Coarse-graining entropy, forces, and structures

Cite as: J. Chem. Phys. **135**, 214101 (2011); https://doi.org/10.1063/1.3663709
Submitted: 01 September 2011 . Accepted: 04 November 2011 . Published Online: 02 December 2011

Joseph F. Rudzinski and W. G. Noid

Coarse-graining entropy, forces, and structures

Joseph F. Rudzinski and W. G. Noid[a)]
Department of Chemistry, The Pennsylvania State University, University Park, Pennsylvania 16802, USA

(Received 1 September 2011; accepted 4 November 2011; published online 2 December 2011)

Coarse-grained (CG) models enable highly efficient simulations of complex processes that cannot be effectively studied with more detailed models. CG models are often parameterized using either force- or structure-motivated approaches. The present work investigates parallels between these seemingly divergent approaches by examining the relative entropy and multiscale coarse-graining (MS-CG) methods. We demonstrate that both approaches can be expressed in terms of an information function that discriminates between the ensembles generated by atomistic and CG models. While it is well known that the relative entropy approach minimizes the average of this information function, the present work demonstrates that the MS-CG method minimizes the average of its gradient squared. We generalize previous results by establishing conditions for the uniqueness of structure-based potentials and identify similarities with corresponding conditions for the uniqueness of MS-CG potentials. We analyze the mapping entropy and extend the MS-CG and generalized-Yvon-Born-Green formalisms for more complex potentials. Finally, we present numerical calculations that highlight similarities and differences between structure- and force-based approaches. We demonstrate that both methods obtain identical results, not only for a complete basis set, but also for an incomplete harmonic basis set in Cartesian coordinates. However, the two methods differ when the incomplete basis set includes higher order polynomials of Cartesian coordinates or is expressed as functions of curvilinear coordinates. © *2011 American Institute of Physics*. [doi:10.1063/1.3663709]

I. INTRODUCTION

Despite tremendous recent advances in computational hardware,[1–4] software,[5–7] and methodology,[8–10] many processes of fundamental interest cannot be effectively simulated with conventional atomically detailed models. These considerations continue to motivate tremendous interest in highly efficient coarse-grained (CG) models that describe systems in reduced detail by grouping atoms into fewer effective interaction sites.[11–17] The potentials governing the site-site interactions are typically parameterized to reproduce either thermodynamic or structural properties of the system. For instance, pioneering studies by the Klein[18–20] and Marrink[21–23] groups have employed thermodynamic data to parameterize potentials that have subsequently provided considerable transferability for modeling a wide range of systems. However, despite considerable progress in this direction,[24,25] the resulting potentials may not necessarily provide quantitative accuracy for modeling the internal structure and flexibility of complex molecules.[23,26]

Alternatively, CG potentials may be parameterized to reproduce structural properties of an atomistic model for the same system. In principle, a many-body potential of mean force is the appropriate potential for a CG model that quantitatively reproduces all structural features of the atomistic model (at the resolution of the CG mapping).[12,27,28] However, in practice, the many-body potential of mean force must be approximated by simpler potentials.[29]

These approximate potentials are often parameterized to reproduce target low-order structural correlation functions (e.g., radial distribution functions) that are determined by the atomistic model and CG mapping. In some cases, these potentials can be determined by directly inverting the corresponding correlation functions.[30] However, if the interactions are coupled, CG potentials determined from direct Boltzmann inversion may not reproduce the atomistic correlation functions.[31–33] Motivated by renormalization group considerations,[34] the seminal early works by Lyubartsev and Laaksonen[35,36] developed an iterative inverse Monte Carlo (IMC) method for parameterizing approximate CG potentials. The IMC method represents the approximate potential as a sum of terms, each of which is the product of a potential parameter and a conjugate function of CG coordinates, i.e., a conjugate order parameter. The IMC approach employs Newton's method to determine the set of potential parameters reproducing the target averages for the conjugate set of order parameters, i.e., the corresponding atomistic correlation functions.[37] Papoian and coworkers have expanded upon these considerations to develop a molecular renormalization group[38,39] approach that has accurately modeled complex molecules such as DNA.[40]

Shell has recently proposed an elegant relative entropy formalism for variationally determining CG potentials.[41] Shell defined the relative entropy as an average of an information function, $\phi(\mathbf{r})$, corresponding to the Kullback-Leibler divergence[42] for discriminating between the ensembles of *atomistic* configurations sampled by atomistic and CG models. Minimizing the relative entropy determines the potential parameters that reproduce target atomistic averages for the conjugate order parameters.[41,43] In principle, these order parameters may be quite general and may correspond to structural or, e.g., energetic properties.[43] Moreover, the relative

a)Electronic mail:wnoid@chem.psu.edu

entropy may prove useful for estimating errors in CG models and for optimizing the CG mapping.[44]

Murtola *et al.* noted[45] that the relative entropy functional is closely related to the density functional employed by Chayes and coworkers[46,47] in earlier mathematical studies of the existence and uniqueness of potentials that reproduce known distribution functions. Furthermore, Murtola *et al.* also demonstrated that applying Newton's method to minimize the relative entropy leads to the IMC method of Lyubartsev and Laaksonen.[45] Consequently, the relative entropy formalism also provides a convenient variational framework for considering several structure-based CG approaches.

Independently, Izvekov and Voth[48,49] pioneered an alternative force-based approach.[50–54] This multiscale coarse-graining (MS-CG) method employs forces sampled from atomistic simulations[55,56] to variationally project the many-body mean force field (i.e., the force field corresponding to the many-body potential of mean force) onto a force field "basis set" that is defined by the form of the approximate CG potential.[50,51] When the variational calculation is performed with a complete basis set, the MS-CG method determines the many-body mean force. Simulations with this force field will quantitatively reproduce all structural correlations of the atomistic model (at the level of the CG mapping).[50,57] When the variational calculation is performed with an incomplete basis set, the MS-CG method determines the force field (within the subspace spanned by the incomplete basis) that is minimum "distance" from the many-body mean force field.[50,51] The MS-CG potential can be determined directly from a system of normal linear equations that are expressed in terms of force and structural correlation functions sampled from atomistic simulations. When re-expressed in terms of structural correlation functions, the resulting equations define a generalized-Yvon-Born-Green (g-YBG) integral equation theory for complex molecules.[58–61]

The relative entropy and MS-CG approaches both determine the same potential (to within an additive constant) when the corresponding variational calculations are performed with a complete basis set.[12,43,50] However, for calculations with an incomplete basis set, the two approaches appear quite divergent. In this case, the relative entropy approach employs multiple simulations to determine the CG potential that reproduces atomistic averages for the conjugate order parameters. In contrast, the MS-CG method does not require iterative simulations and employs atomistic force information to directly project the many-body mean force field onto the approximate basis set, but the resulting model is not guaranteed to reproduce any particular atomistic correlation functions. (It should be emphasized, though, that for many complex systems the MS-CG model does quantitatively or semi-quantitatively reproduce the corresponding atomistic distribution functions.[62–66]) Ruhle *et al.* have performed an insightful study that explicitly compared the potentials obtained for structure- and force-motivated models of water, methanol, and hexane.[67] In addition, Krishna and Larini have investigated mean field formulations of systematic coarse-graining methods, including the MS-CG and relative entropy approaches.[68] Nevertheless, the general relationship between structure- and force-based methods and the resulting potentials remains relatively obscure.

The present work explores the relationship between structure- and force-based CG approaches by examining the relative entropy and MS-CG approaches. This work identifies several striking parallels between the two approaches, including the relationship between their basis sets, the variational functionals, and the uniqueness of the resulting potentials. A major conclusion of the present work is that both the relative entropy and the MS-CG functional can be expressed in terms of an information function, $\Phi(\mathbf{R})$, that discriminates between the ensembles of CG configurations sampled by the atomistic and CG models. The relative entropy corresponds to the average of Φ, while the MS-CG functional corresponds to the average of $|\nabla\Phi|^2$. In addition, we generalize the well-known result of Henderson[69] by identifying conditions for the uniqueness of a set of potentials that reproduce a given set of structural correlation functions. We show that these conditions are closely related to the conditions for the uniqueness of MS-CG potentials. In the course of this analysis, we investigate the mapping entropy and generalize the MS-CG and generalized-Yvon-Born-Green theory for more complex potentials. Finally, we also present numerical calculations that clarify these relationships for incomplete basis sets and particularly simple systems. We demonstrate that relative entropy and MS-CG calculations not only agree for calculations with a complete basis set, but also for calculations with a highly incomplete harmonic basis set expressed in Cartesian coordinates. In contrast, the two methods differ either for higher order Cartesian basis sets or for incomplete basis sets that are functions of non-Cartesian coordinates.

II. PRELIMINARIES

This section defines key quantities for the following analysis. We consider atomistic and CG models for a given system in the canonical ensemble and assume that neither model includes rigid constraints. Lower and upper case symbols correspond to atomistic and CG quantities, respectively. The notation largely follows previous work.[50,58,70]

A. Atomistic model

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

$$p_r(\mathbf{r}) \propto \exp[-u(\mathbf{r})/k_BT]. \tag{1}$$

The nonideal configurational entropy for this model is given by

$$s_{\mathbf{r}} = -k_B \int d\mathbf{r}\; p_r(\mathbf{r}) \ln[V^n p_r(\mathbf{r})]. \tag{2}$$

B. Coarse-grained model

The configuration of the CG model is similarly defined by the Cartesian coordinates, $\mathbf{R}$, for N sites in a volume, V,

that interact according to a potential, $U(\mathbf{R})$. The CG configuration distribution, $P_R(\mathbf{R}|U)$, depends upon the CG potential according to

$$P_R(\mathbf{R}|U) = V^{-N} \exp[-(U(\mathbf{R}) - F[U])/k_BT], \quad (3)$$

where

$$\exp[-F[U]/k_BT] = V^{-N} \int d\mathbf{R} \exp[-U(\mathbf{R})/k_BT] \quad (4)$$

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

The present work considers CG potentials of the following form:

$$U(\mathbf{R}) = \sum_\zeta \sum_\lambda U_\zeta(\psi_\zeta(\{\mathbf{R}\}_\lambda)), \quad (5)$$

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

To emphasize the similarity between CG methods and classical density functional theories, we represent the approximate potential as a sum of terms, each of which is a product of a potential, U_ζ, and a conjugate order parameter or density, $\hat{\rho}_\zeta$:

$$U(\mathbf{R}) = \sum_\zeta \int dx\ U_\zeta(x) \hat{\rho}_\zeta(\mathbf{R}; x), \quad (6)$$

where

$$\hat{\rho}_\zeta(\mathbf{R}; x) = \sum_\lambda \delta(\psi_{\zeta\lambda}(\mathbf{R}) - x), \quad (7)$$

and $\psi_{\zeta\lambda}(\mathbf{R}) \equiv \psi_\zeta(\{\mathbf{R}\}_\lambda)$. Equation (6) assumes that the integrals are evaluated over the appropriate domain for each potential and that these potentials obey suitable smoothness properties and boundary conditions. The density fields defined in Eq. (7) form an incomplete basis that spans a linear vector space of potentials of the form given by Eq. (6). The functions, $U_\zeta(x)$, serve as constant coefficients that identify a potential, U, in this space.

