Reference state for the generalized Yvon–Born–Green theory: Application for coarse-grained model of hydrophobic hydration

J. W. Mullinax and W. G. Noid^{a)}*Department of Chemistry, The Pennsylvania State University, University Park, Pennsylvania 16802, USA*

(Received 15 June 2010; accepted 2 August 2010; published online 29 September 2010)

Coarse-grained (CG) models provide a computationally efficient means for investigating phenomena that remain beyond the scope of atomically detailed models. Although CG models are often parametrized to reproduce the results of atomistic simulations, it is highly desirable to determine accurate CG models from experimental data. Recently, we have introduced a generalized Yvon–Born–Green (g-YBG) theory for directly (i.e., noniteratively) determining variationally optimized CG potentials from structural correlation functions. In principle, these correlation functions can be determined from experiment. In the present work, we introduce a reference state potential into the g-YBG framework. The reference state defines a fixed contribution to the CG potential. The remaining terms in the potential are then determined, such that the combined potential provides an optimal approximation to the many-body potential of mean force. By specifying a fixed contribution to the potential, the reference state significantly reduces the computational complexity and structural information necessary for determining the remaining potentials. We also validate the quantitative accuracy of the proposed method and numerically demonstrate that the reference state provides a convenient framework for transferring CG potentials from neat liquids to more complex systems. The resulting CG model provides a surprisingly accurate description of the two- and three-particle solvation structures of a hydrophobic solute in methanol. This work represents a significant step in developing the g-YBG theory as a useful computational framework for determining accurate CG models from limited experimental data. © 2010 American Institute of Physics. [doi:10.1063/1.3481574]

I. INTRODUCTION

Atomistic molecular dynamics (MD) simulations provide exquisite insight into biological and condensed phase processes that evolve on nanometer length scales and nanosecond time scales.^{1,2} However, despite outstanding progress in hardware, software, and methodology, the computational expense of atomically detailed models significantly limits their utility for studying processes, e.g., peptide aggregation or membrane dynamics, that evolve on significantly longer length and time scales.³ These considerations have motivated renewed interest in lower resolution coarse-grained (CG) models that eliminate (or more precisely, integrate out) uninteresting atomistic details.^{4,5} By representing molecular systems with fewer effective interaction sites, CG models have provided highly efficient means for investigating complex processes that remain beyond the scope of all-atom models.^{6,7} Nevertheless, CG models may be misleading unless they have been carefully parametrized to accurately describe the system or phenomenon of interest.^{8,9} Accordingly, CG models are often parametrized to reproduce the properties of atomically detailed models.¹⁰ However, extensive atomistic simulations may then be required to adequately characterize the atomistic model. Moreover, the resulting CG model will reflect any biases or deficiencies in the original atomistic force field.^{11–13} Consequently, despite potential

practical limitations in the accuracy and availability of experimental data, it is clearly desirable to develop methods for determining CG models directly from experiment.^{14,15}

Indeed, the pioneering studies of Klein and co-workers^{16–18} and Marrink and co-workers^{19,20} have parametrized CG models to accurately reproduce experimental thermodynamic data, e.g., surface tension or partitioning free energies. The resulting models have proven to be remarkably transferable and, in some cases, have provided a semiquantitative model of molecular structure.¹⁹ However, in general, the resulting models are not optimized to provide an accurate model of molecular structure and it may prove challenging to adapt these potentials for modeling the conformational flexibility of complex biomolecules, e.g., proteins.^{21,22}

Considerable research has also focused on developing CG models that accurately describe molecular structure. The many-body potential of mean force (PMF) is the appropriate potential for a CG model that quantitatively reproduces the structural properties of a particular atomistic model.^{10,23} In general, though, the many-body PMF is prohibitively difficult to calculate, represent, and simulate. Consequently, a number of groups have developed systematic approaches to determine accurate approximations to the PMF that can be readily employed in subsequent CG MD simulations.

For instance, if the interactions considered by the CG model are only weakly coupled, then direct Boltzmann inversion determines a simple and accurate approximation to the many-body PMF, while only requiring simple structural

^{a)}Electronic mail: wnoid@chem.psu.edu.

correlation functions.^{24–26} Alternatively, iterative Boltzmann inversion,^{27–29} reverse (or inverse) Monte Carlo,^{30,31} relative entropy minimization,³² and molecular renormalization group^{33,34} methods have been developed for modeling systems with more strongly correlated interactions. These latter methods employ multiple simulations to iteratively determine a CG potential that accurately reproduces low order (typically pair) structural correlation functions that, at least in principle, can be determined experimentally.^{35,36} In addition to reproducing, e.g., radial distribution functions (RDFs), the resulting potentials also correspond to the solution of several variational principles relating the free energies of atomistic and CG models.^{31,32,37} Allen and Rutledge^{38,39} have demonstrated that these methods can be generalized with density-dependent potentials to also consider the chemical potential. Moreover, density functional theory analyses have demonstrated that, in many cases, the pair potential reproducing a particular RDF exists and is unique.^{40–42} However, in practice, it may prove challenging to iteratively refine the CG potentials for some more complex systems.⁴³

Izvekov and Voth^{44,45} have independently developed a force-based methodology for directly (i.e., noniteratively) determining CG potentials that provide a different approximation to the many-body PMF. This multiscale coarse-graining (MS-CG) method determines a variationally optimal approximation to the PMF by projecting the many-body mean force field onto a “basis set” specified by a particular form for the approximate CG potential.^{46,47} Although this approximate potential does not necessarily reproduce any particular structural correlation functions, the MS-CG method does not require iterative simulations to refine the CG potential, it directly addresses many-body correlations in an accurate and intuitive manner,^{48,49} it has been consistently formulated in both the isothermal⁴⁶ and isothermal-isobaric ensemble,⁵⁰ and it is also compatible with several approaches for determining potentials that can be transferred between temperatures⁵¹ and between systems.⁵² However, because it relies upon force information sampled from atomistic simulations, the MS-CG method cannot be immediately exploited to address experimental structural or thermodynamic data.

Very recently, Mullinax and Noid^{53,54} have demonstrated that the MS-CG method can be reformulated as a generalization of the venerable Yvon–Born–Green (YBG) integral equation theory.⁵⁵ This generalized-YBG (g-YBG) theory provides an exact relationship between a set of inter- and intramolecular potentials and a corresponding set of equilibrium structural correlation functions. Conversely, when given a collection of structural correlation functions, the g-YBG relation determines the set of potentials that provide a variationally optimal approximation (i.e., the MS-CG approximation) to the PMF. Consequently, the g-YBG theory determines the MS-CG potentials directly from experimentally determined structural data. In previous publications, we have employed the g-YBG theory to accurately determine MS-CG potentials from the structural correlations of liquid mixtures,⁵⁴ and also to quantitatively recover the underlying potentials from a collection of structures for a simple protein model.⁵³

In principle, the g-YBG theory is limited only by the quantity and quality of available structural data. However, in previous applications, this approach has employed extensive structural data. Consequently, in the present work, we introduce a “reference state potential” into the g-YBG framework to simplify the calculation of the CG potential. The reference potential is treated as a known contribution to the total potential. The remaining terms in the potential are determined such that, in combination with the fixed reference potential, they provide a variationally optimal approximation to the PMF. The treatment of the reference potential reduces the computational complexity and structural requirements of the method.

