Transferability of Local Density-Assisted Implicit Solvation Models for Homogeneous Fluid Mixtures

David Rosenberger, *,† Tanmoy Sanyal, ‡ M. Scott Shell, *,‡ and Nico F. A. van der Vegt †

†Eduard Zintl Institut für Anorganische und Physikalische Chemie, Technische Universität Darmstadt, Darmstadt, Germany
‡Department of Chemical Engineering, University of California Santa Barbara, Santa Barbara, California 93106, United States

ABSTRACT: The application of bottom-up coarse grained (CG) models to study the equilibrium mixing behavior of liquids is rather challenging, since these models can be significantly influenced by the density or the concentration of the state chosen during parametrization. This dependency leads to low transferability in density/concentration space and has been one of the major limitations in bottom-up coarse graining. Recent approaches proposed to tackle this shortcoming range from the addition of thermodynamic constraints, to an extended ensemble parametrization, to the addition of supplementary terms to the system Hamiltonian. To study fluid phase equilibria with bottom-up CG models, the application of local density (LD) potentials appears to be a promising approach, as shown in previous work by Sanyal and Shell [T. Sanyal, M. S. Shell, J. Phys. Chem. B, 2018, 122, 5678]. Here, we want to further explore this method and test its ability to model a system which contains structural inhomogeneities only on the molecular scale, namely, solutions of methanol and water. We find that a water–water LD potential improves the transferability of an implicit-methanol CG model toward high water concentration. Conversely, a methanol–methanol LD potential does not significantly improve the transferability of an implicit-water CG model toward high methanol concentration. These differences appear due to the presence of cooperative interactions in water at high concentrations that the LD potentials can capture. In addition, we compare two different approaches to derive our CG models, namely, relative entropy optimization and the Inverse Monte Carlo method, and formally demonstrate under which analytical and numerical assumptions these two methods yield equivalent results.

1. INTRODUCTION

Our understanding of the driving forces behind processes in soft condensed matter has greatly benefited from computer simulations and from molecular dynamics (MD) simulations in particular. This technique allows an atomistic view into complex systems but is limited by computational overhead to modeling length and time scales of tens of nanometers and microseconds, respectively. This hurdle can be overcome with coarse-grained (CG) particle models whose number of degrees of freedom (DOFs) and interactions to be evaluated are significantly smaller compared with fine-grained (FG) models. Apart from their computational advantages in computer simulations, CG models, and methods to derive them, provide important additional merits. Efforts to make models as simple as possible provide additional insight into emergent driving forces that may not be as easily obtained based on FG models alone. Systematic removal of DOFs involved in deriving a CG model based on its FG counterpart is the basis for bottom-up or systematic coarse graining. 1−3 To reduce DOFs, the high resolution, or FG, configuration space is projected onto a CG configuration space in resolution. In order to evaluate the corresponding free energy surface, the so-called multibody potential of mean force (PMF) must be calculated,

\[ W(R) = -k_B T \ln \int_V d\mathbf{r} \exp(-\beta U_{FG}(\mathbf{r})) \delta(\mathbf{R} - \mathbf{M}(\mathbf{r})) \] (1)

where \( W(R) \) is the multibody PMF and \( \beta = 1/k_B T \) with \( k_B \) the Boltzmann constant and \( T \) the temperature. \( \mathbf{M} \) is the projection, or mapping, operator, which relates a FG configuration \( \mathbf{r} \) to a CG one \( \mathbf{R} \). Due to the highly multibody nature of the integral, \( W(R) \) is too complicated to compute exactly in practice, and only approximate solutions are possible. Solving an inverse problem is one way to obtain such an approximate solution to the multibody PMF. The goal is then to find a CG model that accurately reproduces one or multiple quantities of the FG model in the CG configuration space by minimizing the difference between the FG and CG configuration space with respect to the quantity chosen. This inverse problem can be solved by means of either a variational principle 4−5 or by application of iterative Newton or quasi-Newton inversion techniques. 10−13

Despite the methodological differences, all bottom-down coarse-grained models suffer from the same two fundamental problems: accurate representability and state point transferability. The representability characterizes the ability of a CG model to simultaneously reproduce multiple properties of the FG system like structure, pressure, and isothermal compressibility. The transferability instead describes the applicability of

Received: November 19, 2018
Published: April 17, 2019
CG models at state points not included in the parametrization. Accurate representability is guaranteed for the target property chosen to solve the inverse problem, but it is not guaranteed for other properties. That is, a match in structure does not automatically guarantee a match in thermodynamics or dynamics, or a match in forces does not guarantee structural agreement.\(^\text{14–18}\) An equally outstanding challenge is the transferability of CG models.\(^\text{19–22}\) The main reason why it is difficult to achieve both is that entropic contributions from particles “lost” upon coarse graining are missing. Therefore, the generated effective pair potentials cannot capture changes in entropy necessary to describe certain thermodynamic properties and necessary to be transferable to different state points.\(^\text{26,23,24}\) Several approaches have been proposed to tackle both representability and transferability. Among the most common are the addition of thermodynamic constraints to account for accurate pressure,\(^\text{22}\) Kirkwood–Buff integrals,\(^\text{25}\) or the surface tension,\(^\text{26,27}\) the application of an extended ensemble parametrization,\(^\text{28,29}\) or the use of a extended Hamiltonian description for the energy of the system.\(^\text{30–35}\)

Allen and Rutledge proposed the idea of local density (LD)-dependent interactions to improve implicit solvent models.\(^\text{20}\) Interactions based on the LD of CG sites have been used to improve the transferability\(^\text{36}\) and the representability\(^\text{37}\) of CG models. The local density of a CG site is simply a weighted local co-ordination number around that site and can be written generally for arbitrary combinations of central and neighboring types of sites. The LD potential can then be cast as a function of the local density and added as a corrective extension to the traditional pairwise form of the Hamiltonian in CG models. Unlike pairwise Hamiltonians, a LD potential incorporates information about the inherently multibody environment around CG sites which contributes to enhanced model transferability. Further, it is a mean-field potential with computational complexity similar to that of pair potentials and thus does not sacrifice computational speed (further implementation details can be found in ref 32). LD potentials have been used to improve the sampling of conformation space in implicit solvent models of superhydrophobic polymers and were found to enhance the model’s transferability to different polymer lengths.\(^\text{32}\) Recently, they have also been shown to improve structural transferability in CG models of liquid mixtures such as benzene in water.\(^\text{44}\)

In this work, we want to test the ability of LD-dependent CG models to quantitatively and qualitatively describe mixtures of water and methanol. These mixtures provide an interesting test case for the LD-dependent potentials, since they show strong microheterogeneities at atomistic length scales as a function of methanol concentration, while remaining miscible at a macroscopic scale.\(^\text{38–43}\) Laaksonen et al. showed that these microheterogeneities are caused by a nonhomogeneous distribution of the two components in the mixture.\(^\text{44}\) This is expressed through structural patterns determined by the dominant component in the system. X-ray emission spectroscopy experiments applied by Guo et al. revealed that the inhomogeneous mixing between methanol and water can be explained by the formation of rings of methanol bridged by water molecules.\(^\text{45}\) In agreement, Perera et al. found methanol molecules forming chain-like structures, caused by water bridging the hydroxyl groups of the methanol molecules.\(^\text{45}\) Further, Pascal and Goddard confirmed the picture of incomplete mixing.\(^\text{46}\) At low methanol concentrations, methanol molecules bury their hydrophobic groups away from water. With increasing methanol concentration, this is no longer possible, and free mixing is observed. At high methanol concentrations, the system is best described as water dissolved in methanol.\(^\text{46}\)

