Eur. Phys. J. B manuscript No.
(will be inserted by the editor)

Energetic and entropic considerations for coarse-graining

Katherine M. Kidder[1,†], **Ryan J. Szukalo**[1,†], **W. G. Noid**[a,1]

[1]Department of Chemistry, The Pennsylvania State University, University Park, Pennsylvania 16802, USA

June 25, 2022

Abstract Molecular dynamics simulations often adopt coarse-grained (CG) models in order to investigate length- and time-scales that cannot be effectively addressed with atomically detailed models. However, the effective potentials that govern CG models are configuration-dependent free energies with significant entropic contributions that have important consequences for the transferability and thermodynamic properties of CG models. This review summarizes recent work investigating the fundamental origin and practical ramifications of these entropic contributions, as well as their sensitivity to the CG mapping. We first analyze the energetic and entropic components of the many-body potential of mean force. By adopting a simple model for protein fluctuations, we examine how these components vary with the CG representation. We then introduce a "dual potential" approach for addressing these entropic considerations in more complex systems, such as ortho-terphenyl (OTP). We demonstrate that this dual approach not only accurately describes the structure and energetic properties of the underlying atomic model, but also accurately predicts the temperature-dependence of the CG potentials. Furthermore, by considering two different CG representations of OTP, we elucidate how these contributions vary with resolution. In sum, we hope this work will prove useful for improving the transferability and thermodynamic properties of CG models for soft materials.

1 Introduction

As detailed in this special issue, molecular dynamics (MD) simulations provide a uniquely powerful tool for investigating a vast range of phenomena. However, despite remarkable advances in computational methods and resources, many interesting phenomena occur on length- and time-scales that cannot be easily addressed with atomically detailed simulations. Consequently, many MD simulations represent systems with low resolution coarse-grained (CG) models that can provide 3-8 orders of magnitude greater computational efficiency.[1–3]

One can envision many approaches for developing CG models with varying goals and purposes. For instance, one can construct simple models to investigate the general consequences of certain basic physicochemical properties, such as the connectivity of polymers or the amphilicity of surfactants. [4–6] Alternatively, one can construct more detailed models to investigate phenomena in specific chemical systems. These more detailed models can be parameterized via "top-down" approaches that use lower resolution information, e.g., macroscopic thermodynamic observables, via "bottom-up" approaches that use higher resolution information, e.g., properties observed in atomically detailed simulations, or via mixed approaches that employ information from a combination of sources.[7] Several reviews provide a useful overview of the vast literature on coarse-graining methods and models.[8–15]

This manuscript seeks to equip readers with certain basic insights for constructing CG models. Specifically, this manuscript summarizes and synthesizes recent results by the present authors and their coworkers that clarify several fundamental aspects of CG models, including (1) the choice of the CG representation; (2) the energetic and entropic properties of CG models; (3) the transferability of effective potentials across different thermodynamic state points; and, most importantly, (4) the inter-relationship between these different aspects. Moreover, we attempt to highlight the ramifications of these fundamental considerations for developing

[a]e-mail:wgn1@psu.edu
[†]Authors contributed equally to this work

approximate CG models in practice. In order to provide a rather focused and didactic presentation, we attempt a balanced and illustrative, although not exhaustive, treatment of the relevant literature. For the same reasons, several interesting and important aspects of coarse-graining, including the dynamics of CG models,[16] are beyond the scope of this review.

We begin by first considering the basic theory of bottom-up coarse-graining.[8, 17–20] This theory allows us to define and analyze the exact potential for perfectly reproducing the structural and thermodynamic properties of an AA model at the resolution of the CG model. We then consider a simple microscopic model for which this exact coarse-graining procedure can be performed. This study illuminates the impact of the CG representation upon the information content and thermodynamic properties of the CG model. We then develop and demonstrate several computational methods for accurately modeling both structural and thermodynamic properties of more realistic AA models for which the exact coarse-graining procedure cannot be performed. Finally, we describe the application and results of these methods from ongoing efforts to develop an accurate and predictive CG model for an archetypical glass former, ortho-terphenyl (OTP). Thus, this manuscript seeks to provide both theoretical and practical insights into the energetic and entropic consequences of coarse-graining complex systems.

2 Exact coarse-graining in theory

This section presents an exact theoretical framework for bottom-up coarse-graining. We focus on the canonical ensemble for a system in a cubic volume $V = L^3$ at temperature T, although this framework can be readily extended to the NPT ensemble.[20] We refer to the high resolution model as an atomistic, or all-atom (AA), model and the low resolution model as a CG model. More generally, this framework can be applied to develop a low resolution model from any classical, particle-based model.

2.1 AA model

We first consider an AA model for a system of n atoms with a configuration $\mathbf{r} = (\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n)$ and a temperature-independent potential, $u(\mathbf{r})$. The force on each atom is $\mathbf{f}_i = -\partial u/\partial \mathbf{r}_i$ and the instantaneous excess pressure is,

$$p_{\rm xs}(\mathbf{r}) = -\left(\frac{\partial u}{\partial V}\right)_{\hat{\mathbf{r}}} = \frac{1}{3V}\sum_i \mathbf{f}_i \cdot \mathbf{r}_i - \left(\frac{\partial u}{\partial V}\right)_{\mathbf{r}}, \tag{1}$$

where $\hat{\mathbf{r}} = (L^{-1}\mathbf{r}_1, L^{-1}\mathbf{r}_2, \ldots, L^{-1}\mathbf{r}_n)$ defines the scaled AA coordinates.

The equilibrium configuration probability density is

$$p_{\rm r}(\mathbf{r}) = z_r^{-1} \exp\left[-\beta u(\mathbf{r})\right], \tag{2}$$

where $\beta = 1/k_B T$ and $z_r = \int_{V^n} d\mathbf{r} \exp\left[-\beta u(\mathbf{r})\right]$ is the AA configuration integral.[21] The excess configurational entropy

$$s_r = -k_B \int_{V^n} d\mathbf{r}\; p_{\rm r}(\mathbf{r}) \ln[V^n p_{\rm r}(\mathbf{r})] \tag{3}$$

quantifies the configuration information present in the equilibrium AA configuration distribution.

The excess Helmholtz potential, $a_r = -k_B T \ln[z_r/V^n]$ is a cumulant generating function such that

$$\left(\frac{\partial(-\beta a_r)}{\partial(-\beta)}\right) = \langle u(\mathbf{r})\rangle \equiv \overline{u}, \tag{4}$$

where the angular brackets denote an average according to $p_{\rm r}(\mathbf{r})$, and $\overline{u}$ denotes the mean potential energy. Similarly, the (excess) specific heat $c_v \equiv \partial\overline{u}/\partial T$ is given by

$$\left(\frac{\partial^2(-\beta a_r)}{\partial(-\beta)^2}\right) = k_B T^2 c_v = \sigma_u^2, \tag{5}$$

where

$$\sigma_u^2 = \int_{V^n} d\mathbf{r}\; p_{\rm r}(\mathbf{r})(u(\mathbf{r}) - \overline{u})^2 \tag{6}$$

is the variance in the AA potential fluctuations.

2.2 Mapped ensemble

We introduce a linear mapping operator, $\mathbf{M}$, to determine the CG configuration $\mathbf{R} = (\mathbf{R}_1, \mathbf{R}_2, \ldots, \mathbf{R}_N)$ as a function of the AA configuration: $\mathbf{R} = \mathbf{M}(\mathbf{r})$.[22] The coordinates of CG site I are determined by:

$$\mathbf{R}_I = \mathbf{M}_I(\mathbf{r}) = \sum_{i=1}^{n} c_{Ii}\mathbf{r}_i. \tag{7}$$

where $c_{Ii} \geq 0$ are constant coefficients. These coefficients must satisfy

$$\sum_{i=1}^{n} c_{Ii} = 1 \quad \forall\, I, \tag{8}$$

such that displacing each atom by $\delta\mathbf{r}$ results in displacing each CG site by $\delta\mathbf{r}$. We refer to an atom i as "involved" in CG site I if $c_{Ii} > 0$. For simplicity, we denote $\{i \in I\}$ as the set of atoms that are involved in site I. In many cases, the CG sites correspond to disjoint atomic groups, i.e., no atom is involved in more than one CG site. However, this is not a requirement. If one or more atoms are involved in multiple sites, then it is useful to introduce a second set of

mapping coefficients, $\{d_{Ii}\}$, describing atoms that are "specific" to a CG site.[22] These coefficients must satisfy three conditions: (1) $d_{Ii} \geq 0 \; \forall \; i$ and I; (2) $\sum_i d_{Ii} = 1 \; \forall \; I$; and (3) if atom i is involved in multiple sites, then $d_{Ii} = 0 \; \forall \; I$, i.e., $d_{Ii} > 0$ only if atom i is specific to site I. These constant coefficients then allow us to define a mapped AA force on each CG site,

$$\mathbf{f}_I(\mathbf{r}) = \sum_{i \in I} \frac{d_{Ii}}{c_{Ii}} \mathbf{f}_i(\mathbf{r}). \tag{9}$$

The d_{Ii} coefficients are not unique and different definitions of the instantaneous force may be adopted as long as they yield the same mean force.[22–25] In the following, we generally assume that the CG sites correspond to disjoint atomic groups. In this case, we can define $d_{Ii} = c_{Ii}$ for all i and I, such that Eq. (9) simplifies to $\mathbf{f}_I(\mathbf{r}) = \sum_{i \in I} \mathbf{f}_i(\mathbf{r})$.

The AA equilibrium distribution, p_r, and CG mapping operator, $\mathbf{M}$, define a "mapped ensemble" which is the target quantity for structure-based coarse-graining methods. In particular, the mapped probability distribution, $p_\mathrm{R}(\mathbf{R})$,

$$p_\mathrm{R}(\mathbf{R}) = \int_{V^n} \mathrm{d}\mathbf{r} \; p_\mathrm{r}(\mathbf{r}) \delta(\mathbf{M}(\mathbf{r}) - \mathbf{R}), \tag{10}$$

is the total probability density for the AA configurations that map to the configuration $\mathbf{R}$. The mapped distribution defines a new excess configurational entropy:

$$s_R = -k_B \int_{V^N} \mathrm{d}\mathbf{R} \; p_\mathrm{R}(\mathbf{R}) \ln[V^N p_\mathrm{R}(\mathbf{R})], \tag{11}$$

which quantifies the configurational entropy in the mapped AA ensemble. The mapping entropy,

$$S_\mathrm{map} = s_r - s_R \leq 0, \tag{12}$$

then quantifies the information that is lost when viewing the AA model at the resolution of the CG model.[26] Notice this mapping entropy is intrinsic to the CG mapping, $\mathbf{M}$, and does not reflect any approximations, e.g., in constructing a CG potential.

For any CG configuration $\mathbf{R}$, one can consider a "subensemble" of AA configurations that map to $\mathbf{R}$. This subensemble is described by the conditioned distribution

$$p_{\mathrm{r}|\mathrm{R}}(\mathbf{r}|\mathbf{R}) = p_\mathrm{r}(\mathbf{r}) \delta(\mathbf{M}(\mathbf{r}) - \mathbf{R}) \, / p_\mathrm{R}(\mathbf{R}), \tag{13}$$

which gives the conditioned probability density for the AA model to sample a configuration $\mathbf{r}$, given that $\mathbf{r}$ maps to $\mathbf{R}$. This conditioned distribution allows us to define the conditioned mean of the AA force, potential, and excess pressure as a function of $\mathbf{R}$,

$$\bar{\mathbf{f}}_I(\mathbf{R}) \equiv \langle \mathbf{f}_I(\mathbf{r}) \rangle_\mathbf{R} \tag{14}$$

$$\bar{u}(\mathbf{R}) \equiv \langle u(\mathbf{r}) \rangle_\mathbf{R} \tag{15}$$

$$\bar{p}_\mathrm{xs}(\mathbf{R}) \equiv \langle p_\mathrm{xs}(\mathbf{r}) \rangle_\mathbf{R}, \tag{16}$$

where the subscripted angular brackets denote a conditioned average according to $p_{\mathrm{r}|\mathrm{R}}(\mathbf{r}|\mathbf{R})$. Finally, we define a very important entropic quantity to describe the mapped subensemble:

$$S_W(\mathbf{R}) = -k_B \left\langle \ln \left[\frac{p_{\mathrm{r}|\mathrm{R}}(\mathbf{r}|\mathbf{R})}{q_{\mathrm{r}|\mathrm{R}}(\mathbf{r}|\mathbf{R})} \right] \right\rangle_\mathbf{R}, \tag{17}$$

where $q_{\mathrm{r}|\mathrm{R}}(\mathbf{r}|\mathbf{R}) = V^{N-n} \delta(\mathbf{M}(\mathbf{r}) - \mathbf{R})$ is the uniform conditioned probability distribution. Note that S_W is the negative of the Kullback-Leibler divergence between $p_{\mathrm{r}|\mathrm{R}}$ and $q_{\mathrm{r}|\mathrm{R}}$.[27] Consequently, $S_W \leq 0$ and only vanishes when $p_{\mathrm{r}|\mathrm{R}} = q_{\mathrm{r}|\mathrm{R}}$. More importantly, S_W, quantifies the information that is lost when we replace the AA conditioned distribution, $p_{\mathrm{r}|\mathrm{R}}$, with the uniform conditioned distribution, $q_{\mathrm{r}|\mathrm{R}}$, i.e., S_W exactly quantifies the information and configurational entropy that is lost when coarse-graining over the subensemble of AA configurations that map to $\mathbf{R}$.[28]

2.3 Potential of mean force

The central quantity in bottom-up coarse graining is the many-body potential of mean force (PMF), W,[8, 17–20, 22, 29]

$$\exp[-\beta W(\mathbf{R})] = V^{-(n-N)} \int_{V^n} \mathrm{d}\mathbf{r} \exp[-\beta u(\mathbf{r})] \delta(\mathbf{M}(\mathbf{r}) - \mathbf{R}), \tag{18}$$

which corresponds to the total Boltzmann weight for the subensemble of AA configurations that map to a single CG configuration $\mathbf{R}$. Note that the mapped distribution may be expressed

$$p_\mathrm{R}(\mathbf{R}) = Z_R^{-1} \exp\left[-W(\mathbf{R})/k_B T\right], \tag{19}$$

where $Z_R = V^{-(n-N)} z_r$. Consequently, a properly thermostatted MD simulation of a CG model that employs W as an interaction potential will perfectly reproduced the mapped configuration distribution that is implied by the AA model and CG mapping.[30] For this reason, W is often referred to as the "exact CG potential."

2.3.1 Excess free energy

The PMF not only reproduces the mapped AA configuration distribution, but also preserves the excess free energy of the AA model such that,

$$V^{-N} \int_{V^N} \mathrm{d}\mathbf{R} \exp[-\beta W(\mathbf{R})] = V^{-n} \int_{V^n} \mathrm{d}\mathbf{r} \exp[-\beta u(\mathbf{r})]. \tag{20}$$

Thus, W is an excess Helmholtz potential in which the CG coordinates are treated as additional mechanical variables.[20] The total differential of this potential is given by

$$\mathrm{d}W = -\sum_I \bar{\mathbf{f}}_I \cdot \mathrm{d}\mathbf{R}_I - \bar{p}_\mathrm{xs} \mathrm{d}V - S_W \mathrm{d}T. \tag{21}$$

Equation (21) has several important implications for modeling AA properties. First, the configuration-dependence of the PMF is determined by a conditioned mean of AA forces. It is for this reason and because it is a function of all N CG coordinates that W is commonly referred to as the many-body PMF. Similarly, the volume-dependence of the PMF is determined by the conditioned mean of the AA excess pressure in the subensemble of AA configurations that map to a given CG configuration. Moreover, the temperature-dependence of the PMF is determined by the entropy lost from this subensemble. Because Eq. (17) implies that $S_W \leq 0$, for every fixed configuration and volume, W necessarily increases with increasing temperature. Note that Maxwell relations relate variations in these conditioned averages,

$$\left(\frac{\partial \bar{\mathbf{f}}_I}{\partial T}\right) = \left(\frac{\partial S_W}{\partial \mathbf{R}_I}\right). \tag{22}$$

Most importantly, Eq. (21) implies that the excess pressure and entropy of the AA model can be determined at the resolution of the CG model from the volume and temperature-dependence of W.

2.3.2 Energetic and entropic contributions

Because W is a configuration-dependent excess Helmholtz potential, it follows that W can be decomposed into energetic and entropic contributions:[28]

$$W(\mathbf{R}) = U_W(\mathbf{R}) - TS_W(\mathbf{R}) \tag{23}$$

The energetic contribution, U_W, is the conditioned average of the atomistic potential, $u(\mathbf{r})$, over the subensemble of AA configurations that map to a fixed CG configuration, $\mathbf{R}$:

$$U_W(\mathbf{R}) \equiv \overline{u}(\mathbf{R}) = \langle u(\mathbf{r})\rangle_{\mathbf{R}}. \tag{24}$$

The entropic contribution, S_W is the Kullback-Leibler divergence between $p_{\mathrm{r}|\mathrm{R}}$ and $q_{\mathrm{r}|\mathrm{R}}$, which was defined earlier in Eq. (17)

$$S_W(\mathbf{R}) = -k_B \left\langle \ln\left[\frac{p_{\mathrm{r}|\mathrm{R}}(\mathbf{r}|\mathbf{R})}{q_{\mathrm{r}|\mathrm{R}}(\mathbf{r}|\mathbf{R})}\right]\right\rangle_{\mathbf{R}}.$$

Thus, $S_W(\mathbf{R})$ quantifies the information loss associated with coarse-graining over this subensemble of AA configurations. Note that W, U_W, and S_W also depend upon V and T, although for simplicity we do not explicitly indicate this dependence.

Since $S_W \leq 0$, it follows that $W \geq U_W$. Consequently, the PMF cannot be used to accurately estimate AA energetics. However, if U_W can be determined, then it can be used to evaluate AA energetics at the resolution of the CG model. In particular, because

$$\int_{V^N} \mathrm{d}\mathbf{R}\; p_{\mathrm{R}}(\mathbf{R}) U_W(\mathbf{R}) = \int_{V^n} \mathrm{d}\mathbf{r}\; p_{\mathrm{r}}(\mathbf{r}) u(\mathbf{r}) = \overline{u}, \tag{25}$$

the thermodynamic average of the AA potential can be determined from the CG model by first sampling configurations with the PMF according to Eq. (19) and by then evaluating U_W for each sampled configuration.

Similarly, S_W can be used to estimate entropic properties of the AA model at the resolution of the CG model. By sampling CG configurations with the PMF according to Eq. (19), one can determine the configurational entropy, s_R, in the mapped ensemble according to Eq. (11). In order to determine the excess entropy of the AA model, s_r, one must also determine the mapping entropy, S_{map}. However, since

$$\int_{V^N} \mathrm{d}\mathbf{R}\; p_{\mathrm{R}}(\mathbf{R}) S_W(\mathbf{R}) = S_{\mathrm{map}}, \tag{26}$$

this mapping entropy can be determined from the CG model by sampling configurations according to the PMF and then evaluating S_W for each sampled configuration.[28] Therefore, while W is the effective potential for sampling configurations, U_W and S_W may be considered "operators" for exactly evaluating energetic and entropic observables, respectively.[31]

2.3.3 Specific heat and energetic fluctuations

The (excess) thermodynamic specific heat $c_v = \partial\overline{u}/\partial T$ is another important thermodynamic property worth briefly considering. Since the specific heat is a thermodynamic observable, in principle it should be possible for the CG model to reproduce c_v. Equation (25) indicates that the exact CG model can reproduce the AA energy at any temperature and, by so doing, reproduce c_v according to $\partial\overline{u}/\partial T$. However, Eq. (5) demonstrates that c_v is also related to the variance, σ_u^2, in the AA potential fluctuations. Since the CG model cannot directly observe the instantaneous AA potential, it is not immediately obvious how the CG model can reproduce the variance in the AA potential fluctuations.[32] This suggests the possibility of a fundamental inconsistency in modeling energetics with CG models. In principle, the CG model can reproduce the temperature dependence of the mean AA potential but not the variance in the AA potential fluctuations.