This potential defines a force on each site $I = 1, \ldots, N$ that may be expressed as

$$\mathbf{F}_I(\mathbf{R}) = \sum_\zeta \int dx\ U_\zeta(x) \boldsymbol{\mathcal{G}}_{I;\zeta}(\mathbf{R}; x), \quad (8)$$

where $\boldsymbol{\mathcal{G}}_{I;\zeta}(\mathbf{R}; x) = -\partial \hat{\rho}_\zeta(\mathbf{R}; x)/\partial \mathbf{R}_I$. Equation (8) may be re-expressed as

$$\mathbf{F} = \sum_\zeta \int dx\ U_\zeta(x) \boldsymbol{\mathcal{G}}_\zeta(x), \quad (9)$$

where both $\mathbf{F}$ and $\boldsymbol{\mathcal{G}}_\zeta(x)$ are the vectors in an abstract vector space of force fields[50,51] that define a Cartesian vector force on each site as a function of the CG configuration $\mathbf{R}$. The vectors $\boldsymbol{\mathcal{G}}_\zeta(x)$ that are included in Eq. (9) form an incomplete basis set that spans a subspace of CG force fields. The functions, $U_\zeta(x)$, serve as constant coefficients that identify a force field, $\mathbf{F}$, in this space.

The functional form of U in Eq. (5) is immediately relevant for molecular mechanics-type force fields. Appendix A demonstrates that this treatment can readily be generalized for the case that U_ζ depends upon multiple variables, $x \to \mathbf{x} = \{x_1, x_2, \ldots\}$. Equations (8) and (9) differ from recent analyses[59,60,70] of the MS-CG method by emphasizing the potential functions, $U_\zeta(x)$, rather than the corresponding force functions, $\phi_\zeta(x) = -dU_\zeta(x)/dx$, as the coefficients of force field basis vectors. In the case that each potential function, $\{U_\zeta\}$, depends upon a single scalar variable and that each potential satisfies appropriate boundary conditions, the two representations are equivalent after integrating Eq. (9) by parts. However, in the case that the potential functions depend upon multiple variables, the force functions corresponding to different variables, e.g., $-\partial U_\zeta(\mathbf{x})/\partial x_i$, can no longer be treated independently in variational calculations. The present formulation is also convenient because it emphasizes the similarity between Eqs. (6) and (9). In particular, the densities, $\hat{\rho}_\zeta$, and corresponding force field vectors, $\boldsymbol{\mathcal{G}}_\zeta$, act as basis vectors for structure- and force-motivated approaches, respectively. It should be noted that (1) the present analysis readily generalizes for the case that the potential, U, is defined by a discrete sum over terms by replacing each integral with a corresponding sum over parameters, $U_{\zeta d}$, and also that (2) the continuous relationships considered in this manuscript are directly derived from the limit of such finite discrete sums.

C. Mapping relationships

We define a mapping $\mathbf{M} : \mathbf{r} \to \mathbf{R} = \mathbf{M}(\mathbf{r})$ that maps an atomistic configuration for the system $\mathbf{r}$ onto a CG configuration $\mathbf{R}$ for the same system. This mapping determines the Cartesian coordinates of each site as a linear combination of atomic coordinates that may correspond to, e.g., the center of mass or center of geometry for an atomic group.[50] This mapping determines a probability distribution, p_R, for the atomistic model to sample a configuration $\mathbf{r}$ that maps to a given CG configuration $\mathbf{R}$,

$$p_R(\mathbf{R}) = \langle \delta(\mathbf{M}(\mathbf{r}) - \mathbf{R}) \rangle . \quad (10)$$

In Eq. (10), and throughout this manuscript, angular brackets denote a canonical average over the atomistic configuration space according to the probability distribution $p_r(\mathbf{r})$ defined in Eq. (1).

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

$$s_\mathbf{R} = -k_B \int d\mathbf{R}\ p_R(\mathbf{R}) \ln[V^N p_R(\mathbf{R})]. \quad (11)$$

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

Previous works[50] have considered a CG model to be consistent (in configuration space) with a particular atomistic model for the same system when the two models

sample the same distribution of configurations in the CG configuration space, i.e., when $P_R(\mathbf{R}) = p_R(\mathbf{R})$ for all $\mathbf{R}$. Consequently, the probability distribution p_R defines the appropriate potential for a consistent CG model, U^0, to within an additive constant,

$$U^0(\mathbf{R}) = -k_B T \ln[V^N p_R(\mathbf{R})] + \text{const.} \tag{12}$$

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

$$\mathbf{F}_I^0(\mathbf{R}) = \langle \mathbf{f}_I(\mathbf{r}) \rangle_{\mathbf{R}}, \tag{13}$$

where $\mathbf{f}_I(\mathbf{r})$ is the atomistic force[50] on site I in configuration $\mathbf{r}$ and, for any quantity $a(\mathbf{r})$,

$$\langle a(\mathbf{r}) \rangle_{\mathbf{R}} = \langle a(\mathbf{r}) \delta(\mathbf{M}(\mathbf{r}) - \mathbf{R}) \rangle \,/\, \langle \delta(\mathbf{M}(\mathbf{r}) - \mathbf{R}) \rangle \tag{14}$$

is a canonically weighted average over the set of atomistic configurations that map to $\mathbf{R}$. A CG model that employs the many body mean force (MF) field, $\mathbf{F}^0$, defined by Eq. (13), as a conservative force field will be consistent with the atomistic model in the sense defined above.[50]

The distribution, p_R, determines several key quantities for parameterizing and analyzing CG models. For each order parameter included in Eq. (6), we define a "target" atomistic structural correlation function,

$$p_\zeta(x) = \int d\mathbf{R} p_R(\mathbf{R}) \hat{p}_\zeta(\mathbf{R}; x). \tag{15}$$

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

$$P_\zeta(x|U) = \int d\mathbf{R} P_R(\mathbf{R}|U) \hat{p}_\zeta(\mathbf{R}; x). \tag{16}$$

As is well known,[35,40] P_ζ is a functional derivative of the CG free energy: $P_\zeta(x|U) = \delta F[U]/\delta U_\zeta(x)$.

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

$$\mathbf{F}^{(1)} \odot \mathbf{F}^{(2)} = \int d\mathbf{R} p_R(\mathbf{R}) \frac{1}{3N} \sum_I \mathbf{F}_I^{(1)}(\mathbf{R}) \cdot \mathbf{F}_I^{(2)}(\mathbf{R}), \tag{17}$$

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

III. VARIATIONAL METHODS

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

A. Relative entropy

Shell introduced[41] the relative entropy in terms of the log likelihood for the CG model to reproduce the distribution of *atomistic* configurations sampled by the atomistic model,

$$S_{\text{rel}}[U] = k_B \int d\mathbf{r} p_r(\mathbf{r}) \ln \left[\frac{p_r(\mathbf{r})}{P_r(\mathbf{r}|U)} \right], \tag{18}$$

where $P_r(\mathbf{r}|U)$ is the probability of sampling the atomistic configuration $\mathbf{r}$ from a CG model with potential U. Kullback and Leibler[42] interpreted the quantity

$$\phi(\mathbf{r}|U) = \ln \left[\frac{p_r(\mathbf{r})}{P_r(\mathbf{r}|U)} \right] \tag{19}$$

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

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

$$P_r(\mathbf{r}|U) = \frac{g(\mathbf{r})}{\Omega(\mathbf{M}(\mathbf{r}))} P_R(\mathbf{M}(\mathbf{r})|U), \tag{20}$$

where

$$\Omega(\mathbf{R}) = \int d\mathbf{r}\, g(\mathbf{r})\, \delta(\mathbf{M}(\mathbf{r}) - \mathbf{R}), \tag{21}$$

and $g(\mathbf{r})$ is a weighting function. This definition for $P_r(\mathbf{r}|U)$ is proportional to the probability for the CG model to sample the configuration $\mathbf{M}(\mathbf{r})$ and also corresponds to defining $g(\mathbf{r})\delta(\mathbf{M}(\mathbf{r}) - \mathbf{R})/\Omega(\mathbf{R})$ as the normalized conditional probability of sampling an atomistic configuration $\mathbf{r}$ given a fixed CG configuration $\mathbf{R}$. Given Eq. (20), the relative entropy is re-expressed as

$$S_{\text{rel}}[U] = k_B \int d\mathbf{r}\; p_r(\mathbf{r}) \ln \left[\frac{V^n}{V^N} \frac{p_r(\mathbf{r})}{P_R(\mathbf{M}(\mathbf{r})|U)} \right] + S_{\text{map}}, \tag{22}$$

where

$$S_{\text{map}} = k_B \left\langle \ln \left[\frac{V^N}{V^n} \frac{1}{g(\mathbf{r})} \Omega(\mathbf{M}(\mathbf{r})) \right] \right\rangle. \tag{23}$$