In the present work, we first numerically validate the proposed framework for a simple model of liquid butane. We demonstrate that the g-YBG method quantitatively recovers the MS-CG potentials directly from structures, as has been shown previously.⁵⁴ We then demonstrate that, when given part of the MS-CG potential as a reference potential, the proposed framework quantitatively determines the remaining MS-CG potentials directly from structures. As a second numerical illustration, we also demonstrate that the reference potential provides a convenient framework for transferring CG potentials parametrized from neat liquids to more complex systems. In this application, we first determine a two-site CG methanol model from structures of the neat liquid. We then employ these potentials as a fixed reference and determine a CG potential for modeling the solvation of a hydrophobic solute in methanol. The resulting CG model reproduces the pair and three-body solvation structure around the hydrophobic solute with surprising accuracy.

The remainder of the manuscript is organized as follows. In Sec. II, we develop the relevant theory for a reference state in the g-YBG framework. In Sec. III, we demonstrate the present framework for simple molecular systems. In Sec. IV, we discuss the present work and suggest future directions. Finally, in Sec. V, we provide concluding remarks.

II. THEORY

In the present section we develop the generalized-YBG theory for optimizing part of a CG potential in combination with a fixed reference potential. As in previous work,^{46,54} the present framework considers the canonical ensembles for high resolution and low resolution models, which shall be referred to as atomistic and CG models, respectively. It will be assumed that neither the atomistic nor the CG model includes rigid constraints.

The configuration of the atomistic model is defined by the Cartesian coordinates of n atoms, $\mathbf{r}=(\mathbf{r}_1, \dots, \mathbf{r}_n)$, that interact according to a potential $u(\mathbf{r})$. We assume that the atomistic configurations are canonically distributed at temperature T with probability, $p_r(\mathbf{r}) \propto \exp[-u(\mathbf{r})/k_B T]$. Similarly, the configuration of the CG model is defined by the Cartesian coordinates of N sites, $\mathbf{R}=(\mathbf{R}_1, \dots, \mathbf{R}_N)$, that interact according to a potential $U(\mathbf{R})$. We assume that the CG configurations are canonically distributed at the same temperature T with probability, $P_R(\mathbf{R}) \propto \exp[-U(\mathbf{R})/k_B T]$.

A set of N functions, $\mathbf{M}=(\mathbf{M}_1(\mathbf{r}), \dots, \mathbf{M}_N(\mathbf{r}))$, map atomistic configurations into CG configurations by determining the Cartesian coordinates of each site as a linear combination of the Cartesian coordinates for a subset of the atoms that are “involved” in that site.⁴⁶ It will be assumed that the coefficients of these linear combinations are appropriately normalized, that each site involves at least one atom, and that no atom is involved in more than one site.⁴⁶ These assumptions are sufficiently weak that the present framework applies for a wide range of CG models, including implicit solvent and mixed resolution models, although not for models in which a single atomic configuration maps to multiple CG configurations or for models in which molecular resolution may vary with time.^{56,57}

The configuration mapping $\mathbf{M}(\mathbf{r})$ and the atomistic configuration distribution $p_r(\mathbf{r})$ determine the probability $p_R(\mathbf{R})$ for sampling an atomistic configuration $\mathbf{r}$ that maps to a particular CG configuration $\mathbf{R}$,

$$p_R(\mathbf{R}) = \langle \delta(\mathbf{R} - \mathbf{M}(\mathbf{r})) \rangle, \quad (1)$$

where the angular brackets denote a canonical ensemble average for the atomistic model and $\delta(\mathbf{R} - \mathbf{M}(\mathbf{r})) = \prod_{I=1}^N \delta(\mathbf{R}_I - \mathbf{M}_I(\mathbf{r}))$.

We adopt the convention that a CG model is “structurally consistent” with a particular atomistic model for the same system if the probability of sampling a configuration for the CG model is equal to the probability of sampling a configuration for the atomistic model that maps to the specified CG configuration, i.e., $p_R(\mathbf{R}) = P_R(\mathbf{R})$.⁴⁶ Consequently, the appropriate potential U^0 for a CG model that is structurally consistent with a particular atomistic model is determined, to within an additive constant,

$$U^0(\mathbf{R}) = -k_B T \ln \langle \delta(\mathbf{R} - \mathbf{M}(\mathbf{r})) \rangle. \quad (2)$$

This potential is a many-body PMF because the corresponding force field $\mathbf{F}^0$ is equal to a conditioned canonical average of the atomistic force field,

$$\mathbf{F}_I^0(\mathbf{R}) = \langle \mathbf{f}_I(\mathbf{r}) \rangle_{\mathbf{R}}, \quad (3)$$

where $\mathbf{f}_I(\mathbf{r})$ is the atomistic force on CG site I in configuration $\mathbf{r}$ and the subscripted angular brackets denote a conditioned canonical ensemble average over the atomistic configurations that map to the specified CG configuration. Simulations that sample the canonical ensemble while using the mean force field as a conservative force field will be quantitatively consistent with the underlying atomistic model, according to the above definition.⁴⁶

As in previous work,^{47,52,53,58} we treat the CG force field, $\mathbf{F}=(\mathbf{F}_1(\mathbf{R}), \dots, \mathbf{F}_N(\mathbf{R}))$, as a vector in an abstract vector space that specifies the force on each CG site I in each CG configuration $\mathbf{R}$. In addition, we define the inner product between two CG force fields, $\mathbf{F}^{(1)}$ and $\mathbf{F}^{(2)}$,

$$\mathbf{F}^{(1)} \odot \mathbf{F}^{(2)} = \left\langle \sum_I \mathbf{F}_I^{(1)}(\mathbf{M}(\mathbf{r})) \cdot \mathbf{F}_I^{(2)}(\mathbf{M}(\mathbf{r})) \right\rangle, \quad (4)$$

and the norm for force fields: $\|\mathbf{F}\| = (\mathbf{F} \odot \mathbf{F})^{1/2}$.

The MS-CG functional^{44–46} is defined for a trial force field $\mathbf{F}$,

$$\chi^2[\mathbf{F}] = \left\langle \frac{1}{3N} \sum_{I=1}^N |\mathbf{F}_I(\mathbf{M}(\mathbf{r})) - \mathbf{f}_I(\mathbf{r})|^2 \right\rangle. \quad (5)$$

Because the mean force field is a projection (i.e., a conditioned expectation value) of the atomistic force field into the vector space of CG force fields,^{47,53,58–61} it follows that

$$\chi^2[\mathbf{F}] = \chi^2[\mathbf{F}^0] + \frac{1}{3N} \|\mathbf{F} - \mathbf{F}^0\|^2. \quad (6)$$

Because $\|\mathbf{F} - \mathbf{F}^0\|^2 \geq 0$ and because $\|\mathbf{F} - \mathbf{F}^0\|^2 = 0$ if and only if $\mathbf{F} = \mathbf{F}^0$, it follows that the mean force field is the unique minimum of χ^2 . Consequently, the MS-CG functional provides a variational framework for determining the mean force field $\mathbf{F}^0$ or optimal approximations to it.⁴⁶ Given a particular linear subspace within the vector space of CG force fields, the particular force field in that subspace that minimizes χ^2 determines the unique force field within that subspace that minimizes $\|\mathbf{F} - \mathbf{F}^0\|^2$. In this sense, the CG force field minimizing χ^2 within a given linear subspace determines a variationally optimal approximation to the mean force field.^{46,47}

