

RESEARCH ARTICLE

Relative Entropy and Optimization-Driven Coarse-Graining Methods in VOTCA

S. Y. Mashayak¹, Mara N. Jochum², Konstantin Koschke², N. R. Aluru¹, Victor Rühle³, Christoph Junghans^{4*}

1 Department of Mechanical Science and Engineering, Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, Urbana, Illinois, 61801, United States of America, **2** Max Planck Institute for Polymer Research, Ackermannweg 10, 55128 Mainz, Germany, **3** Department of Chemistry, University of Cambridge, Cambridge, United Kingdom, **4** Computer, Computational, and Statistical Sciences Division, Los Alamos National Laboratory, Los Alamos, NM 87545, United States of America

* junghans@lanl.gov



CrossMark
click for updates

OPEN ACCESS

Citation: Mashayak SY, Jochum MN, Koschke K, Aluru NR, Rühle V, Junghans C (2015) Relative Entropy and Optimization-Driven Coarse-Graining Methods in VOTCA. PLoS ONE 10(7): e0131754. doi:10.1371/journal.pone.0131754

Editor: Xuhui Huang, Hong Kong University of Science and Technology, HONG KONG

Received: November 12, 2014

Accepted: June 5, 2015

Published: July 20, 2015

Copyright: This is an open access article, free of all copyright, and may be freely reproduced, distributed, transmitted, modified, built upon, or otherwise used by anyone for any lawful purpose. The work is made available under the [Creative Commons CC0](https://creativecommons.org/licenses/by/4.0/) public domain dedication.

Data Availability Statement: The code with which the simulation has been performed can be found here: <https://code.google.com/p/votca/source/browse/?repo=csg>. The input files to the simulations shown in the paper have been published here: <https://code.google.com/p/votca/source/browse/?repo=csg-tutorials>.

Funding: SYM and NRA acknowledge financial support by the NSF under grant Nos. 480 1264282 and 1420882 and AFOSR under grant No. 11157642. CJ was financially supported by SFB 625 in the framework of the multiscale modeling initiative of the Max-Planck Society (M3). This research was

Abstract

We discuss recent advances of the VOTCA package for systematic coarse-graining. Two methods have been implemented, namely the downhill simplex optimization and the relative entropy minimization. We illustrate the new methods by coarse-graining SPC/E bulk water and more complex water-methanol mixture systems. The CG potentials obtained from both methods are then evaluated by comparing the pair distributions from the coarse-grained to the reference atomistic simulations. In addition to the newly implemented methods, we have also added a parallel analysis framework to improve the computational efficiency of the coarse-graining process.

1 Introduction

In recent years, coarse-grained simulations have become an important tool for investigating systems on larger time- and length-scales [1–6]. Here, we focus on the bottom-up approaches of coarse-graining (CG), which create a systematic link between two particle based descriptions of a given system by building a lower resolution model based on a reference higher resolution model.

In a coarse-grained representation, several atoms are grouped into a single CG unit and the effective interactions between them are determined. Such coarse-graining leads to a significant speed-up in the molecular dynamics (MD) simulations due to i) a reduced number of degrees of freedom and hence, the number of interactions to compute, ii) smoother interaction potentials among CG particles that enable larger time steps, and most importantly, iii) a speedup in the intrinsic dynamics of the system which leads to faster diffusion and thus, shorter equilibration times [7].

Although coarse-grained simulations allow one to reach longer time- and length-scales, they also, by definition, lose some information from the underlying reference system. For instance, when replacing a water molecule by a single sphere, one loses information about the

supported in part by the National Science Foundation under grant number NSF PHY11-25915. CJ thanks LANL for a Director's fellowship and the MPG for hospitality in 2013 at MPI-P. K. Koschke acknowledges funding by the Nestlé Research Center. MNJ acknowledges DFG Emmy Noether program for financial support. Assigned: LA-UR-13-26878. LANL is operated by Los Alamos National Security, LLC, for the National Nuclear Security Administration of the U.S. DOE under Contract DE-AC52-06NA25396. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: Nestlé Research Center partly funded this study. There are no patents, products in development or marketed products to declare. This does not alter the authors' adherence to all of the PLOS ONE policies on sharing data and materials.

molecule's orientation. Generally, CG systems are limited in their representability and transferability, i.e., they cannot simultaneously reproduce all of the properties of interest and they may not be applicable for thermodynamic states far away from the one at which the system was parametrized [8, 9].

There exist, however, several coarse-graining techniques of varying complexities and accuracies for modeling CG systems and optimizing their interactions. These range from structure-based methods, such as Boltzmann inversion, iterative Boltzmann inversion (IBI) [10], inverse monte carlo (IMC) [11], and relative entropy [12], to force-based methods like force matching (FM) [13, 14]. Structure-based methods target (multi-body) distribution functions and use the relation between the distribution functions and the potential of mean force (PMF) [10, 14] to derive effective CG interactions. In contrast, the FM method tries to reproduce forces on CG sites [14]. Detailed descriptions about the various aspects and methods for determining CG potentials can be found in ref. [1, 3, 14–16].

Given such a wide range of methods of varying complexity and properties that they target, the question arises as which method is the most suitable for a specific system in a desired application. This can be best answered by a comparative analysis of the CG potentials obtained from different methods. To facilitate the determination of these potentials and to allow for a direct comparison of the methods, the Versatile Object-Oriented Toolkit for Coarse-graining Applications (VOTCA) was developed by Rühle et. al [16], which is available under an open source software license.

In the first release of VOTCA, Boltzmann inversion, IBI, IMC, and FM had been implemented. In ref. [16], the VOTCA package was used to compare IBI, IMC, and FM by coarse-graining four test systems, namely SPC/E water, methanol, liquid propane, and a single chain of hexane. For example, it was observed that the IMC update function is more efficient than that of IBI method. IMC, however, requires significantly more statistics, and hence longer trajectories, and can be sensitive to finite size effects. On the other side, FM can have representability issues if the set of coarse-grained interactions is incomplete. For such cases, three hybrid schemes to combine FM for non-bonded and the Boltzmann inversion for bonded interactions have been discussed in ref. [17] and applied to liquid hexane.

Recently, we added two new methods to the VOTCA package, namely the simplex method [18] and relative entropy method [12].

The simplex algorithm is a general optimization method, which can be used to optimize coarse-grained interactions. As such, it is difficult to target properties other than structure, e.g. thermodynamic quantities. Nevertheless, extensions for pressure correction [10] or the Kirkwood-Buff integral (KBI) [19] have been proposed but lack in a sound theoretical basis. However, structure-based methods, pressure correction, or KBI, are based on the relation between the CG interactions and target properties. By comparison, the simplex algorithm makes no *a priori* assumptions about the relation between the effective interactions and the target properties but instead requires the definition of an objective function which allows to fit arbitrary properties using a given functional form for the potentials. We call this approach targeted coarse-graining. In that sense, structure-based methods could be categorized as a subclass of targeted coarse-graining methods which target structural quantities and use a specialized optimizers like IMC or IBI.

Relative entropy is a method which quantifies the extent of the configurational phase-space overlap between two molecular ensembles [20]. It can be used as a measure of the discrepancies between various properties of the CG system's and the target all-atom (AA) ensemble. It has been shown by Shell S. [12] that one can minimize the relative entropy metric between the model CG system and the target AA system to optimize CG potential parameters such that the CG ensemble would mimic the target AA ensemble.

The main objective of this paper is to describe and demonstrate the two newly implemented methods in VOTCA. In the sections 2.1 and 2.2, we describe the formulations of the simplex-based and the relative entropy-based methods, respectively. Implementation details of the simplex and relative entropy methods are given in the sections 3.1 and 3.2, respectively. In the sections 4.1 and 4.2, we illustrate the simplex and relative entropy methods for a three-site SPC/E bulk water and a more complex water-methanol bulk mixture. To evaluate the accuracy of the CG potentials, we then compare the two-body distribution functions obtained from the CG MD system (CG-MD) with the reference AA MD system (AA-MD) simulations. In addition to these new methods, we have also made several improvements and other additions to VOTCA as described in the sections 3.4 and 3.5. For example, we have added a parallel analysis framework, which improves the performance of VOTCA by allowing it to utilize multi-core CPUs. These performance improvements are discussed in the section 3.4.

2 Methods

2.1 Downhill simplex algorithm

The downhill simplex algorithm [18] is a derivative-free optimization procedure designed to minimize an objective (or penalty) function, $y(x_1, x_2, \dots, x_n)$, with n degrees of freedom.

In the algorithm, the current state is described by $n+1$ vertices $\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_{n+1}$, with their corresponding penalty values from the objective function, which make up the geometric shape of a polytope. This polytope can then be transformed successively using basic operations such as a reflection, an expansion, a contraction, or a reduction in order to move towards the global minimum, i.e., minimize the objective function. The nature of progression of the algorithm depends on the combination of these transformations. In the following notation, we assume that the vertices at each step i are sorted according to their penalty values in ascending order.