Shell's seminal work quite naturally defined $g(\mathbf{r}) = 1$ so that P_r gives equal weight to all atomistic configurations that map to the same CG configuration.[41] $\Omega(\mathbf{R})$ then corresponds to the total unweighted volume element of atomistic configuration space that maps to $\mathbf{R}$. Given $g = 1$, Eqs. (22) and (23) are equivalent to the earlier results of Shell aside from dimensional constants.[41,43]

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

$$P_r(\mathbf{r}|U) = \frac{p_r(\mathbf{r})}{p_R(\mathbf{M}(\mathbf{r}))} P_R(\mathbf{M}(\mathbf{r})|U). \tag{24}$$

Equation (24) does not give equal weight to all $\mathbf{r}$ that map to the same $\mathbf{R}$, but weights them according to their atomistic Boltzmann weight and has the property that $P_r = p_r$ for a consistent CG model, which is not the case if $g = 1$. Given $g(\mathbf{r}) = p_r(\mathbf{r})$, the entropy of mapping, S_{map}, becomes the difference between the entropy of the atomistic model when viewed from the atomistic and the CG configuration space,

$$S_{\text{map}} = s_{\mathbf{r}} - s_{\mathbf{R}}. \tag{25}$$

This mapping entropy may also be expressed as

$$S_{\text{map}} = k_B \left\langle \ln \left[\frac{V^N}{V^n} \Omega_1(\mathbf{M}(\mathbf{r})) \right] \right\rangle - k_B \left\langle \ln \left[\frac{p_r(\mathbf{r})}{\bar{p}_r(\mathbf{M}(\mathbf{r}))} \right] \right\rangle, \tag{26}$$

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

$$\bar{p}_r(\mathbf{R}) = \int d\mathbf{r}\, p_r(\mathbf{r}) \delta(\mathbf{M}(\mathbf{r}) - \mathbf{R}) / \Omega_1(\mathbf{R}) \tag{27}$$

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

Equation (26) decomposes S_{map} into a geometric component and a component describing the smoothing of the probability distribution. The second term in Eq. (26) quantifies the fluctuations in the Boltzmann weight, p_r, that are smoothed out by integrating over all atomistic configurations that map to the same CG configuration. By the Gibbs inequality, the average is non-negative and only vanishes if $p_r(\mathbf{r}) = \bar{p}_r(\mathbf{M}(\mathbf{r}))$ everywhere, i.e., if and only if all the configurations $\mathbf{r}$ that map to a given CG configuration $\mathbf{R}$ have equal Boltzmann weight in the atomistic model. This places an upper bound upon the entropy difference between the two descriptions of the atomistic model,

$$s_{\mathbf{r}} - s_{\mathbf{R}} \leq k_B \left\langle \ln \left[\frac{V^N}{V^n} \Omega_1(\mathbf{M}(\mathbf{r})) \right] \right\rangle. \tag{28}$$

The volume element $\Omega_1(\mathbf{R})$ can be directly evaluated in terms of the mapping coefficients,[71] although the expression is somewhat complex and periodic boundary conditions may introduce additional complications. Nevertheless, it seems intuitively reasonable that Ω_1 should be independent of configuration for a system with periodic boundary conditions. It also seems clear from geometric arguments that the average in Eq. (28) is non-negative. In particular, the average in Eq. (28) vanishes for any mapping that associates each site with a single atom so that $s_{\mathbf{r}} - s_{\mathbf{R}} \leq 0$, e.g., for an atomically detailed implicit solvent model.

It should be emphasized that the entropy difference, $s_{\mathbf{r}} - s_{\mathbf{R}}$, is a direct consequence of the CG mapping reducing the dimensionality of the configuration space and the associated averaging of p_r. This effect is completely independent of the CG potential. In particular, even if a CG model perfectly reproduces the configuration distribution implied by the atomistic model, the resulting model may not reproduce the configurational entropy of the atomistic model and the entropy difference between the two models will satisfy the inequality, Eq. (28). Consequently, S_{map} may provide a useful quantitative metric for optimizing the CG mapping.

Equation (24) allows for a convenient expression for the relative entropy as an average over the CG configuration space,

$$S_{\text{rel}}[U] = k_B \int d\mathbf{R} p_R(\mathbf{R}) \ln \left[\frac{p_R(\mathbf{R})}{P_R(\mathbf{R}|U)} \right]. \tag{29}$$

The relative entropy has now been expressed in analogy to Eq. (18) and without reference to a mapping entropy. In a very recent paper that considered the relative entropy for modeling CG dynamics, Español and Zúñiga derived Eq. (29) for the relative entropy as the log-likelihood for the atomistic and CG models to sample the same distribution of *CG* configurations.[72] The present analysis clarifies the relationship between these two conventions.

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

$$\frac{\delta}{\delta U_\zeta(x)} S_{\text{rel}}[U] = \Delta P_\zeta(x|U)/T, \tag{30}$$

where

$$\Delta P_\zeta(x|U) = p_\zeta(x) - P_\zeta(x|U) \tag{31}$$

is the difference between the target distribution and the distribution generated by simulations with the potential U. Thus, when considering potentials of the form given by Eq. (5), the relative entropy is minimized when the CG potential reproduces the atomistic distribution functions for the corresponding density operators, i.e., when $P_\zeta(x|U) = p_\zeta(x)$.

As noted earlier by Murtola *et al.*,[45] the IMC method determines the CG potentials that minimize the relative entropy by employing Newton's method to find the parameters for which $\Delta P_\zeta(x|U)$ vanishes. Newton's method leads to a system of coupled equations for iteratively updating the potentials $\{U_\zeta(x)\} \to \{U_\zeta(x) + \delta U_\zeta(x)\}$,

$$\Delta P_\zeta(x|U) = \sum_{\zeta'} \int dx'\; M_{\zeta\zeta'}(x, x'|U) \delta U_{\zeta'}(x'), \tag{32}$$

where

$$\begin{aligned} M_{\zeta\zeta'}(x, x'|U) &\equiv \frac{\delta P_\zeta(x|U)}{\delta U_{\zeta'}(x')} \\ &= -\frac{1}{k_B T} \int d\mathbf{R}\; P_R(\mathbf{R}|U) \\ &\quad \times \delta\hat{\rho}_\zeta(\mathbf{R}; x|U) \delta\hat{\rho}_{\zeta'}(\mathbf{R}; x'|U) \end{aligned} \tag{33}$$

is a susceptibility matrix describing correlated fluctuations of the order parameters in the CG model with $\delta\hat{\rho}_\zeta(\mathbf{R}; x|U) = \hat{\rho}_\zeta(\mathbf{R}; x) - P_\zeta(x|U)$. We emphasize that, because g is independent of U, Shell's earlier work[41,43] leads to the same system of equations for the CG potentials, i.e., Eqs. (30) and (33).

B. Multiscale coarse-graining

The multiscale coarse-graining method[48,49] considers a force-matching[55,56] functional,

$$\chi^2[U] = \frac{1}{3N} \left\langle \sum_I |\mathbf{f}_I(\mathbf{r}) - \mathbf{F}_I(\mathbf{M}(\mathbf{r})|U)|^2 \right\rangle \tag{34}$$

$$= \chi^2[U^0] + ||\mathbf{F}^0 - \mathbf{F}[U]||^2, \tag{35}$$

where $\mathbf{F}$ is the CG force field determined from the approximate potential, U, and $\mathbf{F}^0$ is the force field determined from the many-body PMF, U^0. The second relation follows from Eq. (13).[50] Because the second term is necessarily non-negative, the many-body PMF determines the unique global minimum[57] of χ^2.

Given the incomplete basis set in Eq. (9), the MS-CG variational principle determines the force field spanned by the basis that is "closest" to the many-body mean force field.[50] Varying χ^2 with respect to the corresponding potential parameters leads to

$$\frac{\delta}{\delta U_\zeta(x)}\chi^2[U] = 2\boldsymbol{\mathcal{G}}_\zeta(x) \odot (\mathbf{F}[U] - \mathbf{F}^0). \tag{36}$$

Therefore, minimizing χ^2 determines the approximate force field by geometrically projecting the MF onto the given basis set.[58,59,73,74] After evaluating Eq. (36) at this minimum and expanding $\mathbf{F}$ according to Eq. (9), the MS-CG potential is directly determined from the normal system[51,57] of linear equations,

$$\sum_{\zeta'} \int dx'\, G_{\zeta\zeta'}(x,x') U_{\zeta'}(x') = b_\zeta(x), \tag{37}$$

where $G_{\zeta\zeta'}(x,x') = \boldsymbol{\mathcal{G}}_\zeta(x) \odot \boldsymbol{\mathcal{G}}_{\zeta'}(x')$ is a metric tensor and $b_\zeta(x) = \boldsymbol{\mathcal{G}}_\zeta(x) \odot \mathbf{F}^0$ is the projection of the many-body mean force onto the corresponding basis vector.[58,59] The metric, $G_{\zeta\zeta'}(x,x')$, is a structural correlation function quantifying correlations between the different interactions and plays an analogous role to $M_{\zeta\zeta'}(x,x'|U)$ in Eq. (33). However, $G_{\zeta\zeta'}$ is independent of U and instead determined directly from the atomistic model and CG mapping. The MS-CG method evaluates $b_\zeta(x)$ by employing forces sampled from atomistic simulations,

$$b_\zeta(x) = \frac{1}{3N}\left\langle \sum_I \mathbf{f}_I(\mathbf{r}) \cdot \boldsymbol{\mathcal{G}}_{I;\zeta}(\mathbf{M}(\mathbf{r});x) \right\rangle, \tag{38}$$

which follows from Eq. (13). However, $b_\zeta(x)$ may also be evaluated in terms of structural correlation functions according to a generalized Yvon-Born-Green equation, so that the MS-CG potential may be calculated without explicit knowledge of atomistic forces.[58–61,70]

As noted earlier, the MS-CG method has previously been discussed in terms of force functions: $\phi_\zeta(x) = -dU_\zeta(x)/dx$. If each U_ζ is a function of a single variable and if these functions satisfy appropriate boundary conditions so that each term is uniquely determined from $\phi_\zeta(x)$, then Eq. (37) is equivalent to previous expressions. In this case, Eq. (37) may be derived from prior expressions by differentiation with respect to x and an integration by parts with respect to x'. In particular, $b_\zeta(x) = d\widetilde{b}_\zeta(x)/dx$ and $G_{\zeta\zeta'}(x,x') = \partial^2 \widetilde{G}_{\zeta\zeta'}(x,x')/\partial x \partial x'$, where $\widetilde{b}_\zeta$ and $\widetilde{G}_{\zeta\zeta'}$ are the correlation functions employed in previous analyses.[50,60] However, in contrast to earlier treatments, the present expressions also readily generalize for the case that U_ζ depends upon multiple variables.

