also vary significantly with density[179,318,323,324] and tend to dramatically overestimate the internal pressure.[217,325] While linear pressure corrections allow structure-based pair potentials to accurately reproduce the internal pressure at a single state point,[142] they often lead to a poor description of the compressibility.[147,233,312] Similarly, bottom-up models often fail to reproduce the coefficient of thermal expansion.[34,326] More basically, bottom-up models generally tend to underestimate the cohesive energy[204,327,328] and configurational entropy of AA models.[88,319] This poor description of thermodynamic properties models stems from "representability issues" due to effective potentials that vary with thermodynamic conditions, as first discussed by Louis and coworkers.[179,329] Accordingly, some studies appear to suggest that these representability issues fundamentally preclude bottom-up methods from reproducing both structural and thermodynamic properties.

We optimistically hypothesize that, by combining rigorous theory with robust computational methods, bottom-up approaches can and will model both structural and thermodynamic properties with predictive accuracy across a range of thermodynamic state points. We anticipate that the many-body PMF, W, holds the key for addressing the transferability and representability limitations of bottom-up models not only in theory, but also in practice. According to Eq. (8), $W(\mathbf{R})$ is the excess Helmholtz potential associated with the subensemble of AA configurations that map to $\mathbf{R}$. Consequently, W depends upon the temperature, volume, composition, and any other relevant thermodynamic variables. If it is known as a function of both configuration and thermodynamic state point, then $W(\mathbf{R}, V, T)$ contains the necessary information for reproducing all structural and thermodynamic properties of the AA model that can be observed at the CG resolution. This simple observation already establishes the fundamental link between representability and transferability challenges. Specifically, if one knows how W varies with thermodynamic state point then one

can resolve representability problems by accounting for this state-point dependence when computing thermodynamic properties. Moreover, from knowledge of this state-point dependence, one can resolve transferability problems by varying U to accurately describe W and, thus, the mapped ensemble at each state point. Consequently, the key to overcoming these challenges lies in developing robust computational methods for accurately calculating and rigorously modeling the state-point dependence of W. It is important to note that this approach fundamentally differs from the traditional approach of varying interaction potentials in order to accurately model thermodynamic properties with conventional textbook approaches.

In order to make this more concrete, let us first consider the total differential of an atomic potential that does not explicitly depend upon the thermodynamic state point:

$$\mathrm{d}u = -\sum_{i=1}^{n} \mathbf{f}_i \cdot (\mathrm{d}\mathbf{r}_i)_V - p_{\mathrm{xs}}\,\mathrm{d}V, \tag{16}$$

where $\mathbf{f}_i$ is the force on atom i and $(\mathrm{d}\mathbf{r}_i)_V$ describes variations in configuration at constant volume.[24] In the isotropic case that the volume changes while the scaled coordinates, $\hat{\mathbf{r}}_i = \mathbf{r}_i/V^{1/3}$, remain fixed, the instantaneous excess pressure of the AA model is given by the standard virial expression

$$p_{\mathrm{xs}} \equiv -\left(\frac{\partial u}{\partial V}\right)_{\hat{\mathbf{r}}} = \frac{1}{3V}\sum_{i=1}^{n} \mathbf{f}_i \cdot \mathbf{r}_i, \tag{17}$$

which describes the force that the system applies to its walls.[25] The two terms in Eq. (16) correspond to mechanical work due to changing the configuration at constant volume and due to isotropically compressing or expanding the system, respectively. Because we assume that the atomic potential is state-point independent this work is independent of T.

[24] If we define scaled coordinates as $\hat{\mathbf{r}}_i = \mathbf{r}_i/V^{1/3}$, then $(\mathrm{d}\mathbf{r}_i)_V = V^{1/3}\mathrm{d}\hat{\mathbf{r}}_i$. Thus, $\mathrm{d}u = -V^{1/3}\sum_{i=1}^{n} \mathbf{f}_i \cdot \mathrm{d}\hat{\mathbf{r}}_i - p_{\mathrm{xs}}\,\mathrm{d}V$.

[25] If u explicitly depends upon V, then $p_{\mathrm{xs}} = \frac{1}{3V}\sum_{i=1}^{n} \mathbf{f}_i \cdot \mathbf{r}_i - (\partial u/\partial V)_{\mathbf{r}}$. Such explicit volume dependence arises in AA simulations, e.g., when accounting for the long-ranged contributions from dispersion or electrostatic interactions.

In contrast, W is the excess Helmholtz potential of the AA model when it is viewed at the resolution of the CG model, i.e., the equilibrium state is specified by $(\mathbf{R}, V, T)$. Consequently, it follows that

$$\mathrm{d}W = -\sum_{I=1}^{N} \overline{\mathbf{f}}_I \cdot (\mathrm{d}\mathbf{R}_I)_V - \overline{p}_{\mathrm{xs}} \mathrm{d}V - S_W \mathrm{d}T. \tag{18}$$

The first two contributions to Eq. (18) are analogous to Eq. (16). However, now $\overline{\mathbf{f}}_I(\mathbf{R}, V, T) \equiv \langle \mathbf{f}_I \rangle_{\mathbf{R}VT}$ and $\overline{p}_{\mathrm{xs}}(\mathbf{R}, V, T) \equiv \langle p_{\mathrm{xs}} \rangle_{\mathbf{R}VT}$ are temperature-dependent averages of the instantaneous force and excess pressure, respectively, evaluated over the conditioned distribution, $p_{\mathrm{r|R}}(\mathbf{r}|\mathbf{R})$, of AA configurations that map to $\mathbf{R}$. Consequently, the first two terms in Eq. (18) correspond to temperature-dependent free energy changes (i.e., the minimum, reversible work) associated with changing the CG configuration and volume. Moreover, $\overline{p}_{\mathrm{xs}}$ generally includes both a virial contribution from the mean forces acting on the volume and also a contribution from the explicit volume-dependence of the PMF:

$$\overline{p}_{\mathrm{xs}}(\mathbf{R}, V, T) = -\left(\frac{\partial W}{\partial V}\right)_{\hat{\mathbf{R}}} = \frac{1}{3V} \sum_{I=1}^{N} \overline{\mathbf{f}}_I \cdot \mathbf{R}_I - \left(\frac{\partial W}{\partial V}\right)_{\mathbf{R}}. \tag{19}$$

Similarly, W now varies with T according to

$$S_W(\mathbf{R}, V, T) \equiv -k_B \mathrm{D}[p_{\mathrm{r|R}} || q_{\mathrm{r|R}}](\mathbf{R}, V, T) = -k_B \int \mathrm{d}\mathbf{r}\; p_{\mathrm{r|R}}(\mathbf{r}|\mathbf{R}) \ln \left[\frac{p_{\mathrm{r|R}}(\mathbf{r}|\mathbf{R})}{q_{\mathrm{r|R}}(\mathbf{r}|\mathbf{R})}\right], \tag{20}$$

which quantifies the excess configurational entropy stored in the subensemble of AA configurations that map to $\mathbf{R}$. Note that the average of $-S_W/k_B$ over the mapped ensemble corresponds to the mapping entropy, H_{map}. Thus, $-k_B H_{\mathrm{map}}$ is the difference between the excess configurational entropy of the AA model and the excess configurational entropy present in the mapped ensemble.

Although it is intuitively obvious and perhaps trivial, Eq. (18) explicitly establishes the fundamental origin and intrinsic duality of representability and transferability issues. As illustrated in Fig. 4, both stem from the transfer of thermodynamic information from the

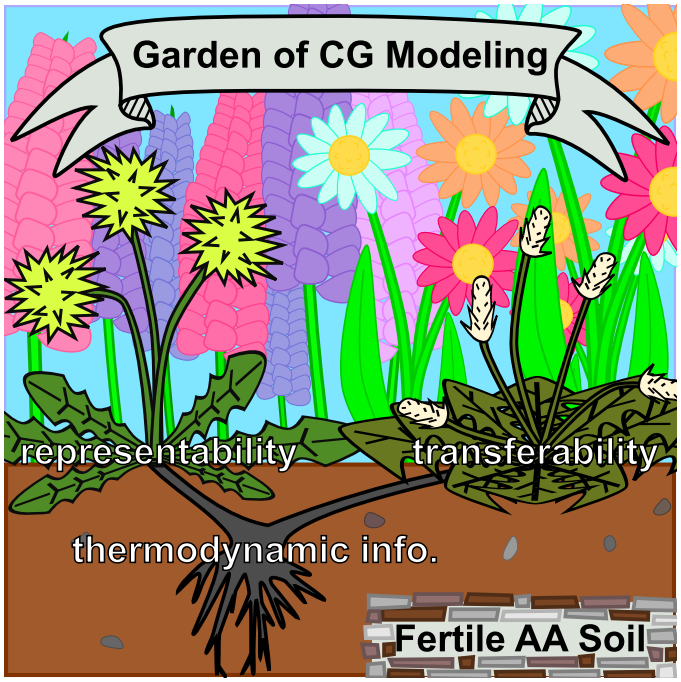


Figure 4: The representability and transferability challenges are two weeds that stem from the same root. Both challenges result from thermodynamic information that is present in the AA configuration distribution but lost from the mapped ensemble. This lost information determines the state-point dependence of the PMF, which must be accounted for when modeling thermodynamic properties or transferring potentials between thermodynamic state points. Reprinted with permission from Dunn, N.J.H.; Foley, T.T.; Noid, W.G. Ref. 30. van der Waals perspective on coarse-graining: Progress toward solving representability and transferability problems. *Acc. Chem. Res.* **2016**, *49*, 2832-2840. Copyright 2016 American Chemical Society.

AA configuration space into the state-point dependence of the PMF.[30] This state-point dependence not only determines how the CG interaction potentials should vary with V and T, but also introduces new contributions to the corresponding conjugate thermodynamic properties, i.e., the pressure and entropy, respectively.

Conventional structure-based approaches focus on reproducing the configuration-dependence of the PMF at a single state point and do not critically consider this volume- or temperature-dependence. Consequently, the resulting potentials cannot be expected to be transferable to other state points or to accurately describe the conjugate thermodynamic properties. Fortunately, recent studies have leveraged the insights from Eq. (18) to develop practical bottom-up approaches for addressing the representability and transferability issues.

5.1 Density-dependence and internal pressure

For instance, consider the density-dependence of the PMF, which at constant composition is equivalent to its volume-dependence. The second term in Eq. (18) implies that, if the approximate potential reproduces this density-dependence,[26] then the CG model will also reproduce the excess pressure of the AA model at the resolution of the CG model.[27] Conversely, the failure of structure-based models to reproduce the AA excess pressure implies that they do not accurately treat the density-dependence of the PMF.[179,318,323] Thus, the representability issues associated with modeling the AA pressure fundamentally stem from transferability issues in modeling the density-dependence of CG interaction potentials.[30,36,329,330]

Perhaps the simplest way to model the density-dependence of the PMF is to introduce a configuration-independent, volume potential, $U_V(V)$.[216] This volume potential directly contributes to the pressure without perturbing the equilibrium configuration distribution at a given volume. By employing FM and RE variational principles, one can variationally optimize U_V to reproduce the average volume-dependence of the PMF, and, consequently, quantitively reproduce the AA pressure-density equation of state.[216–218] Alternatively, one can adopt an "active" approach, in which bottom-up effective pair potentials explicitly vary with volume and introduce new contributions to the pressure.[166,331,332] It may be possible to employ pressure-matching variational principles to predict this volume dependence.

Local-density (LD) potentials have recently emerged as a particularly promising avenue for modeling the density-dependence of the PMF. LD potentials were first introduced to describe non-ideal solutions in top-down, many-body dissipative particle dynamics (DPD) models.[333] By defining the local density, ρ_I, around each site, I, with pair-additive weighting

[26] According to Eq. (19), the density-dependence of the PMF includes both an implicit virial contribution from the mean forces acting on the volume, $\frac{1}{3V}\sum_{I=1}^{N}\overline{\mathbf{f}}_I \cdot \mathbf{R}_I$, and also a contribution due to the explicit density-dependence of the PMF, $-\left(\partial W/\partial V\right)_{\mathbf{R}}$. Efforts to calculate the pressure with CG models almost always adopt a standard virial expression that neglects this second explicit contribution.

[27] The ideal contribution to the thermodynamic pressure scales trivially with the number of missing degrees of freedom. Correcting for this ideal contribution slightly increases the internal pressure of CG models. However, this ideal contribution is much smaller than the virial pressure predicted by structure-based potentials.[30]

functions, LD potentials generate pair-additive forces and retain much of the computational efficiency of conventional pair-additive potentials. Recent bottom-up studies have parameterized LD potentials for implicit solvent models of hydrophobic self-assembly,[334–336] as well as for polymer melts,[337,338] explosive materials,[339] liquid mixtures,[340] liquid interfaces,[283,341–343] and liquid-liquid phase equilibria.[344]

While they can enhance structural fidelity,[336] LD potentials have an even greater impact upon the thermodynamic properties and transferability of bottom-up models. When the local density is defined over sufficiently long distances, LD potentials function similarly to volume potentials and allow bottom-up models to quantitatively reproduce AA pressure-density equations of state, but tend to introduce artifacts at interfaces.[341,342] Conversely, when the local density is defined over shorter distances, LD potentials can accurately describe liquid-solid and liquid-vapor interfaces, while still providing a nearly quantitative description of the pair structure and pressure-density equation of state for bulk liquids.[283,341–343] This outstanding transferability between bulk and interfacial regions has proven challenging for conventional pair-additive bottom-up potentials.[345–347] Moreover, LD potentials can be related to internal energies that are employed in energy-conserving DPD methods.[348–350] This suggests it may be possible to develop bottom-up CG models for accurately simulating shock-waves, chemical reactions, and other non-equilibrium phenomena.[339,351,352] Finally, a recent bottom-up model[353] introduced a potential that depends upon the square of the gradient in the local density, i.e., $|\nabla\rho_I|^2$, which may prove useful for modeling highly inhomogeneous systems and for connecting with classical density functional theories.[354]

5.2 Temperature-dependence, energy, and entropy

Similarly, the last term in Eq. (18) equates the temperature-dependence of the PMF, $(\partial W/\partial T)_{\mathbf{R},V}$, with $-S_W$, which quantifies the excess configurational entropy of the AA subensemble that maps to $\mathbf{R}$. In order for a CG model to properly describe AA entropies, it must not only account for the configurational entropy present in the mapped ensemble, but also account

for this entropic contribution from the AA degrees of freedom that have been eliminated from the mapped ensemble. The observed temperature-dependence of bottom-up potentials, U, that approximate W allows one to approximate $S_W \equiv -(\partial W/\partial T)_{\mathbf{R},V} \approx -(\partial U/\partial T)_{\mathbf{R},V}$. By employing this approximation, Voth and coworkers modeled entropic properties of liquid methanol and chloroform that reflect AA details absent from the CG representation.[319]

Equation (20) provides useful insight into S_W. For high resolution CG representations that associate sites with small, rigid atomic groups, one expects that $p_{\mathrm{r|R}}$ will be quite sharply peaked and will vary relatively little with configuration or thermodynamic conditions. In this case, one anticipates S_W will be rather small[28] and relatively constant across single-phase regions of the phase diagram. Indeed, while nonlinear models have been proposed,[314,315,355–357] recent studies indicate that bottom-up potentials for high resolution models often vary linearly across a rather wide temperature range.[316–320,358] For instance, by assuming that IMC pair potentials were temperature-independent and that the volume potential varied linearly with temperature, Rosenberger and van der Vegt parameterized a CG model for hexane that accurately modeled its structure, thermal expansion, and compressibility over a temperature range of 140 K and well into the super-cooled liquid regime.[326] Conversely, for lower resolution models in which CG sites correspond to more flexible atomic groups, one expects that $p_{\mathrm{r|R}}$ will be broader, more complex, and possibly include multiple peaks corresponding to distinct internal conformations. In this case, one expects that S_W will be larger and will demonstrate more complex dependence upon both configuration and temperature. Quite generally, one expects that CG potentials will vary nonlinearly with temperature across phase boundaries and whenever $p_{\mathrm{r|R}}$ significantly varies.

Because W is a configuration-dependent excess Helmholtz potential, it follows that

$$W(\mathbf{R},V,T) = E_W(\mathbf{R},V,T) - TS_W(\mathbf{R},V,T), \tag{21}$$

[28] According to Eq. (20), $S_W(\mathbf{R})$ increases as $p_{\mathrm{r|R}}(\mathbf{r}|\mathbf{R})$ becomes less uniform. Nevertheless, because H_{map} is the average of $-S_W/k_B$ and because H_{map} systematically increases with coarsening, one expects that $|S_W(\mathbf{R})|$ will be relatively small (on average) for high resolution mappings.

where $E_W(\mathbf{R},V,T) = \langle u(\mathbf{r})\rangle_{\mathbf{R}VT}$ is the conditioned average of the AA potential energy[29] for the subensemble of AA configurations that map to $\mathbf{R}$.[21,36,84] Since $-TS_W \geq 0$, structure-based potentials, U, that accurately describe the configuration-dependence of W cannot be naïvely employed for estimating atomic energetics. In particular, note that $-TS_W$ systematically increases as details are eliminated from the CG model. In contrast, because E_W is a conditioned average, it does not systematically increase with coarsening. Consequently, the PMF becomes increasingly entropic with coarsening, as illustrated in Fig. 5.[84]

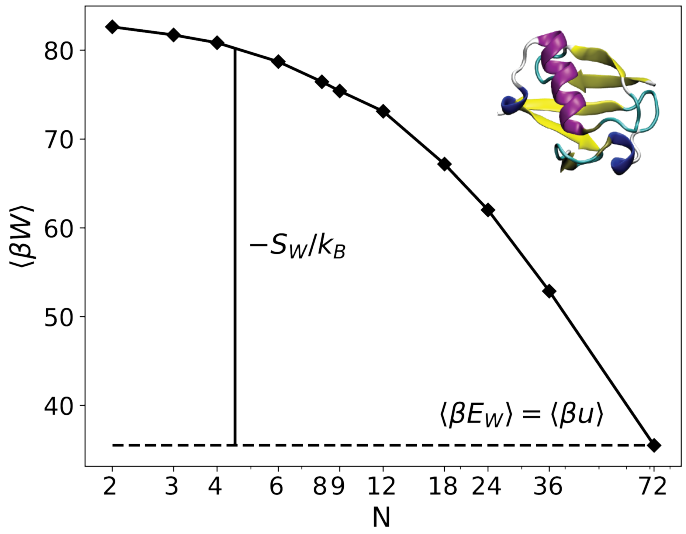


Figure 5: The impact of resolution upon the average energetic and entropic contributions to the many-body PMF, i.e., $\langle\beta W\rangle = \langle\beta E_W\rangle - \langle S_W\rangle/k_B$. The symbols plot the average of the dimensionless PMF, $\langle\beta W\rangle$, as a function of the number of CG sites, N, for CG models of ubiquitin. In order to perform exact calculations of the PMF,[84] the AA model for ubiquitin is a GNM that represents each of its 72 amino acids with the corresponding α carbon. When $N = 72$, the PMF corresponds to the AA potential, i.e., $W = u$, such that $E_W = u$ and $S_W/k_B = 0$. The dashed horizontal line indicates that $\langle E_W\rangle = \langle u\rangle$ for all resolutions. The PMF systematically increases with decreasing resolution because $-\langle S_W\rangle/k_B \geq 0$ increases as details are eliminated from the mapped ensemble. Adapted with permission from Ref. 47. Copyright 2021 Springer Nature.