In the present work we employ the MS-CG variational principle to optimize an approximate CG potential of the following form:⁴⁷

$$U(\mathbf{R}) = \sum_{\zeta} \sum_{\lambda} U_{\zeta}(\psi_{\zeta}(\{\mathbf{R}\}_{\lambda})) + U^R(\mathbf{R}), \quad (7)$$

where ζ indicates a particular type of potential (e.g., a dihedral angle potential) that is a function of a particular scalar variable ψ_{ζ} (e.g., a dihedral angle) that is itself a function of the Cartesian coordinates for a particular set λ of sites and U^R is a reference potential that specifies a fixed contribution to the approximate CG potential.^{47,53,58} In general, U^R may be a linear combination of simple potentials or may include more complex many-body interactions. Our objective is to determine the set of interaction potentials $\{U_{\zeta}\}$ that provide an optimal approximation to the PMF when combined with the reference potential U^R . The approximate CG potential in Eq. (7) determines a force on each site I ,

$$\mathbf{F}_I(\mathbf{R}) = \sum_{\zeta} \sum_{\lambda} \phi_{\zeta}(\psi_{\zeta}(\{\mathbf{R}\}_{\lambda})) \frac{\partial \psi_{\zeta}(\{\mathbf{R}\}_{\lambda})}{\partial \mathbf{r}_I} + \mathbf{F}_I^R(\mathbf{R}), \quad (8)$$

where $\psi_{\zeta\lambda}(\mathbf{R}) \equiv \psi_{\zeta}(\{\mathbf{R}\}_{\lambda})$,

$$\phi_{\zeta}(z) = - \frac{dU_{\zeta}(z)}{dz} \quad (9)$$

is a force function corresponding to the potential U_{ζ} , and $\mathbf{F}_I^R$ is the force from the reference potential U^R . Employing the vector representation discussed above, the CG force field in Eq. (8) may then be expressed as

$$\mathbf{F} = \sum_{\zeta} \int dz \phi_{\zeta}(z) \mathcal{G}_{\zeta}(z) + \mathbf{F}^R, \quad (10)$$

where the force function $\phi_{\zeta}(z)$ is a scalar coefficient for the force field vector $\mathcal{G}_{\zeta}(z)$ that specifies a force on each site in each configuration according to

$$\mathcal{G}_{I,\zeta}(\mathbf{R};z) = \sum_{\lambda} \frac{\partial \psi_{\zeta\lambda}(\mathbf{R})}{\partial \mathbf{R}_I} \delta(\psi_{\zeta\lambda}(\mathbf{R}) - z). \quad (11)$$

A particular reference force field and a set of force field vectors then specify a linear (affine) subspace of CG force fields of the form given by Eq. (10). Within this subspace, there exists a unique force field that minimizes the distance $\|\mathbf{F} - \mathbf{F}^0\|^2$ and defines an optimal approximation (among the force fields within the subspace) to the many-body mean force field.⁴⁷ This optimal approximate CG force field corresponds to the projection^{53,58,61} of the many-body mean force field into the linear subspace defined by Eq. (10) and can be obtained from the MS-CG variational principle.⁴⁷ If we define the inner product of the mean force field with the basis vector $\mathcal{G}_{\zeta}(z)$ as

$$b_{\zeta}(z) \equiv \mathcal{G}_{\zeta}(z) \odot \mathbf{F}^0, \quad (12)$$

then it follows that this optimal approximation $\mathbf{F}$ satisfies the relation

$$\mathcal{G}_{\zeta}(z) \odot \mathbf{F} = b_{\zeta}(z) \quad (13)$$

for each force field vector $\mathcal{G}(z)$ included in Eq. (10). After expanding $\mathbf{F}$ according to Eq. (10), we obtain a system of linear integral equations for each force function, $\phi_{\zeta}(z)$,

$$\sum_{\zeta'} \int dz' G_{\zeta\zeta'}(z, z') \phi_{\zeta'}(z') = b_{\zeta}(z) - b_{\zeta}^R(z), \quad (14)$$

where

$$G_{\zeta\zeta'}(z, z') = \left\langle \sum_{I=1}^N \mathcal{G}_{I,\zeta}(\mathbf{M}(\mathbf{r}); z) \cdot \mathcal{G}_{I,\zeta'}(\mathbf{M}(\mathbf{r}); z') \right\rangle \quad (15)$$

is a metric corresponding to the inner product $\mathcal{G}_{\zeta}(z) \odot \mathcal{G}_{\zeta'}(z')$ (Refs. 47, 53, and 58) and $b_{\zeta}^R(z) = \mathcal{G}_{\zeta}(z) \odot \mathbf{F}^R$ describes the correlation between the forces resulting from the reference potential and the basis vector $\mathcal{G}_{\zeta}(z)$. Because the mean force field corresponds to a projection of the atomistic force field, the inner product $b_{\zeta}(z)$ can be directly evaluated from atomistic force correlation functions according to^{53,58}

$$b_{\zeta}(z) = \left\langle \sum_{I=1}^N \mathcal{G}_{I,\zeta}(\mathbf{M}(\mathbf{r}); z) \cdot \mathbf{f}_I(\mathbf{r}) \right\rangle. \quad (16)$$

Voth and co-workers⁴⁷ have previously derived the system of equations, Eqs. (14)–(16), for the MS-CG force field using atomistic force information to determine $b_{\zeta}(z)$.

Mullinax and Noid^{53,58} have recently demonstrated that $b_{\zeta}(z)$ can be evaluated directly from structural correlation functions describing CG configurations,

$$b_{\zeta}(z) = k_B T \left(\frac{d\bar{g}_{\zeta}(z)}{dz} - L_{\zeta}(z) \right), \quad (17)$$

where

$$L_{\zeta}(z) = \left\langle \sum_{\lambda} \nabla^2 \psi_{\zeta\lambda}(\mathbf{M}(\mathbf{r})) \delta(\psi_{\zeta\lambda}(\mathbf{M}(\mathbf{r})) - z) \right\rangle, \quad (18)$$

$$\bar{g}_{\zeta}(z) = \left\langle \sum_{\lambda} |\nabla \psi_{\zeta\lambda}(\mathbf{M}(\mathbf{r}))|^2 \delta(\psi_{\zeta\lambda}(\mathbf{M}(\mathbf{r})) - z) \right\rangle, \quad (19)$$

$\nabla^2 = \sum_I (\partial/\partial \mathbf{R}_I)^2$ is the Laplacian, and $|\nabla \psi_{\zeta\lambda}(\mathbf{R})|^2 = \sum_I (\partial \psi_{\zeta\lambda}(\mathbf{R})/\partial \mathbf{R}_I)^2$. As noted previously, Eq. (17) is rigorously valid away from singularities in the potential where relevant Jacobian factors vanish.^{53,58} These singularities have not presented practical difficulties in numerical calculations of the CG potential from structures.

By substituting Eq. (17) in Eq. (14), we obtain the central result of this work—a g-YBG integral equation for determining the remaining unknown potentials given a fixed reference potential,

$$\sum_{\zeta'} \int dz' G_{\zeta\zeta'}(z, z') \phi_{\zeta'}(z') = k_B T \left(\frac{d\bar{g}_{\zeta}(z)}{dz} - L_{\zeta}(z) \right) - b_{\zeta}^R(z). \quad (20)$$

This system of linear integral equations employs structural correlation functions to directly determine the force field of the form specified by Eq. (10) that minimizes $\|\mathbf{F} - \mathbf{F}^0\|^2$. The reference potential enters into the calculation only through the correlation function $b_{\zeta}^R(z)$, which can be readily evaluated from structures because the reference potential is known (by definition). The treatment of the reference state significantly simplifies the calculation of the remaining potentials because the metric tensor $G_{\zeta\zeta'}(z, z')$ only considers correlations among basis vectors corresponding to the unknown potentials and does not consider correlations involving the reference potential.

