JCTC Journal of Chemical Theory and Computation

Versatile Object-Oriented Toolkit for Coarse-Graining Applications

Victor Rühle, Christoph Junghans, Alexander Lukyanov, Kurt Kremer, and Denis Andrienko*

Max Planck Institute for Polymer Research, Ackermannweg 10, 55128 Mainz, Germany

Received July 17, 2009

Abstract: Coarse-graining is a systematic way of reducing the number of degrees of freedom representing a system of interest. Several coarse-graining techniques have so far been developed, such as iterative Boltzmann inversion, force-matching, and inverse Monte Carlo. However, there is no unified framework that implements these methods and that allows their direct comparison. We present a versatile object-oriented toolkit for coarse-graining applications (VOTCA) that implements these techniques and that provides a flexible modular platform for the further development of coarse-graining techniques. All methods are illustrated and compared by coarse-graining the SPC/E water model, liquid methanol, liquid propane, and a single molecule of hexane.

1. Introduction

Computational materials science deals with phenomena covering a wide range of length- and time-scales from Ångstrøms (typical bond lengths) and femtoseconds (bond vibrations) to micrometers (crack propagation) and milliseconds (a single polymer chain relaxation). Depending on the characteristic time- and length-scales involved, the system description can vary from first principles and atomistic force fields to coarse-grained models and continuum mechanics. The role of bottom-up coarse-graining, in a broad sense, is to provide a systematic link between these levels of description.

Here we focus on coarse-graining techniques that link two particle-based descriptions with a different number of degrees of freedom. The system with the larger number of degrees of freedom we denote as the reference system. The system with the reduced number of the degrees of freedom is referred to as the coarse-grained system. An example is an all-atom (reference) and a united-atom (coarse-grained) molecular representation, where the number of the degrees of freedom is reduced by embedding hydrogens into heavier atoms.[55] Another example, which is treated in detail here, is an all-atom (three sites) and a single site model of water. Other examples can be readily found in the literature.[1−12]

We also assume that the following prerequisites are satisfied:

(i) Both the reference and the coarse-grained descriptions are represented by a set of point sites, $\boldsymbol{r}=\{\boldsymbol{r}_i\}$, $i = 1, 2, ..., n$, in case of the reference system, and $\boldsymbol{R}=\{\boldsymbol{R}_j\}$, $j = 1, 2, ..., N$, in case of the coarse-grained system.[56]

(ii) A mapping scheme, i.e., a relation between $\boldsymbol{r}$ and $\boldsymbol{R}$, can be expressed as $\boldsymbol{R} = \hat{\mathbf{M}}\boldsymbol{r}$, where $\hat{\mathbf{M}}$ is a $n \times N$ matrix.[57]

(iii) For the reference system, we have the coordinates and the forces of a trajectory that samples a canonical ensemble (or that part of it we are interested in reproducing on a coarse-grained level).

Then the prime task of systematic coarse-graining is to devise a potential energy function of the coarse-grained system, $U(\boldsymbol{R})$.

To do this, one can use several coarse-graining approaches. From the point of view of implementation, these approaches can be divided into iterative and noniterative methods. Boltzmann inversion is a typical example of a noniterative method.[1] In this method, which is exact for independent degrees of freedom, coarse-grained interaction potentials are calculated by inverting the distribution functions of the coarse-grained system. Another example of a noniterative method is force matching, where the coarse-grained potential

* Corresponding author. E-mail:denis.andrienko@mpip-mainz.mpg.de.

10.1021/ct900369w CCC: $40.75
Published on Web 11/09/2009

is chosen in such a way that it reproduces the forces on the coarse-grained beads.[5,13] Configurational sampling,[14] which matches the potential of mean force, also belongs to this category. Boltzmann inversion and force matching only require a trajectory for a reference system.[58] Once that is known, coarse-grained potentials can be calculated for any mapping matrix $\hat{\mathbf{M}}$.

Iterative methods refine the coarse-grained potential $U(\boldsymbol{R})$ by reiterating coarse-grained simulations and by calculating corrections to the potential on the basis of the reference and the coarse-grained observables (e.g., radial distribution function or pressure). The simplest example is the iterative Boltzmann inversion method,[15] which is an iterative analogue of the Boltzmann inversion method. More sophisticated (in terms of the update function) is the inverse Monte Carlo approach.[16]

One can also classify systematic coarse-graining approaches by the micro- and macroscopic observables they use to derive the coarse-grained potential, such as structure-,[1,16,17] force-,[5,13,18] and potential-based approaches,[19] where the name identifies the observable used for coarse-graining. Note that hybrids of these methods are also possible.[3,12]

With a rich zoo of methods plus their combinations available at hand, it is natural to ask about an optimal method for a specific class of systems. On a more fundamental level, one might question whether the different methods provide the same coarse-grained potential and whether it is possible to formulate a set of (even empirical) rules favoring one method with respect to another. It is obvious this is a difficult task to be treated analytically, especially for realistic systems. To assess the quality of a particular coarse-graining technique, one needs to apply all available methods to a certain number of systems and to compare and quantify the degree of discrepancy between the coarse-grained and the reference descriptions. This is, however, cumbersome due to the absence of a single package where all these methods are implemented with the same accuracy and same level of technical detail.

The main aim of this work is to introduce such a coarse-graining package. The paper is organized as follows: We first describe the basic ideas behind each method, paying special attention to the technical issues one has to overcome when implementing them. We then illustrate these methods by coarse-graining systems of different complexities: a three-site SPC/E water, methanol, propane, and hexane.

2. Methods

Before starting with brief recapitulations of the coarse-graining methods, we refer the reader to a (far from complete) list of reviews which cover various aspects of generating coarse-grained potentials.[20−26]

2.1. Boltzmann Inversion. Boltzmann inversion is the simplest method one can use to obtain coarse-grained potentials.[1] It is mostly used for bonded potentials, such as bonds, angles, and torsions. Boltzmann inversion is structure-based and only requires positions of atoms.

The idea of Boltzmann inversion stems from the fact that in a canonical ensemble independent degrees of freedom q obey the Boltzmann distribution, i. e.:

$$P(q) = Z^{-1}\exp[-\beta U(q)] \tag{1}$$

where $Z = \int \exp[-\beta U(q)]\, \mathrm{d}q$ is a partition function, $\beta = 1/k_{\mathrm{B}}T$. Once $P(q)$ is known, one can invert eq 1 and obtain the coarse-grained potential, which, in this case, is a potential of mean force:

$$U(q) = -k_{\mathrm{B}}T \ln P(q) \tag{2}$$

Note that the normalization factor Z is not important since it would only enter the coarse-grained potential $U(q)$ as an irrelevant additive constant.

In practice, $P(q)$ is computed from the trajectory of the reference system, which is sampled either by Monte Carlo, molecular dynamics, stochastic dynamics, or any other integrator that ensures a canonical distribution of states.

Boltzmann inversion is simple to implement, however, one has to be careful with the rescaling of the probability P due to orientational entropy as well as computational issues. The probability rescaling can be explained on a particular example of coarse-graining of a single polymer chain by beads with bond, angle and torsion potentials. In this case the coarse-grained potential U depends on three variables, bond length r, angle θ, and torsion angle φ.

Assuming, as before, a canonical distribution and independence of the coarse-grained degrees of freedom, we can write:

$$P(r, \theta, \varphi) = \exp[-\beta U(r, \theta, \varphi)] \tag{3}$$

$$P(r, \theta, \varphi) = P_r(r)P_\theta(\theta)P_\varphi(\varphi) \tag{4}$$

If we now compute the histograms for the bonds $H_r(r)$, angle $H_\theta(\theta)$, and torsion angle $H_\varphi(\varphi)$, then we must rescale them in order to obtain the volume normalized distribution functions.[59]

$$P_r(r) = \frac{H_r(r)}{4\pi r^2}, \quad P_\theta(\theta) = \frac{H_\theta(\theta)}{\sin\theta}, \quad P_\varphi(\varphi) = H_\varphi(\varphi) \tag{5}$$

The coarse-grained potential can then be calculated by Boltzmann inversion of the distribution functions:

$$\begin{aligned} U(r, \theta, \varphi) &= U_r(r) + U_\theta(\theta) + U_\varphi(\varphi) \\ U_q(q) &= -k_{\mathrm{B}}T \ln P_q(q), \qquad q = r, \theta, \varphi \end{aligned} \tag{6}$$

On the technical side, the implementation of the Boltzmann inversion method requires smoothing of $U(q)$ to provide a continuous force. Splines can be used for this purpose. Poorly and unsampled regions, that is regions with high $U(q)$, shall be extrapolated. Since the contribution of these regions to the canonical density of states is small, the exact shape of the extrapolation is less important.