IV. SIMILARITIES IN THE VARIATIONAL PRINCIPLES

A. Functionals

In analogy to Eq. (19), we define

$$\Phi(\mathbf{R}|U) = \ln\left[\frac{p_R(\mathbf{R})}{P_R(\mathbf{R}|U)}\right] \tag{39}$$

as the information content in configuration $\mathbf{R}$ for discriminating between the distribution of CG configurations generated by the atomistic and CG models.[42] We note that, when considering variations with an incomplete basis set,

$$\frac{\delta\Phi(\mathbf{R}|U)}{\delta U_\zeta(x)} = \delta\hat{\rho}_\zeta(\mathbf{R};x|U)/k_BT, \tag{40}$$

where $\delta\hat{\rho}_\zeta(\mathbf{R};x|U) = \hat{\rho}_\zeta(\mathbf{R};x) - P_\zeta(x|U)$ as in Eq. (33), while variations with respect to Cartesian coordinates may be expressed as

$$\frac{\partial\Phi(\mathbf{R}|U)}{\partial\mathbf{R}_I} = \left(\mathbf{F}_I^0(\mathbf{R}) - \mathbf{F}_I(\mathbf{R}|U)\right)/k_BT, \tag{41}$$

where $\mathbf{F}_I^0$ is the many body MF defined in Eq. (13).

According to Eq. (29), the relative entropy may be expressed as an average of $\Phi(\mathbf{R}|U)$ evaluated according to p_R,

$$S_{\text{rel}}[U] = k_B \int d\mathbf{R}\, p_R(\mathbf{R})\Phi(\mathbf{R}|U). \tag{42}$$

Moreover, as a direct consequence of Eqs. (35) and (41), the MS-CG functional may also be expressed in terms of $\Phi(\mathbf{R}|U)$,

$$\chi^2[U] = \frac{(k_BT)^2}{3N}\int d\mathbf{R}\, p_R(\mathbf{R})\,|\nabla\Phi(\mathbf{R}|U)|^2 + \chi^2[U^0], \tag{43}$$

where $|\nabla A(\mathbf{R})|^2 = \sum_I (\partial A(\mathbf{R})/\partial\mathbf{R}_I)^2$. The second term in Eq. (43) depends only upon the atomistic model and the CG mapping, i.e., it is independent of U, and does not influence the variational calculation of the MS-CG potential.

The parallel between Eqs. (42) and (43) is a central result of the present work. The relative entropy formalism determines the CG potential that minimizes the average of Φ; the MS-CG formalism determines the CG potential that minimizes the average of $|\nabla\Phi|^2$. Although Φ may be either positive or negative, the Gibbs inequality implies that $S_{\text{rel}} \geq 0$ and vanishes if and only if U is the PMF, for which $\Phi(\mathbf{R}|U^0) = 0$. Similarly, χ^2 is also non-negative and is minimized by the PMF for which $|\nabla\Phi(\mathbf{R}|U^0)|^2 = 0$. $\chi^2[U^0]$ and S_{map} serve somewhat analogous roles in the MS-CG and relative entropy approaches, respectively. While $\chi^2[U^0]$ quantifies fluctuations in the atomistic force field for a given $\mathbf{R}$, according to Eq. (26), S_{map} quantifies the fluctuations in the atomistic probability distribution for a given $\mathbf{R}$. The parallel between the relative entropy and MS-CG formalisms demonstrated in Eqs. (42) and (43) is independent of the functional form for the CG potential.

Minimizing S_{rel} with respect to terms, $U_\zeta(x)$, in the approximate CG potential defined by Eq. (5), leads to the condition

$$\int d\mathbf{R}\, p_R(\mathbf{R})\delta\hat{\rho}_\zeta(\mathbf{R};x|U) = 0, \tag{44}$$

which is equivalent to the condition of Eq. (30), i.e., the relative entropy is minimized with respect to a set of potentials when the conjugate distributions generated by the CG model match those determined by the atomistic model. This is a self-consistent condition because it requires knowledge of $P_\zeta(x|U)$, which depends upon the potential-dependent normalization, $\exp[-F[U]/k_BT]$, and can only be determined from simulations with the CG potential U.

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

$$\int d\mathbf{R} p_R(\mathbf{R}) (\nabla\Phi(\mathbf{R}|U) \cdot \nabla) \delta\hat{\rho}_\zeta(\mathbf{R}; x|U) = 0. \quad (45)$$

As a result of Eq. (41) and the relationship between $\hat{\rho}_\zeta(\mathbf{R}; x)$ and $\mathcal{G}_\zeta(x)$, Eq. (45) is equivalent to Eq. (37). Equation (45) can be solved directly. Because $P_\zeta(x|U)$ is a configuration-independent average, the gradient operator eliminates the potential dependence from $\delta\hat{\rho}_\zeta(\mathbf{R}; x|U)$. Clearly, the distinction between the two variational principles hinges upon the operator $\nabla\Phi(\mathbf{R}|U) \cdot \nabla$, which trivially vanishes when U is the many-body PMF. The parallel between the relative entropy and MS-CG formalisms demonstrated in Eqs. (44) and (45) can be generalized for a wide range of CG potentials, including potentials that depend upon multiple order parameters, as in Appendix A, or for potentials that may be expressed as a linear combination of terms with discrete weights.

B. Uniqueness

Having investigated parallels in the variational principles underlying structure-based and force-based CG approaches, we now investigate their respective uniqueness properties. Henderson[69] followed the earlier analyses of Hohenberg, Kohn,[75] and Mermin[76] to demonstrate that, in the case of simple liquids, the pair potential reproducing a given radial distribution function is unique. Johnson *et al.* subsequently extended this result for liquids with site-site pair potentials.[77] The present analysis further extends this result, analyzes its generality, and also observes remarkable similarities with the conditions necessary for the uniqueness of MS-CG potentials.

1. Structure-based uniqueness

We consider two CG potentials U_A and U_B that may be expressed in terms of a set of potentials $\{U_{\zeta A}(x)\}$ and $\{U_{\zeta B}(x)\}$ according to Eq. (6). We shall assume that both sets of potentials satisfy suitable boundary conditions, e.g., that non-bonded potentials vanish at long distances and that bonded potentials vanish at their minima. These two sets of potentials determine two sets of structural correlation functions, $\{P_{\zeta A}(x) = P_\zeta(x|U_A)\}$ and $\{P_{\zeta B}(x) = P_\zeta(x|U_B)\}$, that are defined by Eq. (16).

Equation (6) expressed the CG potential as a linear combination of density fields, $\{\hat{\rho}_\zeta(\mathbf{R}; x)\}$. We define a set of fields as **linearly independent**, if

$$\sum_\zeta \int dx c_\zeta(x) \hat{\rho}_\zeta(\mathbf{R}; x) = 0 \quad (46)$$

for all $\mathbf{R}$ implies that $c_\zeta(x) = 0$ for all ζ and x. Conversely, a set of fields is linearly dependent if any field in the set is a linear combination of the other fields. If the set of fields in Eq. (6) is linearly dependent, then U no longer determines a unique set of coefficients, $\{U_\zeta(x)\}$. It should also be noted that unless suitable boundary conditions are required, the potentials are only unique to within an additive constant.

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

2. Force-based uniqueness

In contrast to the relative entropy and related iterative variational methods, the MS-CG approach is not guaranteed to reproduce any particular structural correlation functions. The MS-CG method directly projects the many-body MF onto a set of vectors employed as an incomplete basis for the CG force field according to Eq. (9). This geometric interpretation implies that the MS-CG potential is unique as long as the vectors included in the incomplete basis are linearly independent in the space of force fields. A set of force field vectors $\{\mathcal{G}_\zeta(x)\}$ is **linearly independent**, if

$$\sum_\zeta \int dx c_\zeta(x) \mathcal{G}_\zeta(\mathbf{R}; x) = 0 \quad (47)$$

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

The force field basis vectors are obtained as gradients of the corresponding density fields according to Eq. (8). Consequently, the linear independence of the MS-CG basis set implies the linear independence of the corresponding set of density fields. Therefore, analysis of the linear independence of the force field basis set via $G_{\zeta\zeta'}$ may prove to be a useful and numerically tractable method for assessing the uniqueness of potentials obtained from relative entropy and other iterative variational methods.

V. NUMERICAL RESULTS

This section considers several numerical examples to further investigate the relationship between force- and structure-based CG approaches. In order to facilitate visual analysis of Φ, we consider the case where the CG potential and the corresponding PMF, U^0, are functions of a single variable. We consider first the case that the variable is a Cartesian coordinate. However, calculations performed for the case that the approximate potential depends upon a curvilinear coordinate (i.e., a distance between two particles) lead to quantitatively similar results. Therefore, despite the simplicity of these

examples, we expect that the resulting insight should generalize to CG models for more complex molecular systems, in which case Φ can no longer be easily visualized. Without loss of generality, we set $k_BT = 1$ in these calculations.

