BOCS: Bottom-up Open-source Coarse-graining Software

Nicholas J. H. Dunn,[†,‡] Kathryn M. Lebold,[†] Michael R. DeLyser,[†] Joseph F. Rudzinski,[†,¶] and W.G. Noid[*,†]

†*Department of Chemistry, The Pennsylvania State University, University Park, Pennsylvania 16802*

‡*Current Address: Minnesota Supercomputing Institute, University of Minnesota, Minneapolis, MN 55455*

¶*Current Address: Max Planck Institute for Polymer Research, 55128 Mainz, Germany*

E-mail: wnoid@chem.psu.edu

Abstract

We present the BOCS toolkit as a suite of open source software tools for parameterizing bottom-up coarse-grained (CG) models to accurately reproduce structural and thermodynamic properties of high resolution models. The BOCS toolkit complements available software packages by providing robust implementations of both the multiscale coarse-graining (MS-CG) force-matching method and also the generalized-Yvon-Born-Green (g-YBG) method. The g-YBG method allows one to analyze and to calculate MS-CG potentials in terms of structural correlations. Additionally, the BOCS toolkit implements an extended ensemble framework for optimizing the transferability of bottom-up potentials, as well as a self-consistent pressure-matching method for accurately modeling the pressure equation of state for homogeneous systems. We illustrate these capabilities by parameterizing transferable potentials for CG models that accurately model the structure, pressure, and compressibility of liquid alkane systems and by quantifying the role of many-body correlations in determining the calculated pair potential for a one-site CG model of liquid methanol.

Introduction

By representing systems in reduced detail, coarse-grained (CG) models provide the necessary computational efficiency for investigating length- and time-scales that cannot be effectively addressed with all-atom (AA) models.[1,2] Of course, CG models must be carefully constructed to faithfully describe the relevant physical forces if they are to provide useful predictions and insight. While one can imagine many approaches for constructing CG models, they are often developed via "top-down" or "bottom-up" approaches.[3–6]

Top-down approaches commonly parameterize relatively simple interaction potentials to reproduce macroscopic thermodynamic properties. Because top-down approaches often address multiple chemical systems and thermodynamic states, the resulting parameters can be used to define a general purpose force field. For instance, the Martini,[7,8] SDK,[9], PLUM,[10,11] and OxDna[12,13] force fields each employ a single set of parameters that is quite transferable, i.e., the parameters reasonably describe thermodynamic properties for a fairly broad range of systems and environments.

In contrast, bottom-up approaches commonly parameterize relatively complex interaction potentials to reproduce the structural properties of a high resolution model for a single chemical system in a single thermodynamic state point. Consequently, bottom-up approaches do not usually provide transferable force fields, but rather system-specific potentials that may require re-parameterization for each system and state-point of interest.[3,4] Accordingly, the practical application of bottom-up methods requires appropriate software for parameterizing these potentials. Fortunately, several software packages[14–17] have been released for implementing bottom-up approaches according to, e.g., iterative Boltzmann Inversion,[18] Inverse Monte Carlo,[19] and the multiscale coarse-graining (MS-CG) methods.[20–24]

Unsurprisingly, bottom-up approaches are currently limited by two common deficiencies. As emphasized above, bottom-up models generally provide limited transferability.[25–36] Similarly, because they often focus on reproducing structural properties, bottom-up approaches generally provide a rather poor description of thermodynamic properties, such as the pres-

sure.[27,37,38] Recently, we have examined the fundamental origin and interrelation between these transferability and representability limitations.[39] Moreover, we have developed rigorous computational methods for addressing these limitations in practice. In particular, the extended ensemble framework provides a principled bottom-up approach for developing potentials that accurately describe multiple chemical systems or thermodynamic states.[31,40] Additionally, self-consistent pressure-matching provides a straight-forward approach for constructing CG models that accurately model the pressure and compressibility of homogeneous systems.[41,42]

In this work, we present the Bottom-up Open-source Coarse-graining Software (BOCS) toolkit to complement the software packages that are currently available for parameterizing bottom-up CG models. The BOCS toolkit includes software written in C, C++, and python for use with the GROMACS[43,44] and LAMMPS[45] simulation packages. The BOCS toolkit includes a robust and stable implementation of the MS-CG force-matching method[20,22] for determining CG potentials directly from atomistic forces. Additionally, the BOCS toolkit implements the generalized Yvon-Born-Green (g-YBG) framework[46,47] for calculating MS-CG potentials directly from structural data. Based upon the g-YBG framework, the BOCS toolkit provides tools for interpreting the physical origin of these potentials in terms of structural correlations generated by the high resolution model.[48] Moreover, the BOCS toolkit implements both the extended ensemble framework[31] and also the self-consistent pressure-matching method.[41,42]

We are releasing the BOCS toolkit as open source software under the GPLv3 license in the hope that the CG modeling community will use and modify these tools according to its needs. Open source software is vital to reproducible computational research, since it facilitates not only the examination of calculations performed with the software, but also of the software itself. The 'many eyes' effect of open source software can help to more quickly identify and correct errors in the software, while also providing opportunity for other researchers to review and improve the underlying algorithms. Finally, open source software

lowers the barrier for researchers entering the field of CG modeling, since new researchers can then leverage and build upon prior work, rather than having to start from scratch.

The remainder of this manuscript is organized as follows. Section II outlines the theoretical basis for the BOCS toolkit, while Section III describes its computational implementation. Section IV illustrates the capabilities of the BOCS toolkit in the context of parameterizing transferable interaction potentials for CG models that reasonably describe the structure and pressure-volume thermodynamics of butane, heptane, decane, and a butane-decane mixture. We also present some diagnostic capabilities of the BOCS toolkit using a one-site model of methanol as a representative example. Finally, Section V presents concluding remarks.

Theory

In this section, we briefly outline the theoretical foundation that is employed by the BOCS software package in parameterizing the potentials for a CG model from a high resolution simulation. The BOCS software package can employ statistics sampled from either the constant NVT or constant NPT ensemble to determine the CG interaction potential. However, CG models will generally require an additional volume-dependent potential to accurately calculate the pressure and to sample the correct density in the constant NPT ensemble.[32,41]

High resolution AA model

We first consider a high resolution model with n particles, $i = 1, \ldots, n$, which we shall refer to as atoms.[49] We indicate the atomic microstate by $(\mathbf{r}, \mathbf{p}, v)$, where the configuration $\mathbf{r} = (\mathbf{r}_1, \ldots, \mathbf{r}_n)$ indicates the Cartesian coordinates of each atom, $\mathbf{p} = (\mathbf{p}_1, \ldots, \mathbf{p}_n)$ indicates the corresponding set of momenta, and v indicates the volume. We assume an atomic Hamiltonian:

$$h(\mathbf{r}, \mathbf{p}, v) = \kappa(\mathbf{p}) + u(\mathbf{r}, v) \tag{1}$$

where the kinetic energy $\kappa(\mathbf{p}) = \sum_i \mathbf{p}_i^2/2m_i$, m_i is the mass of atom i, and the potential, $u(\mathbf{r}, v)$, may explicitly depend upon v, e.g., due to long-ranged interactions.[50–52] The potential determines a force $\mathbf{f}_i = -\left(\partial u/\partial \mathbf{r}_i\right)_v$ on each atom i and also a force on the wall, i.e., the instantaneous excess pressure, $p_{\rm xs} = -\left(\partial u/\partial v\right)_{\hat{\mathbf{r}}}$, where the latter partial derivative is performed at constant scaled coordinates, $\hat{\mathbf{r}}$. The fluctuating internal pressure of the AA model is then[49,53]

$$p_{\rm int}(\mathbf{r}, \mathbf{p}, v) = \frac{2}{3v}\kappa(\mathbf{p}) + p_{\rm xs}(\mathbf{r}, v). \tag{2}$$

Low resolution CG model

We next consider a low resolution model with $N \leq n$ particles, $I = 1, \ldots, N$, which we shall refer to as CG sites. We indicate the CG microstate by $(\mathbf{R}, \mathbf{P}, V)$, where the configuration $\mathbf{R} = (\mathbf{R}_1, \ldots, \mathbf{R}_N)$ indicates the Cartesian coordinates of each site, $\mathbf{P} = (\mathbf{P}_1, \ldots, \mathbf{P}_N)$ indicates the corresponding set of momenta, and V indicates the volume. We assume a CG Hamiltonian:

$$H(\mathbf{R}, \mathbf{P}, V) = \mathscr{K}(\mathbf{P}) + U(\mathbf{R}, V) \tag{3}$$

where the kinetic energy $\mathscr{K}(\mathbf{P}) = \sum_I \mathbf{P}_I^2/2M_I$, M_I is the mass of site I, and the potential, $U(\mathbf{R}, V)$, may depend upon both $\mathbf{R}$ and also V, as indicated below.[32,41,54] The potential determines a force $\mathbf{F}_I = -\left(\partial U/\partial \mathbf{R}_I\right)_V$ on each site I and also the instantaneous excess pressure, $P_{\rm xs} = -\left(\partial U/\partial V\right)_{\hat{\mathbf{R}}}$, where the latter partial derivative is performed at constant scaled CG coordinates, $\hat{\mathbf{R}}$. The fluctuating internal pressure of the CG model is then

$$P_{\rm int}(\mathbf{R}, \mathbf{P}, V) = \frac{2}{3V}\mathscr{K}(\mathbf{P}) + P_{\rm xs}(\mathbf{R}, V). \tag{4}$$

Mapped Ensemble

We intend for the CG model to reproduce the structural and thermodynamic properties of the AA model that can be observed at the resolution of the CG model. Accordingly, we

define a mapped ensemble by mapping each AA microstate $(\mathbf{r}, \mathbf{p}, v)$ to a CG microstate $(\mathbf{R}, \mathbf{P}, V)$. The mapping preserves the volume of the AA microstate, i.e., $V = v$.[32,41] The mapped configuration, $\mathbf{R} = \mathbf{M}(\mathbf{r})$, and momenta, $\mathbf{P} = \mathbf{M}_P(\mathbf{p})$, are specified by determining the Cartesian coordinates, $\mathbf{R}_I$, and momenta, $\mathbf{P}_I$, of each site, I, as a linear combination of atomic coordinates, $\mathbf{r}_i$, and momenta, $\mathbf{p}_i$:

$$\mathbf{R}_I = \mathbf{M}_I(\mathbf{r}) = \sum_i c_{Ii}\mathbf{r}_i \tag{5}$$

$$\mathbf{P}_I = \mathbf{M}_{PI}(\mathbf{p}) = M_I \sum_i c_{Ii}\mathbf{p}_i/m_i. \tag{6}$$

Note that Eq. (6) is equivalent to employing the same linear coefficients for mapping both the coordinates and the velocities.[22]

In principle, the mapping coefficients can be arbitrary positive constants that are appropriately normalized, $\sum_i c_{Ii} = 1$ for each $I = 1, \ldots, N$.[22] This normalization ensures that if each atom is displaced by a constant vector, then each CG site is displaced by the same vector. However, for simplicity, the BOCS package requires that each atom is associated with at most one CG site, i.e., for each atom i, c_{Ii} is non-zero for at most one CG site I. Given this restriction, the mapped atomistic force on each CG site may be expressed

$$\mathbf{f}_I(\mathbf{r}) = \sum_{i \in I} \mathbf{f}_i(\mathbf{r}) \tag{7}$$

where the sum is performed over all atoms i that are "involved" in CG site, I, i.e., the atoms i for which $c_{Ii} > 0$.[22]

Consistency and the many-body potential of mean force

It is straightforward to ensure that the CG model samples the mapped momentum distribution. (Of course, this does not imply that the CG model accurately describes any other dynamical property.[55,56]) Because we have assumed that the CG sites correspond to

disjoint atomic groups, the mapped CG momenta are statistically independent Gaussian random variables.[57] The CG model will be consistent with this mapped distribution if the site masses are given by

$$M_I{}^{-1} = \sum_{i \in I} c_{Ii}^2 m_i{}^{-1}, \tag{8}$$

which corresponds to ensuring that the Boltzmann distribution for the CG momenta has the appropriate variance.[22] Note that, if the mapping coefficients c_{Ii} determine the CG coordinates as the mass center of each corresponding atomic group, then $M_I = \sum_{i \in I} m_i$.