The first transformation performed on the initial polytope is a *reflection*, in which the vertex with the highest objective value, \mathbf{p}_{n+1} , is reflected through the centroid of all remaining points, $\bar{\mathbf{p}}$, such that

$$\mathbf{p}_r = \bar{\mathbf{p}} + \alpha(\bar{\mathbf{p}} - \mathbf{p}_{n+1}), \quad \alpha > 0. \tag{2.1}$$

Here, \mathbf{p}_r is the reflected point and α is the reflection coefficient. This transformation preserves the volume of the simplex. If the penalty of the reflected point, y_r is lower than the best value, i.e., $y_r < y_1$, the algorithm tries to increase the step size along the line of the reflection. This is known as an *expansion*,

$$\mathbf{p}_e = \bar{\mathbf{p}} + \gamma(\bar{\mathbf{p}} - \mathbf{p}_{n+1}), \quad \gamma > 1, \tag{2.2}$$

where \mathbf{p}_e is the expanded point and γ is the expansion coefficient, which is the distance from \mathbf{p}_r through the centroid, $\bar{\mathbf{p}}$, to \mathbf{p}_e . In contrast, if the value of the reflected point is worse than the highest point, i.e., $y_r > y_{n+1}$, the move is too large. Thus, a *contraction* is performed to reduce the step size and find a point which decreases the objective function, where

$$\mathbf{p}_c = \bar{\mathbf{p}} + \rho(\bar{\mathbf{p}} - \mathbf{p}_{n+1}), \quad 0 > \rho > 1. \tag{2.3}$$

Here, \mathbf{p}_c is the contracted point and ρ is the contraction coefficient. If no improvement can be achieved upon contraction, i.e., $y_c > y_{n+1}$, a *reduction* is performed. This happens in situations close to a minimum by contracting the whole simplex around the lowest point, \mathbf{p}_1 , such that

$$\mathbf{p}_i \rightarrow \sigma(\mathbf{p}_i + \mathbf{p}_1), \tag{2.4}$$

where σ is the reduction coefficient. Subsequently, the procedure is restarted with a reflection.

In all other cases, the highest point, p_{n+1} , is first replaced by the newly reflected, expanded, or contracted point, all points are resorted according to their penalty values, and the procedure is then resumed.

The above described algorithm is applicable to a wide range of applications. In the context of coarse-graining, the evaluation of the objective function involves performing a coarse-grained simulation with a given set of parameters and subsequently analyzing its trajectory to compute the properties of interest. Here, the objective function calculates a single value, which is a cumulative measure of how well the target properties are reproduced. Hence, the algorithm constitutes a coarse-graining tool to find the optimal coarse-grained interaction parameters for nearly any measurable simulation property, even the optimal mapping for the CG model itself.

One of the benefits of the downhill simplex algorithm is that it does not require any derivatives of the objective function which are often difficult to obtain. However, problems with convergence may arise when poor initial guesses are provided which yield high penalty values or if the objective function contains too many parameters, such as $n > 10$. Furthermore, since the objective function is a cumulative measure of the properties of interest, providing an appropriate form which accounts for each of these equally is often challenging.

In this work, we set $\alpha = 1$, $\gamma = 2$, and $\rho, \sigma = 0.5$. However, as the performance of the simplex algorithm heavily depends on the type of system and the properties of interest studied, adapting these transformation coefficients may lead to a better performance than reported here.

2.2 Relative entropy

Relative entropy, S_{rel} , is defined as [12]

$$S_{\text{rel}} = \sum_i p_{\text{AA}}(r_i) \ln \left(\frac{p_{\text{AA}}(r_i)}{p_{\text{CG}}(M(r_i))} \right) + \langle S_{\text{map}} \rangle_{\text{AA}}, \quad (2.5)$$

where the sum is over all the configurations of the reference AA system, $r = \{r_i\} (i = 1, 2, \dots)$, M is the mapping operation to generate a corresponding CG configuration, R_i , from a AA configuration, r_i , i.e., $R_i = M(r_i)$, p_{AA} and p_{CG} are the configurational probabilities based on the AA and CG potentials, respectively, and $\langle S_{\text{map}} \rangle_{\text{AA}}$ is the mapping entropy due to the average degeneracy of AA configurations mapping to the same CG configuration, given by

$$S_{\text{map}}(R_i) = \ln \sum_i \delta_{R_i, M(r_i)}, \quad (2.6)$$

where δ is the Kronecker delta function. From Eq (2.6), it can be shown that the mapping entropy, $\langle S_{\text{map}} \rangle_{\text{AA}}$, does not depend on the CG interactions, but instead it is a unique function of the mapping operation, M , and the AA configurational weights. The relative entropy is a metric borrowed from the field of information theory, which quantifies the extent of the configurational phase-space overlap between two molecular ensembles [20]. The log-likelihood based derivation of the relative entropy for molecular systems, as defined in Eq (2.5), is given in ref. [12]. Physically, S_{rel} can be interpreted as the log probability that one test configuration of the model CG ensemble is representative of the target AA ensemble, and when the likelihood is a maximum, S_{rel} is at a minimum. Hence, the numerical minimization of S_{rel} with respect to the parameters of the CG model can be used to optimize the CG model.

Comparisons between relative entropy and other coarse-graining methods are made in ref. [14] and [15]. Chaimovich and Shell [15] have shown that for certain CG models relative entropy minimization produces the same CG potentials as other methods, e.g., it is equivalent to the IBI when CG interactions are modeled using finely tabulated pair additive potentials, and to the FM when a CG model is based on N -body interactions, where N is the number of

degrees of freedom in the CG model. However, there are some advantages of using relative entropy based coarse-graining. Relative entropy method allows to use analytical function forms for CG potentials, which are desired in theoretical treatments, such as parametric study of CG potentials, whereas, methods, like IBI, use tabulated potentials. Recently Lyubartsev et al [21] have shows how to use IMC with an analytical function form, too. BI, IBI, and IMC methods are based on pair correlations and hence, they are only useful to optimize 2-body CG potentials, whereas, relative entropy uses more generic metric which offers more flexibility in modeling CG interactions and not only 2-body, but also 3-body and N-body CG potentials can be optimized. In addition to the CG potential optimization, the relative entropy metric can also be used to optimize an AA to CG mapping operator.

In a canonical ensemble, substituting canonical configurational probabilities into Eq (2.5), the relative entropy simplifies to

$$S_{rel} = \beta \langle U_{CG} - U_{AA} \rangle_{AA} - \beta(A_{CG} - A_{AA}) + \langle S_{map} \rangle_{AA}, \tag{2.7}$$

where $\beta = 1/k_B T$, k_B is the Boltzmann constant, T is the temperature, U_{CG} and U_{AA} are the total potential energies from the CG and AA potentials, respectively, A_{CG} and A_{AA} are the configurational part of the Helmholtz free energies from the CG and AA potentials, respectively, and all the averages are computed in the reference AA ensemble.

Consider a model CG system defined by the CG potentials between various CG sites such that the CG potentials depend on the parameters $\lambda = \{\lambda_1, \lambda_2, \dots, \lambda_n\}$. As described above, in the relative entropy based coarse-graining, the CG potential parameters, λ , are optimized by the relative entropy minimization.

In this work, we use the Newton-Raphson strategy for the relative entropy minimization described in ref. [15]. In this strategy, the CG potential parameters, λ , are refined iteratively as

$$\lambda^{k+1} = \lambda^k - \chi \mathbf{H}^{-1} \cdot \nabla_{\lambda} S_{rel}, \tag{2.8}$$

where k is the iteration index, $\chi \in (0..1)$ is the relaxation parameter that can be adjusted to ensure convergence, $\nabla_{\lambda} S_{rel}$ is the vector of the first derivatives of S_{rel} with respect to λ , which can be computed from Eq (2.7) as

$$\nabla_{\lambda} S_{rel} = \beta \left\langle \frac{\partial U_{CG}}{\partial \lambda} \right\rangle_{AA} - \beta \left\langle \frac{\partial U_{CG}}{\partial \lambda} \right\rangle_{CG}, \tag{2.9}$$

and \mathbf{H} is the Hessian matrix of S_{rel} given by

$$\begin{aligned} \mathbf{H}_{ij} = & \beta \left\langle \frac{\partial^2 U_{CG}}{\partial \lambda_i \partial \lambda_j} \right\rangle_{AA} - \beta \left\langle \frac{\partial^2 U_{CG}}{\partial \lambda_i \partial \lambda_j} \right\rangle_{CG} \\ & + \beta^2 \left\langle \frac{\partial U_{CG}}{\partial \lambda_i} \frac{\partial U_{CG}}{\partial \lambda_j} \right\rangle_{CG} \\ & - \beta^2 \left\langle \frac{\partial U_{CG}}{\partial \lambda_i} \right\rangle_{CG} \left\langle \frac{\partial U_{CG}}{\partial \lambda_j} \right\rangle_{CG}. \end{aligned} \tag{2.10}$$

To compute $\nabla_{\lambda} S_{rel}$ and \mathbf{H} from Eqs (2.9) and (2.10), we need average CG energy derivatives in the AA and CG ensembles. For the averages in the AA ensemble, first a single AA system simulation can be performed and its AA configurations can be saved, then the average CG energy derivatives can be computed by processing the mapped CG configurations of the saved AA configurations using the CG potentials at each iteration. For the averages in the CG ensemble, since the CG ensemble changes with the CG parameters, λ , a short CG simulation can be performed at each iteration to generate corresponding CG configurations. An alternative

approach, which does not require a CG simulation at every iteration, to obtain the CG ensemble averages is to reweight the initial CG configurations obtained from λ^0 [15]. In this work, we implemented the first approach based on performing a short CG simulation at every iteration.