Finally, in closing this section, we note that the present result readily generalizes to the case that the CG force field depends on a linear combination of discrete parameters and also to the case that the CG force field depends on both discrete and continuous parameters. In the latter case, the CG force field can be represented as

$$\mathbf{F} = \sum_{\zeta} \int dz \phi_{\zeta}(z) \mathcal{G}_{\zeta}(z) + \sum_{\zeta d} \phi_{\zeta d} \mathcal{G}_{\zeta d} + \mathbf{F}^R, \quad (21)$$

where the first summation is over interactions described by continuous force functions and the second summation is over interactions described by discrete parameters. The above analysis follows for this case, except that the discrete version of the generalized YBG theory must be employed to determine $b_{\zeta d} = \mathcal{G}_{\zeta d} \odot \mathbf{F}^0$ in terms of structural correlation functions, as shown earlier by Mullinax and Noid.^{53,58}

III. RESULTS

The present section numerically illustrates the approach derived in Sec. II. We first validate the accuracy of the method for a united atom CG model of butane. We demonstrate that if part of the MS-CG potential is employed as a reference potential, then the remaining terms in the potential can be quantitatively recovered directly from structures. We then illustrate this framework for developing a CG model of a hydrophobic solute in methanol. We employ a reference potential defined for pure methanol and determine the remaining solute-solvent potentials directly from structures.

The resulting CG model reproduces both the pair and three-body solvation structure with remarkable accuracy.

All MD simulations were performed with GROMACS 3.3.3 (Refs. 62 and 63) using periodic boundary conditions.¹ The atomistic (AA) models did not employ rigid constraints and included all hydrogens. All atomistic interactions were modeled with the OPLS-AA force field⁶⁴ using a cutoff of 1.2 nm for all van der Waals interactions. Electrostatics were modeled with the particle mesh Ewald method⁶⁵ using a real space cutoff of 1.2 nm. The Nose–Hoover thermostat^{66,67} was employed to sample the canonical ensemble for each system, while isothermal-isobaric ensembles were sampled by employing both the Nose–Hoover thermostat and the Parrinello–Rahman barostat.⁶⁸ A 1 fs time step was used to propagate all equations of motions.

In each case, the CG mapping determined the coordinates for each site from the center of mass of the atoms involved in the site. The CG forces were calculated in several different ways. Calculations using the MS-CG method^{44–47} employed Eq. (16) to determine $b_{\zeta}(z)$ using forces sampled from atomistic simulations and determined the CG forces by solving Eq. (14). Calculations using the g-YBG method^{53,58} employed Eq. (17) to determine $b_{\zeta}(z)$ using structures sampled from atomistic simulations and determined the CG forces by solving Eq. (20). In calculations with the g-YBG method, the derivative $d\bar{g}_{\zeta}(z)/dz$ was determined by first evaluating the structural correlation function $\bar{g}_{\zeta}(z)$ on a grid analogous to that used for the corresponding potential and then evaluating the derivative from centered finite differences. The function $\bar{g}_{\zeta}$ was not smoothed, and in some cases this procedure introduced considerable noise into the g-YBG force functions. Calculations without a reference potential were performed by setting $b_{\zeta}^R(z)=0$. In each case, $G_{\zeta\zeta'}$ was evaluated according to Eq. (15) using structures sampled from the relevant canonical ensemble. The relevant system of linear integral equations was discretized to obtain a corresponding system of linear algebraic equations. The equations corresponding to unsampled rows of $G_{\zeta\zeta'}$ were eliminated and the remaining linear system of equations was solved via LU decomposition after row preconditioning.^{69–71} The calculated pair forces were linearly extrapolated into the hard core region. These forces were interpolated on a finer grid and integrated using the trapezoid rule to determine potentials for subsequent CG MD simulations.

A. Validation for neat butane

A system of 380 atomically detailed butane molecules was equilibrated in the constant NPT ensemble at $P=1.0$ bar and $T=272$ K. The average volume of these simulations was $V=(3.943 \text{ nm})^3$, giving an average density of 0.5983 g/cm^3 . This density agrees quite well with both the previously published density of 0.589 g/cm^3 for the OPLS-AA butane model at $T=272.6$ K (Ref. 64) and also the experimentally measured density of 0.6011 g/cm^3 at $T=273$ K.⁷² The system was then simulated in the constant NVT ensemble at $T=272$ K and $V=(3.943 \text{ nm})^3$, generating an average pressure of 1.34 bar with a root mean squared deviation (RMSD) of 606 bar. Structures and forces from

FIG. 1. Comparison of force functions for a united atom CG model of butane calculated from atomistic forces via the MS-CG method (solid black curves) and from structures via the g-YBG method (dashed red curves). All calculations were performed without a reference potential. (a)–(c) correspond to the intermolecular pair forces between CH3–CH3, CH3–CH2, and CH2–CH2 pairs, respectively. (d)–(f) correspond to the CH3–CH2 bond stretch, CH3–CH2–CH2 bond angle, and CH3–CH2–CH2–CH3 dihedral angle force functions, respectively. The solid green curves in (d)–(f) correspond to the smoothed force functions employed in defining the intramolecular reference potential. The force functions in (a)–(d) are presented in units of kJ/mol nm. The force functions in (e) and (f) are presented in units of kJ/mol rad.

5000 configurations were sampled from these simulations. These atomistic configurations and forces were then mapped to a CG representation by representing each methyl group with a CH3 site and each methylene group with a CH2 site.

A potential for the CG model was then calculated from these mapped configurations using both the MS-CG method and the g-YBG method. In both calculations, the intramolecular potential included bond stretch, bond angle, and dihedral angle potentials, while the intermolecular potential included pair potentials between each pair of sites in distinct molecules using a cutoff of 1.2 nm. Each potential was represented using a delta function basis, i.e., using a discrete grid to tabulate the potential.⁴⁷ The bond stretch, bond angle, dihedral angle, and nonbonded pair potentials were represented with grid spacings of $\delta d=0.001$ nm, $\delta\theta=0.5^\circ$, $\delta\psi=1.0^\circ$, and $\delta r=0.005$ nm, respectively.

Figure 1 compares the force functions calculated via the MS-CG method using forces (solid black curves) with the force functions calculated via the g-YBG method using structures (dashed red curves). Both calculations were performed without a reference potential. Figures 1(a)–1(c) compare the nonbonded pair forces calculated from the MS-CG and g-YBG methods. The two sets of pair forces quantitatively agree, although the g-YBG potentials appear to be somewhat more noisy. Figures 1(d)–1(f) compare the bond stretch, bond angle, and dihedral angle forces calculated from the MS-CG and g-YBG methods. The results obtained for the CH2–CH2 bond force function are similar to those presented in Fig. 1(d) and will not be presented. In the case of the bond stretch and bond angle forces, the g-YBG and MS-CG methods agree quantitatively over the range of distances and angles that are accurately sampled during the ato-

FIG. 2. Differences between the force functions calculated via the g-YBG method with a reference potential and the force functions calculated via the g-YBG method without a reference potential (i.e., the force functions presented in the corresponding panels of Fig. 1). (a)–(c) present the differences in the intermolecular pair forces when using the intramolecular force functions [presented as the solid green curves in Figs. 1(d)–1(f)] as a reference force field. (d)–(f) present the differences in the intramolecular force functions when using the intermolecular pair forces [presented as the dashed red curves in Figs. 1(a)–1(c)] as a reference force field. The errors in (a)–(d) are presented in units of kJ/mol nm. The errors in (e)–(f) are presented in units of kJ/mol rad.

mistic simulations. The g-YBG bond (angle) forces deviate significantly from the MS-CG potentials only at exceptionally large and small bond lengths (angles). In the case of the dihedral angle force function, both the MS-CG and g-YBG forces are noisy, although the g-YBG forces are somewhat more noisy, especially for dihedral angles $\psi \approx 0$ that were rarely sampled during the simulation. Nevertheless, despite the increased noise present in the g-YBG forces, it is clear that the g-YBG method quite accurately recovers the CG forces obtained from the MS-CG method, but only requires structural information.