Another crucial issue is the cross-correlation of the coarse-grained degrees of freedom. Independence of the coarse-grained degrees of freedom is the main assumption that allows factorization of the probability distribution, eq 4, and the potential, eq 6, hence, one has to carefully check whether this assumption holds in practice. This can be done by performing coarse-grained simulations and by comparing cross-correlations for all pairs of degrees of freedom in

atomistic and coarse-grained resolution, e.g., using a two-dimensional histogram, analogous to a Ramachandran plot.[60]

2.2. Iterative Boltzmann Inversion. Iterative Boltzmann inversion (IBI) is a natural extension of the Boltzmann inversion method. Since the goal of the coarse-grained model is to reproduce the distribution functions of the reference system as accurately as possible, one can also iteratively refine the coarse-grained potentials using some numerical scheme. Depending on the update function, this can be done by using either the iterative Boltzmann inversion[15] or the inverse Monte Carlo[16,17] method. We will first discuss the iterative Boltzmann inversion method.

In the iterative Boltzmann inversion, the coarse-grained potential is refined according to the following scheme:[61]

$$\begin{aligned} U^{(n+1)} &= U^{(n)} + \Delta U^{(n)} \\ \Delta U^{(n)} &= k_B T \ln \frac{P^{(n)}}{P_{\text{ref}}} = U^{\text{ref}}_{\text{PMF}} - U^{(n)}_{\text{PMF}} \end{aligned} \tag{7}$$

One can easily see that convergence is reached as soon as the distribution function $P^{(n)}$ matches the reference distribution function P_{ref}, or, in other words, the potential of mean force, $U^{(n)}_{\text{PMF}}$ converges to the reference potential of mean force.

IBI can be used to refine both bonded and nonbonded potentials. It is primarily used for simple fluids with the aim of reproducing the radial distribution function of the reference system in order to obtain nonbonded interactions.[15] It can have convergence problems for multicomponent systems, since it does not account for cross-correlation correction terms, that is the updates for P_{AA}, P_{AB}, and P_{BB} are not coupled (the subscript enumerates a single component in a multicomponent system). For such systems, the inverse Monte Carlo method works better. The scheme can be stabilized by multiplying the update function, $\Delta U^{(n)}$, by a factor $\eta \in [0..1]$.

On the implementation side, IBI has the same issues as the inverse Boltzmann method, i.e., smoothing and extrapolation of the potential must be implemented.

We shall also mention that, according to the Henderson theorem,[25,27] which is a classical analogue of the Hohenberg−Kohn theorem, the pairwise coarse-grained potential $U(r)$ is unique up to an additive constant and exists,[28,29] which, in principle, states that all structure-based iterative methods must converge to the same coarse-grained potential, provided that their aim is to exactly reproduce pair correlation functions of the reference system. As we will see later, this is often not the case in practice, since small changes in the radial distribution function often lead to big changes in the pair potential, i.e., it is difficult to control systematic errors during the calculation of the potential update.

Another issue of coarse-graining is that coarse-grained models cannot reproduce all the statistical or thermodynamic properties of the reference system. Pressure, compressibility, or viscosity[30] are often very different from those of the reference system. In some cases, however, one can correct for some of these. For example, the viscosity can be adjusted by tuning the parameters of the thermostat,[31] and the pressure can be corrected iteratively by adding a linear term to the nonbonded potential:

$$\Delta U^{\text{pressure}} = A\left(1 - \frac{r}{r_{\text{cut}}}\right) \tag{8}$$

where A is either a constant, e.g., $-0.1 k_B T$,[15] or can be estimated from the virial expansion.[32] Compressibility and pressure, however, cannot be corrected simultaneously.

2.3. Inverse Monte Carlo. Inverse Monte Carlo (IMC) is another iterative procedure that refines the coarse-grained potentials until the coarse-grained model reproduces a set of reference distribution functions. It is very similar to IBI except that the update of the potential, ΔU, is calculated using rigorous thermodynamic arguments.

The name "inverse Monte Carlo" is somehow confusing and is due to the fact that the original algorithm was combined with Monte Carlo sampling of the phase space.[16] However, practically any sampling method can be used (e.g., molecular or stochastic dynamics) as long as it provides a canonical sampling of the phase space.

A detailed derivation of the IMC method can be found in ref 16. Here we briefly recapitulate the more compact version for nonbonded interactions, which is outlined in ref 25 emphasizing technical problems encountered during implementation and application of the method.

The idea of IMC is to express the potential update ΔU in a thermodynamically consistent way in terms of measurable statistical properties, e.g., radial distribution function $g(r)$. Considering a single-component system as an example, we can write the Hamiltonian of the system as

$$H = \sum_{i,j} U(r_{ij}) \tag{9}$$

where $U(r_{ij})$ is the pair potential, and we assume that all interactions depend only on the distance, r_{ij}, between particles i and j. We further assume that this potential is short-ranged, i.e., $U(r_{ij}) = 0$, if $r_{ij} \geq r_{\text{cut}}$.

The next step is to tabulate the potential $U(r)$ on a grid of M points, $r_\alpha = \alpha \Delta r$, where $\alpha = 0, 1, ..., M$, and $\Delta r = r_{\text{cut}}/M$ is the grid spacing. Then the Hamiltonian, eq 9, can be rewritten as

$$H = \sum_{\alpha} U_\alpha S_\alpha \tag{10}$$

where S_α is the number of particle pairs with interparticle distances $r_{ij} = r_\alpha$, which correspond to the tabulated value of the potential U_α.

On one hand, the average value of S_α is related to the radial distribution function $g(r)$:

$$\langle S_\alpha \rangle = \frac{N(N-1)}{2} \frac{4\pi r_\alpha^2 \Delta r}{V} g(r_\alpha) \tag{11}$$

where N is the number of atoms in the system, $((1/2)N(N-1)$ is then the number of all pairs), Δr is the grid spacing, r_{cut}/M, and V is the total volume of the system.

On the other hand, $\langle S_\alpha \rangle$ is a function of the potential U_α and, hence, can be expanded in a Taylor series with respect to small perturbations of U_α, ΔU_α

$$\Delta\langle S_\alpha\rangle = \sum_\gamma \frac{\partial\langle S_\alpha\rangle}{\partial U_\gamma}\Delta U_\gamma + \mathcal{O}(\Delta U^2) \tag{12}$$

The derivatives $\partial\langle S_\alpha\rangle/\partial U_\gamma$ can be obtained by using the chain rule:

$$\begin{aligned} A_{\alpha\gamma} &= \frac{\partial\langle S_\alpha\rangle}{\partial U_\gamma} \\ &= \frac{\partial}{\partial U_\gamma}\frac{\int dq S_\alpha(q)\exp[-\beta\sum_\lambda U_\lambda S_\lambda(q)]}{\int dq \exp[-\beta\sum_\lambda U_\lambda S_\lambda(q)]} \\ &= \beta(\langle S_\alpha\rangle\langle S_\gamma\rangle - \langle S_\alpha S_\gamma\rangle) \end{aligned} \tag{13}$$

Equations 11−13 allow us to calculate the correction for the potential by solving a set of linear equations:

$$\langle S_\alpha\rangle - S_\alpha^{\mathrm{ref}} = A_{\alpha\gamma}\Delta U_\gamma \tag{14}$$

where S_α^{ref} is given by the target radial distribution function. The procedure is then repeated until convergence is reached.

A clear advantage of the IMC compared to the IBI method is that the update of the potential is rigorously derived using statistical mechanics, and hence, the iterative procedure shall converge faster with the IMC update than with the empirical IBI update. Another advantage is that, in the case of multicomponent mixtures, IMC takes into account correlations of observables, that is updates for U_{AA}, U_{AB}, and U_{BB} are interdependent (A and B denote different particle types). In the IBI method, these updates are independent which often leads to convergence problems for multicomponent systems.

The advantages come, of course, at a computational cost. As it is clear from eq 13, one has to calculate cross-correlations of S_α. This requires much longer runs to get statistics that are good enough to calculate the potential update to a similar accuracy as IBI. The accuracies of the update functions of IMC and IBI methods are compared in Section 4.1 for the case of a coarse-grained model of water.

Another issue of the IMC method is the stability of the scheme. Several factors can influence it: the first, and rather technical, point is that $g^{\mathrm{ref}}(r_\alpha)$ has to be calculated using exactly the same convention for the grid as S_α (e.g., the function value should be assigned to the middle of the interval), otherwise the scheme becomes unstable. Second, inversion of $A_{\alpha\gamma}$ requires that it shall be well-defined. This means that one has to remove the regions which are not sampled, such as those at the beginning of the radial distribution function. The convergence can be significantly improved if a smoothing of the potential update ΔU is used. Note that it is better to do smoothing of the update function, not the potential itself, since the latter has more features which can be lost due to too aggressive smoothing. The convergence can also be improved by introducing a multiplicative prefactor for the update function or by using a regularization procedure by adding thermodynamic constraints.[33]

Finally, we have also noticed that the systematic error in $\langle S_\alpha S_\gamma\rangle$ is always higher in the vicinity of the cutoff, which leads to a shift in the tail of the interaction potential and, as a result, to a large offset of pressure. The cross-correlation term $\langle S_\alpha S_\gamma\rangle$ is also very sensitive to the box size, and special care must be taken in order to converge the results with respect to system size. Finite size effects are discussed in detail in Section 4.2, where we coarse-grain liquid methanol.