Indeed, the CG resolution only allows one to directly observe fluctuations in U_W, i.e. fluctuations in the conditioned mean of the AA potential:

$$\Sigma_{U_W}^2 = \int_{V^N} \mathrm{d}\mathbf{R}\; p_{\mathrm{R}}(\mathbf{R}) \left|U_W(\mathbf{R}) - \overline{u}(\mathbf{r})\right|^2. \tag{27}$$

The CG model cannot directly observe the potential fluctuations in the subensemble of AA configurations that map to a given CG configuration:

$$\sigma_{u|\mathrm{R}}^2(\mathbf{R}) = \int_{V^n} \mathrm{d}\mathbf{r}\; p_{\mathrm{r}|\mathrm{R}}(\mathbf{r}|\mathbf{R}) \left|u(\mathbf{r}) - U_W(\mathbf{R})\right|^2 \tag{28}$$

Clearly, $\sigma^2_{u|\mathrm{R}}(\mathbf{R}) \geq 0$ and $\sigma^2_{u|\mathrm{R}}(\mathbf{R}) = 0$ only if $u(\mathbf{r})$ is a constant for all AA configurations that map to $\mathbf{R}$. Moreover, it can be shown that

$$\sigma^2_u = \Sigma^2_{U_W} + \int_{V^N} \mathrm{d}\mathbf{R}\; p_{\mathrm{R}}(\mathbf{R})\sigma^2_{u|\mathrm{R}}(\mathbf{R}), \tag{29}$$

i.e., the total variance in the AA potential fluctuations can be decomposed into the fluctuations in the conditioned mean potential, $\Sigma^2_{U_W}$, plus the average of the fluctuations within the subensemble of the AA configurations that map to each CG configuration, $\sigma^2_{u|\mathrm{R}}(\mathbf{R})$.[32] Since $\sigma^2_{u|\mathrm{R}} > 0$ for any non-trivial coarse-graining, it then follows that $\Sigma^2_{U_W} < \sigma^2_u$ and, thus, the CG representation of Eq. 6 does not hold:

$$k_B T^2 c_v \neq \Sigma^2_{U_W}. \tag{30}$$

This apparent inconsistency can be resolved by accounting for the temperature-dependence of the PMF. It is convenient to define a configuration-dependent CG specific heat,

$$C_W(\mathbf{R}) = (\partial U_W(\mathbf{R})/\partial T) = T\,(\partial S_W(\mathbf{R})/\partial T)\,. \tag{31}$$

Since W is a configuration-dependent Helmholtz potential, it also acts as a cumulant generating function:

$$\left(\frac{\partial^2(-\beta W(\mathbf{R}))}{\partial(-\beta)^2}\right) = k_B T^2 C_W(\mathbf{R}) = \sigma^2_{u|\mathrm{R}}(\mathbf{R}). \tag{32}$$

Thus, the temperature-dependence of W and, more precisely, C_W, directly quantifies the energetic fluctuations hidden by coarse-graining. Using Eqs. (29) and (32), we obtain a generalized fluctuation theorem for computing the AA specific heat in terms of energetic fluctuations that are observable at the CG resolution:

$$k_B T^2 c_v = \Sigma^2_{U_W} + \int_{V^N} \mathrm{d}\mathbf{R}\; p_{\mathrm{R}}(\mathbf{R}) C_W(\mathbf{R}). \tag{33}$$

Thus, in principle, CG models are capable of exactly and consistently reproducing the AA heat capacity, either from the temperature-dependence of the mean AA potential (i.e., according to Eq. (25)) or from the generalized fluctuation relation (i.e., according to Eq. (33)).

2.4 Insights from a study of exact coarse-graining

While the preceding described an exact coarse-graining procedure, the exact PMF cannot be determined for most systems of interest. However, it is possible to derive the exact PMF for certain simple models. In such simple cases, one can analyze the exact PMF in order to gain general, qualitative insight into the impact of resolution and the details of specific mappings upon the information content and thermodynamic properties of CG models. In the following, we illustrate the key findings from the study of Foley, Shell, and Noid,[28] as well as the more recent work by Foley, Kidder, Shell, and Noid.[33] Importantly, these studies directly assess the intrinsic qualities of CG mappings without introducing confounding effects due to approximating the many-body PMF.

2.4.1 High resolution GNM

These studies adopted the Gaussian Network Model (GNM) as a simple high resolution model for proteins.[34–36] In the GNM, each amino acid of a protein is represented as a bead located at its α carbon. The α carbons within a specified cutoff distance of each other are then connected by isotropic linear springs. Consequently, the GNM potential can be decomposed into independent components acting in each direction

$$u_{\mathrm{GNM}}(\mathbf{r}|\mathbf{r_0}) = \frac{1}{2}\Gamma\,\delta\mathbf{r}^\dagger\boldsymbol{\kappa}\delta\mathbf{r}, \tag{34}$$

where $\delta\mathbf{r} = \mathbf{r} - \mathbf{r}_0$, $\mathbf{r}$ is a vector that specifies the coordinates of the α carbons in one direction, e.g. $\mathbf{r} = (x_1, x_2, \ldots, x_n)$, $\mathbf{r}_0$ specifies the equilibrium coordinates in this same direction, Γ is a dimensional scaling factor, and $\boldsymbol{\kappa}$ is the Kirchhoff matrix. The entries of the Kirchhoff matrix are given by $\kappa_{ij} = \delta_{ij} n_i - \theta_{ij}$, where δ_{ij} is the Kronecker delta, n_i is the number of springs connected to atom i, and $\boldsymbol{\theta}$ is the contact matrix of the GNM, i.e., $\theta_{ij} = 1$ if there exists a spring between atoms i and j and $\theta_{ij} = 0$ otherwise. Note that $\boldsymbol{\theta}$ corresponds to the adjacency matrix of a "protein interaction network" in which each amino acid is represented by a vertex and an edge is drawn between contacting residues. $\boldsymbol{\kappa}$ then is the graph Laplacian of this protein interaction network. Also note that the GNM potential is invariant if each atom is displaced by δx. Consequently, $\mathbf{J}_n = (1, \ldots, 1)^\dagger$ is an eigenvector of $\boldsymbol{\kappa}$ with eigenvalue 0 that corresponds to uniformly translating all n atoms. Assuming that the protein is connected, $\mathbf{J}_n$ is the unique 0 eigenvector of the GNM. Therefore $\boldsymbol{\kappa}$ is a symmetric semi-positive definite matrix with a 1-dimensional nullspace. It is also convenient to introduce a vibrational projection operator, $\mathbf{Q}_n = \mathbb{1}_n - n^{-1}\mathbf{J}_n\mathbf{J}_n^\dagger$, where $\mathbb{1}_n$ is the $n \times n$ identity matrix. Then $\delta\mathbf{r}_{\mathrm{v}} = \mathbf{Q}_n\delta\mathbf{r}$ describes the atomic vibrational motion.

The configurational probability density of the high resolution GNM is given by a Gaussian distribution:

$$p_{\mathrm{r}}(\mathbf{r}) \propto \exp\left[-\frac{1}{2}\beta\Gamma\,\delta\mathbf{r}^\dagger\boldsymbol{\kappa}\delta\mathbf{r}\right]. \tag{35}$$

The excess configurational entropy of the high resolution GNM is

$$s_r/k_B = \frac{1}{2}(n-1)(1 + \ln[2\pi/(\beta\Gamma V^2)]) - \frac{1}{2}\ln t_{\boldsymbol{\kappa}}, \tag{36}$$

where $V = L$ is the "volume" in one-dimension and $t_{\boldsymbol{\kappa}} = \frac{1}{n}\det{}_1\boldsymbol{\kappa}$. Since $\boldsymbol{\kappa}$ is singular we denote the product of its

$n-1$ positive eigenvalues by $\det_1 \boldsymbol{\kappa}$. According to Kirchhoff's matrix-tree theorem, $t_{\boldsymbol{\kappa}}$ corresponds to the number of spanning trees in the protein interaction network defined by the GNM.[37] A spanning tree is a fully connected graph that contains all of the vertices in the protein interaction network, but edges have been removed such that there are no loops.[37] Since the first term in Eq. (36) depends trivially on n and is independent of the protein identity, we define the "non-trivial information" in the high resolution model as $h = \frac{1}{2}\ln t_{\boldsymbol{\kappa}}$.

The (vibrational) covariance matrix of the high resolution GNM is $\mathbf{c} = \langle \delta\mathbf{r}_{\mathrm{v}} \delta\mathbf{r}_{\mathrm{v}}^{\dagger} \rangle = (\beta\Gamma\boldsymbol{\kappa})^{\mathrm{I}}$ where the superscript I denotes the Moore-Penrose pseudo-inverse. We define the vibrational power of the GNM as the sum of the mass-weighted mean square displacements, $\sigma = m\mathrm{Tr}\mathbf{c} = \left\langle \sum_i m\delta r_{\mathrm{v};i}^2 \right\rangle$, where m is the atomic mass.

2.4.2 Exact CG model

We then determine a CG representation by mapping the coordinates of the high resolution GNM according to Eq. (7): $\mathbf{R} = \mathbf{M}(\mathbf{r})$. We require that the mapping coefficients are properly normalized according to Eq. (8) and linearly independent such that the rank of $\mathbf{M}$ is equal to the number of CG sites, N. Because a linear combination of Gaussian variables is also Gaussian,[38] it then follows that the mapped probability density for the CG coordinates is given by

$$p_{\mathrm{R}}(\mathbf{R}) \propto \exp\left[-\frac{1}{2}\beta\Gamma\delta\mathbf{R}^{\dagger}\mathbf{K}\delta\mathbf{R}\right], \tag{37}$$

where $\mathbf{K}^{\mathrm{I}} = \mathbf{Q}_N\mathbf{M}\boldsymbol{\kappa}^{\mathrm{I}}\mathbf{M}^{\dagger}\mathbf{Q}_N$ is the CG Kirchhoff matrix, $\mathbf{Q}_N = \mathbb{1}_N - N^{-1}\mathbf{J}_N\mathbf{J}_N^{\dagger}$ is the vibrational projector for CG displacements, $\mathbb{1}_N$ is the $N\times N$ identity matrix, and $\mathbf{J}_N = (1,\ldots,1)^{\dagger}$ spans the nullspace of $\mathbf{K}$.[28] Note that $\mathbf{K}$ is not a true Kirchhoff matrix in the sense that $K_{IJ} \neq \delta_{IJ}N_I - \theta_{IJ}$ because integrating over the coordinates introduces non-negative off diagonal elements.[39] In the following, we will require that the CG sites correspond to disjoint sets of n/N atoms and that the site coordinates are given by the mass centers of the corresponding atomic groups. In this case, $\mathbf{K}^{\mathrm{I}} = \mathbf{M}\boldsymbol{\kappa}^{\mathrm{I}}\mathbf{M}^{\dagger}$, which is readily interpreted as a condition ensuring that p_{R} has the proper covariance matrix.

The excess configurational entropy in the mapped ensemble is given by Eq. (11) and may be explicitly evaluated:

$$s_R/k_B = \frac{1}{2}(N-1)(1+\ln[2\pi/(\beta\Gamma V^2)]) - \frac{1}{2}\ln T_{\mathbf{K}}, \tag{38}$$

where $T_{\mathbf{K}} = \frac{1}{N}\det_1\mathbf{K}$ and again $\det_1\mathbf{K}$ corresponds to the product of the positive eigenvalues of $\mathbf{K}$. The first term of Eq. (38) is independent of protein identity and the CG mapping. Accordingly, we define the non-trivial information in the mapped ensemble as $H(\mathbf{M}) = \frac{1}{2}\ln T_{\mathbf{K}}$. The covariance matrix in the mapped ensemble is $\mathbf{C} = \left\langle \delta\widehat{\mathbf{R}}_{\mathrm{v}}(\mathbf{r})\delta\widehat{\mathbf{R}}_{\mathrm{v}}^{\dagger}(\mathbf{r}) \right\rangle = (\beta\Gamma\mathbf{K})^{I}$, where $\delta\widehat{\mathbf{R}}_{\mathrm{v}}(\mathbf{r}) = \mathbf{Q}_N\mathbf{M}\delta\mathbf{r}$. We define the vibrational power in the mapped ensemble by $\Sigma(\mathbf{M}) = M\mathrm{Tr}\mathbf{C} = \left\langle \sum_I M\delta R_{\mathrm{v};I}^2 \right\rangle$, where $M = (n/N)m$ is the CG mass. Importantly, both H and Σ depend on the CG mapping $\mathbf{M}$.

As discussed in Section 2.3.2, the PMF can be decomposed into an energetic and an entropic component, $W(\mathbf{R}) = U_W(\mathbf{R}) - TS_W(\mathbf{R})$. These components are particularly simple for the linear GNM. The energetic component is given by

$$U_W(\mathbf{R}) = \frac{1}{2}k_BT(n-N) + \frac{1}{2}\Gamma\delta\mathbf{R}^{\dagger}\mathbf{K}\delta\mathbf{R}. \tag{39}$$

The first term depends linearly upon temperature, is independent of protein identity or the details of the CG mapping, and simply corresponds to the equipartition result for the mean energy of the $(n-N)$ harmonic degrees of freedom that have been integrated out of the CG model. The second term is temperature-independent, but depends upon the protein identity and the details of the mapping. Interestingly, the mapped covariance of the GNM is completely determined by the energetic contribution to the PMF. The entropic component is given by

$$S_W(T)/k_B = \frac{1}{2}(n-N)(1+\ln[2\pi/(\beta\Gamma V^2)]) + \frac{1}{2}\ln(T_{\mathbf{K}}/t_{\boldsymbol{\kappa}}). \tag{40}$$

Again, the first term only reflects the excess entropy lost from the $n-N$ harmonic degrees of freedom that have been integrated out of the CG model. The second term in Eq. (40) reflects the Kirchoff matrix of the underlying protein and the CG mapping. Because the entropic component is configuration independent, $S_W = \langle S_W \rangle = S_{\mathrm{map}}$ and explicitly illustrates the general result from Eq. (12): $S_{\mathrm{map}} = s_r - s_R$. Equations (39) and (40) also illustrate the general result that $C_W(\mathbf{R}) = \partial U_W(\mathbf{R})/\partial T = T\partial S_W(\mathbf{R})/\partial T$. In the case of the GNM, C_W is simply $\frac{1}{2}(n-N)k_B$, which corresponds to the specific heat of the harmonic degrees of freedom that have been integrated out of the CG model.

2.4.3 Block maps

As a preliminary study, we first considered the impact of resolution upon the information content and thermodynamic properties of CG models for the GNM.[28] In this section, we present results for the PDBID 1UBQ (i.e., ubiquitin), which consists of 72 amino acids after the last four residues of the "floppy tail" have been trimmed. The inset of Fig. 1 presents a ribbon representation of the folded, equilibrium structure for 1UBQ. The original work of Foley et al. demonstrated very similar results for 7 other proteins with rather diverse structures.[28]

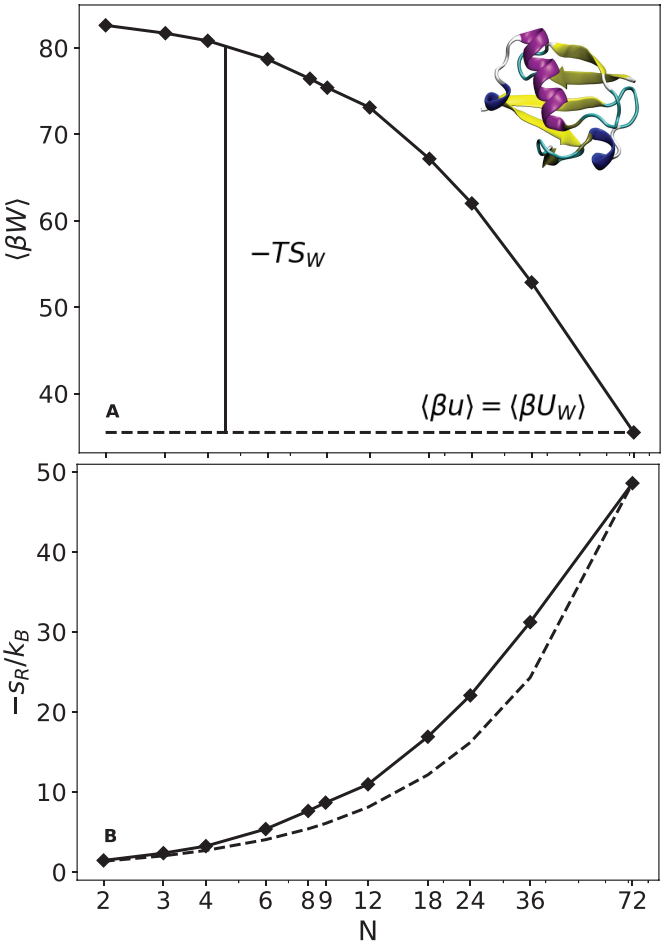

Fig. 1 The impact of resolution upon the PMF and mapped ensemble for block CG representations of 1UBQ. The ribbon structure is colored according to secondary structure assignments. Panel A presents $\langle \beta W \rangle$ as a function of the number of CG sites, N. Panel B presents the excess configurational entropy in the mapped ensemble, $-s_R$ (solid) along with a naïve scaling by N/n of the high resolution excess configurational entropy, $-s_r$ (dashed). Both panels employ a log scale for the x-axis. For these calculations we have employed $\beta \Gamma V^2 = 10$.

Specifically, this first study considered "block maps," which associated CG sites with equally sized, contiguous segments along the protein backbone. For example, the $N = 2$ block map for 1UBQ represents the first 36 amino acids with one site and represents amino acids 37-72 with a second site.

Figure 1 illustrates the impact of resolution upon CG models for 1UBQ. Panel A presents the average of the PMF, while panel B presents the excess configurational entropy retained in the mapped ensemble. At the highest resolution, $N = n = 72$ and the CG model is equivalent to the high resolution model. At this highest resolution, the PMF is simply the high resolution potential, which has an average of $\frac{1}{2} k_B T (n-1)$ according to the equipartition theorem.[40] Moreover, all of the excess configurational entropy is stored in the high resolution configuration distribution, i.e., $s_r = s_R$ and $S_{\text{map}} = 0$. As the CG resolution decreases, the information in the mapped configuration distribution systematically decreases. In order to preserve the excess free energy of the high resolution model, the PMF systematically increases to account for the lost configurational entropy. However, because it is a conditioned average of the high resolution potential, the average energetic contribution to the PMF does not vary with resolution. Consequently, the PMF becomes increasingly dominated by entropic contributions as the resolution decreases. Interestingly, the entropy loss varies less rapidly with resolution than might be naïvely expected. The dashed curve in panel B presents the naïve expectation, $-s_r/(n/N)$, which is always less than $-s_R$.

The impact of CG resolution upon configurational entropy has also been investigated for more realistic AA models.[41, 42] In particular, the recent work of Bernhardt and coworkers utilized the 2PT method[43] to estimate the rotational, translational, and vibrational contributions to the configurational entropy of AA models for several small molecules.[44] Importantly, the 2PT method allowed them to estimate the configurational entropy in the mapped ensemble of these molecules. They observed that the configurational entropy of the mapped ensemble systematically decreases with decreasing CG resolution, as expected. Additionally, they also applied the 2PT method to estimate the configurational entropy for approximate CG models that were parameterized via the conditional reversible work (CRW) method.[45] They observed that, in some cases, the translational entropy of these CRW models actually increased with decreasing resolution. This presumably results from the failure of the pair additive CRW potentials to accurately describe the many-body correlations present in the mapped ensemble. Indeed, early work by Lyubartsev and Laaksonen demonstrated that pair additive potentials have maximum entropy among all potentials that reproduce the same pair correlations.[46] Moreover, this artificially enhanced translational entropy of approximate CG models may be interpreted as a "smoothing" of the many-body PMF and likely contributes to the accelerated dynamics observed in CG models.[16, 47, 48]

2.4.4 Systematic study of mapping space

The previous section considered a single block map at each resolution. For proteins and other large molecules, though, there exist many different CG representations with the same resolution, i.e., the same number of CG sites, N. These different representations correspond to distinct groupings of the underlying atoms. This section illustrates some of the key findings from the recent study by Foley, Kidder, Shell, and Noid that more systematically investigated the space of CG representations.[33] We defined two metrics to quantitatively assess the intrinsic quality of a given CG mapping, $\mathbf{M}$. Specifically, we defined the information content, I, of the

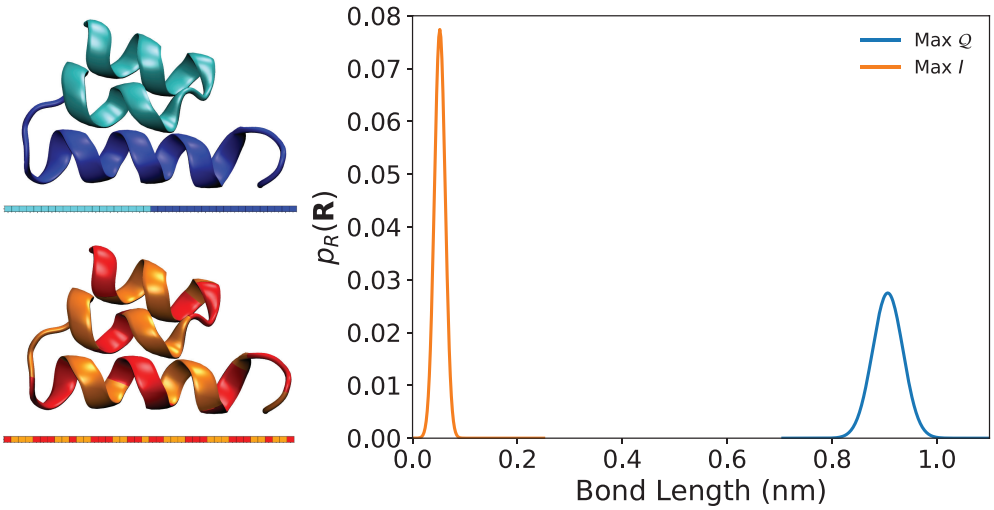

Fig. 2 Effect of the mapping on the mapped ensemble of 2ERL. Left: Illustration of the CG maps that maximize $\mathcal{Q}$ (top) and I (bottom). The ribbon structures are colored according to the assignment of α carbons to the two sites. The colored bar underneath the ribbon structure describes the sequence of the backbone, which has also been colored according to the CG site assignment of the residue. Right: The normalized probability distributions of the bond length, $x_1 - x_2$, in the mapped ensembles defined by two maps for 2ERL. The map with maximal spectral fitness, $\mathcal{Q}$, which is presented in blue, and the map with maximal information content, I, which is presented in orange, were identified via Monte Carlo sampling.[33]

mapping,

$$I(\mathbf{M}) = H(\mathbf{M})/h, \tag{41}$$

as the fraction of non-trivial information in the high resolution model that is preserved by the CG mapping. Because h and H reflect the protein- and the mapping-dependent contributions to s_r and s_R, it follows that maximizing I is equivalent to minimizing the (absolute) magnitude of the mapping entropy S_{map}. We also defined the spectral quality,

$$\mathcal{Q}(\mathbf{M}) = \Sigma(\mathbf{M})/\sigma, \tag{42}$$

as the fraction of the vibrational power in the high resolution model that is preserved by the CG mapping. Intuitively, I quantifies the information preserved by the CG model, while $\mathcal{Q}$ quantifies how well the mapping preserves the low frequency, large scale motions of the high resolution model. Note that I is determined by the eigenvalues of $\mathbf{K}$, while $\mathcal{Q}$ is determined by the eigenvalues of $\mathbf{K}^{\text{I}}$. Most importantly, these two metrics assess the intrinsic properties of the CG mapping and do not reflect any approximations, e.g., in constructing an effective potential.