Classical MD simulations of methanol–water mixtures have almost exclusively been based on all-atom force field models.\(^\text{38,44–46}\) It is, however, interesting to ask if CG particle models can equally well be used to describe the structural properties of these systems. This question may provide insights to modeling large-scale phenomena driven by the interplay between hydrophobic and hydrophilic interactions in complex systems not amenable to all-atom models. In this study, we investigate the possibility to study this interplay with simple CG models. We do this by explicitly accounting for LD effects and examine if LD potentials can effectively describe the microheterogeneities observed in water–methanol mixtures with a simple single site CG model for both liquids, water, and methanol in an implicit solvent environment. We derive two different CG models: (I) CG methanol in implicit water and (II) CG water in implicit methanol. Further, we compare two different methods to generate bottom-up CG models, namely, Inverse Monte Carlo (IMC)\(^\text{11}\) and relative entropy optimization,\(^\text{12}\) and we show analytically and numerically under which assumptions these two methods are equivalent.

The remainder of the article is structured as follows: first, the basic theoretical background on IMC and relative entropy optimization is given. Second, we prove analytically under which assumptions IMC and the relative entropy method are equivalent. Next, we briefly discuss the extension of the relative entropy method to LD potentials, followed by the details of the numerical calculations performed. Adjacent, we present the main results of this study along with a detailed discussion, followed by the conclusion.

2. METHODS

2.1. Inverse Monte Carlo. The Inverse Monte Carlo (IMC) Method, or Newton inversion method (introduced by Lyubartsev and Laaksonen), aims to derive a CG force field (FF) that reproduces the pairwise structure, i.e., the radial distribution function (RDF), of the underlying atomistic or fine grained (FG) system.\(^\text{41}\) The CG FF is estimated initially as the two-body potential of mean force (PMF, \(U^0(r_j)\)) acting between two particles \(i\) and \(j\) along the distance \(r_j\) obtained from the corresponding RDF of the FG system \(g_0^0(r_j)\):

\[
U^0(r_j) = -k_BT \ln g_0^0(r_j)
\]

(2)

In many cases, the PMF does not accurately resemble the effective pair potential in the CG configuration space due to the relevance of higher order correlations. Thus, the potential is updated a series of times \(n\), solving a set of linear equations until the difference in the RDF is minimized. This leads to numerical pair potentials.

The set of linear equations is given by eq 3, where \(N_p\) is the number of particle pairs at a distance \(\alpha\) either in the CG system or in the mapped reference system \((N_\alpha^0)\), \(J\) is a Jacobian matrix, \(AU\) is the potential update, and \(\gamma\) is a particle pair distance > \(\alpha\). The Jacobian matrix is defined in eq 4, where \(N_p^0\) is the number of particle pairs separated by a distance \(\gamma\). Finally, the update of the PMF gets computed according to eq 5.

\[
\langle N_\alpha^0 \rangle_{\text{CG}} - N_\alpha^0 = J_{\alpha,\gamma} \Delta U_\gamma
\]

(3)
\[ J = \frac{\partial (N^\alpha_{\text{MC}})}{\partial U^\alpha} = -\beta (\langle N_a N_r \rangle - \langle N_a \rangle \langle N_r \rangle) \] (4)

\[ U^\alpha(q) = U^\alpha(q) - \frac{1}{2} \sum_{\alpha=1}^{N} \sum_{\beta=1}^{N} q_{\alpha} q_{\beta} g^\alpha(\epsilon_{\alpha}) \] (5)

\[ \langle N_a \rangle = \frac{N(N-1)}{2} \frac{4\pi r^3}{V} g^\alpha(r) \] (6)

where \( V \) is the volume of the system, \( N \) the number of particles, and \( \Delta r \) is the discretization grid spacing. The Jacobian defined in eq 4 explicitly contains cross correlations between the number of particle pairs at different distances. This provides several advantages in converging toward the final effective pair potentials and some disadvantages in terms of numerical stability of IMC, compared to the similar Iterative Boltzmann Inversion method as discussed in the literature.

2.2. Relative Entropy Optimization for Pair Potentials. Another way to determine a CG FF is through relative entropy optimization, as proposed by Shell. The relative entropy is a quantity that measures the information loss upon reducing the FG system to the CG model. It is defined as

\[ S_{\text{rel}} = \sum p(i) \ln \left( \frac{p(i)}{p_{\text{CG}}(i)} \right) + S_{\text{map}} \] (7)

where \( p \) is the probability to observe a certain configuration \( i \) determined either by the FG FF \( p_0 \) or by the CG FF \( p_{\text{CG}} \), and \( S_{\text{map}} \) is a mapping entropy which accounts for the degeneracy of atomic states in the CG configuration space. In the canonical ensemble, the relative entropy can be expressed as

\[ S_{\text{rel}} = \beta (U_{\text{CG}}(\xi) - U_{\text{FG}}) - \beta (A_{\text{CG}}(\xi) - A_{\text{FG}}) + S_{\text{map}} \] (8)

where \( U \) is the potential energy, \( A \) is the Helmholtz free energy, and \( \xi \) is a vector that contains all parameters of the CG FF. The idea in relative entropy optimization is to find the optimal set of parameters \( \xi \) that minimize the information loss between the FG and CG system. Similar to IMC, where the pair potential \( U \) is updated, in relative entropy optimization the FF parameters \( \xi \) are updated until a minimum in the relative entropy is reached. Using a Newton–Raphson approach, the parameter update scheme is

\[ \xi^k = \xi^{k-1} - H^{-1} \nabla S_{\text{rel}} \] (9)

In eq 9, \( H \) is the Hessian matrix, and \( \nabla S_{\text{rel}} \) is the gradient of the relative entropy. Sequential iterations successively bring the parameters to a local \( S_{\text{rel}} \) minimum. This scheme is only applicable if the Hessian is positive definite, otherwise a steepest descent or conjugate gradient optimization scheme is applied. It is interesting to note that if \( U \) consists of splines or tabulated potentials (as in IMC), then the relative entropy has a single global minimum and eq 9 always applies.

2.3. Equivalence between Inverse Monte Carlo and Relative Entropy Optimization. The two methods, IMC and relative entropy optimization (when using a Newton–Raphson update scheme), are equivalent for the derivation of effective pair potentials if the following applies:

\[ \nabla S_{\text{rel}} = \left( \langle N_a \rangle_{\text{CG}} - \langle N_a \rangle_{\text{FG}} \right) \] (10)

\[ H^{-1} = J^{-1} \] (11)

For simplicity, we assume a system with only one component, but the equations can be easily extended to multicomponent systems as well.