Since E_W is a conditioned average of u, Lebold and Noid repurposed a least squares energy-matching variational principle[360] to determine an energetic operator, E, that provides an optimal approximation to E_W.[47,358,359,361] In this "dual" approach, CG models are simulated with conventional structure-based potentials in order to accurately sample config-

[29]Prior studies referred to E_W as U_W.[30,84,359]

urational space. Atomic energetics are then estimated (at the resolution of the CG model) by evaluating E for the sampled configurations. This approach effectively implements the early suggestion of Louis and coworkers in addressing the most basic of representability issues, i.e., that separate pair functions were necessary for reproducing the rdf and for modeling atomic energetics.[179,329] Furthermore, given a structure-based potential, U, and an energetic operator, E, that accurately approximate the configuration-dependence of W and E_W, respectively, one can then approximate $S_W = (E_W - W)/T$ with $S \equiv (E - U)/T$. This then provides a predictive estimate for the temperature dependence of the bottom-up potentials, i.e., $(\partial U/\partial T)_{\mathbf{R},V} \approx (\partial W/\partial T)_{\mathbf{R},V} = -S_W \approx -S$. Initial studies for various molecular liquids, including water, methanol, and ortho-terphenyl, suggest that this dual approach can not only reasonably model atomic energetics, but also accurately predict the temperature-dependence of bottom-up potentials.[358,359] Moreover, Lebold and Noid introduced a configuration-dependent specific-heat, C_W, to model the temperature-dependence of E_W and S_W, as well as to estimate the variance in the atomic energy fluctuations for each CG configuration.[361] This framework provides an internally consistent framework for bottom-up models to reproduce the atomic specific heat.

As mentioned above, Pretti and Shell recently proposed an elegant complementary approach based upon modeling the joint distribution, $p_{\mathrm{RE}}(\mathbf{R}, E)$, of CG configurations and energies that are sampled by the AA model.[187] Rather than approximating W and its derivatives, they employed a RE variational principle to determine an optimal approximation for

$$\Omega(\mathbf{R}, E) \propto \int_{V^N} \mathrm{d}\mathbf{R}\, \delta(\mathbf{M}(\mathbf{r}) - \mathbf{R})\delta(u(\mathbf{r}) - E), \tag{22}$$

which corresponds to a microcanonical partition function for the AA model.[350] If $\Omega(\mathbf{R}, E)$ is known for all E, then one can determine E_W, S_W, and W for any temperature. Conversely, if $W(\mathbf{R}, T)$ is known for all T, then at least in principle one should be able to infer $\Omega(\mathbf{R}, E)$ via inverse Laplace transform, although this may prove more challenging in practice.

5.3 Composition-dependence

Unfortunately, Eq. (18) provides no insight into how the PMF varies with chemical composition. Moreover, relatively few studies have investigated the transferability and thermodynamic properties of CG potentials as a function of composition. An interesting study by Deichmann and van der Vegt investigated the accuracy of various bottom-up approaches for modeling hexane-perfluorohexane mixtures.[362] Because they minimize environmental contributions, the CRW and EF-CG potentials appeared independent of composition and, moreover, quite reasonably described both structural and thermodynamic properties of the mixtures. In contrast, Potter and Wilson demonstrated that IBI and MS-CG potentials for liquid mixtures of octane and benzene vary significantly as a function of composition and, consequently, provide rather poor transferability.[34]

One approach for improving the transferability of bottom-up potentials is to determine a single set of potential functions, i.e., a force field, that provides an optimal compromise for modeling a range of systems and thermodynamic state points. In particular, the multistate IBI (ms-IBI) method determines transferable pair potentials that optimally match a set of rdfs obtained from various simulation conditions.[363] Indeed, Potter and Wilson demonstrated that ms-IBI pair-potentials provided high structural fidelity across the entire range of octane-benzene mixtures.[34] McCabe and coworkers have recently employed ms-IBI to develop a transferable force field for stratum corneum membranes comprised of sphingolipids and cholesterol.[364–366] Similarly, Hall and coworkers employed ms-IBI to parameterize a transferable force field for phospholipids based upon simulated rdfs for multiple temperatures.[367]

Alternatively, one can employ an "extended ensemble" variational principle to determine a set of transferable potentials that optimally describes the PMF for multiple systems across multiple state points.[234] For instance, Rudzinski et al. employed the extended ensemble approach to develop transferable potentials for modeling poly(ethyleneoxide)-based ionomers with varying degrees of sulfonation at a range of temperatures.[321] Sanyal and Shell em-

ployed the extended ensemble framework to derive a transferable backbone potential that, when combined with a Gō-based potential for stabilizing native contacts,[368] accurately folded a rather remarkable diversity of complex protein structures.[369] Kanekal and coworkers employed the extended ensemble framework in an automated high-throughput study that employed simulation data from hundreds of simulations to parameterize transferable potentials for modeling the chemical space defined by 3,000 distinct isomers of C_7O_2.[370] They observed that, in this case, the extended ensemble approach not only improved the transferability but also the structural accuracy of the effective potentials. They concluded that averaging over systems acted to "regularize" the model by smoothing over atomic features that could not be accurately described by the approximate CG potential. In contrast, Shen et al. developed a similar extended ensemble framework for the RE variational principle and demonstrated that including statistics from multiple state points did not improve the transferability of soft sphere models for methanol-water mixtures.[371] Rather, the authors concluded that it was more important to ensure that the parameterization ensemble contained information about correlated composition fluctuations. As discussed earlier, they hypothesized that the Fisher information metric could be used to identify maximally informative ensembles for improving the transferability of bottom-up models. Finally, Malfreyt and coworkers recently developed an analogous extension of the statistical trajectory matching method in order to parameterize transferable potentials for CG models of copolymer systems.[372]

One expects that LD potentials may also significantly improve the transferability of bottom-up models across mixture compositions.[335,340,344] In particular, bottom-up models with LD potentials quite accurately describe the liquid-liquid phase separation of benzene and water mixtures in two-phase regions of the phase diagram.[344] Interestingly, a seemingly diverse set of potentials performed equally well. However, it is quite likely that these potentials were actually equivalent (or at least nearly so) because there exist families of distinct pair and LD potentials[30] that generate equivalent forces and distributions.[342] Future studies

[30]More explicitly, consider a pair potential, $U_2(r)$, and a LD potential, $U_\rho(\rho_I)$. If one adds a linear term to the pair potential, $U_2(r) \to U_2(r) + kr$, then one can add a corresponding correction to the LD potential

should further explore the use of LD potentials for modeling complex mixtures.

5.4 Further considerations

One expects that it should be possible to determine observable operators, $A(\mathbf{R})$, for describing any AA property, $a(\mathbf{r})$, at the resolution of the CG model as a conditioned average over the AA configurations that map to the CG configuration, i.e., $A(\mathbf{R}) = \langle a(\mathbf{r}) \rangle_{\mathbf{R}}$.[29] Recent studies have developed operators for estimating various thermodynamic observables, including the energy, pressure, surface tension, specific heat, and chemical potential, at the resolution of the CG model.[29,214,328,359,361,373] However, the "predictive" framework of Koutsourelakis and coworkers indicates the even more exciting possibility of probabilistically modeling atomic properties that would seem beyond the resolution of the CG model.[186] Pretti and Shell achieved a significant step in this direction by explicitly modeling the distribution of AA energies for a given CG configuration, $\mathbf{R}$.[187] The electronic coarse-graining method takes this approach one step further by employing ML approaches to model the distribution of electronic energies that map to $\mathbf{R}$.[374,375] As demonstrated by recent back-mapping approaches that modeled the conditioned distribution, $p_{\mathrm{r|R}}$,[132,135] ML approaches may prove particularly useful for modeling conditioned distributions, $p_{\mathrm{a|R}}(a|\mathbf{R})$, describing arbitrary AA observables, $a(\mathbf{r})$. In particular, Rotskoff and coworkers recently proposed a "weak" consistency criterion that may prove generally useful for parameterizing such predictive CG models.[230]

There is an important distinction between "observable operators" and "effective potentials." If operators are only used to estimate properties by post-processing simulated configurations, then they can not ensure that simulations sample the correct fluctuations in the conjugate thermodynamic variables. For instance, in order to properly sample volume fluctuations in the constant NPT ensemble, the simulated effective potential and, in particular, the barostat equation of motion must accurately account for the volume dependence

such that the net force on each site is unchanged.

of the exact PMF.[30,216] Simply applying an operator to estimate the internal pressure from sampled configurations will not ensure that the CG simulation samples the correct volume distribution. When considering equilibrium statistics, it may be possible to reweight sampled configurations according to observable operators. However, this requires that the operator gives a relatively small correction to the observable, such that there is good overlap between the simulated and reweighted distributions. Moreover, statistical reweighting cannot be applied when considering dynamical or nonequilibrium processes. For instance, while energetic operators can provide insight in post-processing,[359,361] CG models must explicitly model the internal energy of CG sites in order to properly describe energy transfer in shock waves or heat conduction.[376,377]

Recent insights may also assist in improving the transferability of ML potentials. Quite generally, one expects that the transferability of any potential critically relies upon identifying the proper collective variables and incorporating the relevant physical principles. For instance, the transferability of atomic ML potential relies upon incorporating relevant symmetries and equivariances into the potential.[378] Similarly, the state-point dependence of the PMF is not arbitrary, but is completely determined by certain observables within the mapped ensemble. Consequently, one expects that the transferability of ML potentials will be dramatically improved by accounting for these observables. For instance, one expects that ML approaches should be able to predict accurate temperature-dependent effective potentials by properly distinguishing and treating the energetic and entropic contributions to the PMF.[187,359,361] By so doing, these ML approaches should be able to also accurately model energetic and entropic properties of the AA model. Moreover, one expects that the transferability of ANN potentials for modeling multiple proteins will rely upon incorporating features that govern transferable physical principles, e.g., the hydrophobic forces driving protein collapse.[379] These hydrophobic forces are not only many-body and long-ranged, but are also probably not well described by pair distances between CG sites.[380] Conversely, by incorporating local densities or other features that directly measure the solvation of amino

acids, it may be possible to develop simpler and more transferable ML potentials for proteins.

6 Conclusion: Retro- and Pro-spectives

6.1 Key advances and insights

Almost ten years ago a perspective review discussed several outstanding challenges and promising directions for bottom-up approaches.[23] It is perhaps useful to revisit this list in order to assess recent progress, identify remaining challenges, and propose promising directions. The prior perspective considered the following areas:

6.1.1 Basis set

The "basis set" refers to the various contributions, U_ζ, in Eq. (9) that define the form and flexibility of the approximate potential, U. Due both to their familiarity from AA force fields and also to their computational efficiency, early CG models relied almost exclusively upon pair-additive site-site potentials for describing intermolecular interactions. The prior perspective[23] noted that

> Bottom-up approaches may greatly benefit from considering more complex potential terms to model, e.g., hydrogen-bonding, electrostatic interactions, solvation forces, and anisotropic interactions. More generally, it would be highly desirable to develop an algorithm or framework for systematically identifying "missing" basis vectors that are computationally efficient to parameterize and simulate and that would also provide significantly improved descriptions of biomolecular structure, dynamics, and thermodynamics.

Recent bottom-up studies have demonstrated remarkable progress in developing more advanced physics-based potentials. These new potentials have significantly improved the structural fidelity, transferability, and thermodynamic properties of bottom-up models. For

instance, a growing number of bottom-up models have adopted orientation-dependent potentials to more accurately describe interactions between anisotropic sites.[213,241–250] Ultra CG and surface hopping approaches provide promising frameworks for modeling hydrogen bonds and for modulating pair interactions to reflect their local environment.[192,208,209,235,266–269] Similarly, bottom-up LD potentials appear promising for modeling many-body solvation forces[334–336,344] and liquid interfaces,[283,343,353] as well as for describing thermodynamic properties.[283,337,338,340–342] Moreover, ML approaches provide entirely new classes of basis functions for more accurately approximating the PMF. In particular, bottom-up models with kernel-based and ANN potentials appear capable of describing many-body structural correlations with unprecedented fidelity.[194,224,285,292]

Future studies should both continue developing these promising directions and also continue inventing new classes of potentials. For instance, bottom-up approaches should develop physically motivated many-body potentials for describing the hierarchical structures of complex biomolecules,[381] as well as more sophisticated approaches for modeling long-ranged electrostatic interactions in highly charged systems.[382,383] Moreover, by combining ultra-CG[192,208,209] or surface-hopping[268,269] potentials along with dynamical CG representations that vary as a function of local environment,[105,230] it may be possible to develop computationally efficient, low resolution models with unprecedented accuracy and transferability.

ML tools provide a promising framework for identifying "missing" basis vectors that more accurately describe many-body correlations and hierarchical structures.[282] Conversely, one hopes that physical insights will inform the architecture and features of ML potentials with improved accuracy and transferability, while also reducing their computational cost and simplifying their training. More generally, one expects that outstanding physical intuition may provide greater computational efficiency and more penetrating insight than the naïve application of ML tools. For instance, the prescient AWSEM model pioneered long ago both the use of knowledge-based "memory" terms, which are strikingly similar to the libraries employed in kernel methods, and also LD potentials for describing water-mediated

and hydrophobic interactions.[289–291] Consequently, we anticipate that physical insight may play a key role in developing computationally efficient bottom-up approaches for accurately modeling long-range electrostatic interactions, many-body solvation forces, and hierarchical structures.

6.1.2 Mapping

The prior perspective[23] noted that

> The mapping significantly impacts structural, dynamic, and thermodynamic properties of a CG model. Unfortunately, ... relatively little theoretical work has addressed its importance. However, it seems intuitively reasonable that the ability of a CG model to describe the correct physics governing a particular system fundamentally relies upon the CG model capturing the key physical features underlying this physics. It also seems intuitively reasonable that "better" mappings will allow for a "simpler" description of this physics.

The prior perspective suggested that "practical, rigorous algorithms that apply this intuition" for determining the CG representation should significantly improve the dynamical, structural, and thermodynamic properties of bottom-up models.[23]

Recent studies have made considerable progress in developing algorithms for optimizing CG representations to capture the "correct physics." For instance, while the ED-CG method remains an important framework for identifying CG representations that preserve the collective motions present in AA simulations,[62] graph-based methods appear capable of applying this intuition without requiring explicit AA simulations or an underlying network model.[77] Graph-based approaches also appear promising for organizing and optimizing CG representations.[76] Similarly, variational autoencoders[97] appear a promising framework for simultaneously optimizing both the CG representation and also a deterministic backmapping operator. More generally, graph-based ML approaches provide a new framework for representing systems in CG detail.[230,295,297]

Perhaps even more importantly, recent studies have provided fundamental insight into the impact of the mapping upon the properties of bottom-up models. For instance, by employing Monte Carlo methods to systematically sample and statistically characterize the space of CG representations, recent studies have begun to provide general insight into the impact of the CG representation upon the mapped ensemble, as well as the differences between "good" and "bad" representations.[85,96] Recent studies also suggest that simplifying the mapped ensemble, p_{R}, will generally increase the structural fidelity of bottom-up models.[204,235,236] In this context, it is important to distinguish the "intrinsic" quality from the "practical" quality of a CG representation. While the intrinsic quality of a CG representation is completely determined by the AA model and CG mapping, its practical quality reflects the various approximations that are made in modeling interactions and in computing observable properties. In practice, one expects that the "optimal" map will depend upon the complexity of the CG potential. It will likely be beneficial to continue exploring both the intrinsic and practical quality of the CG mapping.

Recent studies have also begun to illuminate the impact of the mapping upon the conditioned distribution, $p_{\mathrm{r}|\mathrm{R}}$, that describes the "lost" subensemble of AA configurations that map to a given CG configuration. The entropy chain rule of Eq. (7) quantifies a fundamental tradeoff in how the mapping, $\mathbf{M}$, partitions the complexity of the underlying AA model between the mapped ensemble and this lost subensemble.[79,84] In particular, mappings that simplify p_{R} generally increase the information content of $p_{\mathrm{r}|\mathrm{R}}$, which may render back-mapping approaches both more feasible and more meaningful.[97] This entropy chain rule also clarifies the influence of $\mathbf{M}$ upon the thermodynamic properties and transferability of bottom-up models. One expects that representations with relatively simple mapped ensembles will correspond to softer effective potentials with relatively large entropic contributions.[84] These large entropic contributions must be properly considered when computing thermodynamic properties and when transferring potentials between different temperatures.

While the entropy chain rule should hold quite generally, it is important to emphasize

that much of our practical intuition relies upon studies of relatively simple models. Most studies of CG mappings consider either liquids of relatively rigid small molecules or simplified protein models that fluctuate about a well-defined equilibrium conformation. Future studies should further investigate the impact of the mapping for modeling more complex molecules, such as polymers and biomolecules that sample more complex hierarchical structures and transition between multiple metastable conformations.[85,105] Moreover, it may be beneficial to investigate the physical content of more abstract graph-based or dynamical CG representations.[230,295,297]

6.1.3 Model optimization/assessment

The prior perspective[23] noted that

> the many-body PMF is often approximated with simple potentials that are parameterized to reproduce the distributions observed in an atomistic model along the corresponding degrees of freedom. For instance, non-bonded pair potentials are often optimized to reproduce the corresponding pair distribution functions. ... Depending upon the CG mapping, an accurate description of local or low order structural properties may not guarantee an accurate description of global higher-order structure, such as protein tertiary structure.

The prior perspective related this to an important early result due to Lyubartsev and Laaksonen (LL) regarding structure-based models that reproduce target rdfs.[237] Specifically, LL proved that, given the family of N-particle equilibrium distributions that are consistent with a set of target rdfs, the distribution with maximum entropy corresponds to a pair-additive potential. In practice, this suggests that a conventional structure-based CG model will generate the most disordered configuration distribution that is consistent with the target AA distributions. Consequently, the prior perspective suggested explicitly considering hierarchical many-body correlations and other global structural properties when parameterizing and assessing bottom-up CG models. The prior perspective also suggested developing new

tools for predicting the structural fidelity of bottom-up models without requiring explicit simulations.