In order for the CG model to sample the mapped distribution for the configuration and volume, the Boltzmann weight for each CG configuration, $\mathbf{R}$, must equal the net Boltzmann weight for the atomic configurations, $\mathbf{r}$, that map to $\mathbf{R}$ at the given volume, V. Accordingly, the appropriate potential is the many-body potential of mean force (PMF), W:

$$\exp[-\beta W(\mathbf{R}, V, T)] = V_0^{N-n} \int_{V^n} \mathrm{d}\mathbf{r} \exp[-\beta u(\mathbf{r}, V)] \; \delta(\mathbf{R} - \mathbf{M}(\mathbf{r})), \tag{9}$$

where V_0 is an arbitrary reference volume that ensures dimensional consistency.[32,41,54,58–60]

The BOCS software package employs two variational principles to determine the potential U for the CG model. The force-[20,61] and pressure-[32,41] matching functionals are defined

$$\chi_1^2[U] = \left\langle \frac{1}{3N} \sum_I \left| \mathbf{f}_I(\mathbf{r}) - \mathbf{F}_I(\mathbf{M}(\mathbf{r})) \right|^2 \right\rangle \tag{10}$$

$$\chi_2^2[U] = \left\langle \left| p_{\mathrm{int}}(\mathbf{r}, \mathbf{p}, v) - P_{\mathrm{int}}(\mathbf{M}(\mathbf{r}), \mathbf{M}_P(\mathbf{p}), v) \right|^2 \right\rangle, \tag{11}$$

where the angular brackets denote an equilibrium ensemble average for the high resolution model. In practice we typically approximate these ensemble averages with configurations sampled from high resolution simulations. By minimizing the functionals χ_1^2 and χ_2^2, the BOCS toolkit determines U to approximate the configuration- and volume-dependence of the PMF, respectively.[22,32,41,62,63]

Das and Andersen (DA) originally proposed weighting each configuration in χ_1^2 by a

factor of $v^{2/3}$ in developing the pressure-matching method for systems in which the volume isotropically fluctuates.[32] Accordingly, the BOCS toolkit allows for the option of including this scaling in χ_1^2. However, this factor has no effect at constant V and appears to have little practical significance for condensed phase systems undergoing isotropic volume fluctuations at constant external pressure. Also, we note that the equivalence of Eq. (11) to the original pressure-matching functional proposed by Das and Andersen requires that the CG masses are consistently treated according to Eq. (8).

Approximate Potentials

We assume the following form for the CG potential:

$$U(\mathbf{R}, V) = U_R(\mathbf{R}) + U_V(V), \tag{12}$$

where the interaction potential, U_R, and volume-dependent potential, U_V, are optimized to approximate the configuration- and volume-dependence of the many-body PMF, respectively.[32,41]

Interaction potential

The interaction potential, U_R, is expressed as a sum of terms corresponding to different interactions, ζ, involving groups of particles, λ, that depend on scalar functions, ψ_ζ, of the corresponding CG coordinates, $\mathbf{R}_\lambda$:

$$U_R(\mathbf{R}) = \sum_\zeta \sum_\lambda U_\zeta(\psi_{\zeta\lambda}(\mathbf{R})) \tag{13}$$

where $\psi_{\zeta\lambda}(\mathbf{R}) = \psi_\zeta(\mathbf{R}_\lambda)$.[23,46] The Appendix illustrates this general potential form for a typical molecular potential. The resulting force on site I is then

$$\mathbf{F}_I(\mathbf{R}) = \sum_\zeta \sum_\lambda F_\zeta(\psi_{\zeta\lambda}(\mathbf{R}))\nabla_I \psi_{\zeta\lambda}(\mathbf{R}), \tag{14}$$

where $F_\zeta(x) = -\mathrm{d}U_\zeta(x)/\mathrm{d}x$ and $\nabla_I = \partial/\partial\mathbf{R}_I$. We represent each force function as a linear combination of basis functions, $f_{\zeta d}(x)$, with constant coefficients $\phi_{\zeta d}$:

$$F_\zeta(x) = \sum_d \phi_{\zeta d} f_{\zeta d}(x). \tag{15}$$

Given this representation of the force functions, we define force field "basis vectors" [23,46]

$$\boldsymbol{\mathcal{G}}_{I;\zeta d}(\mathbf{R}) = \sum_\lambda f_{\zeta d}(\psi_{\zeta\lambda}(\mathbf{R}))\nabla_I \psi_{\zeta\lambda}(\mathbf{R}) \tag{16}$$

such that the force on each site may be expressed:

$$\mathbf{F}_I(\mathbf{R}) = \sum_\zeta \sum_d \phi_{\zeta d} \boldsymbol{\mathcal{G}}_{I;\zeta d}(\mathbf{R}) = \sum_D \phi_D \boldsymbol{\mathcal{G}}_{I;D}(\mathbf{R}) \tag{17}$$

where, in the last expression, D is a "super-index" that specifies a combination ζd. Given Eq. (17) for the CG forces, χ_1^2 becomes a simple quadratic form in the force field parameters. The parameters that minimize χ_1^2 and, thus, provide an optimal approximation to the configuration-dependence of the PMF are determined by solving the normal system of linear equations[23,47,63]

$$\sum_{D'} G_{DD'}\phi_{D'} = b_D \tag{18}$$

where

$$b_D = \left\langle \frac{1}{3N} \sum_I \mathbf{f}_I(\mathbf{r}) \cdot \boldsymbol{\mathcal{G}}_{I;D}(\mathbf{M}(\mathbf{r})) \right\rangle \tag{19}$$

$$G_{DD'} = \left\langle \frac{1}{3N} \sum_I \boldsymbol{\mathcal{G}}_{I;D}(\mathbf{M}(\mathbf{r})) \cdot \boldsymbol{\mathcal{G}}_{I;D'}(\mathbf{M}(\mathbf{r})) \right\rangle . \tag{20}$$

Equation (18) can be interpreted as the projection of either the atomic force field or the many-body PMF (more precisely, the corresponding force field) onto the space of force fields spanned by the basis defined by Eq. (17).[22,23,46,62,64] Note that, if χ_1^2 scales each configuration by $v^{2/3}$, then b_D and $G_{DD'}$ both inherit this scaling in Eqs. (18)-(20). In practice we then divide b_D and $G_{DD'}$ by $\left\langle v^{2/3} \right\rangle$ in order to preserve their original scale and dimensions.

Volume-dependent potential

According to Eq. (12), the pressure of the CG model may be expressed:

$$P_{\text{int}}(\mathbf{R}, \mathbf{P}, V) = P^0_{\text{int}}(\mathbf{R}, \mathbf{P}, V) + F_V(V) \tag{21}$$

where

$$P^0_{\text{int}}(\mathbf{R}, \mathbf{P}, V) = \frac{2}{3V} \mathscr{K}(\mathbf{P}) - \left(\frac{\partial U_R(\mathbf{R})}{\partial V} \right)_{\hat{\mathbf{R}}} \tag{22}$$

includes the kinetic and virial contributions to the pressure from U_R, and $F_V(V) = -\mathrm{d}U_V(V)/\mathrm{d}V$ is a "pressure correction" for the CG model. Since U_R is optimized without regard to the pressure, P^0_{int} will tend to dramatically overestimate the pressure of the underlying atomistic model.[21,32,38,41,65] Consequently, U_V can be adjusted to ensure that the CG model provides appropriate Boltzmann weight for each volume and, equivalently, that it accurately reproduces the pressure of the atomistic model. Importantly, U_V does not impact the configuration distribution at a fixed volume.[66]

Das and Andersen[32] suggested representing the volume-dependent potential as a sum of basis functions:

$$U_V(V) = \sum_d \psi_d u_{Vd}(V). \tag{23}$$

where ψ_d act as parameters for U_V, u_{Vd} are basis functions of the form

$$u_{Vd}(V) \;=\; \begin{cases} N(V/\bar{v}), & \text{for } d = 1 \\ N(V/\bar{v} - 1)^d, & \text{for } d \geq 2 \end{cases} \tag{24}$$

and $\bar{v}$ is the average volume of the reference AA ensemble. The BOCS toolkit can also employ other basis functions for representing U_V. However, in practice Eq. (24) is quite convenient, since often only two basis functions are required to accurately model equilibrium density fluctuations at constant external pressure.[32,41] The two coefficients then correspond to corrections for the pressure and the compressibility:

$$\Delta P_{\text{int}} \;=\; -N\psi_1/\bar{v} \tag{25}$$

$$\Delta \kappa_T^{-1} \;=\; 2N\psi_2/\bar{v}. \tag{26}$$

Given the interaction potential, U_R, determined from Eq. (18), U_V is then determined by minimizing the pressure-matching functional χ_2^2 in Eq. (11). Given Eqs. (21)-(23) for the pressure of the CG model, this pressure-matching variational principle reduces to a linear least squares problem for the parameters ψ_d, which is then solved by a normal system of equations analogous to Eq. (18). The resulting U_V significantly reduces the pressure of the CG model and will often provide a qualitatively reasonable description of the AA pressure equation of state.[32,41]

The BOCS toolkit implements an iterative self-consistent pressure-matching method to further refine U_V such that the CG model quantitatively reproduces the AA pressure equation of state, $p_{\text{int}}(V)$.[41] In this method, one first simulates the CG model in the constant NPT

ensemble with a fixed interaction potential, U_R, and a trial estimate for U_V. This simulation provides a local estimate of the CG equation of state, $P_{\text{int}}(V)$. The discrepancy between the CG and AA equations of state then determines a correction to $F_V(V)$ in analogy to iterative Boltzmann inversion: $\delta F_V(V) = p_{\text{int}}(V) - P_{\text{int}}(V)$. In practice this procedure often quickly converges such that $p_{\text{int}}(V) \approx P_{\text{int}}(V)$ quite accurately. This procedure corresponds to determining U_V by minimizing a relative entropy[41,67,68] describing the overlap of AA and CG distributions for the constant NPT ensemble:

$$S_{\text{rel}}[U] = \int \mathrm{d}V \int_V \mathrm{d}\mathbf{R}\, p_{RV}(\mathbf{R}, V) \ln\left[p_{RV}(\mathbf{R}, V)/P_{RV}(\mathbf{R}, V; U)\right] \tag{27}$$

where p_{RV} and P_{RV} are the distributions for the mapped ensemble and for the CG model, respectively.

g-YBG formulation

In the canonical ensemble at constant volume, the normal equations for the MS-CG potential parameters are equivalent to a generalization of the Yvon-Born-Green equation from liquid state theory.[69,70] This can be seen by representing the CG force field with a continuous set of basis functions such that Eq. (17) can be expressed[46,47]

$$\mathbf{F}_I(\mathbf{R}) = \sum_\zeta \int \mathrm{d}x\, F_\zeta(x)\, \boldsymbol{\mathcal{G}}_{I;\zeta}(\mathbf{R}; x). \tag{28}$$