In the case of a CG model, in which CG interactions are modeled by a two-body pair potential, u_{CG} , between CG sites, the ensemble averages of the CG energy derivatives can be computed as

$$\begin{aligned} \left\langle \left(\frac{\partial^a U_{CG}}{\partial \lambda^a} \right)^b \right\rangle_{AA} &= \left\langle \left(\sum_{i<j} \frac{\partial^a u_{CG}(r_{ij})}{\partial \lambda^a} \right)^b \right\rangle_{AA} \\ \left\langle \left(\frac{\partial^a U_{CG}}{\partial \lambda^a} \right)^b \right\rangle_{CG} &= \left\langle \left(\sum_{i<j} \frac{\partial^a u_{CG}(r_{ij})}{\partial \lambda^a} \right)^b \right\rangle_{CG}, \end{aligned} \tag{2.11}$$

where the sum is performed over all the CG site pairs (i, j) , a stands for the 1st, 2nd, ... derivatives and b stands for the different powers, i.e., $b = 1, 2, \dots$

3 Implementation

3.1 Simplex

To be versatile, we have implemented a generic optimization engine embedded in the existing iterative script framework of VOTCA. This optimizer uses a state file to communicate the potential energy parameters for the next MD simulation to be performed. Here, potential energy tables are generated from these parameters which are then used in the simulation. After the simulation is complete, it can be analyzed to evaluate the objective function and again communicate back the next set of parameters to the optimizer via the state file.

The state-based approach described above allows one to easily swap different optimization algorithms. In this work, we have implemented the downhill simplex algorithm (in Perl) and an interface to the existing covariance matrix adaptation (CMA) evolution strategy optimizer [22] (written in Python). CMA is reported to have a better convergence if a large number of parameters are to be optimized.

3.2 Relative entropy

In the VOTCA package, we implemented the relative entropy-based coarse-graining method using the iterative workflow framework described in ref. [16]. Required inputs are the pair distributions of the CG sites in the reference AA ensemble, the initial guess for the CG potential parameters, the CG ensemble simulation set up files, and the option file describing the modeling options for the CG interactions. Furthermore, the user can provide the Newton-Raphson iteration parameters, such as the relaxation parameter, χ , the convergence check criteria, etc.

To describe the CG potentials, we use the uniform cubic B-spline (CBSPL) form given by

$$u_{CBSPL}(r) = [1 \quad t \quad t^2 \quad t^3] \frac{1}{6} \begin{bmatrix} 1 & 4 & 1 & 0 \\ -3 & 0 & 3 & 0 \\ 3 & -6 & 3 & 0 \\ -1 & 3 & -3 & 1 \end{bmatrix} \begin{bmatrix} c_k \\ c_{k+1} \\ c_{k+2} \\ c_{k+3} \end{bmatrix}, \tag{3.1}$$

where $\{c_0, c_1, c_2, \dots, c_m\}$ are the spline knot values tabulated for m evenly spaced intervals of size $\Delta r = r_{cut}/(m - 2)$ along the separation distance $r_i = i \times \Delta r$ with the cut-off r_{cut} , and t is given by

$$t = \frac{r - r_k}{\Delta r}, \tag{3.2}$$

where index k is determined such that $r_k \leq r < r_{k+1}$. We choose CBSPL form because it exhibits remarkable flexibility, and it can represent various complex functional characteristics of pair potentials for sufficiently large number of knots.

To ensure the stability of the relative entropy minimization, some precautionary measures are taken. For the Newton-Raphson update to converge towards a minimum, the Hessian, \mathbf{H} , must be positive definite at each step. With a good initial guess for the CG parameters and by adjusting the value of the relaxation parameter, χ , stability of the Newton-Raphson method can be ensured. One approach to initialize the CG parameters can be to fit them to PMF obtained by inverting the pair distributions of the CG sites obtained from the reference AA ensemble. For the CBSPL form, which is linear in its parameters, the second derivative of S_{rel} is never negative, hence the minimization converges to a single global minimum. However, due to locality property of the CBSPL form, i.e., update to c_i affects only the value of the potential near r_i , and the poor sampling of the very small separation distances in the high repulsive core, the rows of \mathbf{H} corresponding to the first few spline knots in the repulsive core may become zero causing \mathbf{H} to be a singular matrix. To avoid this singularity issue, we specify a minimum separation distance, r_{min} , for each CG pair interaction and remove the spline knots corresponding to the $r \leq r_{\text{min}}$ region from the Newton-Raphson update. Once the remaining knot values are updated, the knot values in the poorly sampled region, i.e., $r \leq r_{\text{min}}$, can be extrapolated. The value of r_{min} can be estimated from the minimum distance at which the reference CG pair distribution is nonzero. Also, to ensure that the CG pair potentials and forces go smoothly to zero near r_{cut} , couple of knot values before and after r_{cut} are fixed to zero.

For the convergence check, we define two types of errors: (i) the CG parameter error, ϵ_λ , given by

$$\epsilon_\lambda^k = \sum_{i=0}^{n-1} (\lambda_i^k - \lambda_i^{k-1})^2, \tag{3.3}$$

where n is the total number of CG parameters to be optimized, k is the index of the iteration step and (ii) the CG potential error, ϵ_u , given by

$$\epsilon_u^k = \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} (u_i^k(r_j) - u_i^{k-1}(r_j))^2, \tag{3.4}$$

where M is the number of CG pair potentials to be optimized, and N is the the number of discrete points used to model CG potentials. Then, the total error, ϵ_{tot} , is defined as

$$\epsilon_{\text{tot}}^k = w_\lambda \epsilon_\lambda^k + w_u \epsilon_u^k, \tag{3.5}$$

where w_λ and w_u are the weights assigned to the CG parameter error and the CG potential error, respectively. Iterations are terminated when ϵ_{tot} is less than the specified tolerance value or the specified limit of maximum number of iterations is reached. Finally, due to stochastic nature of the CG simulations, optimal parameters are computed by evaluating their average over the last few iterations.

3.3 Computational cost

In the simplex and relative entropy algorithms, CG-MD simulations and analysis of the CG ensemble configurations at each iteration step are the main computationally expensive stages. The costs of the CG-MD simulation and analysis of the CG ensemble configurations increase with the size of the system, i.e., number of particles, and the length of the simulation, i.e., number of time steps. To ensure the convergence of the coarse-graining iterations, statistically

reliable CG ensemble configurations are required, which can be obtained by selecting sufficiently large system size and long simulation length. An optimal system size and simulation length to reduce the cost of coarse-graining can be determined by doing few trial iterations.

In VOTCA, to perform coarse-graining in a computationally efficient manner, we use external MD simulation packages, such as GROMACS, LAMMPS, etc. (see Sec. 3.5), which are specifically developed for efficient MD simulations. For efficiently analyzing CG configurations, we implement parallel analysis framework in VOTCA, which is explained in the following subsection 3.4.

3.4 Parallel Analysis

Analysis routines, such as the calculation of radial distribution functions, can be computationally expensive. In fact, post-processing of MD simulation data may take as much time as the actual simulation itself. Usually, the cost of an analysis increases with the number of particles in the system as well as the number of time steps. In the previous serial versions of VOTCA, an analysis is performed in the sequential order on a frame-by-frame basis. In this work, we implement a thread-based parallel framework in order to speed up the post-processing of MD simulation data by making use of the multi-core architecture of modern processors.

The thread-based parallelization is implemented using a straight-forward embarrassingly parallel approach. Post-processing of trajectories is usually performed by applying the same calculations independently at all time steps. Therefore, straight-forward parallelization can be achieved by dividing a trajectory into time intervals, i.e., frames, and distributing these among cores or threads. In contrast, another possibility would be to divide of a single frame into chunks, such as domain-decomposition schemes per frame, followed by the distribution of these chunks to cores. A domain-decomposition per frame based parallelization, although not applicable to all algorithms, would lead to a performance benefit independent of the number of frames to process. For the thread-based parallelization of the VOTCA analysis tools, we use the *POSIX* thread library. A thread-based implementation offers the advantage of shared-memory architectures, where memory can be accessed from all working nodes directly. The evaluation of each frame is now distributed among threads. For this purpose, the distribution or forking of the trajectory among threads, as well as the merging of the data from different threads has to be defined. However, the analysis part of the code remains unchanged as in the serial version. Single frames are then processed individually by each thread. By default, frames are evaluated according to their original order and output also follows the same order. Mutexes are used to impose the correct order of the frames and to handle locking of the input and output routines. A mutex, short for mutual exclusion, is typically used to serialize data access, where the parallel access of concurrent threads would lead to an undefined behavior. A mutex allows only one thread to access a block of code, while other threads have to wait until that first thread finishes this part of the code. The implementation, due to its simplicity, does not require the user to know much about parallel concepts. Most analysis algorithms can be parallelized with little code modifications, such as adjusting forking and merging details.

3.5 Additional Improvements

In addition to the new methods, we added support for coarse-graining bonded interactions by IBI [23], an extension to the IBI process to refine the potential to match the Kirkwood-Buff integrals [19], and iterative refinement of the thermodynamic force [24], which is needed for adaptive resolution simulations [25]. In addition to the technical improvements, we added interfaces to LAMMPS [26], Espresso [27], Espresso++ [28], HOOMD-blue [29] and DLPOLY [30], to complement the existing GROMACS [31] interface, which was updated to support

GROMACS 5.0 [32]. This approach is a major advantage over other coarse-graining packages [33, 34], where an internal, very specialized, MD or MC engine is used for sampling and hence, often has limited simulation capabilities in comparison with the other very general MD codes mentioned above. Minor changes in CG mapper have been implemented to ensure compatibility with the STOCK package [35]. Last but not least, we also added the support for using hybrid methods, like a combination of FM and Boltzmann inversion [17].