In order to illustrate the impact of the CG mapping upon these two metrics, we consider two distinct 2-site maps for the protein 2ERL. These two maps maximized $\mathcal{Q}$ and I among the 2-site maps that we obtained from Monte Carlo sampling. [33] The left panel of Fig. 2 compares the atomic groupings of the 2 maps, while the right panel compares the probability distributions for the bond length, i.e., $x_1 - x_2$, in the corresponding mapped representations. Note that the mapping that maximizes $\mathcal{Q}$ decomposes the protein structure into two compact regions that are separated by a relatively large equilibrium distance, $d_0 = 0.906$ nm. Conversely, the mapping that maximizes I forms atomic groups that are distributed almost uniformly through the protein sequence such that the equilibrium distance between the two sites is only 0.052 nm. The bond distribution is much more localized for the most informative map and is much broader for the map with maximum spectral quality. Thus, although both CG representations employ only 2 sites, they result in very different mapped ensembles.

As noted above, our initial studies with the GNM have considered a restricted class of CG mappings in which each site corresponds to an equal number of connected atoms. Even given these restrictions, the number of possible maps is too large to exhaustively enumerate. Consequently, we employed Monte Carlo (MC) methods to statistically investigate the space of CG mappings.[49] We sampled each map, $\mathbf{M}$, according to the probability distribution $p(\mathbf{M}) = \exp[-\beta_{\mathcal{E}}\mathcal{E}(\mathbf{M})]$ with $\mathcal{E}(\mathbf{M}) = 1 - \mathcal{Q}(\mathbf{M})$ or $\mathcal{E}(\mathbf{M}) = 2H(\mathbf{M})$, where $\beta_{\mathcal{E}}$ is a corresponding fictitious inverse temperature. After performing MC simulations at a range of temperatures, we employed the Multistate Bennet Acceptance Ratio method to estimate a density of states quantifying the number of maps with a given quality.[50] While the following presents results for the small helical protein 2ERL, Ref. [33] provides similar results for several additional proteins.

We first consider the maps that maximize $\mathcal{Q}$ and I. While Fig. 2 presents the 2-site maps that maximize $\mathcal{Q}$ and I, Fig. 3 presents the corresponding optimal maps with $N = 4$ and 8 sites. Maps that maximize $\mathcal{Q}$ tend to associate CG sites with compact, highly connected atomic groups that correspond to structural motifs, such as helical segments or loops, in order to preserve low frequency, large scale motions. These maps are consistent with various approaches that seek to associate

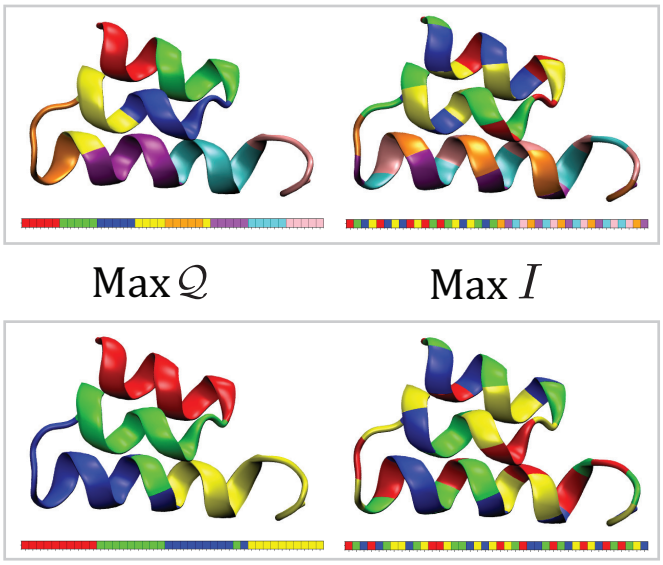

Fig. 3 Maps with maximal $\mathcal{Q}$ (left) and maximal I (right) among CG representations with $N = 4$ (bottom) and 8 (top) sites.

CG sites with rigid regions that move coherently.[51–54] In particular, our numerical studies indicate that $\mathcal{Q}$ is highly (anti-)correlated with the metric, $\chi^2_{\mathrm{EDCG}}(\mathbf{M})$, that is minimized in the essential dynamics coarse-graining method.[52, 53] Moreover, $\mathcal{Q}$ is closely related to the modularity and several other metrics that have been employed in spectral methods for clustering graphs.[55–58] Conversely, maps that maximize I tend to associate CG sites with loosely connected atomic groups that are often scattered throughout the protein structure in order to preserve the many and relatively informative high frequency protein motions.

We next consider statistical properties of the space of CG mappings. The top and left panels of Fig. 4 present the (log) density of states (DoS) $\ln\Omega(\mathcal{Q})$ and $\ln\Omega(I)$, respectively, which quantify the number of maps with a given spectral quality, $\mathcal{Q}$, and information content, I, for a given number of CG sites. Since $\ln\Omega$ is characterized by a single dominant maximum at each resolution, the mapping space for each resolution appears to be dominated by a large number of "typical" maps with a characteristic spectral quality and information content. These characteristic values systematically decrease with decreasing resolution. The DoS $\ln\Omega(I)$ remains similarly narrow at each resolution, indicating that the information content of these maps is primarily determined by the number of CG sites. In contrast, $\ln\Omega(\mathcal{Q})$ is considerably broader, indicating greater variation in the spectral quality of maps with a fixed N. Moreover, $\ln\Omega(\mathcal{Q})$ broadens with decreasing resolution and develops a "fat tail" of very rare maps with surprisingly high spectral quality. In particular, one observes rare 4-site maps that provide spectral quality that is comparable to 10-site maps of relatively low spectral quality. Indeed, similar overlap can also be observed in the $\ln\Omega(I)$ densities of states. Consequently, increasing resolution does not necessarily increase the quality of the CG representation.

The center panel in Fig. 4 presents an intensity plot of the natural logarithm of the joint 2-dimensional DoS, $\ln\Omega(\mathcal{Q},I)$, which quantifies the correlation between the spectral quality and information content of CG mappings. These 2D DoS are similarly sharply peaked with a single dominant maximum. At the highest resolution (i.e., N =20 such that each site corresponds to 2 amino acids), the spectral quality and information content appear almost statistically independent. However, at lower resolutions the two metrics appear quite anti-correlated.

Although this may initially appear surprising, this result can be directly traced to the definition of the two metrics and may be a rather general property of CG mappings. The information content, I, is defined above by the ratio of $H \propto \ln N^{-1} \det{}_1\, \mathbf{K}$ and $h \propto \ln n^{-1} \det{}_1\, \boldsymbol{\kappa}$, while the spectral fitness, $\mathcal{Q}$, is defined by the ratio of $\Sigma \propto \mathrm{Tr}\mathbf{M}\mathbf{K}^{\mathrm{I}}$ and $\sigma \propto \mathrm{Tr}\boldsymbol{\kappa}^{\mathrm{I}}$. Consequently, I emphasizes high frequency vibrations, while $\mathcal{Q}$ emphasizes low frequency vibrations. Moreover, the vibrational DoS of soft materials often include many more high frequency modes than low frequency modes. These relatively few low frequency modes are relatively uninformative but often describe physically important global motions. On the other hand, the high frequency modes are relatively informative but often describe unimportant localized fluctuations. This suggests that maps with maximal $\mathcal{Q}$ likely preserve the few uninformative low frequency modes at the expense of the many informative high frequency modes. Conversely, maps with maximal I likely preserve informative high frequency modes at the expense of the relatively few uninformative low frequency modes.

The recent study of Giulini and coworkers provides additional, useful insight into these considerations.[59] This study considered a fully anharmonic molecular mechanics model for proteins and focused on decimation mappings in which each site corresponded to a single atom. Giulini and coworkers determined mappings in order to minimize $|S_{\mathrm{map}}|$, which they approximated based upon configurations sampled from MD simulations.[59] Since $|S_{\mathrm{map}}|$ corresponds to the information lost by the CG representation, this approach is similar to maximizing the information content, I, of the CG mapping. Giulini and coworkers observed that CG mappings that preserved only backbone atoms were less informative than randomly sampled maps. Assuming that these backbone-based maps provide a relatively good description of low frequency fluctuations, this observation appears consistent with the anti-correlation between $\mathcal{Q}$ and I that is observed for the GNM. Moreover, Giulini and coworkers also observed that more informative maps identified biologically important amino acids. This suggests that the selection of a CG mapping should not only consider generic metrics, but

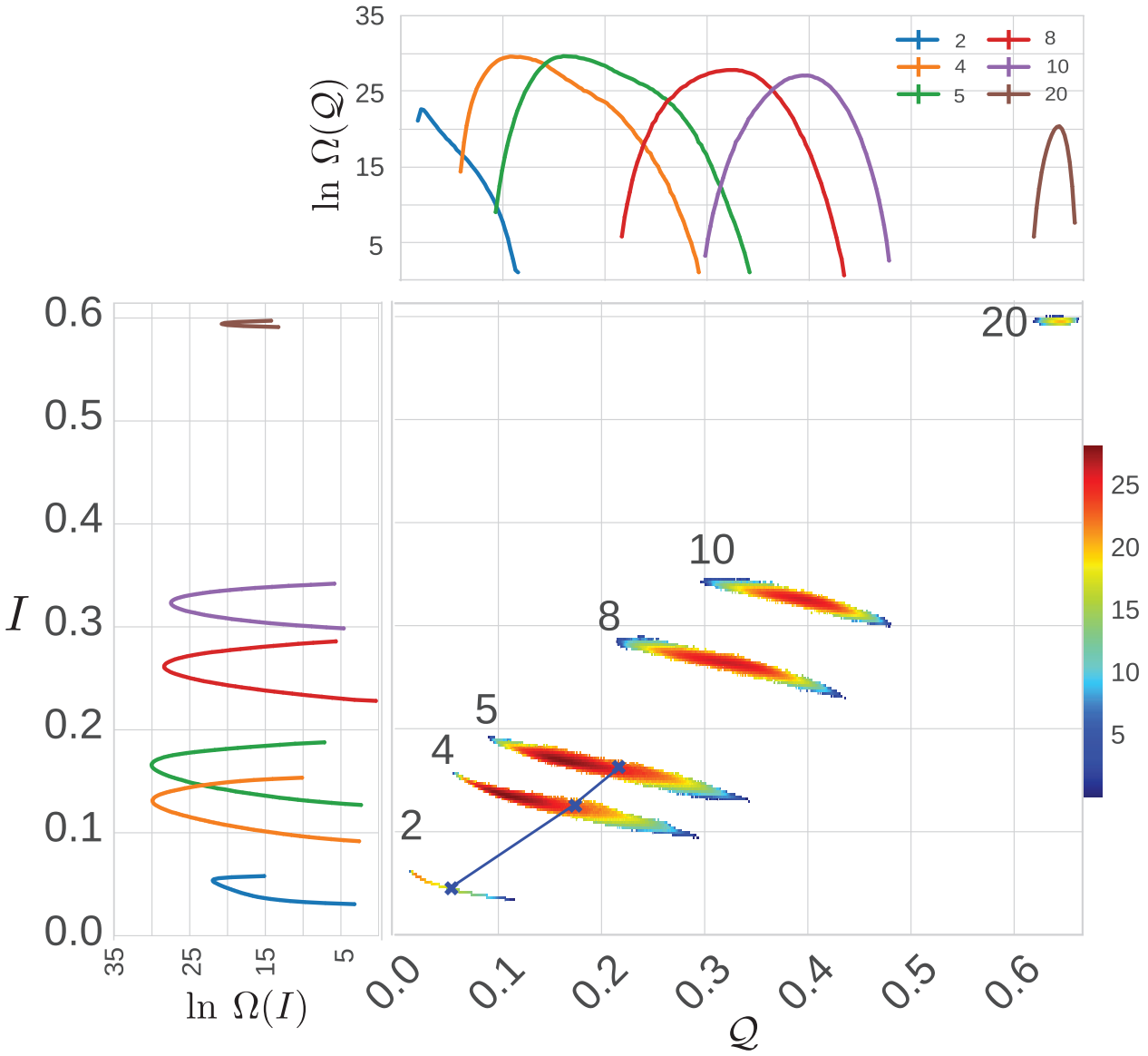

Fig. 4 Densities of state characterizing the statistics of CG mappings. The top panel presents the log density of states with respect to $\mathcal{Q}$, $\ln \Omega(\mathcal{Q})$. The left panel presents the log density of states with respect to I, $\ln \Omega(I)$. The legend indicates the number of sites, N, considered in each DoS. The middle panel presents the 2D log density of states with respect to $\mathcal{Q}$ and I, $\ln \Omega(\mathcal{Q}, I)$. These 2D DoSs are labelled by N. The blue x's indicate the separatrix for the apparent phase transitions that are observed at sufficiently low resolution.

also the biological, chemical, and physical properties that are relevant to the system and phenomena of interest.

One other aspect of the calculated DoS deserves further mention. At sufficiently low resolutions, $\ln \Omega(\mathcal{Q})$ exhibits regions of alternating curvature. Similar "back-bending" in the densities of states for mechanical models signals the existence of a phase transition in a finite system.[60] Therefore, the inflection points in $\ln \Omega(\mathcal{Q})$ suggest the possibility of a "phase transition" in the space of CG representations. In order to investigate this possibility, Fig. 5 presents both the mean and variance among sampled 4-site maps as a function of the fictitious temperature $T_{\mathcal{E}}$. As expected for a first order phase transition, the mean in $\mathcal{Q}$ and I both transition over a small temperature window in which the fluctuations in $\mathcal{Q}$ and I attain a maximum. Moreover, this transition is only observed at sufficiently low resolutions, which are indicated by the blue X's in Fig. 4. Consequently, these results suggest the rather intriguing possibility that there exists a "critical resolution" for viewing physical systems. Below this resolution a phase transition qualitatively distinguishes between "good" and "bad" maps, while above this critical point no such qualitative distinction exists.

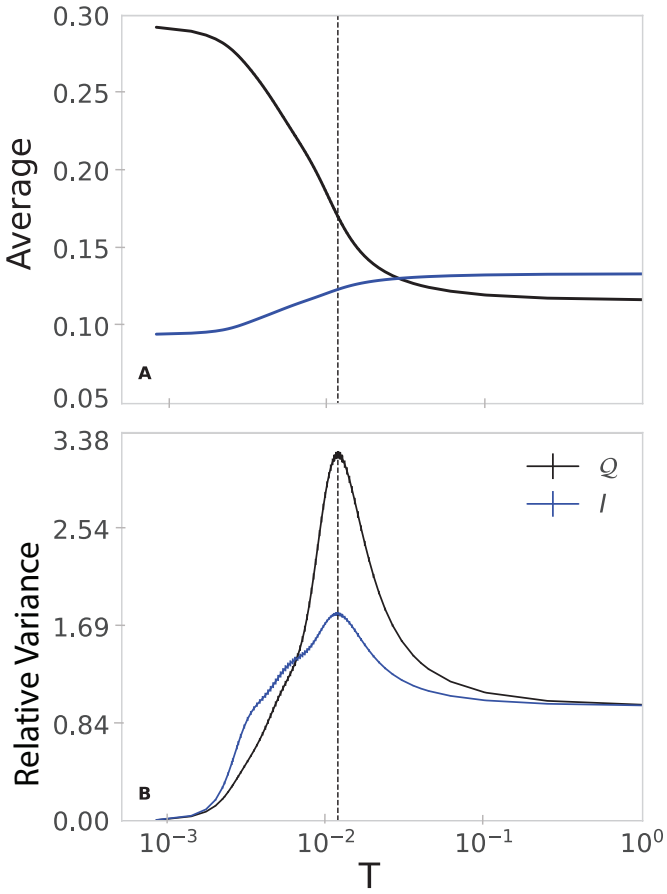

Fig. 5 Average (Panel A) and variance relative to the $T \to \infty$ limit (Panel B) of $\mathcal{Q}$ and I for 4-site representations of 2ERL as a function of the fictitious temperature $T = T_{\mathcal{E}}$. The vertical dashed line represents the identified transition temperature. Note that the x-axis is presented on a log scale.

2.4.5 Considerations for more realistic models

In closing this section, it is important to emphasize that the preceding results were obtained for the GNM, which is a particularly simple and analytically tractable microscopic model. Because the GNM corresponds to a Gaussian microscopic probability distribution, the mapped distribution is also Gaussian and depends only trivially upon the physical temperature, T. Consequently, it is important to briefly speculate upon the generality of the preceding results for more realistic microscopic models. In such cases, the microscopic configuration distribution may be much more complex and the mapped distribution may demonstrate nontrivial temperature-dependence.

To begin with, we anticipate that the qualitative trends of Fig. 1 will be robust because the underlying energy-entropy decomposition of the PMF is quite general. Specifically, because U_W is a conditioned mean of u, $\langle U_W \rangle = \langle u \rangle$ at all resolutions for any microscopic model. Conversely, the entropic contribution, $-TS_W$, to the PMF should always systematically increase as the model resolution decreases, although S_W may potentially demonstrate more complex temperature dependence for more realistic models. Similarly, we anticipate that the mapped ensemble will quite generally depend upon the choice of the CG mapping, as illustrated by Fig. 2. Moreover, because I and $\mathcal{Q}$ emphasize opposite ends of the vibrational spectrum, we suspect that one will often, though not necessarily always, find significant differences between representations with maximal information and maximal spectral fitness, as illustrated in Figs. 2 and 3.

Conversely, we expect that the physical characteristics of optimal maps, the statistics of mapping space, and the existence of a critical resolution, which were considered in Figs. 3 - 5, may be more sensitive to the specific microscopic model and its temperature-dependent configuration distribution. We expect that the qualitative insights obtained for the GNM will remain relatively robust as long as the underlying microscopic ensemble is characterized by fluctuations about a single dominant equilibrium structure. In this case, we expect that maps with maximal spectral fitness will be characterized by compact, densely connected structural subunits that are present in the equilibrium structure. Moreover, we expect that there may exist a critical resolution below which a phase transition distinguishes good and bad maps. However, the situation is likely more complex when the microscopic ensemble features two or more interconverting metastable conformations. Assuming that these conformational transitions are described by relatively few, low frequency modes, one expects that maps with maximal spectral fitness will associate CG sites with structural subunits that are present in multiple conformations and that move coherently during conformational transitions between the conformations. Alternatively, in this case, it may be more appropriate to associate different maps with different regions of configuration space, as suggested by Boninsegna and coworkers.[61] Finally, we consider the case that the microscopic potential energy surface includes many competing or degenerate minima, as in the case of a molecular liquid or polymer melt. In this case, we speculate that maps with high spectral quality will associate CG sites with the structural subunits that are preserved among the various minima.

We anticipate that the studies with the GNM have provided important insights into basic properties of CG representations. However, these first studies strongly motivate future studies that investigate the generality of their findings for more complex microscopic models. In such cases, the mapped distribution cannot be analytically characterized. These studies should develop accurate approximations for metrics, such as I and $\mathcal{Q}$, that quantitatively characterize the mapped ensemble itself, rather than the accuracy of the approximate potential employed to simulate the CG model. Given accurate approximations for such metrics, MC methods could be again applied to statistically characterize the space of representations. Such studies would clearly provide useful insight

for constructing particle-based CG models. Moreover, since the CG mapping is a linear example of more general, nonlinear order parameters, these studies may provide insight for developing general order parameters and non-linear dimensional reduction techniques.