The normal MS-CG equations may then be expressed:

$$b_\zeta(x) = \bar{g}_\zeta(x)\, F_\zeta(x) + \sum_{\zeta'} \int \mathrm{d}x'\, \bar{G}_{\zeta\zeta'}(x, x')\, F_{\zeta'}(x') \tag{29}$$

where

$$b_\zeta(x) \;=\; \frac{1}{3N} \left\langle \sum_\lambda \left(\sum_I \mathbf{f}_I(\mathbf{r}) \cdot \nabla_I \psi_{\zeta\lambda}(\mathbf{M}(\mathbf{r})) \right) \delta\left(\psi_{\zeta\lambda}(\mathbf{M}(\mathbf{r})) - x\right) \right\rangle \tag{30}$$

may be interpreted as an average atomic force along the ψ_ζ order parameter, while

$$\bar{g}_\zeta(x) = \frac{1}{3N}\left\langle \sum_\lambda \left(\sum_I |\nabla_I \psi_{\zeta\lambda}(\mathbf{M}(\mathbf{r}))|^2 \right) \delta\left(\psi_{\zeta\lambda}(\mathbf{M}(\mathbf{r})) - x\right) \right\rangle \tag{31}$$

$$\begin{aligned} \bar{G}_{\zeta\zeta'}(x,x') = \frac{1}{3N}\Bigg\langle \sum_{\lambda\neq\lambda'} \left(\sum_I \nabla_I \psi_{\zeta\lambda}(\mathbf{M}(\mathbf{r})) \cdot \nabla_I \psi_{\zeta'\lambda'}(\mathbf{M}(\mathbf{r})) \right) \\ \times \delta\left(\psi_{\zeta\lambda}(\mathbf{M}(\mathbf{r})) - x\right) \delta\left(\psi_{\zeta'\lambda'}(\mathbf{M}(\mathbf{r})) - x'\right) \Bigg\rangle \end{aligned} \tag{32}$$

are ensemble averages describing equilibrium structural correlations. Eq. (29) can provide insight into the physical origin of the calculated potential, $U_\zeta(x)$, since it decomposes $b_\zeta(x)$ into a direct contribution from $U_\zeta(x)$ and correlated indirect contributions from every other interaction in the system.[48,70]

Moreover, we have previously demonstrated that $b_\zeta(x)$ can be directly calculated from structures[46,47]

$$b_\zeta(x) = k_B T \left[\mathrm{d}\bar{g}_\zeta(x)/\mathrm{d}x - L_\zeta(x) \right] \tag{33}$$

where

$$L_\zeta(x) = \frac{1}{3N}\left\langle \sum_\lambda \left(\sum_I \nabla_I^2 \psi_{\zeta\lambda}(\mathbf{M}(\mathbf{r})) \right) \delta\left(\psi_{\zeta\lambda}(\mathbf{M}(\mathbf{r})) - x\right) \right\rangle. \tag{34}$$

In particular, if $U_\zeta(x)$ is a central pair potential, then

$$b_\zeta(r) = -(2r^2/c_\zeta) w'_\zeta(r) g_\zeta(r) \tag{35}$$

where $g_\zeta(r)$ is the radial distribution function, $w_\zeta(r) = -k_B T \ln g_\zeta(r)$ is the corresponding pair potential of mean force,[71] and c_ζ is a dimensioned normalization constant. In this simple case, Eq. (29) may be re-expressed

$$-\mathrm{d}w_\zeta(r)/\mathrm{d}r = F_\zeta(r) + \sum_{\zeta'} \int \mathrm{d}x\; \bar{g}_\zeta^{-1}(r) \bar{G}_{\zeta\zeta'}(r,x) F_{\zeta'}(x), \tag{36}$$

in direct analogy to the YBG equation.[70]

These results also hold in the constant NPT ensemble as long as b_ζ and $G_{\zeta\zeta'}$ are defined according to Eqs. (19) and (20), respectively. However, if b_ζ and $G_{\zeta\zeta'}$ include the $v^{2/3}$ rescaling proposed in Ref. 32, then this analysis only approximately holds in the constant NPT ensemble.

Extended Ensemble Formulation

The extended ensemble approach provides a simple framework for determining interaction potentials that are transferable to multiple systems.[31] An extended ensemble is defined as a collection of multiple conventional ensembles that may differ in chemical identity or in thermodynamic conditions. We assign a label, γ, and a probability, p_γ, for each ensemble. We define extended ensemble averages

$$\langle a_\gamma(\mathbf{r}_\gamma)\rangle = \sum_\gamma p_\gamma \left\langle a_\gamma(\mathbf{r}_\gamma)\right\rangle_\gamma \tag{37}$$

where $\mathbf{r}_\gamma$ indicates a configuration for ensemble γ and $\langle\cdots\rangle_\gamma$ indicates the corresponding conventional equilibrium ensemble average. In practice, we simply assign equal weight to each γ included in the extended ensemble. For each ensemble, γ, we define a CG representation, $\Gamma = \mu(\gamma)$, and a corresponding configuration mapping: $\mathbf{R}_\Gamma = \mathbf{M}_\gamma(\mathbf{r}_\gamma)$. This mapping then determines a weight, $p_\Gamma = \sum_\gamma p_\gamma \delta_{\gamma,\mu(\gamma)}$, and also a many-body potential of mean force, W_Γ, for each Γ. In practice, the CG representation typically provides a one-to-one relationship between the atomistic and CG ensembles, i.e., each CG ensemble Γ corresponds to a single atomistic ensemble γ_Γ and $p_\Gamma = p_{\gamma_\Gamma}$

We seek to determine potentials U_Γ that provide an optimal approximation to W_Γ for each Γ. The MS-CG force-matching variational principle can be readily extended for this purpose by simply interpreting Eqs. (10) and (18)-(20) in terms of extended ensemble averages. If the potentials U_Γ are treated independently for each Γ, then the extended ensemble approach determines independent MS-CG models for each Γ. However, if the potentials

share transferable parameters, then these parameters are determined to provide an optimal approximation across the entire extended ensemble.

Computational Methods

The BOCS toolkit provides software tools for parameterizing the potential, $U(\mathbf{R}, V)$, for a CG model based upon information from an AA trajectory. Table 1 summarizes the primary input and output for these tools. Figure 1 outlines the workflow for determining the interaction potential, U_R, while Fig. 2 outlines the workflow for determining the volume-dependent potential, U_V.

The cgmap tool generates a mapped ensemble as the CG representation of an AA ensemble. The cgmap tool requires an AA trajectory file that contains atomically detailed configurations and, optionally, the corresponding velocities and forces. The cgmap tool also requires 1) a plain text file that determines the mapping coefficients, $\{c_{Ii}\}$, by specifying the CG representation for each type of molecule in the system; and 2) a CG topology file that specifies the type and connectivity of the sites in the CG model. Based upon the specified mapping coefficients, the cgmap tool determines the CG representation of each AA configuration according to Eq. (5). If the AA trajectory file includes velocity and force information, the cgmap tool determines the mapped velocities and forces according to Eqs. (6) and (7). The cgmap tool then provides a mapped CG trajectory file that can be analyzed using standard GROMACS tools.

The cgff tool calculates the parameters for U_R from the mapped CG trajectory file. The cgff tool requires a plain text input file to specify the types of potentials, U_ζ, included in U_R and also the basis functions, $f_{\zeta d}$, employed to represent each U_ζ. The cgff tool also requires a CG topology file to specify the instances, λ, of each interaction. Assuming that the mapped CG trajectory contains explicit force information, the cgff tool calculates the force correlation function, b_D, and structural correlation function, $G_{DD'}$, according to Eqs. (19) and (20),

respectively, for each pair of basis functions $D = \zeta d$ and $D' = \zeta' d'$. If forces are not present in the mapped trajectory, the cgff tool calculates b_D directly from structural information according to Eq. (33). Although force-based calculations (i.e., via Eq. (19)) require less sampling to accurately determine b_D, structure-based calculations (i.e., via Eq. (33)) yield equivalent results for sufficiently well sampled systems[31,47,72] and have proven quite useful for several applications.[40,73] The cgff tool then solves the normal system of linear equations, Eq. (18), for the potential parameters, ϕ_D. Finally, the cgff tool outputs these parameters, as well as, b_D, $G_{DD'}$, and additional supplemental files that characterize the system and provide diagnostic information about the calculation.

The cgff tool treats a fairly wide range of CG potentials that can be represented according to Eq. (13), i.e., bond-stretch potentials, bond-angle potentials, dihedral potentials, and short-ranged pair potentials. The cgff tool can represent each of these potentials with either piecewise constant functions, piecewise linear functions, or B-spline functions. The cgff tool also implements several standard analytic functional forms, including harmonic bond-stretch or bond-angle potentials, Fourier-series dihedral potentials, and Lennard-Jones-type pair potentials. Additionally, the cgff tool allows for fixed "reference potentials," U_R^{Ref}, that are specified by the user and can be of arbitrary complexity.[23,74] In this case, the user must supply an additional trajectory file specifying the resulting reference force, $\mathbf{F}_I^{\text{Ref}}$, on each CG site in each mapped AA configuration. The cgff tool computes a corresponding contribution to each force projection, b_D^{Ref}, from the reference potential, i.e., using $\mathbf{F}_I^{\text{Ref}}$ in the place of $\mathbf{f}_I$ in Eq. (19). The cgff tool then optimizes the remaining terms in U_R to match the remainder of each force projection, $\delta b_D = b_D - b_D^{\text{Ref}}$. In particular, if Coulombic or other long-ranged potentials are defined as reference potentials, then the cgff tool will determine the short-ranged potentials that, when combined with the specified long-ranged potentials, provide an optimal approximation to the many-body PMF.

The cgff tool provides several additional options for the calculation of ϕ_D and the resulting output. The cgff tool can precondition the normal equations, Eq. (18), by normalizing the

$G_{DD'}$ matrix according to the norm of each column, the max of each row, the total variance in each column, or the variance in b_D. The cgff tool can solve these normal equations via single value, Cholesky, UU, or LU decomposition.[75] The cgff tool can regularize these methods according to Bayesian inference[76] or a simpler uncertainty estimation.[77] The cgff tool also provides several options for specialized diagnostic output, including error estimates, eigendecomposition of $\bar{G}_{DD'}$, and also decomposition of b_D into contributions from different interactions according to the g-YBG theory, i.e., Eq. (29).

Because it quantifies many-body structural correlations, the calculation of $G_{DD'}$ can be quite time-consuming for large systems with many interacting CG sites. As indicated by Eq. (32), the cgff tool calculates the correlation between the forces generated on each site, I, from each pair, λ and λ', of nonbonded interactions. The cgff tool performs this calculation by looping over all triples of interacting particles. For a CG model with N sites, this calculation scales as $O(N^3)$. We have expedited this calculation by exploiting the symmetry of this loop and by employing the OpenMPI framework to distribute the frames of the mapped trajectory over multiple processors. This parallelization scales perfectly because each frame is treated independently in calculating $G_{DD'}$ and because this nested triple loop typically dominates the time required for calculating ϕ_D.

