THE JOURNAL OF PHYSICAL CHEMISTRY B LETTERS

2005, *109,* 2469−2473
Published on Web 01/22/2005

A Multiscale Coarse-Graining Method for Biomolecular Systems

Sergei Izvekov and Gregory A. Voth*
Department of Chemistry and Center for Biophysical Modeling and Simulation, University of Utah, 315 South 1400 E., Rm. 2020, Salt Lake City, Utah 84112-0850

Received: November 24, 2004; In Final Form: December 29, 2004

A new approach is presented for obtaining coarse-grained (CG) force fields from fully atomistic molecular dynamics (MD) trajectories. The method is demonstrated by applying it to derive a CG model for the dimyristoylphosphatidylcholine (DMPC) lipid bilayer. The coarse-graining of the interparticle force field is accomplished by an application of a force-matching procedure to the force data obtained from an explicit atomistic MD simulation of the biomolecular system of interest. Hence, the method is termed a "multiscale" CG (MS−CG) approach in which explicit atomistic-level forces are propagated upward in scale to the coarse-grained level. The CG sites in the lipid bilayer application were associated with the centers-of-mass of atomic groups because of the simplicity in the evaluation of the forces acting on them from the atomistic data. The resulting CG lipid bilayer model is shown to accurately reproduce the structural properties of the phospholipid bilayer.

The development of accurate, reliable, and computationally efficient force fields is a central focus of molecular modeling. Computational efficiency is one of the dominant considerations in choosing particular models for interactions in studies of complex (e.g., biomolecular) systems. Generally speaking, models based on effective pairwise forces are considered to be the most computationally inexpensive. Unfortunately, even simple pairwise atom−atom potentials are still too computationally costly for many applications of practical interest. The need to reach larger simulated time scales and system sizes requires a further simplification of atomistic models, even those based on empirical potentials. Several so-called "coarse-grained" (CG) approaches have been recently developed with such goal in mind.[1−10] The philosophy of these CG approaches is generally the same: to achieve a simpler description of the effective interactions in a given system while not losing the ability of the resulting models to predict the properties of interest.

One way of achieving a CG mapping of a system into a structurally less detailed level is, for example, by grouping the atoms into fewer interaction sites.[1−8] Technically, the implementation of such coarse-graining models can be divided into two distinct phases. The first is a partitioning of the system into the larger structural units to reduce the system complexity, while the second is the construction of an effective force field to describe the interactions between the CG structural units. The second stage, which is the primary focus of the present work, represents a major difficulty in the application of CG approaches. There has generally been no systematic strategy for the development of CG potentials. Typically, CG potentials of a pre-selected analytical form are parameterized either to reproduce average structural properties seen in atomistic simulations, for example using an iterative adjustment of potential parameters starting from an approximation based on potentials of mean force[5,9] or the inverse Monte Carlo technique,[10] or they are parameterized to match thermodynamic properties.[5,8] These approaches are not directly based on the underlying atomistic-scale forces.

The CG approach may be further complicated due to the fact that CG potentials are expected to have less transferability compared to atomistic ones. This is because an effective interaction between the structural units intended for coarse-graining is defined by the average structure (e.g., average orientations, distances) within the complexes formed by those units in a particular phase. Rigorously speaking, the CG potential is a potential of mean force (PMF), which is a configurational *free energy* in a reduced phase space. Therefore, the structural properties and thus the CG potentials even for the same phase can in principle be sensitive to variations in temperature and other thermodynamic conditions. A poor transferability of a CG potential will undermine the reliability of simulations in which the same CG potential was used to simulate different system phases.[6] Ideally, a CG potential should be fit (or refit) to a system under the same thermodynamic conditions at which the system is intended to be simulated. This fact makes the availability of computationally efficient methods for the development of CG potentials critical. In this communication, we present a significantly different approach for developing CG potentials from an underlying explicit atomistic molecular dynamics (MD) simulation. Our approach is called "multiscale coarse-graining" (MS−CG) because the CG potential parameters are systematically derived from atomistic-level interactions.

The basis of the MS−CG approach is a new method for force matching[11,12] which has been used to build effective empirical force fields for complex molecular condensed-phase systems. This force-match (FM) method is an extension of the least-squares force-match approach originally suggested by Ercolessi and Adams.[13] However, our FM method can determine a pairwise effective force field from given trajectory and force data regardless of their origin (e.g., it may be data obtained

10.1021/jp044629q CCC: $30.25

from ab initio MD simulation,[11] path-integral MD simulation,[12] or from coarse-graining of atomistic data as in the present work). The use of the FM in the MS−CG approach is, however, different from the previous applications in that it is used here to define a PMF between CG sites from an underlying atomistic model.