3 Approximate coarse graining

While the preceding section employed an exact coarse-graining procedure to analyze general properties of CG models, this section considers the practical ramifications of this analysis for bottom-up models of complex systems. In practice, the first task in constructing a bottom-up CG model is determining a particular CG representation or mapping, $\mathbf{M}$. This mapping establishes the fundamental link between the AA and CG models and, in particular, defines the mapped ensemble. After defining the mapping, the next task is to define an effective potential, U, for modeling the interactions between the CG sites in canonical MD simulations, such that the resulting equilibrium CG distribution is $P_{\mathrm{R}}(\mathbf{R};U) \propto \exp[-\beta U(\mathbf{R})]$. Ideally U would equal W and the simulated distribution would perfectly reproduce the mapped distribution. Unfortunately, in most cases of interest, W is too complex to calculate or simulate. Instead, bottom-up approaches often determine U as some systematic approximation to W. In many cases, the approximate potential, U can be expressed in a simple molecular mechanics form[22]

$$U(\mathbf{R}) = \sum_{\zeta}\sum_{\lambda} U_{\zeta}(\psi_{\zeta\lambda}(\mathbf{R})). \tag{43}$$

Here, ζ indicates a particular type of interaction; U_{ζ} is the corresponding potential governing this interaction; $\psi_{\zeta\lambda}$ is a scalar variable that depends upon the coordinates, $\{\mathbf{R}\}_{\lambda}$; and λ identifies a set of CG sites. For example, if ζ corresponds to a non-bonded pair interaction, then U_{ζ} is a pair potential, λ identifies a particular pair of sites, and $\psi_{\zeta\lambda}(\mathbf{R})$ is the corresponding pair distance in the configuration $\mathbf{R}$. This potential commonly has terms that describe inter- and, possibly, intra-molecular pair non-bonded interactions, as well as intramolecular bonded, angle, and dihedral interactions. Note that the CG representation influences which terms should be included in the CG potential. For instance, Eq. (43) will only include intramolecular terms if the CG model represents each molecule with two or more sites.

3.1 Mapping

While the CG representation is most commonly determined by the researcher's physico-chemical intuition, a number of recent studies have investigated the importance of the CG mapping for the accuracy of CG models. Several studies have adopted graph theoretic[54, 62, 63] and machine learning[64, 65] methods to automate the selection of optimal CG mappings. In particular, Goméz-Bombarelli and coworkers have trained variational autoencoders to determine CG mappings that allow for accurate reconstruction of atomically detailed configurations, while employing a regularization based upon the instantaneous CG force in order to ensure that the CG potential is smooth.[64, 65] This approach may represent a promising compromise between maximizing the information content and spectral fitness of the CG mapping. Interestingly, the variational autoencoders often predict maps that are "intuitively obvious," i.e., highly consistent with chemical intuition. Consequently, White and coworkers have trained a neural network to predict good mappings based upon a library of CG mappings constructed by human experts[66].

However, other studies have demonstrated that the CG mapping can have more significant and more surprising influence upon the accuracy of bottom-up models. These studies have often focused on specific molecules and adopted a particular bottom-up approach to determine the approximate CG potential. For instance, Chakraborty et al. considered various small hydrocarbons and compared symmetric CG representations that preserve the symmetry of the underlying molecule with asymmetric representations that did not preserve this symmetry.[67] For each CG representation, they determined an approximate interaction potential via the multiscale coarse-graining (MS-CG) method,[22, 68–70] which will be discussed shortly. Chakraborty et al. observed that asymmetric CG models sometimes provided a more accurate description of the mapped radial distribution function (rdf) and velocity auto-correlation function, which they attributed to the additional potentials that were introduced to describe the additional site types that arise in the asymmetric model.[67] Similarly, Dallavelle and van der Vegt employed the CRW method to model various alkanes and observed that lower resolution models can reproduce certain properties better than higher resolution models.[71] Conversely, Khot and coworkers found that for several small molecules, the choice of mapping had relatively little impact upon the accuracy of the CG models.[72]

These findings likely reflect the impact of the CG mapping upon the complexity of the mapped AA ensemble and, consequently, the accuracy with which simple molecular mechanics potentials can accurately approximate the many-body PMF. Since the simple molecular mechanics potential in Eq. (43) severely limits the many-body correlations that can be reproduced by the approximate CG potential, CG mappings that simplify the mapped ensemble will likely improve the structural fidelity of the CG model. The work of Kremer and coworkers has clearly illuminated these considerations for modeling the conformations of polymers and peptides.[73–75] In particular, CG models almost invariably describe the bonded geometry of CG sites with additive bond, angle, and dihedral potentials. This approximation precludes the CG

model from reproducing any nontrivial cross-correlations between these degrees of freedom that may arise in the mapped ensemble. Accordingly, Kremer and coworkers designed CG mappings in order to minimize these cross-correlations. The cross-correlations between 1-4 distances and torsional degrees of freedom in C-alpha protein models can have similar effects for modeling the helix-coil transitions of peptides.[76] Rudzinski and Noid provided complementary insight into these considerations by partitioning the AA and mapped AA conformational spaces into discrete "internal states" associated with intramolecular degrees of freedom that are explicitly governed by AA and CG potentials.[77] In particular, poor maps result in CG models sampling "forbidden" conformational states that are excluded in the mapped ensemble by complex cross-correlations between CG degrees of freedom that cannot be reproduced by additive bond, angle, and dihedral potentials. Conversely, by avoiding such complex cross-correlations, good maps preserve a 1-to-1 relation between the internal states that are sampled by AA and approximate CG models.

These considerations are also important for the intermolecular structure of CG models. In general, the many-body PMF reflects complex many-body intermolecular correlations, which cannot always be reproduced by simple pair additive nonbonded potentials.[78] These considerations are particularly important for the MS-CG method, which determines approximate potentials based upon the assumption that the approximate potential can reproduce the corresponding cross-correlations in the mapped ensemble.[79] For instance, an early study by Izvekov and Voth demonstrated that 1-site MS-CG models of methanol provide a more accurate description of intermolecular structure than more detailed 2-site MS-CG models.[68] More recently, Voth and coworkers have demonstrated that simple pair additive MS-CG potentials accurately describe the mapped ensemble for 1-site representations of acetic acid, but poorly describe the mapped ensemble for higher resolution mappings that explicitly describe the carboxylic acid group.[80] Quite generally, one expects that the lower the density of CG sites in the mapped AA ensemble, the more accurately simple pair potentials will reproduce the mapped structure.[79] Moreover, although iterative structure-based approaches can accurately reproduce pair correlations of, e.g., water, they may do so at the expense of poorly describing three-body and higher order correlations.[81–85]

3.2 Approximating the PMF

One of the basic goals of bottom-up CG methods is to reproduce the structural properties of the mapped ensemble. According to Eq. (18), the CG model will reproduce the mapped distribution if the approximate potential, U, reproduces the configuration-dependence of the PMF, i.e., if $U(\mathbf{R}) = W(\mathbf{R}) + \text{const}$. In practice, Eq. (43) provides limited accuracy for reproducing W. Consequently, bottom-up approaches often focus on parameterizing approximate potentials that accurately reproduce the pair structural correlations in the mapped ensemble, i.e., mapped rdfs. Two variational principles unify a wide range of bottom-up approaches, clarify the connection to the PMF, and provide a procedure for systematically improving the structural fidelity of CG models.

Shell pioneered a relative entropy variational principle for determining an optimal approximate potential by minimizing the relative entropy,[86–88]

$$S_{\text{rel}}[U] = k_B \int_{V^N} \mathrm{d}\mathbf{R}\; p_{\text{R}}(\mathbf{R}) \ln\left[\frac{p_{\text{R}}(\mathbf{R})}{P_{\text{R}}(\mathbf{R};U)}\right], \tag{44}$$

which corresponds to the Kullback-Liebler divergence between the mapped distribution, $p_{\text{R}}(\mathbf{R})$, and the equilibrium distribution, $P_{\text{R}}(\mathbf{R};U)$, for a CG model with an approximate potential U.[27] According to the Gibbs inequality,[89] $S_{\text{rel}}[U] \geq 0$ and $S_{\text{rel}}[U] = 0$ only if $P_R(\mathbf{R};U) = p_{\text{R}}(\mathbf{R})$, i.e., when the CG model perfectly reproduces the mapped ensemble and $U(\mathbf{R}) = W(\mathbf{R}) + \text{const}$.[86, 87] In practice, one cannot determine this global minimum, but instead minimizes S_{rel} with respect to the various terms included in the approximate potential given by Eq. (43). In this case, S_{rel} is minimized with respect to U_ζ when the CG model reproduces the mean of the conjugate operator in the mapped AA ensemble.[8, 90] In particular, S_{rel} is minimized with respect to a pair potential when the CG model reproduces the corresponding mapped AA rdf. Thus, the relative entropy framework provides a variational basis for structure-based approaches such as Iterative Boltzmann Inversion[91] (IBI) and Inverse Monte Carlo[92] (IMC), in which the approximate potential U is iteratively refined to reproduce AA structural correlation functions.[8, 26, 87, 93] Similarly, the relative entropy framework can be extended to other ensembles to match various thermodynamic properties. For example, if one extends Eq. (44) to the constant NPT ensemble, then $S_{\text{rel}}[U]$ is minimized with respect to a volume potential, $U_V(V)$, when the CG model reproduces the AA pressure-volume equation-of-state.[94]

The relative entropy formalism can also provide insight into the CG mapping.[86] Specifically, one can define a probability $P_{\text{r}}(\mathbf{r};U)$ that a CG model with the potential U determines an AA configuration, $\mathbf{r}$. This requires defining a conditional probability for determining $\mathbf{r}$ from a sampled CG configuration, $\mathbf{R}$. If one defines this conditional probability in proportion to the AA equilibrium probability distribution, $p_{\text{r}}(\mathbf{r})$, then

$$P_{\text{r}}(\mathbf{r};U) = \frac{p_{\text{r}}(\mathbf{r})}{p_{\text{R}}(\mathbf{M}(\mathbf{r}))} P_R(\mathbf{M}(\mathbf{r});U). \tag{45}$$

Combining Eqs. 44 and 45, one obtains

$$S_{\text{rel}}[U] = k_B \int_{V^n} \mathrm{d}\mathbf{r}\; p_{\text{r}}(\mathbf{r}) \ln\left[V^{n-N} \frac{p_{\text{r}}(\mathbf{r})}{P_R(\mathbf{M}(\mathbf{r});U)}\right] + S_{\text{map}},$$

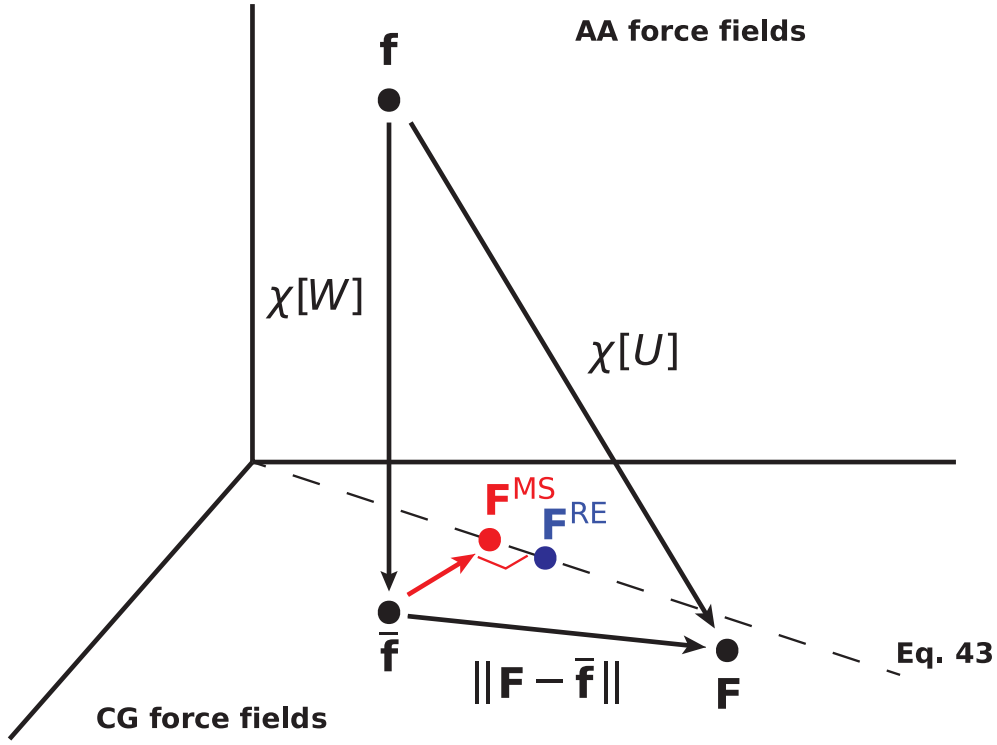

Fig. 6 Geometric interpretation of the MS-CG variational principle. The AA force field, $\mathbf{f}$, corresponds to a point in a vector space of force fields. Within this vector space lies a vector subspace of CG force fields, $\mathbf{F}$, that depend upon only upon CG coordinates. The many body mean force field, $\bar{\mathbf{f}}$, is one particular element in the space of CG force fields. The dashed line indicates a smaller vector subspace of force fields that correspond to Eq. (43), including the force fields determined by the relative entropy variational principle, $\mathbf{F}^{\mathrm{RE}}$, and the MS-CG variational principle, $\mathbf{F}^{\mathrm{MS}}$. The force-matching functional defines a "distance," $\chi[U]$, between the AA force field, $\mathbf{f}$, and the force field, $\mathbf{F}$, specified by an arbitrary CG potential, U. According to this definition, the mean force field, $\bar{\mathbf{f}}$, is closest to $\mathbf{f}$ and can be considered the geometric projection of $\mathbf{f}$ into the space of CG force fields. The corresponding distance, $\chi[W]$, corresponds to the intrinsic noise in the AA force field about the mean forces. Moreover, according to Eq. (48), $\chi^2[U]$, can be decomposed into the noise in the AA force field, $\chi^2[W]$, and the error in the CG force field, $||\mathbf{F}-\bar{\mathbf{f}}||^2$, according to the Pythagorean theorem. Finally, note that, by definition, $\mathbf{F}^{\mathrm{FM}}$ is the closest CG force field to the exact mean force field. However, it is likely that in many cases $||\mathbf{F}^{\mathrm{MS}}-\bar{\mathbf{f}}|| \approx ||\mathbf{F}^{\mathrm{RE}}-\bar{\mathbf{f}}|| \ll \chi[W]$.

(46)

where $S_{\mathrm{map}} = s_r - s_R$ quantifies the intrinsic information lost due to the CG mapping and is independent of the approximate potential, U.[26] Accordingly, one may employ S_{rel} to optimize both the approximate potential and the CG mapping.[59, 86]

Independently, Voth and coworkers pioneered a multiscale coarse-graining (MS-CG) variational principle[22, 68–70] that is based upon minimizing a force-matching functional[95–97]:

$$\chi^2[U] = \frac{1}{3N}\left\langle \sum_I |\mathbf{F}_I(\mathbf{M}(\mathbf{r})) - \mathbf{f}_I(\mathbf{r})|^2 \right\rangle \tag{47}$$

$$= \chi^2[W] + \frac{1}{3N}\int_{V^N} \mathrm{d}\mathbf{R}\; p_{\mathrm{R}}(\mathbf{R}) \sum_I \left|\, \mathbf{F}_I(\mathbf{R}) - \bar{\mathbf{f}}_I(\mathbf{R}) \,\right|^2 \tag{48}$$

where $\mathbf{F}_I(\mathbf{R}) = -\partial U(\mathbf{R})/\partial \mathbf{R}_I$ and $\bar{\mathbf{f}}_I$ is the AA mean force defined in Eq. (14). Because $\chi^2[W]$ is independent of U, the global minimum of χ^2 corresponds to the condition that the CG forces reproduce the mean AA forces, i.e., when $U(\mathbf{R}) = W(\mathbf{R}) + \mathrm{const}$, which implies that the resulting model will perfectly reproduce the mapped ensemble. Importantly, although χ^2 employs AA forces, the objective of force-matching is not to reproduce these forces. Instead, the objective is to reproduce the conditioned mean forces, $\bar{\mathbf{f}}_I$, which directly quantify the configuration-dependence of the many-body PMF. The AA forces provide a direct, albeit noisy, estimator for the mean forces.[22, 64]

Figure 6 provides an intuitive, geometric interpretation of the MS-CG variational principle. $\chi[U]$ can be interpreted as a "distance", $||\mathbf{F}-\mathbf{f}||$, between the AA force field, $\mathbf{f}$, and the CG force field, $\mathbf{F}$, that is determined by the approximate potential U. The conditioned mean force, $\bar{\mathbf{f}}$, can be interpreted as the geometric projection of the AA force field into the space of CG force fields,[79, 97] which is schematically depicted by a 2-dimensional plane in Fig. 6. Thus, Eq. (48) can be interpreted as a statement of the Pythagorean theorem in the space of force fields:

$$||\mathbf{f}-\mathbf{F}||^2 = \left|\left|\mathbf{f}-\bar{\mathbf{f}}\right|\right|^2 + \left|\left|\bar{\mathbf{f}}-\mathbf{F}\right|\right|^2. \tag{49}$$

The first term in Eq. (48) or (49) quantifies the fluctuations in the AA forces about their conditioned mean and may

be interpreted as the “noise" in the AA force field when viewed at a given CG resolution.[64] This noise is somewhat analogous to S_{map} within the relative entropy formalism, which quantifies the loss of configurational information when viewing the AA distribution at the CG resolution.[26, 86] Both $\chi^2[W]$ and S_{map} are direct consequences of the CG mapping and are independent of the CG potential. The second term in Eq. (48), which can be interpreted as the “error” in the CG potential, quantifies the difference in the configuration-dependence of the approximate potential and the exact many-body PMF.[64] Recent numerical calculations have demonstrated that the force-matching functional varies slightly with resolution and has similar values for rather different potentials, suggesting that the large noise due to the fluctuations in the AA force field may often dominate $\chi^2[U]$.[72, 84]

Although they appear rather distinct, the relative entropy and MS-CG variational principles are in fact remarkably similar.[26, 87] Most importantly, the many-body PMF is the global minimizer of both S_{rel} and χ^2.[86, 98] When the approximate potential depends linearly upon its parameters, both result in a convex optimization problem with a unique solution.[26, 87] Furthermore, both can be readily extended to other ensembles.[20, 99, 100] Moreover, although it is perhaps not immediately obvious, both can be related to the information quantity $\ln[p_{\text{R}}(\mathbf{R})/P_{\text{R}}(\mathbf{R};U)]$ and both can be expressed in terms of structural correlations.[8, 26] Consequently, the two variational principles can be employed to determine CG potentials that accurately approximate the exact many-body PMF.

In particular, the minimizing condition for χ^2 can be expressed entirely in terms of structural information and corresponds to a generalization of the Yvon-Born-Green (YBG) integral equation.[98, 101, 102] However, the g-YBG equation that minimizes χ^2 is not self-consistent.[79] Therefore, although the MS-CG potential can be directly calculated from structural data, it is not guaranteed to reproduce any particular structural observable.[79, 98] In contrast, the minimization condition for S_{rel} is a self-consistent equation that requires the CG model to reproduce certain AA structural observables, but generally requires multiple CG simulations to determine the potential.[26, 86, 103, 104]

The Noid group typically employs the MS-CG variational principle to determine the configuration-dependence of the CG potential, U. We often invoke the relative entropy variational principle to determine a volume potential, $U_V(V)$, that quantitatively reproduces the AA pressure-volume equation of state.[20] As noted above, the major disadvantage of the MS-CG method is that the CG potential is not guaranteed to accurately reproduce the AA structure. However, in practice the resulting models often quite accurately reproduce structural properties, although water is a particular challenging system for MS-CG pair potentials.[84, 105] Conversely, an advantage of the MS-CG approach is that the approximate potential is easily calculated directly from AA simulation data without iteration. Consequently, the structural fidelity of the CG model directly assesses the accuracy of the approximate CG potential.

3.3 Transferability and representability problems

One major strength of AA potentials is that they are rather transferable.[106] Having been parameterized once, these potentials can be directly applied (i.e., transferred) to model a wide range of systems and thermodynamic state points without re-parameterization. Another strength of AA potentials is the existence of an established framework for accurately calculating thermodynamic properties.[30, 40] In contrast to AA potentials, which are assumed to be temperature-independent energies, the exact CG potential (the many-body PMF) is a temperature-dependent excess Helmholtz potential.[28] This has important ramifications for the transferability of approximate CG potentials and also leads to new challenges in modeling thermodynamic properties, which are often referred to as “representability” issues.[20, 31, 107–113]

3.3.1 Transferability issues

The relative entropy[86] and MS-CG variational principles[22] both determine a unique effective potential that optimally approximates the many-body PMF at a single thermodynamic state point. However, since the PMF varies with state point, the optimal approximate potential should also vary with state point. Consequently, an approximate potential that accurately reproduces the structural and thermodynamic properties of an AA model at one state point may provide a much less accurate description at others.[7, 28] In practice, most bottom-up approaches employ AA simulations at one state point to determine a single interaction potential that is then treated as independent of thermodynamic conditions.[3, 114–122] The extended ensemble and multistate iterative Boltzmann inversion approaches provide a framework for considering a range of thermodynamic conditions when determining this single potential.[123–128] In either case, though, the transferability of this single potential is then assessed based upon its ability to accurately model a range of state points.