As just discussed, recent bottom-up studies have developed a wide range of advanced potentials for explicitly modeling many-body correlations. Although the RE and FM variational principles were introduced about 15 years ago,[82,188,190] they remain powerful frameworks for rigorously parameterizing these potentials to optimally approximate the many-body PMF. In the limit of sufficient sampling and a sufficiently flexible basis set, both variational principles will determine potentials that exactly reproduce the configuration-dependence of the PMF. In particular, the FM variational principle has been widely adopted for parameterizing kernel-based and ANN potentials that reproduce many-body correlations with exceptional accuracy.[194,224,285,292] It may be highly beneficial to further explore the intriguing relationship between FM and score-matching.[305–308]

ML approaches provide a new set of promising tools for optimizing and assessing bottom-up models. For instance, ML methods can be used to identify relatively simple local order parameters that distinguish higher-order structures.[282] These order parameters can then be used both to assess structural fidelity and to identify novel potentials for accurately reproducing these features. Similarly, methods that are employed to "explain" ML calculations[384] may prove useful for identifying and understanding errors in CG models.[303] As suggested in the prior review, it would be highly desirable to gain general insight into the relationship between simple interaction potentials and the higher-order structural features that they can (and cannot) reproduce.[280]

As discussed earlier, recent studies have provided considerable insight into the practical challenges that arise when developing CG models for systems with complex free energy surfaces that reflect distinct conformational states, e.g., unfolded, partially misfolded, and folded conformations of proteins.[223–225] Even if an approximate potential is capable of reproducing the many-body structural correlations within each conformational state, it remains unclear if, when, and how existing bottom-up approaches can parameterize the potential to

give proper statistical weight to these states. Moreover, even if the CG potential is sufficiently flexible to accurately approximate the PMF over the entire conformational space, the accuracy of the CG model will depend upon the sampling available in the mapped ensemble. Recent studies indicate that the structural fidelity of the FM approach may be quite sensitive to errors in modeling the barrier regions between conformational basins. Conversely, the RE approach may be less sensitive to the quality of the mapped ensemble, but will generally require more computational resources to parameterize the CG potential. Future studies should further investigate how more global properties of the mapped ensemble are treated by the RE and FM variational principles. Contrastive or discriminative divergence approaches, as well as approaches based upon variational classification, may prove useful for parameterizing bottom-up potentials that accurately describe both many-body correlations and global properties of the mapped ensemble.[159–162] Similarly, maximum entropy methods,[385–387] PSO,[151,155,156] and other ML tools[158] appear promising for parameterizing CG models that reproduce experimentally determined thermodynamic properties and higher order structural features present in AA simulations. It would be highly desirable to rigorously relate these new data-based approaches to physics-based frameworks for bottom-up coarse-graining.

It would also be highly desirable to develop a unified, rigorous bottom-up approach for simultaneously optimizing both the CG mapping and the interaction potential for self-consistency with the mapped ensemble.[230,238] It may be possible to employ information-theoretic ideas[388] or ML tools[154] to predictively assess this self-consistency without explicit simulations of the CG model. More pragmatically, it may be beneficial to consider simultaneously optimizing the CG mapping and potential not only for accuracy and transferability, but also for computational efficiency and simplicity.

Finally, bottom-up approaches may benefit from leveraging recent advances in the enhanced sampling community. The CG mapping corresponds to a rather simple, albeit extremely large, set of order parameters, while the PMF is the corresponding free energy surface

(FES).[281] Thus, the parameterization of bottom-up potentials corresponds to an approximate free energy calculation in a very high dimensional space. Consequently, bottom-up approaches may significantly benefit from variational approaches that simultaneously optimize both these order parameters and the corresponding FES.[389,390] Conversely, fundamental insights into the CG mapping and the PMF may possibly prove useful for improving enhanced sampling methods.[46]

6.1.4 Representability and transferability limitations

The prior perspective[23] treated the representability and transferability problems as two related, but distinct challenges for bottom-up approaches:

> As a consequence of averaging over atomic structures, the many-body PMF incorporates significant entropic effects from the "hidden" atomistic degrees of freedom. Consequently, thermodynamic properties cannot be represented in their conventional manner.

This motivated two challenges: (1) "formulat[ing] a consistent treatment of thermodynamic properties" and (2) "accurately reproduc[ing] phase transitions." The prior perspective also commented that

> the many-body PMF necessarily depends upon the system and thermodynamic state point for which it is defined. Similarly, a potential that is optimized to approximate the PMF will likely also depend upon system and thermodynamic state point. However, this state point dependence remains poorly understood.

This motivated the challenge of developing predictive methods for determining potentials that accurately approximate the system- and state-point dependence of the PMF. The perspective also posed the challenge of developing "accurate and efficient CG models that accurately treat changes in the local environment."

The bottom-up community has achieved outstanding progress in addressing these challenges. Careful analyses have provided important insight into the temperature- and density-dependence of the PMF.[29,84] This insight not only clarified the fundamental origin and intrinsic duality of representability and transferability challenges,[30] but has also lead to robust computational methods for rigorously addressing these challenges. For instance, the dual and microcanonical approaches accurately describe AA energetics and also predict the temperature-dependence of bottom-up potentials.[187,358,359,361] By properly treating the lost subensemble of AA fluctuations, these methods provide an internally consistent treatment of energetic fluctuations and the specific heat.[187,361] Voth and coworkers have demonstrated that the temperature-dependence of bottom-up potentials can be employed to predict entropic properties that would seem beyond the resolution of the CG model.[319] Volume[216–218,326] and LD potentials[337,338,341,342] provide robust tools for accurately approximating the density-dependence of the PMF and, consequently, reproducing the AA pressure-density equation of state. Moreover, LD potentials,[283,341–343] ultra-CG models,[235,266,320] and surface-hopping approaches[268,269] modulate interactions to reflect their local environment and, in particular, demonstrate outstanding transferability between bulk and interfacial environments. Future studies should further investigate practical methods for predicting the density-dependence of bottom-up pair potentials.[179,318,323,324] We anticipate that an analogous dual approach may prove useful.

Given these advances, we optimistically anticipate that bottom-up approaches will soon completely resolve the representability and transferability challenges associated with modeling one-component systems of molecular liquids or polymeric melts. Specifically, we anticipate that practical bottom-up approaches will accurately describe both structural features and also the pressure-density equation of state across the entire liquid region of the phase diagram. Moreover, these bottom-up models should reproduce energetic and entropic properties, as well as the compressibility, bulk density, and coefficient of thermal expansion. This suggests the somewhat ironic possibility that rigorous bottom-up approaches may potentially

improve the thermodynamic properties of pragmatic top-down models. For instance, if one considers the Martini potential as a pragmatic approximation to the PMF,[25–27] then one could imagine employing the dual approach to decompose Martini potentials into energetic and entropic contributions in order to properly distinguish energetic and entropic driving forces.[391,392]

Several challenges remain for completely resolving the representability and transferability problems that have long plagued bottom-up models. Since much of the recent progress has been realized for relatively simple liquid systems, these advances need to be extended to more complex systems. While LD potentials appear promising for describing the coexistence and interfaces between liquid and vapor phases,[341] future studies need to further investigate their transferability between condensed phases[393] and their ability to reproduce corresponding phase boundaries. Similarly, although bottom-up models can quite accurately reproduce the liquid-vapor interfacial profile, it is not clear that these models accurately reproduce the surface tension.[283,353] Future studies should also investigate how the PMF reflects spatial inhomogenities.[343,345,346] More fundamentally, relatively little progress has been achieved in understanding or modeling the composition-dependence of the PMF.[34,340,344,371] From our perspective, the most pressing representability/transferability challenges involve accurately modeling the chemical potentials of complex mixtures and achieving predictive transferability as a function of chemical composition.

6.2 Emergent challenges and promising directions

Recent studies also indicate several new challenges and promising directions that have emerged.

6.2.1 Beyond the resolution of CG models

Bottom-up approaches have achieved remarkable advances in treating the subensemble of AA configurations that map to a given CG configuration. For instance, recent studies have

developed variational approaches for optimizing observable operators, $A(\mathbf{R})$, that model an arbitrary AA property, $a(\mathbf{r})$, at the resolution of the CG model as a conditioned average, i.e., $A(\mathbf{R}) \simeq \langle a(\mathbf{r}) \rangle_{\mathbf{R}}$, over this subensemble.[29,214,328,359,361,373] Even more remarkably, recent studies have developed probabilistic models, $p_{\mathrm{a|R}}(a|\mathbf{R})$, for AA properties,[186,230] such as classical[187] and quantum mechanical[374,375] energies, that would seem fundamentally beyond the resolution of the CG model. Similarly, new back-mapping approaches have harnessed ML tools to directly sample this subensemble according to the conditioned distribution, $p_{\mathrm{r|R}}$.[132,135] These are extremely exciting advances that dramatically enhance the predictive power of CG models by describing fluctuating microscopic properties that seemed completely inaccessible only a few years ago. We anticipate that the combination of probabilistic inference with ML tools for modeling complicated dependencies may enable CG models to accurately describe increasingly complex high resolution observables.

6.2.2 ML potentials

Classical ML potentials provide a very accurate tool for calculating many quantum mechanical properties. Recently, ML potentials have also emerged as a powerful tool for improving the structural fidelity of bottom-up models for both liquids and miniproteins.[159,194,203,224,285] We anticipate it would be beneficial to further investigate the factors that influence their structural accuracy, including the impact of the data, optimization strategy, and hyperparameters involved in training the potential.[304,394] As discussed earlier, the accuracy of bottom-up ML potentials may be more limited by the quality of the training data than by the flexibility of the potential.[225] Moreover, it would be interesting to consider the impact of regularization upon the coarse-graining formalism. It will certainly be beneficial for future studies to develop ML potentials for increasingly complex systems,[224] as well as to more carefully consider their transferability and thermodynamic properties.[294]

However, the application of ML potentials for bottom-up CG models presents new challenges and considerations. Classical ML potentials accurately describe quantum mechanical

properties at a fraction of the computational cost of, e.g., density functional theory (DFT) methods. In particular, while conventional DFT methods scale quite poorly with system size (e.g., as N_{el}^3 where N_{el} is the number of electrons), classical ML potentials for atomically detailed models scale nearly linearly with the number of atoms, n, in the system (i.e., as $n \ln n$ if one employs efficient methods to treat long-ranged electrostatics). In the context of developing CG models, though, the computational advantages of ML potentials become less clear. While one expects that the computational cost of CG ML potentials will scale nearly linearly with the number of CG sites, N, in the system, this scaling comes with a very large prefactor due to the complexity of the ML potential. Given the remarkable accuracy, simplicity, transferability, and computational efficiency of AA models, the cost of CG ML potentials may limit their application to particularly coarse representations.[31] Such coarse representations may suffer from particularly acute representability and transferability challenges.

Because ML potentials adopt an extremely flexible form with a very large number of parameters, they can accurately reproduce many-body correlations and free energy surfaces for complex systems. However, this high structural fidelity does not necessarily guarantee that the ML potential is actually capturing the physical principles that are necessary for achieving transferability between different systems, environments, or thermodynamic states. For instance, deep ML potentials for bulk water can very accurately describe structural and thermodynamic properties over a wide temperature range, yet fail to describe liquid-vapor interfaces or phase coexistence.[395] In this case, the ML potential captures the total aggregate potential of bulk water but fails to capture the physics that determines this potential, e.g., in terms of 2- and 3-body interactions, as well as polarization effects. As CG models adopt increasingly abstract representations and increasingly complex ML potentials, one expects they may increasingly suffer from over-fitting and, thus, limited transferability.[24]

[31] For simplicity, let us assume that the cost of AA and CG models scale nearly linearly with system size, i.e., as kn and KN, respectively. Then CG models will only provide significant efficiency gains if the degree of coarsening, n/N, is much greater than the ratio of computational prefactors, K/k, i.e., $n/N \gg K/k$. While $K/k \sim 1$ for conventional physics-based potentials, this ratio may be much larger for ML potentials.

Moreover, the use of complex black-box ML potentials also endangers the conceptual advantages of CG models. Consequently, it may be useful to investigate methods for transforming many-body ML potentials into comparatively simpler, but still many-body, physics-based potentials.[286,287] Alternatively, it may be beneficial to describe CG potentials with simpler ML architectures that are more transparent and, presumably, also more computationally efficient. Similarly, it may be useful to develop more concise descriptors for modeling the local many-body environment. In particular, local densities, which bear resemblance to radial symmetry functions commonly employed in ANN potentials,[396] may prove useful as features for efficiently modeling many-body hydrophobic forces. Finally, we anticipate that recent insights into the representability and transferability challenges may also prove useful for ML potentials. For instance, it may be beneficial to develop ML architectures that explicitly account for the energetic and entropic contributions to the PMF. We anticipate that physical insight and rigorous theory may play key roles in improving the accuracy, efficiency, transferability, and utility of ML potentials.

6.2.3 Modeling Complexity

Recent studies have reported remarkable progress in developing fundamental insight and practical computational approaches for resolving prior limitations of bottom-up approaches. Many of these advances have been achieved by considering relatively simple systems, such as molecular liquids or short peptides, that are not intrinsically interesting per se. Nevertheless, because AA models for these systems can be readily characterized with great statistical precision, they provide an ideal environment for elucidating general principles and rigorously assessing new computational methods. We anticipate these simple systems will continue to provide an important testbed for developing theories and computational methods for bottom-up models.

As recent advances become increasingly mature, though, we hope that they will be increasingly transferred to model more complex systems and phenomena of greater interest.

This presents several new practical challenges. Because they fundamentally rely upon statistical properties of high resolution models, there are significant computational and logistical difficulties associated with applying bottom-up methods to complex multi-component systems. For instance, there are logistical challenges associated with obtaining and storing large quantities of AA simulation data. There are sampling challenges associated with ensuring that this data adequately characterizes the AA configurational space. Moreover, while bottom-up methods provide rigorous frameworks for determining complex potentials with many parameters, there are practical challenges associated with optimizing the corresponding high dimensional objective functions. We anticipate that continuing advances in computational resources, high-throughput methods, automated workflows and infrastructure,[397–399] as well as ML approaches may prove useful for surmounting many of these practical difficulties. In particular, bottom-up methods will greatly benefit from standardized repositories that provide access to large scale AA simulation data[400,401] and, thus, eliminate a key activation barrier in developing and testing bottom-up methods for complex systems. However, new fundamental challenges will also arise because the intuition and approximations that apply well for simple systems may not necessarily transfer to the more complex systems that are of interest. For instance, although recent physics-based and ML potentials appear extremely promising, it remains challenging to accurately model hierarchical biomolecular structures without invoking elastic network or Gō type potentials.[32,143]

As suggested in Fig. 1, bottom-up methods are often perceived as appetizing, but impractical approaches that require resources, capabilities, and experience that are not accessible to "non-expert" users. It is worth noting that almost ten years the pragmatic Martini model was already employed to study the organization of lipid membranes consisting of more than 60 distinct types of lipids.[402] While Voth and coworkers recently employed bottom-up methods in developing a CG model for the coronavirus capsid[5] and Lyubartsev and coworkers recently developed a bottom-up model for nucelosomal self-assembly,[143] one may reasonably ask when — and for what systems — the broader simulation community will be able to

savor bottom-up delicacies. Consequently, it is important for the bottom-up community to surmount these practical challenges in order to provide useful, predictive tools for modeling such complex systems and investigating pressing research questions.

It is perhaps instructive to briefly consider the electronic structure field in this context. In some sense, semi-empirical quantum chemical methods and highly parameterized DFT methods are loosely analogous to pragmatic top-down approaches. Similarly, wave function and truly ab initio methods are perhaps loosely analogous to rigorous bottom-up approaches. Because they are computationally efficient and readily accessible, semi-empirical quantum methods have become standard tools that are widely employed by both expert and non-expert users. Conversely, more rigorous quantum methods can provide much greater accuracy, e.g., for transition metal complexes, but are computationally expensive and have historically been rather inaccessible to non-experts.[403,404] The electronic structure community has greatly benefited from an understanding of the regimes in which pragmatic methods are reliable, as well as the identification, critical assessment, and rigorous analysis of simple test cases, e.g., the dissociation of H_2, for which computationally efficient, semiempirical methods fail.[405] Moreover, benchmarking studies have played an important role both for establishing standardized protocols and also for quantitatively comparing various methods for a range of representative systems.[404,406,407] Furthermore, the electronic structure community has greatly benefited from the recent development of high quality, user-friendly, open source software that allows non-experts to expertly apply rigorous quantum methods with confidence.[408,409]

Similarly, we anticipate that pragmatic top-down and rigorous bottom-up approaches will provide important, complementary frameworks for simulating soft materials. We anticipate that the simulation community will benefit from greater understanding of the regimes in which top-down approaches are reliable, as well as the identification, critical assessment, and rigorous analysis of simple test cases for which pragmatic top-down approaches fail. While the bottom-up community should continue developing diverse approaches, we believe

it will also be useful to perform comprehensive benchmarking studies of existing bottom-up methods for a diverse range of model systems with varying characteristics and complexity.[410] These benchmarking studies should not only compare the computational cost and performance of different bottom-up methods but also standardize protocols for applying bottom-up methods to various classes of soft materials. In so doing, these benchmarking studies will make the "zoo" of bottom-up approaches both more comprehensible and also more accessible to the simulation community. One important step in this direction is the further development of new and existing[146,220,411–415] open source software for bottom-up methods that is not only rigorous and robust, but also well documented and easily used by other groups. As in much of computational chemistry, the community will greatly benefit from the development and distribution of automated work-flows for applying bottom-up methods.

Another approach to addressing these practical challenges may be collaborations between groups that develop bottom-up methods and those that focus on applications to particular systems. More generally, the bottom-up community must continue to attract, develop, and retain talented new researchers in order to continue advancing the field. It is essential to convey to new researchers both the practical relevance and the intrinsic elegance of this field, which represents a unique confluence at the frontiers of fundamental statistical mechanics, modern computational methods, and contemporary chemical physics.

Despite the challenges, we anticipate that efforts to extend rigorous bottom-up methods to more complex systems will be well worth the effort. As argued at the outset, bottom-up models provide efficiency that far exceeds AA models and realism that surpasses existing top-down models, while providing insight into fluctuations and interactions that cannot be resolved with analytic or mean field theories. Consequently, bottom-up models hold unique promise for, e.g., investigating the liquid-liquid phase separation of intrinsically disordered proteins,[416] designing improved organic semiconductors,[417] and answering many other compelling molecular questions.