In the conventional FM method, the classical forces $\mathbf{F}_i^{\mathrm{p}}(g_1, ..., g_M)$ of a preselected analytical form, which are dependent on the set of M parameters $g_1, ..., g_M$, are optimized by minimizing the average over the whole configuration data functional $|\mathbf{F}_i^{\mathrm{ref}} - \mathbf{F}_i^{\mathrm{p}}(g_1, ..., g_M)|^2$, where $\mathbf{F}_i^{\mathrm{ref}}$ is the reference force (e.g., supplied by ab initio simulation). The fitting rapidly becomes intractable as the number of parameters grows. For biological molecules, for example, the problem becomes extreme because of the large variety of atomic species found in the system. However, if the force field to be fit depends linearly on the fitting parameters, such as those that can be often achieved using a suitable (e.g., spline) interpolation, the least-squares problem can be written in the form of an overdetermined system of linear equations. The least-squares solution of such system can be then efficiently found using, for example, orthogonal matrix triangularization (QR decomposition).[14]

A detailed description of the algorithmic development of our FM method is given in ref 11. In essence this new fitting approach involves a solution of a following system of equations

$$\sum_{\beta=1}^{K}\sum_{j=1}^{N_\beta}\left(-f(r_{\alpha il,\beta jl},\{r_{\alpha\beta,\kappa}\},\{f_{\alpha\beta,\kappa}\},\{f''_{\alpha\beta,\kappa}\}) - \frac{q_{\alpha\beta}}{r^2_{\alpha il,\beta jl}}\right)\mathbf{n}_{\alpha il,\beta jl} = \mathbf{F}^{\mathrm{ref}}_{\alpha il}$$

with respect to $\{f_{\alpha\beta,\kappa}, f''_{\alpha\beta,\kappa}, q_{\alpha\beta}\}$ parameters, where $\alpha = 1, ..., K$, $i = 1, ..., N_\alpha$, $l = 1, ..., L$. The first terms in the sum are a spline representation (which is a linear form of the f, f'' parameters) of the short-ranged part of the force and the second term is the Coulomb part. The subscript $\{\alpha il\}$ labels the ith atom of kind α in the lth atomic configuration; $q_{\alpha\beta}$ is a product of partial charges q_α, q_β; N_β, K are, respectively, the number of atoms of each kind β and the total number of atomic types in the system; and L is the number of distinct atomic configurations selected along the trajectories. The latter number should be large enough to make eq 1 overdetermined. The spline mesh $\{r_{\alpha\beta,\kappa}\}$ can be different for different pairs $\{\alpha\beta\}$ of atomic kinds. Standard equations which are linear with respect to $\{f_{\alpha\beta,\kappa}, f''_{\alpha\beta,\kappa}\}$ must be included into eq 1 to ensure that the derivative $f'(r)$ is continuous across the boundary between two intervals.[15]

The linear dependence of the force field on the fitting parameters allows one to carry out the fit in a manner which is more consistent with the determination of the mean CG force field (PMF). In this case, eq 1 has to be solved for different smaller sets of atomic configurations (sampled along short pieces of system MD trajectories) and then the solutions so obtained are averaged over a large number of such sets. Typically the number of atomic configurations used to build each set of eq 1 is sought to be the smallest possible to ensure that the equations overdetermine the force parameters. The process of averaging is one of the major differences between our method and previous FM approaches, in which the forces were fit to whole trajectory configuration and force data (with the block averaging being absent).

To obtain trajectory and force data for a CG image of the system, the trajectories and forces output from an atomistic MD simulation must be transformed accordingly. It is convenient to associate CG sites with the center-of-mass (CM) of atomic groups because the force acting on the CM of an atomic group can be straightforwardly evaluated from the atomistic MD data. The FM procedure then applied to these data yields the effective interaction between the CG sites *as it is present in the underlying atomistic simulation*. However, the CG site need not necessarily be placed onto the CM of the atomic groups, but instead it may be the geometrical center (i.e., the CM of the system assuming that all atoms are of the same mass). The geometrical center (GC) as a place for a CG site might be a better choice for atomic groups with a very uneven distribution of atomic masses so the CM is close to heavy atoms of the group (like water which have heavy oxygen and light hydrogens). One can also mix CM and GC coarse-graining sites in the present methodology. The feature that makes our approach markedly different from all previous methods for the development of CG force fields is that it provides an explicit multiscale bridge to the underlying atomistic potentials.