We note that the MSCGFM code[15] implements the normal equations, Eq. (18), as well as several other numerical methods for minimizing χ_1^2 to determine the MS-CG force field. Lu et al. have provided an excellent discussion of various numerical methods for minimizing χ_1^2, including methods for solving an over-determined system of linear equations with a block-averaging approximation.[15] In comparison to this block-averaging approach, the normal system of equations is more time consuming, due to the nested triple loop discussed above, and also requires the numerical inversion of a matrix with a relatively high condition number. Nevertheless, we find that, with proper choices of solution method, preconditioning, and regularization, our implementation performs well and, in test cases that we can rigorously test, accurately determines the MS-CG potential. Additionally, because they cor-

respond to a g-YBG integral equation that is explicitly expressed in terms of equilibrium ensemble averages, the normal equations facilitate molecular insight into the system and the resulting CG potentials.[48,70] Moreover, the normal equations allow for the calculation of these potentials directly from structural information.[40,46,47,72]

The cgff tool separates the calculated potential parameters, $\{\phi_D\}$, into files corresponding to different interactions. For interactions represented with simple functional forms, such as bond-stretch interactions represented with harmonic potentials, the resulting parameters can be immediately employed as input for CG simulations. However, for potentials represented with more flexible functional forms, such as non-bonded interactions represented with spline functions, the calculated parameters may require additional processing. The tables tool performs the necessary smoothing, extrapolation, and interpolation to generate input files for use in GROMACS simulations.[43,44] The lammps_tables.py script converts these files for use in LAMMPS simulations.[45]

The cgff tool also implements the extended ensemble framework[31] to determine transferable potentials that provide an optimal approximation to the many-body PMF's for multiple mapped ensembles, Γ, that correspond to distinct chemical systems or distinct thermodynamic state points. In this case, the cgff tool requires a mapped CG trajectory file for each AA ensemble, as well as plain text and CG topology files that specify the contributions to the interaction potential, U_Γ, for modeling each Γ. The cgff tool also requires the user specify the weight, p_γ, for each AA ensemble, γ, included in the extended ensemble. Given this input, the cgff tool calculates b_D and $G_{DD'}$ as extended ensemble averages and determines the optimal potential parameters, ϕ_D, from Eq. (18), as in the case of a single system.

The cgmap and cgff tools have been historically developed for use with GROMACS and currently employ several functions and data structures from the GROMACS libraries.[43,44] In particular, we currently employ GROMACS functionality to read and write GROMACS trajectory and topology files, as well as for some aspects of the user interface employed by the cgmap tool. In order to buffer these tools from the GROMACS source code and in

order to facilitate future compatibility, we developed an interface that wraps all references to GROMACS functions and addresses changes to relevant GROMACS libraries and files. The BOCS toolkit is currently natively compatible with GROMACS 4.5.x, 4.6.x, 5.0.x, and 5.1.x.

The BOCS toolkit also provides tools for determining U_V in order to simulate CG models that sample isotropic volume fluctuations under constant external pressure. The first step in this process is to estimate U_V via pressure-matching.[32,41] This calculation requires a fixed CG interaction potential, U_R, and a mapped CG trajectory file containing the mapped configuration, $\mathbf{M}(\mathbf{r}_t)$, mapped momentum, $\mathbf{M}_P(\mathbf{p}_t)$, and volume, V_t, for each time t. We then evaluate, for each t, the pressure, P^0_{int}, that is defined by Eq. (22) and accounts for the kinetic and interaction contributions to the instantaneous pressure of the CG model. In practice, this can be done by post-processing the mapped CG trajectory file using the '-rerun' option with the standard GROMACS mdrun tool. (Note that, if the CG potential includes table files, then these files must be specified in the topology files for this post-processing calculation and for subsequent CG simulations with GROMACS, as indicated by * in Fig. 2.) Given the resulting set of CG pressures, $\{P^0_{\text{int}}(t)\}$, as well as the corresponding AA pressures and volumes, $\{p_{\text{int}}(t), V_t\}$, the pmatch tool then determines U_V to minimize χ^2_2.

The resulting CG potential, $U(\mathbf{R}, V) = U_R(\mathbf{R}) + U_V(V)$, can then be simulated with lmp_pmatch, which is a modification of the LAMMPS distribution[45] from 17 June 2013 that includes the contributions from U_V in the barostat equation of motion. These simulations determine an estimate for the pressure-volume equation of state, $P_{\text{int}}(V)$, for the CG model. In practice, this CG model does not perfectly reproduce the pressure equation of state, $p_{\text{int}}(V)$, of the AA model.[32,41,42] This discrepancy presumably arises due to differences between the mapped and simulated configurational distributions at each V. Consequently, if necessary, we perform iterative self-consistent pressure-matching in order to refine U_V.[41,42] The CG and AA pressure equations of state are provided as input to the pmatch tool, which then

estimates the necessary correction for $U_V(V)$. This process can be iterated until the CG model adequately reproduces the AA pressure equation of state. In practice, this usually requires fewer than 10 iterations.[41,42]

There is no special workflow for determining U_V for transferable potentials obtained via the extended ensemble approach. In practice, we perform self-consistent pressure-matching to determine a separate potential $U_{V\Gamma}$ for each mapped ensemble, Γ. In principle, it may be possible to generalize the extended ensemble approach to determine a transferable pressure correction for modeling multiple state points or chemically distinct systems with similar interaction potentials. However, we have not yet tested this possibility.

Results and Discussion

In this section we illustrate the capabilities of the BOCS toolkit for parameterizing bottom-up CG models. In particular, we determine system-specific MS-CG potentials that accurately describe the structure of butane, heptane, and decane. We employ the extended ensemble (XN) approach to determine a single set of transferable XN potentials for modeling the structure of all three liquids. Additionally, we determine volume potentials, U_V, for accurately modeling the pressure-volume behavior of each alkane system. Finally, we also employ the BOCS toolkit to characterize many-body correlations in liquid methanol and to investigate their contribution to the pair potential of mean force.

We performed atomistic MD simulations of three alkane systems with 267 butane, heptane, or decane molecules in order to parameterize three corresponding system-specific MS-CG potentials as well as a single set of transferable XN potentials. We also performed an atomistic MD simulation of a mixture with 134 butane molecules and 134 decane molecules in order to assess the predictive capability of the XN potential. We performed these simulations according to the procedures described in Ref. 42, which we briefly summarize in the following. We performed all atomistic simulations with GROMACS 4.5.3[43], while using

double-precision and a 1.0 fs timestep. We employed the OPLS-AA force field[78] to describe all interactions and employed the particle mesh Ewald method with a grid spacing of 0.08 nm to model electrostatic interactions.[79] In order to equilibrate these systems, we first heated each system to 1000 K and then cooled the system back to room temperature at constant volume. We next equilibrated each system at constant pressure, while employing the Berendsen thermostat and barostat.[80] Finally, we simulated each system at 1.0 bar pressure and an external temperature of 300 K, using the Parrinello-Rahman barostat[81] and the stochastic dynamics thermostat[82] with an inverse friction constant of 0.1 ps. The production runs of the pure systems were 45 ns in duration, while the production run of the mixture system was 70 ns. We note that, although we performed these simulations in double precision, BOCS can parameterize CG potentials from either single- or double-precision simulations.

We first employed the cgmap tool to map these AA trajectories to their CG representation. Figures 3a, 3b, and 3c present the CG representations for butane, heptane, and decane molecules, respectively. In each case, we represented terminal CH_2CH_3 groups with 'CT' sites and internal $CH_2CH_2CH_2$ groups with 'CM' sites. We employed a standard molecular mechanics CG potential to model each system. The intramolecular potentials included bond-stretch and bond-angle potentials between each pair and triple, respectively, of consecutive sites in the same molecule. The intermolecular potentials included short-ranged pair potentials between each pair of sites in distinct molecules. Table 2 lists the interactions included in the CG models for each liquid. The interactions that are highlighted in bold font were described by transferable potentials in the XN models, i.e., the XN models employed the same potential function for modeling these interactions in each alkane system. Note that the CG sites were not charged and that the intramolecular potential for the CG model of decane did not include a dihedral potential.

We next employed the cgff tool to determine system-specific MS-CG potentials[20,22] for each pure alkane system. Additionally, we also defined a parameterization extended ensemble by assigning a weight $p_\gamma = 1/3$ to each pure alkane system. We then employed the cgff tool

to determine a single set of transferable XN potentials for optimally approximating the many-body PMF for all three systems. We note that we employed the $v^{2/3}$ rescaling in these calculations, although this appears to have minimal impact upon the resulting potentials. The Supporting Information section presents the calculated intramolecular potentials.

Figure 4 presents the calculated nonbonded pair potentials for CT-CT, CT-CM, CM-CM pairs in panels a, b, and c, respectively. The red, blue, and green solid curves in Fig. 4 indicate the system-specific MS-CG pair potentials for butane, heptane, and decane, respectively. Each MS-CG potential reflects two characteristic distances of approximately 0.5 nm and 0.8 nm. The CM-CM and CT-CM MS-CG potentials demonstrate relatively weak attraction and are quite similar for heptane and decane. The XN potentials are quite similar to the MS-CG potentials for these interactions. In comparison to the CM-CM and CM-CT potentials, the CT-CT potentials tend to be much more attractive and demonstrate much greater variation between different liquids. In particular, the CT-CT MS-CG potentials for butane and heptane are much more attractive than the CT-CM or CM-CM MS-CG potentials. The XN CT-CT potential is most similar to the corresponding MS-CG potential for butane.

We then employed the pmatch tool to determine the volume potential, U_V, via pressure-matching.[32,41] In particular, for each of the three pure liquid alkane systems, we determined two distinct volume potentials for compatibility with the system-specific MS-CG potential and the transferable XN potential. In each case, we represented U_V according to Eq. (24) with two basis functions that correspond to corrections for the mean pressure and the compressibility according to Eq. (25) and (26). The resulting potentials, U_V, provided a qualitative, but not quantitative description of the AA pressure-volume fluctuations. Consequently, we employed the self-consistent pressure-matching approach described in Section III to iteratively refine U_V.[41,42] Table 3 expresses the final parameters for U_V in terms of corrections to the mean pressure and compressibility. Table 3 also presents the number of iterations required to optimize U_V for each potential and each system. In almost all cases, self-consistent

pressure-matching converged within 6 iterations.

However, butane required special treatment during this pressure matching procedure. Because the CG model adopts a particularly high resolution for butane, the necessary pressure correction is quite small and requires special care. In particular, the first 10 iterations of self-consistent pressure-matching did not converge upon a pressure correction for the MS-CG butane model that simultaneously reproduced both the mean pressure and the compressibility of the AA model. Consequently, we selected the ψ_1 and ψ_2 coefficients from two different iterations that accurately modeled the mean pressure and the compressibility, respectively. Because the XN potential for butane is more attractive than the corresponding MS-CG potential, the XN butane model requires an even smaller pressure correction. Indeed, given the XN interaction potential, the volume potential that minimized χ_2^2 resulted in the XN butane model vaporizing. Consequently, in order to accurately reproduce the AA pressure-volume behavior with the XN butane model, we discarded the parameters $\{\psi_1, \psi_2\}$ obtained directly from pressure matching and performed iterative pressure matching starting from the trial potential $U_V = 0$. Starting from this trial potential, the iterative pressure-matching determined a satisfactory pressure correction with a single iteration.

All simulations of CG models were performed with the lmp_pmatch program included in the BOCS toolkit. These CG simulations employed the MTTK barostat[51,83] and Nose-Hoover chain thermostat[84] with the default chain length of 3. Otherwise, these simulations employed equivalent parameters to the AA simulations. Figures 5-7 quantify the equilibrium structure and pressure-volume behavior of the CG models for the pure alkane liquids. The Supporting Information more exhaustively compares the AA and CG models.

The system-specific MS-CG and transferable XN potentials reasonably describe the equilibrium structure for each pure liquid. Panels a, b, and c of Fig. 5 present the CT-CT nonbonded radial distribution functions for butane, heptane, and decane, respectively. In each panel, the dashed line presents results for the mapped AA ensemble, while the solid lines present results for the CG models. The MS-CG models reproduce the AA CT-CT rdfs with

nearly quantitative accuracy. In particular, the MS-CG models describe the asymmetry in the first peak of the AA butane rdf and also accurately reproduce the increasing structure in the rdf that is observed with increasing chain length. Importantly, although the XN models employ the same transferable potentials for modeling each liquid, the XN models also reproduce the AA CT-CT rdfs with nearly quantitative accuracy. The Supporting Information demonstrates that the MS-CG and XN models provide a slightly less accurate, although still very satisfactory, description of the AA rdfs for CM-CM and CM-CT pairs.