4 Results

To illustrate functionality of the downhill simplex and relative entropy-based coarse-graining methods implemented in the VOTCA package, we have coarse-grained two systems: an SPC/E [36] bulk water system and a system consisting of a methanol-water mixture. In previous work, Mashayak and Aluru [37] have already used the relative entropy method in its VOTCA implementation to determine CG potentials for water confined inside of graphene slit channels at different thermodynamic states.

4.1 Coarse-graining of bulk water

In the past, water has already been studied extensively from the point of view of both all-atom and coarse-grained representations [8, 9, 13, 38–41]. CG potentials for bulk water obtained by IBI, IMC, and FM methods using the VOTCA package can be found in ref. [16]. In this work, we coarse-grain SPC/E bulk water at a thermodynamic state of 300 K temperature and 1 bar pressure using the simplex and relative entropy methods. In the coarse-grained model, we represent one water molecule by one CG bead positioned at its center of mass (COM), such that the CG beads solely interact via an isotropic two-body potential. Note, that representing multiple water molecules within one CG bead is also possible [42, 43]. Previously, Shell [12] and Chaimovich and Shell [44] have already employed the relative entropy method to optimize CG potentials for SPC/E water at various thermodynamic states, providing an appropriate reference for comparison. Also, recently Lu et al. [45] derived a series of coarse-grained potentials for various water models, TIP3P, SPC/E, TIP4P-Ew, and TIP4P/2005, using the relative entropy method and systematically compared the ability of these CG potentials to reproduce various structural, dynamic, and thermodynamic properties of water.

For the reference all-atom ensemble, we use the same bulk water all-atom configurations as in the NVT ensemble generated in ref. [16]. The all-atom system consisted of 2180 water molecules in a cubic box of size 4.031 nm. The coarse-grained system consisted of 2180 CG beads which are obtained by mapping the all-atom configuration using a COM mapping scheme.

For the simplex method, we use a modified Lennard-Jones (LJ) potential [46] as the functional form for the CG water interactions:

$$U_{\text{CKD}}(r) = \begin{cases} 4\epsilon \left(\frac{\sigma}{r^{12}} - \frac{\sigma}{r^6} + \frac{1}{4} \right) & : r < r_{\text{c,LJ}} \\ -\epsilon \left(\cos^2 \left(\frac{\pi(r - r_{\text{c,LJ}})}{2w_c} \right) \right) & : r_{\text{c,LJ}} < r < r_{\text{c,LJ}} + w_c \\ 0 & : r_{\text{c,LJ}} + w_c < r \end{cases} \quad (4.1)$$

with an additional Gaussian,

$$U_{\text{CKDg}}(r) = U_{\text{CKD}}(r) + h e^{-\frac{(r-p)^2}{2s^2}} - h e^{-\frac{(r_c-p)^2}{2s^2}}. \quad (4.2)$$

In Eqs (4.1) and (4.2), σ , ϵ and $r_{\text{c,LJ}}$ are the usual LJ parameters, w_c is the smoothness parameter

for the attractive part of the CKD potential, and h , p , s , and r_c are the gaussian's amplitude, mean, width, and cutoff radius, respectively. Hence, there are a total of 6 parameters, i.e., σ , ϵ , w_c , h , p , and s , to optimize. We choose the modified LJ plus Gaussian form because a previous study [16] implied that the water-water CG interaction requires two potential energy minima. Initial parameters are obtained by fitting to the potential from ref. [16] and varying them slightly to obtain 7 sets of initial parameters.

We determine 2 sets of parameters for the CKDg functional form in Eq (4.2). The first set of the CKDg parameters is determined by using the reference AA ensemble water-water COM radial distribution as a target property. The second set of the CKDg parameters is determined by attempting to reproduce the radial distribution as well as the pressure of the AA ensemble. For the simplex algorithm, the penalty value, y , is defined as

$$y = \sum_{r_i=0.0}^{0.6} |g_{CG}(r_i) - g_{AA}(r_i)| + a|p_{CG} - p_{AA}|, \quad (4.3)$$

where $g_{CG}(r_i)$ and $g_{AA}(r_i)$ are the CG and the target AA ensemble RDF, respectively, r_i s are the bin positions used to determine RDFs, p_{CG} and p_{AA} are the CG and the target AA ensemble pressure values, respectively, and a is the tunable parameter used to switch between only the RDF-based and RDF+pressure-based coarse-graining. For the latter we used $a = 0.01/\text{bar}$. In all of the simplex iterations, the 200 ps CG simulations are started from the same initial configuration, but energy minimized before the actual sampling, neglecting the first 20 ps. The minimization is necessary due to the fact that in the case of an expansion move, the initial CG configurations can be far away from any stable structure.

Using the relative entropy method, we optimized the CBSPL functional form, Eq (3.1), for the water-water coarse-grained interaction. For the CBSPL form, we have used a cutoff distance of 0.9 nm, fixed the grid spacing to 0.02 nm, and set $r_{\min} = 0.24$ nm, i.e., only the knot values corresponding to the region, $r > r_{\min}$, are optimized, and the knot values in the poorly sampled region, $r \leq r_{\min}$, are extrapolated. Therefore, there are 48 CG parameters for the water-water CG potential. The initial guess for the CG potential parameters was obtained by a least-square fitting of the CBSPL functional form to the PMF obtained by inverting the water-water COM pair distribution function from the reference AA ensemble. At each iteration, a CG simulation of 200 ps is performed with the GROMACS simulation software. For the first iteration, the initial configuration for the CG simulation is obtained by mapping the last configuration of the reference AA ensemble trajectory. For all subsequent iterations, the final CG configuration from the previous step is used as an initial configuration for the CG simulation. Furthermore, at each iteration, CG configurations corresponding to the first 50 ps are discarded as equilibration stage and only the configurations of the last 150 ps stored at 1 ps intervals are used to compute the update for the CG potential parameters.

Fig 1a shows the optimized CG potentials for bulk water. It can be seen that, although the CG potentials from the relative entropy minimization and simplex are quantitatively different, they exhibit a similar core-softened double-well-type shape, which is a very characteristic of the water-water CG potential [16, 44].

Fig 1b shows the water-water RDFs from the CG simulations along with the reference RDF from the AA simulation. One observes that the relative entropy-based CG potential is able to predict the water-water RDF accurately. The accuracy of the RDF from the relative entropy-based CG potential is consistent with the analysis made in ref. [14] and [15]. In these references, it is demonstrated that, when CG potentials are modeled using a finely tabulated functional form such as CBSPL, the relative entropy minimization would result in the CG potentials similar to IBI and IMC, which reproduce the target AA ensemble pair distributions.

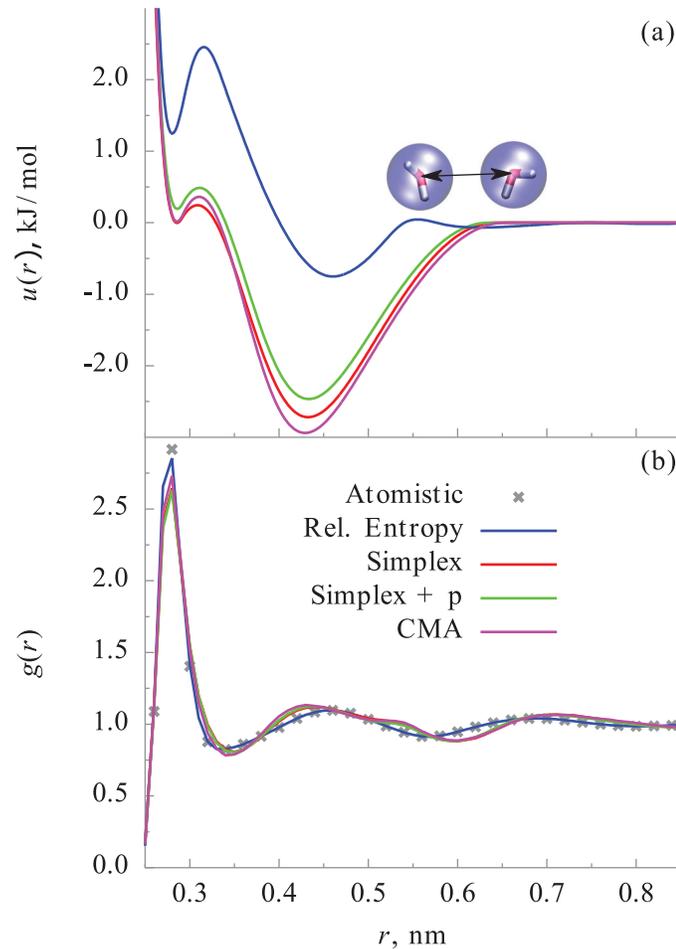


Fig 1. Comparisons of CG potentials (a) and RDFs (b) for bulk water. All methods fit the target RDF sufficiently well and lead to potential of similar shape. The relative entropy method has a slightly smaller error due the fact that more parameter are available in the potential form.

doi:10.1371/journal.pone.0131754.g001

The simplex-based optimization approach is able to fit the radial distribution reasonably well considering the fact that there are only 6 parameters in contrast to the 48 parameters for the CBSPL form. Essentially simplex and CMA lead to the same result, which is not surprising for such a simple system where a local minimum is relatively easy to find. Fitting the RDF and pressure simultaneously leads to an interaction similar in shape to that of RDF-only, but shifted towards lower energies, which is in agreement with earlier findings [9].