2.4. Force Matching. Force matching (FM) is another approach to evaluate corse-grained potentials.[5,13,34] In contrast to the structure-based approaches, its aim is not to reproduce various distribution functions, but instead try to match forces on coarse-grained beads as closely as possible.[62] FM is a noniterative method and, hence, is less computationally demanding.

The method works as follows: we first assume that the coarse-grained force field (and hence the forces) depends on M parameters $g_1, ..., g_M$. These parameters can be prefactors of analytical functions, tabulated values of the interaction potentials, or coefficients of splines used to describe these potentials.

In order to determine these parameters, the reference forces on coarse-grained beads are calculated by properly reweighting the forces on the atoms:

$$\boldsymbol{f}_i^{\mathrm{ref}} = M_i\sum_\alpha \frac{w_\alpha \boldsymbol{f}_\alpha}{m_\alpha} \tag{15}$$

where $M_i = (\sum_\alpha w_\alpha^2/m_\alpha)^{-1}$ is the mass of the bead i, index α numbers all atoms belonging to this bead, $\boldsymbol{f}_\alpha$ is the force on the atom α, m_α is its mass, w_α are mapping coefficients used to obtain the position of the coarse-grained bead, $\boldsymbol{R}_i = \sum_\alpha w_\alpha \boldsymbol{r}_\alpha$. If the center of mass is used in the mapping, then eq 15 simplifies to the sum of the forces.

By calculating the reference forces for L snapshots, we can write down $N \times L$ equations:

$$\boldsymbol{f}_{il}^{\mathrm{cg}}(g_1, ..., g_M) = \boldsymbol{f}_{il}^{\mathrm{ref}}, \quad i = 1, ..., N, \quad l = 1, ..., L \tag{16}$$

Here $\boldsymbol{f}_{il}^{\mathrm{ref}}$ is the force on the bead i, $\boldsymbol{f}_{il}^{\mathrm{cg}}$ is the coarse-grained representation of this force. Index l enumerates snapshots picked for coarse-graining. By running the simulations long enough one can always ensure that $M < N \times L$. In this case, the set of eqs 16 is overdetermined and can be solved in a least-squares sense.

Though the underlying idea of FM is very simple, implementation-wise it is the most complicated method. Here we briefly outline the problems, which are then discussed in more detail in Appendix A.

Going back to the set of eqs 16, one can see that $\boldsymbol{f}_{il}^{\mathrm{cg}}$ is, in principle, a nonlinear function of its parameters $\{g_i\}$. It is, therefore, useful to represent the coarse-grained force field in such a way that eqs 16 become linear functions of $\{g_i\}$. This can be done using splines to describe the functional form of the forces.[5]

An adequate sampling of the system requires a large number of snapshots L. Hence, the applicability of the method is often constrained by the amount of available memory. To remedy the situation, one can split the trajectory into blocks, find the coarse-grained potential for each block and then perform averaging over the blocks. More details

on the technical implementation of force matching using cubic splines is given in Appendix A.

3. Implementation

3.1. Coarse-Graining Engine. In a nutshell, coarse-graining is nothing more than an analysis of the canonical ensemble of a reference (high resolution) system. In addition to this analysis, iterative methods require canonical sampling of the coarse-grained system, which can be done using either molecular dynamics (MD), stochastic dynamics (SD), or Monte Carlo (MC) techniques. The latter are implemented in many standard simulation packages. Rather than implementing its own MD/SD/MC modules, the toolkit allows swift and flexible integration of existing programs in such a way that sampling is performed in the program of choice. Only the analysis needed for systematic coarse-graining is done using the package tools.

The tools include calculations of probability distributions of bonded and nonbonded interactions, correlation and autocorrelation functions, and updates for the coarse-grained pair potential. Analysis tools of the MD package can also be integrated into the coarse-graining workflow, if needed.

The package offers a flexible framework for reading, manipulating, and analyzing of MD/SD/MC topologies and trajectories. Its core is modular, and new file formats can be integrated without changing the existing code. At the moment, an interface for GROMACS[35] topologies and trajectories is provided. An interface to ESPResSo++[63] is planned.

The coarse-graining procedure itself is controlled by several extensible markup language (XML) input files, which contain mapping and other options required for the workflow control. In the mapping, it is possible to select groups of interactions which will be used for coarse-graining or analysis.

3.2. Iterative Workflow Control. The workflowchart is shown in Figure 1. The workflow is implemented as a shell script which can, in principle, be run on all available operating systems and provides the flexibility needed to call external (or overload existing) scripts and programs written in other programming languages. An interface to read values from the steering XML files in C++, Perl, and shell is also provided.

During the global initialization, the initial guess for the coarse-grained potential is calculated from the reference radial distribution function or converted from a given potential guess to the internal format. The actual iterative step starts with an iteration initialization. It searches for possible checkpoints and copies and converts files from the previous step and the base directory. Then the simulation run is prepared by converting potentials to the format required by the external sampling program, and actual sampling is performed. Currently, an interface with GROMACS[35] is implemented, and an extension to other packages is straightforward. After sampling the phase space, potential update ΔU is calculated. Often the update requires postprocessing, such as smoothing, interpolation, extrapolation, or fitting to an analytical form. A simple pressure correction[15] can also be seen as a postprocessing of ΔU due to the fact that it only adds a linear interparticle separation function. Finally,

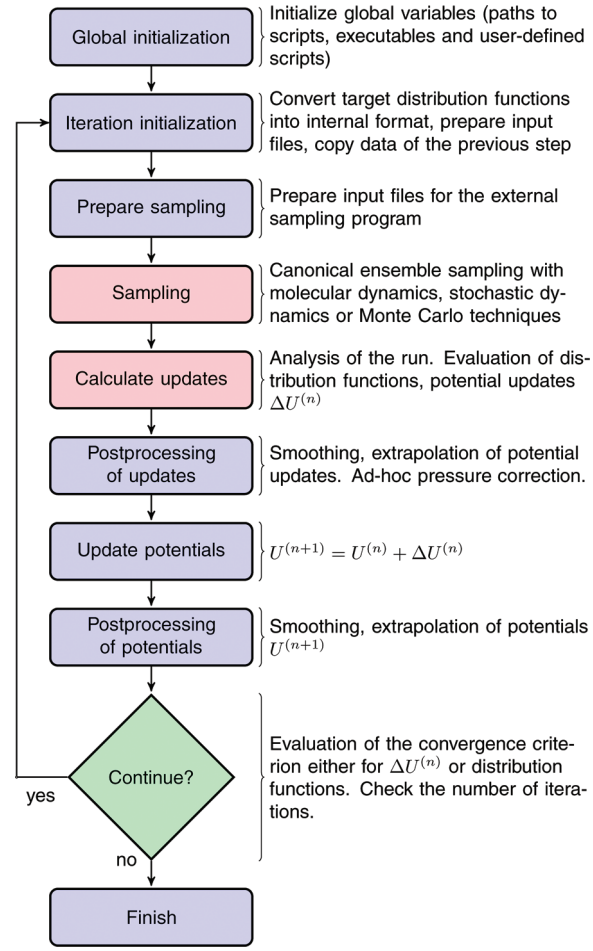

Figure 1. Block-scheme of the workflow control for the iterative methods. The most time-consuming parts are marked in red.

the new potential is determined and postprocessed. If the iterative process continues, then the next iterative step starts to initialize.

4. Examples

We illustrate the package functionality using four systems: SPC/E water, liquid methanol, liquid propane, and a single chain of hexane. The systems are chosen in such a way that the corresponding coarse-grained potentials have already been obtained using one or more techniques, providing a good reference point for comparison.

4.1. Coarse-Graining of Water. Water is one of the most studied liquids from the point of view of both all-atom representations and coarse-grained models.[36,37] Here we coarse-grain one of the all-atom models of water, the SPC/E[38,39] water model. The corresponding parameters of this three-site model are given in the caption to Figure 2. Note that this is a rigid model, i.e., the distances between two hydrogens as well as oxygen and hydrogens are constrained during the molecular dynamics runs. For the coarse-grained representation, we use a one-site representation with a pair potential $U(R_{ij})$, where R_{ij} connects the centers of mass of water molecules i and j.