Figure 6 compares simulated distributions of the radius of gyration, R_{G}, for each pure alkane system. In each case, the MS-CG and XN models generate almost identical distributions. In the case of butane, R_{G} corresponds to the bond between CG sites, which is accurately described by the CG models. In the cases of heptane and decane, the CG models reasonably reproduce the overall shape of the AA distributions and, moreover, reproduce the average R_{G} of the AA models to within approximately 1% error. However, the CG models fail to reproduce the fine details of the AA distributions. In particular, the AA distributions are multimodal with relatively sharp peaks at large R_{G}, which correspond to all atomic torsions sampling trans conformations, and long tails toward more compact conformations. In contrast, the CG distributions are simpler unimodal distributions and, in particular, fail to reproduce the sharp peaks of extended conformations. This discrepancy reflects the tendency of the CG models to sample smaller angles (between triples of bonded sites) than the AA models, as seen in Supporting Figures 7b and 8c. Ultimately, this error reflects the inability of the simple molecular mechanics potential to capture correlations between the bond-stretch and bond-angle in the mapped ensemble.[77,85] Interestingly, as the alkane chains become progressively longer, one expects that the AA distribution will become increasingly simple as more dihedral angles contribute to R_{G} and, consequently, more similar to the CG distribution.

Figure 7 presents the average internal pressure of each model as a function of the volume. As a consequence of the iterative self-consistent pressure-matching approach, the CG models

quantitatively reproduce the AA pressure-volume relations.

We briefly assessed the predictive power of the XN approach by considering a 50:50 butane:decane mixture, which was not considered in parameterizing the XN potential. Figure 8 presents the intermolecular CT-CT rdfs obtained from AA simulations and from CG simulations with the XN potential as the dashed black and solid red curves, respectively. Panels a, b, and c of Fig. 8 correspond to CT sites from butane-butane pairs, from butane-decane pairs, and from decane-decane pairs. Although the XN potential was parameterized without information about the interactions or packing in butane-decane mixtures, the XN model describes the structure of this mixture quite accurately. The XN model overestimates the AA CT-CT rdf for butane-butane pairs, but almost quantitatively reproduces the AA CT-CT rdfs for butane-decane and decane-decane pairs. The Supporting Information demonstrates that the XN model also accurately reproduces the AA CT-CM rdf for butane-decane and decane-decane pairs, as well as the CM-CM rdf for decane-decane pairs. Figure 9 presents the results of self-consistent pressure matching for this mixture. The CG model accurately reproduces the pressure-volume behavior of the AA model by construction.

In addition to determining the interaction potential, U_R, the cgff tool also characterizes many-body correlations in the mapped AA ensemble and quantifies their contribution to U_R. In order to illustrate these features, we consider a system of 968 methanol molecules. As illustrated in Fig. 3d, we represent each methanol with a single site that corresponds to its mass center. We choose this smaller molecule and simpler representation for convenience, since the many-body correlations in the mapped ensemble are then simpler to analyze and interpret. We performed AA simulations for the methanol system in the same manner as described above for the alkane systems, except that the AA production simulation lasted only 5 ns. We did not simulate the resulting CG potential, although previous studies have demonstrated that the MS-CG 1-site model quite accurately describes the structure of liquid methanol.[21]

Panel a of Fig. 10 employs Eq. (36) to decompose the pair mean force, $-w'(r)$, between

methanol molecules into the direct force, $F(r) = \phi(r)$, between the pair and an indirect "three-body" contribution from correlated interactions with other particles in the environment. The pair mean force can be directly calculated from the pair potential of mean force, $w(r) = -k_B T \ln g(r)$, while the direct force is determined via force-matching. The cgff tool uses Eq. (36) to decompose the indirect contribution to the pair mean force into contributions from every other type of interaction in the system.[70] Note that the 3-body contribution is attractive at short ranges, indicating that the environment forces particles closer together once the pair approaches 0.6 nm of one another. It is also interesting that the 2-body MS-CG force function includes a relatively large repulsion corresponding to a desolvation barrier near 0.4 nm that is not so pronounced in the pair mean force. This desolvation barrier in the 2-body force function is partially offset by the contributions of correlated interactions from the environment, as described by the the metric tensor, $\bar{G}_{\zeta\zeta'}(r, z)$. We note that, although we included the $v^{2/3}$ rescaling in calculating b_ζ and $\bar{G}_{\zeta\zeta'}$, we find that Eq. (35) remains valid to within the numerical precision of the calculations.

Because the one-site CG methanol model considers only one type of interaction, the metric tensor reduces to a single block matrix that depends upon the distances, r and r', of a pair of sites from a single central site. Panel b of Fig. 10 presents an intensity plot of this metric tensor, $\bar{G}(r, r')$. As defined by Eq. (32), $\bar{G}(r, r')$ describes the contribution to the pair mean force at r from correlations with particles a distance r' away. In particular, $\bar{G}(r, r')$ corresponds to the average cosine of the angle formed between such triplets of particles.[48] Red and blue regions of this intensity plot indicate positive and negative elements of $\bar{G}$, which in turn correspond to acute and obtuse angles between triplets, respectively. As previously described,[48] the negative blue band along the diagonal $r' \approx r$ indicates the tendency of equidistant particles to form obtuse angles due to their excluded volume. The positive red off-diagonal stripes along $r' \approx r \pm \sigma$ correspond to correlated forces arising from molecules in adjacent solvation shells about a central molecule, where σ characterizes the size of the molecules. The alternating red and blue bands moving out from the diagonal reflect the

successive solvation shells of methanol molecules.

Conclusions

We are releasing the BOCS toolkit as open source software for parameterizing bottom-up CG models. As we illustrated for alkane mixtures, the BOCS toolkit provides a robust implementation of both the MS-CG and g-YBG methods for determining interaction potentials. In principle, the g-YBG approach may be used for determining potentials directly from experimentally determined structure ensembles.[72] In this context, the g-YBG framework may prove useful for interpreting and possibly improving the reference states employed in knowledge based potentials that are empirically inferred from known protein structures.[70,86,87] Moreover, the BOCS toolkit implements an extended ensemble approach for optimizing the transferability of these potentials and also a self-consistent pressure-matching method for accurately modeling isotropic volume fluctuations at constant external pressure. We have recently demonstrated that the resulting volume potential can also be adapted[88] as a function of the local density[89–92] in order to model inhomogeneous systems. Finally, the BOCS toolkit provides unique capabilities for interpreting CG potentials and their relation to many-body correlations in condensed phases.

At the same time, it is worth noting several limitations of the BOCS toolkit. First and most fundamentally, in contrast to iterative methods, such as Iterative Boltzmann Inversion,[18] the Inverse Monte Carlo method,[19] or relative entropy minimization,[67,68] the MS-CG[20–24] and g-YBG methods[46,47] do not guarantee that the CG interaction potential will necessarily reproduce any particular structural features of the underlying mapped ensemble.[93] In practice, the MS-CG and g-YBG models often provide a very good description of intermolecular structure, as illustrated in this work. More generally, though, the structural fidelity of MS-CG and g-YBG models depends upon the adequacy of the approximate potential to account for the relevant many-body correlations in the mapped ensemble.[70,77,85]

Consequently, we intend in future work to implement more complex potentials into the BOCS toolkit and also develop more predictive tools for identifying appropriate CG representations. Furthermore, it may be fruitful to develop an iterative wrapper for the cgff tool in order to take advantage of iterative versions of the MS-CG/g-YBG method that can provide improved accuracy for modeling complex structure ensembles.[77,94–96] Similarly, while the BOCS toolkit is currently useful for accurately modeling the pressure equation of state for homogeneous systems, we anticipate developing tools for modeling other thermodynamic properties. Additionally, the BOCS toolkit is currently limited by the requirement for simple CG representations in which sites correspond to disjoint atomic groups and by the restriction to systems that are either at constant volume or that sample isotropic volume fluctuations. These limitations clearly motivate future work to further develop the BOCS toolkit. Moreover, in future work we envision implementing more efficient methods for calculating the $G_{DD'}$ matrix, as well as checkpointing methods for saving the results of partial calculations.

Finally, we note that the current version of the BOCS toolkit is incompatible with the most recent versions of GROMACS and LAMMPS, as well as with the trajectory formats of other MD engines. However, we are currently developing the next version of the BOCS toolkit, which will eliminate all GROMACS dependencies from the cgmap and cgff codebase. Instead, these tools will employ a simpler topology file format and be compatible with both plain text and binary trajectory file formats. These formats can then be readily translated for use with Gromacs2016 or other MD engines. Moreover, we are also developing software for employing barostats with CG pressure corrections in current and future distributions of the LAMMPS package. These developments should significantly extend the utility of the BOCS toolkit.

Nevertheless, despite the aforementioned limitations, we hope that the BOCS toolkit will provide a useful complement to the software already available for developing bottom-up CG models. The source code, as well as documentation and tutorials, for the BOCS toolkit are available for download at `https://github.com/noid-group/BOCS` under the terms of the

GPLv3 license.

Appendix

The theory section employs rather abstract notation in order to address a correspondingly general class of interaction potentials, U_R. This appendix provides a more concrete and explicit treatment of U_R for a common molecular mechanics potential with contributions from bond-stretch, bond-angle, dihedral, and pair potentials. The potential in configuration $\mathbf{R}$ may be expressed

$$\begin{aligned} U_R(\mathbf{R}) &= \sum_{\zeta}\sum_{\lambda} U_\zeta(\psi_{\zeta\lambda}(\mathbf{R})) \\ &= \sum_{\alpha}^{\text{bonds}} U^{(b)}_{t_b(\alpha)}(b_\alpha) + \sum_{\alpha}^{\text{angles}} U^{(\theta)}_{t_\theta(\alpha)}(\theta_\alpha) + \sum_{\alpha}^{\text{dihedrals}} U^{(\psi)}_{t_\psi(\alpha)}(\psi_\alpha) + \sum_{(I,J)}^{\text{pairs}} U^{(2)}_{t_2(I,J)}(R_{IJ}). \end{aligned} \quad (38)$$

The first term in Eq. (38) describes all contributions from bond-stretch interactions. In this first term, α is a label indexing each bond, the sum ranges over all bonds, $t_b(\alpha)$ indicates the type of bond α, $U^{(b)}_{t_b(\alpha)}$ is the bond-stretch potential governing all bonds of type $t_b(\alpha)$, and b_α indicates the length of bond α in configuration $\mathbf{R}$. The second and third sums in Eq. (38) describe similar contributions from bond-angles and dihedral angles with α indexing the bond-angles and dihedral angles, respectively. Finally, the fourth term describes all non-bonded contributions from pair potentials. In this fourth term, (I, J) indicates a particular pair of sites, the sum is performed over all non-bonded pairs, $t_2(I, J)$ specifies the particular non-bonded potential, $U^{(2)}_{t_2(I,J)}$, describing the interaction between the pair, and R_{IJ} is the distance between the pair in configuration $\mathbf{R}$. Given this potential, the force on each site K

may be expressed

$$\begin{aligned}
\mathbf{F}_K(\mathbf{R}) &= \sum_{\zeta}\sum_{\lambda} F_{\zeta}(\psi_{\zeta\lambda}(\mathbf{R}))\frac{\partial\psi_{\zeta\lambda}(\mathbf{R})}{\partial\mathbf{R}_K} \\
&= \sum_{\alpha}^{\text{bonds}} F^{(b)}_{t_b(\alpha)}(b_\alpha)\frac{\partial b_\alpha}{\partial\mathbf{R}_K} + \sum_{\alpha}^{\text{angles}} F^{(\theta)}_{t_\theta(\alpha)}(\theta_\alpha)\frac{\partial\theta_\alpha}{\partial\mathbf{R}_K} \\
&+ \sum_{\alpha}^{\text{dihedrals}} F^{(\psi)}_{t_\psi(\alpha)}(\psi_\alpha)\frac{\partial\psi_\alpha}{\partial\mathbf{R}_K} + \sum_{(I,J)}^{\text{pairs}} F^{(2)}_{t_2(I,J)}(R_{IJ})\frac{\partial R_{IJ}}{\partial\mathbf{R}_K},
\end{aligned} \tag{39}$$

where $F^{(b)}_{t_b(\alpha)}(x) = -\mathrm{d}U^{(b)}_{t_b(\alpha)}(x)/\mathrm{d}x$ is the bond-force function, while $F^{(\theta)}_{t_\theta(\alpha)}(x)$, $F^{(\psi)}_{t_\psi(\alpha)}(x)$, and $F^{(2)}_{t_2(I,J)}(x)$ are corresponding force functions governing angles, dihedrals, and pair non-bonded interactions, respectively. Each of these force functions is represented by a linear combination of basis functions. For instance, if t_b specifies a particular type of bond governed by the potential function $U^{(b)}_{t_b}$, then the corresponding bond-force function is represented

$$F^{(b)}_{t_b}(x) = \sum_d \phi^{(b)}_{t_b d}\, f^{(b)}_{t_b d}(x), \tag{40}$$

where d indexes parameters, $\phi^{(b)}_{t_b d}$, that describe the bond force function $F^{(b)}_{t_b}(x)$, while $f^{(b)}_{t_b d}(x)$ indicates the corresponding basis function of a single variable. Similar expansions are adopted for the angle, dihedral, and non-bonded force functions.