The treatment of the reference potential in the g-YBG framework was then validated for the butane model. The bond and angle forces calculated from the g-YBG theory were interpolated using a cubic spline on a finer grid ($\delta R = 0.0001$ nm and $\delta\theta = 0.008$ rad, respectively) and smoothed with a running average. The resulting force functions are presented as the green curves in Figs. 1(d) and 1(e). The dihedral force function calculated from the g-YBG theory was fitted to a Ryckaert–Bellemans form⁷³ shown by the green curve in Fig. 1(f). These intramolecular force functions defined a reference potential for calculating the intermolecular pair forces. Figures 2(a)–2(c) present the differences between the pair forces calculated from the g-YBG method with a reference potential and the pair forces calculated from the g-YBG method without a reference potential [i.e., the pair forces presented in Figs. 1(a)–1(c)]. It is clear that, for interparticle separations $R > 0.35$ nm, the calculation with the reference potential quantitatively recovered the pair forces calculated without a reference potential. Interparticle separations $R < 0.35$ nm correspond to the excluded volume “hard core” of each site and the errors in this region are relatively insignificant in comparison to the strong repulsion between sites at these separations.

An analogous test was then performed using the intermolecular pair forces as a reference force field for calculating the intramolecular force functions. In these calculations, the pair forces were directly used for the reference forces without any smoothing. Figures 2(d)–2(f) present the differences between bond stretch, bond angle, and dihedral angle force functions calculated from the g-YBG method with and without a reference potential. The differences in the bond stretch and bond angle forces are particularly small in comparison to the forces presented in Figs. 1(d) and 1(e). The differences in the dihedral forces are also quite small and are largest in the most noisy regions of the force function. In summary, Fig. 2 demonstrates that, when given an accurate reference potential, the remaining terms in the CG potential can be obtained directly from structures with quantitative accuracy.

B. Solvation structure in methanol

A system of 1000 atomically detailed methanol molecules was equilibrated in the constant NPT ensemble at $P = 1.0$ bar and $T = 298$ K. The average volume of these simulations was $V = (4.085 \text{ nm})^3$, giving an average density of 0.7805 g/cm^3 . This density is similar to both the previously published density of 0.779 g/cm^3 for the OPLS-AA methanol model at $T = 298$ K (Ref. 64) and also the experimentally measured density of 0.7913 g/cm^3 at $T = 293$ K.⁷² The system was then simulated in the constant NVT ensemble at $T = 298$ K and $V = (4.0850 \text{ nm})^3$, generating an average pressure of 0.340 bar with a RMSD of 596 bar. Structures for 1000 atomically detailed configurations were sampled from these simulations and then mapped to CG configurations by representing each methyl group with a CGC site and each hydroxyl group with a CGO site. The CG potential was directly calculated from the sampled structures according to the g-YBG theory without a reference potential. The intramolecular potential included a harmonic bond stretch potential, while the intermolecular potential included nonbonded pair potentials between each pair of sites in distinct molecules. The calculated spring constant was $\kappa = 2.483 \times 10^5 \text{ kJ/mol nm}^2$ with an equilibrium displacement of $d_0 = 0.1508$ nm. The calculated pair potentials were represented with discrete delta functions using a grid spacing of $\delta r = 0.005$ nm and a cutoff of 1.4 nm. These pair potentials are presented in Figs. 3(a)–3(c) and are similar to those obtained previously for pure methanol.^{45,51}

An atomistic model was then constructed for a single hydrophobic solute solvated in the system of 1000 methanol molecules. The solute was modeled as a purely repulsive sphere by applying the Weeks–Chandler–Andersen (WCA) transformation to a Lennard-Jones potential with $\sigma = 0.340$ nm and $\epsilon = 0.3598 \text{ kJ/mol}$.⁷⁴ The methanol-methanol interactions were modeled using the OPLS-AA force field⁶⁴ as before. Geometric combination rules following the GROMOS approach^{73,75} were used to determine WCA potentials for interactions between the hydrophobic solute and the C, HC, and O atoms of methanol. These potentials are presented as the dashed green and dotted blue curves in Figs. 3(d) and 3(e).

FIG. 3. Pair potentials employed in a CG model for hydrophobic solvation in methanol. (a)–(c) present the methanol intermolecular site-site potentials for CGC-CGC, CGO-CGO, and CGC-CGO site pairs, respectively. (d) and (e) present the site-site potentials for the hydrophobic solute site (SPH) with methanol CGC and CGO sites, respectively. (d) and (e) also present the WCA potentials employed to model the interaction of the hydrophobic sphere with the C, HC, and O atoms in all-atom simulations of the solute in methanol. All potentials are presented in units of kJ/mol.

The solute-solvent system was equilibrated at $T = 298$ K and $P = 1$ atm using the Nose–Hoover thermostat^{66,67} and Parrinello–Rahman barostat.⁶⁸ Structures for 10 000 atomically detailed configurations were then sampled from the canonical ensemble at $T = 298$ K and $V = (4.088 \text{ nm})^3$. These configurations were mapped to CG configurations by representing each methanol molecule with two (CGC and CGO) sites as before and the solute with a single site (SPH). Figure 4 presents the site-site RDFs calculated from these mapped configurations (solid black curves). Figure 4(a) indicates that the CGC sites are packed relatively uniformly in the atomistic model, while Fig. 4(b) indicates strong orientational dependent H-bonding interactions between methanol molecules. Figures 4(d) and 4(e) present the

FIG. 4. Comparison of site-site RDFs calculated from atomistic (solid black) and CG (dashed green) simulations of a single hydrophobic sphere in methanol. (a)–(c) present the CGC-CGC, CGO-CGO, and CGC-CGO RDFs, respectively. (d) and (e) present the SPH-CGC and SPH-CGO RDFs, respectively.

FIG. 5. Illustration of the geometry defining the angle θ used to characterize the orientation of methanol molecules about the hydrophobic sphere. This image was made with VMD (Ref. 93).

RDFs between the single SPH solute site and the methanol CGC and CGO sites, respectively. These figures demonstrate that the CGC sites, which represent the hydrophobic methyl group, preferably associate with the surface of the hydrophobic solute. Figure 4(d) also demonstrates several successive equally spaced peaks, indicating relatively long-ranged ordered packing of methyl groups about the hydrophobic solute. In contrast, Fig. 4(e) demonstrates that the CGO sites, which represents the polar hydroxy group, are relatively depleted from the solute surface and demonstrate little order after a small peak in the RDF at $r \approx 0.5$ nm.

The solvation structure of the atomistic model was further investigated by considering the three-particle orientational correlations of methanol molecules about the spherical solute in these mapped configurations. The methanol molecules in each configuration were partitioned into shells of thickness $\delta r = 0.2$ nm based on the distance of the CGC site from the solute. Figure 5 illustrates the angle θ defined by the SPH-CGC intermolecular vector and the CGC-CGO intramolecular vector. Figure 6 presents the distribution of $\cos \theta$ calculated in each shell as the solid black and solid green curves. Although the results are somewhat noisy from limited statistics, Fig. 6(a) indicates that methanol molecules in the first solvation shell have a strong preference to orient the hydroxy group away from the solute (i.e., $\cos \theta > 0$ implies $\theta < \pi/2$). However, this preference is considerably diminished in the second partitioning ($0.4 \text{ nm} < r < 0.6 \text{ nm}$) and any alignment is not statistically significant (at least for the present data set of 10^4 configurations) in farther partitions. (Note the difference in the scales of the first and second panels.)