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

$$U(x) = \sum_{d=1}^{m} U_d x^d, \tag{48}$$

where the potential basis functions are $\rho_d(x) = x^d$, and U_d is the corresponding parameter. Because the $d = 0$ term is not considered and because the ρ_d vanish at the origin, the parameters will be unique for both force- and structure-based CG methods. According to Eq. (30), the relative entropy functional will be minimized with respect to U_d when the CG model reproduces the dth moment of the atomistic distribution for x. In contrast, the parameters that minimize the MS-CG functional satisfy the discrete analog of the normal equations (Eq. (37)). In general, the resulting MS-CG model is not guaranteed to reproduce any moments of the atomistic distribution. Consequently, we expect that the force- and structure-based models will differ for this incomplete basis set.

A. Cartesian coordinates

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

$$p(x) = A_1 \exp\left[-(x - x_1)^2/2\sigma_1^2\right] + A_2 \exp\left[-(x - x_2)^2/2\sigma_2^2\right], \tag{49}$$

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

1. Harmonic approximation

We first consider a harmonic approximation to U^0, i.e., $m = 2$ in Eq. (48). In this case, the parameters for a structure-based model can directly be determined by the standard relations for the mean and variance of a Gaussian distribution: $\langle x \rangle = -U_1/2U_2$ and $\langle (x - \langle x \rangle)^2 \rangle = 1/2U_2$, where the angular brackets denote averages according to the atomistic distribution, $p(x)$. The parameters for the force-based model are determined by solving the normal MS-CG equations. Remarkably, for this case, the MS-CG potential also reproduces the mean and variance of the atomistic distribution. More generally, if the approximate potential is expressed as a quadratic function of Cartesian coordinates (in any number of dimensions), the MS-CG normal equations determine the parameters appropriate for reproducing the means and covariances determined by the atomistic model. Consequently, as long as the approximate CG potential is a quadratic function of Cartesian coordinates, the relative entropy and MS-CG methods determine identical potentials. To the best of our knowledge, this corresponds to the first reported instance where force- and structure-based approaches yielded identical results for an incomplete basis set.

Figure 1 presents an example for which the atomistic distribution, $p(x)$, is a bimodal distribution with overlapping Gaussian peaks of similar variance. The solid black curve in Fig. 1(a) corresponds to the atomistic distribution. The dotted black curve corresponds to $P_2(x)$, the distribution for the CG model determined by both force- and structure-based approaches when $m = 2$. Table I compares the first ten moments for $p(x)$ and $P_2(x)$. The dotted black curve in Fig. 1(b) presents the corresponding information function $\Phi_2(x) = \ln p(x)/P_2(x)$. Because P_2 smooths over the bimodal features of p, Φ_2 is relatively small near the center of the distribution, has stationary points near the stationary points of p, and becomes increasingly negative in the wings of the distribution. Although Φ_2 assumes negative values, its average with respect to p is the relative entropy, which is necessarily positive and equals $S_{\text{rel}} = 7.0949 \times 10^{-2}$ for P_2. The dotted black curve in Fig. 1(c) presents $|d\Phi_2(x)/dx|^2$. The average of this quantity with respect to p determines $\chi^2 = 7.2927 \times 10^{-1}$. (For simplicity, the factor $3N$ is set to 1 and the contribution $\chi^2[U^0]$ set to 0 in this calculation and in the remainder of the section, since neither impacts the calculated potential or resulting distribution.) Figure 1(d) directly compares the "match" between the resulting harmonic force field (dotted black curve) with the anharmonic "many-body" mean force field, $F^0(x) = -dU^0(x)/dx$ (solid black curve). The dotted black curve in Fig. 1(c), which was calculated directly from the derivative of the information function, is the square of the difference between these two force fields. Visual inspection confirms that the dotted black curve in Fig. 1(d) provides the best linear approximation to the many-body PMF when the weighting is performed with respect to p.

2. Anharmonic approximation

We next consider an anharmonic approximation, i.e., $m = 4$, to U^0. The potential parameters determined by structure-based approaches will reproduce the first four moments of $p(x)$. These parameters can no longer be computed analytically and were instead iteratively determined by implementing the analog of Eq. (32) in MATHEMATICA.[78] As before, the potential parameters for the MS-CG model were directly determined by the normal MS-CG equations.

In this case, the force- and structure-based CG approaches no longer determine identical potentials. Figure 1(a) compares the resulting CG distributions with the underlying atomistic distribution, $p(x)$. In Fig. 1(a) and in the following

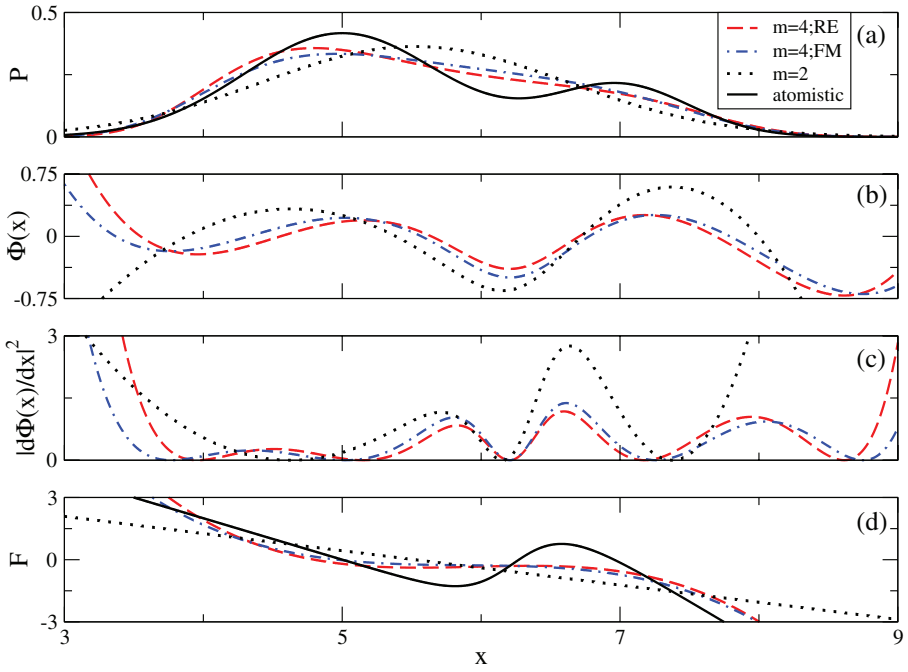

FIG. 1. Analysis of distributions for structure- and force-based models with approximate potentials corresponding to polynomials of order $m = 2$ and $m = 4$. The solid black curves correspond to the underlying high resolution distribution, $p(x)$, that is defined by Eq. (49) with $\sigma_1^2 = 0.5$, $\sigma_2^2 = 0.25$, and $A_1/A_2 = 2$. The dotted black curves correspond to results for CG models with harmonic approximate potentials. The dashed red and dashed-dotted blue curves correspond to the structure-based and force-based models, respectively, using approximate CG potentials with $m = 4$. Panel (a) compares the atomistic and CG distribution functions; panel (b) presents the corresponding information functions, $\Phi(x)$; panel (c) presents its squared gradient, $|d\Phi(x)/dx|^2$; panel (d) presents the corresponding force functions, $-dU(x)/dx$.

figures, the dashed red and dashed-dotted blue curves correspond to structure- and force-based models, respectively, using approximate potentials with $m = 4$. Figures 1(b) and 1(c) compare the corresponding information functions, Φ_4, and the (squared) magnitude of its derivative, $|d\Phi_4/dx|^2$.

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

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

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

TABLE I. Comparison of distributions for the CG coordinates determined from the high resolution (AA) model and also from CG models determined by force- (FM) and structure- (RE) based approaches using approximate potentials with $m = 2$ and $m = 4$ in Eq. (48). The first ten rows present the first ten moments for the high resolution distribution and also the percent error in the corresponding moments for each CG model. The last three rows present χ^2 (with $3N = 1$ and $\chi^2[U^0] = 0$), S_{rel}, and the entropy of each model ($S = -\langle \ln P \rangle$).

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

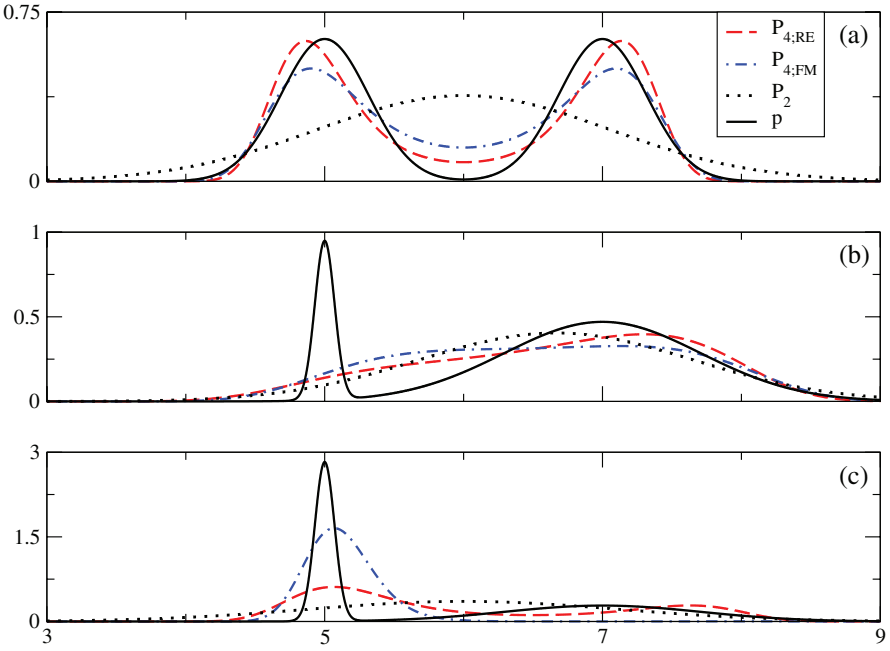

FIG. 2. Comparison of distributions for a high resolution model (defined by Eq. (49)) and for CG models determined by minimizing either S_{rel} (structure-based models) or χ^2 (force-based models). The solid black and dotted black curves correspond to the underlying atomistic distribution, $p(x)$, and to distributions for CG models with harmonic potentials, $P_2(x)$, respectively. The dashed red and dashed-dotted blue curves correspond to structure-based and force-based CG models, respectively, using approximate potentials corresponding to polynomials of order $m = 4$. The three panels correspond to different parameter sets for $p(x)$ – panel (a): $\sigma_1^2 = 0.1$, $\sigma_2^2 = 0.1$, $A_1/A_2 = 1$; panel (b): $\sigma_1^2 = 0.01$, $\sigma_2^2 = 1$, $A_1/A_2 = 2$; panel (c): $\sigma_1^2 = 0.01$; $\sigma_2^2 = 1$; $A_1/A_2 = 10$.

overlapping peaks of similar variance. In this case, the difference between the structure and force-based approaches is subtle, but can be understood in terms of the variational principles underlying the two approaches. However, the resulting models can deviate more if the peaks in the underlying atomistic distribution function no longer overlap. Figure 2 presents three examples for this case. As in Figure 1, the solid black curve corresponds to the atomistic distribution and the dotted black curve corresponds to the distribution obtained from both force- and structure-based models when $m = 2$. The dashed red and dashed-dotted blue curves correspond to distributions obtained from structure- and force-based approaches, respectively, when $m = 4$.