Many studies have demonstrated that a state-point independent CG potential often provides limited transferability.[115, 118, 121, 129–132] However, in some cases a single potential parameterized at a fortuitous state point can provide surprising transferability.[114, 133, 134] For instance, the early study by Ghosh and Faller employed IBI to parameterize CG models for ortho-terphenyl (OTP) at $T = 230$ K and 300 K, which are below and above the OTP glass transition temperature, respectively. They demonstrated that the

potential parameterized at $T = 300$ K failed to reproduce the pair structure at $T = 230$ K, although the potential parameterized at $T = 230$ K reproduced the pair structure at $T = 300$ K quite accurately.[114] More recently, Guo and coworkers employed IBI to parameterize the potential for a CG polyimide model at $T = 800$ K.[134] The resulting potential provided a reasonable transferability down to $T = 300$ K.[134]

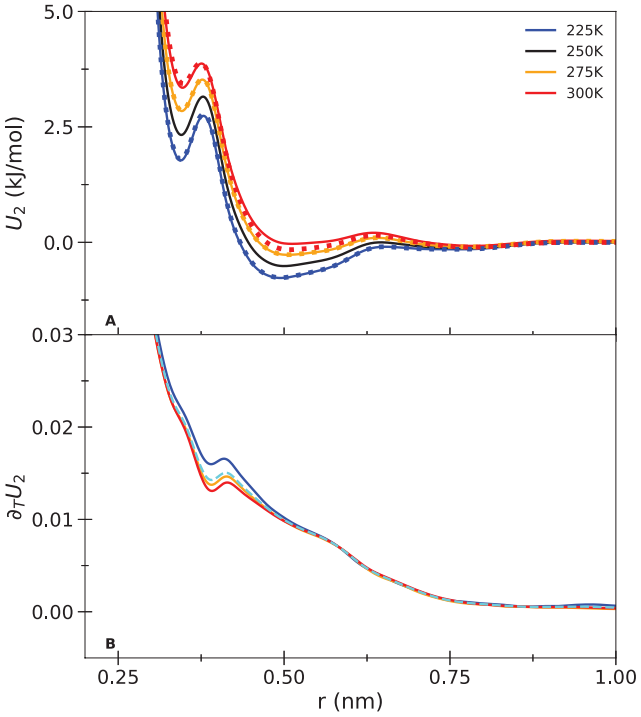

Fig. 7 Impact of temperature upon MS-CG pair potentials for 1-site CG models of methanol at constant density. Panel A presents MS-CG pair potentials. The solid curves present pair potentials calculated from constant NVT AA simulations at a fixed density $\rho_0 = 792.3$ kg/m^3 and the indicated temperatures. The dotted curves present predicted pair potentials calculated via linear extrapolation using the mean finite difference approximation of the temperature-dependence according to: $U_2(T) = U_2(T_0) + \partial_T U_2(T_0)\Delta T$ where $T_0 = 250$ K and $\Delta T = T - T_0$. Panel B presents estimates for the temperature derivative of the pair potential at constant density, $(\partial U_2/\partial T)_\rho$. The solid curves present the finite differences, $\partial_T U_2$, calculated at each temperature according to Eq. (50). The cyan curve corresponds to the mean of these finite differences, which is employed as $\partial_T U(T_0)$ in the linear extrapolation. The blue, black, orange, and red curves correspond to $T = 225$ K, 250 K, 275 K, and 300 K, respectively.

A growing number of studies have adopted a different approach to transferability issues. Rather than trying to determine a single state-point independent potential that provides a compromise for approximating the PMF over a range of state points, these studies have attempted to directly quantify and then model the variations in the PMF with state point.[80, 131, 135–147] According to Eq. (21), $\partial W(\mathbf{R})/\partial T = -S_W(\mathbf{R})$.[28] Thus, the temperature-dependence of the PMF in a given configuration, $\mathbf{R}$, is determined by the configurational entropy lost from the subensemble of AA configurations that map to $\mathbf{R}$. For high resolution models, one expects that $S_W \approx 0$, which suggests that the corresponding approximate potentials should be predominantly energetic and demonstrate little temperature variation.[28, 148] Indeed, Section 2.4.2 demonstrates that $S_W \to 0$ with increasing resolution. Additionally, W will vary linearly over the temperature range for which S_W is approximately constant. Consequently, one expects that the CG potentials for 1-site models of methanol and other small molecular liquids will likely vary linearly with temperature within single-phase regions of the phase diagram. Indeed, a rather large body of work now confirms this expectation.[32, 80, 108, 131, 137–139, 146–150]

For instance, an early study by Müller-Plathe and coworkers modeled the temperature-dependence of IBI potentials for liquid hexane based upon linear interpolation between the IBI potentials explicitly calculated for two extreme temperatures.[138] They demonstrated that this simple linear interpolation more accurately described the temperature-dependent IBI potentials than an empirical nonlinear model motivated by Boltzmann weights.[136] Since then, many studies have employed simple two-point linear interpolation schemes to model effective potentials across a range of temperatures.[131, 142, 146, 147, 151]

The calculations of Lebold and Noid, which are illustrated in Fig. 7, demonstrate this simple linear trend.[146] The solid curves in Fig. 7A present MS-CG pair potentials calculated for 1-site models of methanol across a range of temperatures at a fixed density $\rho_0 = 792.3$ kg/m^3. Note that the potentials become increasingly repulsive with increasing temperature at the constant density, which is consistent with $S_W \leq 0$ according to Eq. (17). The curves in Fig. 7B present finite difference estimates for the temperature-derivative of the MS-CG pair potentials, which are calculated according to

$$\partial_T U_2(T) = \frac{U_2(T) - U_2(T_0)}{T - T_0}, \tag{50}$$

where $T_0 = 250$ K. Clearly, the pair potentials for 1-site MS-CG models of methanol vary almost linearly with temperature at constant density.

Accordingly, a growing number of studies have adopted an intuitive free energy decomposition of approximate CG potentials to model their temperature-dependence[80, 131, 139, 146, 149, 150]

$$U_2(T) = U_2(T_0) - S_2(T_0)\Delta T, \tag{51}$$

where $S_2(T_0)$ is estimated as a finite difference according to Eq. (50) and $\Delta T = T - T_0$. The dotted curves presented in Fig. 7A demonstrate that this linear extrapolation works

nicely for the 1-site MS-CG methanol model.[146] Moreover, Voth and coworkers have demonstrated that this finite difference estimate for the entropic contribution to CG pair potentials, S_2, can then be used to estimate entropic properties of the AA model.[149] Recent studies have also demonstrated that simple linear interpolation can also accurately predict the temperature-dependence of volume potentials.[32, 142]

We close this subsection with two brief comments. In the case of lower resolution CG models, CG sites often represent atomic groups with considerable flexibility. In such cases, S_W will likely be relatively large in magnitude and demonstrate more complex temperature-dependence due to conformational transitions that are internal to each CG site.[152] Secondly, it is important to emphasize that S_W describes the temperature variation in the many-body PMF at constant density.[146, 147] Indeed, Fig. 7a demonstrates that the MS-CG pair potentials increase with temperature at constant density. However, to quantify temperature variations at constant pressure, one must also consider the corresponding density variations.

As an illustration, Fig. 8A presents the corresponding 1-site methanol pair potentials that are calculated from constant NPT simulations at varying temperatures and a constant external pressure of 1 bar. In striking contrast to Fig. 7A, the MS-CG pair potentials now become increasingly attractive with increasing temperature due to the density-dependence of the pair potentials.[108, 131, 146] Figure 8B presents the corresponding finite difference estimates for this density-dependence

$$\partial_\rho U_2(T) = \frac{U_2(T,\rho_T) - U_2(T,\rho_0)}{\rho_T - \rho_0}, \tag{52}$$

where ρ_T is the mean density of the constant NPT AA simulation at the temperature T and $\rho_0 = 792.3$ kg/m^3. The resulting finite difference estimates indicate that the CG pair potentials also vary nearly linearly with density. Indeed, the symbols in Fig. 8A indicate that the MS-CG pair potentials obtained under different conditions can be accurately described by linearly modeling both their temperature- and density-dependence. The recent work of Voth and coworkers suggests that the temperature-dependence at constant external pressure may be naturally interpreted by considering a Legendre transform of the PMF from constant volume to constant pressure.[131]

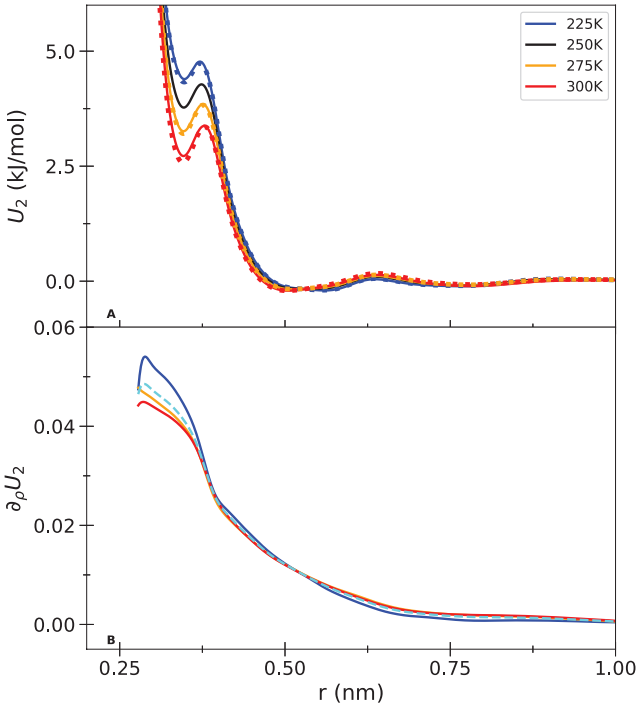

Fig. 8 Impact of temperature upon MS-CG pair potentials for 1-site CG models of methanol at constant external pressure. Panel A presents MS-CG pair potentials calculated from constant NPT AA simulations at the indicated temperatures and an external pressure of 1 bar. The solid curves present MS-CG pair potentials explicitly calculated from AA simulations at each state point. The dotted curves present predicted pair potentials calculated via linear extrapolation using the finite difference approximations for the temperature- and density-dependence according to: $U_2(T) = U_2(T_0) + \partial_T U_2(T_0)\Delta T + \partial_\rho U_2(T_0)\Delta\rho_T$ where $T_0 = 250$ K, $\rho_0 = 792.3$ kg/m^3, $\Delta T = T - T_0$, $\Delta\rho_T = \rho_T - \rho_0$, ρ_T is the mean density of the constant NPT AA simulation at the temperature T, and $\partial_T U(T_0)$ is the estimate for the temperature-derivative given by the cyan curve in Fig. 7B. Panel B presents estimates for the density derivative of the pair potential at constant temperature, $(\partial U_2/\partial\rho)_T$. The solid curves present the finite differences, $\partial_\rho U_2$, calculated for $T =$ 225 K, 275 K, and 300 K according to Eq. (52). The cyan curve corresponds to the mean of these finite differences, which is employed as $\partial_\rho U_2(T_0)$ to model the density-dependence for the dotted curves in panel A. The blue, black, orange, and red curves correspond to $T =$ 225 K, 250 K, 275 K, and 300 K, respectively.

3.3.2 Representability issues

Many prior discussions of representability issues have focused on difficulties in modeling both structure and pressure-volume equations of state with bottom-up pair potentials.[18, 94, 100] Since structure-based pair potentials tend to be highly repulsive, many studies introduce corrections to the CG pair potential in order to reduce the internal pressure and reproduce the correct density, although the resulting models tend to overestimate the compressibility.[153, 154] Following the insights of Das and Andersen[99], Dunn and Noid demonstrated that volume potentials can be self-consistently determined to quantitatively reproduce pressure-volume equations of state for bulk systems.[94, 100] Because these volume potentials depend only upon the instantaneous volume and are otherwise independent of configuration, they preserve the structural fidelity of the CG model. Similarly, local density potentials, i.e., one-body potentials of the local density around each CG site, also provide a significantly improved description of pressure-volume thermodynamics, while providing remarkable transferability between bulk and

interfacial systems.[155–161] Since these considerations have been discussed previously - and in order to provide a simple, didactic presentation - the present work focuses on the representability issues that result from the temperature-dependence of the PMF and, specifically, the difficulty of modeling atomic energetics.

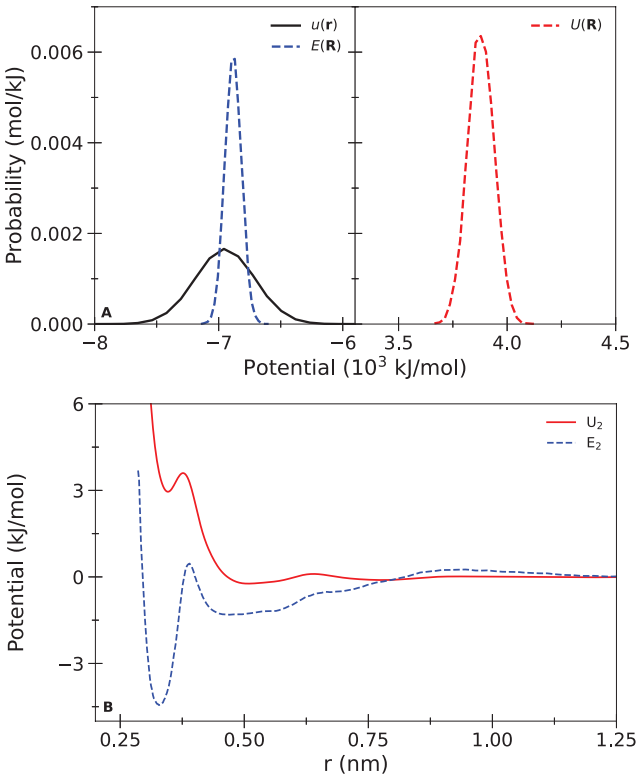

Fig. 9 Energetics of methanol models. Panel A presents distributions of intermolecular energies obtained from constant NVT simulations at $T_0 = 280$ K and $\rho_0 = 791.9$ kg/m^3. The solid black curve presents the distribution of intermolecular potential energies, u_{Inter}, sampled by an AA simulation. The dashed red and blue curves present distributions obtained from simulations of the 1-site CG model that employed the MS-CG potential, $U(\mathbf{R}) = \sum_{(I,J)} U_2(R_{IJ})$, to model interactions. The red curve presents the distribution obtained by evaluating the MS-CG potential, $U(\mathbf{R})$, for the sampled configurations, while the blue curve presents the distribution obtained by evaluating the energetic operator, $E(\mathbf{R}) = \sum_{(I,J)} E_2(R_{IJ})$, for the same configurations. Panel B compares the MS-CG pair potential, U_2, (solid red) and the pair energy function, E_2 (dashed blue).

The most straightforward approach for modeling energetics is to simply treat the approximate CG potential, U, as a potential energy. Figure 9 illustrates this naïve approach to evaluating energetics for a one-site CG model of methanol at $T_0 = 280$ K. The black curve in Fig. 9A presents the distribution of intermolecular potential energies, u_{Inter}, sampled by an AA simulation at this temperature. As expected, the sampled intermolecular potentials are large and negative, reflecting the cohesive energy stabilizing the liquid phase in the AA model. The red curve in Fig. 9B presents the pair potential, U_2, calculated by applying the MS-CG variational principle to this AA simulation. CG simulations with this MS-CG potential nicely reproduce the AA pair structure. However, the red curve in Fig. 9A presents the distribution that results from evaluating the MS-CG potential, U, over the configurations sampled by this CG MD simulation. Because the MS-CG pair potential is almost completely repulsive, the resulting energies are large and positive. This dramatic discrepancy reflects the same entropic contributions to approximate bottom-up potentials that also give rise to their temperature-dependence. Consequently, bottom-up potentials cannot be naïvely treated as potential energies.

3.3.3 Dual potential approach

Both transferability and representability issues arise from the same root cause, i.e., the state-point dependence of the many-body PMF, which suggests that a single approach can simultaneously address both issues.[20] Based upon this reasoning, Lebold and Noid recently proposed a dual potential approach that accurately models AA energetics and simultaneously provides a quantitative estimate of the entropic contribution to the many-body PMF.[32, 148] In this approach, conventional bottom-up methods are employed to determine an effective potential, U, that optimally approximates W at a single state point. Simulations that employ U as a conservative potential should accurately model the equilibrium configuration distribution of the underlying atomistic model. However, as demonstrated in Section 2.3.2 and 3.3.2, U cannot be used to accurately model energetic properties. Moreover, due to the state point variation of the PMF, U may provide a poor approximation to W at other state points. Thus, we adopt a second, "energy-matching" variational principle[162] to determine an energetic operator, E, that optimally approximates the energetic component, U_W, of the PMF. AA energetics can then be accurately modeled by evaluating E for the sampled CG configurations. Furthermore, the knowledge of U and E determines an estimate for the entropic component, S_W, of the PMF:

$$S(\mathbf{R}) \equiv T^{-1}\left(E(\mathbf{R}) - U(\mathbf{R})\right). \tag{53}$$

To the extent that U approximates W and E approximates U_W, it follows that $S \approx S_W$. The entropic function, $S(\mathbf{R})$, can then be used to model entropic quantities and to predict the temperature-dependence of $U(\mathbf{R})$,[146, 149] i.e.,

$$\partial U(\mathbf{R})/\partial T \approx \partial W(\mathbf{R})/\partial T = -S_W(\mathbf{R}) \approx -S(\mathbf{R}). \tag{54}$$

In practice,[32, 148, 150] U_W is approximated with a simple molecular mechanics form analogous to Eq. (43):

$$E(\mathbf{R}) = \sum_{\zeta}\sum_{\lambda} E_{\zeta}(\psi_{\zeta\lambda}(\mathbf{R})), \tag{55}$$

where E_ζ may be considered the energetic contribution to U_ζ. For example, if U_ζ is an effective pair potential, then E_ζ describes the energetic contribution to this pair interaction. For each interaction type in the CG model, the energy function is represented by a linear combination of basis functions, $\{f_{\zeta d}(x)\}$, with constant coefficients, $\{c_{\zeta d}\}$:

$$E_\zeta(x) = \sum_d c_{\zeta d} f_{\zeta d}(x). \tag{56}$$

In particular, we typically use "hat" or spline functions as basis functions for E_ζ. The energetic operator may then be expressed

$$E(\mathbf{R}) = \sum_\zeta \sum_d A_{\zeta d}(\mathbf{R}) c_{\zeta d}, \tag{57}$$

where

$$A_{\zeta d}(\mathbf{R}) = \sum_\lambda f_{\zeta d}(\psi_{\zeta\lambda}(\mathbf{R})). \tag{58}$$

The set $\{A_{\zeta d}(\mathbf{R})\}$ included in Eq. (57) form a basis spanning a linear space of approximate energy operators, while the coefficients $\{c_{\zeta d}\}$ serve as parameters for optimizing E.

The optimal coefficients are determined by minimizing the energy-matching functional[148]

$$\chi^2_{\mathrm{E}}[E] = \left\langle |E(\mathbf{M}(\mathbf{r})) - u(\mathbf{r})|^2 \right\rangle. \tag{59}$$

Because U_W is a conditioned mean of u, the energy-matching functional can be decomposed analogously to the MS-CG functional [22, 32, 96, 148]

$$\chi^2_{\mathrm{E}}[E] = \chi^2_{\mathrm{E}}[U_W] + \int_{V^N} \mathrm{d}\mathbf{R}\, p_R(\mathbf{R}) \, |E(\mathbf{R}) - U_W(\mathbf{R})|^2 \geq \chi^2_{\mathrm{E}}[U_W]. \tag{60}$$

Consequently, given a space of approximate energy operators, the operator that minimizes $\chi^2_{\mathrm{E}}[E]$ provides the optimal approximation to U_W in a least squares sense. While Tóth originally employed χ^2_{E} to determine an effective potential, U, that approximated W for modeling interactions in MD simulations,[162] the dual potential employs χ^2_{E} to determine an operator, E, that approximates U_W for estimating the energetics of sampled configurations.

In order to minimize χ^2_E, we approximate the ensemble average in Eq. (59) with a set of atomic configurations, $\{\mathbf{r}_t\}$, sampled by an AA MD simulation. The minimizing condition may be expressed as an over-determined system of linear equations:

$$u(\mathbf{r}_t) \cong E\left(\mathbf{M}(\mathbf{r}_t)\right) = \sum_\zeta \sum_d A_{\zeta d}\left(\mathbf{M}(\mathbf{r}_t)\right) c_{\zeta d} \tag{61}$$

for each sampled configuration, $\mathbf{r}_t$. The symbol '$\cong$' indicates that this system of over-determined equations is solved in a least squares sense. For 1-site CG models, $U(\mathbf{R}) = \sum_{(I,J)} U_2(R_{IJ})$ and $E(\mathbf{R}) = \sum_{(I,J)} E_2(R_{IJ})$. In this case, E_2 is determined to reproduce the intermolecular AA potential by solving Eq. (61) with $u \equiv u_{\mathrm{Inter}}$. For multi-site CG models that treat both intra- and intermolecular interactions, we decompose the AA potential, $u = u_{\mathrm{Intra}} + u_{\mathrm{Inter}}$, and energetic operator, $E = E_{\mathrm{Intra}} + E_{\mathrm{Inter}}$, into corresponding intra- and inter-molecular contributions. In this case, we consider two independent over-determined systems of equations. Specifically, we determine E_{Intra} by solving Eq. (61) with $u \equiv u_{\mathrm{Intra}}$, while determining E_{Inter} by solving Eq. (61) with $u \equiv u_{\mathrm{Inter}}$. It is worth noting that Dannenhoffer-Lafage and coworkers have proposed alternative approaches for optimizing E, as well as other observable operators, that may prove more computationally efficient than the simple energy-matching approach described above.[163]

Lebold and Noid first introduced this dual potential approach and demonstrated its compatibility with two distinct bottom-up approaches for parameterizing the interaction potential, U: IBI for 1-site models water and MS-CG for 1-site models of methanol.[148] Subsequently, they considered implicit solvent models of Lennard-Jones fluids, which contain significantly larger entropic contributions, and also extended the dual potential approach to the constant NPT ensemble.[32] More recently, Szukalo and Noid extended the dual potential approach to treat intramolecular degrees of freedom in 3-site CG models of ortho-terphenyl (OTP).[150] Here we briefly summarize results for a 1-site model methanol, while Section 4 discusses our ongoing work with OTP.