As proposed by Lyubartsev and Laaksonen, we start with a discretized Hamiltonian \( U \) to describe the potential energy of the system

\[ U = \sum a \langle N_a \rangle_{\zeta_a} \] (12)

where \( \zeta_a \) is the pair potential, and \( N_a \) is the exact number of particle pairs at distance \( \alpha \), given by

\[ \langle N_a \rangle = \int dq N_a \prod_i \exp(-\beta N_a(q) \zeta_a) \] (13)

The Jacobian in eq 4 is then given by

\[ \frac{\partial \langle N_a \rangle}{\partial \zeta_a} = \frac{\partial}{\partial \zeta_a} \left( \int dq N_a \prod_i \exp(-\beta N_a(q) \zeta_a) \right) \] (14)

which by application of the chain rule results in

\[ J = \frac{\partial \langle N_a \rangle}{\partial \zeta_a} = -\beta (\langle N_a N_r \rangle - \langle N_a \rangle \langle N_r \rangle) \] (15)

By inserting eq 15 in eq 3, the potential update, \( \Delta U_{\alpha} \) in IMC is computed by solving

\[ \langle N_a \rangle_{\text{CG}} - \langle N_a \rangle_{\text{FG}} = -\beta (\langle N_a N_r \rangle - \langle N_a \rangle \langle N_r \rangle) \Delta U_{\alpha} \] (16)

where we define the left-hand side (l.h.s.) as

\[ \Delta N_{a}^{\text{IMC}} \equiv \langle N_a \rangle_{\text{CG}} - \langle N_a \rangle_{\text{FG}} \] (17)

Next, we evaluate the relative entropy in eq 8. Its derivative can be written as

\[ \nabla S_{\text{rel}} = \beta \left( \frac{\partial U}{\partial \xi} \right)_{\text{FG}} - \beta \left( \frac{\partial U}{\partial \xi} \right)_{\text{CG}} \] (18)

Substituting \( U \) defined by eq 12, we obtain

\[ \nabla S_{\text{rel}} = \langle N_a \rangle_{\text{CG}} - \langle N_a \rangle_{\text{FG}} \equiv \Delta N_{a}^{\text{rel}} \] (19)

In comparison with eq 17, the result of eq 19 reveals the following relation:

\[ -\Delta N_{a}^{\text{rel}} = \Delta N_{a}^{\text{IMC}} \] (20)

The Hessian matrix in eq 9 is given by

\[ H = \left( \frac{\partial^2 U}{\partial \xi_a \partial \xi_b} \right)_{\text{FG}} - \left( \frac{\partial^2 U}{\partial \xi_a \partial \xi_b} \right)_{\text{CG}} + \beta \left( \frac{\partial U}{\partial \xi_a} \frac{\partial U}{\partial \xi_b} \right)_{\text{CG}} \] (21)

where the first two terms vanish due to the linearity of \( U \) in the parameters. The linearization that is exploited here results from using cubic splines for the effective pair potentials in the relative entropy method. The fact that the CG pair potential is
linear in its parameters (spline knots) is crucial to the success of a simple scheme like Newton–Raphson descent in discovering a global minimum on the relative entropy surface which then theoretically guarantees robust representability at least for the pair correlations. Using the Hamiltonian defined in eq 12, the remainder is

$$H = \beta(N_f N_f) - \beta(N_f) (N_f)$$

(22)

which leads to a similar relation as in eq 20

$$H^{-1} = -J^{-1}$$

(23)

Taking these results and applying the definition for the potential update (see eqs 5 and 9) we end with the following relation:

$$H^{-1}(-\Delta N_a^{\text{SG}}) = -J^{-1} \Delta N_a^{\text{IMC}}$$

(24)

$$H^{-1} \Delta N_a^{\text{SG}} = J^{-1} \Delta N_a^{\text{IMC}}$$

(25)

This proves that under the assumption of a discretized Hamiltonian (tabulated in IMC and represented with splines in the relative entropy method) quasi-Newton optimization strategies, like Newton–Raphson, lead to an exact equivalence of the pair potentials obtained through IMC and relative entropy. Moreover, this shows that relative entropy optimization leads to a match in the RDF between the FG and CG systems without directly using it as a target quantity in the optimization process. It is important to note that the agreement between IMC and relative entropy minimization is likely to also hold when very knot-dense spline potentials are employed in the latter (instead of discretized ones), although the necessary knot density and discretization to observe quantitative agreement may be high.

2.4. Relative Entropy Optimization for Local Density Potentials. Recently Sanay and Shell applied the relative entropy approach not only for pair potentials but also for so-called local density (LD) potentials,32,34 LD potentials ($U_{LD}$) account for the effect of neighboring particles on the effective pair potential ($u_{pair}$). This additional contribution changes the total CG potential energy to

$$U_{CG} = \sum_{i<j} u_{pair} + U_{LD}$$

(26)

Here, $U_{LD}$ is a sum over an unspecified function of the local density around each particle $i$ in the system

$$U_{LD} = \sum_i f(\rho_i)$$

(27)

where $f(\rho_i)$ is practically represented using cubic B splines. Here, the local density $\rho_i$, $\rho_c$ is the total number of neighboring particles within a specified and smoothed cutoff ($r_c$)

$$\rho_i = \sum_{j \neq i} \phi(r_{ij})$$

(28)

where the indicator function ($\phi$) adopts a value of 1 below an inner cutoff $r_0$ but continuously and quickly decays to 0 at $r_c$. The shape of this function is chosen to be computationally convenient and does not require the calculation of the absolute distance between pairs of particles (or any square root operations).

The difference between $r_0$ and $r_c$ are called, $\Delta$ and is of the order of 0.1–0.12 nm. The exact form of the coefficients $\epsilon$ as well as more details on the LD potentials can be found in the original work by Sanay and Shell and the recent extension to binary mixtures.32,34

3. SIMULATION DETAILS

FG and IMC CG simulations are performed with the Gromacs-5.1.2 MD engine.52,53 IMC potentials are generated with the VOTCA coarse-graining package (version 1.4).47,54 The relative entropy optimization is achieved with an in-house code. To account for local density potentials, CG simulations are executed with a modified version of the LAMMPS simulation package that includes a custom local density potential.55

3.1. Fine-Grained Simulations. All systems studied contain 5000 molecules in total, with different methanol mole fractions, $x_{methanol}$ of interest, namely, 0.1, 0.5, and 0.9 in SPC/E water.56,57 For FG methanol, a Kirkwood–Buff-based force field for united atoms is used.36,66 Newton’s equations of motion are integrated based on a leapfrog algorithm with a time step of 1 fs. All systems are equilibrated for 2 ns at a constant pressure of 1 bar and at a constant temperature of 300 K (NPT condition). For both the barostat and thermostat, the weak coupling method of Berendsen is applied with a coupling constant of $\tau = 1$ ps for the barostat and $\tau = 0.5$ ps for the thermostat.59 The barostat compressibility is $4.5 \times 10^{-3}$ bar$^{-1}$. For the short-range van der Waals interactions, a cutoff of 1.2 nm is applied with a long-range dispersion correction. Electrostatic interactions are treated with the particle-mesh-Ewald method60 with a real space cutoff of 1.2 nm and a grid size of 0.12 nm. Bonded interactions in methanol are constrained with the LINCS algorithm.61,62 Periodic boundary conditions are applied in the $x$, $y$, and $z$ directions. The short equilibration is followed by a 10 ns run under NPT conditions replacing the Berendsen barostat and thermostat with the Parrinello–Rahman barostat63 and the Nosé–Hoover thermostat,64 respectively. The coupling constants are set to $\tau = 1$ ps for the barostat and $\tau = 0.5$ ps thermostat. All other parameters are kept the same during the short equilibration run. The average volume of the 10 ns NPT simulation is then used for the final production run under constant volume (NVT) conditions. Besides the volume constraint, all other parameters are the same as during the 10 ns NPT run.