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

Figures 2(b) and 2(c) consider cases for which the atomistic distribution consists of non-overlapping peaks with different heights and variance. In these cases, neither approach reproduces the atomistic distribution very well. In Fig. 2(b), the first peak of $p(x)$ is ten times more narrow and twice as tall as the second peak. Both approaches determine distributions that span both peaks in $p(x)$ and emphasize the second wider peak. However, because of the large derivative associated with the more narrow peak, the force-based model weights the first narrow peak slightly more. The force-based model more accurately reproduces the tails of the atomistic distribution, but the structure-based model more accurately reproduces the second wider peak.

In Fig. 2(c), the first peak of $p(x)$ is ten times more narrow and also ten times larger than the second peak. Because of the large slope of the first peak, the force-based model shifts even more towards this peak and completely disregards the majority of the second peak. In contrast, in order to reproduce the low order moments of the atomistic distribution, the structure-based model more accurately describes the second peak and, consequently, models the larger peak less accurately.

B. Curvilinear coordinates

In addition, we also performed numerical calculations for the case that the approximate CG potential depends upon the distance between a pair of particles. In this case, force- and structure-based CG approaches no longer give identical results for the case that the potential is a harmonic function of distance. However, the differences between the resulting harmonic potentials are numerically insignificant for the cases that we considered. Moreover, we also performed calculations corresponding to Fig. 2 for the case that the approximate potential is an anharmonic function of interparticle distance. The results are virtually identical to those presented in Fig. 2 for the case that the CG potential is a function of Cartesian coordinates.

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

VI. DISCUSSION

CG models provide a highly efficient means for investigating complex processes that cannot be effectively studied with more detailed models. These models are frequently parameterized to reproduce structural properties that are determined from either experiments or from high resolution simulations. Structure-based approaches, such as the relative entropy method,[41,43] employ multiple CG simulations to reproduce target structural correlation functions. In contrast, the MS-CG method[48–50,54] employs forces sampled from atomistic simulations to directly project the many-body MF onto a given basis set for the CG force field, but the resulting force field is not guaranteed to reproduce any particular target correlation functions.[50] If the CG potential is represented with a complete basis set, then both approaches determine the same potential, i.e., the many-body PMF, and will quantitatively reproduce all structural features of the atomistic model.[12] In practice, though, the CG potential is represented with a highly incomplete basis set, typically of a molecular mechanics form. In this case, the MS-CG and relative entropy methods will generally determine different approximations to the many-body PMF.[67]

Despite their obvious differences, the present work reveals remarkable similarities in the framework underlying force- and structure-based approaches to CG modeling. Both the MS-CG and relative entropy approaches determine CG potentials through variational calculations that are typically performed in a linear space spanned by a highly incomplete basis set. The force field basis vectors employed in the MS-CG approach correspond to gradients of the potential basis functions employed in the relative entropy approach. Most significantly, the functionals employed in the two variational calculations can both be expressed in terms of an information function, $\Phi(\mathbf{R})$, that discriminates between the distributions of configurations sampled by the atomistic and CG models. Shell originally defined an equivalent relative entropy as the average of an analogous information function defined on the atomistic configuration space.[41] The present work demonstrates that the MS-CG method minimizes the average $|\nabla\Phi(\mathbf{R})|^2$.

The present work generalizes the well-known result of Henderson[69] by identifying conditions for the uniqueness of CG potentials that reproduce a set of low order correlation functions. The potentials obtained from the structure- and force-based methods can only be unique if the corresponding basis vectors are linearly independent. The linear independence of basis functions for structure-based methods may be relatively difficult to test *a priori*. However, the linear independence of the MS-CG basis vectors may be directly tested by the metric tensor $G_{\zeta\zeta'}$, which can be readily calculated from sampled atomistic configurations.[51] Moreover, the linear independence of the MS-CG basis vectors implies the linear independence of the corresponding basis functions for structure-based potentials.

The present numerical calculations further probe the relationship between force- and structure-based methods. These calculations considered particularly simple models for which the atomistic distribution can be readily visualized and the consequences of an incomplete basis set investigated. Our calculations and analysis demonstrate that force- and structure-based approaches determine the same potential, not only in the limit of a complete basis set, but also in the case that the potential is a quadratic function of Cartesian coordinates. However, the methods may obtain different results if the basis set is expanded to include higher order polynomials or if the approximate potential depends upon non-Cartesian coordinates, e.g., interparticle distances. Given a basis set described by mth-order polynomials, the relative entropy method determines the parameters that quantitatively reproduce the first m moments of the atomistic distribution. Neither method is guaranteed to reproduce any higher order moments. Our results suggest that, depending upon the atomistic distribution, either the MS-CG or the relative entropy method may provide a more accurate description of these higher order moments. The MS-CG approach appears to more accurately reproduce the wings and rapidly varying regions of the atomistic distribution, but may be less sensitive to more slowly varying regions of the distribution. Nevertheless, our numerical investigations suggest that the two approaches lead to similar results for a wide range of atomistic distributions.

We emphasize that these numerical examples considered a particularly simple case for which the "many-body" PMF corresponded to a function of a single variable. In this case, we considered the effect of an incomplete basis set by approximating the PMF with low order polynomials that neglected higher order modes. In CG models for more complex molecular systems, however, the many-body PMF is generally a highly complex function of the coordinates for all the CG sites. In this case, the many-body PMF must be approximated by an incomplete basis set that is typically defined by relatively simple molecular mechanics potentials. These approximate potentials couple nonbonded pairs of particles with central pair potentials and neglect many-body interactions. (Although, we note that these effective pair potentials approximately incorporate the effects of these many-body correlations.[57,79]) Consequently, the physical significance of the missing basis vectors is very different in these two cases. Nevertheless, we anticipate that the principal conclusions from this study should hold quite generally for CG models of more complex systems. Both theoretical analysis and simple numerical calculations indicate that the relative entropy approach determines the CG potential that minimizes

the average of the information function, Φ, that discriminates between the distribution of configurations sampled by atomistic and CG models, while the MS-CG approach determines the CG potential that minimizes the average of its square gradient, $|\nabla\Phi|^2$. This suggests that the relative entropy approach most accurately describes the slowly varying regions of the many-body PMF, while the MS-CG approach may be most accurate in describing the rapidly varying regions of the PMF. Weeks-Chandler-Andersen theory emphasizes that local structural properties are most sensitive to the rapidly varying short-range part of the potential.[80] Consequently, this observation may partially explain the remarkable capability of the MS-CG method to accurately reproduce structural properties of an atomistic model without requiring iterative refinement of the CG potential.

The present work also addressed several other aspects of force- and structure-based CG approaches. We reformulated and generalized the MS-CG and g-YBG approaches for more complex potentials that are functions of multiple order parameters. We demonstrated the equivalence of two definitions for the relative entropy that have been expressed as averages over the atomistic[41] or CG[72] configuration space. This analysis suggested an intriguing physical significance for the mapping entropy as the difference in the configurational entropy of the atomistic model when viewed from either the atomistic or CG configuration space. This entropy difference has a rigorous upper bound that is likely non-negative and this bound vanishes for the case that each site corresponds to a single atom.

We decomposed this entropy difference into two contributions. One contribution describes the geometry of the CG mapping and may be calculated analytically as a function of mapping coefficients. The second contribution results from the averaging of fluctuations in the atomistic distribution due to integrating out degrees of freedom. This second contribution is non-negative and only vanishes when all the configurations that map to the same CG configuration have equal weight in the atomistic model. We emphasize that this entropy difference is independent of the CG potential and applies, in particular, for a perfectly consistent CG model, i.e., one for which $p_R = P_R$. Because of its significance for transferability,[66,81] representability,[79] and phase transitions, the mapping entropy may be a useful metric for optimizing CG mappings.

As mentioned above, the present work generalized the noted uniqueness result of Henderson[69] by identifying conditions for the uniqueness of structure-derived potentials. However, this result assumes the existence of CG potentials reproducing a set of distribution functions. It is much more difficult to prove this existence and relatively little progress has been achieved in these directions.[46,47] Furthermore, although the present result is exact, it provides little immediate insight into the sensitivity of the potentials to relatively small changes in the corresponding distribution functions.[82] A number of studies have demonstrated that CG distribution functions are quantitatively similar for a wide range of potentials that differ in, e.g., the long-ranged tail of nonbonded potentials, and this insensitivity has been exploited to optimize thermodynamic properties.[33,83] We also note that the present work specifically addresses the canonical ensemble and does not directly address other ensembles or issues of transferability. Finally, we reiterate that the continuous results presented in this work were derived as the continuum limit of discrete results.

VII. CONCLUDING REMARKS

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

The present work also indicates several directions for future study. The intriguing duality between the relative entropy and MS-CG formalisms is suggestive of scalar potential and vector field formulations for the statistical thermodynamics of CG modeling. Further, this relationship suggests that the relative entropy and MS-CG functionals might be used to define the first two terms in a more general functional for parameterizing CG potentials. Finally, it suggests the possibility for broader application of the MS-CG formalism. The relative entropy formalism is closely connected to log-likelihood methods that are quite widely employed for model optimization in a range of fields.[41] The present analysis suggests that it may also be possible to extend and relate the MS-CG formalism for model optimization in fields other than particle-based modeling.