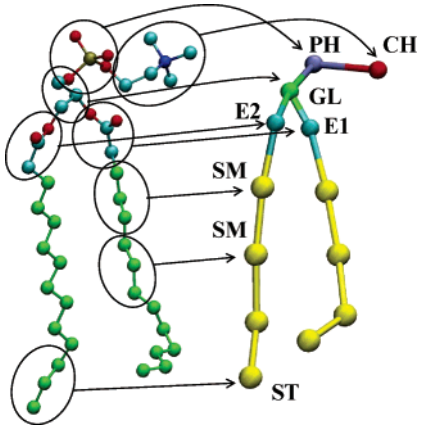

Figure 1. Atomistic (left) and coarse-grain (right) representation of a DMPC molecule. Single coarse-grain (CG) interaction sites were associated with centers-of-mass of the choline (CH), the phosphate (PH), the glycerol backbone (GL: $CH_2-CH-CH_2$), the ester groups (E1 and E2: O_2CCH_2), the triplets of carbon atoms of alkane chains [SM: $(CH_2)_3$], and the tail of alkane chains [ST: $(CH_2)_2-CH_3$].

In the present work, the new MS−CG method was employed to obtain a CG model of a dimyristoylphosphatidylcholine (DMPC) lipid bilayer. The force fields describing interactions between all CG sites in the system (water−water, water−DMPC, and DMPC−DMPC) were derived using atomistic trajectory and force data from a single MD simulation of the bilayer. The atomistic MD simulation was carried out for the bilayer system consisting of 64 DMPC molecules fully hydrated by 1312 water molecules under constant NPT conditions. The DMPC molecules were modeled using a united atom force field.[16] For water, the rigid TIP3P model[17] was used. The temperature of the system was kept constant at $T = 308$ K using the Nosé-Hoover thermostat with a relaxation time of 0.2 ps. The electrostatic interactions were calculated via the particle mesh Ewald summation and the van der Waals interactions were cut off at 1 nm. The system was integrated with a time step of 2 fs, and the initial structure of the DMPC bilayer[18] was equilibrated for 6 ns. The equilibrium volume of the supercell was 107.83 nm^3 with an area per headgroup of 0.61 nm^2.

The water molecules were mapped into a single CG interaction site associated with their geometrical center. With such a choice the FM force field produced water structure which compares better to atomistic simulation than that of a simulation using the FM field obtained with the CG site positioned on the CM of the water molecule. This effect is related to the fact mentioned earlier that the mass is distributed unevenly over water atomic sites. The DMPC molecules were coarse-grained as depicted in Figure 1 in a manner similar to the models of

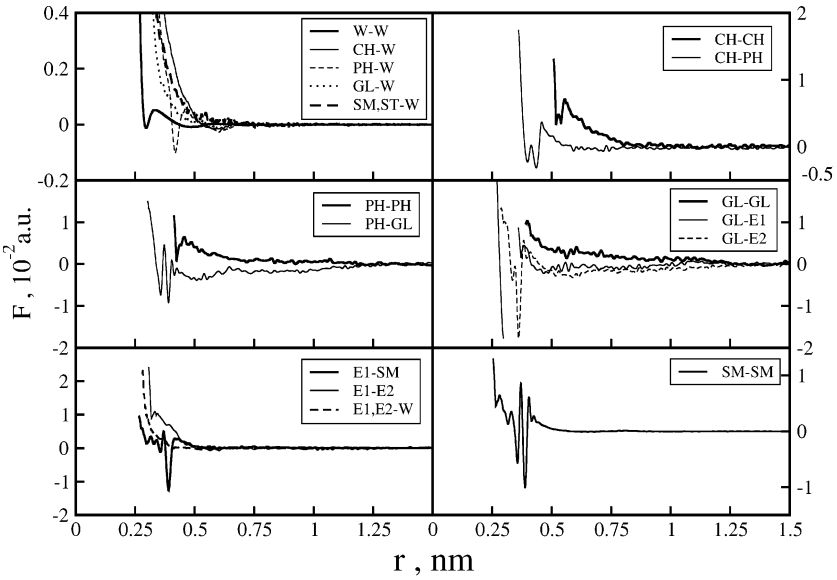

Figure 2. Effective pairwise forces between selected coarse-grained (CG) interaction sites as a function of intersite separation as calculated by the force-matching method. The CG sites are shown in Figure 1 and "W" stands for water.