The all-atom system consisting of 2180 water molecules was first equilibrated in the NPT ensemble at 300K and 1

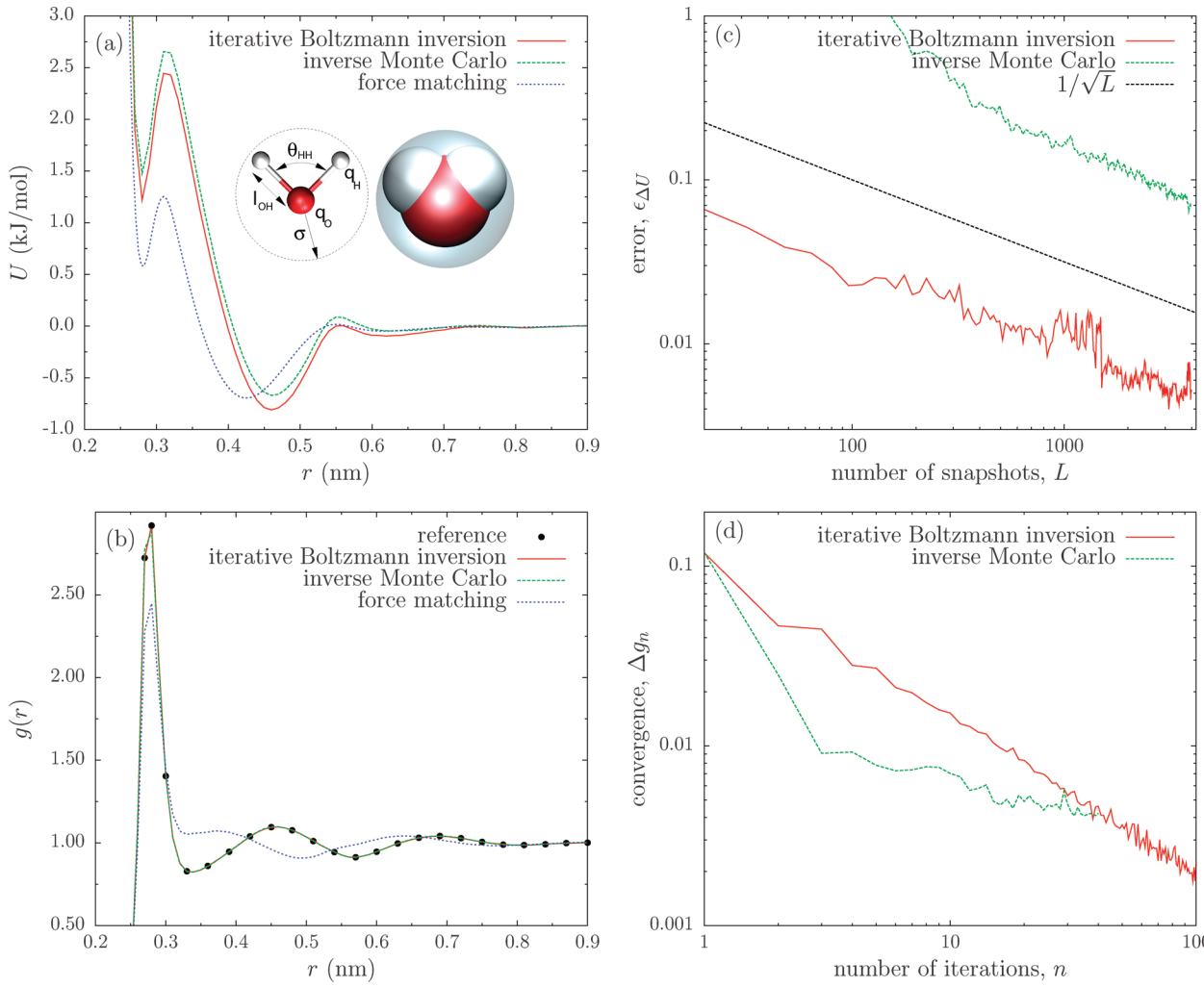

Figure 2. Water: (a) Coarse-grained potentials for SPC/E water obtained using different coarse-graining techniques. (b) Corresponding radial distribution functions. (c) Average error of the potential update function versus number of snapshots used for calculating the update function. (d) Root-mean-square deviation of reference and current radial distribution function versus iteration step. One can see that IMC converges faster than that of IBI. Inset of (a) shows van der Waals excluded volume and coarse-grained representations of a single water molecule as well as parameters used: $\sigma = 3.166$ Å, $\varepsilon = 0.650$ kJ mol^{-1}, $l_{OH} = 1.0000$ Å, $q_H = +0.4238e$, $q_O = -0.8476e$, $\theta_{HH} = 109.47°$.

bar for 100 ns using the Berendsen thermostat and barostat.[40] The last 80 ns were used to determine the equilibrium box size of 4.031 nm, which was then fixed during the 45 ns production run in the NVT ensemble using a stochastic dynamics algorithm.[41] For all further analysis, only the last 40 ns were used. The radial distribution function was calculated using a 0.01 nm grid spacing. The snapshots were output every 0.4 ps.

Force matching potentials were calculated using blocks of six snapshots each. Spline grid spacing of 0.02 nm was used in the interval from 0.24 to 1 nm. For the iterative procedures, the potential of mean force was taken as an initial guess for the interaction potential. The coarse-grained box had the same system size as in the atomistic simulations. Simulations of the coarse-grained liquid were done using a stochastic dynamics algorithm.[41] When using IBI, 300 iterations of 100 ps each were performed. For IMC, we used 10 iterations of 500 ps each. Additionally, two iterations of triangular smoothing were applied to the IMC potential update, ΔU. The cutoff was chosen at 0.9 nm with a grid spacing of 0.01 nm.

The reference radial distribution function, $g^{ref}(r)$, coarse-grained potentials, and corresponding radial distribution functions are shown in Figure 2a,b. IBI and IMC give practically the same interaction potential. Although the force-matched potential has a very similar structure with two minima, the corresponding radial distribution function is very different from the target one. Possible reasons for these discrepancies are discussed in refs 23, 25, and 34, and stem from the fact that FM aims to reproduce the many-body potential of mean force, which does not necessarily guarantee perfect pairwise distribution functions, considering the fact that the basis sets in the coarse-grained force field may be limited.

Note that all three methods lead to a different pressure of the coarse-grained system: 8000 bar (IBI), 9300 bar (IMC), and 6500 bar (FM). Different pressures for the iterative methods are due to a different accuracy of the potential update. Indeed, small changes of pressure can significantly affect the potential, especially its long tail.[15,42] However, they hardly change the radial distribution function due to the small compressibility of water. One can improve the

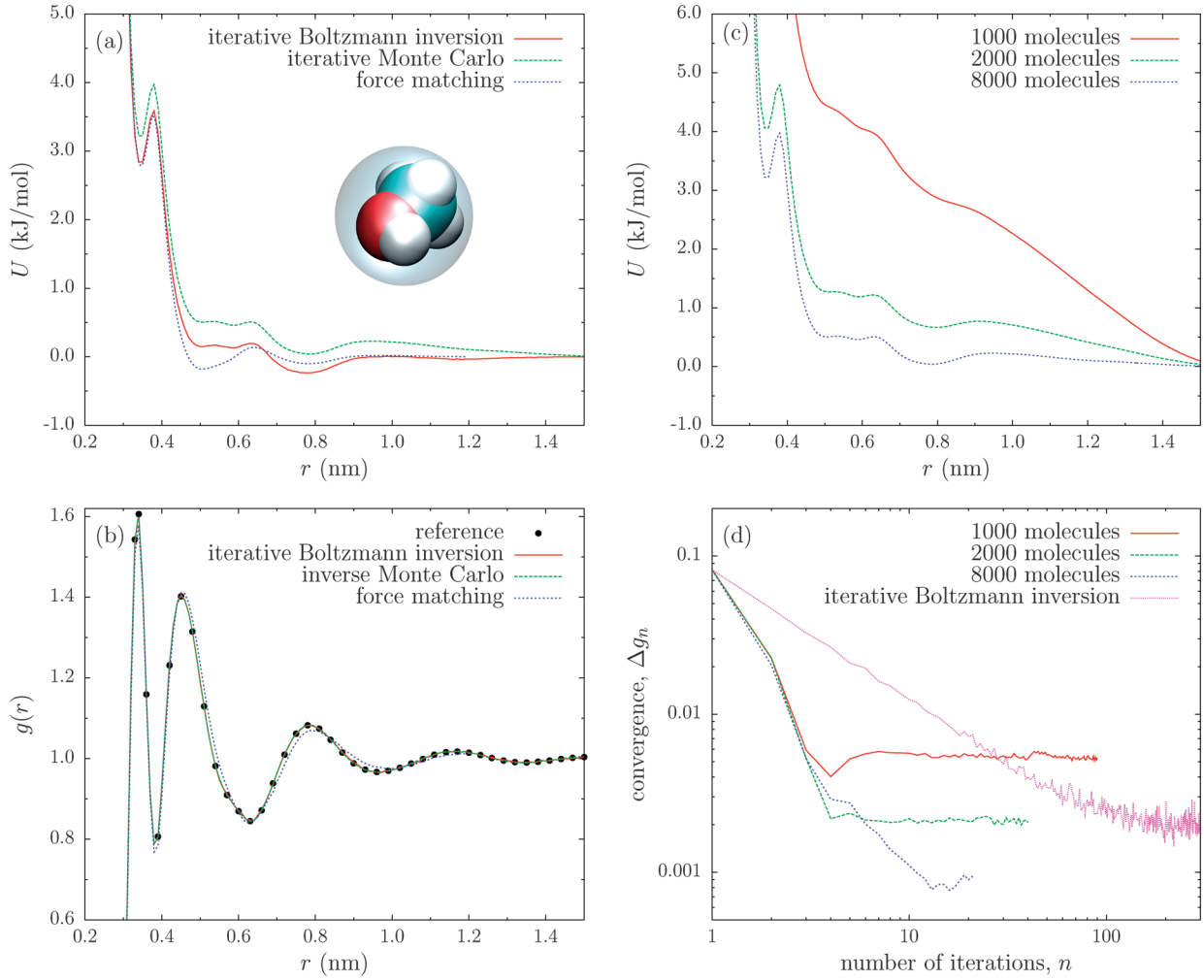

Figure 3. Methanol: (a) Coarse-grained potentials. (b) Corresponding radial distribution functions. (c) coarse-grained potentials using 10 IMC iterations for simulation boxes with 1000, 2000, and 8000 methanol molecules (box size 4.09, 5.15308, and 8.18 nm) equilibrated at the same density. (d) Root-mean-square deviation of reference and the current radial distribution function versus number of iterations. Similar to liquid water, IMC converges faster than IBI. The convergence saturates and the saturation error strongly depends on the system size. The inset of (a) shows the van der Waals excluded volume and coarse-grained representations of a methanol molecule.

agreement between the iterative methods by using pressure correction terms for the update.