3.2. Implicit Solvent Model. To generate the CG model for the united atom model of methanol in implicit SPC/E water56,57 and for SPC/E water66,57 in implicit methanol, a 3 to 1 mapping scheme is applied, where each molecule is mapped to its center of mass as illustrated in Figure 1.

3.2.1. Inverse Monte Carlo Optimization. On the basis of the FG RDFs between the centers of mass of water and water (WW) and methanol and methanol (MM) at $x_{methanol} = 0.5$, implicit solvent models for CG water and CG methanol are derived with the IMC method. As an initial guess for the iterative procedure, the PMF is taken (see eq 2). To generate the CG configuration of the atomistic model, water and methanol are mapped as illustrated in Figure 1, and depending on the case—
Newton’s equations of motion are integrated according to the velocity-Verlet algorithm with a time step of 1 fs. A Langevin thermostat is applied with a coupling constant of 0.1 ps. The cutoff value for the REO potentials as well as the values for the outer cutoff \( (r_c^i, \text{eq 29}) \) and the difference between the outer and inner cutoff of the LD potentials (\( \Delta \)) are listed in Table 1.

<table>
<thead>
<tr>
<th>interaction</th>
<th>type</th>
<th>( r_c ) (nm)</th>
<th>( \Delta ) (nm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>methanol–methanol (MM)</td>
<td>REO pair potentials</td>
<td>1.5</td>
<td></td>
</tr>
<tr>
<td>water–water (WW)</td>
<td>REO LD potentials</td>
<td>0.63</td>
<td>0.1</td>
</tr>
<tr>
<td>methanol–methanol (MM)</td>
<td>REO pair potentials</td>
<td>1.2</td>
<td></td>
</tr>
<tr>
<td>water–methanol (WM)</td>
<td>REO LD potentials</td>
<td>0.34</td>
<td>0.1</td>
</tr>
</tbody>
</table>

3.2.3. Coarse-Grained Simulations. On the basis of the derived CG models, MD simulations are performed for 10 ns under NVT conditions at three different mole fractions of methanol \( x_M = 0.1, x_M = 0.5, \) and \( x_M = 0.9 \). The simulation parameters are the same as the ones used for the MD production phases in the iterative optimization approaches for IMC and relative entropy optimization.

4. RESULTS AND DISCUSSION

4.1. Kirkwood–Buff Analysis of the Fine-Grained System. In order to compare the derived CG models with their parent reference systems beyond structural accuracy, we also compute Kirkwood–Buff integrals (KBIs). KBIs relate local structure and thermodynamic properties like activity coefficients, solvation free energies, or the isothermal compressibility of stable mixtures.\(^{65}\) For mixture components \( i \) and \( j \), they are defined as\(^ {66}\)

\[
G_{ij} = 4\pi \int_0^\infty [g_{ij}(s) - 1]s^2ds
\]  

KBIs can be interpreted as the excess coordination number of particles \( j \) around a central particle \( i \). This means that the larger the KBI value is, the higher is the affinity between particles \( i \) and \( j \). To evaluate eq 30 in computer simulations, a thermodynamic limit needs to be taken, while the system should be open with respect to its components. Notwithstanding the latter requirement, KBIs can be calculated in computer simulations of closed (NVT or NPT) systems. To this end, the integral in eq 30 is usually truncated at an upper integration limit \( r < L/2 \) (with \( L \) the linear simulation box dimension) where \( G_{ij}(r) \) (the running KBI or RKBI) is observed to oscillate around a mean plateau value. In recent work, it has been shown that this mean value corresponds to the thermodynamic limiting value of eq 30. A more detailed discussion on the issue of finite size and ensemble effects is beyond the scope of this work and can be found elsewhere.\(^ {67–69}\)

By applying eq 30, one critical problem occurs that should not be ignored: RDFs do not strictly approach a limiting value of 1 in closed systems. This leads to a drift in the asymptotic behavior of the RKBIs. The drift is caused by depletion or accumulation of particles \( j \) around a particle \( i \) at local scales. This local depletion or accumulation is then compensated by a positive or negative excess of particles \( j \) at long distances, since the total number of particles \( j \) is constant. This leads to
incorrect limiting behavior \((r \to \infty)\) of \(g_j(r)\), which must be corrected. We here use the empirical correction introduced by Ganguly and van der Vegt\textsuperscript{22}:

\[
g_{ij}^{corr}(r) = g_{ij}(r) \frac{N\left(1 - \frac{(4/3)\pi r^3}{V}\right)}{N\left(1 - \frac{(4/3)\pi r^3}{V}\right) - \Delta N_j(r) - \delta_{ij}} \quad (31)
\]

where \(N_j\) is the number of particles \(j\), \(V\) is the volume of the system, \(\Delta N_j\) is the excess number of particles \(j\) around a particle \(i\) within a sphere of radius \(r\), and \(\delta_{ij}\) is the Kronecker delta. KBIs obtained from \(g_{ij}^{corr}(r)\) are named Ganguly (GKBI) in the following.

Figure 2 shows the RKBI of the FG simulations at \(x_M = 0.5\). One clearly sees a drift in the asymptotic behavior of the water–water (WW) RKBI (solid black line) and how the empirical correction of Ganguly (solid red line) shifts the RKBI to larger values at longer distances. This is also observed for methanol–methanol (MM) (dotted lines). In case of methanol–water (MW, dashed lines), one observes that the RKBI are already well converged without any correction. Here, the empirical correction of Ganguly introduces a small shift in the tails of the KBIs toward lower values. Now, if one compares the RKBI values averaged between 1.0 and 1.5 nm, the influence of the Ganguly correction is fairly small, and the effect of the empirical correction becomes important only at larger distances.

The average values for all KBIs obtained from FG simulations are presented in Table 2. The averages are taken between 1.0 and 1.5 nm, and the errors are calculated by averaging over five independent simulations according to

\[
\text{err} = \frac{\sqrt{\sum_{i=1}^{N} (G_i - \langle G \rangle)^2}}{N(N-1)} \quad (32)
\]

where \(N = 5\) is the number of simulations. As one sees, all KBI values are nearly the same no matter if the Ganguly correction is applied or not. The largest discrepancy between the RKBI and GKBI occurs for WW at \(x_M = 0.9\). Further, one observes a minimum in \(G_{MM}\) at \(x_M = 0.1\), which means that at low methanol concentrations the methanol molecules are mutually