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

ACKNOWLEDGMENTS

The present work has been financially supported by a CAREER award (Grant No. MCB 1053970) from the National Science Foundation and also by start-up funds from the Pennsylvania State University. J.F.R. was also partially supported by The Pennsylvania State University Academic

Computing Fellowship. Numerical calculations were performed using support and resources from Research Computing and Cyberinfrastructure, a unit of Information Technology Services at Penn State. W.G.N. gratefully acknowledges Dr. Irina Navrotskaya for her early contributions to understanding the uniqueness of structure-based potentials, Professor Maria Isabel Bueno Cachadina for directing us to Sylvester's equation, Professor Qiang Du, Professor Markus Deserno, Professor Scott Milner, Professor Dave Allara, and Professor M. Scott Shell for helpful conversations, and also Dr. Irina Navrotskaya and Dr. Ard Louis for helpful comments on this manuscript.

APPENDIX A: MORE GENERAL POTENTIALS

Recent analyses of the MS-CG and g-YBG approaches have treated χ^2 as a functional of CG force functions.[60,70] While this treatment is appropriate when the terms in the CG potential depend upon a single scalar variable, it is not adequate for the case that the potential includes terms depending upon two or more independent variables. If a term in the CG potential depends upon multiple independent variables, e.g., an angle and a bond distance, then partial derivatives with respect to these variables lead to distinct force functions that can no longer be varied independently. In this case, it is more convenient to perform variational calculations with the CG potential functions as the independent variables. This appendix briefly demonstrates how the MS-CG and g-YBG formalisms may be readily generalized for potential functions that depend upon multiple variables. For convenience, we assume that the CG potential includes only one type of interaction. Distinct types of interactions can be readily treated by introducing an additional subscript (ζ) to distinguish them as in Eq. (5). We note that Voth and coworkers have previously applied the MS-CG method to potentials depending upon the coordinates of three non-bonded particles.[84]

We consider the following potential:

$$V(\mathbf{R}) = \sum_{\lambda} U(\boldsymbol{\psi}(\{\mathbf{R}\}_\lambda)) = \int d\boldsymbol{x}\; U(\boldsymbol{x})\hat{\rho}(\mathbf{R};\boldsymbol{x}), \quad \text{(A1)}$$

where

$$\hat{\rho}(\mathbf{R};\boldsymbol{x}) = \sum_{\lambda} \delta(\boldsymbol{\psi}_\lambda(\mathbf{R}) - \boldsymbol{x}), \quad \text{(A2)}$$

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

$$\mathbf{F} = \int d\boldsymbol{x}\; U(\boldsymbol{x})\; \boldsymbol{\mathcal{G}}(\boldsymbol{x}), \quad \text{(A3)}$$

where the basis vectors have elements

$$\boldsymbol{\mathcal{G}}_I(\mathbf{R};\boldsymbol{x}) = -\nabla_I \hat{\rho}(\mathbf{R};\boldsymbol{x}) \quad \text{(A4)}$$

$$= \sum_{\lambda}\sum_{\alpha} (\nabla_I \psi_{\lambda\alpha}(\mathbf{R}))\, \frac{\partial}{\partial x_\alpha} \delta(\boldsymbol{\psi}_\lambda(\mathbf{R}) - \boldsymbol{x}), \quad \text{(A5)}$$

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

$$b(\boldsymbol{x}) = \int d\boldsymbol{x}' G(\boldsymbol{x},\boldsymbol{x}') U(\boldsymbol{x}'), \quad \text{(A6)}$$

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

$$b(\boldsymbol{x}) = \frac{1}{3N}\sum_{\alpha} \frac{\partial}{\partial x_\alpha} \int d\mathbf{r}\, p_r(\mathbf{r}) \sum_{\lambda}\sum_{I} (\mathbf{f}_I(\mathbf{r}) \cdot \nabla_I \psi_{\lambda\alpha}(\mathbf{M}(\mathbf{r}))) \times \delta(\boldsymbol{\psi}_\lambda(\mathbf{M}(\mathbf{r})) - \boldsymbol{x}), \quad \text{(A7)}$$

$$G(\boldsymbol{x},\boldsymbol{x}') = \frac{1}{3N}\sum_{\alpha}\sum_{\alpha'} \frac{\partial^2}{\partial x_\alpha \partial x'_{\alpha'}} \int d\mathbf{R}\; p_R(\mathbf{R}) \times \sum_{\lambda}\sum_{\lambda'} (\nabla\psi_{\lambda\alpha}(\mathbf{R}) \cdot \nabla\psi_{\lambda'\alpha'}(\mathbf{R}))\delta(\boldsymbol{\psi}_\lambda(\mathbf{R}) - \boldsymbol{x}) \times \delta(\boldsymbol{\psi}_{\lambda'}(\mathbf{R}) - \boldsymbol{x}'), \quad \text{(A8)}$$

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

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

$$b(\boldsymbol{x}) = \frac{k_B T}{3N} \int d\mathbf{R}\; p_R(\mathbf{R}) \sum_I \left(-\frac{\partial}{\partial \mathbf{R}_I} \cdot \boldsymbol{\mathcal{G}}_I(\mathbf{R};\boldsymbol{x}) \right) \quad \text{(A9)}$$

$$= k_B T[\bar{g}(\boldsymbol{x}) - L(\boldsymbol{x})], \quad \text{(A10)}$$

in terms of

$$\bar{g}(\boldsymbol{x}) = \frac{1}{3N}\sum_{\alpha,\alpha'} \frac{\partial^2}{\partial x_\alpha \partial x_{\alpha'}} \left\langle \sum_{\lambda} (\nabla\psi_{\lambda\alpha}(\mathbf{M}(\mathbf{r})) \cdot \nabla\psi_{\lambda\alpha'}(\mathbf{M}(\mathbf{r}))) \times \delta(\boldsymbol{\psi}_\lambda(\mathbf{M}(\mathbf{r})) - \boldsymbol{x}) \right\rangle, \quad \text{(A11)}$$

$$L(\boldsymbol{x}) = \frac{1}{3N}\sum_{\alpha} \frac{\partial}{\partial x_\alpha} \left\langle \sum_{\lambda} \nabla^2\psi_{\lambda\alpha}(\mathbf{M}(\mathbf{r}))\delta(\boldsymbol{\psi}_\lambda(\mathbf{M}(\mathbf{r})) - \boldsymbol{x}) \right\rangle. \quad \text{(A12)}$$

Equations (A6) and (A10) determine a g-YBG equation as in previous work.[58,60,70]

APPENDIX B: PROOF OF UNIQUENESS FOR STRUCTURE-BASED POTENTIALS

We consider two CG potentials U_A and U_B expressed in terms of a set of potentials $\{U_{\zeta A}(x)\}$ and $\{U_{\zeta B}(x)\}$ (that satisfy suitable boundary conditions and that correspond to linearly independent fields) and corresponding distribution functions $\{P_{\zeta A}(x)\}$ and $\{P_{\zeta B}(x)\}$. In Subsection IV B 1, we asserted that if these sets of distribution functions are equal, then the two sets of potentials are equal. In this appendix, we shall prove this result by proving its contrapositive, i.e., that if the two

potentials differ in any term, $U_\zeta(x)$, then the resulting distribution functions also differ.

The Gibbs inequality[85] implies that

$$\int d\mathbf{R} P_R(\mathbf{R}|U_A)\,(U_A(\mathbf{R}) - U_B(\mathbf{R})) \leq F[U_A] - F[U_B], \tag{B1}$$

where $F[U]$ is the configurational free energy defined in Eq. (4). Adding this result to the symmetric inequality that is obtained from averaging with respect $P_R(\mathbf{R}|U_B)$ leads to

$$\Psi = \sum_\zeta \int dx \Delta U_\zeta(x) \Delta P_\zeta(x) \leq 0, \tag{B2}$$

where $\Delta U_\zeta(x) = U_{\zeta A}(x) - U_{\zeta B}(x)$ and $\Delta P_\zeta(x) = P_{\zeta A}(x) - P_{\zeta B}(x)$. The equality is obtained in Eqs. (B1) and (B2), if and only if $P_R(\mathbf{R}|U_A) = P_R(\mathbf{R}|U_B)$ for all $\mathbf{R}$.

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

Thus, if the two potentials differ in some particular term, then the resulting distribution functions differ. The contrapositive must also hold true, which proves the assertion in Subsection IV B 1: Assuming suitable boundary conditions and the linear dependence of the density fields included in the approximate potential, the set of potentials that reproduce a given set of conjugate correlation functions is unique.