Figure 10 illustrates the results of applying the energy-matching variational method for the 1-site model of methanol at the state point $T_0 = 280$ K and $\rho_0 = 791.9$ kg/m^3. Specifically, the dashed blue curve in Fig. 10B presents the calculated pair energy function, E_2. Interestingly, the MS-CG pair potential, U_2, and the pair energy function, E_2, demonstrate similar features on similar length-scales. However, the pair energy function is significantly more attractive and exhibits more pronounced features. The dashed blue curve in Fig. 10A presents the energetic distribution obtained by evaluating the energetic operator, E, over the set of CG configurations sampled by simulations with the MS-CG interaction potential, U. The resulting energy distribution quantitatively reproduces the mean AA potential, but underestimates the variance in the AA potential distributions. This discrepancy fundamentally reflects the AA fluctuations that are hidden by the CG mapping. According to Sect. 2.3.3, these fluctuations give rise to a configuration-dependent specific heat C_W.

Given the MS-CG interaction potential, $U(T_0)$, and energetic operator, $E(T_0)$, calculated at the reference temperature $T_0 = 280$ K, the entropic component of the many-body PMF was approximated according to Eq. (53): $S(T_0) = T_0^{-1}\left(E(T_0) - U(T_0)\right) \approx S_W(T_0)$. The solid curve in Fig. 10A presents the resulting pair entropy function, $S_2(T_0)$, while

the dashed curve presents an empirical estimate for the pair entropy function based upon finite difference estimates for the temperature derivative of the calculated MS-CG pair potentials according to Eq. (50). The two calculations of the pair entropy functions agree quite well.

The pair entropy function that is inferred by applying the dual potential approach at the single temperature, $T_0 =$ 280 K, can then be employed to predict MS-CG pair potentials at other temperatures and the same density according to Eq. (51). The dotted lines in Fig. 10B present the predicted pair potentials for the 1-site CG model for methanol. The predicted potentials agree remarkably well with the MS-CG pair potentials that were independently calculated for each temperature from corresponding AA simulations. Thus, the dual potential approach not only accurately reproduces the energetics of the AA model for methanol, but also accurately predicts the temperature-dependence of the MS-CG potential.

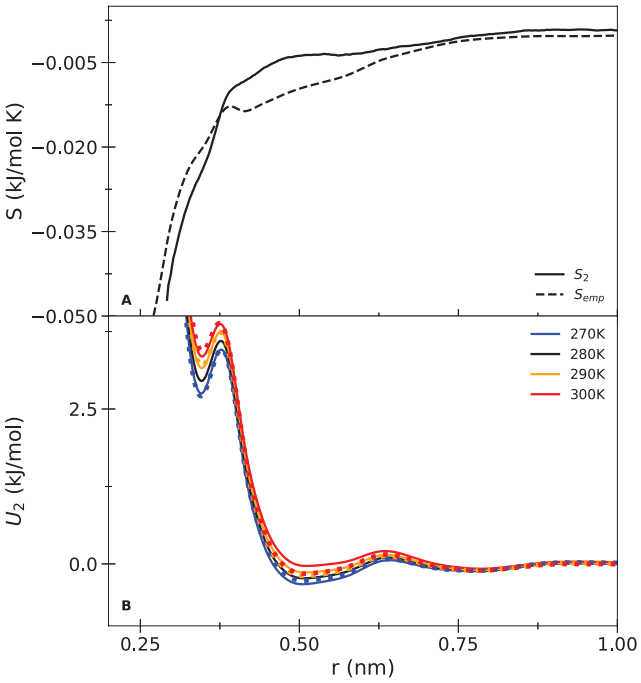

Fig. 10 Entropic predictions of the dual potential approach. Panel A compares the predicted and inferred pair entropy functions. The solid curve presents the pair entropy function calculated by the dual potential approach based upon constant NVT simulations at ($T_0 = 280$ K, $\rho_0 = 791.9$ kg/m^3). The dashed curve presents the empirical estimate of the pair entropy function determined from explicitly calculating the MS-CG pair potential from constant NVT simulations with $\rho_0 =$ 791.9 kg/m^3 and $T =$ 270 K, 290 K, and 300 K. Panel B assess the accuracy of the dual potential predictions for the temperature-dependence of the MS-CG pair potentials. The solid curves present pair potentials calculated from constant NVT AA simulations at $\rho_0 = 792.3$ kg/m^3 and several temperatures. The dotted curves present predictions for these pair potentials based entirely upon information at the reference state point ($T_0 = 280$ K, $\rho_0 = 791.9$ kg/m^3).

4 Coarse graining in practice

The preceding section demonstrated that the dual-potential approach provides a rigorous and robust framework for addressing the temperature-dependence of the many-body PMF for relatively simple systems, such as liquid methanol. This section summarizes our recent and ongoing work applying the dual-potential approach to OTP, which is a model glass-former and a considerably more complex system. Glass transitions are particularly interesting for bottom-up structure-based approaches because the dynamic and thermodynamic properties of the liquid dramatically change, while the local structure remains relatively unchanged.[164, 165] An early study by Ghosh and Faller investigated the transferability of IBI effective potentials across this glass transition.[114] Douglas and coworkers have also examined the dynamical properties of CG models for OTP near the glass transition.[166] More recently, we developed MS-CG models for OTP and systematically analyzed the temperature- and density-dependence of the effective potentials across the glass transition, using the techniques discussed in Sect. 3.3.1.[147] Interestingly, the linear trends observed for methanol and other simple molecular liquids may break down across the glass transition.[147]

In this section, we present dual potential results for modeling OTP in the constant NVT ensemble at a single fixed density. By comparing results for two different CG representations, we gain insight into the practical ramifications of resolution upon approximate CG models for complex systems. Additionally, we outline future work for extending this approach to develop a simple model for state-point dependent potentials that can accurately and predictively describe the structure and thermodynamic properties of OTP across the entire liquid region of its phase diagram and, hopefully, also for supercooled states below its glass transition.

4.1 Mapping considerations

As discussed in Section 3.1, the first step in constructing the CG model is to determine an appropriate mapping operator, $\mathbf{M}(\mathbf{r})$. Figure 11A presents the molecular structure of OTP, as well as two mappings of OTP that are consistent with our chemical intuition.

The 1-site mapping represents the entire molecule with a single sphere at the molecular mass center. A 1-site mapping corresponds to the lowest resolution representation that still explicitly describes each molecule. (Note that other one-site mappings are possible that employ different definitions of the CG site location.) Assuming a pair-additive approximation to the PMF, the approximate potential, U, is then completely determined by a single pair potential, U_2, which is likely the most computationally efficient model that explicitly models each OTP molecule. (Note that it is possible that

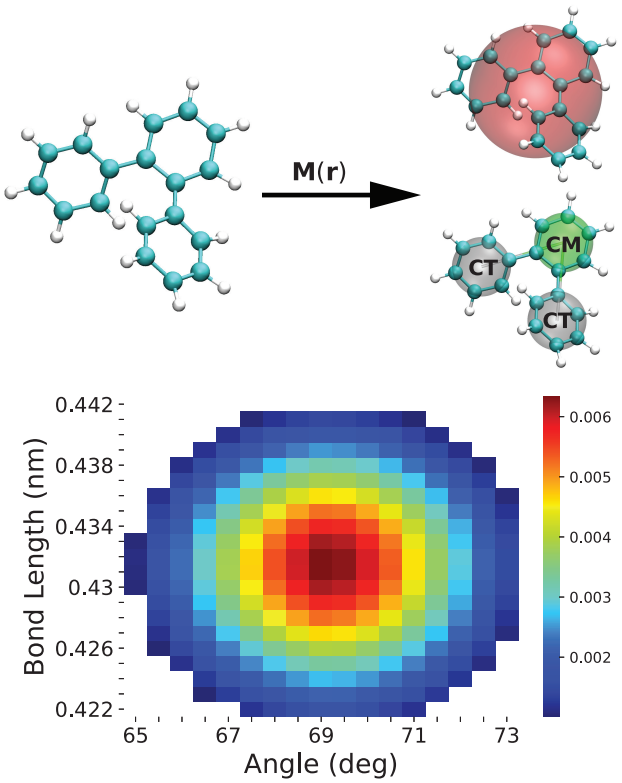

Fig. 11 Mapping considerations for OTP. Top: AA representation of the OTP molecular structure, as well as 1- and 3-site CG representations. The 1-site mapping represents all OTP atoms with a single CG site located at the molecular mass center, which is shown in red. The 3-site mapping represents the central benzene ring with a CM site (shown in green) and the terminal benzene rings with equivalent CT sites (shown in grey). Bottom: Joint probability distribution, $P(\theta, r_b)$, quantifying the correlation between the OTP bond angle and bond lengths in the mapped ensemble for the 3-site CG representation.

higher resolution models may be more computationally efficient if they adopt a sufficiently shorter cut-off.) A possible limitation of this 1-site mapping for OTP, though, is that it does not properly describe the molecular symmetry of OTP. In particular, since OTP is relatively flat, an isotropic spherical potential may not be able to accurately reproduce the intermolecular packing of the underlying AA model.

Accordingly, the second 3-site mapping is perhaps more consistent with chemical intuition. Here, the central benzene ring is represented with a CM site (green) and the terminal benzene rings are represented with two equivalent CT sites (grey), while each site is located at the mass center of the corresponding atomic group. Two earlier CG studies of OTP also adopted this 3-site mapping,[114, 167] while the recent work by Wang and Goméz-Bombarelli demonstrated that this mapping allowed for an optimal reconstruction of AA configurations.[25] The approximate potential for this CG representation describes the molecular geometry with independent bond, U_b, and angle, U_θ, potentials, while describing intermolecular interactions with 3 independent pair potentials: $U_{\mathrm{CM-CM}}$, $U_{\mathrm{CM-CT}}$, and $U_{\mathrm{CT-CT}}$. As discussed in Section 3.1, the 3-site representation may possibly introduce complex cross-correlations between bond and angle degrees of freedom that may impact the accuracy of the CG model. However, the joint probability distribution in Fig. 11B indicates that the CG bond and angle degrees of freedom are statistically independent. Consequently, the simple approximate potential in Eq. (43) appears sufficient for describing the conformation of OTP.

4.2 MS-CG effective potential

We adopted the OPLS-AA model[168] as a simple high resolution model for OTP. Our initial simulation study indicated that this AA model undergoes a glass transition at 1 bar external pressure near $T_g \approx 343$K.[147] Consequently, in order to develop a predictive CG model for the liquid phase, we performed constant NPT simulations of the OPLS-AA model at 1 bar pressure and $T_0 = 400$ K, which resulted in an equilibrium density of $\rho_0 = 1033$ kg/m^3. We treat the thermodynamic state point (T_0, ρ_0) as a reference state for investigating the temperature-dependence, as well as the energetic and entropic properties of CG models for OTP at constant density.

We mapped this reference AA simulation to the 1- and 3-site CG representations illustrated in Fig. 11. The solid curves in Fig. 12A and B present the mapped AA rdfs for the 1- and 3-site representations, respectively. As expected, the mapped rdf demonstrates more pronounced structural features for the higher resolution 3-site representation than for the lower resolution 1-site representation. The first peak of the 3-site rdf occurs at $r \approx 0.5$ nm and, presumably, corresponds to an intermolecular stacking interaction between benzene rings. Interestingly, this interaction gives rise to a shoulder in the first peak of the 1-site rdf, which reaches a relatively larger maximum at a considerably greater distance of $r \approx 0.8$ nm.

Given these mapped ensembles, we employed BOCS v 4.0[169] to determine the MS-CG potential, U, that provides a variationally optimal approximation for the configuration-dependence of the many-body PMF, W, at each resolution. The dashed black curve in Fig. 12C presents the pair potential, U_2, for the 1-site model, while the dashed blue curve presents the CM-CM pair potential, $U_{\mathrm{CM-CM}}$, for the 3-site model. Note that the CM-CM pair potential is plotted on a much smaller scale. For simplicity, we will not present results for the remaining inter- or intra-molecular interactions in the 3-site MS-CG model, although they are presented and analyzed in the original study.[150]

For both resolutions, the pair potentials are almost entirely repulsive. However, the 1-site pair potential is much more repulsive and larger in magnitude than the 3-site pair potentials. This is consistent with the expectation that the

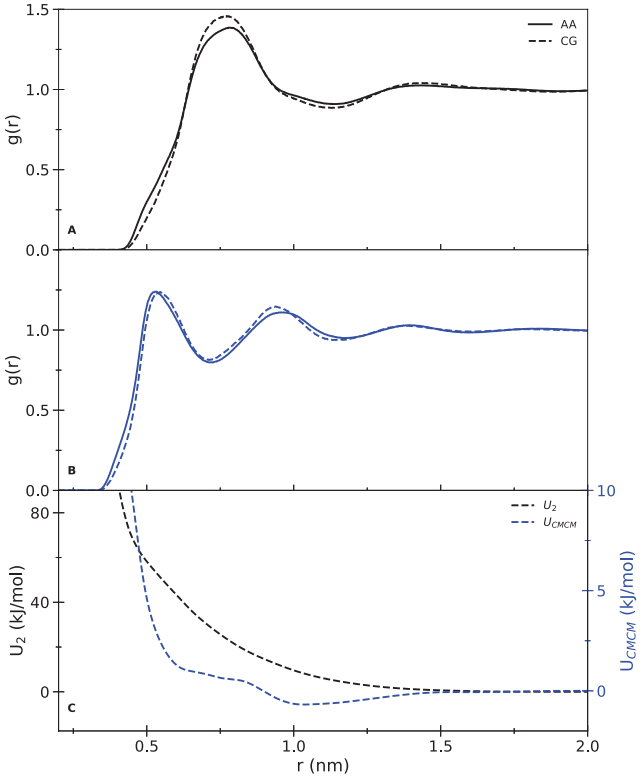

Fig. 12 Pair structure and MS-CG pair potentials for liquid OTP at a reference state point with $T_0 = 400$ K and $\rho_0 = 1033$ kg/m^3. Panel A presents rdfs for the 1-site model, while panel B presents CM-CM rdfs for the 3-site model. In panels A and B, the solid curves present mapped rdfs obtained from AA simulations, while the dashed curves present rdfs calculated from CG simulations which employed the corresponding MS-CG potential as an interaction potential. Panel C presents the MS-CG pair potential, U_2 for the 1-site model (black) and the CM-CM pair potential, $U_{\mathrm{CM-CM}}$, for the 3-site model (blue). Note that the 1-site pair potential is plotted with respect to the left scale, while the 3-site CM-CM pair potential is plotted with respect to the much smaller right scale.

potentials for the lower resolution model should include a larger, repulsive entropic contribution.

We then performed constant NVT simulations of both CG models at the reference state point (T_0, ρ_0), while employing the corresponding MS-CG potentials to model interactions. The dashed curves in Fig. 12A and B present the resulting site-site rdfs. The 1-site model understimates the shoulder and slightly overestimates the first peak of the AA rdf, while the 3-site model reproduces the AA rdf with nearly quantitative accuracy. Although not shown here, the 3-site model also accurately reproduced the CM-CT and CT-CT mapped rdfs, as well as the intramolecular bond and angle distributions.[150] These results suggest that the 3-site mapping provides a better description of the shape and packing of OTP molecules than the 1-site model.[64, 150] Nevertheless, both models reproduce the mapped AA ensemble with relatively high fidelity. Since the MS-CG models are not guaranteed to reproduce rdfs, the observed agreement indicates that the MS-CG potential quite accurately describes the configuration-dependence of the PMF at each resolution.

4.3 Modeling AA energetics

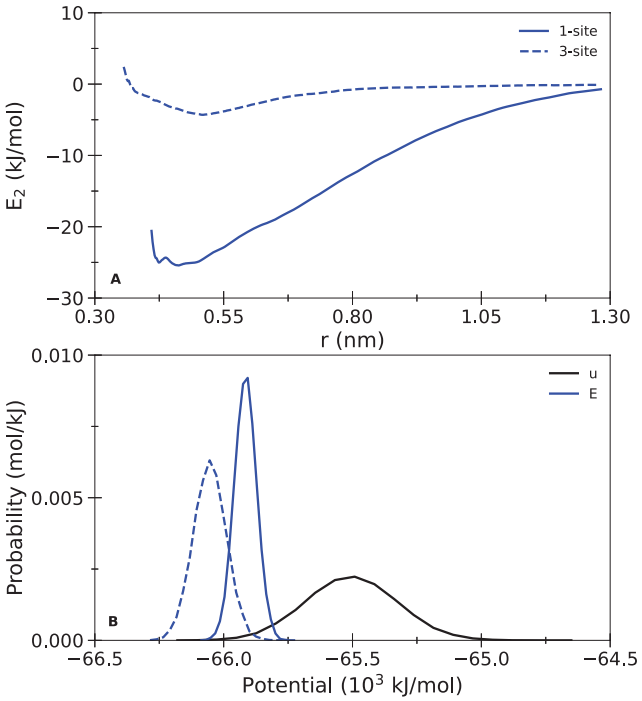

Fig. 13 Energetic pair functions and energetic distributions for OTP at the reference state point with $T_0 = 400$ K and $\rho_0 = 1033$ kg/m^3. Panel A presents the pair energy functions for the 1-site model, E_2 (solid curve), and also the CM-CM pair energy function, $E_{\mathrm{CM-CM}}$ (dashed curve), for the 3-site model. Panel B presents distributions of intermolecular potential energies. The solid black curve corresponds to the AA energy distribution. The blue curves present energy distributions obtained by evaluating E for the configurations sampled by CG simulations that employ U as the interaction potential. The solid blue curve presents the distribution for the 1-site model, while the dashed blue curve presents the distribution obtained for the 3-site model.

Given the configurations and potential energies sampled by the constant NVT AA simulation at the reference state point, (T_0, ρ_0), we employed the energy-matching variational principle to determine an energetic operator, E, that optimally approximates the energetic contribution, U_W, to the many-body PMF at each CG representation. The 1-site energetic operator is completely specified by a single pair energy function, E_2, which we determined from the AA intermolecular energies according to Eq. (61). The solid blue curve in Fig. 13A presents the resulting pair energy function for the 1-site model. The 3-site energetic operator included pair energy functions for the CM-CM, CM-CT, and CT-CT intermolecular pairs, as well as intramolecular bond and angle energy functions. The intra- and inter-molecular energy

functions were determined to match the AA intra- and intermolecular potential energies, respectively. The dashed blue curve in Fig. 13A presents the resulting CM-CM pair energy function.

In stark contrast to the highly repulsive MS-CG pair potentials, the pair energy functions are highly attractive. The 1-site pair energy function, E_2, is much more attractive than the CM-CM pair energy function, $E_{\mathrm{CM-CM}}$. This is presumably because the 3-site model can distribute the total atomic potential across many more site-site interactions. It is also interesting that, while the 1- and 3-site rdfs reach maxima at rather different distances, the 1- and 3-site pair energy functions both reach a minimum near $r = 0.5$ nm. This distance corresponds to the first peak of the 3-site CM-CM rdf and presumably reflects direct stacking of benzene rings. Since the peak of the 1-site rdf occurs at considerably greater distance, this coincidence of the pair energy minima further emphasizes that the effective MS-CG pair potentials incorporate both energetic and entropic contributions.

Figure 13B assesses the accuracy of the dual potential approach for calculating energetic properties at a given CG resolution. The solid black curve presents the AA intermolecular potential energy distribution. The solid and dashed blue curves present CG intermolecular potential energy distributions obtained from evaluating the 1- and 3-site energetic operators, respectively, for the set of configurations sampled from CG simulations that employed the corresponding MS-CG potentials from Fig. 12 as interaction potentials. Both CG models reproduce the mean AA intermolecular potential to within a few percent, although both models underestimate the variance in the AA potential distribution due to the loss of atomic fluctuations. As expected, since the 3-site model preserves more conformational information and thus, greater energetic fluctuations, the 3-site energy distribution (dashed curve) is somewhat broader than the 1-site energy distribution (solid curve). Although it is not shown, the intramolecular energy operator for the 3-site model also reproduces the AA distribution of intramolecular potential energies with similar accuracy.

4.4 Entropic operators

As described in Section 3.3.3, the approximate MS-CG potential, U, and the approximate energetic operator, E, determine an approximation, S, for the entropic contribution, S_W, to the exact many-body PMF, W. Figure 14 presents the pair entropy operators determined from Eq. (53). The solid black curve presents the pair entropy function, S_2, for the 1-site model. The dashed, dotted, and dashed-dotted blue curves present the pair entropy functions corresponding to the CM-CM, CM-CT, and CT-CT interactions for the 3-site model. For comparison, the solid green curve presents the corresponding dual potential estimate for the entropic contribution to the MS-CG pair potential for a 1-site model of methanol at the same temperature $T_0 = 400\ K$ and density $\rho_0 = 737.31$ kg/m^3.