Given this expansion the total force on site K may be expressed

$$\begin{aligned}
\mathbf{F}_K(\mathbf{R}) &= \sum_D \phi_D \boldsymbol{\mathcal{G}}_{K;D}(\mathbf{R}) \\
&= \sum_{t_b}^{b-\text{types}}\sum_d \phi^{(b)}_{t_b d}\boldsymbol{\mathcal{G}}_{K;t_b d}(\mathbf{R}) + \sum_{t_\theta}^{\theta-\text{types}}\sum_d \phi^{(\theta)}_{t_\theta d}\boldsymbol{\mathcal{G}}_{K;t_\theta d}(\mathbf{R}) \\
&+ \sum_{t_\psi}^{\psi-\text{types}}\sum_d \phi^{(\psi)}_{t_\psi d}\boldsymbol{\mathcal{G}}_{K;t_\psi d}(\mathbf{R}) + \sum_{t_2}^{\text{pair}-\text{types}}\sum_d \phi^{(2)}_{t_2 d}\boldsymbol{\mathcal{G}}_{K;t_2 d}(\mathbf{R}).
\end{aligned} \tag{41}$$

In Eq. 41, the first double sum describes contributions from bond-stretch forces. In this term, the first sum is over all types, t_b, of bonds, while the second sum ranges over the

parameters $\phi^{(b)}_{t_b d}$ describing the potential for bonds of type t_b. The corresponding force field basis vectors may be expressed

$$\boldsymbol{\mathcal{G}}_{K;t_b d}(\mathbf{R}) = \sum_{\alpha \in t_b} f^{(b)}_{t_b d}(b_\alpha) \frac{\partial b_\alpha}{\partial \mathbf{R}_K}, \tag{42}$$

where the sum is performed over all bonds α of type t_b. The remaining terms represent corresponding contributions from bond-angle, dihedral, and pair potentials.

Supporting Information Available

Details of simulated potentials, as well as additional comparison of AA and CG models.

Acknowledgement

The authors gratefully acknowledge financial support from the National Science Foundation (NSF Grant Nos. MCB-1053970, CHE-1565631), from the Alfred P. Sloan Foundation, and from a Camille Dreyfus Teacher-Scholar Award. This work was also supported by ACS PRF under Grant No. 52100-ND6. We gratefully acknowledge the Donors of the American Chemical Society Petroleum Research fund for support of this research. This work was partially supported by funding from the Penn State Materials Computation Center. Portions of this research were conducted with Advanced CyberInfrastructure computational resources provided by The Institute for CyberScience at The Pennsylvania State University (`http://ics.psu.edu`). Figure 3 employed VMD. VMD is developed with NIH support by the Theoretical and Computational Biophysics group at the Beckman Institute, University of Illinois at Urbana-Champaign.

References

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

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

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

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

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

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

(7) Marrink, S. J.; Risselada, H. J.; Yefimov, S.; Tieleman, D. P.; de Vries, A. H. The MARTINI force field: coarse grained model for biomolecular simulations. *J. Phys. Chem. B* **2007**, *111*, 7812–7824.

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

(9) Shinoda, W.; Devane, R.; Klein, M. L. Multi-property fitting and parameterization of a coarse grained model for aqueous surfactants. *Mol. Simul.* **2007**, *33*, 27–36.

(10) Bereau, T.; Deserno, M. Generic coarse-grained model for protein folding and aggregation. *J. Chem. Phys.* **2009**, *130*, 235106.

(11) Wang, Z. J.; Deserno, M. A systematically coarse-grained solvent-free model for quantitative phospholipid bilayer simulations. *J. Phys. Chem. B* **2010**, *114*, 11207–11220.

(12) Ouldridge, T. E.; Louis, A. A.; Doye, J. P. K. Structural, mechanical, and thermodynamic properties of a coarse-grained DNA model. *J. Chem. Phys.* **2011**, *134*, 085101.

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

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

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

(16) Karimi-Varzaneh, H. A.; Qian, H.-J.; Chen, X.; Carbone, P.; Müller-Plathe, F. IBIsCO: a molecular dynamics simulation package for coarse-grained simulation. *J. Comput. Chem.* **2011**, *32*, 1475–1487.

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

(18) Müller-Plathe, F. Coarse-graining in polymer simulation: From the atomistic to the mesoscopic scale and back. *ChemPhysChem* **2002**, *3*, 755 – 769.

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

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

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

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

(23) Noid, W. G.; Liu, P.; Wang, Y. T.; Chu, J.-W.; Ayton, G. S.; Izvekov, S.; Andersen, H. C.; Voth, G. A. The multiscale coarse-graining method. II. Numerical implementation for molecular coarse-grained models. *J. Chem. Phys.* **2008**, *128*, 244115.

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

(25) Louis, A. A.; Bolhuis, P. G.; Hansen, J. P.; Meijer, E. J. Can polymer coils be modeled as "soft colloids"? *Phys. Rev. Lett.* **2000**, *85*, 2522–2525.

(26) Murtola, T.; Falck, E.; Karttunen, M.; Vattulainen, I. Coarse-grained model for phospholipid/cholesterol bilayer employing inverse Monte Carlo with thermodynamic constraints. *J. Chem. Phys.* **2007**, *126*, 075101.

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

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

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

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

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

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

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

(34) Izvekov, S. Towards an understanding of many-particle effects in hydrophobic association in methane solutions. *J. Chem. Phys.* **2011**, *134*, 034104.

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

(36) Mirzoev, A.; Lyubartsev, A. P. Effective solvent mediated potentials of Na^+ and Cl^- ions in aqueous solution: temperature dependence. *Phys. Chem. Chem. Phys.* **2011**, *13*, 5722–5727.

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

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

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

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

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

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

(43) Pronk, S.; Pall, S.; Schulz, R.; Larsson, P.; Bjelkmar, P.; Apostolov, R.; Shirts, M. R.; Smith, J. C.; Kasson, P. M.; van der Spoel, D. et al. GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* **2013**, *29*, 845–854.

(44) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **2015**, *12*, 19 – 25.

(45) Plimpton, S. Fast parallel algorithms for short-range molecular dynamics. *J. Comp. Phys.* **1995**, *117*, 1 – 19.

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

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

(48) Rudzinski, J. F.; Noid, W. G. The role of many-body correlations in determining po-

tentials for coarse-grained models of equilibrium structure. *J. Phys. Chem. B* **2012**, *116*, 8621–8635.

(49) Tuckerman., M. E. *Statistical Mechanics: Theory and Molecular Simulation*; Oxford University Press: Oxford, Great Britain, 2013.

(50) Allen, M. P.; Tildesley, D. P. *Computer Simulation of Liquids*; Oxford Press: New York, NY USA, 1987.

(51) Martyna, G. J.; Tobias, D. J.; Klein, M. L. Constant pressure molecular dynamics algorithms. *J. Chem. Phys.* **1994**, *101*, 4177–4189.

(52) Hummer, G.; Gro/nbech-Jensen, N.; Neumann, M. Pressure calculation in polar and charged systems using Ewald summation: Results for the extended simple point charge model of water. *J. Chem. Phys.* **1998**, *109*, 2791–2797.

(53) Andersen, H. C. Molecular dynamics simulations at constant pressure and/or temperature. *J. Chem. Phys.* **1980**, *72*, 2384–2393.

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

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

(56) Davtyan, A.; Dama, J. F.; Voth, G. A.; Andersen, H. C. Dynamic force matching: A method for constructing dynamical coarse-grained models with realistic time dependence. *J. Chem. Phys.* **2015**, *142*, 154104.

(57) Wang, M. C.; Uhlenbeck, G. E. On the theory of the Brownian motion II. *Rev. Mod. Phys.* **1945**, *17*, 323–342.

(58) Liwo, A.; Ołdziej, S.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Scheraga, H. A. A united-residue force field for off-lattice protein-structure simulations. I. Functional

forms and parameters of long-range side-chain interaction potentials from protein crystal data. *J. Comp. Chem.* **1997**, *18*, 849–873.

(59) Akkermans, R. L. C.; Briels, W. J. A structure-based coarse-grained model for polymer melts. *J. Chem. Phys.* **2001**, *114*, 1020–1031.

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

(61) Ercolessi, F.; Adams, J. B. Interatomic potentials from first-principles calculations: The force-matching method. *Europhys. Lett.* **1994**, *26*, 583–588.

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

(63) Noid, W. G.; Chu, J.-W.; Ayton, G. S.; Voth, G. A. Multiscale coarse-graining and structural correlations: Connections to liquid state theory. *J. Phys. Chem. B* **2007**, *111*, 4116–4127.

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

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

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

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

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

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

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

(71) Hill, T. L. *An Introduction to Statistical Thermodynamics*; Addison Wesley Publishing Company, 1987.

(72) Mullinax, J. W.; Noid, W. G. Recovering physical potentials from a model protein databank. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 19867–19872.

(73) Lu, K.; Rudzinski, J. F.; Noid, W. G.; Milner, S. T.; Maranas, J. K. Scaling behavior and local structure of ion aggregates in single-ion conductors. *Soft Matter* **2014**, *10*, 978–989.

(74) Mullinax, J. W.; Noid, W. G. Reference state for the generalized Yvon-Born-Green theory: Application for a coarse-grained model of hydrophobic hydration. *J. Chem. Phys.* **2010**, *133*, 124107.

(75) Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P. *Numerical Recipes in FORTRAN: The Art of Scientific Computing*; Cambridge University Press: New York, NY USA, 1992.

(76) Liu, P.; Shi, Q.; Daume, H.; Voth, G. A. A Bayesian statistics approach to multiscale coarse graining. *J. Chem. Phys.* **2008**, *129*, 214114.

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

(78) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and testing of the OPLS All-Atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.* **1996**, *118*, 11225–11236.

(79) Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald: An $N \log(N)$ method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98*, 10089–10092.

(80) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **1984**, *81*, 3684–3690.

(81) Parrinello, M.; Rahman, A. Crystal structure and pair potentials: A molecular-dynamics study. *Phys. Rev. Lett.* **1980**, *45*, 1196–1199.