4.2 Coarse-graining of water-methanol mixtures

To further test robustness of the simplex and relative entropy methods and their implementation within the VOTCA package, we consider a much more complex system, namely a water-methanol mixture. Methanol is the smallest alcohol and its structural properties feature a winding hydrogen-bonded chains with an average of approximately 2 hydrogen bonds per molecule [47]. Coarse-grained potentials for pure liquid methanol system have been determined by IBI, IMC, and FM methods in ref. [16]. In addition, a water-methanol mixture is a suitable system for studies of several structural aspects of solvation in aqueous mixtures [48].

Table 1. Simulated water-methanol mixtures.

| | I | II | III |
|----------------------------|-------|-------|-------|
| number of H ₂ O | 3752 | 2000 | 248 |
| number of MeOH | 248 | 2000 | 3752 |
| X_m | 0.062 | 0.5 | 0.938 |
| ρ g/cm ³ | 0.97 | 0.885 | 0.80 |

doi:10.1371/journal.pone.0131754.t001

In this study, we considered 3 different water-methanol mixtures with methanol mole fractions, X_m , of 0.062 (diluted), 0.5 (equimolar), and 0.938 (concentrated), similar to the ones used in ref [48]. Reference AA simulations were performed in the NVT ensemble using the GROMACS simulation software [31]. Water is modeled using the SPC/E model, whereas the OPLS [49] force field was used to model methanol, and the LJ interaction parameters, namely C_{12} and C_6 , for the cross-interactions between water and methanol are determined using a geometric mean rule. The number of molecules and average densities of the simulated solutions are given in Table 1. All three mixtures were simulated in a cubic box of length 5.05691 nm with periodic boundary conditions, at 300 K temperature maintained using the Nosé-Hoover thermostat [50]. Systems were equilibrated for 5 ns each, followed by production runs of 20 ns. Reference radial distributions have been computed using snapshots at every 1 ps and a bin size of 0.01 nm.

In the CG model for the water-methanol mixtures, water and methanol molecules are represented by CG beads positioned at their COM and the interactions between CG beads are modeled via isotropic two-body potentials. The number of water and methanol molecules in the CG simulations are the same as in the reference AA simulations (see Table 1).

The simplex setup is very similar to the above mentioned bulk-water setup. We have used the CKDg potential (Eq (4.1)) to model the coarse-grained interactions, yielding a total of 18 parameters to optimize.

For the relative entropy-based coarse-graining, we have used the CBSPL form, Eq (3.1), to model all three CG interactions, i.e., water-water, water-methanol, and methanol-methanol interactions. For the water-water CG potential, a cut-off distance of 1.0 nm was used with the grid spacing of 0.01 nm and $r_{\min} = 0.24$ nm. For the water-methanol and methanol-methanol CG potentials, a cut-off distance of 1.32 nm was used with the grid spacing of 0.02 nm. r_{\min} for the water-methanol and methanol-methanol CG potentials was set to 0.27 and 0.3 nm, respectively. Therefore, there are total of 241 (103 for water-water, 69 for water-methanol and methanol-methanol) CG parameters to be optimized. At each iteration, a CG simulation of 500 ps is performed with the GROMACS and the CG configurations corresponding to the first 100 ps are discarded as equilibration stage and the configurations of the last 400 ps stored at 1 ps intervals are used to compute the update for the CG potential parameters.

The CG potentials obtained from the simplex optimization and the relative entropy minimization for the three different water-methanol mixtures are shown in Fig 2 along with the corresponding RDFs obtained from the CG simulations. Observations about the accuracies of the CG potentials from the simplex and relative entropy methods are similar to that of the bulk water case. Despite the limited flexibility of the CKDg form, the CG potentials from the simplex optimization are able to predict the RDFs reasonably well. As expected, due to the finely tabulated nature of the CBSPL functional form, the CG potentials from the relative entropy optimization are able to predict the water-water, water-methanol, and methanol-methanol RDFs as

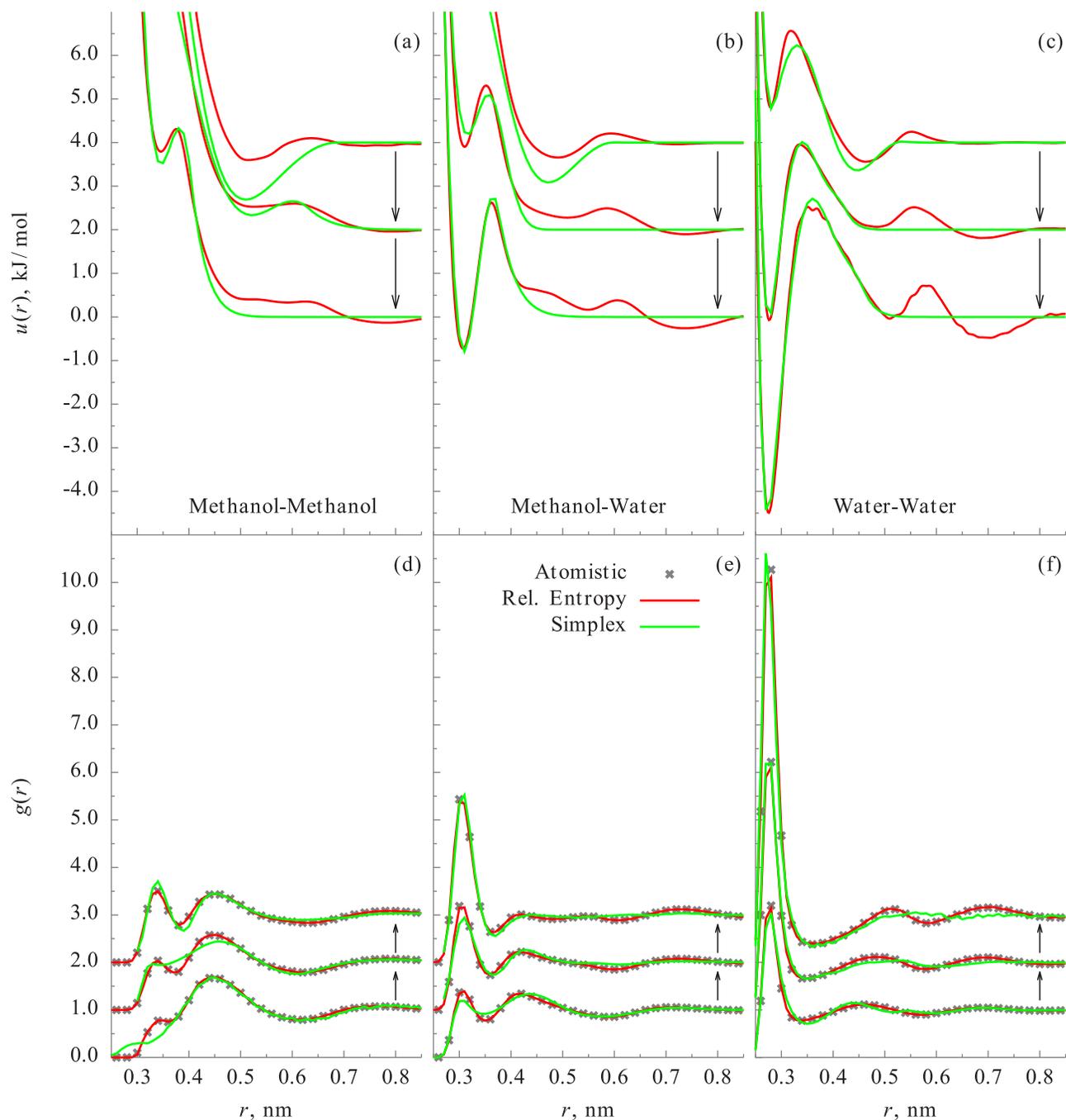


Fig 2. Comparisons of the CG potentials and RDFs of methanol-water: Mixture at different methanol mole fractions, $X = 0.062, 0.5, 0.938$ are shown. Arrow indicates the direction of increasing X . All methods fit the target RDF very well for all mole fractions.

doi:10.1371/journal.pone.0131754.g002

Table 2. Average computational cost per iteration step (in seconds) for simplex and relative entropy-based coarse-graining. For the simplex methods the number in the bracket denote the time spend in energy minimization before the actual molecular dynamics part.

| | Total | CG-MD | CG analysis |
|----------------------------|--------|----------------|-------------|
| Water | | | |
| Simplex | 16.3 | 4.3(+ 0.08) | 11.92 |
| RE | 435 | 124.5 | 310.5 |
| Water-methanol (X = 0.062) | | | |
| Simplex | 2140.2 | 2109.9(+ 5.29) | 25.01 |
| RE | 5076 | 2137.4 | 2938.6 |
| Water-methanol (X = 0.5) | | | |
| Simplex | 1555.2 | 1522.3(+ 0.76) | 32.14 |
| RE | 4176.9 | 1631.8 | 2545.1 |
| Water-methanol (X = 0.938) | | | |
| Simplex | 1165.8 | 1130.2(+ 0.87) | 34.73 |
| RE | 2796.7 | 1151.3 | 1645.4 |

doi:10.1371/journal.pone.0131754.t002

accurately as the reference AA simulations. We note that the CG potentials for the water-methanol mixture system are different for different mole-fractions. This is not surprising, because it is well-known that the CG potentials depend on the thermodynamic state of the reference system [8, 9]. However, it is possible to optimize CG potentials for multiple state-points simultaneously [51], but it is beyond the scope of this work.