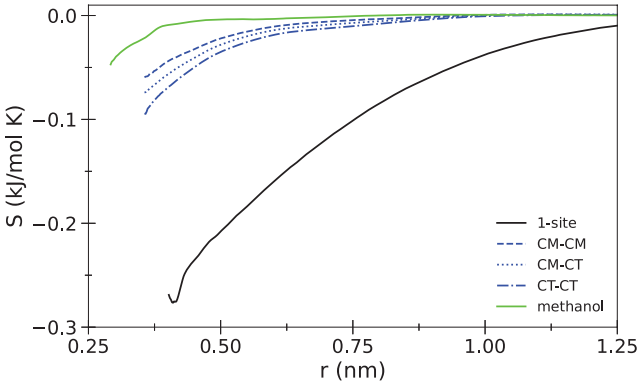

Fig. 14 Entropic pair functions, $S_2(r) = (E_2(r) - U_2(r))/T_0$, for the 1-site and 3-site models of OTP calculated via the dual potential approach at the reference state point with $T_0 = 400$ K and $\rho_0 = 1033$ kg/m^3. The black and blue curves present the pair entropy function for the 1-site and 3-site model, respectively. The dashed, dotted, and dashed-dotted blue curves correspond to entropic pair functions calculated for each non-bonded interaction in the 3-site model: $S_{\mathrm{CM-CM}}$, $S_{\mathrm{CM-CT}}$, and $S_{\mathrm{CT-CT}}$, respectively. For comparison, the green curve presents the pair entropy function for a 1-site MS-CG model of methanol at the same temperature $T_0 = 400$ K and the density $\rho_0 = 737.31$ kg/m^3.

Figure 14 provides general insight into the entropic components of effective potentials for CG models of complex systems and, moreover, how these entropic components vary with resolution. As expected from Eq. (17), the pair entropy functions are almost entirely negative. The absolute value of the pair entropy functions reach a maximum near contact and then generally decay towards 0 with increasing distance. Because the CG sites in the 1-site methanol model only represent two heavy atoms, the corresponding pair entropy function is relatively small and rapidly decays to zero. Because the CG sites in the 3-site OTP model represent 6 heavy atoms, the corresponding pair entropy functions are considerably larger and act over considerably longer distances. Because the CG sites in the 1-site OTP model represent 18 heavy atoms, the corresponding pair entropy function is even much larger and acts over a much longer distance. Thus, as suggested earlier in Section 3.3.1, the entropic contribution to approximate effective potentials appears to grow in magnitude and range with decreasing resolution.

4.5 Predicted temperature-dependence

In addition to providing an approximate operator for modeling AA entropies,[149] the approximate entropy functions also provide a predictive estimate for the temperature-dependence of the MS-CG potentials. Given the calculated MS-

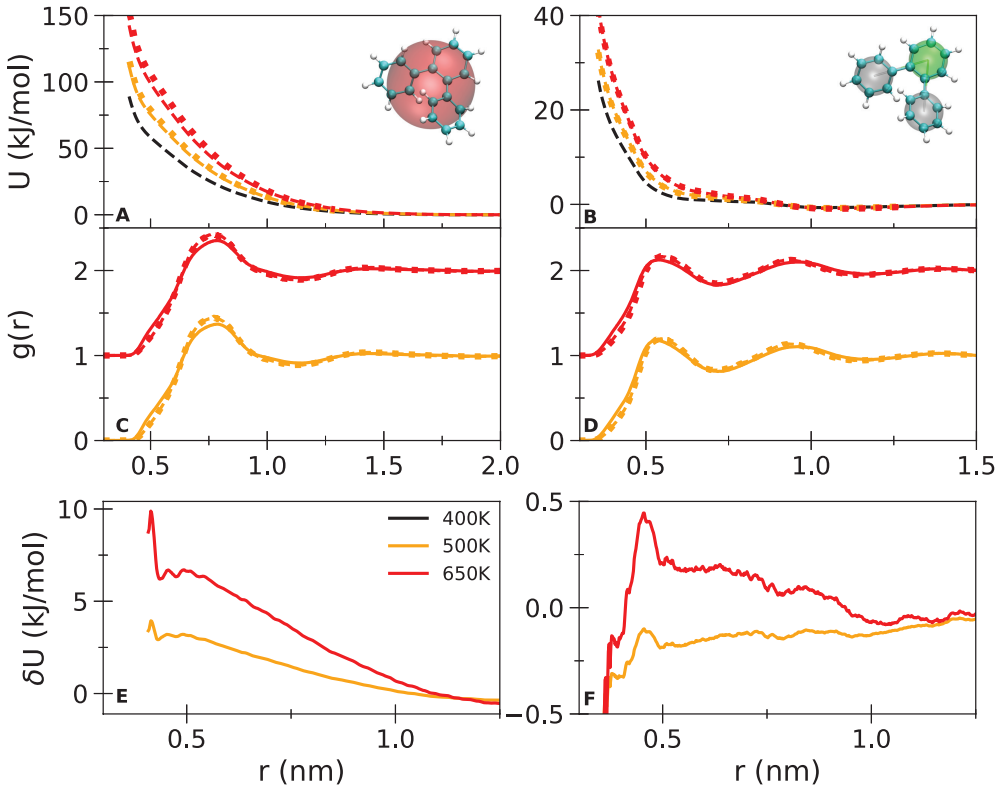

Fig. 15 Pair structure and MS-CG pair potentials characterizing the 1-site (left column) and 3-site (right column) CG models for liquid OTP at constant density, $\rho_0 = 1033$ kg/m^3. Panels A and B present the MS-CG pair potential for the 1-site and 3-site CM-CM pair interactions, respectively. The dashed curves present the "optimized" MS-CG pair potentials, which were calculated from the AA simulations at the corresponding state point. The dotted curves present the "predicted" MS-CG pair potentials, which were calculated via the dual potential approach and only employed AA simulation data from the reference state point ($T_0 = 400$ K, $\rho_0 = 1033$ kg/m^3). Panels C and D present the rdfs for the 1-site and 3-site CM-CM pair interaction, respectively, which were obtained from constant NVT simulations at the fixed density ρ_0 and indicated temperature. The solid curves present mapped rdfs obtained from AA simulations. The dashed and dotted curves present rdfs calculated from CG simulations that employed the optimized and predicted MS-CG potentials, respectively, as an interaction potential. Panels E and F quantify the difference, δU, between the optimized and predicted MS-CG pair potentials. The black, orange, and red curves correspond to results for $T_0 = 400$ K, $T = 500$ K, and $T = 650$ K, respectively.

CG potential, $U(T_0)$, and inferred entropic operator, $S(T_0)$, at the reference state point (T_0, ρ_0), the dual potential approach predicts the temperature-dependence of the MS-CG potential according to Eq. (51). Note that this extrapolation assumes constant density and that $S(T_0)$ is approximately constant between T_0 and T.

Figure 15 presents results of this simple dual potential approach for predicting the MS-CG potentials for OTP at $T = 500$ K and 650 K based upon information sampled at $T_0 = 400$ K, while maintaining a constant density $\rho_0 = 1033$ kg/m^3. In this figure, the left panels present results for the 1-site model, while the right panels present results describing the CM-CM pair in the 3-site model. The dashed black curves in panels A and B present the reference pair potentials employed in this extrapolation for the 1- and 3-site model, respectively. The dotted curves in panels A and B present the predicted pair potential, $U_2(T)$, for the 1-site model and the predicted CM-CM pair potential, $U_{\mathrm{CM-CM}}(T)$, for the 3-site model, respectively. The dotted curves in panels C and D present the corresponding site-site rdfs obtained from constant NVT simulations with the predicted CG potentials.

In order to assess the accuracy of the predicted CG potentials, we performed additional constant NVT AA simulations at $T = 500$ K and 650 K with the fixed density $\rho_0 = 1033$ kg/m^3. The solid curves in panels C and D present the resulting mapped AA rdfs. Clearly, the predicted CG models very accurately reproduced the pair structure of the AA simulations.

From these additional AA simulations, MS-CG potentials were calculated for the 1- and 3-site representations of OTP. The dashed curves in panels A and B present the resulting pair potentials for the 1-site model and the CM-CM pair potentials for the 3-site model for comparison with the potentials predicted via the dual potential approach. Indeed, the dual potential approach predicts the temperature-dependent MS-CG potentials with remarkable accuracy. The simple linear prediction slightly overestimates the temperature-dependence of the 1-site pair potential, but almost quantita-

tively reproduces the temperature-dependence of the 3-site potentials.

Panels E and F quantify the difference between the MS-CG pair potentials specifically optimized for each temperature and the pair potentials predicted via the dual potential approach. Relative to the magnitude of the effective potentials, this discrepancy is quite small for both resolutions. This error appears to increase systematically with temperature in the 1-site model. This systematic error suggests that the simple linear extrapolation in Eq. (51) fails to account for the temperature-dependence in the pair entropy function, S_2, for the one-site model. Conversely, the error in the predictions for the 3-site model are approximately an order of magnitude smaller and also less systematic. This suggests that the pair entropy functions for the 3-site model are approximately constant over this temperature range, while the pair entropy function for the 1-site model varies more rapidly with temperature.

In order to investigate this hypothesis, we also applied the dual potential approach at $T = 500$ K and 650 K. We first performed energy matching to determine approximate energetic operators for each model based upon AA simulation data at each temperature. We inferred the corresponding entropic operators based upon the difference between the MS-CG potentials and the energy operators calculated for each temperature. We compared these inferred entropy functions with the empirical entropy functions determined from the observed temperature-dependence of the MS-CG potentials specifically calculated for each temperature. Figure 16 presents the results of this analysis. As in Fig. 15, the left column presents the results for the 1-site model, while the right column presents results for the CM-CM interaction in the 3-site model.

Figure 16 largely confirms our hypothesis. The pair energy and entropy functions for the 1-site model do indeed systematically increase with temperature. This temperature variation then determines an approximation to the configuration-dependent specific heat, C_W, which is discussed in Sect. 2.3.2. By accounting for this configuration-dependent specific heat, the dual potential approach should more accurately predict the temperature-dependence of the MS-CG pair potentials. Conversely, with the exception of the CM-CM pair energy function at $T = 650$ K, the energetic and entropic pair functions for the 3-site model appear almost temperature-independent. This is consistent with the expectations of Section 3.3.1, i.e., the entropic contributions to effective pair potentials should not only systematically increase with decreasing resolution, but also demonstrate increasing temperature-dependence. In particular, the entropic contributions to the 3-site pair potential appear almost independent of temperature. Consequently, the predictions of the dual approach for the 3-site model should be independent of reference state point over this temperature range.

We also assess the accuracy of the energetic operators for reproducing the mean AA potentials at each CG resolution. Table 1 presents the resulting average energies calculated at each temperature for the 1-site and 3-site models. As expected, the approximate energy functions, E, reproduce the mean AA energies within a few percent. In particular, the intramolecular operators for the 3-site model, E_{Intra}, exactly reproduce the average AA intramolecular potential at all three state points.

4.6 Outlook

The preceding work represents the first step towards predictive CG models for the liquid phase of OTP. It will clearly be important to extend this work for modeling OTP at different densities. Lebold and Noid previously applied the dual potential approach to develop implicit solvent models for simulating the constant NPT ensemble at a range of temperatures. [32] Following their study, the first step will be to perform constant NPT AA simulations at a range of external pressures. From these simulations, we will determine the density-dependence of the MS-CG interaction potentials, $U(\mathbf{R})$, as well as the volume potential, $U_V(V)$, that is necessary to reproduce the AA pressure-density equations of state. We will then perform energy-matching to estimate the energetic contributions to the interaction and volume potentials, from which the corresponding entropic contributions can then be inferred. Given these energetic and entropic contributions, as well as their density dependence, it should be possible to predict pair and volume potentials for modeling the entire liquid region of the phase diagram. As observed in Section 4.5, it is likely that simple linear predictions may prove less successful for the lower resolution 1-site model. In this case, it may prove necessary to account for the configuration-dependent specific heat, C_W. Ultimately, it is hoped that this work will provide a prototype for efficiently establishing predictive CG models for reproducing the structure and thermodynamics of molecular liquids across large regions of their phase diagram.

5 Conclusion

While bottom-up models have been plagued by transferability and representability difficulties, our analysis of the many-body PMF clarifies their fundamental origin. More importantly, this analysis suggests computational methodologies for addressing these difficulties in practice. Specifically, by properly distinguishing the energetic and entropic contributions to the PMF, the dual potential approach provides a simple and rigorous approach for addressing the transferability and representability issues that stem from the temperature-dependence of the PMF. Similarly, by properly accounting

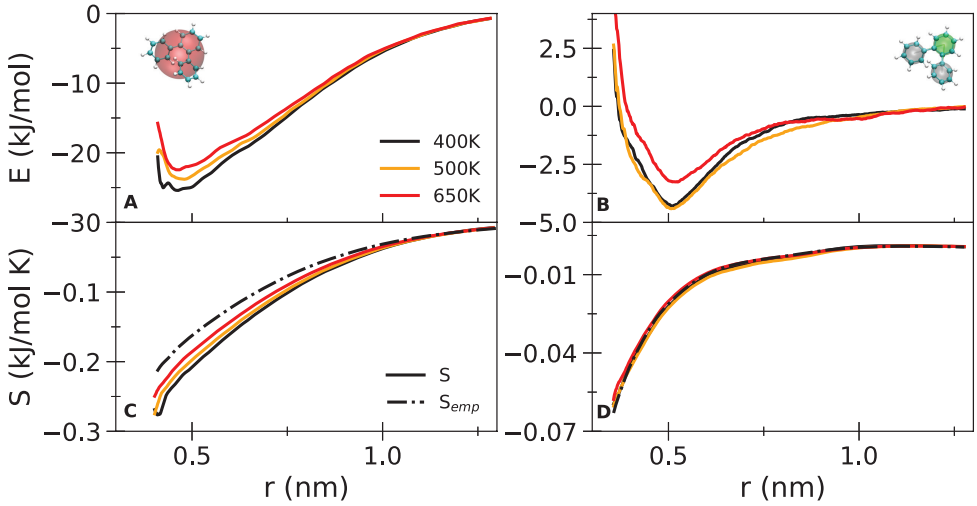

Fig. 16 Application of the dual potential approach to the 1-site and 3-site CG models of OTP at a fixed density, $\rho_0 = 1033$ kg/m^3. As before, the left panels present results for the 1-site model, while the right panels present results for the CM-CM pair in the 3-site model. Panels A and B present the resulting pair energy functions calculated via the energy-matching variational principle, which employed the energies sampled by AA simulations at the corresponding state point. Panels C and D present the corresponding pair entropy potentials inferred for each temperature. The black dashed-dotted curve presents an empirical estimate for S_2, which was calculated as a finite difference according to Eq. (50). The black, orange, and red curves correspond to results for $T_0 = 400$ K, $T = 500$ K, and $T = 650$ K, respectively.

Table 1 Table summarizing various average energies for AA and CG models of OTP at a fixed density, $\rho_0 = 1033$ kg/m^3. The temperatures are reported in K and the energies are reported in 10^3 kJ/mol.

1-site				3 -site						
T	u_{Inter}	U	E	T	u_{Inter}	U_{Inter}	E_{Inter}	u_{Intra}	U_{Intra}	E_{Intra}
400	-64.6	131.1	-65.7	400	-64.6	38.6	-66.0	196.5	126.9	196.8
500	-62.3	183.8	-63.5	500	-62.3	74.6	-63.6	222.7	88.2	227.1
650	-59.4	242.1	-60.6	650	-59.4	100.5	-61.2	272.3	27.3	272.7

for the atomic fluctuations that are hidden by the CG resolution, this approach can accurately and consistently model the atomic specific heat. The present numerical results demonstrate that the dual potential approach is not only simple and predictive, but also remarkably accurate for modeling rather complex fluids, such as OTP. While additional work is necessary to treat the density-dependence of CG potentials with similar rigor and to treat more complex polymeric systems, we hope that our ongoing work with OTP will establish a robust computational framework for developing predictive CG models that accurately model both structural and thermodynamic properties across wide ranges of the phase diagram. Finally, our analysis of CG mappings for the GNM and for OTP elucidate the importance of the mapping upon the information content, thermodynamic properties, and accuracy of CG models. Consequently, we hope that this work will provide fundamental insights and useful guidance for developing and analyzing CG models of complex systems - both in theory and in practice.

Acknowledgements The authors gratefully acknowledge the essential contributions of Tommy Foley, M. Scott Shell, and Kate Lebold to the prior studies that are explicitly discussed herein. The authors also acknowledge former group members Michael DeLyser, Nick Dunn, Joe Rudzinski, and Wayne Mullinax, who made important contributions to the computational methods employed in these studies. The authors gratefully acknowledge financial support from the National Science Foundation (Grant Nos. MCB-1053970, CHE-1565631, CHE-1856337) that made this work possible, as well as a fellowship to Kate Lebold from the Molecular Sciences Software Institute under NSF Grant No. ACI-1547580. Portions of this research were conducted with Advanced CyberInfrastructure computational resources provided by The Institute for Computational and Data Sciences at The Pennsylvania State University (http://icds.psu.edu). In addition, parts of this research were conducted with XSEDE resources awarded by Grant TG - CHE170062. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by the National Science Foundation (Grant ACI-1548562). Figs. 1-3, 11, and 15, and 16 employed VMD.[170] VMD is developed with NIH support by the Theoretical and Computational Biophysics group at the Beckman Institute, University of Illinois at Urbana - Champaign.

Author Contributions

K.K. and R.S. contributed equally to this work under the supervision of W.N. K.K. and R.S. both performed origi-

nal calculations and analyzed previously published results. K.K., R.S., and W.N. wrote and edited the manuscript.

Data Availability Statement

The data that support the findings of this study will be openly available via Jupyter Notebooks at the following URL: https://www.datacommons.psu.edu

References

1. M.L. Klein, W. Shinoda, Science **321**(5890), 798 (2008). DOI 10.1126/science.1157834
2. C. Peter, K. Kremer, Faraday Discuss. **144**, 9 (2010)
3. M. Guenza, M. Dinpajooh, J. McCarty, I. Lyubimov, J. Phys. Chem. B **122**(45), 10257 (2018)
4. M. Muller, K. Katsov, M. Schick, Phys. Rep. **434**(5-6), 113 (2006). DOI 10.1016/j.physrep.2006.08.003
5. F. Schmid, Macromol. Rapid Comm. **30**(9-10), 741 (2009). DOI 10.1002/marc.200800750
6. M. Deserno, Macromol. Rapid Comm. **30**(9-10), 752 (2009). DOI 10.1002/marc.200900090
7. W.G. Noid, J. Chem. Phys. **139**(9), 090901 (2013). DOI http://dx.doi.org/10.1063/1.4818908
8. W.G. Noid, Methods Mol Biol **924**, 487 (2013). DOI 10.1007/978-1-62703-017-5_19
9. M.G. Saunders, G.A. Voth, Annu. Rev. Biophys. **42**, 73 (2013). DOI 10.1146/annurev-biophys-083012-130348
10. E. Brini, E.A. Algaer, P. Ganguly, C. Li, F. Rodríguez-Ropero, N.F.A. van der Vegt, Soft Matter **9**, 2108 (2013). DOI 10.1039/C2SM27201F
11. R. Potestio, C. Peter, K. Kremer, Entropy **16**(8), 4199 (2014). DOI 10.3390/e16084199
12. H.I. Ingólfsson, C.A. Lopez, J.J. Uusitalo, D.H. de Jong, S.M. Gopal, X. Periole, S.J. Marrink, Wiley Interdiscip. Rev. Comput. Mol. Sci. **4**(3), 225 (2014). DOI 10.1002/wcms.1169
13. S. Kmiecik, D. Gront, M. Kolinski, L. Wieteska, A.E. Dawid, A. Kolinski, Chem. Rev. **116**, 7898 (2016)
14. S.Y. Joshi, S.A. Deshmukh, Mol. Simu. pp. 1–18 (2020)
15. M. Giulini, M. Rigoli, G. Mattiotti, R. Menichetti, T. Tarenzi, R. Fiorentini, R. Potestio, Frontiers in Molecular Biosciences **8**, 460 (2021). DOI 10.3389/fmolb.2021.676976
16. J.F. Rudzinski, Comput. **7**(3), 42 (2019)
17. A. Liwo, C. Czaplewski, J. Pillardy, H.A. Scheraga, J. Chem. Phys. **115**, 2323 (2001)
18. C.N. Likos, Phys. Rep. **348**(4–5), 267 (2001). DOI http://dx.doi.org/10.1016/S0370-1573(00)00141-1
19. R.L.C. Akkermans, W.J. Briels, J. Chem. Phys. **114**(2), 1020 (2001). DOI 10.1063/1.1330744
20. N.J.H. Dunn, T.T. Foley, W.G. Noid, Acc. Chem. Res. **49**(12), 2832 (2016)
21. M.E. Tuckerman., Statistical Mechanics: Theory and Molecular Simulation (Oxford University Press, Oxford, Great Britain, 2013)
22. W.G. Noid, J.W. Chu, G.S. Ayton, V. Krishna, S. Izvekov, G.A. Voth, A. Das, H.C. Andersen, J. Chem. Phys. **128**, 244114 (2008)