<table>
<thead>
<tr>
<th>(G_{WW})</th>
<th>(x_M)</th>
<th>RKBI (nm(^3))</th>
<th>GKBI (nm(^3))</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.5</td>
<td>-0.025 ± 0.042 (10^{-3})</td>
<td>-0.025 ± 0.055 (10^{-3})</td>
</tr>
<tr>
<td>0.5</td>
<td>0.5</td>
<td>0.021 ± 0.033 (10^{-3})</td>
<td>0.025 ± 0.055 (10^{-3})</td>
</tr>
<tr>
<td>0.9</td>
<td>0.5</td>
<td>-0.02 ± 0.018 (10^{-3})</td>
<td>0.01 ± 0.050 (10^{-3})</td>
</tr>
<tr>
<td>(G_{MM})</td>
<td>0.1</td>
<td>-0.039 ± 0.034 (10^{-3})</td>
<td>-0.043 ± 0.057 (10^{-3})</td>
</tr>
<tr>
<td>0.5</td>
<td>0.5</td>
<td>0.041 ± 0.034 (10^{-3})</td>
<td>0.043 ± 0.057 (10^{-3})</td>
</tr>
<tr>
<td>0.9</td>
<td>0.5</td>
<td>-0.016 ± 0.030 (10^{-3})</td>
<td>-0.016 ± 0.040 (10^{-3})</td>
</tr>
<tr>
<td>(G_{MM})</td>
<td>0.1</td>
<td>-0.16± 0.003 (10^{-2})</td>
<td>-0.14 ± 0.030 (10^{-2})</td>
</tr>
<tr>
<td>0.5</td>
<td>0.5</td>
<td>-0.078 ± 0.003 (10^{-2})</td>
<td>-0.077 ± 0.040 (10^{-2})</td>
</tr>
<tr>
<td>0.9</td>
<td>0.5</td>
<td>-0.075 ± 0.002 (10^{-2})</td>
<td>-0.075 ± 0.020 (10^{-2})</td>
</tr>
</tbody>
</table>

stronger depleted than at higher concentrations. This is in agreement with the work of Laaksonen et al.\textsuperscript{44} who showed that at low \(x_M\) the methanol molecules are separated by larger distances. Going from \(x_M = 0.1\) to \(x_M = 0.5\), \(G_{WW}\) and \(G_{MM}\) become larger (less negative). \(G_{WW}\) even changes sign and becomes positive. This indicates stronger water–water and methanol–methanol association in the \(x_M = 0.5\) mixture. Upon further increasing the methanol content from \(x_M = 0.5\) to \(x_M = 0.9\), the values of \(G_{MM}\) in turn become larger, while \(G_{WW}\) becomes smaller. Thus, at high mole fractions of methanol, the water molecules preferentially interact with methanol molecules. This observation agrees with the work of Pascal and Goddard.\textsuperscript{46} The increase in \(G_{WW}\) upon raising the methanol concentration is also supported by other studies, which found that small water aggregates are formed when more methanol molecules are present in the system.\textsuperscript{44--46} We note that the values of \(G_{MM}\) at \(x_M = 0.1\), and of \(G_{WW}\) at \(x_M = 0.9\), suffer from large uncertainties. This indicates that the KBIs are not fully converged, and much longer sampling is needed.\textsuperscript{22} For the purpose of this study, we compare the FG model with the different CG models within the range of these uncertainties, being aware that a quantitative match might be difficult to achieve.

4.2. Effective Potentials for Different Implicit Solvent Models. The implicit solvent models are generated as described in the simulation details. The final CG potentials are presented in Figure 3. Figure 3a shows the final effective pair potentials for the methanol–methanol (MM) interaction. As theoretically derived in Section 2.3 on the equality between IMC and REO, REO (dashed green line) and IMC (solid red line) lead to very nearly the same set of pair potentials. The REO LD model shows a slightly smaller second maximum in the pair potential, which is illustrated through the dashed blue curve in Figure 3a. This shift is caused by the additional MM LD potential, presented in Figure 3c. This potential is small in magnitude but weakly attractive and lowers the potential barrier in the pair potential. Notably, the small magnitude of the LD potential reveals only a weak multibody nature of the MM interaction at this state point.

The CG potentials for water in implicit methanol are shown in Figure 3b. Similar to CG methanol in implicit water, the effective pair potentials obtained from IMC (solid red line) and from the pair-only relative entropy optimization (dashed green line) nearly overlap, with a small difference between the first maximum and the second minimum. Interestingly, the pair

Figure 2. Influence of the Ganguly correction on the convergence of the Kirkwood–Buff integrals in the AA system for water–water (solid lines), methanol–water (dashed lines), and methanol–methanol (dotted lines) at a mole fraction of \(x_M = 0.5\);: uncorrected RKBI (black) and the RDF correction of Ganguly, GKBI, (red).
potential of the REO LD model (dashed blue line) lacks the inner potential well present in the other two models, but the absence of an attractive well is compensated by the additional WW LD potential (Figure 3d). This points toward a strong coupling between the pair and LD potentials, which may mean that the attractive interactions are more naturally captured at a mean-field multibody level (LD potential), but this result may also be an outcome of an overlap in the function space of the pair and LD potentials whereby either can compensate for the another in a manner to which the CG optimization procedure is insensitive. Indeed, similar compensation effects were reported by Sanyal and Shell34 and more recently by Scherer and Andrienko for the coupling between two- and three-body interactions.70 The WW LD potential is larger in magnitude compared to the MM one and saturates at a minimum of four neighboring water molecules, which no doubt corresponds to the preference for tetrahedral coordination in liquid water. The LD potential reaches a plateau from four neighbors onward, which is by design of the relative entropy optimization. At some point, the system no longer explores local densities beyond a maximal value (see for example Figure 4c) where the largest is \( \approx 15 \). So, beyond these values, there is no information in the reference simulation that can be used to tune the potential. The relative entropy algorithm thus extrapolates a constant value thereafter. Noticeably, the WW IMC potentials needed 10 iterations more to converge than the MM ones (15 vs 5 iterations).

4.3. Structural and Thermodynamic Representability. To evaluate the representability of the derived CG models, we compute the RDF between CG water in implicit methanol and between CG methanol in implicit water, both at a mole fraction of \( x_M = 0.5 \). In Figure 4a, the RDFs between methanol molecules for the different CG models in comparison with the center of mass RDFs between water molecules (WW). Representability analysis of the LD distributions for the REO implicit solvent models at \( x_M = 0.5 \): (c) Comparison of the LD distributions of methanol molecules (MM). (d) Comparison of the LD distributions of water molecules (WW). The FG model is illustrated through the solid black line, the IMC model through the dashed red line, the REO model through the dotted green line, and the REO LD model through the dotted blue line.

Figure 3. (a) Effective pair potentials for the methanol–methanol (MM) interaction at 50% methanol and (b) for the water–water (WW) interaction obtained from IMC (solid red lines), REO (dashed green line), and REO LD optimization (dotted blue line). (c) Local density potentials for the MM interaction derived at 50% methanol and (d) for the WW interaction.

Figure 4. Representability analysis of the RDFs for the implicit solvent models at \( x_M = 0.5 \): (a) Comparison of center of mass RDFs between methanol molecules (MM). (b) Comparison of center of mass RDFs between water molecules (WW). Representability analysis of the LD distributions for the REO implicit solvent models at \( x_M = 0.5 \): (c) Comparison of the LD distributions of methanol molecules (MM). (d) Comparison of the LD distributions of water molecules (WW). The FG model is illustrated through the solid black line, the IMC model through the dashed red line, the REO model through the dotted green line, and the REO LD model through the dotted blue line.
LD distributions obtained from the REO and REO LD model are shown in comparison with the FG reference. As shown, the REO model (dashed green line) already captures the FG LD distribution (solid black line) quite accurately, and the additional LD potential improves the match only slightly (dashed blue line). This corresponds to the similarity in the underlying pair potentials and the small contribution of the LD potentials and further indicates only a weak multibody nature of the MM interaction.