Ultimately, both the premise and the promise of bottom-up methods rely upon the many-

body PMF. Importantly, the PMF is not simply an abstract or mathematical quantity. Rather, it is a physical quantity that describes the consequences of molecular interactions when perceived at the CG resolution. If the CG representation is properly chosen, the PMF should reflect relatively simple, transferable physical principles. In this case, it should be possible to accurately approximate the PMF with relatively simple interaction potentials that properly describe these physical principles. Moreover, one expects that these interactions potentials should be as widely as transferable as the underlying physical principles. Consequently, if one can identify the proper CG representation and develop appropriate interaction potentials, then bottom-up models should be both accurate and predictive. We anticipate that the remarkable progress achieved in recent years, as well as the analysis and directions outlined herein will lead to bottom-up models that realize this tremendous promise.

6.3 Closing thoughts

The bottom-up community has achieved remarkable progress in recent years. Recent studies have provided fundamental insight into the CG representation, as well as the representability and transferability challenges that have long plagued bottom-up approaches. While there still remain fundamental challenges for treating, e.g., the composition- and density-dependence of bottom-up potentials, we anticipate that existing bottom-up methods will very soon provide predictive accuracy and transferability for modeling liquids and polymers. Simultaneously, the development of advanced physics-based and ML potentials have dramatically improved the structural fidelity of bottom-up models for proteins and other complex biological systems. We hope that future studies will not only increasingly extend these recent advances for increasingly complex systems, but also make these advances more accessible to the broader scientific community. We anticipate that the combination of rigorous theory and modern computational methods will continue to rapidly propel the field into the future. Consequently, we anticipate that rigorous bottom-up methods are approaching the threshold of

Figure 6: At the threshold of predictive bottom-up CG models. [Photograph courtesy of A. Noid.]

providing predictive accuracy and transferability for modeling complex phenomena in soft materials. We hope this perspective will contribute usefully to this progress.

Acknowledgements

This work greatly benefited from many generous contributors. WGN gratefully acknowledges his current group members, Katie Kidder, Maria Lesniewski, and Ryan Szukalo, who contributed greatly in revising the manuscript and developing figures, as well as his former group members Wayne Mullinax, Joseph Rudzinski, Nick Dunn, Tommy Foley, Kathryn Lebold, and Michael DeLyser, who all contributed greatly to this perspective. WGN also gratefully acknowledges many extremely helpful comments on this manuscript from Tristan Bereau,

Markus Deserno, Nick Jackson, Siewert-Jan Marrink, Lukas Muechler, Christine Peter, Raffaello Potestio, Joseph Rudzinski, and Scott Shell, as well as from Cecilia Clementi, Frank Noé, and their group members. WGN is particularly grateful to the following for generously sharing their artistry in this work: Maria Lesniewski for the table of contents image and for help with revising the bottom panel of Figure 1, Amy Tong and uTry.it for permission to use the top photograph in Fig. 1, https://www.vecteezy.com/free-vector/human for use of the bottom image in Fig. 1, Siewert-Jan Marrink for sharing his vision for Fig. 1, and A. Noid for sharing the photograph in Fig. 6. WGN acknowledges financial support from the National Science Foundation (Grant Nos. CHE-1856337 and CHE-2154433). Figures 2, 3, and 5 employed VMD.[418] VMD is developed with NIH support by the Theoretical and Computational Biophysics group at the Beckman Institute, University of Illinois at Urbana-Champaign. Finally, WGN expresses his appreciation to Martin Zanni for his patience with this work.

References

(1) Levitt, M. A simplified representation of protein conformations for rapid simulation of protein folding. J. Mol. Biol. **1976**, 104, 59–107.

(2) Baschnagel, J.; Binder, K.; Doruker, P.; Gusev, A. A.; Hahn, O.; Kremer, K.; Mattice, W. L.; Muller-Plathe, F.; Murat, M.; Paul, W. et al. Bridging the gap between atomistic and coarse-grained models of polymers: Status and perspectives. Adv. Poly. Sci. **2000**, 152, 41–156.

(3) Klein, M. L.; Shinoda, W. Large-scale molecular dynamics simulations of self-assembling systems. Science **2008**, 321, 798–800.

(4) Zimmerman, M. I.; Porter, J. R.; Ward, M. D.; Singh, S.; Vithani, N.; Meller, A.; Mallimadugula, U. L.; Kuhn, C. E.; Borowsky, J. H.; Wiewiora, R. P. et al. SARS-CoV-2 simulations go exascale to predict dramatic spike opening and cryptic pockets across the proteome. Nat. Chem. **2021**, 13, 651–659.

(5) Yu, A.; Pak, A. J.; He, P.; Monje-Galvan, V.; Casalino, L.; Gaieb, Z.; Dommer, A. C.; Amaro, R. E.; Voth, G. A. A multiscale coarse-grained model of the SARS-CoV-2 virion. Biophys. J. **2021**, 120, 1097–1104.

(6) Deserno, M. Mesoscopic Membrane Physics: Concepts, Simulations, and Selected Applications. Macromol. Rapid Comm. **2009**, 30, 752–771.

(7) Bowman, G. R. Accurately modeling nanosecond protein dynamics requires at least microseconds of simulation. J. Comp. Chem. **2016**, 37, 558–566.

(8) Girard, M.; Bereau, T. Finite-size transitions in complex membranes. Biophys. J. **2021**, 120, 2436–2443.

(9) Gupta, C.; Sarkar, D.; Tieleman, D. P.; Singharoy, A. The ugly, bad, and good stories

of large-scale biomolecular simulations. Current Opinion in Structural Biology **2022**, 73, 102338.

(10) Kanekal, K. H.; Bereau, T. Resolution limit of data-driven coarse-grained models spanning chemical space. J. Chem. Phys. **2019**, 151, 164106.

(11) Menichetti, R.; Kanekal, K. H.; Bereau, T. Drug–Membrane Permeability across Chemical Space. ACS Central Science **2019**, 5, 290–298.

(12) Shi, G.; Liu, L.; Hyeon, C.; Thirumalai, D. Interphase human chromosome exhibits out of equilibrium glassy dynamics. Nat. Commun. **2018**, 9, 3161.

(13) Statt, A.; Casademunt, H.; Brangwynne, C. P.; Panagiotopoulos, A. Z. Model for disordered proteins with strongly sequence-dependent liquid phase behavior. J. Chem. Phys. **2020**, 152, 075101.

(14) Hyeon, C.; Thirumalai, D. Capturing the essence of folding and functions of biomolecules using coarse-grained models. Nat. Commun. **2011**, 2, 487.

(15) Walker, C. C.; Meek, G. A.; Fobe, T. L.; Shirts, M. R. Using a Coarse-Grained Modeling Framework to Identify Oligomeric Motifs with Tunable Secondary Structure. J. Chem. Theory Comput. **2021**, 17, 6018–6035.

(16) Bereau, T. In Handbook of Materials Modeling: Theory and Modeling; Andreoni, W., Yip, S., Eds.; Springer International Publishing, 2020; pp 1459–1470.

(17) Goldenfeld, N.; Kadanoff, L. P. Simple Lessons from Complexity. Science **1999**, 284, 87–89.

(18) Carroll, L. Sylvie and Bruno Concluded; MacMillan and Co., 1893.

(19) Borges, J. L. On exactitude in science; 1998; Vol. 3.

(20) Diggins, P.; Liu, C.; Deserno, M.; Potestio, R. Optimal Coarse-Grained Site Selection in Elastic Network Models of Biomolecules. J. Chem. Theory Comput. **2019**, 15, 648–664.

(21) Muller, M.; Katsov, K.; Schick, M. Biological and synthetic membranes: What can be learned from a coarse-grained description? Phys. Rep. **2006**, 434, 113–176.

(22) Schmid, F. Toy amphiphiles on the computer: What can we learn from generic models? Macromol. Rapid Comm. **2009**, 30, 741–751.

(23) Noid, W. G. Perspective: coarse-grained models for biomolecular systems. J. Chem. Phys. **2013**, 139, 090901.

(24) Fröhlking, T.; Bernetti, M.; Calonaci, N.; Bussi, G. Toward empirical force fields that match experimental observables. The Journal of Chemical Physics **2020**, 152, 230902.

(25) Marrink, S. J.; Tieleman, D. P. Perspective on the Martini model. Chem. Soc. Rev. **2013**, 42, 6801–6822.

(26) Souza, P. C. T.; Alessandri, R.; Barnoud, J.; Thallmair, S.; Faustino, I.; Grünewald, F.; Patmanidis, I.; Abdizadeh, H.; Bruininks, B. M. H.; Wassenaar, T. A. et al. Martini 3: a general purpose force field for coarse-grained molecular dynamics. Nature Methods **2021**, 18, 382–388.

(27) Marrink, S. J.; Monticelli, L.; Melo, M. N.; Alessandri, R.; Tieleman, D. P.; Souza, P. C. T. Two decades of Martini: Better beads, broader scope. WIREs Computational Molecular Science **2022**,

(28) Noid, W. G. Systematic methods for structurally consistent coarse-grained models. Methods Mol Biol **2013**, 924, 487–531.

(29) Wagner, J. W.; Dama, J. F.; Durumeric, A. E. P.; Voth, G. A. On the representability

problem and the physical meaning of coarse-grained models. J. Chem. Phys. **2016**, 145, 044108.

(30) Dunn, N. J. H.; Foley, T. T.; Noid, W. G. van der Waals perspective on coarse-graining: progress toward solving representability and transferability problems. Acc. Chem. Res. **2016**, 49, 2832–2840.

(31) Minhas, V.; Sun, T.; Mirzoev, A.; Korolev, N.; Lyubartsev, A. P.; Nordenskiöld, L. Modeling DNA Flexibility: Comparison of Force Fields from Atomistic to Multiscale Levels. The Journal of Physical Chemistry B **2020**, 124, 38–49.

(32) Pak, A. J.; Voth, G. A. Advances in coarse-grained modeling of macromolecular complexes. Current Opinion in Structural Biology **2018**, 52, 119–126.

(33) Ghosh, J.; Faller, R. State point dependence of systematically coarse-grained potentials. Mol. Simu. **2007**, 33, 759–767.

(34) Potter, T. D.; Tasche, J.; Wilson, M. R. Assessing the transferability of common top-down and bottom-up coarse-grained molecular models for molecular mixtures. Phys. Chem. Chem. Phys. **2019**, 21, 1912–1917.

(35) Marrink, S.-J. CECAM workshop on "New frontiers in particle-based multiscale and coarse-grained modeling" 17-19 Sep 2018.

(36) Likos, C. N. Effective interactions in soft condensed matter physics. Phys. Rep. **2001**, 348, 267 – 439.

(37) Murtola, T.; Bunker, A.; Vattulainen, I.; Deserno, M.; Karttunen, M. Multiscale modeling of emergent materials: Biological and soft matter. Phys. Chem. Chem. Phys. **2009**, 11, 1869–92.

(38) Peter, C.; Kremer, K. Multiscale simulation of soft matter systems. Faraday Discuss. **2010**, 144, 9–24.

(39) Riniker, S.; Allison, J. R.; van Gunsteren, W. F. On developing coarse-grained models for biomolecular simulation: a review. Phys. Chem. Chem. Phys. **2012**, 14, 12423–12430.

(40) Lu, L.; Voth, G. A. The multiscale coarse-graining method. Adv. Chem. Phys. **2012**, 149, 47–81.

(41) Brini, E.; Algaer, E. A.; Ganguly, P.; Li, C.; Rodríguez-Ropero, F.; van der Vegt, N. F. A. Systematic coarse-graining methods for soft matter simulations - a review. Soft Matter **2013**, 9, 2108–2119.

(42) Li, Y.; Abberton, B. C.; Kröger, M.; Liu, W. K. Challenges in Multiscale Modeling of Polymer Dynamics. Polymers **2013**, 5, 751–832.

(43) Saunders, M. G.; Voth, G. A. Coarse-graining methods for computational biology. Annu. Rev. Biophys. **2013**, 42, 73–93.

(44) Potestio, R.; Peter, C.; Kremer, K. Computer Simulations of Soft Matter: Linking the Scales. Entropy **2014**, 16, 4199–4245.

(45) Shell, M. S. Adv. Chem. Phys.; John Wiley & Sons, Inc., 2016; pp 395–441.

(46) Giulini, M.; Rigoli, M.; Mattiotti, G.; Menichetti, R.; Tarenzi, T.; Fiorentini, R.; Potestio, R. From System Modeling to System Analysis: The Impact of Resolution Level and Resolution Distribution in the Computer-Aided Investigation of Biomolecules. Front. Mol. Biosci. **2021**, 8, 676976.

(47) Kidder, K. M.; Szukalo, R. J.; Noid, W. G. Energetic and entropic considerations for coarse-graining. Eur. Phys. J. B **2021**, 94, 153.

(48) Jin, J.; Pak, A. J.; Durumeric, A. E. P.; Loose, T. D.; Voth, G. A. Bottom-up Coarse-Graining: Principles and Perspectives. Journal of Chemical Theory and Computation **2022**, 18, 5759–5791.

(49) Sun, T.; Minhas, V.; Korolev, N.; Mirzoev, A.; Lyubartsev, A. P.; Nordenskiöld, L. Bottom-Up Coarse-Grained Modeling of DNA. Frontiers in Molecular Biosciences **2021**, 8, 645527.

(50) Liwo, A.; Czaplewski, C.; Sieradzan, A. K.; Lipska, A. G.; Samsonov, S. A.; Murarka, R. K. Theory and Practice of Coarse-Grained Molecular Dynamics of Biologically Important Systems. Biomolecules **2021**, 11, 1347.

(51) Durumeric, A. E.; Charron, N. E.; Templeton, C.; Musil, F.; Bonneau, K.; Pasos-Trejo, A. S.; Chen, Y.; Kelkar, A.; Noé, F.; Clementi, C. Machine learned coarse-grained protein force-fields: Are we there yet? Current Opinion in Structural Biology **2023**, 79, 102533.

(52) Gartner, T. E.; Jayaraman, A. Modeling and Simulations of Polymers: A Roadmap. Macromolecules **2019**, 52, 755–786.

(53) Joshi, S. Y.; Deshmukh, S. A. A review of advancements in coarse-grained molecular dynamics simulations. Mol. Simu. **2020**, 47, 786–803.

(54) Dhamankar, S.; Webb, M. A. Chemically specific coarse-graining of polymers: Methods and prospects. Journal of Polymer Science **2021**, 59, 2613–2643.

(55) Ye, H.; Xian, W.; Li, Y. Machine Learning of Coarse-Grained Models for Organic Molecules and Polymers: Progress, Opportunities, and Challenges. ACS Omega **2021**, 6, 1758–1772.

(56) Schmid, F. Understanding and Modeling Polymers: The Challenge of Multiple Scales. ACS Polymers Au **2023**, 3, 28–58.

(57) Rudzinski, J. F. Recent progress towards chemically-specific coarse-grained simulation models with consistent dynamical properties. Comput. **2019**, 7, 42.

(58) Klippenstein, V.; Tripathy, M.; Jung, G.; Schmid, F.; van der Vegt, N. F. A. Introducing Memory in Coarse-Grained Molecular Simulations. The Journal of Physical Chemistry B **2021**, 125, 4931–4954.

(59) Schilling, T. Coarse-grained modelling out of equilibrium. Physics Reports **2022**, 972, 1–45.

(60) Arkhipov, A.; Freddolino, P.; Schulten, K. Stability and dynamics of virus capsids described by coarse-grained modeling. Structure **2006**, 129, 1767–77.

(61) Gohlke, H.; Thorpe, M. F. A natural coarse graining for simulating large biomolecular motion. Biophys. J. **2006**, 91, 2115–20.

(62) Zhang, Z. Y.; Lu, L. Y.; Noid, W. G.; Krishna, V.; Pfaendtner, J.; Voth, G. A. A Systematic Methodology for Defining Coarse-Grained Sites in Large Biomolecules. Biophys. J. **2008**, 95, 5073–5083.

(63) Li, Z.; Wellawatte, G. P.; Chakraborty, M.; Gandhi, H. A.; Xu, C.; White, A. D. Graph neural network based coarse-grained mapping prediction. Chem. **2020**, 11, 9524–9531.

(64) Bereau, T.; Kremer, K. Automated Parametrization of the Coarse-Grained Martini Force Field for Small Organic Molecules. J. Chem. Theory Comput. **2015**, 11, 2783–2791.

(65) Potter, T. D.; Barrett, E. L.; Miller, M. A. Automated Coarse-Grained Mapping Algorithm for the Martini Force Field and Benchmarks for Membrane–Water Partitioning. Journal of Chemical Theory and Computation **2021**, 17, 5777–5791.

(66) Buslaev, P.; Gushchin, I. Effects of Coarse Graining and Saturation of Hydrocarbon Chains on Structure and Dynamics of Simulated Lipid Molecules. Sci. Rep. **2017**, 7, 11476.

(67) Amadei, A.; Linssen, A. B. M.; Berendsen, H. J. C. Essential dynamics of proteins. Proteins **1993**, 17, 412 – 425.

(68) Madsen, J. J.; Sinitskiy, A. V.; Li, J.; Voth, G. A. Highly Coarse-Grained Representations of Transmembrane Proteins. J. Chem. Theory Comput. **2017**, 13, 935–944.

(69) Li, M.; Zhang, J. Z. H.; Xia, F. A new algorithm for construction of coarse-grained sites of large biomolecules. J. Comp. Chem. **2016**, 37, 795–804.

(70) Wu, Z.; Zhang, Y.; Zhang, J. Z.; Xia, K.; Xia, F. Determining Optimal Coarse-Grained Representation for Biomolecules Using Internal Cluster Validation Indexes. J. Comp. Chem. **2020**, 41, 14–20.

(71) Tirion, M. M. Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis. Phys. Rev. Lett. **1996**, 77, 1905–1908.

(72) Bahar, I.; Lezon, T. R.; Bakan, A.; Shrivastava, I. H. Normal Mode Analysis of Biomolecular Structures: Functional Mechanisms of Membrane Proteins. Chem. Rev. **2010**, 110, 1463–1497.

(73) Goldenfeld, N. Lectures on Phase Transitions and the Renormalization Group; Westview Press, 1992.

(74) Orioli, S.; Faccioli, P. Dimensional reduction of Markov state models from renormalization group theory. J. Chem. Phys. **2016**, 145, 124120.

(75) Koehl, P.; Poitevin, F.; Navaza, R.; Delarue, M. The renormalization group and its applications to generating coarse-grained models of large biological molecular systems. J. Chem. Theory Comput. **2017**, 13, 1424–1438.

(76) Chakraborty, M.; Xu, C.; White, A. D. Encoding and selecting coarse-grain mapping operators with hierarchical graphs. J. Chem. Phys. **2018**, 149, 134106.

(77) Webb, M. A.; Delannoy, J.-Y.; de Pablo, J. J. Graph-Based Approach to Systematic Molecular Coarse-Graining. J. Chem. Theory Comput. **2018**,

(78) Kullback, S.; Leibler, R. A. On Information and Sufficiency. Ann. Math. Stat. **1951**, 22, 79–86.