The g-YBG framework with a reference potential was then employed to calculate a potential for a CG model of the methanol-solute system. The site-site pair potentials calculated for pure methanol and presented in Figs. 3(a)–3(c) were interpolated and smoothed on a grid with spacing $\delta r = 0.002$ nm as described above for butane. These pair potentials, along with the harmonic bond potential calculated for pure methanol, were used as a fixed reference potential. The solute-solvent pair potentials were directly calculated using 5000 structures sampled from the canonical ensemble for the

FIG. 6. Analysis of methanol orientation at different distances from the hydrophobic solute in atomistic and CG simulations. Molecules were partitioned into different spherical shells based on the distance of the CGC site from the solute. The angle θ corresponds to the angle formed by the two vectors illustrated in Fig. 5. Each panel presents distributions of $\cos \theta$ calculated from atomistic (solid) and CG (dashed) simulations in two successive shells with the distributions from the nearer (farther) shell represented by the black (green) curve.

solute-solvent system and using the g-YBG theory with the methanol reference potential. The calculated SPH-CGC and SPH-CGO potentials are presented as the solid black curves in Figs. 3(d) and 3(e), respectively. Figure 3(d) demonstrates that the excluded volume distance calculated for SPH and CGC sites is similar to the corresponding distance between the WCA sphere and the C atom in the atomistic model, although the SPH-CGC excluded volume interaction is somewhat more repulsive than the WCA-C interaction. The SPH-CGC pair potential also demonstrates a shallow long-ranged ($R \approx 1$ nm) attraction that is entirely absent (by definition) from the atomistic WCA-C pair potential. Figure 3(e) demonstrates that, although the WCA-O and SPH-CGO repulsions are quite similar at sufficiently short interparticle separations, the excluded volume distance of the SPH-CGO interaction is not well-defined. Instead, the SPH-CGO pair potential demonstrates a long-ranged effective repulsion over the same distance as the SPH-CGC attraction.

The methanol potentials calculated for pure methanol, along with these solute-methanol pair potentials, define a potential for a CG model of the hydrophobic sphere solvated in methanol. This CG model was simulated for 10 ns in the constant NVT ensemble at $T=298$ K and $V=(4.088 \text{ nm})^3$. Figure 4 compares the site-site RDFs obtained from this CG MD simulation (dashed green curves) with those obtained from mapping atomistic configurations (solid black curves). Figure 4(b) indicates that the CG model slightly overestimates the first peak and underestimates the second peak of the CGO-CGO RDF. Figure 4(c) indicates that the CG model does not reproduce the first peak or shallow first minimum of the CGC-CGO RDF. Overall, though, the CG model quite accurately reproduces the methanol site-site RDFs. More significantly, Figs. 4(d) and 4(e) indicate that the present CG model reproduces the solute-solvent RDFs with near quantitative accuracy. Furthermore, and much more impressively, Fig. 6 demonstrates that the CG model also accurately repro-

duces the orientation of methanol molecules about the hydrophobic solute. In particular, in the CG model and in quantitative agreement with the atomistic model, methanol molecules in the first partition direct the hydroxy group away from the solute, while in subsequent partitions this alignment rapidly decays with distance.

IV. DISCUSSION

The computational efficiency of reduced models has motivated tremendous interest in the development of systematic methods for parametrizing CG potentials that provide an accurate description of a particular molecular system or phenomenon of interest.^{9,10} Although many studies have employed atomistic simulations to parametrize these potentials, it is highly desirable to develop approaches for exploiting experimental data. While the Klein^{16–18} and Marrink^{19–21} groups have pioneered the use of thermodynamic data for parametrizing CG models, the resulting models do not necessarily provide an accurate description of structural fluctuations in complex biomolecules that possess significant conformational flexibility. Consequently, several other leading approaches have focused on developing CG models that accurately reproduce structural correlation functions.¹⁰ In particular, several methods iteratively refine CG potentials to reproduce low order structural correlation functions that can be, in many cases, experimentally determined.^{27–31,33,34} As a complement to these methods, Voth and co-workers^{44–46} have developed a MS-CG approach that directly determines the CG potential as a variationally optimal approximation to the many-body PMF. Because the MS-CG approach employs atomistic force information, it would appear that the MS-CG potential cannot be obtained from experimental data. However, we have recently introduced a generalized-YBG theory^{53,58} for determining this MS-CG potential directly from structural correlation functions that, at least in principle, can be determined from experiment.^{53,54} Moreover, we have previously demonstrated that this g-YBG theory determines CG potentials with quantitative accuracy, although in some cases this calculation has employed large amounts of structural data.⁵⁸

The present work introduces a reference state potential into the g-YBG framework as a means to simplify the calculation and reduce the structural data employed in determining the CG potential. The reference potential defines a fixed contribution to the total CG potential. The remaining terms in the CG potential are calculated such that, in combination with the reference potential, they provide a variationally optimal approximation to the many-body PMF.

The treatment of the reference potential significantly simplifies the calculation of the CG force field. The contribution from the reference potential can be readily subtracted from the mean force field by using structures to evaluate b_{ζ}^R . Importantly, the reference potential does not appear in the metric, $G_{\zeta\zeta'}$. This metric is defined by many-body structural distribution functions that quantify the correlation between the potential functions that are variationally optimized in the calculation of the CG force field. Because it reduces the number of potential functions that have to be optimized, the

reference potential reduces the size and complexity of $G_{\zeta\zeta'}$. This has significant advantages both because it reduces the hardware requirements involved in solving the system of equations for the force field parameters and also because it reduces the number of many-body correlation functions that must be sampled in $G_{\zeta\zeta'}$. A judicious choice of the reference potential may significantly reduce the amount of structural information required to accurately determine the remaining potentials. The cost of the reference potential is that it can limit the accuracy of the resulting approximation to the many-body PMF by restricting the space of force fields considered by the variational calculation. Consequently, it will be important to develop physical intuition and accurate *a priori* descriptions of the reference potential, e.g., from polymer physics or integral equation theories.

The present work demonstrates that the reference potential can be accurately implemented into the g-YBG framework. We considered a united atom CG model of butane and demonstrated that, if part of the MS-CG potential is specified as a reference potential, the remaining contributions to the MS-CG potential can be accurately determined directly from structures. We also demonstrated that the reference potential provides a convenient framework for transferring potentials between different systems. In particular, we calculated a CG potential for pure methanol and then employed this potential as a reference state for determining the CG potentials describing the interactions between methanol and a single hydrophobic solute.

The latter CG model provided a remarkably accurate description of both the pair and orientational three-site solvation structure around the hydrophobic solute. These calculations also demonstrated that even a relatively small degree of coarse-graining can generate effective potentials that radically differ from the potentials of the original all-atom model. In particular, starting from a purely repulsive WCA solute-solvent interaction and simply integrating out the hydrogen atoms from the original atomistic model resulted in a long-ranged effective attraction (repulsion) between the hydrophobic solute and the methyl (hydroxy) CG sites. These results indicate the care with which effective CG potentials must be treated.^{76,77} Although we have not yet considered these potentials in great detail, it is clear that they are highly sensitive to the resolution of the CG mapping and they are likely to be highly state point dependent. In the future, it may be interesting to further consider the transferability of these potentials within the extended ensemble framework.⁵²

We note in passing that the very simple systems and mappings considered in the present numerical calculations do not reflect limitations upon the present framework, but were instead chosen for computational convenience. For example, while the present approach could be easily applied for a single site model of butane, this work considered a united atom CG model because it included the complexity relevant for typical molecular mechanics force fields with, e.g., torsional potentials. As a consequence of these high resolution mappings, the present “CG” models are only slightly more computationally efficient than all-atom models. However, the

proposed approaches readily generalize to more complex molecular systems, e.g., proteins, for which CG models can provide considerable computational savings.