(82) Bussi, G.; Donadio, D.; Parrinello, M. Canonical sampling through velocity rescaling. *J. Chem. Phys.* **2007**, *126*, 014101.

(83) Martyna, G. J.; Tuckerman, M. E.; Tobias, D. J.; Klein, M. L. Explicit reversible integrators for extended systems dynamics. *Mol. Phys.* **1996**, *87*, 1117–1157.

(84) Martyna, G. J.; Klein, M. L.; Tuckerman, M. NoséHoover chains: The canonical ensemble via continuous dynamics. *J. Chem. Phys.* **1992**, *97*, 2635–2643.

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

(86) Sippl, M. J. Calculation of conformational ensembles from potentials of mean force: An approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol.* **1990**, *213*, 859–883.

(87) Skolnick, J. In quest of an empirical potential for protein structure prediction. *Curr. Opin. Struct. Biol.* **2006**, *16*, 166–171.

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

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

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

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

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

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

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

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

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

Tables

Table 1: Tools included in the BOCS toolkit with their primary inputs and outputs

Tool	Purpose	Input	Output
cgmap	Maps AA trajectory	AA trajectory, CG and map topologies	Mapped CG trajectory
cgff	Determines interaction potential, U_R	Mapped CG trajectory, CG potential definition	Interaction potential parameters, ϕ_D
tables	Converts CG potentials to GROMACS format	CG potential parameters	GROMACS table files
translate_table.py	Converts CG potentials to LAMMPS format	GROMACS table files	LAMMPS table files
pmatch	Determines volume potential, U_V	AA, CG pressures and volumes	Volume potential parameters, ψ_d
lmp_pmatch	Simulates CG model with $U = U_R + U_V$	LAMMPS table files, pressure correction	Simulated CG trajectory

Table 2: Contributions included in the interaction potential for each alkane system. Highlighted interactions correspond to XN potential functions that are employed in multiple alkane systems.

Molecule	Bonds	Angles	Nonbonded
Butane	CT-CT	-	**CT-CT**
Heptane	**CT-CM**	CT-CM-CT	**CT-CT**
	-	-	**CT-CM**
	-	-	**CM-CM**
Decane	**CT-CM**	CT-CM-CM	**CT-CT**
	CM-CM	-	**CT-CM**
	-	-	**CM-CM**

Table 3: Average corrections for the pressure and inverse compressibility, as well as the number of iterations required by self-consistent pressure-matching. Pressures and inverse compressibilities are given in units of 10^3 bar. The asterisk (*) indicates that the pressure correction did not converge within 10 iterations and was manually determined according to the procedure described in Section .

	$\langle F_V \rangle$		$\Delta\kappa_T^{-1}$		N_{Iter}	
System	MS-CG	XN	MS-CG	XN	MS-CG	XN
Butane	-0.36	0.033	-0.86	-0.67	*	1
Heptane	-0.77	-1.59	-1.57	-2.93	6	6
Decane	-3.15	-2.46	-6.23	-6.23	4	6
But/Dec Mix	-	-1.55	-	-3.50	-	3

Figures

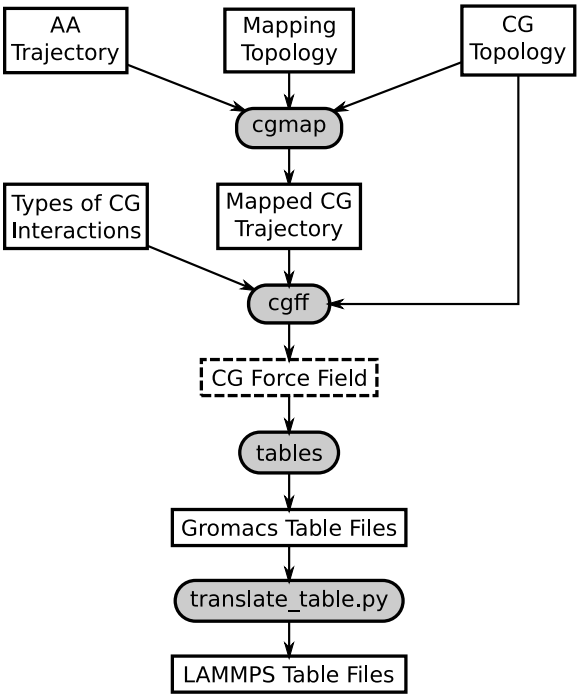

Figure 1

Workflow for the force-matching/g-YBG component of the BOCS toolkit. Boxes with sharp corners denote files, while boxes with rounded corners indicate operations performed on these files. Boxes filled with gray represent software tools provided in the BOCS toolkit. The dashed box indicates the major output of this workflow: the CG interaction potential, U_R.

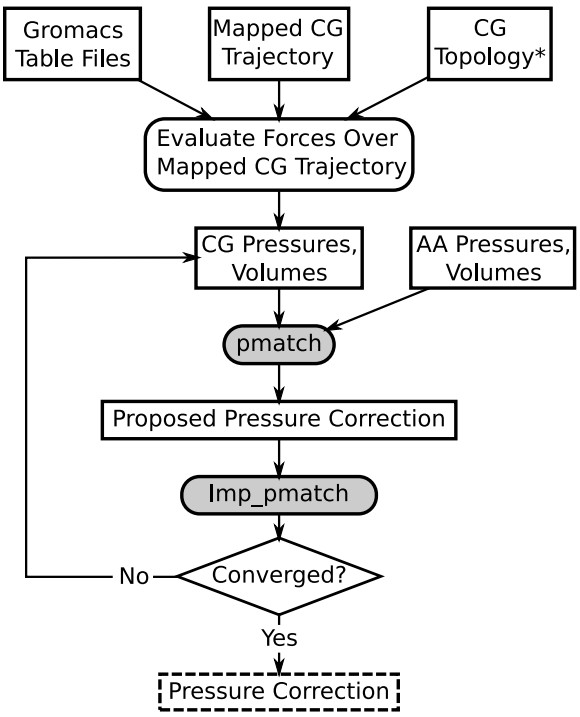

Figure 2
Workflow for the pressure-matching component of the the BOCS toolkit. See legend of Figure 1 for the meaning of the box shapes and outlines. The dashed box indicates the major output of this workflow: the CG volume potential, U_V.

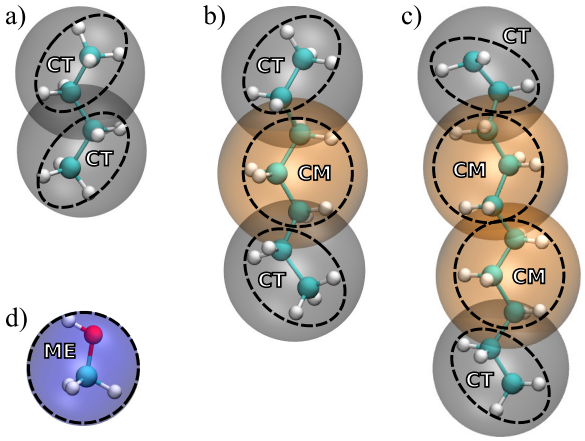

Figure 3
Mapping schemes for CG models superimposed upon the corresponding all-atom models, which are indicated in ball-and-stick representation. The CG sites (transparent spheres) are associated with the mass centers for the corresponding atomic groups, which are enclosed by the dashed circles. The size of the CG spheres indicates the distance at which the corresponding site-site radial distribution function vanishes, providing an estimate of the excluded volume for each site.

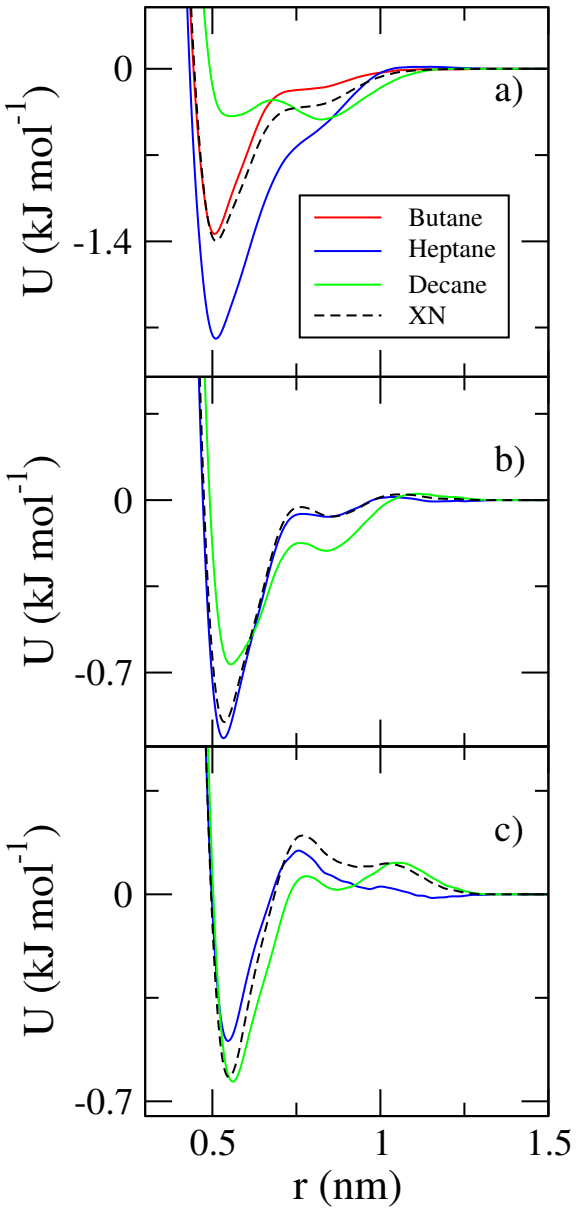

Figure 4
Calculated nonbonded potentials for a) CT-CT, b) CT-CM, c) CM-CM pair interactions. The solid red, blue, and green curves present MS-CG potentials calculated for butane, heptane, decane, respectively. The dashed black curves present the transferable XN potentials.

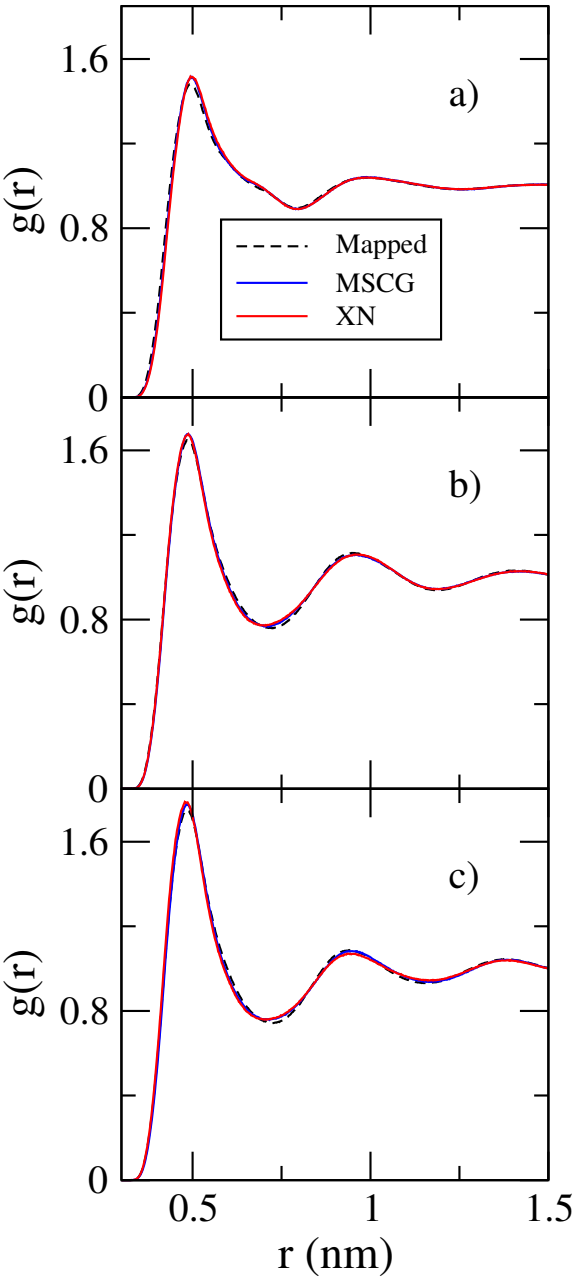

Figure 5
Radial distribution functions for the CT-CT pair interactions in a) butane, b) heptane, and c) decane. The dashed black, solid blue, and solid red curves present results for the mapped atomistic ensemble, the system-specific MS-CG model, and the transferable XN model, respectively.