Shelley[5] and Marrink.[8] However, the CG mapping is not fully equivalent as they used the positions of atoms identified as those closest to the geometrical center of the atomic groups.

For the purpose of force matching, the atomistic MD simulation was carried out for 400 ps in the constant NVT ensemble using the equilibrium volume obtained from the constant NPT simulation. The trajectory and force data were sampled at an interval of 0.1 ps, so the number of configurations was 4000. Then, these trajectories and forces were collapsed into trajectories and forces of the CG sites, and the resulting CG trajectory and force data were used as an input into the FM algorithm. Despite the fact that some CG sites are charged (e.g., CH and PH in Figure 1), it is possible to omit the Coulomb term in eq 1). This is because the FM including the Coulomb term yielded effective charges on the CG sites much smaller than those inferred from the atomistic charges. This is a result of screening effects from the environment (water as well as polar and charged DMPC headgroups) which are captured by the MS−CG force field. Therefore, the FM carried out without the Coulomb term resulted in a short-ranged term which seemed to effectively account for the missing Coulomb term, when using a fairly large cutoff for the short-ranged interaction. The reader is referred to the Supporting Information for more detail on the implementation of the FM procedure.

Selected CG force profiles are shown in Figure 2. To carry out a MS−CG MD simulation the force fields identified as nonbonded and their corresponding potentials, which were obtained by integration of the forces and then shifting the potential to zero at the cutoff radius, were tabulated on a fine mesh and the DL_POLY computer program[19] was used. An input in DL_POLY format with tabulated forces and potentials (TABLE file) is available upon request. The unevennesses in the force profiles for the DMPC-DMPC interactions (seen in Figure 2) are caused by limited statistics and were not a source of difficulty in the MS−CG MD simulation. This is because statistical noise is smoothed out in the spline fit.

For simplicity, the intra-lipid CG sites were linked by harmonic bonds with their force constants determined using a least-squares linear fit of the respective FM force profiles. This was done for small separations at which the forces are more likely to originate from the intramolecular degrees of freedom. Because of the uncertainty in the partitioning of the FM forces into bonded and nonbonded parts for the interactions which involve the SM sites, the SM-E1/E2, SM-SM, SM-ST bonded forces were evaluated separately from vacuum lipid simulations as explained in the Supporting Information. As seen in Figure 2, for some CG site pairs (e.g., PH-GL) the FM bonded forces are noticeably nonlinear. The harmonic approximation was chosen to better reproduce the slope of the FM profile at a close distance when the bonded force is zero (i.e., to have a better harmonic frequency). However, the harmonic approximation to such forces may not be the best choice and may contribute to the small disagreement between properties from the MS−CG and atomistic MD simulations. This aspect of the MS−CG method will be the topic of future research.

The intralipid sites CH, PH, GL, E1, and E2 shown in Figure 1 interacted only through harmonic bonds in the MS−CG model. The proper stiffness and overall length of the alkane chains were therefore maintained by connecting the nearest-neighbor E1, E2, SM, and ST sites by harmonic bonds as described in the Supporting Information. The nonbonded sites between the lipid molecules interacted via nonbonded FM forces that helped to maintain their proper molecular geometry. The highest bond frequency in CG model of the lipids is about 400 cm^{-1}.

The system in the MS−CG simulation was of the same size as that in the atomistic MD simulation. The MS−CG MD was carried out using a constant NVT ensemble with an equilibrium system volume taken from the atomistic MD simulation. The use of the constant NVT conditions were dictated by the fact that the bare MS−CG water force field for the TIP3P model was unable to reproduce the correct water density at constant NPT conditions. However, the structural properties of MS−CG TIP3P water under constant NVT conditions were in satisfactory agreement with the atomistic MD simulation. It is possible to perform additional optimization of the water MS−CG force field to give a correct pure water density without changing the structural properties,[11] in which case the NPT ensemble can be used to simulate the lipid bilayer.