(79) Cover, T. M.; Thomas, J. A. Elements of Information Theory, 2nd ed.; Wiley Interscience, 2006.

(80) Isihara, A. Gibbs-Bogoliubov inequality. J. Phys. A: Math., Nucl., Gen. **1968**, 1, 539–548.

(81) Touchette, H. The large deviation approach to statistical mechanics. Physics Reports **2009**, 478, 1–69.

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

(83) Rudzinski, J. F.; Noid, W. G. Coarse-graining entropy, forces, and structures. J. Chem. Phys. **2011**, 135, 214101.

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

(85) Giulini, M.; Menichetti, R.; Shell, M. S.; Potestio, R. An Information-Theory-Based Approach for Optimal Model Reduction of Biomolecules. J. Chem. Theory Comput. **2020**, 16, 6795–6813.

(86) Lin, S.-T.; Blanco, M.; Goddard, W. Two-phase thermodynamic method for accurate free energies for liquids directly from molecular dynamics simulations. J. Chem. Phys. **2003**, 119, 11792 – 11805.

(87) Lin, S.-T.; Maiti, P. K.; Goddard, W. A. Two-Phase Thermodynamic Model for Efficient and Accurate Absolute Entropy of Water from Molecular Dynamics Simulations. The Journal of Physical Chemistry B **2010**, 114, 8191–8198.

(88) Bernhardt, M. P.; Dallavalle, M.; Van der Vegt, N. F. Application of the 2PT model to understanding entropy change in molecular coarse-graining. Soft Mater. **2020**, 18, 274–289.

(89) Tozzini, V. Coarse-grained models for proteins. Curr. Opin. Struct. Biol. **2005**, 15, 144 – 150.

(90) Maupetit, J.; Gautier, R.; Tufféry, P. SABBAC: online Structural Alphabet-based protein BackBone reconstruction from Alpha-Carbon trace. Nucl. Acids Res. **2006**, 34, W147–51.

(91) Rotkiewicz, P.; Skolnick, J. Fast procedure for reconstruction of full-atom protein models from reduced representations. J. Comp. Chem. **2008**, 29, 1460–5.

(92) Holtzman, R.; Giulini, M.; Potestio, R. Making sense of complex systems through resolution, relevance, and mapping entropy. Phys. Rev. E **2022**, 106, 044101.

(93) Mele, M.; Covino, R.; Potestio, R. Information-theoretical measures identify accurate low-resolution representations of protein configurational space. Soft Matter **2022**, 18, 7064–7074.

(94) Song, J.; Marsili, M.; Jo, J. Resolution and relevance trade-offs in deep learning. Journal of Statistical Mechanics: Theory and Experiment **2018**, 2018, 123406.

(95) Delvenne, J.-C.; Yaliraki, S. N.; Barahona, M. Stability of graph communities across time scales. Proc. Natl. Acad. Sci. U.S.A. **2010**, 107, 12755–12760.

(96) Foley, T. T.; Kidder, K. M.; Shell, M. S.; Noid, W. Exploring the landscape of model representations. Proc. Natl. Acad. Sci. U.S.A. **2020**, 117, 24061–24068.

(97) Wang, W.; Gómez-Bombarelli, R. Coarse-graining auto-encoders for molecular dynamics. npj Comput. Mat. **2019**, 5, 125.

(98) Ruza, J.; Wang, W.; Schwalbe-Koda, D.; Axelrod, S.; Harris, W. H.; Gómez-Bombarelli, R. Temperature-transferable coarse-graining of ionic liquids with dual graph convolutional neural networks. J. Chem. Phys. **2020**, 153, 164501.

(99) Flory, P. J.; Gordon, M.; McCrum, N. G. Statistical Thermodynamics of Random Networks [and Discussion]. Proc. Roy. Soc. Lond. A: Math. Phys. Sci. **1976**, 351, 351–380.

(100) Haliloglu, T.; Bahar, I.; Erman, B. Gaussian Dynamics of Folded Proteins. Phys. Rev. Lett. **1997**, 79, 3090–3093.

(101) Girvan, M.; Newman, M. E. J. Community structure in social and biological networks. Proc. Natl. Acad. Sci. U.S.A. **2002**, 99, 7821–7826.

(102) Newman, M. E. J.; Girvan, M. Finding and evaluating community structure in networks. Phys. Rev. E **2004**, 69, 026113, Publisher: American Physical Society.

(103) Fortunato, S. Community detection in graphs. Phys. Rep. **2010**, 486, 75–174.

(104) Menichetti, R.; Giulini, M.; Potestio, R. A journey through mapping space: characterising the statistical and metric properties of reduced representations of macromolecules. The European Physical Journal B **2021**, 94, 204.

(105) Boninsegna, L.; Banisch, R.; Clementi, C. A Data-Driven Perspective on the Hierarchical Assembly of Molecular Structures. J. Chem. Theory Comput. **2018**, 14, 453–460, Publisher: American Chemical Society.

(106) Coifman, R. R.; Kevrekidis, I. G.; Lafon, S.; Maggioni, M.; Nadler, B. Diffusion Maps, Reduction Coordinates, and Low Dimensional Representation of Stochastic Systems. Multiscale Model. Simul. **2008**, 7, 842–864.

(107) Husic, B. E.; Pande, V. S. Markov State Models: From an Art to a Science. Journal of the American Chemical Society **2018**, 140, 2386–2396.

(108) Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y. B. et al. Atomic-Level Characterization of the Structural Dynamics of Proteins. Science **2010**, 330, 341–346.

(109) Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E. How Fast-Folding Proteins Fold. Science **2011**, 334, 517–520.

(110) Panchenko, A. R.; Luthey-Schulten, Z.; Wolynes, P. G. Foldons, protein structural modules, and exons. Proc. Natl. Acad. Sci. U.S.A. **1996**, 93, 2008–2013.

(111) Ayton, G. S.; Noid, W. G.; Voth, G. A. Multiscale Modeling of biomolecular systems: In serial and in parallel. Curr. Opin. Struct. Biol. **2007**, 17, 192–8.

(112) Hunkler, S.; Lemke, T.; Peter, C.; Kukharenko, O. Back-mapping based sampling: Coarse grained free energy landscapes as a guideline for atomistic exploration. The Journal of Chemical Physics **2019**, 151, 154102.

(113) Thaler, S.; Praprotnik, M.; Zavadlav, J. Back-mapping augmented adaptive resolution simulation. The Journal of Chemical Physics **2020**, 153, 164118.

(114) Praprotnik, M.; Delle Site, L.; Kremer, K. A macromolecule in a solvent: Adaptive resolution molecular dynamics simulation. J. Chem. Phys. **2007**, 126, 134902.

(115) Potestio, R.; Fritsch, S.; Español, P.; Delgado-Buscalioni, R.; Kremer, K.; Everaers, R.; Donadio, D. Hamiltonian adaptive resolution simulation for molecular liquids. Phys. Rev. Lett. **2013**, 110, 108301.

(116) Heath, A. P.; Kavraki, L. E.; Clementi, C. From coarse-grain to all-atom: Toward multiscale analysis of protein landscapes. Proteins **2007**, 25, 646–661.

(117) Wassenaar, T. A.; Pluhackova, K.; Böckmann, R. A.; Marrink, S. J.; Tieleman, D. P. Going Backward: A Flexible Geometric Approach to Reverse Transformation from Coarse Grained to Atomistic Models. Journal of Chemical Theory and Computation **2014**, 10, 676–690.

(118) Shimizu, M.; Takada, S. Reconstruction of Atomistic Structures from Coarse-Grained Models for Protein–DNA Complexes. Journal of Chemical Theory and Computation **2018**, 14, 1682–1694.

(119) Li, M.; Teng, B.; Lu, W.; Zhang, J. Z. Atomic-level reconstruction of biomolecules by a rigid-fragment- and local-frame-based (RF-LF) strategy. Journal of Molecular Modeling **2020**, 26, 31.

(120) Isralewitz, B.; Gao, M.; Schulten, K. Steered molecular dynamics and mechanical functions of proteins. Current Opinion in Structural Biology **2001**, 11, 224–230.

(121) Poblete, S.; Bottaro, S.; Bussi, G. A nucleobase-centered coarse-grained representation for structure prediction of RNA motifs. Nucleic Acids Research **2018**, 46, 1674–1683.

(122) Poblete, S.; Bottaro, S.; Bussi, G. Effects and limitations of a nucleobase-driven backmapping procedure for nucleic acids using steered molecular dynamics. Biochemical and Biophysical Research Communications **2018**, 498, 352–358.

(123) Krajniak, J.; Pandiyan, S.; Nies, E.; Samaey, G. Generic Adaptive Resolution Method for Reverse Mapping of Polymers from Coarse-Grained to Atomistic Descriptions. Journal of Chemical Theory and Computation **2016**, 12, 5549–5562.

(124) Krajniak, J.; Zhang, Z.; Pandiyan, S.; Nies, E.; Samaey, G. Reverse mapping method for complex polymer systems. Journal of Computational Chemistry **2018**, 39, 648–664.

(125) Zhang, G.; Chazirakis, A.; Harmandaris, V. A.; Stuehn, T.; Daoulas, K. C.; Kremer, K.

Hierarchical modelling of polystyrene melts: from soft blobs to atomistic resolution. Soft Matter **2019**, 15, 289–302.

(126) Peng, J.; Yuan, C.; Ma, R.; Zhang, Z. Backmapping from Multiresolution Coarse-Grained Models to Atomic Structures of Large Biomolecules by Restrained Molecular Dynamics Simulations Using Bayesian Inference. Journal of Chemical Theory and Computation **2019**, 15, 3344–3353.

(127) An, Y.; Deshmukh, S. A. Machine learning approach for accurate backmapping of coarse-grained models to all-atom models. Chemical Communications **2020**, 56, 9312–9315.

(128) Louison, K. A.; Dryden, I. L.; Laughton, C. A. GLIMPS: A Machine Learning Approach to Resolution Transformation for Multiscale Modeling. Journal of Chemical Theory and Computation **2021**, 17, 7930–7937.

(129) Yang, W.; Zhang, X.; Tian, Y.; Wang, W.; Xue, J.-H.; Liao, Q. Deep Learning for Single Image Super-Resolution: A Brief Review. IEEE Transactions on Multimedia **2019**, 21, 3106–3121.

(130) Li, W.; Burkhart, C.; Polińska, P.; Harmandaris, V.; Doxastakis, M. Backmapping coarse-grained macromolecules: An efficient and versatile machine learning approach. The Journal of Chemical Physics **2020**, 153, 041101.

(131) Christofi, E.; Chazirakis, A.; Chrysostomou, C.; Nicolaou, M. A.; Li, W.; Doxastakis, M.; Harmandaris, V. A. Deep convolutional neural networks for generating atomistic configurations of multi-component macromolecules from coarse-grained models. The Journal of Chemical Physics **2022**, 157, 184903.

(132) Stieffenhofer, M.; Wand, M.; Bereau, T. Adversarial Reverse Mapping of Equilibrated Condensed-Phase Molecular Structures. Machine Learning: Science and Technology **2020**, 1, 045014.

(133) Creswell, A.; White, T.; Dumoulin, V.; Arulkumaran, K.; Sengupta, B.; Bharath, A. A. Generative Adversarial Networks: An Overview. IEEE Signal Processing Magazine **2018**, 35, 53–65.

(134) Stieffenhofer, M.; Bereau, T.; Wand, M. Adversarial reverse mapping of condensed-phase molecular structures: Chemical transferability. APL Materials **2021**, 9, 031107.

(135) Wang, W.; Xu, M.; Cai, C.; Miller, B. K.; Smidt, T.; Wang, Y.; Tang, J.; Gómez-Bombarelli, R. Generative Coarse-Graining of Molecular Conformations. **2022**, arXiv:2201.12176 [physics], doi: 10.48550/ARXIV.2201.12176.

(136) Shmilovich, K.; Stieffenhofer, M.; Charron, N. E.; Hoffmann, M. Temporally Coherent Backmapping of Molecular Trajectories From Coarse-Grained to Atomistic Resolution. The Journal of Physical Chemistry A **2022**, 126, 9124–9139.

(137) Kirkwood, J. G. Statistical mechanics of fluid mixtures. J. Chem. Phys. **1935**, 3, 300–313.

(138) Liwo, A.; Ołdziej, S.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Scheraga, H. A. A united-residue force field for off-lattice protein-structure simulations. I. Functional forms and parameters of long-range side-chain interaction potentials from protein crystal data. J. Comp. Chem. **1997**, 18, 849–873.

(139) Akkermans, R. L. C.; Briels, W. J. Coarse-grained interactions in polymer melts: A variational approach. J. Chem. Phys. **2001**, 115, 6210–6219.

(140) Tóth, G. Interactions from diffraction data: historical and comprehensive overview of simulation assisted methods. Journal of Physics: Condensed Matter **2007**, 19, 335220.

(141) Lyubartsev, A. P.; Laaksonen, A. Calculation of effective interaction potentials from radial distribution functions: a reverse Monte Carlo approach. Phys. Rev. E Stat. Phys. Plasmas Fluids Relat. Interdiscip. Topics **1995**, 52, 3730–3737.

(142) Müller-Plathe, F. Coarse-graining in polymer simulation: From the atomistic to the mesoscopic scale and back. ChemPhysChem **2002**, 3, 754 – 769.

(143) Sun, T.; Minhas, V.; Mirzoev, A.; Korolev, N.; Lyubartsev, A. P.; Nordenskiöld, L. A Bottom-Up Coarse-Grained Model for Nucleosome–Nucleosome Interactions with Explicit Ions. Journal of Chemical Theory and Computation **2022**, 18, 3948–3960.

(144) Peter, C.; Delle Site, L.; Kremer, K. Classical simulations from the atomistic to the mesoscale and back: coarse graining an azobenzene liquid crystal. Soft Matter **2008**, 4, 859–869.

(145) Hadley, K. R.; McCabe, C. A structurally relevant coarse-grained model for cholesterol. Biophys. J. **2010**, 99, 2896–905.

(146) Bernhardt, M. P.; Hanke, M.; van der Vegt, N. F. Stability, Speed, and Constraints for Structural Coarse-Graining in VOTCA. Journal of Chemical Theory and Computation **2023**, acs.jctc.2c00665.

(147) Wang, H.; Junghans, C.; Kremer, K. Comparative atomistic and coarse-grained study of water: What do we lose by coarse-graining? Eur. Phys. J. E: Soft Matter Biol. Phys. **2009**, 28, 221–229.

(148) Bernhardt, M. P.; Hanke, M.; van der Vegt, N. F. A. Iterative integral equation methods for structural coarse-graining. J. Chem. Phys. **2021**, 154, 084118.

(149) Bejagam, K. K.; Singh, S.; An, Y.; Deshmukh, S. A. Machine-Learned Coarse-Grained Models. J. Phys. Chem. Lett. **2018**, 9, 4667–4672.

(150) Bejagam, K. K.; Singh, S.; An, Y.; Berry, C.; Deshmukh, S. A. PSO-Assisted Development of New Transferable Coarse-Grained Water Models. J. Phys. Chem. B **2018**, 122, 1958–1971.

(151) An, Y.; Bejagam, K. K.; Deshmukh, S. A. Development of New Transferable Coarse-Grained Models of Hydrocarbons. J. Phys. Chem. B **2018**, 122, 7143–7153.

(152) McDonagh, J. L.; Shkurti, A.; Bray, D. J.; Anderson, R. L.; Pyzer-Knapp, E. O. Utilizing Machine Learning for Efficient Parameterization of Coarse Grained Molecular Force Fields. J. Chem. Inf. Model. **2019**, 59, 4278–4288.

(153) Chan, H.; Cherukara, M. J.; Narayanan, B.; Loeffler, T. D.; Benmore, C.; Gray, S. K.; Sankaranarayanan, S. K. R. S. Machine learning coarse grained models for water. Nature Communications **2019**, 10, 379.

(154) Shireen, Z.; Weeratunge, H.; Menzel, A.; Phillips, A. W.; Larson, R. G.; Smith-Miles, K.; Hajizadeh, E. A machine learning enabled hybrid optimization framework for efficient coarse-graining of a model polymer. npj Computational Materials **2022**, 8, 224.

(155) Empereur-Mot, C.; Pesce, L.; Doni, G.; Bochicchio, D.; Capelli, R.; Perego, C.; Pavan, G. M. *Swarm-CG* : Automatic Parametrization of Bonded Terms in MARTINI-Based Coarse-Grained Models of Simple to Complex Molecules *via* Fuzzy Self-Tuning Particle Swarm Optimization. ACS Omega **2020**, acsomega.0c05469.

(156) Empereur-mot, C.; Capelli, R.; Perrone, M.; Caruso, C.; Doni, G.; Pavan, G. M. Automatic multi-objective optimization of coarse-grained lipid force fields using *SwarmCG*. The Journal of Chemical Physics **2022**, 156, 024801.

(157) Panaretos, V. M.; Zemel, Y. Statistical Aspects of Wasserstein Distances. Annual Review of Statistics and Its Application **2019**, 6, 405–431.

(158) Thaler, S.; Zavadlav, J. Learning neural network potentials from experimental data via Differentiable Trajectory Reweighting. Nature Communications **2021**, 12, 6884.

(159) Lemke, T.; Peter, C. Neural Network Based Prediction of Conformational Free Energies - A New Route toward Coarse-Grained Simulation Models. J. Chem. Theory Comput. **2017**, 13, 6213–6221.

(160) Durumeric, A. E. P.; Voth, G. A. Adversarial-residual-coarse-graining: Applying machine learning theory to systematic molecular coarse-graining. J. Chem. Phys. **2019**, 151, 124110.

(161) Ding, X.; Zhang, B. Contrastive Learning of Coarse-Grained Force Fields. Journal of Chemical Theory and Computation **2022**, 18, 6334–6344.

(162) Jumper, J. M.; Faruk, N. F.; Freed, K. F.; Sosnick, T. R. Trajectory-based training enables protein simulations with accurate folding and Boltzmann ensembles in cpu-hours. PLOS Computational Biology **2018**, 14, e1006578.

(163) Lyubartsev, A.; Mirzoev, A.; Chen, L. J.; Laaksonen, A. Systematic coarse-graining of molecular models by the Newton inversion method. Faraday Discuss. **2010**, 144, 43–56.

(164) Hansen, J.-P.; McDonald, I. R. Theory of Simple Liquids, 2nd ed.; Academic Press: San Diego, CA USA, 1990.

(165) Guenza, M.; Dinpajooh, M.; McCarty, J.; Lyubimov, I. Accuracy, transferability, and efficiency of coarse-grained models of molecular liquids. J. Phys. Chem. B **2018**, 122, 10257–10278.