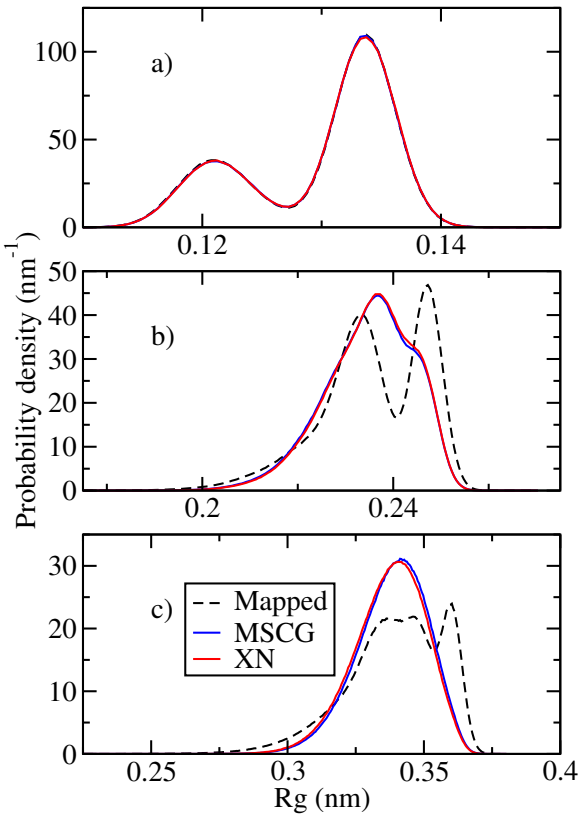

Figure 6
Probability distributions for the radius of gyration in a) butane, b) heptane, and c) decane. The dashed black, solid blue, and solid red curves present results for the mapped atomistic ensemble, the system-specific MS-CG model, and the transferable XN model, respectively.

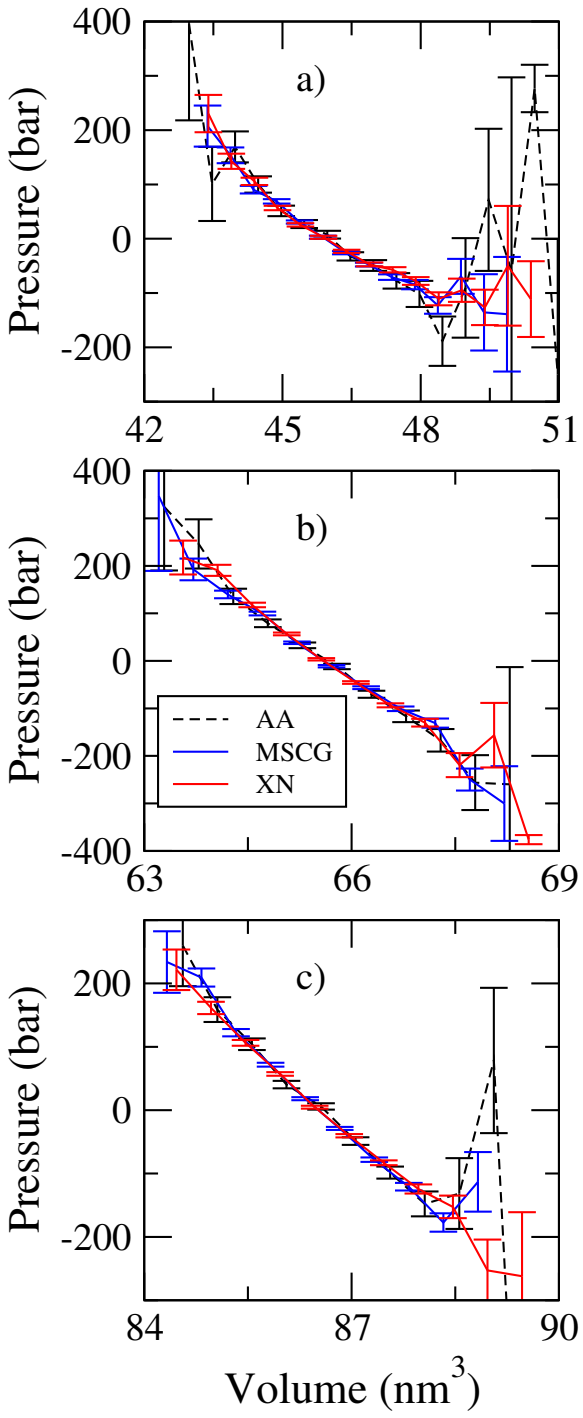

Figure 7
Simulated pressure-volume equations of state for a) butane, b) heptane, and c) decane. The error bars indicate the standard error of the corresponding bin. The dashed black, solid blue, and solid red curves present results for the mapped atomistic ensemble, the system-specific MS-CG model, and the transferable XN model, respectively.

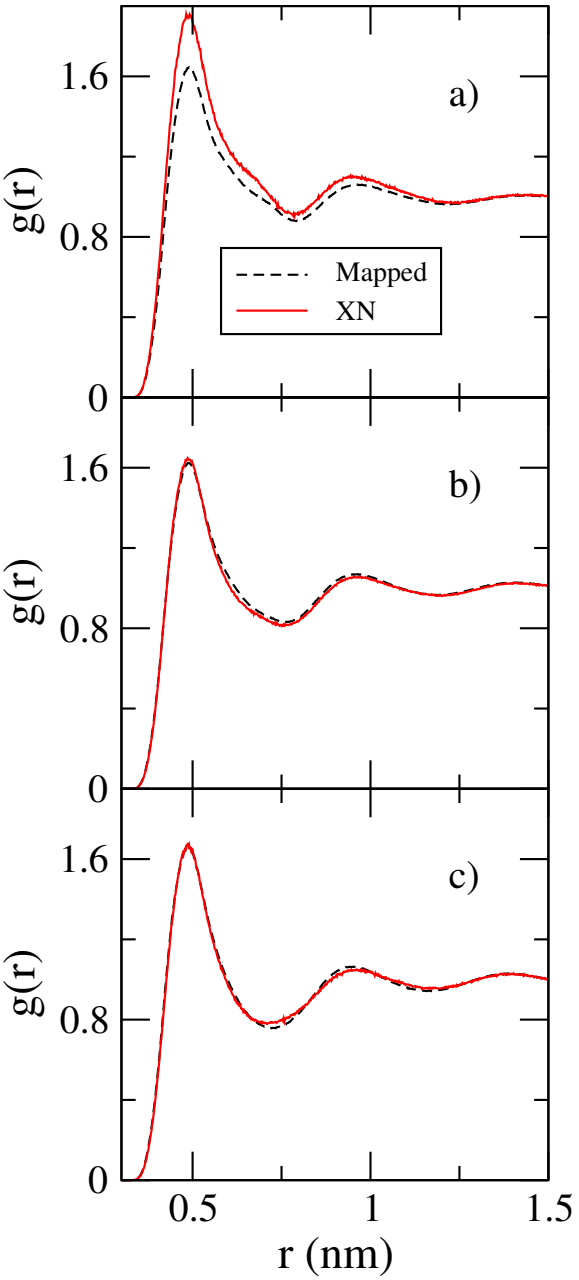

Figure 8
CT-CT radial distribution functions in the 50:50 butane-decane mixture for CT sites in a) butane-butane, b) butane-decane, and c) decane-decane pairs. The dashed black and solid red curves present results for the atomistic model and for the extended ensemble CG model, respectively.

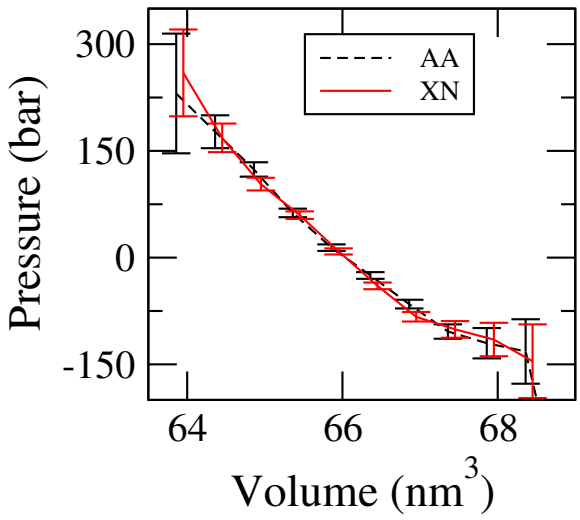

Figure 9
Pressure-volume equations of state for 50:50 butane-decane mixture. The error bars indicate the standard error of the corresponding bin. The dashed black and solid red curves present results for the atomistic model and for the extended ensemble CG model, respectively.

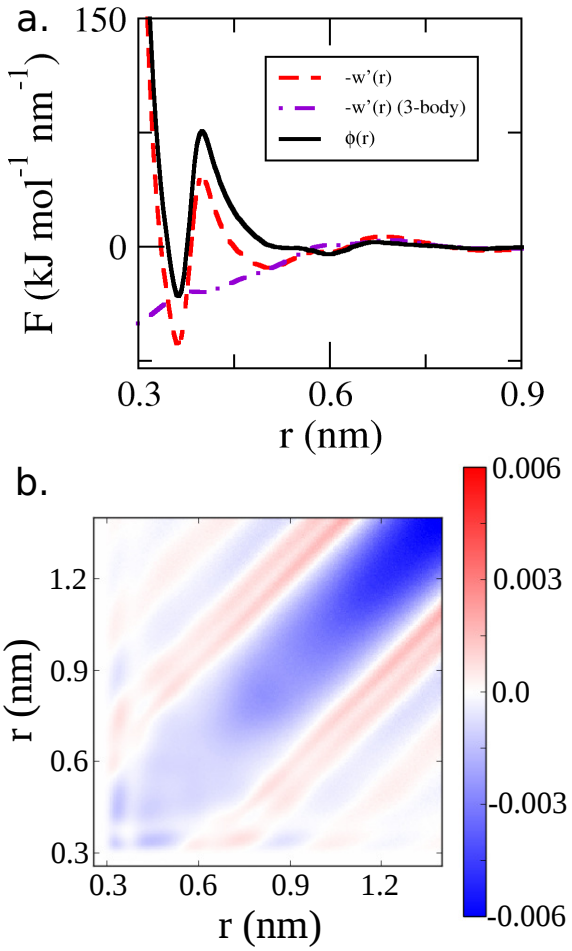

Figure 10

a) Contributions to the nonbonded ME-ME pair mean force for methanol. The solid black curve presents the MS-CG pair force, $F(r) = \phi(r)$, that minimizes χ_1^2. The dashed red curve presents the corresponding pair mean force, $-w'(r)$. The dashed-dotted purple curve presents the 3-body (indirect) contributions to the pair mean force. Panel b) presents the 3-body contributions to the metric tensor, $\bar{G}(r, r')$. Red and blue regions indicate positive and negative values, respectively.

TOC Graphic