The performance of the iterative methods depends on two factors: (i) the average (over all bins) error of the potential update $\varepsilon_{\Delta U}$; and (ii) the number of iterations required for convergence. We define the average error as

$$\varepsilon_{\Delta U} = \frac{1}{N}\sum_{i=0}^{N} \varepsilon(\Delta U(r_i)) \tag{17}$$

where N is the number of bins and $\varepsilon(\Delta U(r_i))$ is the error of the update function at a separation r_i. $\varepsilon(\Delta U(r_i))$ was calculated using a Jackknife analysis.[43]

The average error of the potential update is shown in Figure 2c as a function of the run length. One can see that, for both methods, the error decreases as $1/\sqrt{L}$, where L is the number of snapshots used for averaging. However, the prefactor for the IBI update error, which is based on the radial distribution function, is at least 10 times smaller than of the IMC update error, which makes use of cross-correlations of S_α. This observation implies that, in order to have the same accuracy of the update function, IMC needs significantly longer sampling.

This disadvantage is, of course, compensated by the efficiency of the update function, which is assessed by computing the root-mean-square deviation, Δg_n, of the current and target radial distribution functions:

$$\Delta g_n^2 = \int [g^{\mathrm{ref}}(r) - g^{(n)}(r)]^2 \, \mathrm{d}r \tag{18}$$

Δg_n is plotted as a function of the number of iterations, n, in Figure 2d. It is clear that IMC converges much faster than IBI, though the root-mean-square deviation saturates after some number of iterations.

4.2. Coarse-Graining of Methanol. Liquid methanol (see the inset in Figure 3) is the second example of coarse-graining of nonbonded interactions that we present here. In fact, FM has already been used to coarse-grain this system,[42] and contrary to water, the liquid structure (radial distribution function) is well reproduced by the FM coarse-grained

potential. In addition, the excluded volume of methanol is larger than that of water, and the undulations of the radial distribution function extend up to 1.5 nm. As we will see, this leads to pronounced finite size effects for IMC, since it has a nonlocal potential update. FM and IBI do not have this problem, since the IBI potential energy update is local, and FM is based on pair forces. The range of the latter is much shorter than the correlation length of structural properties (such as undulations of the radial distribution function), which may propagate over the boundaries for small boxes.

Simulation parameters were taken from ref 42, and OPLS[44,45] all-atom force field was used. Atomistic simulations were performed with 1000 methanol molecules in a cubic box (4.09 nm box size) at 300K using the Nosé−Hoover thermostat.[46,47] The system was equilibrated for 2 ns followed by a production run of 18 ns. The reference radial distribution function was calculated using snapshots every 0.5 ps and is shown in Figure 3b.

The FM potential was calculated using blocks of six frames each and using a spline grid of 0.02 nm. With this potential, coarse-grained simulations were performed using a stochastic dynamics integrator and using 1000 beads with the same box size and the same temperature as in the atomistic simulations. The system was equilibrated for 40 ps followed by a production run of 160 ps. Snapshots were stored every 5 ps and used to calculate the radial distribution function.

For the iterative procedures, the potential of mean force was taken as an initial guess. The cutoff was chosen at 1.54 nm with a grid spacing of 0.01 nm. For IBI, 300 iterations were performed using stochastic dynamics with the same parameters used in the FM-based procedure. The IMC iterations were performed with 8000 molecules and a box size of 8.18 nm. The total length of the run was 1 ns, and snapshots were stored every 0.2 ps. Two smoothing steps were used at each iteration for the potential update, ΔU.

The coarse-grained potentials for all methods are shown in Figure 3a. In spite of small differences between the coarse-grained potentials, the agreement between the reference and the coarse-grained radial distribution functions is excellent, as can be seen from Figure 3b.

It is important to mention that the IMC method, which has a nonlocal update, is prone to systematic errors due to finite size effects and, hence, requires much larger simulation boxes in order to calculate the potential update. This is due to artificial cross-correlations of S_α at large distances, which lead to a small difference of tails between the coarse-grained and the reference radial distribution functions, and, as a consequence, to a much higher pressure of the coarse-grained system and a significantly different coarse-grained potential. In contrast, IBI and FM work well with system sizes of the order of two radial distribution function cutoff lengths.

To illustrate this point, we prepared simulation boxes of three different sizes, with 1000, 2000, and 8000 methanol molecules (box size of 4.09, 5.15308, and 8.18 nm and simulation times of 3, 2, and 1 ns, respectively). The IMC iterative procedure was repeated until the potentials converged, and these are shown in Figure 3c. One can see that the potentials significantly differ from each other. These differences lead to small deviations in the tail of the radial distribution function, which, however, vanish in a systematic way for bigger boxes, as illustrated in Figure 3d where we plot the integral of the difference of the reference and the current distribution functions.[64]

To summarize, IMC should be used with care for small systems. The potential update (or the coarse-grained potential) must be converged with respect to the simulation box size. In the case of methanol coarse-graining, a box of size three times the radial distribution function cutoff was not enough to achieve the converged potential for IMC, even though this is sufficient for IBI and FM methods.

4.3. Liquid Propane: From an All- To an United-Atom Description. So far we have illustrated coarse-graining of nonbonded degrees of freedom using liquid water and methanol as examples. Here we show how bonded interactions can be coarse-grained by deriving a united-atom model (i.e., hydrogens embedded into heavier atoms) from an all-atom model of liquid propane.[65] The mapping scheme as well as the bonded coarse-grained variables (two bonds, b, and one angle, θ) are shown in the inset of Figure 4. Note that this coarse-graining scheme has two different bead types: an inner bead, of type B, with two hydrogens, and two outer beads, of type A, with three hydrogens. As a result, three types of nonbonded interactions, U_{AA}, U_{BB}, and U_{AB} must be determined.

As before, atomistic simulations were performed using the OPLS all-atom force field.[44,45] A box of liquid propane was first equilibrated at 200K and 1 bar in the NPT ensemble for 10 ns, using the Berendsen thermostat and barostat.[40] The equilibrated box of the size $4.96337 \times 5.13917 \times 4.52386$ nm^3 was then simulated for 10 ns in the NVT ensemble at 200K using velocity rescaling.[48] No bond constraints were used during the simulations, and hence, the integration time step was 1 fs. Snapshots were written every 1 ps.

In the case of iterative methods, the bonded potentials (bond and angle) were calculated by Boltzmann-inverting the corresponding distribution functions of a single molecule in vacuum, according to eq 5. The propane molecule in vacuum was simulated in an stochastic dynamics run[41] for 100 ns with snapshots stored every 2 ps. Nonbonded potentials were iteratively refined by using IBI with a grid spacing of 0.01 nm and a cutoff of 1.36 nm (1.38 nm) for A−A, A−B (B−B) interaction types, respectively. The run length for each iteration was 50 ps with snapshots written every 0.5 ps. At every iteration step, only one interaction type was corrected. When using the FM method, both bonded and nonbonded potentials were obtained at the same time, since FM does not require the explicit separation of bonded and nonbonded interactions.

The obtained potentials are shown in Figure 4a, c, and d. FM and Boltzmann inversion-derived bond and angle potentials (Figure 4c and d) perfectly agree with each other. The nonbonded potentials, shown in Figure 4a, are not completely identical but have similar shapes and barrier heights. This, of course, results in a good reproducibility of the propane liquid structure by the FM-based coarse-grained potentials, as can bee seen from the radial distribution functions shown in Figure 4b. Again, as expected, IBI reproduces the reference radial distribution functions exactly.