[1] M. Shirts and V. S. Pande, Science **290**, 1903 (2000).
[2] J. Mervis, Science **293**, 1235 (2001).
[3] N. Okimoto, N. Futatsugi, H. Fuji, A. Suenaga, G. Morimoto, R. Yanai, Y. Ohno, T. Narumi, and M. Taiji, PLOS Comput. Biol. **5**, e1000528 (10 2009).
[4] D. E. Shaw, P. Maragakis, K. Lindorff-Larsen, S. Piana, R. O. Dror, M. P. Eastwood, J. A. Bank, J. M. Jumper, J. K. Salmon, Y. B. Shan, and W. Wriggers, Science **330**, 341 (2010).
[5] D. A. Case, I. T. E. Cheatham, T. Darden, H. Gohlke, R. Luo, J. K. M. Merz, A. Onufriev, C. Simmerling, B. Wang, and R. Woods, J. Comput. Chem. **26**, 1668 (2005).
[6] J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kale, and K. Schulten, J. Comput. Chem. **26**, 1781 (2005).
[7] B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl, J. Chem. Theory Comput. **4**, 435 (2008).
[8] M. R. Shirts and J. D. Chodera, J. Chem. Phys. **129**, 124105 (2008).
[9] F. Noe, C. Schutte, E. Vanden-Eijnden, L. Reich, and T. R. Weikl, Proc. Natl. Acad. Sci. U.S.A. **106**, 19011 (2009).
[10] G. R. Bowman, V. A. Voelz, and V. S. Pande, J. Am. Chem. Soc. **133**, 664 (2011).
[11] *Coarse-Graining of Condensed Phase and Biomolecular Systems*, edited by G. A. Voth (CRC, Boca Raton, FL, 2009).
[12] T. Murtola, A. Bunker, I. Vattulainen, M. Deserno, and M. Karttunen, Phys. Chem. Chem. Phys. **11**, 1869 (2009).
[13] C. Peter and K. Kremer, Faraday Discuss. **144**, 9 (2010).
[14] G. A. Voth, J. Chem. Theory Comput. **2**(3), 463 (2006), introduction to partial special issue on "Coarse-Graining in Molecular Modeling and Simulation."
[15] Themed issue, "Coarse-Grained Modeling of Soft Condensed Matter," Phys. Chem. Chem. Phys. **11**(12), 1853 (2009).
[16] Themed issue, "Modeling Soft Matter," Soft Matter **2**(22), 4341 (2009).
[17] C. Peter and K. Kremer, "Multiscale Simulations of Soft Matter Systems," Faraday Discuss. **144**, 9 (2010).
[18] J. C. Shelley, M. Y. Shelley, R. C. Reeder, S. Bandyopadhyay, and M. L. Klein, J. Phys. Chem. B **105**, 4464 (2001).
[19] S. O. Nielsen, C. F. Lopez, G. Srinivas, and M. L. Klein, J. Chem. Phys. **119**, 7043 (2003).
[20] W. Shinoda, R. Devane, and M. L. Klein, Mol. Simul. **33**, 27 (2007).
[21] S. J. Marrink, A. H. de Vries, and A. E. Mark, J. Phys. Chem. B **108**, 750 (2004).
[22] S. J. Marrink, H. J. Risselada, S. Yefimov, D. P. Tieleman, and A. H. de Vries, J. Phys. Chem. B **111**, 7812 (2007).
[23] L. Monticelli, S. K. Kandasamy, X. Periole, R. G. Larson, D. P. Tieleman, and S.-J. Marrink, J. Chem. Theory Comput. **4**, 819 (2008).
[24] G. Rossi, L. Monticelli, S. R. Puisto, I. Vattulainen, and T. Ala-Nissila, Soft Matter **7**, 698 (2011).
[25] K. A. Maerzke and J. I. Siepmann, J. Phys. Chem. B **115**, 3452 (2011).
[26] R. DeVane, W. Shinoda, P. B. Moore, and M. L. Klein, J. Chem. Theory Comput. **5**, 2115 (2009).
[27] J. G. Kirkwood, J. Chem. Phys. **3**, 300 (1935).
[28] A. Liwo, S. Oldziej, M. R. Pincus, R. J. Wawak, S. Rackovsky, and H. A. Scheraga, J. Comput. Chem. **18**, 849 (1997).
[29] A. Liwo, C. Czaplewski, J. Pillardy, and H. A. Scheraga, J. Chem. Phys. **115**, 2323 (2001).
[30] J. Baschnagel, K. Binder, P. Doruker, A. A. Gusev, O. Hahn, K. Kremer, W. L. Mattice, F. Muller-Plathe, M. Murat, W. Paul, S. Santos, U. W. Suter, and V. Tries, Adv. Polym. Sci. **152**, 41 (2000).
[31] W. Schommers, Phys. Rev. A **28**, 3599 (1983).
[32] A. K. Soper, Chem. Phys. **202**, 295 (1996).
[33] F. Müller-Plathe, ChemPhysChem **3**, 754 (2002).
[34] R. H. Swendsen, Phys. Rev. Lett. **42**, 859 (1979).
[35] A. P. Lyubartsev and A. Laaksonen, Phys. Rev. E **52**, 3730 (1995).
[36] A. P. Lyubartsev and A. Laaksonen, Phys. Rev. E **55**, 5689 (1997).
[37] A. Lyubartsev, A. Mirzoev, L. J. Chen, and A. Laaksonen, Faraday Discuss. **144**, 43 (2010).
[38] A. Savelyev and G. A. Papoian, J. Phys. Chem. B **113**, 7785 (2009).
[39] A. Savelyev and G. A. Papoian, Biophys. J. **96**, 4044 (2009).
[40] A. Savelyev and G. A. Papoian, Proc. Natl. Acad. Sci. U.S.A. **107**, 20340 (2010).
[41] M. S. Shell, J. Chem. Phys. **129**, 144108 (2008).
[42] S. Kullback and R. A. Leibler, Ann. Math. Stat. **22**, 79 (1951).
[43] A. Chaimovich and M. S. Shell, J. Chem. Phys. **134**, 094112 (2011).
[44] A. Chaimovich and M. S. Shell, Phys. Rev. E **81** (2010).
[45] T. Murtola, M. Karttunen, and I. Vattulainen, J. Chem. Phys. **131**, 055101 (2009).
[46] J. T. Chayes, L. Chayes, and E. H. Lieb, Commun. Math. Phys. **93**, 57 (1984).
[47] J. T. Chayes and L. Chayes, J. Stat. Phys. **36**, 471 (1984).
[48] S. Izvekov and G. A. Voth, J. Phys. Chem. B **109**, 2469 (2005).
[49] S. Izvekov and G. A. Voth, J. Chem. Phys. **123**, 134105 (2005).
[50] W. G. Noid, J.-W. Chu, G. S. Ayton, V. Krishna, S. Izvekov, G. A. Voth, A. Das, and H. C. Andersen, J. Chem. Phys. **128**, 244114 (2008).
[51] W. G. Noid, P. Liu, Y. T. Wang, J.-W. Chu, G. S. Ayton, S. Izvekov, H. C. Andersen, and G. A. Voth, J. Chem. Phys. **128**, 244115 (2008).
[52] A. Das and H. C. Andersen, J. Chem. Phys. **132**, 164106 (2010).
[53] L. Y. Lu, S. Izvekov, A. Das, H. C. Andersen, and G. A. Voth, J. Chem. Theory Comput. **6**, 954 (2010).
[54] S. Izvekov, P. W. Chung, and B. M. Rice, J. Chem. Phys. **133**, 064109 (2010).
[55] F. Ercolessi and J. B. Adams, Europhys. Lett. **26**, 583 (1994).
[56] S. Izvekov, M. Parrinello, C. J. Burnham, and G. A. Voth, J. Chem. Phys. **120**, 10896 (2004).
[57] W. G. Noid, J.-W. Chu, G. S. Ayton, and G. A. Voth, J. Phys. Chem. B **111**, 4116 (2007).
[58] J. W. Mullinax and W. G. Noid, Phys. Rev. Lett. **103**, 198104 (2009).
[59] J. W. Mullinax and W. G. Noid, J. Phys. Chem. C **114**, 5661 (2010).
[60] J. W. Mullinax and W. G. Noid, J. Chem. Phys. **133**, 124107 (2010).
[61] J. W. Mullinax and W. G. Noid, Proc. Natl. Acad. Sci. U.S.A. **107**, 19867 (2010).
[62] S. Izvekov, A. Violi, and G. A. Voth, J. Phys. Chem. B **109**, 17019 (2005).
[63] S. Izvekov and G. A. Voth, J. Chem. Theory Comput. **2**, 637 (2006).
[64] J. Zhou, I. F. Thorpe, S. Izvekov, and G. A. Voth, Biophys. J. **92**, 4289 (2007).
[65] P. Liu, S. Izvekov, and G. A. Voth, J. Phys. Chem. B **111**, 11566 (2007).
[66] J. W. Mullinax and W. G. Noid, J. Chem. Phys. **131**, 104110 (2009).

[67]V. Ruhle, C. Junghans, A. Lukyanov, K. Kremer, and D. Andrienko, J. Chem. Theory Comput. **5**, 3211 (2009).
[68]V. Krishna and L. Larini, J. Chem. Phys. **135**, 124103 (2011).
[69]R. L. Henderson, Phys. Lett. A **49**, 197 (1974).
[70]C. R. Ellis, J. F. Rudzinski, and W. G. Noid, Macromol. Theory Sim. **20**, 478 (2011).
[71]D. M. Bradley and R. C. Gupta, Ann. Inst. Stat. Math. **54**, 689 (2002).
[72]P. Español and I. Zúñiga, Phys. Chem. Chem. Phys. **13**, 10538 (2011).
[73]A. J. Chorin, Multiscale Model. Simul. **1**, 105 (2003).
[74]A. J. Chorin and O. H. Hald, *Stochastic Tools in Mathematics and Science* (Springer, New York, 2006).
[75]P. Hohenberg and W. Kohn, Phys. Rev. **136**, B864 (1964).
[76]N. D. Mermin, Phys. Rev. **137**, A1441 (1965).
[77]M. E. Johnson, T. Head-Gordon, and A. A. Louis, J. Chem. Phys. **126**, 144509 (2007).
[78]MATHEMATICA version 8.0 Wolfram Research, Inc., Champaign, IL, 2010).
[79]A. A. Louis, J. Phys.: Condens. Matter **14**, 9187 (2002).
[80]H. C. Andersen, D. Chandler, and J. D. Weeks, Adv. Chem. Phys. **34**, 105 (1976).
[81]V. Krishna, W. G. Noid, and G. A. Voth, J. Chem. Phys. **131**, 024103 (2009).
[82]J. D. Weeks, J. Stat. Phys. **110**, 1209 (2003).
[83]H. Wang, C. Junghans, and K. Kremer, Eur. Phys. J. E **28**, 221 (2009).
[84]L. Larini, L. Y. Lu, and G. A. Voth, J. Chem. Phys. **132**, 164107 (2010).
[85]J.-P. Hansen and I. R. McDonald, *Theory of Simple Liquids*, 2nd ed. (Academic, San Diego, CA, 1990).