23. G. Ciccotti, R. Kapral, E. Vanden-Eijnden, ChemPhysChem **6**, 1809 (2005)
24. L. Zhang, J. Han, H. Wang, R. Car, W. E, J. Chem. Phys. **149**(3), 034101 (2018). DOI 10.1063/1.5027645
25. W. Wang, R. Gómez-Bombarelli, npj Comput. Mat. **5**(1), 125 (2019). DOI 10.1038/s41524-019-0261-5
26. J.F. Rudzinski, W.G. Noid, J. Chem. Phys. **135**(21), 214101 (2011). DOI 10.1063/1.3663709
27. S. Kullback, R.A. Leibler, Ann. Math. Stat. **22**(1), 79 (1951)
28. T.T. Foley, M.S. Shell, W.G. Noid, J. Chem. Phys. **143**, 243104 (2015)
29. J.G. Kirkwood, J. Chem. Phys. **3**(5), 300 (1935)
30. D. Frenkel, B. Smit, Understanding Molecular Simulation: From Algorithms to Applications, 2nd edn. (Academic Press, San Diego, CA USA, 2002)
31. J.W. Wagner, J.F. Dama, A.E.P. Durumeric, G.A. Voth, J. Chem. Phys. **145**(4), 044108 (2016). DOI http://dx.doi.org/10.1063/1.4959168
32. K.M. Lebold, W.G. Noid, J. Chem. Phys. **151**(16), 164113 (2019)
33. T.T. Foley, K.M. Kidder, M.S. Shell, W. Noid, Proc. Natl. Acad. Sci. U.S.A. **117**(39), 24061 (2020)
34. P.J. Flory, M. Gordon, N.G. McCrum, Proc. Roy. Soc. Lond. A: Math. Phys. Sci. **351**(1666), 351 (1976). DOI 10.1098/rspa.1976.0146
35. T. Haliloglu, I. Bahar, B. Erman, Phys. Rev. Lett. **79**, 3090 (1997)
36. I. Bahar, T.R. Lezon, L.W. Yang, E. Eyal, Annu. Rev. Biophys. **39**, 23 (2010). DOI 10.1146/annurev.biophys.093008.131258
37. J.M. Harris, J.L. Hirst, M.J. Mossinghoff, Combinatorics and graph theory (Springer, 2010)
38. M.C. Wang, G.E. Uhlenbeck, Rev. Mod. Phys. **17**, 323 (1945)
39. T.R. Lezon, I. Bahar, PLoS Comput. Biol. **6**(6), e1000816 (2010). DOI 10.1371/journal.pcbi.1000816
40. D.A. McQuarrie, Statistical Mechanics (University Science Books, 2000)
41. R. Baron, A.H. de Vries, P.H. Hünenberger, W.F. van Gunsteren, J. Phys. Chem. B **110**(16), 8464 (2006). DOI 10.1021/jp055888y
42. R. Baron, V. Molinero, J. Chem. Theory Comput. **8**(10), 3696 (2012). DOI 10.1021/ct300121r
43. S.T. Lin, M. Blanco, W. Goddard, J. Chem. Phys. **119**, 11792 (2003)
44. M.P. Bernhardt, M. Dallavalle, N.F. Van der Vegt, Soft Mater. pp. 1–16 (2020)
45. E. Brini, V. Marcon, N.F.A. van der Vegt, Phys. Chem. Chem. Phys. **13**(22), 10468 (2011). DOI 10.1039/c0cp02888f
46. A.P. Lyubartsev, A. Laaksonen, Phys. Rev. E **55**, 5689 (1997). DOI 10.1103/PhysRevE.55.5689
47. G.G. Rondina, M.C. Böhm, F. Müller-Plathe, J. Chem. Theory Comput. **16**(3), 1431 (2020)
48. M.K. Meinel, F. Müller-Plathe, J. Chem. Theory Comput. **16**(3), 1411 (2020)
49. M.E.J. Newman, G.T. Barkema, Monte Carlo Methods in Statistical Physics (Clarendon Press, 1999)
50. M.R. Shirts, J.D. Chodera, J. Chem. Phys. **129**(12), 124105 (2008)
51. H. Gohlke, M.F. Thorpe, Biophys. J. **91**, 2115 (2006)
52. Z.Y. Zhang, L.Y. Lu, W.G. Noid, V. Krishna, J. Pfaendtner, G.A. Voth, Biophys. J. **95**(11), 5073 (2008)
53. Z.Y. Zhang, G.A. Voth, J. Chem. Theory Comput. **6**(9), 2990 (2010). DOI 10.1021/ct100374a
54. P. Koehl, F. Poitevin, R. Navaza, M. Delarue, J. Chem. Theory Comput. **13**(3), 1424 (2017)
55. M. Girvan, M.E.J. Newman, Proc. Natl. Acad. Sci. U.S.A. **99**(12), 7821 (2002). DOI 10.1073/pnas.122653799
56. J. Reichardt, S. Bornholdt, Phys. Rev. E **74**(1) (2006). DOI 10.1103/PhysRevE.74.016110
57. S. Fortunato, Phys. Rep. **486**(3-5), 75 (2010). DOI 10.1016/j.physrep.2009.11.002
58. M.T. Schaub, J.C. Delvenne, S.N. Yaliraki, M. Barahona, PLoS ONE **7**(2), e32210 (2012). DOI 10.1371/journal.pone.0032210
59. M. Giulini, R. Menichetti, M.S. Shell, R. Potestio, J. Chem. Theory Comput. **16**(11), 6795 (2020)
60. D.H.E. Gross, Microcanonical Thermodynamics (WORLD SCIENTIFIC, 2001). DOI 10.1142/4340
61. L. Boninsegna, R. Banisch, C. Clementi, J. Chem. Theory Comput. **14**(1), 453 (2018). DOI 10.1021/acs.jctc.7b00990. Publisher: American Chemical Society
62. M.A. Webb, J.Y. Delannoy, J.J. de Pablo, J. Chem. Theory Comput. (2018). DOI 10.1021/acs.jctc.8b00920
63. M. Chakraborty, C. Xu, A.D. White, J. Chem. Phys. **149**(13), 134106 (2018)
64. J. Wang, S. Olsson, C. Wehmeyer, A. Pérez, N.E. Charron, G. De Fabritiis, F. Noé, C. Clementi, ACS Cent. Sci. **5**(5), 755 (2019)
65. J. Ruza, W. Wang, D. Schwalbe-Koda, S. Axelrod, W.H. Harris, R. Gómez-Bombarelli, J. Chem. Phys. **153**(16), 164501 (2020)
66. Z. Li, G.P. Wellawatte, M. Chakraborty, H.A. Gandhi, C. Xu, A.D. White, Chem. **11**(35), 9524 (2020)
67. M. Chakraborty, J. Xu, A.D. White, Phys. Chem. Chem. Phys. **22**(26), 14998 (2020)
68. S. Izvekov, G.A. Voth, J. Phys. Chem. B **109**, 2469 (2005)
69. S. Izvekov, G.A. Voth, J. Chem. Phys. **123**, 134105 (2005)

70. W.G. Noid, P. Liu, Y.T. Wang, J.W. Chu, G.S. Ayton, S. Izvekov, H.C. Andersen, G.A. Voth, J. Chem. Phys. **128**, 244115 (2008)
71. M. Dallavalle, N.F. van der Vegt, Phys. Chem. Chem. Phys. **19**(34), 23034 (2017)
72. A. Khot, S.B. Shiring, B.M. Savoie, J. Chem. Phys. **151**(24), 244105 (2019)
73. V.A. Harmandaris, D. Reith, N.F.A. Van der Vegt, K. Kremer, Macromol. Chem. Phys. **208**, 2109 (2007). DOI 10.1002/macp.200700245
74. O. Bezkorovaynaya, A. Lukyanov, K. Kremer, C. Peter, J. Comp. Chem. **33**(9), 937 (2012). DOI 10.1002/jcc.22915
75. T. Ohkuma, K. Kremer, Polymer **130**, 88 (2017)
76. J.F. Rudzinski, W.G. Noid, J. Chem. Theory Comput. **11**(3), 1278 (2015)
77. J.F. Rudzinski, W.G. Noid, J. Phys. Chem. B **118**(28), 8295 (2014)
78. J.F. Rudzinski, W.G. Noid, J. Phys. Chem. B **116**(29), 8621 (2012). DOI 10.1021/jp3002004
79. J.F. Rudzinski, W.G. Noid, Eur. Phys. J.: Spec. Top. **224**, 2193 (2015)
80. J. Jin, G.A. Voth, J. Chem. Theory Comput. **14**, 2180 (2018)
81. A. Chaimovich, M.S. Shell, Phys. Rev. E **89**(2), 022140 (2014)
82. C. Scherer, D. Andrienko, Phys. Chem. Chem. Phys. **20**(34), 22387 (2018)
83. J.I. Monroe, M.S. Shell, J. Chem. Phys. **151**(9), 094501 (2019)
84. J. Jin, Y. Han, A.J. Pak, G.A. Voth, J. Chem. Phys. **154**(4), 044104 (2021)
85. J. Jin, A.J. Pak, Y. Han, G.A. Voth, J. Chem. Phys. **154**(4), 044105 (2021)
86. M.S. Shell, J. Chem. Phys. **129**, 144108 (2008)
87. A. Chaimovich, M.S. Shell, J. Chem. Phys. **134**(9), 094112 (2011)
88. M.S. Shell, Coarse-Graining with the Relative Entropy (John Wiley & Sons, Inc., 2016), pp. 395–441. DOI 10.1002/9781119290971.ch5
89. A. Isihara, J. Phys. A: Math., Nucl., Gen. **1**(5), 539 (1968)
90. A. Lyubartsev, A. Mirzoev, L.J. Chen, A. Laaksonen, Faraday Discuss. **144**, 43 (2010)
91. D. Reith, M. Pütz, F. Müller-Plathe, J. Comp. Chem. **24**, 1624 (2003)
92. A.P. Lyubartsev, A. Laaksonen, Phys. Rev. E Stat. Phys. Plasmas Fluids Relat. Interdiscip. Topics **52**, 3730 (1995)
93. T. Murtola, M. Karttunen, I. Vattulainen, J. Chem. Phys. **131**, 055101 (2009)
94. N.J.H. Dunn, W.G. Noid, J. Chem. Phys. **144**, 204124 (2016)
95. F. Ercolessi, J.B. Adams, Europhys. Lett. **26**, 583 (1994)
96. A.J. Chorin, Multiscale Model. Simul. **1**, 105 (2003)
97. A.J. Chorin, O.H. Hald, Stochastic Tools in Mathematics and Science (Springer, New York, NY USA, 2006)
98. W.G. Noid, J.W. Chu, G.S. Ayton, G.A. Voth, J. Phys. Chem. B **111**, 4116 (2007)
99. A. Das, H.C. Andersen, J. Chem. Phys. **132**, 164106 (2010)
100. N.J.H. Dunn, W.G. Noid, J. Chem. Phys. **143**(24), 243148 (2015)
101. J.W. Mullinax, W.G. Noid, Phys. Rev. Lett. **103**, 198104 (2009)
102. J.W. Mullinax, W.G. Noid, J. Phys. Chem. C **114**, 5661 (2010)
103. S.P. Carmichael, M.S. Shell, J. Phys. Chem. B **116**(29), 8383 (2012). DOI 10.1021/jp2114994
104. S.Y. Mashayak, M.N. Jochum, K. Koschke, N.R. Aluru, V. Rühle, C. Junghans, PLOS ONE p. 20 (2015)
105. L. Larini, L.Y. Lu, G.A. Voth, J. Chem. Phys. **132**(16), 164107 (2010)
106. J.A. Harrison, J.D. Schall, S. Maskey, P.T. Mikulski, M.T. Knippenberg, B.H. Morrow, App. Phys. Rev. **5**(3), 031104 (2018). DOI 10.1063/1.5020808
107. A.A. Louis, J. Phys.: Condens. Matter **14**, 9187 (2002)
108. M.E. Johnson, T. Head-Gordon, A.A. Louis, J. Chem. Phys. **126**, 144509 (2007)
109. F.H. Stillinger, H. Sakai, S. Torquato, J. Chem. Phys. **117**(1), 288 (2002). DOI 10.1063/1.1480863
110. Y.T. Wang, W.G. Noid, P. Liu, G.A. Voth, Phys. Chem. Chem. Phys. **11**(12), 2002 (2009). DOI 10.1039/b819182d
111. A.J. Clark, J. McCarty, I.Y. Lyubimov, M.G. Guenza, Phys. Rev. Lett. **109**, 168301 (2012). DOI 10.1103/PhysRevLett.109.168301
112. J. McCarty, A.J. Clark, J. Copperman, M.G. Guenza, J. Chem. Phys. **140**(20), 204913 (2014)
113. G. D'Adamo, A. Pelissetto, C. Pierleoni, J. Chem. Phys. **138**(23), 234107 (2013). DOI 10.1063/1.4810881
114. J. Ghosh, R. Faller, Mol. Simu. **33**, 759 (2007)
115. P. Carbone, H.A.K. Varzaneh, X. Chen, F. Müller-Plathe, J. Chem. Phys. **128**, 064904 (2008)
116. D.M. Huang, R. Faller, K. Do, A.J. Moule, J. Chem. Theory Comput. **6**(2), 526 (2010). DOI 10.1021/ct900496t. PMID: 26617308
117. G. Megariotis, A. Vyrkou, A. Leygue, D.N. Theodorou, Ind. Eng. Chem. Res. **50**, 546 (2011)
118. B. Mukherjee, L. Delle Site, K. Kremer, C. Peter, J. Phys. Chem. B **116**(29), 8474 (2012)
119. A. Mirzoev, A.P. Lyubartsev, Phys. Chem. Chem. Phys. **13**, 5722 (2011). DOI 10.1039/C0CP02397C

120. Q. Xiao, H. Guo, Phys. Chem. Chem. Phys. **18**, 29808 (2016). DOI 10.1039/C6CP03753D
121. T.D. Potter, J. Tasche, M.R. Wilson, Phys. Chem. Chem. Phys. **21**, 1912 (2019). DOI 10.1039/C8CP05889J
122. S. Mortezazadeh, Y. Jamali, H. Naderi-Manesh, A.P. Lyubartsev, PLoS One **14**, e0214673 (2019)
123. J.W. Mullinax, W.G. Noid, J. Chem. Phys. **131**, 104110 (2009)
124. T.C. Moore, C.R. Iacovella, C. McCabe, J. Chem. Phys. **140**(22), 224104 (2014)
125. J.F. Rudzinski, K. Lu, S.T. Milner, J.K. Maranas, W.G. Noid, J. Chem. Theory Comput. **13**(5), 2185 (2017)
126. T. Sanyal, J. Mittal, M.S. Shell, J. Chem. Phys. **151**(4), 044111 (2019)
127. K. Shen, N. Sherck, M. Nguyen, B. Yoo, S. Köhler, J. Speros, K.T. Delaney, G.H. Fredrickson, M.S. Shell, J. Chem. Phys. **153**(15), 154116 (2020)
128. J.F. Rudzinski, T. Bereau, J. Chem. Phys. **153**(21), 214110 (2020)
129. J. Zhang, H. Guo, J. Phys. Chem. B **118**(17), 4647 (2014). DOI 10.1021/jp411615f
130. F. Cao, H. Sun, J. Chem. Theory Comput. **11**, 4760 (2015). DOI 10.1021/acs.jctc.5b00573
131. J. Jin, A. Yu, G.A. Voth, J. Chem. Theory Comput. **16**(11), 6823 (2020)
132. J. Jin, Y. Han, A.J. Pak, G.A. Voth, J. Chem. Phys. **154**(4), 044104 (2021)
133. J. Xia, Q. Xiao, H. Guo, Polymer **148**, 284 (2018)
134. C. Hu, T. Lu, H. Guo, J. Chem. Inf. Model. **59**(5), 2009 (2019). DOI 10.1021/acs.jcim.8b00887
135. T. Vettorel, H. Meyer, J. Chem. Theory Comput. **2**, 616 (2006)
136. H.J. Qian, P. Carbone, X. Chen, H.A. Karimi-Varzaneh, C.C. Liew, F. Müller-Plathe, Macromolecules **41**(24), 9919 (2008)
137. A. Chaimovich, M.S. Shell, Phys. Chem. Chem. Phys. **11**(12), 1901 (2009). DOI 10.1039/b818512c
138. K. Farah, A.C. Fogarty, M.C. Böhm, F. Müller-Plathe, Phys. Chem. Chem. Phys. **13**(7), 2894 (2011). DOI 10.1039/c0cp01333a
139. L. Lu, G.A. Voth, J. Chem. Phys. **134**(22), 224107 (2011). DOI 10.1063/1.3599049
140. A. Liwo, M. Khalili, C. Czaplewski, S. Kalinowski, S. Ołdziej, K. Wachucik, H.A. Scheraga, J. Phys. Chem. B **111**(1), 260 (2007). DOI 10.1021/jp065380a
141. G. Deichmann, M. Dallavalle, D. Rosenberger, N.F. van der Vegt, J. Phys. Chem. B **123**(2), 504 (2018)
142. D. Rosenberger, N.F.A. van der Vegt, Phys. Chem. Chem. Phys. **20**, 6617 (2018). DOI 10.1039/c7cp08246k
143. C. Hu, T. Lu, H. Guo, J. Chem. Inf. Model. **59**(5), 2009 (2019)
144. Y. Li, V. Agrawal, J. Oswald, J. Polym. Sci., Part B: Polym. Phys. **57**(6), 331 (2019)
145. M. King, S. Pasler, C. Peter, J. Phys. Chem. C **123**(5), 3152 (2019)
146. K.M. Lebold, W.G. Noid, J. Chem. Phys. **150**(1), 014104 (2019)
147. R.J. Szukalo, W. Noid, Soft Mater. **18**, 1 (2020)
148. K.M. Lebold, W.G. Noid, J. Chem. Phys. **150**, 234107 (2019)
149. J. Jin, A.J. Pak, G.A. Voth, J. Phys. Chem. Lett. **10**(16), 4549 (2019)
150. R.J. Szukalo, W.G. Noid, J. Phys.: Condens. Matter **33**(15), 154004 (2021). DOI 10.1088/1361-648x/abdff8
151. A.P. Lyubartsev, A. Laaksonen, Phys. Rev. E **55**(5), 5689 (1997)
152. J.F. Dama, A.V. Sinitskiy, M. McCullagh, J. Weare, B. Roux, A.R. Dinner, G.A. Voth, J. Chem. Theory Comput. **9**, 2466 (2013). DOI 10.1021/ct4000444
153. F. Müller-Plathe, ChemPhysChem **3**, 754 (2002)
154. H. Wang, C. Junghans, K. Kremer, Eur. Phys. J. E: Soft Matter Biol. Phys. **28**(2), 221 (2009)
155. I. Pagonabarraga, D. Frenkel, J. Chem. Phys. **115**(11), 5015 (2001)
156. T. Sanyal, M.S. Shell, J. Chem. Phys. **145**(3), 034109 (2016). DOI http://dx.doi.org/10.1063/1.4958629
157. M.R. DeLyser, W.G. Noid, J. Chem. Phys. **147**, 134111 (2017)
158. T. Sanyal, M.S. Shell, J. Phys. Chem. B **122**, 5678 (2018)
159. M.R. DeLyser, W. Noid, J. Chem. Phys. **151**(22), 224106 (2019)
160. M. DeLyser, W. Noid, J. Chem. Phys. **153**(22), 224103 (2020)
161. N. Shahidi, A. Chazirakis, V. Harmandaris, M. Doxastakis, J. Chem. Phys. **152**(12), 124902 (2020). DOI 10.1063/1.5143245
162. G. Tóth, J. Phys.: Condens. Matter **19**(33), 335222 (2007). DOI 10.1088/0953-8984/19/33/335222
163. T. Dannenhoffer-Lafage, J.W. Wagner, A.E.P. Durumeric, G.A. Voth, J. Chem. Phys. **151**(13), 134115 (2019). DOI 10.1063/1.5116027
164. M.D. Ediger, C.A. Angell, S.R. Nagel, J. Chem. Phys. **100**(31), 13200 (1996)
165. L. Berthier, G. Biroli, Rev. Mod. Phys. **83**(2), 587 (2011)
166. X. Song, M. Jensen, V. Jogini, R.A. Stein, C.H. Lee, H.S. Mchaourab, D.E. Shaw, E. Gouaux, Nature **556**(7702), 515 (2018). DOI 10.1038/s41586-018-0039-9
167. W. Xia, J. Song, N.K. Hansoge, F.R. Phelan Jr, S. Keten, J.F. Douglas, J. Phys. Chem. B **122**(6), 2040 (2018)

168. W.L. Jorgensen, D.S. Maxwell, J. Tirado-Rives, J. Am. Chem. Soc. **118**, 11225 (1996)
169. N.J.H. Dunn, K.M. Lebold, M.R. DeLyser, J.F. Rudzinski, W.G. Noid, J. Phys. Chem. B **122**(13), 3363 (2018). DOI 10.1021/acs.jpcb.7b09993
170. W. Humphrey, A. Dalke, K. Schulten, J. Mol. Graph. **14**, 33 (1996)