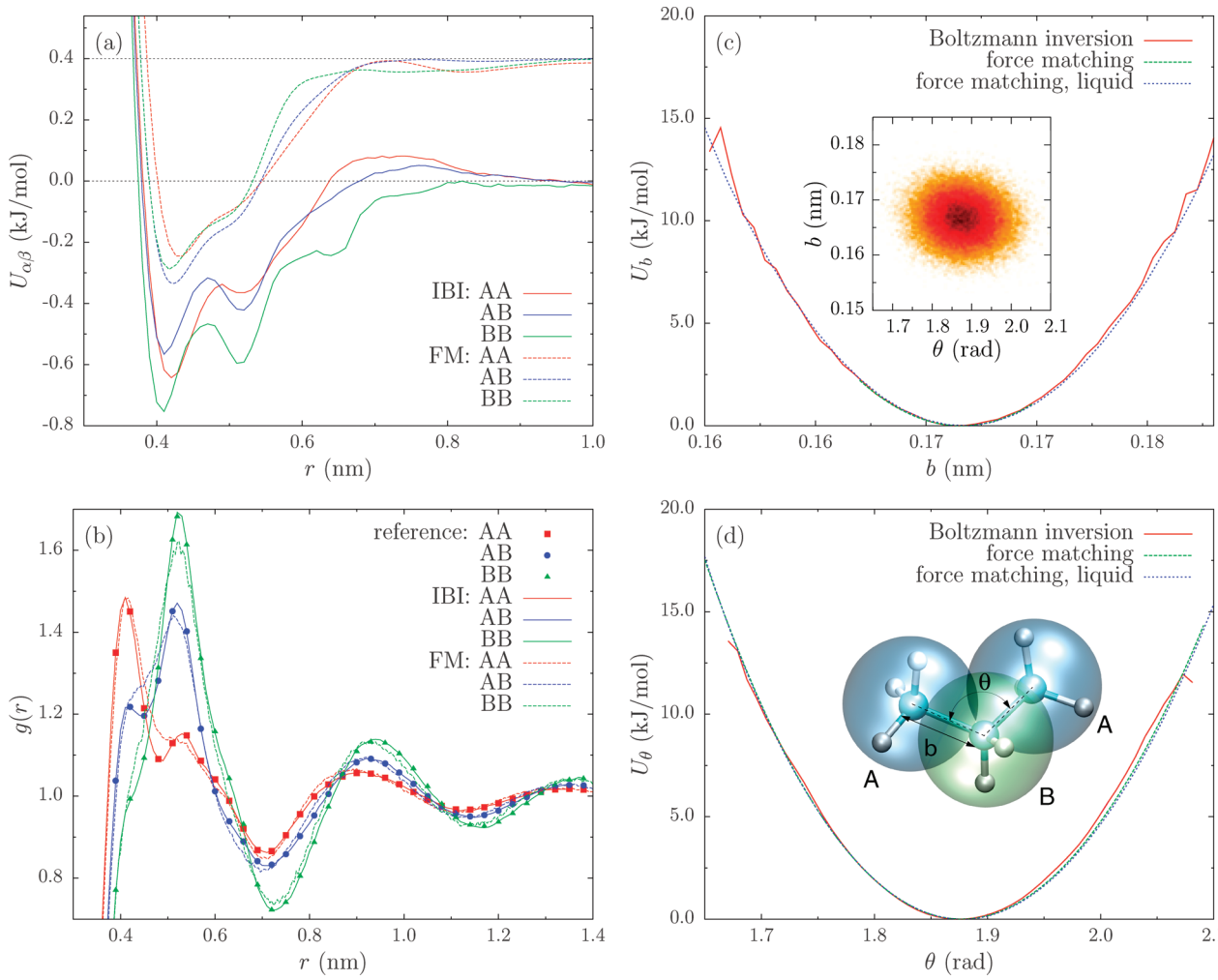

Figure 4. Propane: (a) Nonbonded interaction potentials U_{AA}, U_{BB}, and U_{AB} obtained with IBI and FM methods. For clarity, FM potentials are offset along the *y*-axis. (b) Corresponding radial distribution functions plotted together with the atomistic radial distribution function. (c) Bond potential obtained for a single molecule in vacuum by Boltzmann-inverting the corresponding distribution function, using FM for a single propane molecule in vacuum and using force matching for liquid propane. (d) Angular coarse-grained potentials. The inset of (c) shows the correlations of b and θ. The inset of (d) shows all-atom and coarse-grained representations of a propane molecule, bead types, and coarse-grained bonded degrees of freedom (bond b and angle θ).

To summarize, the united-atom model of liquid propane is an ideal example of coarse-graining where the structure- and force-based methods result in similar bonded and nonbonded interaction potentials. As we will see later, this is due to: (i) the completeness of the basis set used to construct the coarse-grained force field; and (ii) independence of bond and angular degrees of freedom. The latter can be understood with the help of a histogram showing the correlation of b and θ, depicted in the inset of Figure 4c.

In the next section, we will look at coarse-graining of a single molecule of hexane, for which this is not the case.

4.4. Angular Potential of a Hexane Molecule. The final example we would like to discuss here is the angular potential of a hexane coarse-grained into a three-bead chain, with two carbon atoms per bead (see the inset in Figure 5a). Atomistic simulations of a single hexane molecule in vacuum were performed using an all-atom OPLS force field and a stochastic dynamics integrator.[41] The run length was 1000 ns, and the snapshots were stored every 2 ps.

The coarse-grained angular potential was again obtained by Boltzmann-inverting the angular distribution function or by using the FM method (we used blocks of 50000 frames each, spline grid of 0.05 nm, and sampling in the $\theta \in [1.6, 3.14]$ interval). Both coarse-grained potentials are shown in Figure 5a. The corresponding distribution functions, together with the reference function obtained from the atomistic simulations, are shown in Figure 5b.

It is obvious that the distribution, which corresponds to simple Boltzmann inversion, is practically identical to the reference distribution, while the FM-based distribution samples small angles much more often, which is a direct consequence of a very deep local minimum in the angular potential at these angles. It is easy to understand why FM fails to predict the relative height of this minimum. On a coarse-grained level, the change of the angle from large to small values corresponds to the reorientation of the dihedral angles at the atomistic level. This reorientation results in instantaneous forces, $\boldsymbol{f}_1$, $\boldsymbol{f}_2$, and $\boldsymbol{f}_3$, on the beads which have

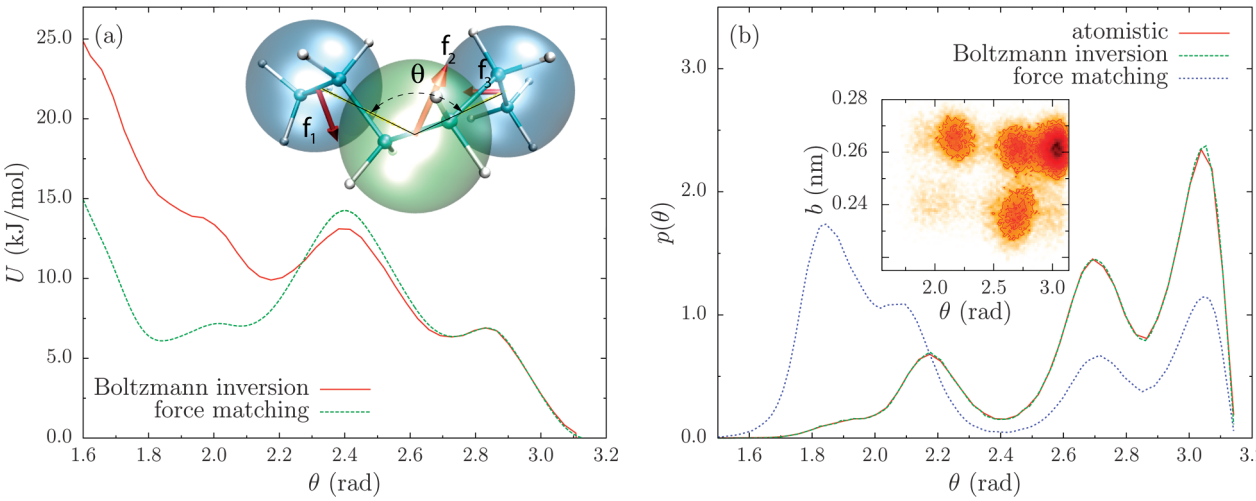

Figure 5. Hexane: (a) Coarse-grained angular potentials obtained using Boltzmann inversion (no iterations) and using FM for a single hexane molecule in vacuum. The inset of (a) shows the hexane molecule and its coarse-grained representation. Arrows indicate the directions of the forces on three beads for a specific snapshot. (b) Probability density (probability distribution normalized by the interval) obtained from the atomistic run as well as from the runs using coarse-grained angular potentials. The inset of (b) shows the correlation of b and θ.

an out of plane component, where the plane is defined by the centers of the beads (see also the inset of Figure 5a). The coarse-grained potential, however, has only an angular term, U_θ, and can only capture forces which lie in the plane in which the angle θ is defined. Hence, only the projections of the forces on this plane are used in FM, and this clearly leads to underestimation of the position of the second minimum, since the work conducted by the out-of-plane forces is completely ignored.[66]

Additionally, this mapping scheme does not have independent variables, e.g., bond and angle degrees of freedom are coupled, as can be seen from the Ramachandran plot shown in the inset of Figure 5b. This means that, even though Boltzmann inversion reproduces correct distributions, sampling of the configurational space is incorrect because of the lack of cross-correlation terms in the coarse-grained potential.

This example clearly shows that coarse-graining shall be used with understanding and caution, the methods should be cross-checked with respect to each other as well as with respect to the reference system.