The water−water (WW) RDFs presented in Figure 4b show a very sharp first peak that is present in all models. This points toward a high probability to find a water molecule next to another one, but due to the narrow width, only approximately four water molecules are found in the first solvation shell, in comparison to icosahedral coordination in simple liquids. This is in agreement with prior experimental and theoretical studies.44,45,71 Despite the similarity in the qualitative appearance of the RDF, the REO LD model (dashed blue line) slightly underestimates this first peak, as shown in Figure 4b. The REO model (dashed green line) also slightly underestimates the first peak, and the IMC model (solid red line), by construction, shows the best agreement. This is surprising since IMC and REO show only minor differences in the underlying pair potentials. One possible reason could be the different interpolation and extrapolation schemes applied by the two coarse-graining software packages during the generation and/or evaluation of the final tabulated potentials. This assumption is supported by Figure 5a, where we compare the raw output of VOTCA, i.e., no interpolation between data points, and the two tabulated potentials used in the simulation. Here, potentials are interpolated. Moreover, the relative entropy approach utilizes piecewise cubic splines to represent interactions that are distinct from the potential interpolation in the IMC approach that uses the Akima interpolation scheme. Another reason for the differences between IMC and REO models may be the distinct convergence strategies, involving iteration in the former and minimization in the latter.

One sees that the repulsive part as well as the potential well in the tabulated IMC potential (solid black line) is slightly softer than for the REO-tabulated potential (red line) and the raw output of VOTCA (green line). On the basis that structure is mainly determined by the short-range part of a potential and given the narrow width of the first peak in the RDF, this small difference between raw output data and final tabulated potentials could cause the structural difference observed. For methanol−methanol, those small differences in the repulsive region are not observed as depicted in Figure 5b.

The methanol−methanol LD distributions seem insensitive to the parametrization method, consistent with the weak LD interactions that emerge in the REO LD model. The water−water LD distribution, on the other hand, reveals that the agreement between the FG model (solid black line) and the REO model (dashed green line) is improved by the LD potential (dashed blue line), as shown in Figure 4d. Nevertheless, the improved agreement in the LD distributions comes with a loss of accuracy in the RDF, so it seems difficult to quantitatively match both at the same time with the current set of potentials. A reason for this may be the expanded parameter space for relative entropy minimization when LD potentials are included, in which spline discretization (i.e., knot density) begins to become important. This issue may also relate to the particular balancing of attractive interactions, which manifest in the LD strategy as a compensation between repulsive pair interactions and attractive LD ones, as discussed above. Indeed, in principle, relative entropy minimization with pair and LD potentials should exactly reproduce both the pair and LD correlation functions, as discussed by Chaimovich and Shell,72 such that observed differences must be due either to algorithm convergence, statistical fluctuations, or the manner by which the potential is approximated (e.g., splines of a chosen knot density).

To quantify the mismatch in the RDF, the root mean squared (RMS) error in the g(r) calculation between the FG and the CG system is computed as

$$\text{RMS error} = \sqrt{\frac{1}{N} \sum_{r=1}^{N} (g_{\text{FG}}(r) - g_{\text{CG}}(r))^2}$$

by performing the summation over a total of N = 5 simulations. The results are presented in Figure 6. The error for the MM interaction, which is depicted in Figure 6a, is less than 0.01 for all the CG models, which further confirms the accurate structural quality of all CG methanol models in implicit water. However, for CG water, the REO LD model (Figure 6b) shows a deviation of about 0.4 at the short distances, which corresponds to the underestimation of the first peak in the g(r)
coincides with the di...consequence, all CG water models are more compressible. The difference between water molecules outside the first solvation shell. As a consequence, all CG water models are more compressible. The difference between the IMC and REO and REO LD model...the determination of the model compared to the REO or IMC model since the number of neighboring molecules is most likely to be <3.

The GKBIs of all CG methanol models, $G_{M0}$, are in good agreement with the underlying FG one. This is in agreement with the structural overlap between all models. All CG water models show larger GKBIs compared to the FG reference. This general increase in the GKBIs for CG water in implicit methanol implies an effective stronger affinity between water molecules outside the first solvation shell. As a consequence, all CG water models are more compressible. The difference between the IMC and REO and REO LD model coincides with the differences observed in the RDFs. The REO and REO LD model show the same value within error bars. In general, inclusion of the LD potentials does not seem to impact the GKBIs values, with respect to relative entropy pair-only models. This is likely due to the emphasis on long-range correlations of the GKBIs integrals and the fact that the LD potential modifies short-range interactions.

The similarity in the GKBIs between the REO and REO LD models might further point toward a dominant role of direct pairwise interactions in the system. To probe this idea, we compute the cluster size distribution that is an emergent property dependent on the collective interactions in the system. If multibody effects are significant, then one expects the cluster size distribution to be impacted by the ability of the CG model to reproduce them. To evaluate the cluster size distributions, we chose the same cutoff values as for the LD potentials to determine whether or not particles are in contact and thus belong to the same cluster. In practice, the cluster size distribution can be very sensitive to the cutoff choice. As shown in Figure 7b, the FG methanol molecules comprise a single system-spanning cluster at $x_M = 0.5$. A corresponding snapshot of the FG system presented in Figure 7a confirms this picture. Note that for visual clarity, we do not show the water molecules, which are explicitly present in the simulation. The network-like structure is similar to the formation of rings and strings, which follows the observations made by Guo et al. and Perera et al. Since all CG models are able to match the cluster size distribution of the underlying FG system, it suggests that multibody effects play a less significant role in forming this percolating structure.

Figure 7b shows the cluster size distribution of water molecules for the FG model and the implicit methanol models. Here, no large water clusters are formed, and the water molecules largely populate isolated clusters as well as doublets and triplets, as indicated from the snapshot presented in the inset (methanol molecules are not depicted in the snapshot for visual clarity). The network-like methanol structure restricts the number of water–water contacts. At this concentration, water molecules are not significantly tetrahedrally coordinated, which supports the picture that the LD potential has a less attractive impact on the performance of the model compared to the REO or IMC model since the number of neighboring molecules is most likely to be <3.

Further, the formation of only small aggregates points toward a less dominant role of multibody effects, at least at the level of water–water interactions, and implies a weakened effect of the LD potentials. That observation most likely explains why the REO LD model is slightly less populated at low coordination numbers and cluster sizes. This observations coincides with the work of Laaksonen et al., who showed a loss in tetrahedral coordination of water molecules from $x_M = 0.5$ onward. The difference between the REO and REO LD model can be explained by an increased attraction between water molecules introduced by the negative tail of the LD potential if the number of neighboring waters is $\geq 3$. This tail has a lower value than the minimum in the REO model and thus introduces stronger attraction.

4.4. Structural and Thermodynamic Transferability. Despite the negligible contribution of LD potentials to the representability of the CG models at $x_M = 0.5$, we now test their effect on the transferability toward different methanol mole fractions. Because the IMC- and REO-optimized pair models are nearly identical, we focus on comparison of the IMC and the REO LD models to study the effect of LD potentials on the transferability of the derived CG models.

Figure 6. RMS error between FG and CG RDFs for the implicit solvent models at $x_M = 0.5$: (a) RMS error in the methanol methanol (MM) RDF and (b) RMS error in the water–water (WW) RDF. The difference between the FG and the IMC model is illustrated through the solid red line, between the FG and the REO model through the solid green line, and between the FG and the REO LD model through the solid blue line.