4.3 Computational cost comparisons

Table 2 shows the comparison of the average computational cost per iteration step for the simplex and relative entropy-based coarse-graining of the water and water-methanol systems. Please note that, an iteration in VOTCA does not necessary correspond to an iteration in the classical definition of a simplex step. In VOTCA one iteration means running one coarse-grained simulation, while in the simplex method one iteration means one transformation of the polytope which can involve multiple coarse-grained simulations. The computational costs are obtained on a desktop machine with 8-cores Intel Xeon W3520 2.67GHz CPU. As described in the subsection 3.3, the CG-MD simulation and CG configuration analysis are the two major computational parts in the simplex and relative entropy-based coarse-graining. Therefore, in Table 2, we show the distribution of the total computational cost into the CG-MD and CG analysis parts. We observe that, for the systems studied in this work, the simplex is faster than the relative entropy, even though a energy minimization is performed in each simplex step, mainly because the CG analysis step of the relative entropy requires more time than the simplex. The CG analysis step of the relative entropy is more computationally expensive than the simplex due to two main reasons: (i) the number of CG parameters in the relative entropy coarse-graining is more than the simplex and (ii) for the relative entropy update of the CG parameters, in addition to processing CG configurations to evaluate the Hessian and the derivative matrices, we need to solve the system of linear equations, i.e., Eq (2.8), which requires $\mathcal{O}(n^3/3)$ computational cost when solved using the Cholesky decomposition. For the methanol-water systems, the CG-MD simulation lengths in the simplex and relative entropy-based coarse-graining are the same, and hence, the CG-MD cost is similar in both the methods. However, for the the bulk water case, the relative entropy CG-MD requires more

time than the simplex, because the length of the CG-MD in the relative entropy is longer (100000 steps) than the simplex (500 steps). We also note that, for both the simplex and relative entropy methods, the computational cost of coarse-graining water-methanol systems is higher than the bulk water because of the higher number of CG parameters, the larger CG-MD system size (see [Table 1](#)), and longer CG-MD simulations at each iteration step.

4.4 Parallel Analysis

We demonstrate the performance of the parallel analysis implementation by calculating the RDF of a Lennard-Jones [52] fluid. Furthermore, we highlight a crucial part of many analysis algorithms: constructing neighbor lists, i.e., the search for the neighbors of a particle within a cut-off distance. Neighbor lists are often required in internal loops, such as the computation of the distance between two particles within a sphere. Depending on the analysis algorithm and system size, the construction of neighbor lists may dominate the total computational cost. We compare two approaches for constructing neighbor lists- (i) a naive approach of the simple search and (ii) the more efficient implementation based on the grid search. The performance of the analysis is measured as a function of the system size and neighbor search cut-off.

The simple search algorithm for constructing neighbor lists is as following. For each particle in a system, loop over all other particles and check the distance between the two particles. If the distance is less than the provided cut-off, tag this pair as neighbors. Therefore, the simple search algorithm requires to loop through all particles twice, and hence, its expected computational cost is $\mathcal{O}(N^2)$, where N is the number of particles.

In contrast, the grid search algorithm loops over all particles once and distributes the particles among a grid based on their positions. The size of the cells in the grid depends on the cut-off. A neighbor lookup will then lead to a constant cost of checking all surrounding cells. The grid search algorithm is expected to have, besides a large prefactor, a cost of $\mathcal{O}(N)$. The performance of the grid search decreases with an increasing cut-off, since each cell in the grid will hold more particles, whereas the cost of the simple search does not depend on the cut-off.

VOTCA supports both the neighbor list algorithms, and the user is given a choice to select one of them.

For the performance analysis, we study three different LJ systems with different sizes: the small system with 5324, the medium with 17687, and the big with 60132 fluid particles. The system temperature and density are set to $T^* = 0.73$ and $\rho^* \approx 0.9$ in the reduced LJ units, respectively.

First, we compare the performance of the simple search and grid search algorithms. For comparison, we measure the computational time for constructing neighbor-list for 1 frame on 1 core only. In [Fig 3](#), the timing results for the neighbor-list creation as a function of the number of particles are shown. The computational times of the simple and grid search algorithms follow the expected scaling of $\mathcal{O}(N^2)$ and $\mathcal{O}(N)$, respectively. The grid search algorithm shows a very good agreement with our expectation. However, the simple search algorithm performs worse than expected for the systems with 262636 and 492038 particles. The performance loss of the simple search algorithm is likely due to the large amount of data the algorithm has to process in the inner loops. For large data, cache memory limits may exceed causing bad scaling. An additional benchmark, not reported here, confirmed the dependence of the grid search algorithm on the neighbor cut-off distance, whereas the simple search is unaffected. However, even after increasing the cut-off from 1.6 to 2.6σ , it is found that the grid search performs an order of magnitude faster than the simple search. The cross-over, where the simple search becomes faster than the grid search depends on the cut-off distance, but is typically too large for productive uses, and hence, the grid search should be preferred over the simple search in general.

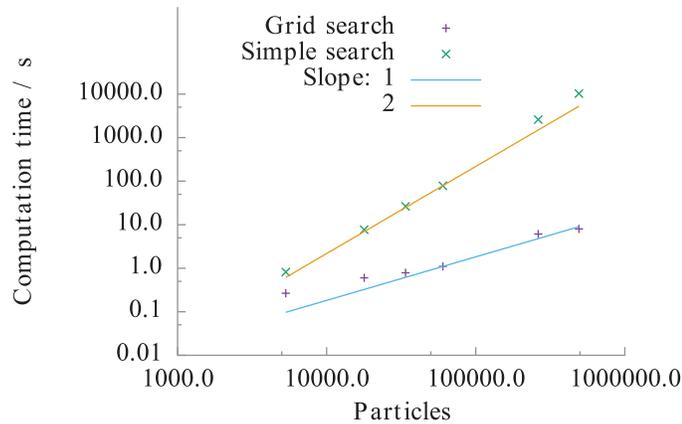


Fig 3. Computation times for the construction of neighbor lists: simple vs. grid search. Shown are the results for the Lennard-Jones fluid as a function of particles in the system. In addition to the data points for the simple and grid search algorithms, lines indicate the scaling law with 2 and 1 as the exponent, respectively. These exponents result from the cost of the simple and grid search algorithm: $\mathcal{O}(N^2)$ and $\mathcal{O}(N)$, respectively. The cut-off for the neighbor search was set to 1.6σ , which roughly corresponds to the first minimum in the radial distribution function.

doi:10.1371/journal.pone.0131754.g003

Next, we study the overall performance of the parallel analysis framework. Ideally, we expect the speed-up of the parallel analysis, i.e., the ratio of the time required for the serial analysis to the time for parallel analysis, to match the number of threads used. Due to the use of shared-memory architecture, communication between threads is not necessary. Hence, the overall performance becomes more efficient with increasing intensity ratio, i.e., computational cost to input-output (IO) cost. However, the performance may be negatively affected due to the overhead of threads, such as the additional definition and initialization needed for the single thread and the handling of mutexes, as well as possible interferences while accessing the shared memory from multiple threads. Some of the mutex-overhead can be saved by explicitly neglecting the original ordering of frames. Both negative effects increase with the number of threads and reduce the overall speed-up.

Fig 4 summarizes the computational times for the calculation of $g(r)$ as a function of the number of threads. It is observed that increasing the number of threads always decreases the total computation time, however, the timings deviate from the ideal scaling line with increasing thread count. Up to 4 threads, the computation time is closer to the ideal scaling, i.e., the parallel speed-up is around 3–4. For 6, 8, and 12 threads, however, the total computation time deviates from the ideal scaling line. The non-ideal time scaling behavior is mainly caused by the constant overhead of the VOTCA input routine and a missing cache optimization. The raw numbers of the computation time for all 960 frames on 1 core are 20s, 68s, and 243s for the small, medium and big system, respectively. For comparison, the GROMACS tool, `g_rdf`, which is based on simple search, takes 740s, 8116s and 82728s to compute $g(r)$ for the small, medium and big system, respectively. In the former versions of VOTCA, the parallelization of `g_rdf` was script-based and handled by `multi_g_rdf`. The trajectory was explicitly split in time into chunks and multiple instances of `g_rdf` were called. The results for the small and medium systems are also shown in Fig 4.

As stated earlier, the overall performance gain is improved if computationally intensive algorithms are used. In other words, when using threads, a computationally expensive algorithm leads to a better scaling behavior. The constant overhead cost becomes more negligible

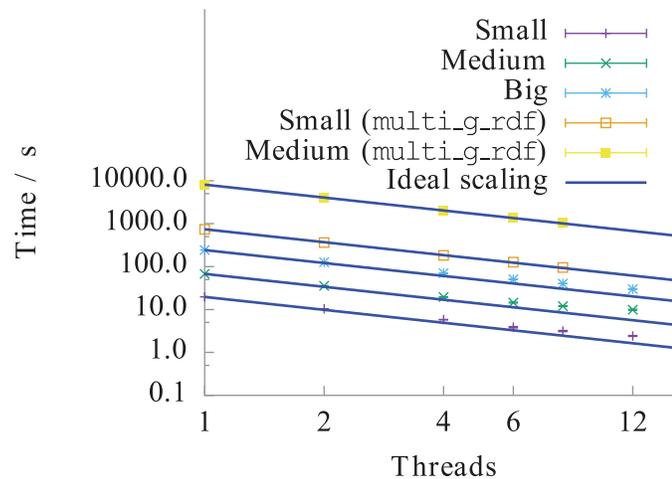


Fig 4. Absolute computation time for the radial distribution function calculation as a function of threads. The small system holds 5324, the medium 17687 and the big system 60132 particles. The dashed line shows the ideal scaling line. Also shown are the results from the script-based parallelization of `multi_g_rdf`.

doi:10.1371/journal.pone.0131754.g004

and cache-misses are less likely to happen. For example, the algorithm that spends most of the CPU time in the inner loops of the evaluation, will perform much better than the lightweight calculation of a RDF. Our tests showed almost-perfect scaling up to 8 threads and only 20% loss of the performance at 12 threads for the algorithm which makes a heavy use of intrinsic CPU commands.