(166) Dinpajooh, M.; Guenza, M. G. On the Density Dependence of the Integral Equation Coarse-Graining Effective Potential. J. Phys. Chem. B **2018**, 122, 3426–3440.

(167) Martin, T. B.; Gartner, T. E.; Jones, R. L.; Snyder, C. R.; Jayaraman, A. pyPRISM: A Computational Tool for Liquid-State Theory Calculations of Macromolecular Materials. Macromolecules **2018**, 51, 2906–2922.

(168) Moradzadeh, A.; Aluru, N. R. Transfer-Learning-Based Coarse-Graining Method for Simple Fluids: Toward Deep Inverse Liquid-State Theory. J. Phys. Chem. Lett. **2019**, 10, 1242–1250.

(169) Berressem, F.; Nikoubashman, A. BoltzmaNN: Predicting effective pair potentials and equations of state using neural networks. J. Chem. Phys. **2021**, 154, 124123.

(170) Wang, Y. T.; Noid, W. G.; Liu, P.; Voth, G. A. Effective force coarse-graining. Phys. Chem. Chem. Phys. **2009**, 11, 2002–2015.

(171) Brini, E.; Marcon, V.; van der Vegt, N. F. A. Conditional reversible work method for molecular coarse graining applications. Phys. Chem. Chem. Phys. **2011**, 13, 10468–74.

(172) Deichmann, G.; van der Vegt, N. F. A. Conditional Reversible Work Coarse-Grained Models with Explicit Electrostatics—An Application to Butylmethylimidazolium Ionic Liquids. J. Chem. Theory Comput. **2019**, 15, 1187–1198.

(173) Dallavalle, M.; van der Vegt, N. F. Evaluation of mapping schemes for systematic coarse graining of higher alkanes. Phys. Chem. Chem. Phys. **2017**, 19, 23034–23042.

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

(175) Carmichael, S. P.; Shell, M. S. A new multiscale algorithm and its application to coarse-grained peptide models for self-assembly. J. Phys. Chem. B **2012**, 116, 8383–93.

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

(177) Murtola, T.; Karttunen, M.; Vattulainen, I. Systematic coarse graining from structure using internal states: Application to phospholipid/cholesterol bilaye. J. Chem. Phys. **2009**, 131, 055101.

(178) Henderson, R. L. A uniqueness theorem for fluid pair correlation functions. Phys. Lett. A **1974**, 49, 197–8.

(179) Johnson, M. E.; Head-Gordon, T.; Louis, A. A. Representability problems for coarse-grained water potentials. J. Chem. Phys. **2007**, 126, 144509.

(180) Frommer, F.; Hanke, M.; Jansen, S. A note on the uniqueness result for the inverse Henderson problem. J. Math. Phys. **2019**, 60, 093303.

(181) Potestio, R. Effective interactions in soft condensed matter physics. J. Unsolved Quest. **2013**, 3, 13.

(182) Wang, H.; Stillinger, F. H.; Torquato, S. Sensitivity of pair statistics on pair potentials in many-body systems. J. Chem. Phys. **2020**, 153, 124106.

(183) Shen, K.; Sherck, N.; Nguyen, M.; Yoo, B.; Köhler, S.; Speros, J.; Delaney, K. T.; Fredrickson, G. H.; Shell, M. S. Learning composition-transferable coarse-grained models: Designing external potential ensembles to maximize thermodynamic information. J. Chem. Phys. **2020**, 153, 154116.

(184) Moradzadeh, A.; Motevaselian, M. H.; Mashayak, S. Y.; Aluru, N. R. Coarse-Grained Force Field for Imidazolium-Based Ionic Liquids. J. Chem. Theory Comput. **2018**, 14, 3252–3261.

(185) Harmandaris, V.; Kalligiannaki, E.; Katsoulakis, M.; Plecháăc, P. Path-space variational inference for non-equilibrium coarse-grained systems. J. Comp. Phys. **2016**, 314, 355–383.

(186) Schöberl, M.; Zabaras, N.; Koutsourelakis, P.-S. Predictive coarse-graining. J. Comp. Phys. **2017**, 333, 49–77.

(187) Pretti, E.; Shell, M. S. A microcanonical approach to temperature-transferable coarse-grained models using the relative entropy. J. Chem. Phys. **2021**, 155, 094102.

(188) Izvekov, S.; Voth, G. A. A multiscale coarse-graining method for biomolecular systems. J. Phys. Chem. B **2005**, 109, 2469 – 2473.

(189) Izvekov, S.; Voth, G. A. Multiscale coarse graining of liquid-state systems. J. Chem. Phys. **2005**, 123, 134105.

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

(191) Ciccotti, G.; Kapral, R.; Vanden-Eijnden, E. Blue Moon Sampling, Vectorial Reaction Coordinates, and Unbiased Constained Dynamics. ChemPhysChem **2005**, 6, 1809–14.

(192) Dama, J. F.; Sinitskiy, A. V.; McCullagh, M.; Weare, J.; Roux, B.; Dinner, A. R.; Voth, G. A. The Theory of Ultra-Coarse-Graining. 1. General Principles. J. Chem. Theory Comput. **2013**, 9, 2466–80.

(193) Kalligiannaki, E.; Harmandaris, V.; Katsoulakis, M. A.; Plechac, P. The geometry of generalized force matching and related information metrics in coarse-graining of molecular systems. J. Chem. Phys. **2015**, 143, 084105.

(194) Wang, J.; Olsson, S.; Wehmeyer, C.; Pérez, A.; Charron, N. E.; De Fabritiis, G.; Noé, F.; Clementi, C. Machine learning of coarse-grained molecular dynamics force fields. ACS Cent. Sci. **2019**, 5, 755–767.

(195) Krämer, A.; Durumeric, A. P.; Charron, N. E.; Chen, Y.; Clementi, C.; Noé, F. Statistically Optimal Force Aggregation for Coarse-Graining Molecular Dynamics. 2023; http://arxiv.org/abs/2302.07071, arXiv:2302.07071 [physics, stat], doi: 10.48550/ARXIV.2302.07071.

(196) Noid, W. G.; Chu, J.-W.; Ayton, G. S.; Voth, G. A. Multiscale coarse-graining and

structural correlations: Connections to liquid state theory. J. Phys. Chem. B **2007**, 111, 4116–4127.

(197) Mullinax, J. W.; Noid, W. G. A generalized Yvon-Born-Green theory for molecular systems. Phys. Rev. Lett. **2009**, 103, 198104.

(198) Mullinax, J. W.; Noid, W. G. A generalized Yvon-Born-Green theory for determining coarse-grained interaction potentials. J. Phys. Chem. C **2010**, 114, 5661–5674.

(199) Rudzinski, J. F.; Noid, W. G. A generalized-Yvon-Born-Green method for coarse-grained modeling. Eur. Phys. J.: Spec. Top. **2015**, 224, 2193–2216.

(200) Rudzinski, J. F.; Noid, W. G. Investigation of coarse-grained mappings via an iterative generalized Yvon-Born-Green method. J. Phys. Chem. B **2014**, 118, 8295–8312.

(201) Chorin, A. J. Conditional expectations and renormalization. Multiscale Model. Simul. **2003**, 1, 105–118.

(202) Chorin, A. J.; Hald, O. H. Stochastic Tools in Mathematics and Science; Springer: New York, NY USA, 2006.

(203) Zhang, L.; Han, J.; Wang, H.; Car, R.; E, W. DeePCG: Constructing coarse-grained models via deep neural networks. J. Chem. Phys. **2018**, 149, 034101.

(204) Khot, A.; Shiring, S. B.; Savoie, B. M. Evidence of information limitations in coarse-grained models. J. Chem. Phys. **2019**, 151, 244105.

(205) Jin, J.; Han, Y.; Pak, A. J.; Voth, G. A. A new one-site coarse-grained model for water: Bottom-up many-body projected water (BUMPer). I. General theory and model. J. Chem. Phys. **2021**, 154, 044104.

(206) Jin, J.; Han, Y.; Voth, G. A. Coarse-graining involving virtual sites: Centers of symmetry coarse-graining. J. Chem. Phys. **2019**, 150, 154103.

(207) Pak, A. J.; Dannenhoffer-Lafage, T.; Madsen, J. J.; Voth, G. A. Systematic Coarse-Grained Lipid Force Fields with Semiexplicit Solvation via Virtual Sites. J. Chem. Theory Comput. **2019**, 15, 2087–2100.

(208) Davtyan, A.; Dama, J. F.; Sinitskiy, A. V.; Voth, G. A. The Theory of Ultra-Coarse-Graining. 2. Numerical Implementation. J. Chem. Theory Comput. **2014**, 10, 5265–5275.

(209) Dama, J. F.; Jin, J.; Voth, G. A. The Theory of Ultra-Coarse-Graining. 3. Coarse-Grained Sites with Rapid Local Equilibrium of Internal States. J. Chem. Theory Comput. **2017**, 13, 1010–1022, PMID: 28112956.

(210) Dequidt, A.; Solano Canchaya, J. G. Bayesian parametrization of coarse-grain dissipative dynamics models. J. Chem. Phys. **2015**, 143, 084122.

(211) Solano Canchaya, J. G.; Dequidt, A.; Goujon, F.; Malfreyt, P. Development of DPD coarse-grained models: From bulk to interfacial properties. J. Chem. Phys. **2016**, 145, 054107.

(212) Kempfer, K.; Devémy, J.; Dequidt, A.; Couty, M.; Malfreyt, P. Development of Coarse-Grained Models for Polymers by Trajectory Matching. ACS Omega **2019**, 4, 5955–5967.

(213) Nguyen, H.; Huang, D. Systematic bottom-up molecular coarse- graining via force and torque matching using anisotropic particles. J. Chem. Phys. **2022**, 156, 184118.

(214) Villet, M. C.; Fredrickson, G. H. Numerical coarse-graining of fluid field theories. J. Chem. Phys. **2010**, 132, 034109.

(215) Sherck, N.; Shen, K.; Nguyen, M.; Yoo, B.; Köhler, S.; Speros, J. C.; Delaney, K. T.; Shell, M. S.; Fredrickson, G. H. Molecularly Informed Field Theories from Bottom-up Coarse-Graining. ACS Macro Lett. **2021**, 576–583.

(216) Das, A.; Andersen, H. C. The multiscale coarse-graining method. V. Isothermal-isobaric ensemble. J. Chem. Phys. **2010**, 132, 164106.

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

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

(219) Español, P.; Zúñiga, I. Obtaining fully dynamic coarse-grained models from MD. Phys. Chem. Chem. Phys. **2011**, 13, 10538–10545.

(220) Rühle, V.; Junghans, C.; Lukyanov, A.; Kremer, K.; Andrienko, D. Versatile object-oriented toolkit for coarse-graining applications. J. Chem. Theory Comput. **2009**, 5, 3211–3223.

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

(222) Wörner, S. J.; Bereau, T.; Kremer, K.; Rudzinski, J. F. Direct route to reproducing pair distribution functions with coarse-grained models via transformed atomistic cross correlations. J. Chem. Phys. **2019**, 151, 244110.

(223) Chen, Y.; Krämer, A.; Charron, N. E.; Husic, B. E.; Clementi, C.; Noé, F. Machine learning implicit solvation for molecular dynamics. The Journal of Chemical Physics **2021**, 155, 084101.

(224) Köhler, J.; Chen, Y.; Krämer, A.; Clementi, C.; Noé, F. Flow-Matching: Efficient

Coarse-Graining of Molecular Dynamics without Forces. Journal of Chemical Theory and Computation **2023**, 19, 942–952.

(225) Thaler, S.; Stupp, M.; Zavadlav, J. Deep coarse-grained potentials via relative entropy minimization. The Journal of Chemical Physics **2022**, 157, 244103.

(226) Potter, T. D.; Walker, M.; Wilson, M. R. Self-assembly and mesophase formation in a non-ionic chromonic liquid crystal: insights from bottom-up and top-down coarse-grained simulation models. Soft Matter **2020**, 16, 9488–9498.

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

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

(229) Rudzinski, J. F.; Noid, W. G. Bottom-up coarse-graining of peptide ensembles and helix-coil transitions. J. Chem. Theory Comput. **2015**, 11, 1278–1291.

(230) Chennakesavalu, S.; Toomer, D. J.; Rotskoff, G. M. Ensuring thermodynamic consistency with invertible coarse-graining. **2022**, arXiv:2210.07882 [cond-mat.stat-mech], doi:10.48550/ARXIV.2210.07882.

(231) Bezkorovaynaya, O.; Lukyanov, A.; Kremer, K.; Peter, C. Multiscale simulation of small peptides: Consistent conformational sampling in atomistic and coarse-grained models. J. Comp. Chem. **2012**, 33, 937–949.

(232) Harmandaris, V. A.; Reith, D.; Van der Vegt, N. F. A.; Kremer, K. Comparison between coarse-graining models for polymer systems: Two mapping schemes for polystyrene. Macromol. Chem. Phys. **2007**, 208, 2109–2120.

(233) Ohkuma, T.; Kremer, K. Comparison of two coarse-grained models of cis-polyisoprene with and without pressure correction. Polymer **2017**, 130, 88–101.

(234) Mullinax, J. W.; Noid, W. G. Extended ensemble approach for deriving transferable coarse-grained potentials. J. Chem. Phys. **2009**, 131, 104110.

(235) Jin, J.; Han, Y.; Voth, G. A. Ultra-Coarse-Grained Liquid State Models with Implicit Hydrogen Bonding. J. Chem. Theory Comput. **2018**, 14, 6159–6174.

(236) Chakraborty, M.; Xu, J.; White, A. D. Is preservation of symmetry necessary for coarse-graining? Phys. Chem. Chem. Phys. **2020**, 22, 14998–15005.

(237) Lyubartsev, A. P.; Laaksonen, A. Osmotic and activity coefficients from effective potentials for hydrated ions. Phys. Rev. E **1997**, 55, 5689–5696.

(238) Chen, Y.-L.; Habeck, M. Data-driven coarse graining of large biomolecular structures. PLOS ONE **2017**, 12, e0183057.

(239) Babadi, M.; Everaers, R.; Ejtehadi, M. R. Coarse-grained interaction potentials for anisotropic molecules. J. Chem. Phys. **2006**, 124, 174708.

(240) Morriss-Andrews, A.; Rottler, J.; Plotkin, S. S. A systematically coarse-grained model for DNA and its predictions for persistence length, stacking, twist, and chirality. J. Chem. Phys. **2010**, 132, 035105.

(241) Haxton, T. K.; Mannige, R. V.; Zuckermann, R. N.; Whitelam, S. Modeling Sequence-Specific Polymers Using Anisotropic Coarse-Grained Sites Allows Quantitative Comparison with Experiment. Journal of Chemical Theory and Computation **2015**, 11, 303–315.

(242) Li, G.; Shen, H.; Zhang, D.; Li, Y.; Wang, H. Coarse-Grained Modeling of Nucleic Acids Using Anisotropic Gay–Berne and Electric Multipole Potentials. Journal of Chemical Theory and Computation **2016**, 12, 676–693.

(243) Tripathy, M.; Agarwal, U.; Kumar, P. B. S. Toward Transferable Coarse-Grained Potentials for Poly-Aromatic Hydrocarbons: A Force Matching Approach. Macromolecular Theory and Simulations **2019**, 28, 1800040.

(244) Ricci, M.; Roscioni, O. M.; Querciagrossa, L.; Zannoni, C. MOLC. A reversible coarse grained approach using anisotropic beads for the modelling of organic functional materials. Phys. Chem. Chem. Phys. **2019**, 21, 26195–26211.

(245) Bellussi, F. M.; Roscioni, O. M.; Ricci, M.; Fasano, M. Anisotropic Electrostatic Interactions in Coarse-Grained Water Models to Enhance the Accuracy and Speed-Up Factor of Mesoscopic Simulations. J. Phys. Chem. B **2021**, 125, 12020–12027.

(246) Goujon, F.; Martzel, N.; Dequidt, A.; Latour, B.; Garruchet, S.; Devémy, J.; Blaak, R.; Munch, E.; Malfreyt, P. Backbone oriented anisotropic coarse grains for efficient simulations of polymers. The Journal of Chemical Physics **2020**, 153, 214901.

(247) Martzel, N.; Dequidt, A.; Devémy, J.; Blaak, R.; Garruchet, S.; Latour, B.; Goujon, F.; Munch, E.; Malfreyt, P. Grain Shape Dynamics for Molecular Simulations at the Mesoscale. Advanced Theory and Simulations **2020**, 2000124.

(248) Cohen, A. E.; Jackson, N. E.; de Pablo, J. J. Anisotropic Coarse-Grained Model for Conjugated Polymers: Investigations into Solution Morphologies. Macromolecules **2021**, 54, 3780–3789.

(249) Tanis, I.; Rousseau, B.; Soulard, L.; Lemarchand, C. A. Assessment of an anisotropic coarse-grained model for *cis* -1,4-polybutadiene: a bottom-up approach. Soft Matter **2021**, 17, 621–636.

(250) Friday, D. M.; Jackson, N. E. Modeling the Interplay of Conformational and Electronic Structure in Conjugated Polyelectrolytes. Macromolecules **2022**, 55, 1866–1877.

(251) Gay, J. G.; Berne, B. J. Modification of the overlap potential to mimic a linear site–site potential. J. Chem. Phys. **1981**, 74, 3316–3319.

(252) Stone, A. The description of bimolecular potentials, forces and torques: the S and V function expansions. Molecular Physics **1978**, 36, 241–256.

(253) Bowen, A. S.; Jackson, N. E.; Reid, D. R.; de Pablo, J. J. Structural Correlations and Percolation in Twisted Perylene Diimides Using a Simple Anisotropic Coarse-Grained Model. Journal of Chemical Theory and Computation **2018**, 14, 6495–6504.

(254) Molinero, V.; Moore, E. B. Water Modeled As an Intermediate Element between Carbon and Silicon. J. Phys. Chem. B **2009**, 113, 4008–4016.

(255) Stillinger, F. H.; Weber, T. A. Computer simulation of local order in condensed phases of silicon. Phys. Rev. B **1985**, 31, 5262–5271.

(256) Larini, L.; Lu, L. Y.; Voth, G. A. The multiscale coarse-graining method. VI. Implementation of three-body coarse-grained potentials. J. Chem. Phys. **2010**, 132, 164107.

(257) Scherer, C.; Andrienko, D. Understanding three-body contributions to coarse-grained force fields. Phys. Chem. Chem. Phys. **2018**, 20, 22387–22394.

(258) King, M.; Pasler, S.; Peter, C. Coarse-Grained Simulation of CaCO3 Aggregation and Crystallization Made Possible by Nonbonded Three-Body Interactions. J. Phys. Chem. C **2019**, 123, 3152–3160.