The MS−CG method is able to reproduce the lipid bilayer structural properties accurately as is shown in Figure 3, where the CG site−site distribution functions are compared with the corresponding exact atomistic MD radial distributions functions (RDFs) of the respective groups (see Figure 1). The observed

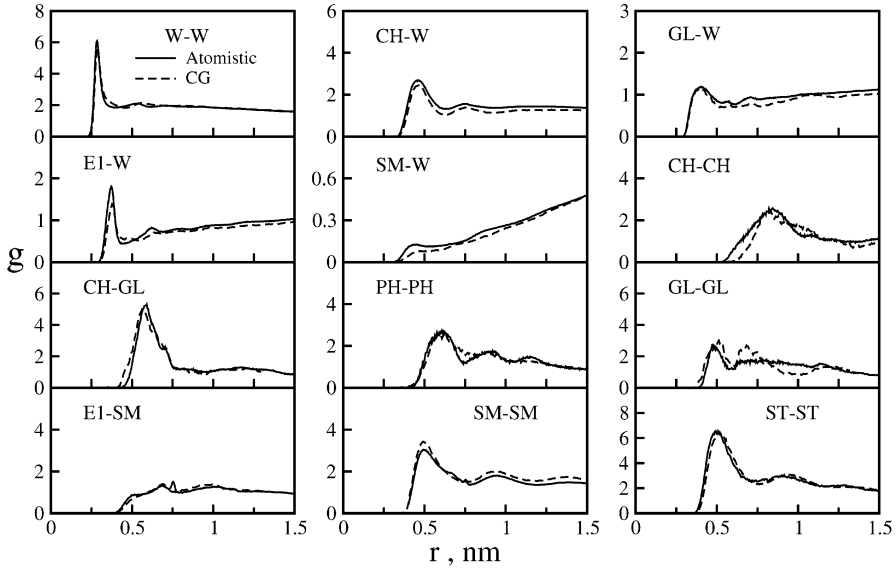

Figure 3. Selected site−site RDFs from atomistic MD simulation (solid lines) compared to those from the MS−CG MD simulation (dashed lines). The peaks from the bonded intramolecular sites are not shown.

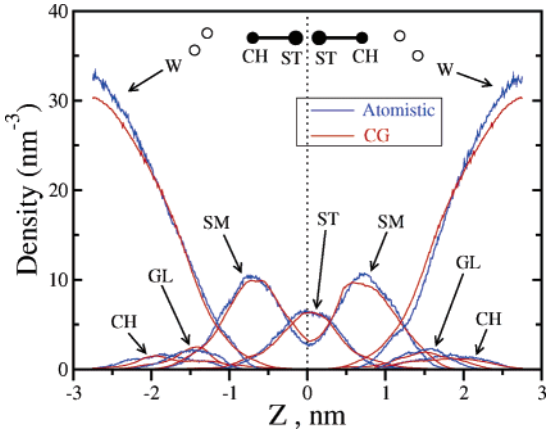

Figure 4. Comparison of the atomistic (blue) and MS−CG (red) density profiles perpendicular to the bilayer (PH, E1, and E2 profiles are not shown).

agreement is quite good despite the fact that some significant approximations were made to the intra-lipid force field. The solvation numbers of the lipid groups in atomistic and MS−CG simulations also agree well. For example, for the CH site the coordination number in the first solvation shell is 15.9 in the atomistic MD and 15.6 in the MS−CG simulation. The analogous numbers for the E1 group are 2.7 and 2.1.

The gain in a speed of the MS−CG simulation for the same underlying atomistic system size and MD time step was about 50 times. However, this speed up is expected to be much greater for larger systems since the MS−CG simulation only contains short-ranged forces which are highly efficient to calculate. Furthermore, it is expected that a significantly larger time step can be used in the MS−CG simulations. Both aspects of the method will be systematically explored in future publications.

Figure 4 compares the *z*-dependent (bilayer normal) densities of the water (W) and the CH, PH, GL, SM, and ST groups from the same simulations as in Figure 3. The MS−CG and atomistic MD water density profiles agree rather well; however, the CG water penetrates the bilayer region slightly more than in the atomistic MD simulation. It seems likely that a more accurate MS−CG model of water will improve this feature, which will also be explored in the future.

The structural properties of the bilayer are also sensitive to the quality of the fit of the W−W and, especially, the W-DMPC interactions. For example, an insufficient cutoff to these interactions in the FM procedure decreased the ability of the system to maintain a proper bilayer structure, i.e., the bilayer was observed to dissolve slowly (on a scale of several nanoseconds).