Finally, in closing, we emphasize that other researchers have previously employed a reference potential for determining CG potentials. As noted above, Voth, Andersen, and co-workers⁴⁷ have treated electrostatic potentials as a reference potential within the MS-CG framework. Moreover, the present work is particularly motivated by the central role of reference potentials in knowledge-based approaches for determining statistical potentials from protein databank structures. Knowledge-based (KB) approaches^{78–80} frequently determine amino-acid interactions, e.g., pair potentials, in CG protein models by comparing statistics compiled from the protein databank with those that would be expected for a fixed reference state.^{81–83} The resulting KB potentials (KBPs) have enjoyed considerable success in a wide range of protein science applications, including protein design,⁸⁴ folding,⁸⁵ and structure prediction.⁸⁶ Nevertheless, the fundamental basis of KBPs has been sharply criticized. KB methods do not correctly treat many-body correlations between interacting amino acids⁸⁷ and relate CG potentials to statistics compiled from unphysical ensembles using equations that are difficult to justify.^{88,89} In principle, these inaccuracies might be addressed using an appropriate reference state, although the identity of this reference state has remained somewhat unclear.^{90–92}

The present work extends the first principles framework that we have been developing for addressing these concerns.⁵⁸ The extended ensemble framework that we have recently proposed⁵² provides a rigorous method for compiling statistics from multiple systems to variationally optimize both the accuracy and transferability of CG potentials. The g-YBG theory exactly treats many-body correlations and determines these optimized potentials directly from structural correlation functions.^{53,54} The present work introduces a reference state into this rigorous framework. In combination, these advances will provide a first principles approach for systematically assessing and variationally improving the reference states employed in popular KB methods.

V. CONCLUSIONS

The present work introduces a reference state potential into the g-YBG framework for determining CG potentials directly from structural information. Given a fixed reference potential, this framework determines the remaining terms in the potential, such that, in combination with the reference potential, they provide a variationally optimal approximation to the many-body PMF. We have numerically demonstrated that, when given an accurate reference potential, this framework quantitatively determines the remaining contributions to the CG potential. We also demonstrated that this approach provides a convenient framework for transferring CG potentials between systems. In particular, we considered a hydrophobic solute in methanol and, using the CG potential for pure methanol as a reference potential, we determined the solute-solvent potentials directly from structures. The resulting CG model reproduced the pair and three-body structure

of the solvation structure with remarkable accuracy. Most importantly, the present work provides a systematic framework for reducing the computational complexity of the g-YBG framework and represents an important step toward developing this approach as a practical methodology for determining accurate CG models directly from structural information.

ACKNOWLEDGMENTS

The present work has been financially supported by start-up funds from the Pennsylvania State University. 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 many helpful conversations with Professor Jayanth Banavar. Figure 5 was made with VMD. VMD is developed with NIH support by the Theoretical and Computational Biophysics group at the Beckman Institute, University of Illinois at Urbana-Champaign.