5. Conclusions

To conclude, we have presented a flexible toolkit for developing and testing coarse-graining methods. Three of them, namely iterative Boltzmann inversion, inverse Monte Carlo, and force matching, have been implemented. With the help of the developed toolkit, we have coarse-grained liquid water, methanol, and propane and a single molecule of hexane. We have also illustrated several advantages as well as shortcomings of the implemented methods. For example, inverse Monte Carlo has an update function which is more efficient than that of the iterative Boltzmann inversion method. On the other hand, inverse Monte Carlo is very sensitive to the system size and the statistical averaging. We have also discussed problems one might encounter when using force matching due to incompleteness of the basis set used to represent the coarse-grained potential energy surface. It should always be kept in mind that the coarse-grained systems are physically different to the reference systems and that the coarse-graining methods cannot be used as a black box and require thorough cross-checking.

We shall also mention that the toolkit has an interface to the fast molecular orbital overlap calculations library and kinetic Monte Carlo code. Combined, these three packages have already been used to study self-assembly and charge transport in organic semiconductors.[49,50]

The source code of VOTCA is available on request and will soon be released under a public license.

Acknowledgment. This work was partially supported by DFG via IRTG program between Germany and Korea, DFG grant AN 680/1-1 and by the Multiscale Modeling Initiative of the Max Planck Society. POLYMAT graduate program (V.R.) and Eurosim Early Stage Training project of Marie Curie actions (A.L.) are acknowledged for the financial support. We are grateful to Christine Peter, Will Noid, Brad Lambeth, Markus Deserno, and Karen Johnston for useful discussions. V.R. is thankful to Alexander Lyubartsev for numerous discussions during his stay at Stockholm University.

Appendix

A. Force Matching Using Cubic Splines. Implementations of force matching using different basis functions (linear splines, cubic splines, and step functions) and different methods for solving the least-squares problem (QR decomposition, singular value decomposition, iterative techniques, and normal matrix approach) are discussed in detail in ref 45.

Here, we outline the implementation using cubic splines as basis functions, QR-decomposition for solving the least-squares problem, and block averaging to sample large trajectories.

In our implementation the force $\boldsymbol{f}_{\gamma_i}(\{\boldsymbol{r}_k\})$ acting on bead i due to an interaction γ_i with the potential U_{γ_i} can be written as

$$\begin{aligned} \boldsymbol{f}_{\gamma i}(\{\boldsymbol{r}_k\}) &= -\nabla_i U(\kappa(\{\boldsymbol{r}_k\})) \\ &= -\frac{\partial U}{\partial \kappa}\nabla_i \kappa(\{\boldsymbol{r}_k\}) \\ &= -f_{\gamma i}\nabla_i \kappa(\{\boldsymbol{r}_k\}) \end{aligned} \tag{19}$$

where $\kappa = r,\ b,\ \theta$, and φ denotes the type of interaction and ∇_i is the gradient with respect to the coordinates $\boldsymbol{r}_i$ of bead i. The variable κ can label nonbonded interactions, bonds, angles, or dihedral angles, which are given by the distance between the two beads, the bond length, and the angle, which depends on three beads or on the dihedral angle defined using four beads, respectively. Now, the total force $\boldsymbol{f}_i^{\mathrm{cg}}$, acting on coarse-grained bead i, can be expressed in terms of the coarse-grained interactions, and eq 16 can be rewritten as

$$\sum_{\gamma_i} f_{\gamma_i}(\kappa)\nabla_i \kappa(\{\boldsymbol{r}_{kl}\}) = \boldsymbol{f}_{il}^{\mathrm{ref}} \tag{20}$$

where γ_i enumerates all interactions acting on bead i.

$f(\kappa)$ is interpolated using cubic splines connecting a set of points $\{\kappa_k\}$:

$$\begin{aligned} S_n(\kappa, \{\kappa_k\}, \{f_k\}, \{f_k''\}) &= A_n(\kappa) f_n \\ &+ B_n(\kappa) f_{n+1} \\ &+ C_n(\kappa) f_n'' \\ &+ D_n(\kappa) f_{n+1}'' \end{aligned} \tag{21}$$

where $\{f_k\}$and $\{f_k''\}$ are tabulations of $f(\kappa)$ and its second derivative on the grid $\{\kappa_k\}$, the parameters $\{f_k\}$ and $\{f_k''\}$ are obtained from the fit, $\kappa \in [\kappa_n, \kappa_{n+1}]$, and the coefficients A_n, B_n, C_n, and D_n have the following form:

$$\begin{aligned} A_n(\kappa) &= 1 - \frac{\kappa - \kappa_n}{h_{n+1}} \\ B_n(\kappa) &= \frac{\kappa - \kappa_n}{h_{n+1}} \\ C_n(\kappa) &= \frac{1}{2}(\kappa - \kappa_n)^2 - \frac{1}{6}\frac{(\kappa - \kappa_n)^3}{h_{n+1}} - \frac{1}{3}h_{n+1}(\kappa - \kappa_n) \\ D_n(\kappa) &= \frac{1}{6}\frac{(\kappa - \kappa_n)^3}{h_{n+1}} - \frac{1}{6}h_{n+1}(\kappa - \kappa_n) \end{aligned} \tag{22}$$

where $h_n = \kappa_{n+1} - \kappa_n$.

An additional requirement on the spline coefficients is the continuity of the first derivatives:

$$\begin{aligned} A_n(\kappa_{n+1})' f_n + B_n(\kappa_{n+1})' f_{n+1} + C_n(\kappa_{n+1})' f_n'' + \\ D_n(\kappa_{n+1})' f_{n+1}'' = A_{n+1}(\kappa_{n+1})' f_{n+1} + B_{n+1}(\kappa_{n+1})' f_{n+2} + \\ C_{n+1}(\kappa_{n+1})' f_{n+1}'' + D_{n+1}(\kappa_{n+1})' f_{n+2}'' \end{aligned} \tag{23}$$

If the total number of grid points is $N + 1$ ($n = 0, 1, ..., N$), then these conditions are specified for the points $n = 0, 1, ..., N - 1$. For nonperiodic potentials, the end points are treated using normal boundary conditions, i.e., $f_0'' = 0$ and $f_N'' = 0$.

Due to the spline fitting, eq 20 simplifies to a set of linear equations with respect to the fitting parameters f_n and f_n''. The complete set of equations to solve, therefore, consists of eq 20 and constraints, eq 23. Strictly speaking, this set of equations cannot be solved in a least-squares sense using simple QR decomposition. The reason is that the constraints shall be satisfied exactly to ensure the continuity of the first derivative of the potential, which is not the case if they are treated in a least-squares sense. To solve the problem, one, in principle, has to use a constrained least-squares solver.[51] From a practical point of view, however, it is simpler to treat the constraints in a least-squares sense for each block. This will only lead to a piecewise smooth potential, but the smoothness can be "recovered" by averaging over the blocks.

References

(1) Tschöp, W.; Kremer, K.; Batoulis, J.; Burger, T.; Hahn, O. *Acta Polym.* **1998**, *49*, 61–74.

(2) Shelley, J.; Shelley, M.; Reeder, R.; Bandyopadhyay, S.; Klein, M. *J. Phys. Chem. B* **2001**, *105*, 4464–4470.

(3) Abrams, C.; Kremer, K. *Macromolecules* **2003**, *36*, 260–267.

(4) Murtola, T.; Falck, E.; Patra, M.; Karttunen, M.; Vattulainen, I. *J. Chem. Phys.* **2004**, *121*, 9156–9165.

(5) Izvekov, S.; Voth, G. *J. Chem. Phys.* **2005**, *123*, 134105.

(6) Sun, Q.; Faller, R. *J.Chem. Theo. Comp.* **2006**, *2*, 607–615.

(7) Harmandaris, V.; Adhikari, N.; van der Vegt, N.; Kremer, K. *Macromolecules* **2006**, *39*, 6708–6719.

(8) Yelash, L.; Müller, M.; Wolfgang, P.; Binder, K. *J. Chem. Theor. Comp.* **2006**, *2*, 588–597.

(9) Shih, A.; Arkhipov, A.; Freddolino, P.; Schulten, K. *J. Phys. Chem. B* **2006**, *110*, 3674–3684.

(10) Lyubartsev, A. *Eur. Biophys. J.* **2005**, *35*, 53–61.

(11) Zhou, J.; Thorpe, I.; Izvekov, S.; Voth, G. *Biophys. J.* **2007**, *92*, 4289–4303.

(12) Villa, A.; van der Vegt, N.; Peter, C. *Phys. Chem. Chem. Phys.* **2009**, *11*, 2068–2076.

(13) Ercolessi, F.; Adams, J. B. *Europhys. Lett.* **1994**, *26*, 583–588.

(14) Hess, B.; Holm, C.; van der Vegt, N. *Phys. Rev. Lett.* **2006**, *96*, 147801.

(15) Reith, D.; Pütz, M.; Müller-Plathe, F. *J. Comput. Chem.* **2003**, *24*, 1624–1636.

(16) Lyubartsev, A.; Laaksonen, A. *Phys. Rev. E: Stat. Phys., Plasmas, Fluids, Relat. Interdiscip. Top.* **1995**, *52*, 3730–3737.