(259) Jin, J.; Pak, A. J.; Han, Y.; Voth, G. A. A new one-site coarse-grained model for water: Bottom-up many-body projected water (BUMPer). II. Temperature transferability and structural properties at low temperature. J. Chem. Phys. **2021**, 154, 044105.

(260) Wu, Z.; Beltran-Villegas, D. J.; Jayaraman, A. Development of a New Coarse-Grained Model to Simulate Assembly of Cellulose Chains Due to Hydrogen Bonding. J. Chem. Theory Comput. **2020**, 16, 4599–4614.

(261) Bianchi, E.; Largo, J.; Tartaglia, P.; Zaccarelli, E.; Sciortino, F. Phase Diagram of Patchy Colloids: Towards Empty Liquids. Phys. Rev. Lett. **2006**, 97, 168301.

(262) Doye, J. P. K.; Ouldridge, T. E.; Louis, A. A.; Romano, F.; Sulc, P.; Matek, C.; Snodin, B. E. K.; Rovigatti, L.; Schreck, J. S.; Harrison, R. M. et al. Coarse-graining DNA for simulations of DNA nanotechnology. Phys. Chem. Chem. Phys. **2013**, 15, 20395–20414.

(263) Ilie, I. M.; den Otter, W. K.; Briels, W. J. A coarse grained protein model with internal degrees of freedom. Application to alpha -synuclein aggregation. J. Chem. Phys. **2016**, 144, 085103.

(264) Sharp, M. E.; Vázquez, F. X.; Wagner, J. W.; Dannenhoffer-Lafage, T.; Voth, G. A. Multiconfigurational Coarse-Grained Molecular Dynamics. Journal of Chemical Theory and Computation **2019**, 15, 3306–3315.

(265) Jin, J.; Voth, G. A. Statistical Mechanical Design Principles for Coarse-Grained Interactions across Different Conformational Free Energy Surfaces. The Journal of Physical Chemistry Letters **2023**, 1354–1362.

(266) Jin, J.; Voth, G. A. Ultra-coarse-grained models allow for an accurate and transferable treatment of interfacial systems. J. Chem. Theory Comput. **2018**, 14, 2180–2197.

(267) Grime, J. M. A.; Dama, J. F.; Ganser-Pornillos, B. K.; Woodward, C. L.; Jensen, G. J.; Yeager, M.; Voth, G. A. Coarse-grained simulation reveals key features of HIV-1 capsid self-assembly. Nat. Commun. **2016**, 7, 11568.

(268) Bereau, T.; Rudzinski, J. F. Accurate Structure-Based Coarse Graining Leads to Consistent Barrier-Crossing Dynamics. Phys. Rev. Lett. **2018**, 121, 256002.

(269) Rudzinski, J. F.; Bereau, T. Coarse-grained conformational surface hopping: Methodology and transferability. J. Chem. Phys. **2020**, 153, 214110.

(270) Tully, J. C. Molecular dynamics with electronic transitions. J. Chem. Phys. **1990**, 93, 1061–1071.

(271) Rzepiela, A. J.; Louhivuori, M.; Peter, C.; Marrink, S. J. Hybrid simulations: combining atomistic and coarse-grained force fields using virtual sites. Physical Chemistry Chemical Physics **2011**, 13, 10437.

(272) Sahrmann, P. G.; Loose, T. D.; Durumeric, A. E. P.; Voth, G. A. Utilizing Machine Learning to Greatly Expand the Range and Accuracy of Bottom-Up Coarse-Grained Models Through Virtual Particles. 2022; http://arxiv.org/abs/2212.04530, arXiv:2212.04530 [physics].

(273) Lafond, P. G.; Izvekov, S. Multiscale Coarse-Graining of Polarizable Models through Force-Matched Dipole Fluctuations. J. Chem. Theory Comput. **2016**, 12, 5737–5750.

(274) Lafond, P. G.; Izvekov, S. Multiscale Coarse-Graining with Effective Polarizabilities: A Fully Bottom-Up Approach. J. Chem. Theory Comput. **2018**, 14, 1873–1886.

(275) Jumper, J. M.; Faruk, N. F.; Freed, K. F.; Sosnick, T. R. Accurate calculation of side chain packing and free energy with applications to protein molecular dynamics. PLOS Computational Biology **2018**, 14, e1006342.

(276) Liwo, A.; Czaplewski, C.; Pillardy, J.; Scheraga, H. A. Cumulant-based expressions for the multibody terms for the correlation between local and electrostatic interactions in the united-residue force field. J. Chem. Phys. **2001**, 115, 2323 – 2347.

(277) Sieradzan, A. K.; Makowski, M.; Augustynowicz, A.; Liwo, A. A general method for the derivation of the functional forms of the effective energy terms in coarse-grained energy functions of polymers. I. Backbone potentials of coarse-grained polypeptide chains. The Journal of Chemical Physics **2017**, 146, 124106.

(278) Lubecka, E. A.; Liwo, A. A general method for the derivation of the functional forms of the effective energy terms in coarse-grained energy functions of polymers. II. Backbone-local potentials of coarse-grained O1→4-bonded polyglucose chains. The Journal of Chemical Physics **2017**, 147, 115101.

(279) Liwo, A.; Sieradzan, A. K.; Lipska, A. G.; Czaplewski, C.; Joung, I.; Żmudzińska, W.; Hałabis, A.; Ołdziej, S. A general method for the derivation of the functional forms of the effective energy terms in coarse-grained energy functions of polymers. III. Determination of scale-consistent backbone-local and correlation potentials in the UNRES force field and force-field calibration and validation. The Journal of Chemical Physics **2019**, 150, 155104.

(280) Sherman, Z. M.; Howard, M. P.; Lindquist, B. A.; Jadrich, R. B.; Truskett, T. M. Inverse methods for design of soft materials. J. Chem. Phys. **2020**, 152, 140902.

(281) Gkeka, P.; Stoltz, G.; Barati Farimani, A.; Belkacemi, Z.; Ceriotti, M.; Chodera, J. D.; Dinner, A. R.; Ferguson, A. L.; Maillet, J.-B.; Minoux, H. et al. Machine Learning Force Fields and Coarse-Grained Variables in Molecular Dynamics: Application to Materials and Biological Systems. Journal of Chemical Theory and Computation **2020**, 16, 4757–4775.

(282) Campos Villalobos, G.; Giunta, G.; Marín-Aguilar, S.; Dijkstra, M. Machine-learning effective many-body potentials for anisotropic particles using orientation-dependent symmetry functions. The Journal of Chemical Physics **2022**, 157, 024902.

(283) Wagner, J. W.; Dannenhoffer-Lafage, T.; Jin, J.; Voth, G. A. Extending the range and physical accuracy of coarse-grained models: Order parameter dependent interactions. J. Chem. Phys. **2017**, 147, 044113.

(284) Friederich, P.; Häse, F.; Proppe, J.; Aspuru-Guzik, A. Machine-learned potentials for next-generation matter simulations. Nature Materials **2021**, 20, 750–761.

(285) John, S. T.; Csányi, G. Many-Body Coarse-Grained Interactions Using Gaussian Approximation Potentials. J. Phys. Chem. B **2017**, 121, 10934–10949.

(286) Glielmo, A.; Zeni, C.; De Vita, A. Efficient nonparametric n -body force fields from machine learning. Physical Review B **2018**, 97, 184307.

(287) Scherer, C.; Scheid, R.; Andrienko, D.; Bereau, T. Kernel-Based Machine Learning for Efficient Simulations of Molecular Liquids. Journal of Chemical Theory and Computation **2020**, 16, 3194–3204.

(288) Wang, J.; Chmiela, S.; Müller, K.-R.; Noé, F.; Clementi, C. Ensemble learning of coarse-grained molecular dynamics force fields with a kernel approach. The Journal of Chemical Physics **2020**, 152, 194106.

(289) Friedrichs, M. S.; Wolynes, P. G. Toward protein tertiary structure recognition by means of associative memory hamiltonians. Science **1989**, 246, 371–3.

(290) Papoian, G. A.; Ulander, J.; Eastwood, M. P.; Luthey-Schulten, Z.; Wolynes, P. G. Water in protein structure prediction. Proc. Natl. Acad. Sci. U.S.A. **2004**, 101, 3352–3357.

(291) Wu, H.; Wolynes, P. G.; Papoian, G. A. AWSEM-IDP: A Coarse-Grained Force Field for Intrinsically Disordered Proteins. The Journal of Physical Chemistry B **2018**, 122, 11115–11125.

(292) Zhang, L.; Han, J.; Wang, H.; Car, R.; E, W. DeePCG: Constructing coarse-grained models via deep neural networks. J. Chem. Phys. **2018**, 10.

(293) Zhang, L.; Han, J.; Wang, H.; Car, R.; E, W. Deep Potential Molecular Dynamics: A Scalable Model with the Accuracy of Quantum Mechanics. Physical Review Letters **2018**, 120, 143001.

(294) Ruza, J.; Wang, W.; Schwalbe-Koda, D.; Axelrod, S.; Harris, W. H.; Gómez-Bombarelli, R. Temperature-transferable coarse-graining of ionic liquids with dual graph convolutional neural networks. The Journal of Chemical Physics **2020**, 153, 164501.

(295) Fu, X.; Xie, T.; Rebello, N. J.; Olsen, B. D.; Jaakkola, T. Simulate Time-integrated Coarse-grained Molecular Dynamics with Geometric Machine Learning. **2022**, arXiv:2204.10348 [physics], doi:10.48550/ARXIV.2204.10348.

(296) Schütt, K. T.; Sauceda, H. E.; Kindermans, P.-J.; Tkatchenko, A.; Müller, K.-R. SchNet – A deep learning architecture for molecules and materials. The Journal of Chemical Physics **2018**, 148, 241722.

(297) Husic, B. E.; Charron, N. E.; Lemm, D.; Wang, J.; Pérez, A.; Krämer, A.; Chen, Y.; Olsson, S.; de Fabritiis, G.; Noé, F. et al. Coarse Graining Molecular Dynamics with Graph Neural Networks. J. Chem. Phys. **2020**, 153, 194101.

(298) Doerr, S.; Majewski, M.; Pérez, A.; Krämer, A.; Clementi, C.; Noe, F.; Giorgino, T.; De Fabritiis, G. TorchMD: A Deep Learning Framework for Molecular Simulations. Journal of Chemical Theory and Computation **2021**, 17, 2355–2363.

(299) Wang, J.; Charron, N.; Husic, B.; Olsson, S.; Noé, F.; Clementi, C. Multi-body effects in a coarse-grained protein force field. J. Chem. Phys. **2021**, 154, 164113.

(300) Papamakarios, G.; Nalisnick, E.; Rezende, D. J.; Mohamed, S.; Lakshminarayanan, B. Normalizing Flows for Probabilistic Modeling and Inference. Journal of Machine Learning Research **2021**, 22, 1–64.

(301) Kobyzev, I.; Prince, S. J.; Brubaker, M. A. Normalizing Flows: An Introduction and Review of Current Methods. IEEE Transactions on Pattern Analysis and Machine Intelligence **2021**, 43, 3964–3979.

(302) Reid, M. D.; Williamson, R. C. Information, Divergence and Risk for Binary Experiments. Journal of Machine Learning Research **2011**, 12, 731–817.

(303) Durumeric, A. E. P.; Voth, G. A. Explaining classifiers to understand coarse-grained models. 2021; arXiv:2109.07337 [physics], doi:10.48550/ARXIV.2109.07337.

(304) Ricci, E.; Giannakopoulos, G.; Karkaletsis, V.; Theodorou, D. N.; Vergadou, N. Developing Machine-Learned Potentials for Coarse-Grained Molecular Simulations: Challenges and Pitfalls. Proceedings of the 12th Hellenic Conference on Artificial Intelligence. Corfu Greece, 2022; pp 1–6.

(305) Hyvarinen, A. Estimation of Non-Normalized Statistical Models by Score Matching. Journal of Machine Learning Research **2005**, 6, 695–709.

(306) Song, Y.; Kingma, D. P. How to Train Your Energy-Based Models. **2021**, arXiv:2101.03288 [cs, stat], doi:10.48550/ARXIV.2101.03288.

(307) Köhler, J.; Chen, Y.; Krämer, A.; Clementi, C.; Noé, F. Flow-matching – efficient coarse-graining of molecular dynamics without forces. **2022**, doi: 10.48550/ARXIV.2203.11167.

(308) Arts, M.; Satorras, V. G.; Huang, C.-W.; Zuegner, D.; Federici, M.; Clementi, C.; Noé, F.; Pinsler, R.; Berg, R. v. d. Two for One: Diffusion Models and Force Fields for Coarse-Grained Molecular Dynamics. 2023; http://arxiv.org/abs/2302.00600, arXiv:2302.00600 [cs].

(309) Johnson, O. T. Information Theory And The Central Limit Theorem; World Scientific, 2004.

(310) Papoian, G. A.; Ulander, J.; Eastwood, M. P.; Luthey-Schulten, Z.; Wolynes, P. G. Water in protein structure prediction. Proceedings of the National Academy of Sciences **2004**, 101, 3352–3357.

(311) McQuarrie, D. A. Statistical Mechanics; University Science Books, 2000.

(312) Hu, C.; Lu, T.; Guo, H. Developing a Transferable Coarse-Grained Model for the Prediction of Thermodynamic, Structural, and Mechanical Properties of Polyimides at Different Thermodynamic State Points. J. Chem. Inf. Model. **2019**, 59, 2009–2025.

(313) Li, Y.; Agrawal, V.; Oswald, J. Systematic coarse-graining of semicrystalline polyethylene. J. Polym. Sci., Part B: Polym. Phys. **2019**, 57, 331–342.

(314) Qian, H.-J.; Carbone, P.; Chen, X.; Karimi-Varzaneh, H. A.; Liew, C. C.; Müller-Plathe, F. Temperature-Transferable Coarse-grained potentials for Ethylbenzene, Polystyrene and their mixtures. Macromolecules **2008**, 41, 9919–29.

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

(316) Farah, K.; Fogarty, A. C.; Böhm, M. C.; Müller-Plathe, F. Temperature dependence of coarse-grained potentials for liquid hexane. Phys. Chem. Chem. Phys. **2011**, 13, 2894–902.

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

(318) Lebold, K. M.; Noid, W. G. Systematic study of temperature and density variations in effective potentials for coarse-grained models of molecular liquids. J. Chem. Phys. **2019**, 150, 014104.

(319) Jin, J.; Pak, A. J.; Voth, G. A. Understanding entropy in coarse-grained systems: Addressing issues of representability and transferability. J. Phys. Chem. Lett. **2019**, 10, 4549–4557.

(320) Jin, J.; Yu, A.; Voth, G. A. Temperature and phase transferable bottom-up coarse-grained models. J. Chem. Theory Comput. **2020**, 16, 6823–6842.

(321) Rudzinski, J. F.; Lu, K.; Milner, S. T.; Maranas, J. K.; Noid, W. G. Extended ensemble approach to transferable potentials for low-resolution coarse-grained models of ionomers. J. Chem. Theory Comput. **2017**, 13, 2185–2201.

(322) Mukherjee, B.; Delle Site, L.; Kremer, K.; Peter, C. Derivation of Coarse Grained Models for Multiscale Simulation of Liquid Crystalline Phase Transitions. J. Phys. Chem. B **2012**, 116, 8474–8484.

(323) Szukalo, R. J.; Noid, W. Investigation of coarse-grained models across a glass transition. Soft Mater. **2020**, 18, 1–15.

(324) Chaimovich, A.; Shell, M. S. Anomalous waterlike behavior in spherically-symmetric water models optimized with the relative entropy. Phys. Chem. Chem. Phys. **2009**, 11, 1901–1915.

(325) Guenza, M. Thermodynamic consistency and other challenges in coarse-graining models. Eur. Phys. J.: Spec. Top. **2015**, 224, 2177–2191.

(326) Rosenberger, D.; van der Vegt, N. F. A. Addressing the temperature transferability of structure based coarse graining models. Phys. Chem. Chem. Phys. **2018**, 20, 6617–6628.

(327) Lu, J.; Qiu, Y.; Baron, R.; Molinero, V. Coarse-Graining of TIP4P/2005, TIP4P-Ew, SPC/E, and TIP3P to Monatomic Anisotropic Water Models Using Relative Entropy Minimization. J. Chem. Theory Comput. **2014**, 10, 4104–4120.

(328) Dannenhoffer-Lafage, T.; Wagner, J. W.; Durumeric, A. E. P.; Voth, G. A. Compatible observable decompositions for coarse-grained representations of real molecular systems. J. Chem. Phys. **2019**, 151, 134115.

(329) Louis, A. A. Beware of density dependent pair potentials. J. Phys.: Condens. Matter **2002**, 14, 9187–9206.

(330) Rowlinson, J. Intermolecular potentials that are functions of thermodynamic variables. Mol. Phys. **1984**, 52, 567–572.

(331) Stillinger, F. H.; Sakai, H.; Torquato, S. Statistical mechanical models with effective potentials: Definitions, applications, and thermodynamic consequences. J. Chem. Phys. **2002**, 117, 288–296.

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

(333) Pagonabarraga, I.; Frenkel, D. Dissipative particle dynamics for interacting systems. J. Chem. Phys. **2001**, 115, 5015–5026.

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

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

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

(337) Agrawal, V.; Peralta, P.; Li, Y.; Oswald, J. A pressure-transferable coarse-grained potential for modeling the shock Hugoniot of polyethylene. J. Chem. Phys. **2016**, 145, 104903.

(338) Shahidi, N.; Chazirakis, A.; Harmandaris, V.; Doxastakis, M. Coarse-graining of polyisoprene melts using inverse Monte Carlo and local density potentials. J. Chem. Phys. **2020**, 152, 124902.

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

(340) Rosenberger, D.; Sanyal, T.; Shell, M. S.; van der Vegt, N. F. A. Transferability of Local Density-Assisted Implicit Solvation Models for Homogeneous Fluid Mixtures. J. Chem. Theory Comput. **2019**, 15, 2881–2895.

(341) DeLyser, M. R.; Noid, W. G. Extending pressure-matching to inhomogeneous systems via local-density potentials. J. Chem. Phys. **2017**, 147, 134111.

(342) DeLyser, M. R.; Noid, W. Analysis of local density potentials. J. Chem. Phys. **2019**, 151, 224106.

(343) DeLyser, M.; Noid, W. Bottom-up coarse-grained models for external fields and interfaces. J. Chem. Phys. **2020**, 153, 224103.

(344) Sanyal, T.; Shell, M. S. Transferable Coarse-Grained Models of Liquid-Liquid Equilibrium Using Local Density Potentials Optimized with the Relative Entropy. J. Phys. Chem. B **2018**, 122, 5678–5693.

(345) Jochum, M.; Andrienko, D.; Kremer, K.; Peter, C. Structure-based coarse-graining in liquid slabs. J. Chem. Phys. **2012**, 137, 064102.