Although it is tempting to ascribe dynamical properties to the MS−CG simulation results, such an effort would be seriously misguided. In fact, all CG models are in essence approaches designed to represent the effective free energy surface (PMF) for the coarse-grained groups. As such, running classical MD on such CG "potentials" cannot yield any real dynamical information which is, for example, consistent with the First and Second Fluctuation−Dissipation Theorems. The CG classical dynamics merely explores the PMF in a (hopefully) more efficient fashion, but the resulting dynamics cannot reflect the actual dynamical behavior of the CG groups as would be seen in an exact atomistic MD simulation. Additional terms are needed in the dynamical equations in order to accomplish such a goal (e.g., generalized Langevin-like friction and random force terms). Such equations are currently under development for the MS−CG model.

In conclusion, in this communication we have presented a new multiscale method for obtaining effective coarse-grained force fields from atomistic MD force and trajectory data. The resulting MS−CG model is seen to reproduce the structural properties of a lipid bilayer from the atomistic simulations quite accurately. Our new approach provides a systematic way to coarse-grain underlying atomistic force fields, and it utilizes only those atomistic interactions as an input. Our MS−CG approach is general and computationally relatively inexpensive. Applications to complex biomolecular systems are currently underway.

Acknowledgment. The computations in this project were supported in part by the United States National Science Foundation (NSF) cooperative agreement No. ACI-9619020 for the computing resources provided by the National Partnership for an Advanced Computational Infrastructure at the San Diego Supercomputer Center. An allocation of computer time from

the Center for High Performance Computing at the University of Utah is gratefully acknowledged.

Supporting Information Available: Details of force-matching procedure. This material is available free of charge via the Internet at http://pubs.acs.org.

References and Notes

(1) Smit, B.; Esselink, K.; Hilbers, P. A. J.; van Os, N. M.; Rupert, L. A. M.; Szleifer, I. *Langmuir* **1993**, *9*, 9.
(2) Palmer, B. J.; Liu, J. *Langmuir* **1996**, *12*, 746.
(3) Goetz, R.; Lipowsky, R. J. *J. Chem. Phys.* **1998**, *108*, 7397.
(4) Shelley, J. C.; Shelley, M. Y. *Curr. Opin. Colloid Interface Sci.* **2000**, *5*, 101.
(5) Shelley, J. C.; Shelley, M. Y.; Reeder, R. C.; Bandyopadhyay, S.; Klein, M. L. *J. Phys. Chem. B* **2001**, *105*, 4464.
(6) Marrink, S.; Mark, A. *J. Am. Chem. Soc.* **2003**, *125*, 15233.
(7) Stevens, M. J.; Hoh, J. H.; Woolf, T. B. *Phys. Rev. Lett.* **2003**, *91*, 188102.
(8) Marrink, S. J.; de Vries, A. H.; Mark, A. E. *J. Phys. Chem. B* **2004**, *108*, 750.
(9) Meyer, H.; Biermann, O.; Faller, R.; Reith, D.; Müller-Plathe, F. *J. Chem. Phys.* **2000**, *113*, 6264.
(10) Murtola, T.; Falck, E.; Patra, M.; Karttunen, M.; Vattulainena, I. *J. Chem. Phys.* **2004**, *121*, 9156.
(11) Izvekov, S.; Parrinello, M.; Burnham, C. J.; Voth, G. A. *J. Chem. Phys.* **2004**, *120*, 10896.
(12) Hone, T.; Izvekov, S.; Voth, G. A. *J. Chem. Phys.* **2005**.
(13) Ercolessi, F.; Adams, J. B. *Europhys. Lett.* **1994**, *26*, 583.
(14) Lawson, C. L.; Hanson, R. J. *Solving Least Squares Problems*; Prentice Hall: Englewood Cliffs, New Jersey, 1974.
(15) De Boor, C. *Practical Guide to Splines*; Springer-Verlag: New York, 1978.
(16) Smondyrev, A. M.; Berkowitz, M. L. *J. Comput. Chem.* **1999**, *20*, 531.
(17) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926.
(18) Smondyrev, A. M.; Berkowitz, M. L. *J. Chem. Phys.* **1999**, *111*, 9864.
(19) Forester, T. R.; Smith, W. *DL_POLY user manual*; CCLRC, Daresbury Laboratory: Daresbury, Warrington, U.K., 1995.