5 Conclusions

In this work, we have extended the versatility of the VOTCA package by implementing two recent coarse-graining methods, namely, the downhill simplex optimization and the relative entropy minimization. We have also demonstrated the applicability of the newly implemented coarse-graining techniques in VOTCA by coarse-graining a bulk water system and water-methanol mixture. We have found that for both the systems, the CG pair potentials described by the CBSPL functional form and optimized by minimizing the relative entropy reproduce the target pair distributions. This result validates the known characteristic of the relative entropy minimization, i.e., when the CG pair potentials are modeled using finely tabulated functional form then at relative entropy minimum the CG pair potentials reproduce the pair distributions of the target ensemble. We have also found that the simplex-based optimization method is effective in minimizing the error between the properties from the CG and the target AA simulations. For the bulk water and water-methanol mixtures, we have shown that the simplex method can effectively optimize the CG potentials modeled using the CKDg functional form with just 6 parameters, such that the target pair distributions and pressure values are predicted reasonably well by CG simulations. The accuracy of the CG potentials optimized by the simplex method can be improved if a more complex functional form with additional parameters is used to model the CG potentials. However, for a large number of parameters, the simplex method becomes too computationally inefficient. Input for these simulation will be made available as tutorials as part of the upcoming VOTCA 1.3. Finally, in addition to the new coarse-graining techniques, we have also improved the computational efficiency of the VOTCA

package by implementing the parallel analysis framework and added support for more MD sampling engines.

Acknowledgments

We would like to thank K. Kremer for bringing the authors of this paper together at the Max-Planck-Institute for Polymer Research in Mainz and for his continuous support of this project over the years. SYM and NRA acknowledge financial support by the NSF under grant Nos. 1264282 and 1420882 and AFOSR under grant No. 11157642. CJ was financially supported by SFB 625 in the framework of the multiscale modeling initiative of the Max-Planck Society (M³) during his time in Mainz in 2011 and acknowledge hospitality at KITP in July 2012, where this research was supported in part by the National Science Foundation under grant number NSF PHY11-25915. CJ thanks LANL for a Director's fellowship and the MPG for hospitality in 2013 at MPI-P. SYM, CJ, and VR acknowledge Shell S. for helpful suggestions during the implementation of the relative entropy method. K. Koschke acknowledges funding by the Nestlé Research Center. MNJ acknowledges C. Peter for fruitful scientific discussions and the DFG Emmy Noether program for financial support. Assigned: LA-UR-13-26878. LANL is operated by Los Alamos National Security, LLC, for the National Nuclear Security Administration of the U.S. DOE under Contract DE-AC52-06NA25396.

Author Contributions

Conceived and designed the experiments: SYM CJ. Performed the experiments: SYM CJ KK. Analyzed the data: SYM CJ KK. Contributed reagents/materials/analysis tools: VR SYM NRA CJ MNJ KK. Wrote the paper: VR SYM CJ MNJ KK.

References

1. Peter C, Kremer K. Multiscale simulation of soft matter systems from the atomistic to the coarse-grained level and back. *Soft Matter*. 2009; 5:4357–4366. doi: [10.1039/b912027k](https://doi.org/10.1039/b912027k)
2. Praprotnik M, Delle Site L, Kremer K. Multiscale Simulation of Soft Matter: From Scale Bridging to Adaptive Resolution. *Annual Review of Physical Chemistry*. 2008; 59:545–571. doi: [10.1146/annurev.physchem.59.032607.093707](https://doi.org/10.1146/annurev.physchem.59.032607.093707) PMID: [18062769](https://pubmed.ncbi.nlm.nih.gov/18062769/)
3. Müller-Plathe F. Coarse-graining in polymer simulation: from the atomistic to the mesoscopic scale and back. *Chemphyschem*. 2002; 3:755–769. PMID: [12436902](https://pubmed.ncbi.nlm.nih.gov/12436902/)
4. Voth GA. *Coarse-Graining of Condensed Phase and Biomolecular Systems*. CRC Press; 2008.
5. Noid WG. Perspective: Coarse-grained models for biomolecular systems. *The Journal of Chemical Physics*. 2013; 139:090901. doi: [10.1063/1.4818908](https://doi.org/10.1063/1.4818908) PMID: [24028092](https://pubmed.ncbi.nlm.nih.gov/24028092/)
6. Hyeon C, Thirumalai D. Capturing the essence of folding and functions of biomolecules using coarse-grained models. *Nature Communications*. 2011; 2:487. doi: [10.1038/ncomms1481](https://doi.org/10.1038/ncomms1481) PMID: [21952221](https://pubmed.ncbi.nlm.nih.gov/21952221/)
7. Fritz D, Koschke K, Harmandaris VA, van der Vegt NFA, Kremer K. Multiscale modeling of soft matter: scaling of dynamics. *Physical Chemistry Chemical Physics*. 2011; 13:10412–10420. doi: [10.1039/c1cp20247b](https://doi.org/10.1039/c1cp20247b) PMID: [21468407](https://pubmed.ncbi.nlm.nih.gov/21468407/)
8. Johnson ME, Head-Gordon T, Louis AA. Representability problems for coarse-grained water potentials. *The Journal of Chemical Physics*. 2007; 126:144509. doi: [10.1063/1.2715953](https://doi.org/10.1063/1.2715953) PMID: [17444725](https://pubmed.ncbi.nlm.nih.gov/17444725/)
9. Wang H, Junghans C, Kremer K. Comparative atomistic and coarse-grained study of water: What do we lose by coarse-graining? *Eur Phys J E*. 2009; 28:221–229. doi: [10.1140/epje/i2008-10413-5](https://doi.org/10.1140/epje/i2008-10413-5) PMID: [19198906](https://pubmed.ncbi.nlm.nih.gov/19198906/)
10. Reith D, Pütz M, Müller-Plathe F. Deriving effective mesoscale potentials from atomistic simulations. *Journal of Computational Chemistry*. 2003; 24:1624–1636. doi: [10.1002/jcc.10307](https://doi.org/10.1002/jcc.10307) PMID: [12926006](https://pubmed.ncbi.nlm.nih.gov/12926006/)
11. Lyubartsev AP, Laaksonen A. Calculation of effective interaction potentials from radial distribution functions: A reverse Monte Carlo approach. *Phys Rev E*. 1995; 52:3730–3737. doi: [10.1103/PhysRevE.52.3730](https://doi.org/10.1103/PhysRevE.52.3730)
12. Shell MS. The relative entropy is fundamental to multiscale and inverse thermodynamic problems. *The Journal of chemical physics*. 2008; 129:144108. doi: [10.1063/1.2992060](https://doi.org/10.1063/1.2992060) PMID: [19045135](https://pubmed.ncbi.nlm.nih.gov/19045135/)