Table 3. Ganguly-Corrected Running Kirkwood–Buff Integrals Averaged between 1.0 and 1.5 nm for Different CG Models

<table>
<thead>
<tr>
<th>$x_M$</th>
<th>model</th>
<th>$G_{WW}$ (nm$^3$)</th>
<th>$G_{M0}$ (nm$^3$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>FG</td>
<td>$0.025 \pm 2.5 \times 10^{-3}$</td>
<td>$-0.077 \pm 0.4 \times 10^{-3}$</td>
</tr>
<tr>
<td></td>
<td>IMC</td>
<td>$0.094 \pm 0.9 \times 10^{-3}$</td>
<td>$-0.078 \pm 0.2 \times 10^{-3}$</td>
</tr>
<tr>
<td></td>
<td>REO</td>
<td>$0.063 \pm 2.2 \times 10^{-3}$</td>
<td>$-0.076 \pm 0.2 \times 10^{-3}$</td>
</tr>
<tr>
<td></td>
<td>REO LD</td>
<td>$0.065 \pm 1.1 \times 10^{-3}$</td>
<td>$-0.074 \pm 0.1 \times 10^{-3}$</td>
</tr>
</tbody>
</table>
In the top panel of Figure 8, the RDFs for the two CG models are shown at a mole fraction of \(x_M = 0.1\) in comparison with FG RDF at \(x_M = 0.1\) and at \(x_M = 0.5\). In Figure 8a, one can see that at \(x_M = 0.1\) the CG models consistently show a decrease in the first peak of the MM RDF relative to the FG one at the same composition and lower than the FG structure at \(x_M = 0.5\). Further, the CG models underestimate the second maximum and overestimate the second minimum relative to the FG simulation at the same composition. Despite this quantitative mismatch, both CG models reproduce the main features (maxima, minima) of the FG structure well and show some aspects of transferability, when compared to the reference structure at \(x_M = 0.5\).

For water, the REO LD model (solid light blue line) closely matches the structure of the FG system at high water content (solid black line), as depicted in Figure 8b. In contrast, the IMC model (solid red line) is closer to the structure at \(x_M = 0.5\) (solid orange line) that was the original parametrization condition. This may point to the role of multibody water−water interactions as relevant to the solution behavior at high water concentrations.

The bottom panel of Figure 8 shows the RDFs at \(x_M = 0.9\). In Figure 8c, the CG model for methanol in implicit water shows no significant difference between the IMC model (solid red line) and the REO LD model (dotted light blue line). Similar to the lower mole fraction case, both of these models cannot reproduce exactly the FG structure (solid black line) but are able to capture the location of the first maximum in the RDF compared to the structure at \(x_M = 0.5\) (solid orange line). In Figure 8d, one sees that the IMC model (solid red line) leads to a decreased first peak in the RDF for CG water in implicit methanol compared to the FG model (solid black line). Further, the RDF obtained from the IMC model overlaps with the RDF of the FG system (dashed orange line) at the reference point \((x_M = 0.5)\). These results suggest that IMC better seems to match the FG structure at \(x_M = 0.5\) (solid orange line) that was the original parametrization condition. This may point to the role of multibody water−water interactions as relevant to the solution behavior at high water concentrations.
qualitatively similar to composed of all methanol molecules in the system, which is concentrations (Figure 9b) there exists only a single cluster methanol molecules are grouped together, whereas at high one sees that at low methanol concentrations only a few nonreference cases, as presented in Figure 9. In Figure 9a, illustrated by the methanol cluster size distribution for upon increasing methanol mole fraction.

In general, it seems that to model the structural transferability, LD-dependent potentials play no significant role for methanol in implicit water. Contrarily, for water in implicit methanol, the presence of a LD potential improves the structural transferability with decreasing methanol concentration. This is in agreement with the observations of Sanyal and Shell, who pointed out that water–water LD potentials lead to transferable CG models of aqueous mixtures only when water–water interactions have a major contribution to the multibody PMF. Further, this corresponds to the work of Laaksonen et al., who found out that the dominant species in the system is mainly responsible for the structural features of the mixture. Following this argument, the decreasing effect of LD contributions with increasing methanol concentration reveals the difficulty of using LD potentials to effectively capture the unusual mixing behavior of methanol and water upon increasing methanol mole fraction.

The negligible effect of multibody contributions is again illustrated by the methanol cluster size distribution for nonreference cases, as presented in Figure 9. In Figure 9a, one sees that at low methanol concentrations only a few methanol molecules are grouped together, whereas at high concentrations (Figure 9b) there exists only a single cluster composed of all methanol molecules in the system, which is qualitatively similar to $x_M = 0.5$. These observations are further supported by the snapshots of the FG system depicted in the inset of the figures, where water molecules are not shown for visual clarity. There again a network-like structure is visible for methanol mole fractions greater than 0.5. Since all CG models perform in the same way, it further indicates that methanol–methanol interactions are not dominated by multibody effects in aqueous solutions.

We now turn to the cluster size distributions in the implicit-methanol systems. In Figure 10a, the distribution for waters at $x_M = 0.1$ shows two distinct cluster sizes for the FG system. On the one hand, a cluster emerges with very few water molecules, while on the other hand one appears with the number of water molecules in the system (i.e., a percolating or system spanning cluster). A snapshot of the FG system (shown in the inset) illustrates this, where the methanol molecules are not depicted for visual clarity. One sees a smaller cluster in the lower left corner and a large, system spanning cluster. It appears that small methanol subdomains, present at low methanol concentration ($x_M = 0.1$, see Figure 9a), prevent that all of the water aggregates into a single cluster. This produces the separation of peaks in the WW cluster distribution (Figure 10a) and is in agreement with the study of Perera et al. The IMC model produces clusters with only few water molecules, that is more similar to the cluster distributions at $x_M = 0.5$. This is in agreement to what we observe for $g(r)$ as well, where the $g(r)$ of the IMC model is more similar to the one of the FG model at $x_M = 0.5$ than at $x_M = 0.1$. Thus, this model is not transferable toward lower methanol concentrations. Interestingly, the REO LD model cluster distribution attains better agreement with the cluster size distribution, in particular, capturing the two-state population at low and high cluster size. In this case, we adjusted the distance cutoff to 0.37 nm to effectively locate water neighbors in this heterogeneous environment. The cluster size distribution at $x_M = 0.1$ clearly shows that the LD potential improves the CG water model. Moreover, it confirms the picture that in water-rich phases multibody effects play a significant role and LD-dependent potentials are effective strategies for capturing such interactions. At high methanol concentrations ($x_M = 0.9$), no large clusters of water are formed, and the LD-dependent potential does not improve the CG model compared to the IMC one. The more isolated behavior of water molecules is again depicted in a snapshot of the corresponding molecules taken from the FG system. This indicates a negligible role of water
the IMC and REO LD model show stronger affinity between methanol molecules than the FG model does. The LD potential even emphasizes this effect. The same trend can be observed at \( x_M = 0.9 \). Here, the REO LD model shows again the largest GKBIs value, whereas the IMC model is closer to the FG system. This suggests further that LD effects do not play a significant role on the behavior of methanol in aqueous solutions.