(17) Soper, A. *Chem. Phys.* **1996**, *202*, 295–306.

(18) Izvekov, S.; Parrinello, M.; Burnham, C. J.; Voth, G. A. *J. Chem. Phys.* **2004**, *120*, 10896–10913.

(19) Toth, G. *J. Phys. Cond. Mat.* **2007**, *19*, 335222.

(20) Baschnagel, J.; Binder, K.; Doruker, P.; Gusev, A. A.; Hahn, O.; Kremer, K.; Mattice, W. L.; Müller-Plathe, F.; Murat, M.; Paul, W.; Santos, S.; Suter, U. W.; Tries, V. *Advances in Polymer Science: Viscoelasticity, Atomistic Models, Statistical Chemistry*; Springer Verlag: Heidelberg, Germany, 2000.

(21) Kremer, K. In *Soft and fragile matter, nonequilibrium dynamics, metastability and flow*; Cates, M. E., Evans, M. R., Eds.; J. W. Arrowsmith Ltd.: Bristol, U.K., 2000.

(22) Müller-Plathe, F. *Chem. Phys. Phys. Chem.* **2002**, *3*, 754–769.

(23) Johnson, M.; Head-Gordon, T.; Louis, A. *J. Chem. Phys.* **2007**, *126*, 144509.

(24) *Coarse-Graining of Condensed Phase and Biomolecular Systems*; Voth, G. A., Ed.; CRC Group: Boca Raton, FL, 2008.

(25) Murtola, T.; Bunker, A.; Vattulainen, I.; Deserno, M.; Karttunen, M. *Phys. Chem. Chem. Phys.* **2009**, *11*, 1869–1892.

(26) Peter, C.; Kremer, K. Soft Matter 2009, accepted. DOI: 10.1039/b912027k.

(27) Henderson, R. *Phys. Lett. A* **1974**, *A49*, 197–198.

(28) Chayes, J.; Chayes, L. *J. Stat. Phys.* **1984**, *36*, 471–488.

(29) Chayes, J.; Chayes, L.; Lieb, E. *Comm. Math. Phys.* **1984**, *93*, 57–121.

(30) Leon, S.; van der Vegt, N.; Delle Site, L.; Kremer, K. *Macromolecules* **2005**, *38*, 8078–8092.

(31) Junghans, C.; Praprotnik, M.; Kremer, K. *Soft Matter* **2008**, *4*, 156–161.

(32) Wang, H.; Junghans, C.; Kremer, K. *Eur. Phys. J. E* **2009**, *28*, 221–229.

(33) Murtola, T.; Falck, E.; Karttunen, M.; Vattulainen, I. *J. Chem. Phys.* **2007**, *126*, 075101.

(34) Noid, W.; Chu, J.; Ayton, G.; Voth, G. *J. Phys. Chem. B* **2007**, *111*, 4116–4127.

(35) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. *J. Chem. Theo. Comp.* **2008**, *4*, 435–447.

(36) Nezbeda, I.; Slovak, J. *Mol. Phys.* **1997**, *90*, 353–372.

(37) Wallqvist, A.; Mountain, R. D. *Rev. Comp. Chem.* **2007**, *13*, 183–247.

(38) Kusalik, P. G.; Svishchev, I. M. *Science* **1994**, *65*, 1219–1221.

(39) Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. *J. Phys. Chem.* **1987**, *91*, 6269–6271.

(40) Berendsen, H. J. C.; Postma, J. P. M.; Vangunsteren, W. F.; Dinola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684–3690.

(41) Gunsteren, W. F. V.; Berendsen, H. J. C. *Mol. Sim.* **1988**, *1*, 173.

(42) Noid, W. G.; Liu, P.; Wang, Y.; Chu, J.; Ayton, G.; Izvekov, S.; Andersen, H. C.; Voth, G. *J. Chem. Phys.* **2008**, *128*, 244115.

(43) Janke, W. *Statistical Analysis of Simulations: Data Correlations and Error Estimation*, Lecture notes; Grotendorst, J., Marx, D., Muramatsu, A., Eds.; John von Neumann Institut für Computing (NIC) Series, Vol. 10; NIC: Jülich, Germany, 2002; pp 423−445.

(44) Jorgensen, W.; Tirado-Rives, J. *J. Chem. Soc., Abstr.* **1998**, *216*, U696–U696.

(45) Jorgensen, W.; Tirado-Rives, J. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6665–6670.

(46) Nosé, S. *Mol. Phys.* **1984**, *52*, 255–268.

(47) Hoover, W. G. *Phys. Rev. A: At., Mol., Opt. Phys.* **1985**, *31*, 1695.

(48) Bussi, G.; Donadio, D.; Parrinello, M. *J. Chem. Phys.* **2007**, *126*, 014101.

(49) Kirkpatrick, J.; Marcon, V.; Nelson, J.; Kremer, K.; Andrienko, D. *Phys. Rev. Lett.* **2007**, *98*, 227402.

(50) Feng, X.; Marcon, V.; Pisula, W.; Hansen, M.; Kirkpatrick, J.; Grozema, F.; Andrienko, D.; Kremer, K.; Müllen, K. *Nat. Mat.* **2009**, *8*, 421–426.

(51) Golub, G. H.; Van Loan, C. F. *Matrix Computations*, 3rd ed.; Johns Hopkins University Press: Baltimore, MD, 1996.

(52) Harmandaris, V. A.; Reith, D.; Van der Vegt, N. F. A.; Kremer, K. *Macromol. Chem. Phys.* **2007**, *208*, 2109–2120.

(53) Villa, A.; Peter, C.; van der Vegt, N. F. A. *Phys. Chem. Chem. Phys.* **2009**, *11*, 2077–2086.

(54) Noid, W. G.; Chu, J.; Ayton, G. S.; Krishna, V.; Izvekov, S.; Voth, G.; Das, A.; Andersen, H. C. *J. Chem. Phys.* **2008**, *128*, 244114.

(55) See also footnote 65.

(56) Note that the coordinates $\{\boldsymbol{R}_I\}$ which are obtained from an atomistic trajectory shall not be confused with the coordinates of a trajectory obtained from coarse-grained simulations.

(57) Note that here we only consider the special case of a linear relation between the $\boldsymbol{r}$ and $\boldsymbol{R}$. $\hat{\mathbf{M}}$ is a block-diagonal matrix and to construct it is enough to specify building blocks for each molecule type. For polymers it is enough to specify $\hat{\mathbf{M}}$ for one repeat unit only.

(58) Note that this is often a "special" trajectory which is designed to decouple the degrees of freedom of interest, e. g., a single polymer chain in vacuum with appropriate exclusions.[1]

(59) Note that, as before, we ignored an irrelevant normalization prefactor Z.

(60) Checking the linear correlation coefficient does not guarantee statistical independence of variables, for example $c(x, x^2) = 0$ if x has a symmetric probability density $P(x) = P(-x)$. This case is often encountered in systems used for coarse-graining.[52,53] The concept is illustrated in section IV for liquid propane and a single molecule of hexane.

(61) Note that eq 7 is nothing else but a numerical scheme that allows one to match the coarse-grained and the reference distribution functions. It can be seen as a firstorder correction to the interaction potential with respect to a gas of non-interacting particles. Indeed, in an ideal gas, the probability of finding two particles at a distance r is $P^{(0)} = 4\pi r^2$, which is equivalent to the statement that the radial distribution function of an ideal gas is 1. Substituting $P^{(0)}$ into eq 7 we obtain the first iteration $U^{(1)} = -k_B T \ln(P_{ref}/4\pi r^2)$, which is the potential of mean force, eq 2.

(62) A formal statistical mechanical framework of force matching applied to a liquid state, or a multiscale coarse-graining method, is given in ref 54.

(63) http://www.espresso-pp.de.

(64) More detailed analyses have shown that, for small boxes, an additional linear term in the potential update at large separations appear. To determine the origin of this term, ΔU was calculated using the full matrix $\mathbf{A}_{\alpha\beta}$ as well as only its diagonal elements. The potential after 50 IBI iterations was taken as an initial guess. Without the off-diagonal elements ΔU was small once the reference and coarse-grained radial distribution functions were matching each other. Inclusion of the off-diagonals elements always resulted in an additional, practically linear, term in the potential update which became smaller for large boxes. Based on this observation we concluded that the off-diagonal elements of the matrix $\mathbf{A}_{\alpha\beta}$ systematically change with the box size.

(65) The united atom model we use here shall not be confused with the united atom models commonly used in the atomistic force-field community, for example OPLS-UA forcefield.[44,45] The latter models map the potentials, which are analytical functions of bonds, angles, and dihedral angles, onto thermo-

dynamic properties of the corresponding substances. In our case coarse-grained potentials are tabulated functions of coarse-grained variables and only the mapping (hydrogens embedded into heavier atoms) is similar to that of the united atom force-fields.

(66) For condensed phase systems, the error introduced by the off-plane component of the force might be compensated by some other pair interactions. In this particular case, however, coarse-graining of liquid hexane with both bonded and non-bonded degrees of freedom treated simultaneously results in a very similar angular distribution to that of a single molecule in vacuum.

CT900369W