13. Izvekov S, Voth GA. A Multiscale Coarse-Graining Method for Biomolecular Systems. *J Phys Chem B*. 2005; 109:2469–2473. doi: [10.1021/jp044629q](https://doi.org/10.1021/jp044629q) PMID: [16851243](https://pubmed.ncbi.nlm.nih.gov/16851243/)
14. Rudzinski JF, Noid WG. Coarse-graining entropy, forces, and structures. *The Journal of Chemical Physics*. 2011; 135:214101. doi: [10.1063/1.3663709](https://doi.org/10.1063/1.3663709) PMID: [22149773](https://pubmed.ncbi.nlm.nih.gov/22149773/)
15. Chaimovich A, Shell MS. Coarse-graining errors and numerical optimization using a relative entropy framework. *The Journal of Chemical Physics*. 2011; 134:094112. doi: [10.1063/1.3557038](https://doi.org/10.1063/1.3557038) PMID: [21384955](https://pubmed.ncbi.nlm.nih.gov/21384955/)
16. Rühle V, Junghans C, Lukyanov A, Kremer K, Andrienko D. Versatile object-oriented toolkit for coarse-graining applications. *Journal of Chemical Theory and Computation*. 2009; 5:3211–3223. doi: [10.1021/ct900369w](https://doi.org/10.1021/ct900369w)
17. Rühle V, Junghans C. Hybrid Approaches to Coarse-Graining using the VOTCA Package: Liquid Hexane. *Macromolecular Theory and Simulations*. 2011; 20:472–477. doi: [10.1002/mats.201100011](https://doi.org/10.1002/mats.201100011)
18. Nelder JA, Mead R. A simplex method for function minimization. *Computer Journal*. 1965; 7:308–313. doi: [10.1093/comjnl/7.4.308](https://doi.org/10.1093/comjnl/7.4.308)
19. Ganguly P, Mukherji D, Junghans C, van der Vegt NFA. Kirkwood-Buff coarse-grained force fields for aqueous solutions. *Journal of Chemical Theory and Computation*. 2012; 8:1802–1807. doi: [10.1021/ct3000958](https://doi.org/10.1021/ct3000958)
20. Wu D, Kofke DA. Phase-space overlap measures. I. Fail-safe bias detection in free energies calculated by molecular simulation. *The Journal of chemical physics*. 2005; 123:054103. doi: [10.1063/1.1992483](https://doi.org/10.1063/1.1992483) PMID: [16108627](https://pubmed.ncbi.nlm.nih.gov/16108627/)
21. Lyubartsev A, Mirzoev A, Chen L, Laaksonen A. Systematic coarse-graining of molecular models by the Newton inversion method. *Faraday discussions*. 2010; 144:43–56. doi: [10.1039/B901511F](https://doi.org/10.1039/B901511F) PMID: [20158022](https://pubmed.ncbi.nlm.nih.gov/20158022/)
22. Hansen N, Ostermeier A, Gawelczyk A. On the Adaptation of Arbitrary Normal Mutation Distributions in Evolution Strategies: The Generating Set Adaptation. In: *ICGA*. Citeseer; 1995. p. 57–64.
23. Bezkorovaynaya O, Lukyanov A, Kremer K, Peter C. Multiscale simulation of small peptides: Consistent conformational sampling in atomistic and coarse-grained models. *Journal of Computational Chemistry*. 2012; 33:937–949. doi: [10.1002/jcc.22915](https://doi.org/10.1002/jcc.22915) PMID: [22298285](https://pubmed.ncbi.nlm.nih.gov/22298285/)
24. Pobleto S, Praprotnik M, Kremer K, Delle Site L. Coupling different levels of resolution in molecular simulations. *J Chem Phys*. 2010; 132:114101. doi: [10.1063/1.3357982](https://doi.org/10.1063/1.3357982) PMID: [20331275](https://pubmed.ncbi.nlm.nih.gov/20331275/)
25. Fritsch S, Junghans C, Kremer K. Structure formation of toluene around C60: Implementation of the Adaptive Resolution Scheme (AdResS) into GROMACS. *Journal of Chemical Theory and Computation*. 2012; 8:398. doi: [10.1021/ct200706f](https://doi.org/10.1021/ct200706f)
26. Plimpton S. Fast parallel algorithms for short-range molecular dynamics. *Journal of Computational Physics*. 1995; 117:1–19. doi: [10.1006/jcph.1995.1039](https://doi.org/10.1006/jcph.1995.1039)
27. Limbach HJ, Arnold A, Mann BA, Holm C. ESPResSo—an extensible simulation package for research on soft matter systems. *Computer Physics Communications*. 2006; 174:704–727. doi: [10.1016/j.cpc.2005.10.005](https://doi.org/10.1016/j.cpc.2005.10.005)
28. Halverson JD, Brandes T, Lenz O, Arnold A, Bevc S, Starchenko V, et al. ESPResSo++: A modern multiscale simulation package for soft matter systems. *Computer Physics Communications*. 2012;.
29. Anderson JA, Lorenz CD, Travesset A. General purpose molecular dynamics simulations fully implemented on graphics processing units. *Journal of Computational Physics*. 2008; 227:5342–5359. doi: [10.1016/j.jcp.2008.01.047](https://doi.org/10.1016/j.jcp.2008.01.047)
30. Smith W, Todorov IT. A short description of DL_POLY. *Molecular Simulation*. 2006; 32:935–943. doi: [10.1080/08927020600939830](https://doi.org/10.1080/08927020600939830)
31. Hess B, Kutzner C, van der Spoel D, Lindahl E. GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *Journal of Chemical Theory and Computation*. 2008; 4:435–447. doi: [10.1021/ct700301q](https://doi.org/10.1021/ct700301q)
32. Apol E, Apostolov R, Berendsen H, van Buuren A, Bjelkmar P, van Drunen R, et al. *Gromacs User Manual Version 5.0*; 2014.
33. Karimi-Varzaneh HA, Qian HJ, Chen X, Carbone P, Müller-Plathe F. IBIsCO: A molecular dynamics simulation package for coarse-grained simulation. *Journal of Computational Chemistry*. 2011; 32:1475–1487. doi: [10.1002/jcc.21717](https://doi.org/10.1002/jcc.21717) PMID: [21425295](https://pubmed.ncbi.nlm.nih.gov/21425295/)
34. Mirzoev A, Lyubartsev AP. MagiC: Software Package for Multiscale Modeling. *Journal of Chemical Theory and Computation*. 2013; 9:1512–1520. doi: [10.1021/ct301019v](https://doi.org/10.1021/ct301019v)
35. Bevc S, Junghans C, Praprotnik M. STOCK: Structure mapper and online coarse-graining kit for molecular simulations. *Journal of computational chemistry*. 2015; 36:467. doi: [10.1002/jcc.23806](https://doi.org/10.1002/jcc.23806) PMID: [25504076](https://pubmed.ncbi.nlm.nih.gov/25504076/)

36. Berendsen HJC, Grigera JR, Straatsma TP. The missing term in effective pair potentials. *Journal of Physical Chemistry*. 1987; 91:6269–6271. doi: [10.1021/j100308a038](https://doi.org/10.1021/j100308a038)
37. Mashayak SY, Aluru NR. Thermodynamic state-dependent structure-based coarse-graining of confined water. *The Journal of Chemical Physics*. 2012; 137:214707. doi: [10.1063/1.4769297](https://doi.org/10.1063/1.4769297) PMID: [23231256](https://pubmed.ncbi.nlm.nih.gov/23231256/)
38. Nezbeda I, Slovak J. A family of primitive models of water: three-, four and five-site models. *Molecular Physics*. 1997; 90:353–372. doi: [10.1080/002689797172471](https://doi.org/10.1080/002689797172471)
39. Wallqvist A, Mountain RD. 4. In: *Molecular Models of Water: Derivation and Description*. John Wiley and Sons, Inc.; 2007. p. 183–247.
40. Molinero V, Moore EB. Water Modeled As an Intermediate Element between Carbon and Silicon? *The Journal of Physical Chemistry B*. 2009; 113:4008–4016. doi: [10.1021/jp805227c](https://doi.org/10.1021/jp805227c) PMID: [18956896](https://pubmed.ncbi.nlm.nih.gov/18956896/)
41. Hadley KR, McCabe C. Coarse-Grained Molecular Models of Water: A Review. *Molecular simulation*. 2012; 38:671–681. doi: [10.1080/08927022.2012.671942](https://doi.org/10.1080/08927022.2012.671942) PMID: [22904601](https://pubmed.ncbi.nlm.nih.gov/22904601/)
42. Marrink SJ, Risselada HJ, Yefimov S, Tieleman DP, de Vries AH. The MARTINI force field: coarse grained model for biomolecular simulations. *The Journal of Physical Chemistry B*. 2007; 111:7812–7824. doi: [10.1021/jp071097f](https://doi.org/10.1021/jp071097f) PMID: [17569554](https://pubmed.ncbi.nlm.nih.gov/17569554/)
43. Nagarajan A, Junghans C, Matysiak S. Multiscale simulation of liquid water using a four-to-one mapping for coarse-graining. *Journal of Chemical Theory and Computation*. 2013; 9:5168. doi: [10.1021/ct400566j](https://doi.org/10.1021/ct400566j)
44. Chaimovich A, Shell MS. Anomalous waterlike behavior in spherically-symmetric water models optimized with the relative entropy. *Physical Chemistry Chemical Physics*. 2009; 11:1901–1915. doi: [10.1039/b818512c](https://doi.org/10.1039/b818512c) PMID: [19280001](https://pubmed.ncbi.nlm.nih.gov/19280001/)
45. Lu J, Qiu Y, Baron R, Molinero V. Coarse-Graining of TIP4P/2005, TIP4P-Ew, SPC/E, and TIP3P to Monatomic Anisotropic Water Models Using Relative Entropy Minimization. *Journal of Chemical Theory and Computation*. 2014; 10:4104–4120. doi: [10.1021/ct500487h](https://doi.org/10.1021/ct500487h)
46. Cooke IR, Kremer K, Deserno M. Tunable generic model for fluid bilayer membranes. *Physical Review E*. 2005; 72:011506. doi: [10.1103/PhysRevE.72.011506](https://doi.org/10.1103/PhysRevE.72.011506)
47. Jorgensen WL. Optimized intermolecular potential functions for liquid alcohols. *Journal of Physical Chemistry*. 1986; 90:1276–1284. doi: [10.1021/j100398a015](https://doi.org/10.1021/j100398a015)
48. Laaksonen A, Kusalik P. Three-dimensional structure in water-methanol mixtures. *The Journal of Physical Chemistry A*. 1997; 101:5910–5918. doi: [10.1021/jp970673c](https://doi.org/10.1021/jp970673c)
49. Jorgensen WL, Tirado-Rives J. Potential energy functions for atomic-level simulations of water and organic and biomolecular systems. *Proceedings of the National Academy of Sciences of the United States of America*. 2005; 102:6665–6670. doi: [10.1073/pnas.0408037102](https://doi.org/10.1073/pnas.0408037102) PMID: [15870211](https://pubmed.ncbi.nlm.nih.gov/15870211/)
50. Nosé S. A unified formulation of the constant temperature molecular dynamics methods. *The Journal of Chemical Physics*. 1984; 81:511–519. doi: [10.1063/1.447334](https://doi.org/10.1063/1.447334)
51. Mullinax JW, Noid WG. Extended ensemble approach for deriving transferable coarse-grained potentials. *The Journal of Chemical Physics*. 2009; 131:104110. doi: [10.1063/1.3220627](https://doi.org/10.1063/1.3220627)
52. Jones JE. On the Determination of Molecular Fields. II. From the Equation of State of a Gas. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*. 1924; 106:463–477. doi: [10.1098/rspa.1924.0081](https://doi.org/10.1098/rspa.1924.0081)