Contrary to that, the REO LD model reproduces the FG WW-GKBIs at low methanol concentrations \( (x_M = 0.1) \). At \( x_M = 0.9 \), the WW-GKBIs is however overestimated by the REO LD model, as indicated by the RDF as well. Thus, the GKBIs analysis follows the same trends observed in the RDFs. This brings us to the following conclusion: in water rich phases, where LD effects have a larger contribution to the water—water multibody PMF, LD potentials improve the transferability of the derived CG models. However, with increasing methanol concentrations, these contributions become less significant, and the CG models do not further improve.

However, it is essential to note that the performance of all models in reproducing GKBIs is purely algorithmic in nature because the IMC, REO, and REO LD methods all formally should, in principle, locate CG models that reproduce exact \( g(r) \) forms upon which the GKBIs rely. Therefore, any difference between the properties in Table 4 must be considered a result of numerical approximations (e.g., tabulated or splined potentials) or convergence (CG method) and the fact that the GKBIs integrals are sensitive to subtle pair correlation effects at large distances.

### 5. CONCLUSION

In this work, we addressed the question of whether single site CG models for methanol and water mixtures, in an implicit solvent environment, can capture the mixing behavior of these two components. We could have also considered mapping methanol to a two-bead model to account for nonpolar and polar sites. But, this would require up to six local density potentials for the combination of central and neighbor type. This would make the model (and its parametrization) significantly more complex, which we wanted to avoid. We derived implicit solvent CG models for various water—methanol mixtures by application of two different coarse graining methods, namely, IMC and relative entropy optimization. We showed analytically and numerically that under the assumptions of a discretized Hamiltonian and the application of a Newton–Raphson scheme the methods are formally equal.

We further investigated if embedding of LD potentials improves the derived CG models in terms of concentration transferability. In agreement with previous work by Sanyal and Shell,\textsuperscript{32,34} we found if water—water LD contributions play a dominant role in the multibody PMF, then LD potentials improve the transferability of water in implicit methanol toward lower methanol concentrations. Unlike the work in ref 32, here we find that LD potentials do not improve transferability in water-lean solutions. The LD-included CG model for methanol in implicit water shows no significant improvement at either concentration. When compared to a CG model with pair potentials only, the presence of a LD potential even emphasizes the affinity between methanol molecules as indicated by the GKBIs. One difference in the present study from ref 34, which showed that LD models were consistently able to improve transferability across composition space in

### Table 4. Average Ganguly-Corrected Running Kirkwood–Buff Integrals at Different Concentrations of Methanol Obtained Either from FG Simulations or from CG Simulations with Previously Generated Potentials at \( x_M = 0.5 \)

<table>
<thead>
<tr>
<th>( x_M )</th>
<th>model</th>
<th>( G_{WW} ) (nm(^3))</th>
<th>( G_{MM} ) (nm(^3))</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>FG</td>
<td>(-0.025 \pm 2.5 \times 10^{-3})</td>
<td>(-0.14 \pm 3.70 \times 10^{-2})</td>
</tr>
<tr>
<td>IMC</td>
<td>(0.900 \pm 1.4 \times 10^{-5})</td>
<td>(-0.091 \pm 1.2 \times 10^{-5})</td>
<td></td>
</tr>
<tr>
<td>REO LD</td>
<td>(-0.026 \pm 0.2 \times 10^{-5})</td>
<td>(-0.054 \pm 0.2 \times 10^{-5})</td>
<td></td>
</tr>
<tr>
<td>0.9</td>
<td>FG</td>
<td>(0.01 \pm 1.05 \times 10^{-2})</td>
<td>(-0.075 \pm 0.3 \times 10^{-7})</td>
</tr>
<tr>
<td>IMC</td>
<td>(0.084 \pm 3.5 \times 10^{-5})</td>
<td>(-0.071 \pm 0.2 \times 10^{-7})</td>
<td></td>
</tr>
<tr>
<td>REO LD</td>
<td>(0.380 \pm 0.6 \times 10^{-5})</td>
<td>(-0.066 \pm 0.2 \times 10^{-7})</td>
<td></td>
</tr>
</tbody>
</table>

Figure 10. Transferability analysis of the cluster size distributions for the implicit solvent models: (a) Comparison between the cluster size distribution of water molecules (WW) at \( x_M = 0.1 \). (b) Comparison between the cluster size distribution of water molecules (WW) at \( x_M = 0.9 \). The FG model is illustrated through black circles, the IMC model through the dashed red line, and the REO LD model through the dashed blue line. The insets depict a snapshot of the corresponding FG trajectories without the methanol molecules for visual clarity.

and multibody effects in methanol-rich phases, which corresponds to the picture of Laaksonen and Perera.\textsuperscript{44,45}

Another prediction of LD potentials is further indicated by the GKBIs presented in Table 4. At low concentrations, both...
benzene–water mixtures, is that here a species is made implicit. It may be that LD potentials are more effective when all species are explicitly present but coarse grained. The interpretation that multibody effects have a less significant effect on the structural arrangement of methanol molecules is further supported by the analysis of the cluster size distributions in the systems. Here, all CG models show similar size distributions for all concentrations independent of the presence of a LD-dependent potential. The CG methanol models show more network-like structures due to large volume occupancy in the systems, whereas the CG water stays mostly isolated, perhaps due to restraints posed by the methanol network. In agreement with our work, Scherer and Andrienko recently found a negligible effect of multibody contributions on methanol–methanol interactions by investigating the impact of three-body potentials on the pair structure of liquid methanol. What seems to be more important than multibody contributions is the ability to form hydrogen bond networks in order to accurately describe water–methanol mixtures as discussed in the literature. Explicit electrostatics or any directionality introduced to the model, as for example done on the basis of point multipole and Gay–Berne potentials, seems to be necessary to accomplish the effect of hydrogen bonding. Without performing extensive analysis on the FG system, we used bottom-up coarse-graining techniques to identify the negligible contribution of LD-dependent potentials on structural changes in alcohol–water mixtures. This does not mean that LD potentials cannot improve the transferability of CG models as nicely demonstrated in the literature. It rather shows that one should carefully consider the underlying physics of the system, specifically the extent of multibody effects, for a successful application of LD potentials. Further, this study highlights the aid of bottom-up coarse graining to identify important degrees of freedom, e.g., here the ability to explicitly form hydrogen bonds, to accurately describe a system in the CG configuration space.

**AUTHOR INFORMATION**

**Corresponding Authors**

*E-mail: rosenberger@cpc.tu-darmstadt.de* (D.R.).

*E-mail: shell@engineering.ucsb.edu* (M.S.S.).

**ORCID**

David Rosenberger: 0000-0001-6620-6499

M. Scott Shell: 0000-0002-0439-1534

Nico F. A. van der Vegt: 0000-0003-2880-6383

**Notes**

The authors declare no competing financial interest.

**ACKNOWLEDGMENTS**

N.F.A.v.d.V. and D.R. acknowledge funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Project number 233630050 - TRR 146 on “Multiscale Simulation Methods for Soft Matter Systems”. T.S. and M.S.S. acknowledge funding from the National Science Foundation through Grant Nos. CHEM-1300770 and CHEM-1800344.

**REFERENCES**


(26) Murtola, T.; Falck, E.; Karttunen, M.; Vattulainen, I. Coarse-grained model for phospholipid/cholesterol bilayer employing inverse...