- ¹M. P. Allen and D. P. Tildesley, *Computer Simulation of Liquids* (Oxford, New York, 1987).
- ²D. Frenkel and B. Smit, *Understanding Molecular Simulation: From Algorithms to Applications*, 2nd ed. (Academic, San Diego, 2002).
- ³G. Brannigan, L. C.-L. Lin, and F. L. H. Brown, *Eur. Biophys. J.* **35**, 104 (2006).
- ⁴M. Levitt, *J. Mol. Biol.* **104**, 59 (1976).
- ⁵*Coarse-Graining of Condensed Phase and Biomolecular Systems*, edited by G. A. Voth (CRC, Boca Raton, FL, 2009).
- ⁶V. Tozzini, *Curr. Opin. Struct. Biol.* **15**, 144 (2005).
- ⁷C. Clementi, *Curr. Opin. Struct. Biol.* **18**, 10 (2008).
- ⁸A. T. Hagler and B. Honig, *Proc. Natl. Acad. Sci. U.S.A.* **75**, 554 (1978).
- ⁹G. S. Ayton, W. G. Noid, and G. A. Voth, *Curr. Opin. Struct. Biol.* **17**, 192 (2007).
- ¹⁰T. Murtola, A. Bunker, I. Vattulainen, M. Deserno, and M. Karttunen, *Phys. Chem. Chem. Phys.* **11**, 1869 (2009).
- ¹¹Y. Mu, D. S. Kosov, and G. Stock, *J. Phys. Chem. B* **107**, 5064 (2003).
- ¹²T. Yoda, Y. Sugita, and Y. Okamoto, *Chem. Phys.* **307**, 269 (2004).
- ¹³C. Maciel, E. E. Fileti, and R. Rivelino, *J. Phys. Chem. B* **113**, 7045 (2009).
- ¹⁴R. DeVane, M. L. Klein, C. C. Chiu, S. O. Nielsen, W. Shinoda, and P. B. Moore, *J. Phys. Chem. B* **114**, 6386 (2010).
- ¹⁵C. C. Chiu, R. DeVane, M. L. Klein, W. Shinoda, P. B. Moore, and S. O. Nielsen, *J. Phys. Chem. B* **114**, 6394 (2010).
- ¹⁶J. C. Shelley, M. Y. Shelley, R. C. Reeder, S. Bandyopadhyay, and M. L. Klein, *J. Phys. Chem. B* **105**, 4464 (2001).
- ¹⁷S. O. Nielsen, C. F. Lopez, G. Srinivas, and M. L. Klein, *J. Chem. Phys.* **119**, 7043 (2003).
- ¹⁸W. Shinoda, R. DeVane, and M. L. Klein, *Mol. Simul.* **33**, 27 (2007).
- ¹⁹S. J. Marrink, A. H. de Vries, and A. E. Mark, *J. Phys. Chem. B* **108**, 750 (2004).
- ²⁰S. J. Marrink, H. J. Risselada, S. Yefimov, D. P. Tieleman, and A. H. de Vries, *J. Phys. Chem. B* **111**, 7812 (2007).
- ²¹L. Monticelli, S. K. Kandasamy, X. Periole, R. G. Larson, D. P. Tieleman, and S.-J. Marrink, *J. Chem. Theory Comput.* **4**, 819 (2008).
- ²²R. DeVane, W. Shinoda, P. B. Moore, and M. L. Klein, *J. Chem. Theory Comput.* **5**, 2115 (2009).
- ²³A. Liwo, S. Oldziej, M. R. Pincus, R. J. Wawak, S. Rackovsky, and H. A. Scheraga, *J. Comput. Chem.* **18**, 849 (1997).
- ²⁴W. Tschöp, K. Kremer, J. Batoulis, T. Burger, and O. Hahn, *Acta Polym.* **49**, 61 (1998).
- ²⁵J. 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. Triess, *Adv. Polym. Sci.* **152**, 41 (2000).
- ²⁶V. A. Harmandaris, N. P. Adhikari, N. F. A. van der Vegt, and K. Kremer, *Macromolecules* **39**, 6708 (2006).
- ²⁷W. Schommers, *Phys. Rev. A* **28**, 3599 (1983).
- ²⁸F. Müller-Plathe, *ChemPhysChem* **3**, 754 (2002).
- ²⁹D. Reith, M. Pütz, and F. Müller-Plathe, *J. Comput. Chem.* **24**, 1624 (2003).
- ³⁰A. P. Lyubartsev and A. Laaksonen, *Phys. Rev. E* **52**, 3730 (1995).
- ³¹A. P. Lyubartsev and A. Laaksonen, *Phys. Rev. E* **55**, 5689 (1997).
- ³²M. S. Shell, *J. Chem. Phys.* **129**, 144108 (2008).
- ³³A. Savelyev and G. A. Papoian, *J. Phys. Chem. B* **113**, 7785 (2009).
- ³⁴A. Savelyev and G. A. Papoian, *Biophys. J.* **96**, 4044 (2009).
- ³⁵R. L. McGreevy and L. Pusztai, *Mol. Simul.* **1**, 359 (1988).
- ³⁶D. A. Keen and R. L. McGreevy, *Nature (London)* **344**, 423 (1990).
- ³⁷T. Murtola, M. Karttunen, and I. Vattulainen, *J. Chem. Phys.* **131**, 055101 (2009).
- ³⁸E. C. Allen and G. C. Rutledge, *J. Chem. Phys.* **130**, 034904 (2009).
- ³⁹E. C. Allen and G. C. Rutledge, *J. Chem. Phys.* **130**, 204903 (2009).
- ⁴⁰R. L. Henderson, *Phys. Lett. A* **49**, 197 (1974).
- ⁴¹J. T. Chayes, L. Chayes, and E. H. Lieb, *Commun. Math. Phys.* **93**, 57 (1984).
- ⁴²J. T. Chayes and L. Chayes, *J. Stat. Phys.* **36**, 471 (1984).
- ⁴³G. Megariotis, A. Vyrkou, A. Leygue, and D. N. Theodorou, "Systematic coarse graining of 4-cyano-4-pentylbiphenyl," *Ind. Eng. Chem. Res.* (in press).
- ⁴⁴S. Izvekov and G. A. Voth, *J. Phys. Chem. B* **109**, 2469 (2005).
- ⁴⁵S. Izvekov and G. A. Voth, *J. Chem. Phys.* **123**, 134105 (2005).
- ⁴⁶W. 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).
- ⁴⁷W. 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).
- ⁴⁸W. G. Noid, J.-W. Chu, G. S. Ayton, and G. A. Voth, *J. Phys. Chem. B* **111**, 4116 (2007).
- ⁴⁹L. Larini, L. Y. Lu, and G. A. Voth, *J. Chem. Phys.* **132**, 164107 (2010).
- ⁵⁰A. Das and H. C. Andersen, *J. Chem. Phys.* **132**, 164106 (2010).
- ⁵¹V. Krishna, W. G. Noid, and G. A. Voth, *J. Chem. Phys.* **131**, 024103 (2009).
- ⁵²J. W. Mullinax and W. G. Noid, *J. Chem. Phys.* **131**, 104110 (2009).
- ⁵³J. W. Mullinax and W. G. Noid, *Phys. Rev. Lett.* **103**, 198104 (2009).
- ⁵⁴J. W. Mullinax and W. G. Noid, *J. Phys. Chem. C* **114**, 5661 (2010).
- ⁵⁵J.-P. Hansen and I. R. McDonald, *Theory of Simple Liquids*, 2nd ed. (Academic, San Diego, CA, 1990).
- ⁵⁶M. Praprotnik, L. Delle Site, and K. Kremer, *J. Chem. Phys.* **126**, 134902 (2007).
- ⁵⁷M. Praprotnik, L. Delle Site, and K. Kremer, *Annu. Rev. Phys. Chem.* **59**, 545 (2008).
- ⁵⁸J. W. Mullinax and W. G. Noid, "Recovering physical potentials from a model protein databank," *Proc. Natl. Acad. Sci. USA* (to be published).
- ⁵⁹A. J. Chorin, O. H. Hald, and R. Kupferman, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 2968 (2000).
- ⁶⁰A. J. Chorin, *Multiscale Model. Simul.* **1**, 105 (2003).
- ⁶¹A. J. Chorin and O. H. Hald, *Stochastic Tools in Mathematics and Science* (Springer, New York, 2006).
- ⁶²E. Lindahl, B. Hess, and D. van der Spoel, *J. Mol. Model.* **7**, 306 (2001).
- ⁶³D. van der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. C. Berendsen, *J. Comput. Chem.* **26**, 1701 (2005).
- ⁶⁴W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives, *J. Am. Chem. Soc.* **118**, 11225 (1996).
- ⁶⁵T. Darden, D. York, and L. Pedersen, *J. Chem. Phys.* **99**, 8345 (1993).
- ⁶⁶S. Nosé, *Mol. Phys.* **52**, 255 (1984).
- ⁶⁷W. G. Hoover, *Phys. Rev. A* **31**, 1695 (1985).
- ⁶⁸M. Parrinello and A. Rahman, *J. Chem. Phys.* **76**, 2662 (1982).
- ⁶⁹W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in FORTRAN: The Art of Scientific Computing* (Cambridge University Press, New York, 1992).
- ⁷⁰J. W. Demmel, *Applied Numerical Linear Algebra* (SIAM, Philadelphia, PA, 1997).
- ⁷¹E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. D. Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen, *LAPACK Users' Guide* (SIAM, Philadelphia, PA, 1999).
- ⁷²J. G. Speight, *Lange's Handbook of Chemistry*, 16th ed. (McGraw-Hill, New York, 2005).
- ⁷³D. van der Spoel, E. Lindahl, B. Hess, A. R. van Buuren, E. Apol, P. J. Meulenhoff, D. P. Tieleman, A. L. T. M. Sijbers, K. A. Feenstra, R. van Drunen, and H. J. C. Berendsen, *GROMACS user manual version 3.3* (2005).
- ⁷⁴H. C. Andersen, D. Chandler, and J. D. Weeks, *Adv. Chem. Phys.* **34**, 105 (1976).
- ⁷⁵C. Oostenbrink, A. Villa, A. E. Mark, and W. F. Van Gunsteren, *J. Comput. Chem.* **25**, 1656 (2004).

- ⁷⁶A. A. Louis, *J. Phys.: Condens. Matter* **14**, 9187 (2002).
- ⁷⁷M. E. Johnson, T. Head-Gordon, and A. A. Louis, *J. Chem. Phys.* **126**, 144509 (2007).
- ⁷⁸S. Tanaka and H. A. Scheraga, *Macromolecules* **9**, 945 (1976).
- ⁷⁹S. Miyazawa and R. L. Jernigan, *Macromolecules* **18**, 534 (1985).
- ⁸⁰M. J. Sippl, *J. Mol. Biol.* **213**, 859 (1990).
- ⁸¹M. J. Sippl, *Curr. Opin. Struct. Biol.* **5**, 229 (1995).
- ⁸²R. L. Jernigan and I. Bahar, *Curr. Opin. Struct. Biol.* **6**, 195 (1996).
- ⁸³J. Moult, *Curr. Opin. Struct. Biol.* **7**, 194 (1997).
- ⁸⁴F. E. Boas and P. B. Harbury, *Curr. Opin. Struct. Biol.* **17**, 199 (2007).
- ⁸⁵C. Wilson and S. Doniach, *Proteins: Struct., Funct., Genet.* **6**, 193 (1989).
- ⁸⁶J. Skolnick, *Curr. Opin. Struct. Biol.* **16**, 166 (2006).
- ⁸⁷P. D. Thomas and K. A. Dill, *J. Mol. Biol.* **257**, 457 (1996).
- ⁸⁸A. Ben-Naim, *J. Chem. Phys.* **107**, 3698 (1997).
- ⁸⁹J. Shimada, A. V. Ishchenko, and E. I. Shakhnovich, *Protein Sci.* **9**, 765 (2000).
- ⁹⁰J. Skolnick, L. Jaroszewski, A. Kolinski, and A. Godzik, *Protein Sci.* **6**, 676 (1997).
- ⁹¹M. R. Betancourt and D. Thirumalai, *Protein Sci.* **8**, 361 (1999).
- ⁹²S. Liu, C. Zhang, H. Y. Zhou, and Y. Q. Zhou, *Proteins: Struct., Funct., Bioinf.* **56**, 93 (2004).
- ⁹³W. Humphrey, A. Dalke, and K. Schulten, *J. Mol. Graphics* **14**, 33 (1996).