(346) Dalgicdir, C.; Sensoy, O.; Peter, C.; Sayar, M. A transferable coarse-grained model for diphenylalanine: How to represent an environment driven conformational transition. J. Chem. Phys. **2013**, 139.

(347) Campos-Villalobos, G.; Siperstein, F. R.; Patti, A. Transferable coarse-grained MARTINI model for methacrylate-based copolymers. Molecular Systems Design & Engineering **2019**, 4, 186–198.

(348) Avalos, J. B.; Mackie, A. D. Dissipative particle dynamics with energy conservation. Europhysics Letters (EPL) **1997**, 40, 141–146.

(349) Español, P. Fluid particle dynamics: A synthesis of dissipative particle dynamics and smoothed particle dynamics. Europhysics Letters (EPL) **1997**, 39, 605–610.

(350) Español, P.; Serrano, M.; Pagonabarraga, I.; Zúñiga, I. Energy-conserving coarse-graining of complex molecules. Soft Matter **2016**, 12, 4821–4837.

(351) Avalos, J. B.; Lísal, M.; Larentzos, J. P.; Mackie, A. D.; Brennan, J. K. Generalised dissipative particle dynamics with energy conservation: density- and temperature-dependent potentials. Physical Chemistry Chemical Physics **2019**, 21, 24891–24911.

(352) Lísal, M.; Larentzos, J. P.; Avalos, J. B.; Mackie, A. D.; Brennan, J. K. Generalized Energy-Conserving Dissipative Particle Dynamics with Reactions. Journal of Chemical Theory and Computation **2022**, 18, 2503–2512.

(353) DeLyser, M.; Noid, W. Coarse-grained models for local density gradients. J. Chem. Phys. **2022**, 156, 034106.

(354) Evans, R. Nature of the liquid-vapor interface and other topics in the statistical-mechanics of nonuniform, classical fluids. Adv. Phys. **1979**, 28, 143–200.

(355) Liwo, A.; Khalili, M.; Czaplewski, C.; Kalinowski, S.; Ołdziej, S.; Wachucik, K.; Scheraga, H. A. Modification and optimization of the united-residue (UNRES) potential energy function for canonical simulations. I. Temperature dependence of the effective energy function and tests of the optimization method with single training proteins. J. Phys. Chem. B **2007**, 111, 260–85.

(356) Xia, W.; Song, J.; Jeong, C.; Hsu, D. D.; Phelan, F. R.; Douglas, J. F.; Keten, S. Energy-Renormalization for Achieving Temperature Transferable Coarse-Graining of Polymer Dynamics. Macromolecules **2017**, 50, 8787–8796.

(357) Huang, H.; Wu, L.; Xiong, H.; Sun, H. A Transferrable Coarse-Grained Force Field for Simulations of Polyethers and Polyether Blends. Macromolecules **2019**, 52, 249–261.

(358) Szukalo, R. J.; Noid, W. G. Investigating the energetic and entropic components of effective potentials across a glass transition. J. Phys.: Condens. Matter **2021**, 33, 154004.

(359) Lebold, K. M.; Noid, W. G. Dual approach for effective potentials that accurately model structure and energetics. J. Chem. Phys. **2019**, 150, 234107.

(360) Tóth, G. Effective potentials from complex simulations: a potential-matching algorithm and remarks on coarse-grained potentials. J. Phys.: Condens. Matter **2007**, 19, 335222.

(361) Lebold, K. M.; Noid, W. G. Dual-potential approach for coarse-grained implicit solvent models with accurate, internally consistent energetics and predictive transferability. J. Chem. Phys. **2019**, 151, 164113.

(362) Deichmann, G.; Dallavalle, M.; Rosenberger, D.; van der Vegt, N. F. Phase Equilibria Modeling with Systematically Coarse-Grained Models—A Comparative Study on State Point Transferability. J. Phys. Chem. B **2019**, 123, 504–515.

(363) Moore, T. C.; Iacovella, C. R.; McCabe, C. Derivation of coarse-grained potentials via multistate iterative Boltzmann inversion. J. Chem. Phys. **2014**, 140.

(364) Moore, T. C.; Iacovella, C. R.; Hartkamp, R.; Bunge, A. L.; MCabe, C. A Coarse-Grained Model of Stratum Corneum Lipids: Free Fatty Acids and Ceramide NS. The Journal of Physical Chemistry B **2016**, 120, 9944–9958.

(365) Moore, T. C.; Iacovella, C. R.; Leonhard, A. C.; Bunge, A. L.; McCabe, C. Molecular dynamics simulations of stratum corneum lipid mixtures: A multiscale perspective. Biochemical and Biophysical Research Communications **2018**, 498, 313–318.

(366) Shamaprasad, P.; Moore, T. C.; Xia, D.; Iacovella, C. R.; Bunge, A. L.; McCabe, C. Multiscale Simulation of Ternary Stratum Corneum Lipid Mixtures: Effects of Cholesterol Composition. Langmuir **2022**, 38, 7496–7511.

(367) Wang, K. W.; Wang, Y.; Hall, C. K. Development of a coarse-grained lipid model, LIME 2.0, for DSPE using multistate iterative Boltzmann inversion and discontinuous molecular dynamics simulations. Fluid Phase Equilibria **2020**, 521, 112704.

(368) Taketomi, H.; Ueda, Y.; Go, N. Studies on protein folding, unfolding, and fluctuations by computer-simulation 1. Int. J. Pept. Protein Res. **1975**, 7, 445–459.

(369) Sanyal, T.; Mittal, J.; Shell, M. S. A hybrid, bottom-up, structurally accurate, G o^{-}-like coarse-grained protein model. J. Chem. Phys. **2019**, 151, 044111.

(370) Kanekal, K. H.; Rudzinski, J. F.; Bereau, T. Broad chemical transferability in structure-based coarse-graining. The Journal of Chemical Physics **2022**, 157, 104102.

(371) Shen, K.; Sherck, N.; Nguyen, M.; Yoo, B.; Köhler, S.; Speros, J.; Delaney, K. T.; Fredrickson, G. H.; Shell, M. S. Learning composition-transferable coarse-grained models: Designing external potential ensembles to maximize thermodynamic information. J. Chem. Phys. **2020**, 153, 154116.

(372) Nkepsu Mbitou, R. L.; Goujon, F.; Dequidt, A.; Latour, B.; Devémy, J.; Blaak, R.; Martzel, N.; Emeriau-Viard, C.; Tchoufag, J.; Garruchet, S. et al. Consistent and Transferable Force Fields for Statistical Copolymer Systems at the Mesoscale. Journal of Chemical Theory and Computation **2022**, 18, 6940–6951.

(373) Jin, J.; Pak, A. J.; Voth, G. A. Understanding Missing Entropy in Coarse-Grained Systems: Addressing Issues of Representability and Transferability. J. Phys. Chem. Lett. **2019**, 10, 4549–4557.

(374) Sivaraman, G.; Jackson, N. E. Coarse-Grained Density Functional Theory Predictions via Deep Kernel Learning. Journal of Chemical Theory and Computation **2022**, 18, 1129–1141.

(375) Maier, J. C.; Jackson, N. E. Bypassing backmapping: Coarse-grained electronic property distributions using heteroscedastic Gaussian processes. The Journal of Chemical Physics **2022**, 157, 174102.

(376) Strachan, A.; Holian, B. L. Energy Exchange between Mesoparticles and Their Internal Degrees of Freedom. Physical Review Letters **2005**, 94, 014301.

(377) Lin, K.-H.; Holian, B. L.; Germann, T. C.; Strachan, A. Mesodynamics with implicit degrees of freedom. J. Chem. Phys. **2014**, 141, 064107.

(378) Noé, F.; Tkatchenko, A.; Müller, K.-R.; Clementi, C. Machine Learning for Molecular Simulation. Annual Review of Physical Chemistry **2020**, 71, 361–390.

(379) Dill, K. A. Dominant forces in protein folding. Biochemistry **1990**, 29, 7133–7155.

(380) Chandler, D. Interfaces and the driving force of hydrophobic assembly. Nature **2005**, 437, 640–7.

(381) Savelyev, A.; Papoian, G. A. Molecular renormalization group coarse-graining of polymer chains: Applications to double-stranded DNA. Biophys. J. **2009**, 96, 4044–52.

(382) Netz, R.; Orland, H. Beyond Poisson-Boltzmann: Fluctuation effects and correlation functions. The European Physical Journal E **2000**, 1, 203.

(383) Castelnovo, M.; Joanny, J.-F. Complexation between oppositely charged polyelectrolytes: Beyond the Random Phase Approximation. The European Physical Journal E **2001**, 6, 377–386.

(384) Belle, V.; Papantonis, I. Principles and Practice of Explainable Machine Learning. Frontiers in Big Data **2021**, 4, 688969.

(385) Pitera, J. W.; Chodera, J. D. On the Use of Experimental Observations to Bias Simulated Ensembles. Journal of Chemical Theory and Computation **2012**, 8, 3445–3451.

(386) Amirkulova, D. B.; White, A. D. Recent advances in maximum entropy biasing techniques for molecular dynamics. Molecular Simulation **2019**, 45, 1285–1294.

(387) Latham, A. P.; Zhang, B. Maximum Entropy Optimized Force Field for Intrinsically Disordered Proteins. Journal of Chemical Theory and Computation **2020**, 16, 773–781.

(388) Chaimovich, A.; Shell, M. S. Relative entropy as a universal metric for multiscale errors. Phys. Rev. E **2010**, 81, 060104.

(389) McCarty, J.; Parrinello, M. A variational conformational dynamics approach to the selection of collective variables in metadynamics. The Journal of Chemical Physics **2017**, 147, 204109.

(390) Ribeiro, J. M. L.; Bravo, P.; Wang, Y.; Tiwary, P. Reweighted autoencoded variational Bayes for enhanced sampling (RAVE). The Journal of Chemical Physics **2018**, 149, 072301.

(391) Wu, Z.; Cui, Q.; Yethiraj, A. A New Coarse-Grained Force Field for Membrane–Peptide Simulations. J. Chem. Theory Comput. **2011**, 7, 3793–3802.

(392) Jarin, Z.; Newhouse, J.; Voth, G. A. Coarse-grained Force Fields from the Perspective

of Statistical Mechanics: Better Understanding the Origins of a MARTINI Hangover. J. Chem. Theory Comput. **2020**, 17, 1170–1180.

(393) Sanyal, T.; Shell, M. S. Transferable coarse-grained models of liquid-liquid equilibrium using local density potentials optimized with the relative entropy. J. Phys. Chem. B **2018**, 122, 5678–5693.

(394) Nasikas, D.; Ricci, E.; Giannakopoulos, G.; Karkaletsis, V.; Theodorou, D. N.; Vergadou, N. Investigation of Machine Learning-based Coarse-Grained Mapping Schemes for Organic Molecules. Proceedings of the 12th Hellenic Conference on Artificial Intelligence. Corfu Greece, 2022; pp 1–8.

(395) Zhai, Y.; Caruso, A.; Bore, S. L.; Luo, Z.; Paesani, F. A "short blanket" dilemma for a state-of-the-art neural network potential for water: Reproducing experimental properties or the physics of the underlying many-body interactions? The Journal of Chemical Physics **2023**, 158, 084111.

(396) Behler, J. Perspective: Machine learning potentials for atomistic simulations. The Journal of Chemical Physics **2016**, 145, 170901.

(397) Tadmor, E. B.; Elliott, R. S.; Phillpot, S. R.; Sinnott, S. B. NSF cyberinfrastructures: A new paradigm for advancing materials simulation. Current Opinion in Solid State and Materials Science **2013**, 17, 298–304.

(398) van der Giessen, E.; Schultz, P. A.; Bertin, N.; Bulatov, V. V.; Cai, W.; Csányi, G.; Foiles, S. M.; Geers, M. G. D.; González, C.; Hütter, M. et al. Roadmap on multiscale materials modeling. Modelling and Simulation in Materials Science and Engineering **2020**, 28, 043001.

(399) Cummings, P. T.; MCabe, C.; Iacovella, C. R.; Ledeczi, A.; Jankowski, E.; Jayaraman, A.; Palmer, J. C.; Maginn, E. J.; Glotzer, S. C.; Anderson, J. A. et al. Open-

source molecular modeling software in chemical engineering focusing on the Molecular Simulation Design Framework. AIChE Journal **2021**, 67.

(400) Wilkinson, M. D.; Dumontier, M.; Aalbersberg, I. J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L. B.; Bourne, P. E. et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data **2016**, 3, 160018.

(401) Scheffler, M.; Aeschlimann, M.; Albrecht, M.; Bereau, T.; Bungartz, H.-J.; Felser, C.; Greiner, M.; Groß, A.; Koch, C. T.; Kremer, K. et al. FAIR data enabling new horizons for materials research. Nature **2022**, 604, 635–642.

(402) Ingólfsson, H. I.; Melo, M. N.; van Eerden, F. J.; Arnarez, C.; Lopez, C. A.; Wassenaar, T. A.; Periole, X.; de Vries, A. H.; Tieleman, D. P.; Marrink, S. J. Lipid Organization of the Plasma Membrane. Journal of the American Chemical Society **2014**, 136, 14554–14559.

(403) Larsson, H. R.; Zhai, H.; Umrigar, C. J.; Chan, G. K.-L. The Chromium Dimer: Closing a Chapter of Quantum Chemistry. Journal of the American Chemical Society **2022**, 144, 15932–15937.

(404) Williams, K. T.; Yao, Y.; Li, J.; Chen, L.; Shi, H.; Motta, M.; Niu, C.; Ray, U.; Guo, S.; Anderson, R. J. et al. Direct Comparison of Many-Body Methods for Realistic Electronic Hamiltonians. Physical Review X **2020**, 10, 011041.

(405) Cohen, A. J.; Mori-Sánchez, P.; Yang, W. Challenges for Density Functional Theory. Chemical Reviews **2012**, 112, 289–320.

(406) LeBlanc, J.; Antipov, A. E.; Becca, F.; Bulik, I. W.; Chan, G. K.-L.; Chung, C.-M.; Deng, Y.; Ferrero, M.; Henderson, T. M.; Jiménez-Hoyos, C. A. et al. Solutions of the Two-Dimensional Hubbard Model: Benchmarks and Results from a Wide Range of Numerical Algorithms. Physical Review X **2015**, 5, 041041.

(407) Eriksen, J. J.; Anderson, T. A.; Deustua, J. E.; Ghanem, K.; Hait, D.; Hoffmann, M. R.; Lee, S.; Levine, D. S.; Magoulas, I.; Shen, J. et al. The Ground State Electronic Energy of Benzene. The Journal of Physical Chemistry Letters **2020**, 11, 8922–8929.

(408) Sayfutyarova, E. R.; Sun, Q.; Chan, G. K.-L.; Knizia, G. Automated Construction of Molecular Active Spaces from Atomic Valence Orbitals. J. Chem. Theory Comput. **2017**, 16.

(409) Sherrill, C. D.; Manolopoulos, D. E.; Martínez, T. J.; Michaelides, A. Electronic structure software. The Journal of Chemical Physics **2020**, 153, 070401.

(410) Fu, X.; Wu, Z.; Wang, W.; Xie, T.; Keten, S.; Gomez-Bombarelli, R.; Jaakkola, T. Forces are not Enough: Benchmark and Critical Evaluation for Machine Learning Force Fields with Molecular Simulations. 2022; http://arxiv.org/abs/2210.07237, arXiv:2210.07237 [physics].

(411) Mashayak, S. Y.; Jochum, M. N.; Koschke, K.; Aluru, N. R.; Rühle, V.; Junghans, C. Relative Entropy and Optimization-Driven Coarse-Graining Methods in VOTCA. PLOS ONE **2015**, 20.

(412) Lu, L. Y.; Izvekov, S.; Das, A.; Andersen, H. C.; Voth, G. A. Efficient, regularized, and scalable algorithms for multiscale coarse-graining. J. Chem. Theory Comput. **2010**, 6, 954–965.

(413) Mirzoev, A.; Lyubartsev, A. P. MagiC: software package for multiscale modeling. J. Chem. Theory Comput. **2013**, 9, 1512–1520.

(414) Mirzoev, A.; Nordenskiöld, L.; Lyubartsev, A. Magic v.3: An integrated software package for systematic structure-based coarse-graining. Computer Physics Communications **2019**, 237, 263–273.

(415) Dunn, N. J. H.; Lebold, K. M.; DeLyser, M. R.; Rudzinski, J. F.; Noid, W. G. BOCS: Bottom-up Open-source Coarse-graining Software. J. Phys. Chem. B **2018**, 122, 3363–3377.

(416) Shea, J.-E.; Best, R. B.; Mittal, J. Physics-based computational and theoretical approaches to intrinsically disordered proteins. Current Opinion in Structural Biology **2021**, 67, 219–225.

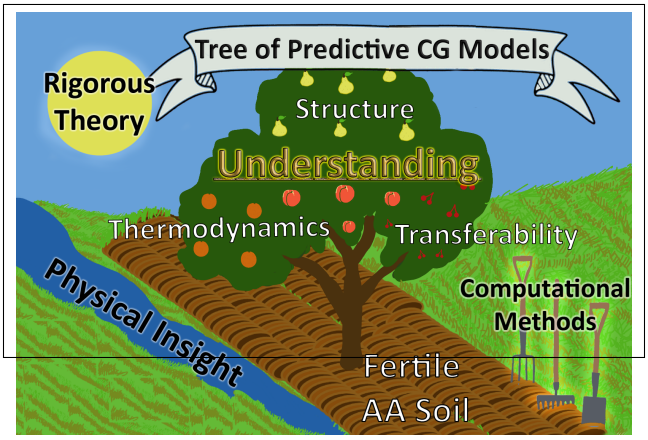
(417) Jackson, N. E. Coarse-Graining Organic Semiconductors: The Path to Multiscale Design. The Journal of Physical Chemistry B **2021**, 125, 485–496.

(418) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual Molecular Dynamics. J. Mol. Graph. **1996**, 14, 33–38.

Biography

W.G. Noid received a B.S. from the University of Tennessee (Knoxville) in 2000. He conducted graduate research with Roger Loring at Cornell University, where he also collaborated with Greg Ezra. He earned his PhD in 2005 for work on classical, semiclassical, and quantum mechanical theories for nonlinear vibrational spectroscopy. He began working on theories for coarse-grained models as a postdoctoral fellow with Greg Voth at the University of Utah (Salt Lake City), where he also collaborated with Hans Andersen. He joined the faculty at Penn State in 2007.

TOC Graphic


Tree of Predictive CG Models
Rigorous Theory
Structure
Understanding
Thermodynamics
Transferability
Computational Methods
Physical Insight
Fertile
AA Soil